Complexity

Data-Enabled Intelligence in Complex Industrial Systems

Lead Guest Editor: Long Wang
Guest Editors: Chao Huang and Jenq-Haur Wang

Data-Enabled Intelligence in Complex Industrial Systems

Complexity

Data-Enabled Intelligence in Complex Industrial Systems

Lead Guest Editor: Long Wang
Guest Editors: Chao Huang and Jenq-Haur Wang

Contents

Hindawi
Complexity
Volume 2022, Article ID 4637939, 17 pages
https://doi.org/10.1155/2022/4637939

Research Article

Real-Time Explainable Multiclass Object Detection for Quality Assessment in 2-Dimensional Radiography Images

Sadra Naddaf-Sh,[1] M-Mahdi Naddaf-Sh,[1] Hassan Zargarzadeh,[1] Maxim Dalton,[2] Soodabeh Ramezani,[2] Gabriel Elpers,[2] Vinay S. Baburao,[2] and Amir R. Kashani[2]

[1]*Phillip M. Drayer Electrical Engineering Department, Lamar University, Beaumont, TX, USA*
[2]*Artificial Intelligence Lab, Stanley Oil & Gas, Stanley Black & Decker, New Britain, CT, USA*

Correspondence should be addressed to Hassan Zargarzadeh; hzargarzadeh@lamar.edu

Received 29 May 2021; Revised 22 August 2021; Accepted 30 August 2021; Published 8 August 2022

Academic Editor: Long Wang

Quality inspection and defect detection play a critical role in infrastructure safety and integrity specially when it comes to aging infrastructure mostly owned by governments around the world. One of the prevalent inspections performed in the industry is nondestructive testing (NDT) using radiography imaging. Growing demand, shortage of experts, diversity of required skills, and specific regional standards with a time-limited requirement of inspection results make automated inspection an urgent need. Therefore, utilizing artificial intelligence- (AI-) based tools as an assistive technology has become a trend for industrial applications, which automates repeated tasks and provides increased confidence before and during the inspection operation. Most of the works in quality assessment are focused on the classification of few categories of defects and mostly performed on public or noncomprehensive research datasets. In this work, a scalable, efficient, and real-time deep learning family of models for detection and classification of 10 various categories of weld characteristics on a real-world industrial dataset is presented. The models are evaluated and compared against each other, various critical hyperparameters and components are optimized, and local explainability of models is discussed. Additionally, AutoAugment for object detection and various techniques are utilized and investigated. The best performance for object detection and classification for 10 class models is reached by mean average precision of 72.4% and top-1 accuracy of 90.2%, respectively. Also, the fastest object detection model is able to evaluate a full 15360×1024 pixels weld image in 0.39 seconds. Finally, the proposed models are deployable on edge-devices to perform as assistant to NDT experts or auditing professionals.

1. Introduction

Inspection and assessment of welded joints are critical in many industries such as marine, aerospace, and chemical, and specifically in oil and gas industries [1]. Welded joints are among vulnerable parts of any industrial infrastructure including pipelines. Hence, preliminary weld inspection during the construction has a crucial role in its longevity as a small discontinuity can grow into an utter failure over time [2, 3]. Moreover, pipeline failures can damage life in large-scale and is a threat to the environment [4]. Furthermore, it is very costly to maintain continuous inspection to track the growth of initial imperfections over time or efforts to restore the surrounding environment or the pipeline when defects are larger than a certain threshold [5, 6]. Thus, weld inspection is the most economical preventive approach specifically at early stages of its construction. Among different nondestructive testing (NDT) technologies at the point of constructions, radiographic testing (RT), ultrasonic testing (UT), and magnetic testing (MT) are of great importance. Currently RT, in which X-ray imaging of the welded part is done, is preferred due to the universal training and accuracy of its technology [7]. Nonetheless, analysis of X-ray images is time-consuming and tedious, and at the end different experts might have different opinions and hence auditing is essential [8]. Thus, automation of these systems is of interest in the industry to certify reliability and safety of the product in various stages of construction, approval, audit, and risk assessment.

In recent decades, many research have been conducted on automation of tasks employing robots [9–12], including robotic platforms automation of welding operation to accelerate the process and reduce human error. As an instance, Figure 1 shows a robotic digital X-ray photographer by Stanley Oil and Gas. The robot autonomously conducts the X-ray imaging that significantly minimizes the human intervention to prevent the operators from exposure [13]. After a robotic imaging process is done, human experts use images generated to inspect the welds. However, recent rapid improvement in machine learning, computer vision, and pattern recognition has opened new roads to provide novel solutions in order to address the challenges regarding ultimate defect diagnosis and complete tractability of discontinuities over the pipeline's life cycle [2, 3, 8, 14, 15]. In the following, a review on related research performed in weld and defect diagnosis is provided.

Previous research works with focus on defect analysis are mainly divided into two smaller subgroups. Before the prevalence of deep learning and convolutional neural networks (CNNs) approaches in the early 2010s, procedures focused on traditional image processing methods for image preprocessing and classification utilizing classical machine learning methods (e.g., support vector machine (SVM)) and training artificial neural networks (ANNs) based on hand-crafted features extracted from image patches (cropped rectangular pieces of a larger image). Among these works, they mainly focused on classifying defect and nondefect images and assigning a single label to an image patch with or without segmentation of defect area. Mery and Berti [16] used texture features to train ANNs and the best result reached 8% false alarm. In [17], gray level co-occurrence matrix (GLCM) texture features were used for multiclass ANNs with 86.1% accuracy and optimized to reach 87.3% by applying Levenberg–Marquardt optimizing function in [18]. A similar approach to classify defects with a combination of statistical and geometric features and utilizing top-hat filtering, thresholding, and morphological smoothing as preprocessing presented in [19] resulted in 91% accuracy in detecting defects and nondefects and 96% in classifying of a hundred of test images containing low contrast images. In [20], Wiener filter is considered the best enhancement as it leads to lower rooted mean square error (RMSE) in comparison with median filtering and contrast enhancement, and also defective segments are obtained from the segmented image using an automatic threshold. Finally, for feature extraction, the lexicographically-ordered one-dimensional signal of the image is generated, and mel-frequency cepstral coefficients (MFCCs) and polynomial coefficients are extracted from the power density spectra (PDSs) of the image and passed into ANN, which reduced false positive rate to 7%. Lim al. [21] employed a multilayer perceptron (MLP) network trained on a simulated dataset of weld radiographic images for classification of the patches.

Zapata et al. in [22] used an adaptive network-based fuzzy inference system (ANFIS) and ANN, in which geometrical and texture features were selected with respect to minimizing computational complexity and reached 82.6% accuracy. Valavanis and Kosmopoulos [23] applied certain classifiers for distinguishing between six types of defects annotated based on British Standards or labeling as nondefect. Preprocessing steps of their research include utilizing local threshold, graph-based segmentation, and then geometric and texture features are used as input for classifiers like ANN, K-nearest neighbor (KNN), and SVM. In [24], a comprehensive review of similar methods is provided. It can be concluded that classical approaches require major preprocessing steps before feature extraction and preprocessing enhancements have direct impact on final accuracy.

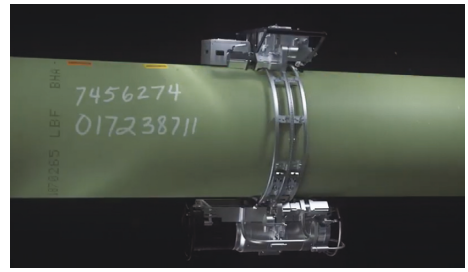


Figure 1: Digital X-ray detector and source on a robotic platform, Stanley Oil and Gas.

On the other hand, a few researches focused on image segmentation to provide a general understanding of defect localization. Carrasco and Mery [25] presented a method for segmenting defects. The method consists of a few steps: median filtering, bottom-hat filter, binary thresholding, and watershed transform. The results suggested an area under curve (AUC) of 93.58% for ten images. In [26], sliding window approach is used for weld object detection based on a large set of features. In [27], Ben Gharsallah and Ben Braiek proposed a method to address nonrobustness of defect segmentation caused by uneven illumination, based on level set active contour guided with an off-center saliency map, in which an energy function gets minimized to achieve segmentation. Despite faster convergence and higher accuracy than local image filtering and contrast enhancement, the method requires further investigation to minimize human intervention in finding region of interest (ROI). In [28], defect segmentation problem is addressed using Gabor filtering and canny edge detector. As more recent research, which is also evaluated on aerospace weld dataset, a novel pixelwise segmentation defect detection system is presented in [8]. Dong et al. [8] described a system to detect weld defects by using random forest instead of Softmax as the classifier of a U-net [29]. The approach is pixelwise labeling of highly similar circular defects, which are prevalent in aerospace industries.

Since the prevalence of deep convolutional neural networks (DCNNs), many works have focused on using these models for feature extraction/selection instead of traditional hand-crafted feature extraction and nonrobust methods. Primarily two general tasks are performed using DCNNs (i.e., classification and object detection task). Furthermore, weld defect dataset has class-imbalance issue, since the number of weld defects might not distribute equally among different classes. Hoe et al. [30] focused on extending three types of datasets using auto encoders to address the

imbalance problem. Next, a few models, including DCNNs and other models based on extracted features are trained to classify four different types of defects and reached accuracy of 97.2%. Ajmi et al. [31] explored two-class (porosity and lack of penetration) classification of weld defects. Data augmentation through horizontal mirroring, translations, and RGB channels modification are applied to boost model performance, and 85.2% accuracy is reported with transfer learning utilizing AlexNet [32] and addition of a few dropout layers as well as modified final layer on GDXray [33]. In [34], a real-time and two-stage method based on images from a 3D laser scanner is proposed. The method performs four-class classification of narrow lap welds. Also, a comparison on classical and deep classification methods is performed with average accuracy of 80% for classical approaches while for deep methods of VGG-16 [35] and ResNet50 [36], 97.1% and 97.8% accuracy are reported, respectively. Wang et al. [37] presented a tutorial for weld defect detection based on DCNNs with implementation provided in PyTorch [38]. The paper provides a step-by-step approach for the data collection, preprocessing, and model designing, training, and testing.

Further investigation is performed for accurate localization of weld characteristics using deep methods. Hou et al. [14] designed a deep learning-based system for weld quality assessment. They used sparse autoencoder (SAE) to extract and use intrinsic features for classifying 32×32 pixels weld patches and finally using a sliding window to classify image pixels as defect or nondefect. The process reaches an accuracy of 91% on GDXray [33], even though the work is a binary class defect classification and the process is time-consuming because of the nature of the sliding window approach and size of full weld images. In [39], extensive experiments with 24 various computer vision-based weld object detection methods (including deep learning methods based on sliding window) are performed and reported. In [40], two-stage detectors (i.e., Faster RCNN [41]) are used whose task is object detection of weld defects in shipbuilding which accounts for 60% of the building process, where radiography testing is used to inspect welded joints. The proposed object detector is trained to detect two general types of porosity and lack of fusion/slag defects. Moreover, the best result is acquired by data augmentation, which reached 53.2 mean average precision (mAP) on Faster RCNN [41] with ResNet50 [36] backbone.

Gau et al. [42] developed a contrast enhancement conditional generative adversarial network (GAN) to address the contrast and class-imbalance issue. There are two separate target networks in their work. The first network accepts a 71×71 pixels patch from weld seam to classify the patch as defect/nondefect. For determining defect type, defective patches are passed into a second classification network. At the end, the sliding window approach is used for localizing defects. Thus, with respect to the two-stage design of the system and the sliding window, the entire system will not perform in real-time for high-resolution images. In [43], a defect localization method based on U-net and augmentation using conditional GAN (cGAN) [44] is presented, and the method is evaluated on GDXray dataset [33]. Although the method shows AUC of 88.4% for defect segmentation, lack of defect classification is discernible. Gantala and Balasubramaniam [45] presented an automatic defect recognition model trained on total focusing method (TFM) imaging dataset and finite element simulated dataset with addition of noise and further expansion of dataset utilizing deep convolutional gaN (DCGAN). Their two-class defect detection model was evaluated with yolov4 [46] and reached 85 average precision (AP) on the noisy dataset.

Although the above research papers are mostly related to employing deep CNN methods to automate the preliminary inspection in construction and welding, studies using deep CNN methods for NDT and defect diagnosis are not limited to radiography images and weld construction. Yan et al. [47] developed deep models for enhanced feature extraction and ultrasonic pattern recognition for inspection gas pipelines. The method uses contact-less dual-mode bulk wave electromagnetic acoustic transduce (EMAT) and interpretations of A-scan signals to detect defects. It leverages continuous wavelet transform (CWT) to extract frequency-time domain features, then a deep CNN model is applied to perform high-end feature extraction, and finally, a pretrained SVM is used for defect/nondefect classification of signals. The method feature extraction ability is verified by comparing to other methods, including discrete wavelet transform (DWT) and statistical features, all of which are outperformed by the CNN model, which achieves 93.75% accuracy on a dataset of pipe with artificially manufactured defects. The work is performed for defect/nondefect classification, and the possibility of defect type classification is to be investigated.

In addition to ultrasonic pattern recognition, deep CNNs are also utilized for thermography crack detection. In [48], Hu et al. explored supervised thermography video sequence metal crack detection and localization. The work uses eddy current pulsed thermography (ECPT), a multi-physics coupling method, to detect turbulence in conductive materials by analyzing thermal patterns. Initially, principal component analysis (PCA) is used to extract thermal sequence components from original data. Then, Faster RCNN [41] is used to perform object detection on images accurately. Finally, the method is compared to traditional detection methods, and it demonstrates 0.97 probability of detection, which outperforms the accurate prior method by 26%. Proposed methods are validated experimentally and have shown significant improvement in their own type of NDT and data acquisition, demonstrating the advantages of using CNN for feature extraction in NDT. While UT and thermography methods (e.g., ECPT) are commonly used for in-line inspection and maintenance purposes and not for weld construction inspections, these methods have their limitations, such as low sensitivity to small defects or internal crack detection [13].

Studies mentioned above are all experimentally evaluated on either (1) a set of images from a private dataset (i.e., usually created for experimental purposes) or (2) GDXray [33] or similar public and noncomprehensive sets. As shown in Figures 2 and 3, there are noticeable differences in images from welded joints at Stanley, and the GDXray dataset. First, GDXray has a limited number of samples. Second, class

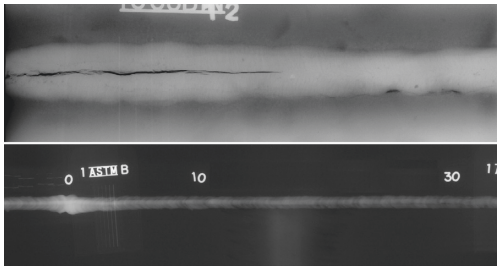


Figure 2: Tow samples of images in the GDXray database (top) and SBD dataset (bottom). The bottom image is cropped to be able to compare with each other. Defects are more visible in the GDXray database than SBD dataset.

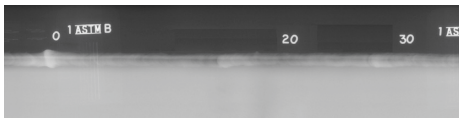


Figure 3: An image sample from SBD database in that two surfaces with different thicknesses are welded.

diversity is limited and also annotations and weld characteristics are based on a different standard [33]. Third, visibility of defects is limited compared to defects at Stanley. Also, in some cases, a single patch contains more than one type of defect, which does not permit the experts to designate a single label for the entire image patch, all of which make classification only or defect/nondefect localization incompatible with real-world industrial requirements and standards. In other words, the detection of non-hand-picked and diverse real-world samples is more of a challenge. On the other hand, since the systems will work as assistant to NDT experts, and there are limitations in hardware for deploying as well as time-constraint processing requirements, scalability is required for efficient and optimized utilization. Considering mentioned reasons, these methods either fail to reach required specifications or do not meet required performance based on industry measures.

This paper aims to address the accuracy and inference time trade-offs by presenting an efficient and scalable set of deep models. Moreover, instead of assigning a single label for each patch, accurate location and label for each discontinuity will be determined. The contributions in this work are as follows: (1) describing an efficient and scalable system for object detection or classification of weld characteristics on long, high-resolution radiography weld images, which is deployable as a real-time assistant for NDT experts, (2) demonstration and analysis of the transferring augmentation strategies during training which can improve the performance of the system on detection of rare small discontinuity which are easier to miss during manual inspection and harder to detect with deep learning methods, (3) analyzing and experimenting with different components of the deep model, such as activation functions, and feature extraction backbones, and (4) comparative analysis on the presented models with base-line models.

The rest of this article is organized as follows. In Sections 2 and 3, an overview of dataset preparation and proposed methods is provided, respectively, as well as description of system architecture. In Section 4, the methods are tested, various models are described, and the augmentation approach and results are evaluated. Finally, conclusions are proposed in Section 5.

2. Dataset

The dataset contains thousands of X-ray images taken with the purpose of NDT of weld construction in preliminary stages. There is little to no material variation in weld construction, which helps developing a model focusing on accuracy and robustness. The majority of the structures are plain carbon steel. The diameter of the pipes ranges from 24 to 56 inches. However, pipes with either 36 inches or 42 inches are mostly common. Moreover, the pipes wall thickness is at least 0.5 inches with the grade of X65 or greater. Finally, all pipes are consistent with API 5L [49] in terms of types, dimensions, material, and grade.

Welded-joint images have various resolutions depending on the exterior diameter of the structure. In this dataset, the resolution of the images is roughly 15360×1024 pixels, with the occurrence of weld discontinuities. As the welded area only covers one-fifth of each weld image's center area, images are cropped into 224×224 patches with 20% overlap. This overlap benefits in two ways. First, it assists in retaining defects lying in between two patches in one patch. Second, as smaller defects shift in two consecutive images, it can be interpreted as data augmentation. Next, experts annotated the images based on API 1104 [50] standards. Most of the defect-free patches are removed from the dataset to prevent overwhelming the network with nondefect images. Finally, Figure 4 shows samples of the dataset, and Table 1 shows the distribution of images for each set. As the dataset reveals, about 75% is used as train set (i.e., 17872 images), and 10% and 15% are used as dev/validation set and test set, respectively. Note that the dataset is collected from welding of various structures and different welding devices. Thus, results obtaining from this dataset can demonstrate the generalizability and robustness of proposed solutions for extensive use as assistant to NDT experts. Figure 5 summarizes preprocessing steps on the dataset. The steps are described in detail in Section 3.1.

3. Method

Addressing robustness, accuracy, and time performance are required for employing a deep convolutional model in production for the task of weld defect object detection. Over recent years, scaling up image resolution, depth and width of the network, and using a larger backbone are widely used to boost the performance of the models [51–54]. However, this costs, in a larger model, higher computation and inference time [51, 52] as well as longer training time. Thus, a robust and efficient design is required to address the accuracy versus time performance. In order to address the trade-off between accuracy and time and achieve efficiency in models, a family of one-stage and scalable models called EfficientDet [52] are exploited. Employing a single compound coefficient,

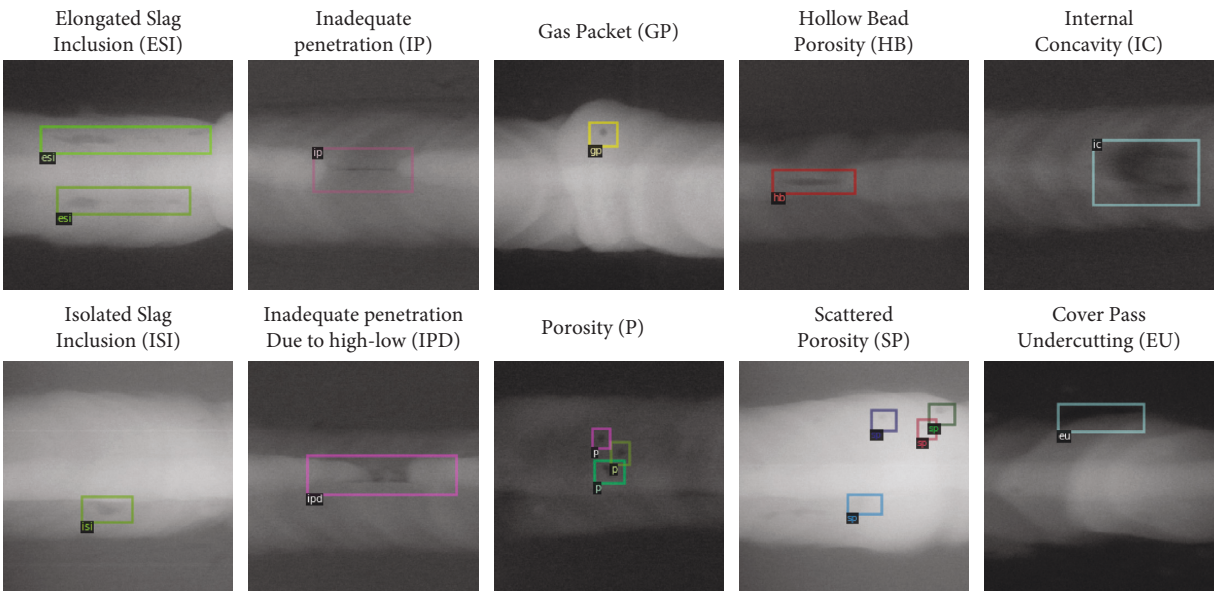


Figure 4: Samples of classes in dataset.

Table 1: Distribution of images and labels.

	Total	Categories									
		ESI	ISI	IP	IPD	GP	P	HB	SP	IC	EU
Train	17872	4424	2948	908	702	1840	598	4062	498	597	1295
Validation	2203	569	338	118	97	280	71	435	51	87	157
Test	3394	891	490	190	139	408	116	707	83	130	240
Total	23469	5884	3776	1216	938	2528	785	5204	632	814	1692

one can scale the architecture to address the trade-off between model size and accuracy of the model, resulting in a model deployable on various end-devices ranging from mobile devices to high-performance GPU clusters.

A two-stage object detection model generally starts with a search on regions of interest (ROI) using the selective search or, in more recent designs, applying region proposal networks (RPNs), and then by passing image to the second stage for feature extraction, classification of the boxes, and refinement of the bounding box are performed [41, 55]. Although the tow-stage methods might lead to higher accuracy, the inference time because of the first stage burden is significant in sight of the additional step (RPN). In contrast, one-stage detectors apply a feature extractor called backbone and then fuse multilevel extracted features. In the end, class/box networks help to extract class labels and regression of bounding boxes. Since the image passes only once through the network, the one-stage detector performs significantly faster than other methods [54]. By utilizing pretrained backbones, the power from classification tasks transfers to these object detectors as employed in [56]. In this section, preprocessing steps, EfficientDet architecture design, and augmentation strategies for object detection as well as system architecture to achieve an accurate model with low time latency are discussed, respectively.

3.1. System Architecture. Figure 5 depicts the required preprocessing steps to generate the dataset, which start with downloading image patches and quality validation. Although the images on the cloud storage are prevalidated for quality, it can be done through a wire IQI tag, which is discernible on the image in Figure 6. As this step is optional and can be done upon uploading the images to the cloud storage, its time burden is disregarded from total system time performance. As the final two steps, brightness correction and contrast leveling as well as slicing of the original 15360×1024 pixels image with 20% overlap are done.

As Figure 5 training depicts, training starts on a scaled model, which depends on a single coefficient for the determination of depth and width of the network. In addition, AutoAugment during training is performed. Procedures of network design, scaling, and augmentation are elaborated in Sections 3.2–3.4. As the next step, based on the type of the trained network in the model, it predicts either label and accurate location of the defects or assigns a single label for the whole patch with explanations on the decision provided. Finally, Figure 5 visualization indicates stitching as the first step of visualization. Since exact slicing points are saved during slicing, relative predicted defect locations of the whole image are calculable. Finally, the full DICONDE image can be visualized through Stanley web-app or mobile-app or saved as DICONDE metadata.

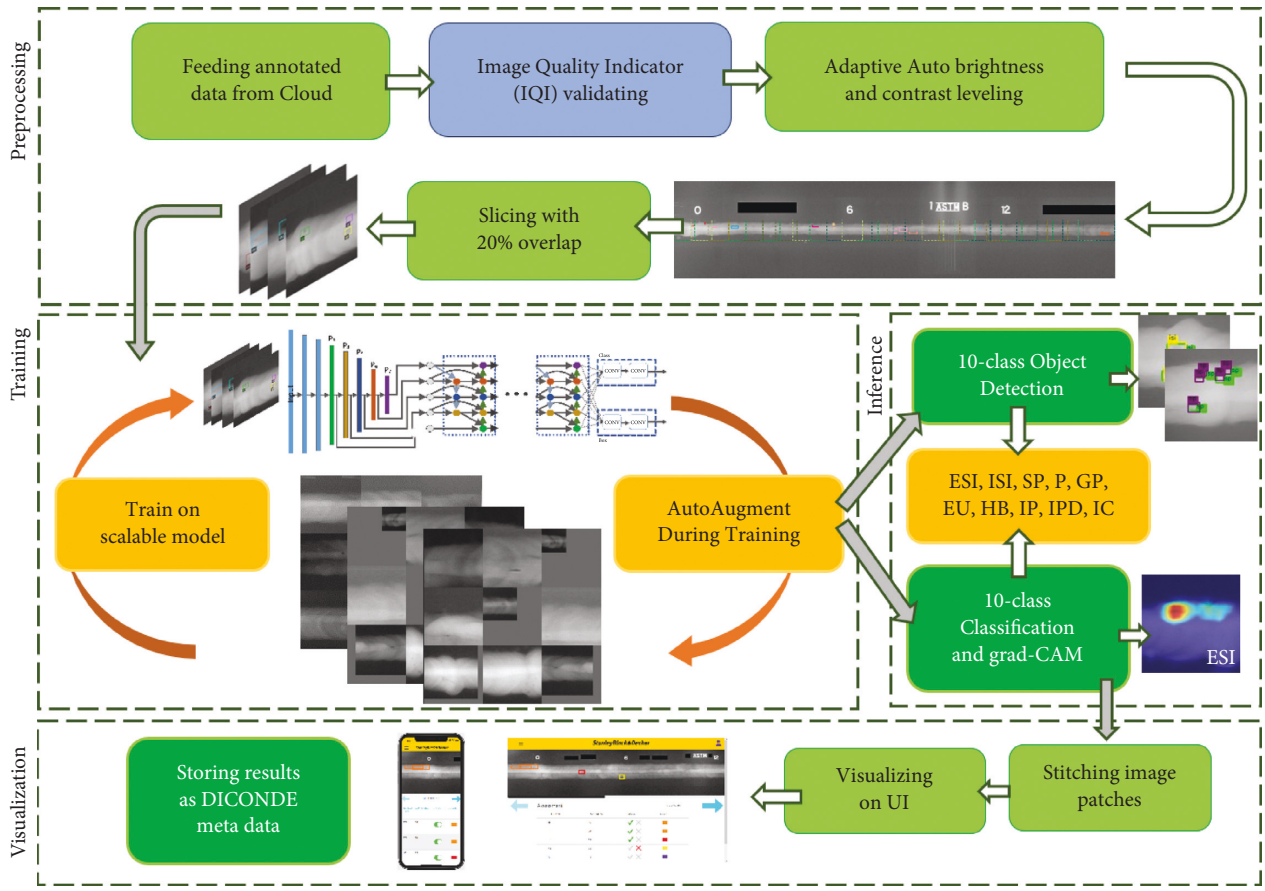


Figure 5: An overview of the system architecture: the box in blue accent is optional and can be done once images are captured before the start of the detection process. The arrows in gray color show transitions to the next stage. In section inference, one of the object detection or classification tasks will be done based on the in-use trained model. In visualization, depending on the need, stitching, visualization, storing, or all of them in once can be done.

3.2. Network. As described in Section 2, the final dataset contains 23469 image patches of size 224×244 pixels. An image patch passes through backbone for feature extraction. In this work, EfficientNet [53] is used as the backbone of object detection models and for feature extraction in classification models. However, for weld quality assessment, different backbone performances are evaluated, and class activation maps are reported. Next, multiscale features from levels P_3 to P_7 pass through a successor of feature pyramid networks (FPNs) [57]. P_i denotes the resolution of the input activation map that is $1/2^i$ of the original input image. In conventional FPN, it is assumed that features from various scales contribute equally to the final detection. A few works have investigated the optimization of feature fusion; e.g., NAS-FPN [51] is an effort to find optimum architecture for cross-scale fusing network through search. However, it takes thousands of GPU hours to find an optimal design and the resulting model is oversized. To address the equal contribution of different scales in fusing features, EfficientDet uses bidirectional FPN (BiFPN). In BiFPN, similar to FPN, a top-down pass is used, and similar to PA-Net [58] bottom-up pass is added. Nonetheless, the bottom-up pass adds a lot of costly additional weights to the network. Thus, nodes with single connections (highest and lowest levels) are removed in view of less contribution in feature fusion to optimize the structure. In addition, a few edges from input to output (similar to skip connections in ResNet [36]) are added, which boost both the training and accuracy processes. Finally, fused features pass through two similar class and box networks used to determine the class label and the bounding box location of detected discontinuities. Similar to backbone and BiFPN, depth of class/box nets gets scaled with a single coefficient.

3.3. Scalability. In this part, the single compound scaling coefficient of the overall architecture is reviewed. EfficientDet family starts from the smallest model D0 and ends with the deepest and largest model D7, where the number stands for the single compound coefficient ϕ, used to scale input image resolution and overall depth and width of the architecture. For backbone, if EfficientNet is used, one of the pretrained networks is applied based on ϕ. Figure 6 shows the architecture, which is similar for all networks. The final input image resolution is determined using the following equation:

$$R_{\text{input}} = 512 + \phi \cdot 128. \quad (1)$$

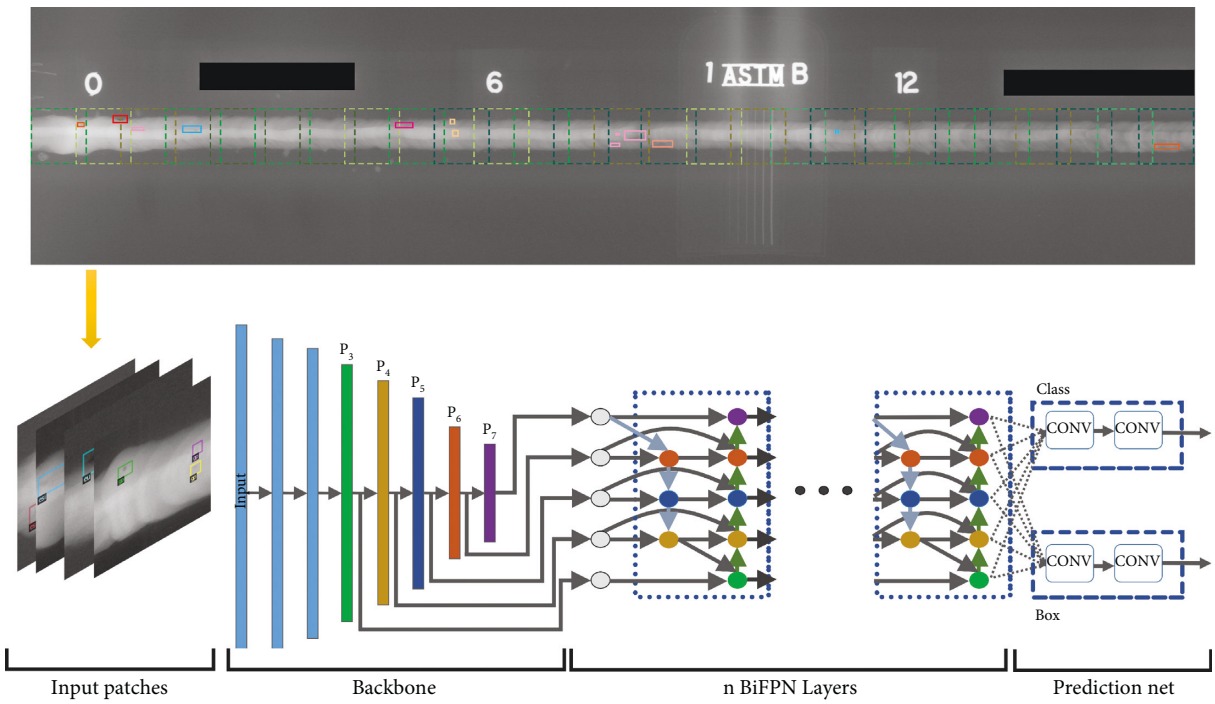


Figure 6: EfficientDet family architecture: images pass through the backbone, and feature scales P3 to P7 get fed into the BiFPN network. Input image resolution is calculated from equation (1). The number of BiFPN repeated blocks extracted using equation (2). Depth of box/class prediction nets is determined using equation (3).

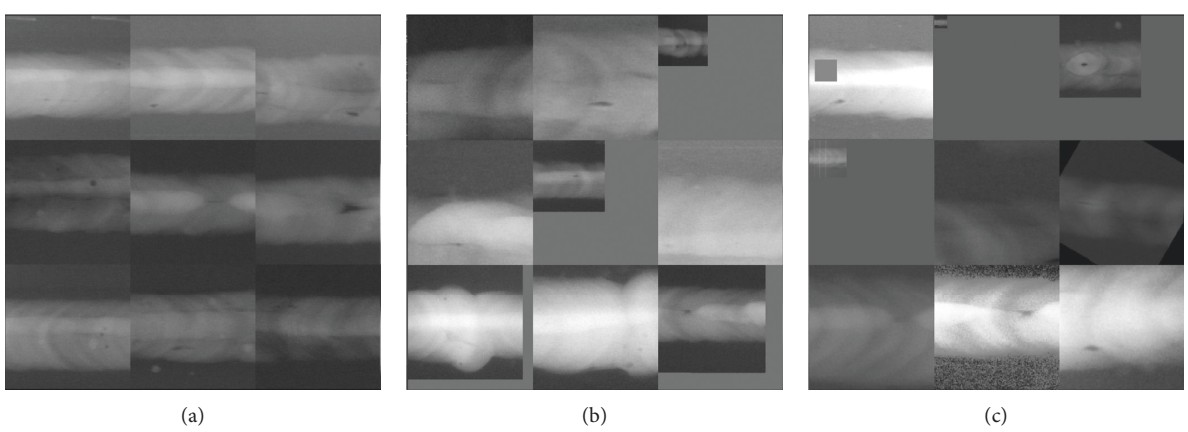


Figure 7: Samples of a training batch with different augmentations: (a) a batch with no augmentation, (b) an augmented batch with a collection of random augmentation named as train-time augmentation, and (c) augmented batch based on policyV3.

Equation (2) shows how the number of channels/layers of the BiFPN is scaled, where each layer is one of the BiFPN repeated blocks starting from 3 for D0 shown in Figure 6. Finally, the number of layers of the class/box is determined through equation (3).

$$W_{\text{bifpn}} = 64 \cdot \left(1.35^{\phi}\right), \quad D_{\text{bifpn}} = 3 + \phi, \tag{2}$$

$$D_{\text{box}} = D_{\text{class}} = 3 + \lfloor \frac{\phi}{3} \rfloor. \tag{3}$$

3.4. Data Augmentation. Many object detection as well as weld quality assessment deep learning approaches employ data augmentation in order to improve both the performance of the network and generalization [31, 40, 43]. The effectiveness of augmentation is shown and evaluated in literature [59]. Nonetheless, there are countless strategies, such as rotation, affine, zoom in/out, flipping, etc., various magnitudes, and also different possible combinations of strategies to be used for augmenting the dataset. One solution is to search through all possible solutions to find the optimal ones. The authors in [60] investigated and searched through the area of 10^{10} different combinations for the

classification task. Similarly, [61] investigated the effectiveness of AutoAugmentation for object detection and extracted a few sets of policies enhancing detection performance the best for the object detection task named as policy V0-3. As searching for optimal augmentation strategies is a time-consuming task, extracted policies are applied and investigated in this work. For this purpose, a base model (D0 with EfficientNet B0 backbone) is trained utilizing each of the policies to find the best policy. Then, best policy is used for training larger models and investigating other effective parameters of the model.

3.5. Evaluating Metrics. Evaluating results is performed through average precision (AP) metrics. Models output a bounding box, a corresponding class label, and confidence for each detection. A detection is considered correct when the area of the ground truth bounding box and the detected box have at least 0.5 intersection over the area of the union of two mentioned boxes, which is called Intersection over Union (IoU). Also, the class labels of both bounding boxes should be the same, which means

$$\mathrm{IoU} = \frac{\mathrm{area}\left(\mathrm{B_{Bp}} \cap \mathrm{B_{gt}}\right)}{\mathrm{area}\left(\mathrm{B_{Bp}} \cup \mathrm{B_{gt}}\right)}. \tag{4}$$

With IoU less than 0.5, the detection is counted as fp. Fn is also the count of nondetected bounding boxes. Therefore, precision and recall are calculated through the following:

$$\mathrm{precision} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}; \mathrm{recall} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}. \tag{5}$$

As recall and precision of a robust object detector do not alter much with varying confidence, it is required to consider multiple confidence thresholds to evaluate the performance of the object detector [62]. Defining all-point interpolation of the area under precision-recall curve obtains accurate results by pruning zig-zag behavior of the curve and utilizing maximum precision (P_{interp} (r) where r is recall level and recall of the point is greater than r_{n+1}) at each recall level, instead of using the precision at that point. The mathematical presentation of this is as follows:

$$\mathrm{AP} = \sum_{n} \left(r_{n+1} - r_n\right) P_{\mathrm{interp}}\left(r_{n+1}\right), \tag{6}$$

where

$$P_{\mathrm{interp}}\left(r_{n+1}\right) = \max_{\tilde{r}:\, \tilde{r} \geq r_{n+1}} P(\tilde{r}). \tag{7}$$

AP has become a standard for comparing model performance in different object detection challenges [63] as well as literature [41, 52, 64, 65].

In Section 4, models are evaluated using mAP (mean AP) (which is equal to mean of AP with IOU threshold ranging from 0.50 to 0.95 and step of 0.05), AP_{50}, AP_{75} (which is equal to ap@iou=0.75), AP_{s} (s stands for small and objects with area $< 32^2$), AP_{m} (m is medium and area of the objects is between 32^2 and 96^2), and AP_{l} (objects with area $> 96^2$).

4. Experiments

In this section, various experiments are designed and performed to investigate a set of scalable models with fast processing time while maintaining high accuracy. In addition to EfficientNet backbones, results are reported utilizing other backbones, namely, MobileNetV3 [66], ResNet50 [36], which is called Resdet50 in detection models, CspResdet50 [67], and Darknet (utilized in Yolov3 [54]). Moreover, standalone object detection models including Yolov3, Yolov4 [46], Yolov5 [68], and RetinaNet [65] are fine-tuned as a basis for comparison. In the following sections, the K-means method is used to extract optimal anchor boxes; analysis and results from applying various AutoAugment policies on models, training setup and hyperparameter tuning, quality assessment with single class labels, effects of using several activation functions, and backbones are elaborated, respectively.

4.1. Anchor Boxes. Similar to [57], EfficientDet uses anchor boxes to detect objects. By default, there are three distinct aspect ratios (0.5, 1.0, and 2.0). K-means clustering is utilized to find the set of optimal aspect ratios for the box prediction network [64]. Moreover, the input image resolution is also considered in optimized aspect ratio calculation. Table 2 demonstrates the effectiveness of using aspect ratios calculated by K-means. The results are reported using AP metrics. New aspect ratios (1.2, 2.14, 3.8) suggest that 99.92% (i.e., equivalent to the percentage of bounding boxes that lie into one of K-means calculated clusters) of the defects are horizontal rectangles, and optimizing helps with a 6.6% increase in AP_{50}.

4.2. Analysis of Augmentation Policies. Since training each model employing all policies V0 to V3 is time-consuming, EfficientDet-D0 is considered the base model for analyzing how transferring augmentation policies affect the detection of characteristics. Table 3 demonstrates the effects of utilizing various policies during training for augmentation, based on AP metrics. In NoAugment, raw images are passed to the network, while in Train-timeAug, two common augmentations for train-time are used. First, images are flipped horizontally with a probability of 50%, and second, images are randomly resized and padded with a random scale between 0.1 and 2.0. Also, bilinear interpolation is used while resizing, and the mean of the dataset is applied for padding when the final image is smaller than 512×512 pixels (as 512×512 pixels image is the target image size for the D0 model). PolicyV0-3 refers to each of 4 policies introduced in [61]. In a similar way, during training of the D0 model using each of these policies, a random set of strategies from the selected policy with a probability of 66% is selected, and the input image is augmented based on it (the probability of not performing any of the strategies is one-third). Moreover, similar augmentation is performed on bounding boxes if any is affected. Based on Table 3, augmentation policies dramatically boost the performance of the network by 3.8 to 6.9 AP. Most policies assist the network detect smaller defects

Table 2: Effect of optimizing aspect ratio of anchor boxes based on bounding box annotations in train dataset. For this experiment, EfficientDet-D0 model is used and the optimized aspect ratios are (1.2, 2.14, 3.8).

Default					K-means optimized (% of improvement)				
mAP	AP_{50}	AP_s	AP_m	AP_l	mAP	AP_{50}	AP_s	AP_m	AP_l
30.7	59.3	24.0	33.9	68.3	34.8 (↑**4.1**)	65.9 (↑**6.6**)	25.4 (↑**1.4**)	35.0 (↑**1.1**)	69.2 (↑**0.9**)

(i.e., AP_s which are of more importance since they are easier to miss during manual inspection and are harder to detect with deep learning methods). For further investigations, train-timeAug, and policyV3 are applied to the images as they resulted the best in these experiments. Figure 8 depicts a sample training batch with mentioned augmentations applied.

4.3. Training and Hyperparameter Tuning. The size of the models and the resolution of input images increase from D0 to D7 gradually using equations (1)–(3). It is not possible to train all the models on GPUs with 16 GB RAM with suitable possible batch size relative to model size. Models with smaller ϕ coefficients (i.e., D0 to D2) are trained on 3 NVIDIA V100 16 GB RAM GPUs with maximum possible batch size though for fitting these models in such GPU memory, a few actions are performed. First, mixed-precision training2 is applied using the Apex package which assists in decreasing memory usage and training time by utilizing half-precision weights and operations if possible. Second, as providing accurate statistics for batch normalization is crucial for the stabilized learning process and high-speed convergence, in distributed training, synchronized batch normalization is used to provide cross-device batch-norm statistics. Nonetheless, these would not help to fit D5 to D7 models in GPU. Thus, results related to those models are not reported. Finally, for comparison, several original Yolo and RetinaNet models are trained. For RetinaNet models, images are resized to 800×800 pixels, and ResNet with 50 or 101 layers are used as the backbone, and for Yolo models, images are resized to 640×640 pixels. Implementation for yolo models can be found in [68], and RetinaNet can be found in detectron2 framework [69].

During training, normalization using precomputed mean and standard deviation values per channel on the entire dataset is performed. Also, each image is first randomly flipped horizontally and/or resized for all experiments (i.e., Train-timeAug explained in Section 3.4). For weight initialization, the weights originally were trained on MS COCO dataset in [52] and are converted to PyTorch in [70]. Thus, all weights of the network are trained to reach maximum performance similar to our previous work [56].

Identical to [52], cosine learning [71] is used. At the beginning of the training process, for the first few epochs (epoch numbers 0 to 5) learning rate increases gradually to the desired point, and from epoch 5 to the end of the training process the learning rate decreases gradually in cosine form. In addition, learning rate noises applied to 30% and 90% of the training process. Moreover, in a few experiments, exponential moving average (EMA) [72] with a weight decay of 0.9998 was applied; however, it was removed as non-EMA training ended up with higher AP. Furthermore, a few optimizers are evaluated and results show in this task Fusedadam [73] optimizer converges faster and reaches a higher accuracy (0.6 mAP). In the following, the impact of different activation functions is discussed.

Table 3: Policies used to train on D0 model on entire train images. Results are reported on the validation set.

Policy name	Base (% of improvement)			
	mAP	AP_s	AP_m	AP_{50}
NoAugment	24.2	21.5	27.2	46.0
Train-timeAug	30.4 (↑**6.2**)	24.4 (↑**2.9**)	39.0 (↑**11.8**)	62.5 (↑**16.5**)
policyV0	28.0 (↑**3.8**)	23.2 (↑**1.7**)	33.0 (↑**5.8**)	56.0 (↑**10.**)
policyV1	30.3 (↑**6.1**)	22.2 (↑**0.7**)	35.1 (↑**7.9**)	61.1 (↑**15.1**)
policyV2	30.2 (↑**6.0**)	22.1 (↑**0.6**)	36.3 (↑**9.1**)	59.7 (↑**13.7**)
policyV3	31.1 (↑**6.9**)	24.9 (↑**3.4**)	37.2 (↑**10.**)	60.9 (↑**14.9**)

4.4. Effect of Different Activation Functions. The performance of the base model (i.e., EfficientDet-D0) is analyzed by testing over different activation functions, namely, Leaky Rectified Linear Unit (Leaky ReLU), Gaussian Error Linear Units (GeLU) [74], Swish [75], Mish [76], and hard Swish [66] in which sigmoid is replaced with $\text{relu6}(x+3)/6$, which is more memory efficient, and hard Mish [77]. Figure 9(a) visualizes mentioned activation functions. Note that the specified activation function is used for BiFPN layers and class/box prediction nets. As shown in Figure 9(b), both hard Swish and Swish outperforms other activation functions based on AP_{50}. The same improvement applies for other AP metrics. However, this did not happen for deeper models of the EfficientDet family. Thus, default Swish is used to maximize model performance, though in view of the memory efficiency of hard Swish, it is a preferable choice for activation function if the model is planned to get deployed on hardware-constraint end-device. For non-EfficientDet family models, the default activation function of the model is used.

4.5. Defect Object Detection without considering Class Labels. In this experiment, all discontinuities are considered with a single *defect* label. Table 4 shows network performance considering a single class for all discontinuities. Although these models only perform localization of the defects and no class label is available, higher accuracy in localization is reached. In addition to EfficientDet Family, several other models are also added for sake of comparison. All models are

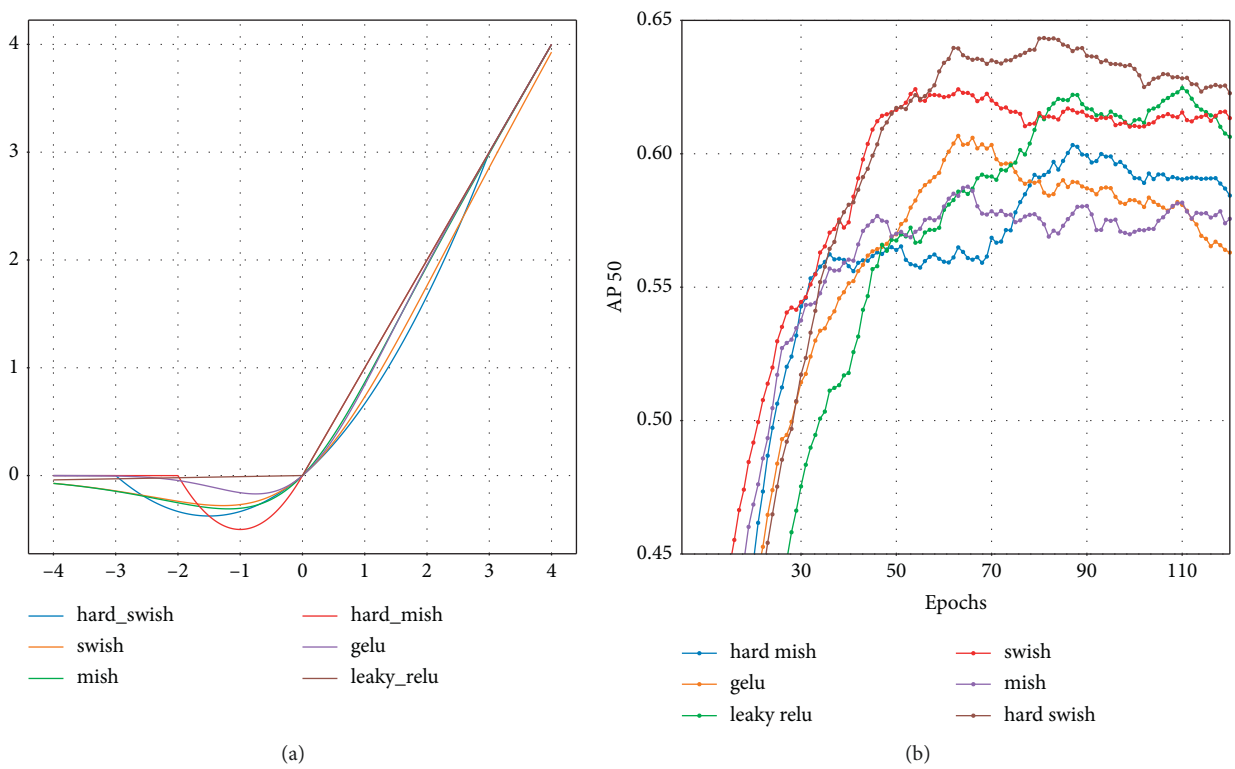


Figure 8: Activation functions used for comparing performance of the models.

trained using base parameters in their original work, except for Yolov3, in which focal loss is preferred to address class imbalance. In Section 4.6, models in Table 4 are discussed.

4.6. Evaluating Results. Tables 4–6 demonstrate models evaluation on 10-class object detection task, 10-class classification, and defect/nondefect object detection task, respectively. All results for object detection models are obtained from the test set. For all experiments, common train-time augmentations and policyV3 (described in Section 3.4) are applied. Although hard Swish improved accuracy for shallower models, the same did not happen for deeper ones (D3 and deeper). Thus, default activation functions of the models are used (i.e., Swish for EfficientDet models, leaky-ReLU for CspResdet50 and darkdet53, and ReLU for Resdet50). As mentioned in the tables, generally, deeper models show more accurate performance. However, little improvement or minor deterioration in larger models is a result of having to use smaller batch size to be able to fit the model into the GPUs (i.e., batch size of 20 per GPU is used to train the D0 versus batch sizes 3 and 1 for D3 and D4 models, respectively), which accounts for inaccurate estimation of statistics of batch normalization and deteriorates training process. Since for task of weld quality assessment and indexing of weld as well as rejecting or accepting, predicting 50% of the discontinuity is acceptable, AP_{50} is used for further model comparison and analysis. AP_l is not reported because defects in welds with area greater than 96^2 pixels are undersampled and uncommon. Therefore, it would not be a reliable measure to evaluate the performance of the models, and it is not reported.

For inference time analysis, a similar GPU that is used for training, NVIDIA V100 16 GB, is exploited. Inference times in Tables 4 and 6 suggest that models are able to perform in real-time performance based on definition of real-time for object detection models [78]. As a result, the fastest and the most accurate models can infer up to 224 and 150 image patches per second with a batch size of 16, respectively. Considering the fact that in the worst case each complete weld image has a length of 15360 pixels and is cropped with 20% overlap, a full image will have about 86 patches, meaning models can infer an image in 385 to 465 ms. Thus, models are able to process weld images in real time with consideration of required preprocessing. Figure 10 summarizes models' latency and floating-point operations (FLOPs) count. Yolo models have a higher number of operations, and the resulting models are larger. In contrast, the EfficientDet family models and models with Bi-FPN Layers enhanced with AutoAugmentation are both smaller and more accurate. Although EfficientDet models have a smaller number of parameters, they perform slower on GPU because of slower execution of separable convolution. Finally, a fusion of Resnet50 with EfficientDet object detection architecture results in best accuracy versus latency, for this task.

In Tables 4 and 6, models reported above double line are trained for the sake of comparison. In Yolo models in addition to Train-timeAug, mosaic augmentation is applied. Although this improved results by 0.5 AP, EfficientDet

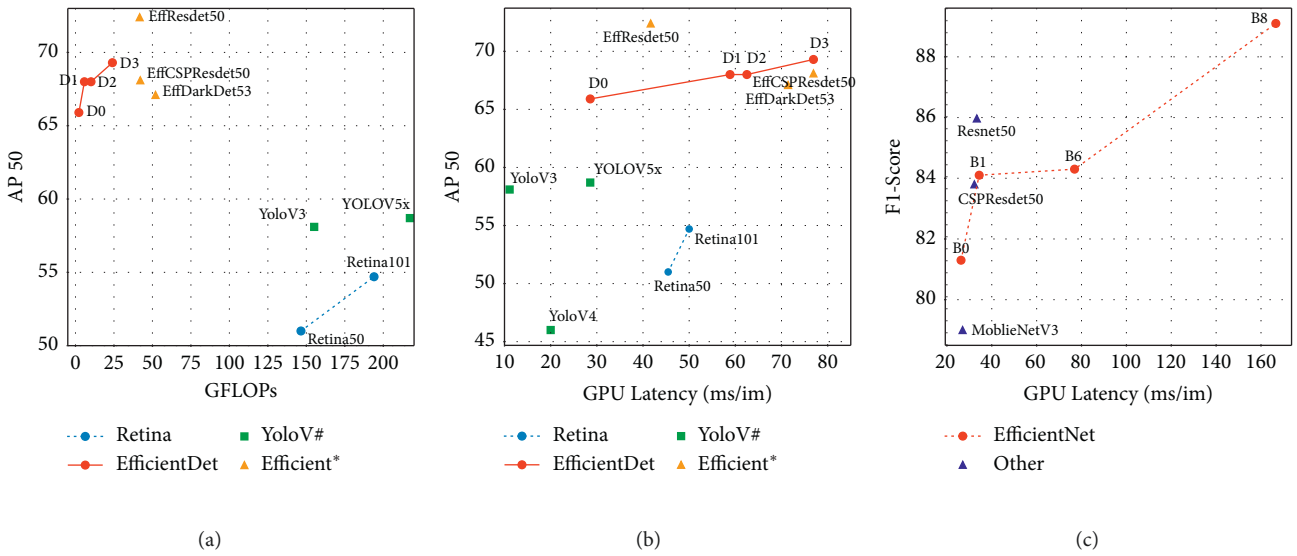


Figure 9: (a) AP_{50} accuracy versus giga floating-point operations (GFLOPs) for 10-class object detection models. (b) AP_{50} versus GPU latency (milliseconds/224 × 224 image patch) for 10-class object detection. (c) F1-score versus GPU latency (ms/224 × 224 image patch) performance for 10-class classification models.

Table 4: Model defect detection performance on test set based on defect/nondefect labels.

Model Name	# of params	Test					Inference time (images/sec)		
		AP	AP_{50}	AP_{75}	AP_s	AP_m	$b = 1$	$b = 8$	$b = 16$
Yolo v3	61.4 M	34.3	81.0	—	—	—	54	65	75
Yolo v4	64.3 M	27.4	61.6	—	—	—	48	130	185
Yolo v5x	87 M	34.8	81.9	—	—	—	35	50	56
RetinaNet-50	37.9 M	29.6	73.6	14.5	29.5	30.5	22	—	—
RetinaNet-101	56.9 M	30.9	75.2	14.2	29.9	31.5	20	—	—
EfficientDet-D0	3.8 M	37.1	87.5	18.5	37.5	37.3	35	130	185
EfficientDet-D3	11.9 M	37.7	88.1	19.7	36.7	38.8	14	56	65
Efficientdarkdet53	11.9 M	36.6	85.7	18.9	38.4	35.5	15	34	37
EfficientCspResDet50	23.6 M	37.6	86.2	20.2	38.8	36.8	15	36	39
EfficientResDet50	**26.9 M**	**38.2**	**89.0**	**21.8**	**37.8**	**40.9**	**27**	**143**	**168**

Inference times are reported based on tests on a V100 GPU card. For EfficientDet models, number D# stands for ϕ coefficient of the model, and depth and input image resolution can be acquired using equations (1)–(3). Also, the number stands for backbone number in [53]. For other models, the same structure of EfficientDet-D1 is used except the backbone, the name of which is determined in Model Name. For models without efficient in their name, implementations from [69] or [68] with common augmentation are employed.

Table 5: 10-class classification backbone performances.

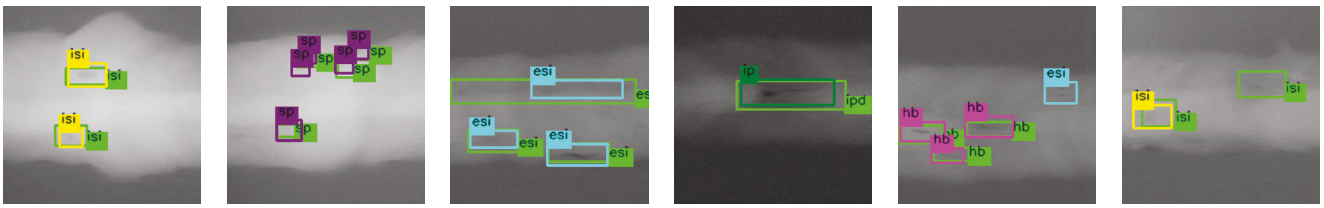
Model Name	Image Size	# of params	Top-1 accuracy	Precision	Recall	F1-score	Inference time (images/sec)	
							$b = 1$	$b = 16$
Resnet50	224	23.5 M	86.83	85.68	86.26	85.97	30	207
CspResnet50	224	20.6 M	85.3	83.89	83.71	83.8	31	266
MobileNetV3	224	4.2 M	79.91	79.41	78.67	79.04	37	267
EfficientNet-b0	224	3.8 M	82.15	81.4	81.36	81.38	38	355
EfficientNet-b1	240	6.5 M	85.4	83.8	84.4	84.1	29	200
EfficientNet-b6	528	40.7 M	85.71	84.5	84.1	84.3	13	26
EfficientNet-b8	**672**	**89.6 M**	**90.2**	**89.5**	**88.72**	**89.11**	**6**	**10**

models are still more accurate. In contrast to large number of parameters in Yolo, they still perform faster than EfficientDet models and the reason is that depth-wise separable convolutions is employed for feature fusion in EfficientDet and they run slower on GPU. However, thanks to lower number parameters, these models will perform better on

Table 6: 10-class defect detection models performance on test set based and inference time with multiple batch sizes.

Model Name	# of params (M)	Test set					Inference time		
		AP	AP_{50}	AP_{75}	AP_s	P_m	$b^{b}=1$	$b=8$	$b=16$
YOLO V3	61.5 M	30.9	58.4	—	—	—	0	150	150
YOLO V4	64.3 M	24.1	46.0	—	—	—	0	120	125
YOLO V5x	87.2 M	30.1	58.1	—	—	—	5	50	56
RetinaNet50	37.9 M	26.7	51.0	18.4	20.	27.6	21	—	—
RetinaNet101	56.9 M	27.7	54.7	18.2	21.	30.7	20	—	—
EfficientDet-D0	3.8 M	34.8	65.9	22.2	2.4	35.0	35	144	224
EfficientDet-D1	6.5 M	34.9	68.0	23.7	27.3	34.4	17	111	148
EfficientDet-D2	8M	35.2	68.0	23.0	26.0	36.5	16	88	106
EfficientDet-D3	11.9 M	34.7	69.3	25.1	25.9	34.9	13	58	66
Efficientdarkdet53	11.9 M	34.4	67.1	25.4	264	33.9	14	33	37
EffcientCspResdet50	23.7 M	34.1	68.1	22.5	25.7	33.6	13	35	39
EfficientResDet50	**27M**	**36.1**	**72.4**	**25.0**	**26.0**	**37.7**	**24**	**132**	**150**

For EfficientDet models, number D# stands for ϕ coefficient of the model and depth and input image resolution can be acquired using equations (1)–(3). Also, the number stands for backbone number in [53]. For other models, the same structure of EfficientDet-D1 is used except the backbone. Note that all models are optimized for maximum AP. b denotes batch size for inference.



(a) (b) (c) (d) (e) (f)

Figure 10: Examples of network prediction: true positives (a, b) (green boxes denote ground truth and others are network prediction), false negatives and false positive as a result of IoU < 0.5 (c), erroneous class prediction because of the similarity of the classes IP and IPD (d), and hard samples which resulted in incorrect prediction as a consequence of not meeting minimum standards (e, f).

CPUs compared with competitors. Results from Table 4, in which for all discontinuities a single label is used, suggest that a portion of false-positive detections are related to detecting correct labels. In the following, erroneous and missed detections of the best model are elaborated upon. Also, the performance of feature extraction backbones both numerically and visually is discussed.

Error analysis: based on Table 6 and inferred images of the test set, erroneous detections of the network with highest AP_{50} belong to one of these subcategories: (1) errors as a consequence of inadequate IoU of detections and ground truth: 0.7% of error cases of test set belong to this category, and they are from 3 classes of IP, ESI, and GP, where 76% of cases are ESI. Figure 11(c) shows a sample from this category. (2) False positives were mostly related to instances that a nondefect bounding box is detected, and it is closely similar to one of the other defect classes, and a nonexpert observer might consider it as a defect. However, it does not meet minimum requirements such as length for slag inclusion, size for porosity, and other criteria to be counted as a discontinuity. Finally, out of 4.5% of instances lying in this category, slag inclusions and HBs had the largest normalized percentage of errors. Mostly, sides of the weld root and also weld toe were falsely predicted as slag inclusions. Figure 11(e) is a sample of this category. As a workaround to reduce the error rate of this type, adding a large number of similar image patches to the train set is suggested. (3) False negatives are where the network does not detect defects. With more than 12% of false negatives, this group contains the largest erroneous behaviour of the model, with HBs and ESIs forming more than 55% of the normalized number of false positive detections. Figure 11(f) is a sample of a false negative. A suggested workaround is to perform online or offline hard example mining for training. Note that by lowering the minimum confidence threshold, most of these are detected concisely by the network. (4) Misclassified samples are when the network detects the object with acceptable IoU, though the class label is incorrect. A sample is shown in Figure 11(d). Finally, Figure 12(a) shows the distribution of misclassified samples from the 10-class object detection model. It is showing that HB class has the most misclassified detections, and it is mostly mistaken with class IC, and the similarity is that both IC and HB create a hallow area in the weld root.

Comparison of backbones: although it is common that multiple discontinuity types appear in a single patch, a part of the dataset (which includes around 80% of each set) that holds image patches with a single defect type is separated and used to evaluate feature extraction and backbone performance, and also to train a classification model. Table 5 shows performance of various backbones. The most accurate backbone is EfficientNet-B8, with 90.2% accuracy on the validation set. A similar training environment and optimizer with object

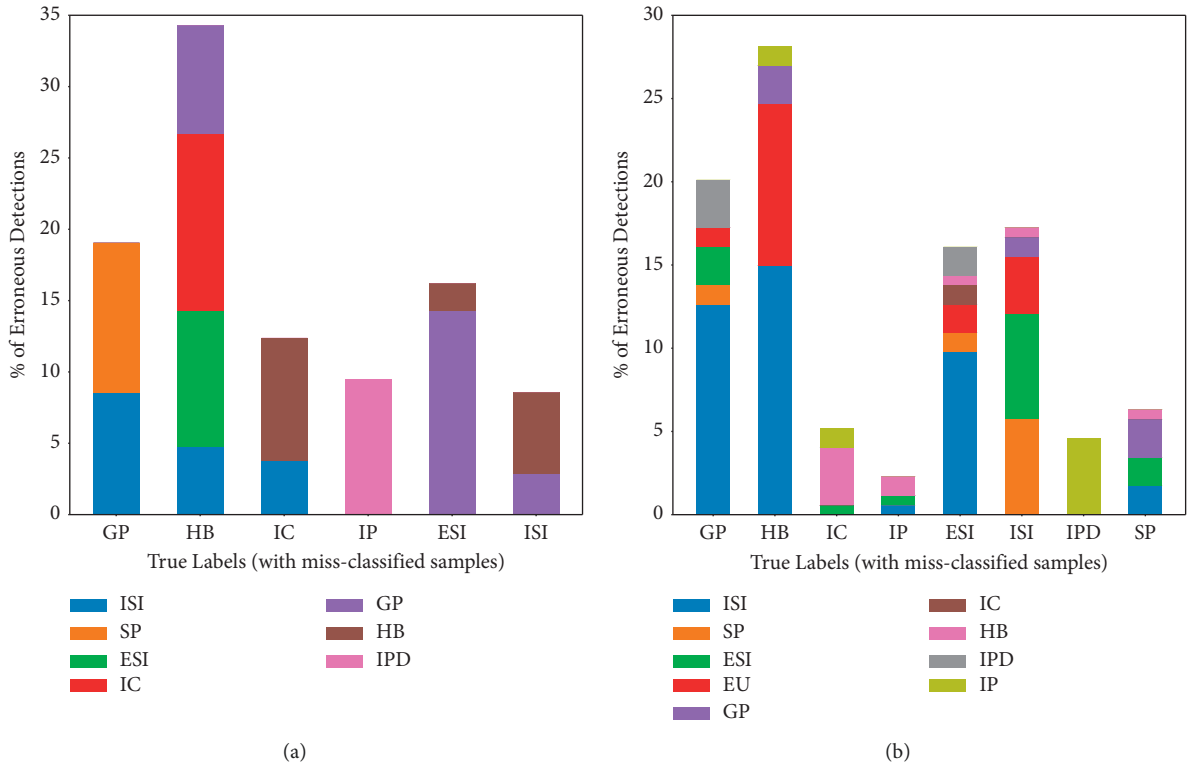


Figure 11: Misclassified detected defects. (a) Misclassification distribution for 10-class object detection model and (b) number of misclassified patches in classification backbone (EfficientNet-b8).

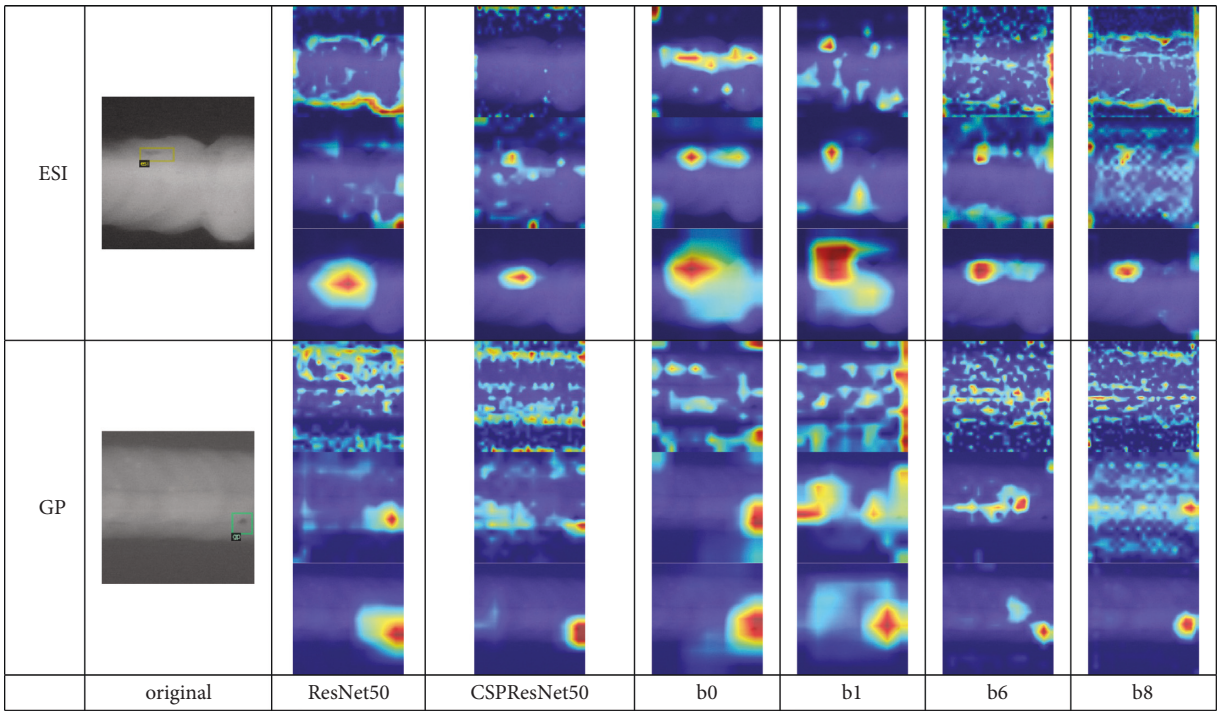


Figure 12: Grad-CAM visualization of different backbones from their three last blocks for classes GP and ESI.

detection models are used for this purpose. Transfer learning is applied and weights were originally trained on ImageNet [77]. Figure 12(b) shows the distribution of erroneous behavior of the classifier. Based on Figure 12(b), most of the misclassified samples are related to HB. Also, erroneous detections are mostly misclassified as ISI, as shown in Figure 12(b).

Explainability using Grad-CAM: gradient-weighted class activation mappings (Grad-CAMs) [79] recognizes the parts of input image with deterministic role in final decision-making of the model. In Grad-CAM, instead of applying global average pooling as ending layers [80] which requires model modification and affects the network performance, back-propagation is utilized to extract feature contributions. Therefore, class activation maps get extracted precisely. In Figure 7, Grad-CAMs of various backbones with different depth are visualized, which provides local explainability for input images. Bottom image of each cell shows final layer Grad-CAM, and upper images are second to last and third from last block output. It shows how network gradually attend to discriminative features of each image.

5. Conclusions

In this paper, a scalable and efficient family of deep models for 10-class weld quality assessment using object detection is presented. A comparative analysis on various models is also performed; several critical elements of the networks such as activation functions and hyperparameters are explored and tuned to achieve state-of-the-art results on the dataset. Moreover, the effects of transferring object detection AutoAugment policies are surveyed. Furthermore, various scenarios such as considering task as a classification only task and defect/nondefect scenarios are also analyzed and models are compared with main-stream object detection models in real-time applications. Finally, model visual explainability is analyzed through employing Grad-CAM and visualizing gradient information for target class. The results are interpreted. They demonstrate that models are able to infer a complete welded joint (15360×1024 resolution X-ray Image) in 385 milliseconds. Although classification task outperforms object detection models, localization of the defect (whether the defect is on root pass, fill pass, or cover pass) is necessary for further indexing of the weld, pass or rejection, and optimization of welding operation.

Traditional computer vision techniques for weld defect detection require several critical preprocessing steps resulting in a nonrobust outcome or human intervention is needed. In contrast, automatic feature extraction approaches and deep learning-based methods require minimum human intervention or preprocessing to achieve state-of-the-art results. The models presented here can be used as assistive defect-recognition systems to facilitate robust defect localization and classification and to reduce both human workload and error. Finally, as experts may have conflicting and personal performance in particular defect detection, provided deep models may train on specific samples and predict defects with a consolidated standard which can also be helpful in training experts.

Future works contain test-time augmentation, model ensemble without sacrificing real-time capability of the system, searching for optimal auto augmentation policies utilizing reinforcement learning since policies were initially extracted from the COCO dataset and the nature of the weld images is not consistent with nature of COCO dataset images. In addition, through time, more samples will be gathered from various sites of different parts of the world, and the dataset will expand in both the number of classes and the number of instances per class.

Data Availability

All open-source implementations used in this paper are referenced in the main body of the article. However, the remaining implementations and dataset are a part of ongoing research and proprietary of Stanley Black & Decker, USA.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was done primarily with the help and support from Michael George, Jeremy Guretzki, Matthew Nelson, Jason Miller, William Aston, Jake Smith, Haresh Ghansyam, Pete Morris, Prashanth Tirumalaseti, Adam Wynne Hughes, and Shengnan Wang as well as great support from Dr. Mark Maybury and Dr. Manish Mehta from the office of CTO at Stanley Black and Decker. This research was conducted at Lamar University and was fully funded by Artificial Intelligence Lab, Stanley Oil & Gas, Stanley Black & Decker, USA.

References

[1] R. Vilara, "An automatic system of classification of weld defects in radiographic images," *NDT & E International*, vol. 42, no. 5, pp. 467–476, 2009.

[2] M.-M. Naddaf-Sh, S. Naddaf-Sh, H. Zargaradeh et al., "Next-generation of weld quality assessment using deep learning and digital radiography," in *Proceedings of the AAAI Spring Symposium Series*, Palo Alto, CA, USA, March 2020.

[3] M.-M. Naddaf-Sh, S. Naddaf-Sh, H. Zargarzadeh et al., "Defect detection and classification in welding using deep learning and digital radiography," in *Fault Diagnosis and Prognosis Techniques for Complex Engineering Systems*vol. 2021, pp. 327–352, 2021.

[4] A. Shahriar, R. Sadiq, and S. Tesfamariam, "Risk analysis for oil & gas pipelines: a sustainability assessment approach using fuzzy based bow-tie analysis," *Journal of Loss Prevention in the Process Industries*, vol. 25, no. 3, pp. 505–523, 2012.

[5] Z. Rui, G. Han, H. Zhang, S. Wang, H. Pu, and K. Ling, "A new model to evaluate two leak points in a gas pipeline," *Journal of Natural Gas Science and Engineering*, vol. 46, pp. 491–497, 2017.

[6] L. T. Popoola, A. S. Grema, G. K. Latinwo, B. Gutti, and A. S. Balogun, "Corrosion problems during oil and gas production and its mitigation," *International Journal of Integrated Care*, vol. 4, no. 1, pp. 1–15, 2013.

[7] S. M. Anouncia and R. Saravanan, "Non-destructive testing using radiographic images? a survey," *Insight-Non-Destructive Testing and Condition Monitoring*, vol. 48, no. 10, pp. 592–597, 2006.

[8] X. Dong, C. J. Taylor, and T. F. Cootes, "Small defect detection using convolutional neural network features and random forests," *Lecture Notes in Computer Science*, vol. 11132, pp. 398–412, 2019.

[9] A. Shukla and H. Karki, "Application of robotics in offshore oil and gas industry-a review part II," *Robotics and Autonomous Systems*, vol. 75, pp. 508–524, 2016.

[10] L. Yang, Y. Liu, and J. Peng, "Advances techniques of the structured light sensing in intelligent welding robots: a review," *International Journal of Advanced Manufacturing Technology*, vol. 110, pp. 1–20, 2020.

[11] S. Habibian, M. Dadvar, B. Peykari et al., "Design and implementation of a maxi-sized mobile robot (karo) for rescue missions," *ROBOMECH Journal*, vol. 8, 2021.

[12] M. Dadvar and S. Habibian, "Contemporary research trends in response robotics," 2021, https://arxiv.org/abs/2105.07812.

[13] Q. Ma, G. Tian, Y. Zeng et al., "Pipeline in-line inspection method, instrumentation and data management," *Sensors*, vol. 21, no. 11, 2021.

[14] W. Hou, Y. Wei, J. Guo, Y. Jin, and C. Zhu, "Automatic detection of welding defects using deep neural network," *Journal of Physics: Conference Series*, vol. 933, no. 1, 2018.

[15] X. Dong, C. J. Taylor, and T. F. Cootes, "A random forest-based automatic inspection system for aerospace welds in x-ray images," *IEEE Transactions on Automation Science and Engineering*, pp. 1–14, 2020.

[16] D. Mery and M. A. Berti, "Automatic detection of welding defects using texture features," *Insight-Non-Destructive Testing and Condition Monitoring*, vol. 45, no. 10, pp. 676–681, 2003.

[17] J. Kumar, R. Anand, and S. Srivastava, "Multi-class welding flaws classification using texture feature for radiographic images," in *Proceedings of the International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 1–4, Vellore, India, 2014.

[18] J. Kumar, R. S. Anand, and S. P. Srivastava, "Flaws classification using ann for radiographic weld images," in *Proceedings of the 2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 145–150, Noida, India, February 2014.

[19] J. Hassan, A. M. Awan, and A. Jalil, "Welding defect detection and classification using geometric features," in *Proceedings of the 2012 10th International Conference on Frontiers of Information Technology*, pp. 139–144, Islamabad, Pakistan, December 2012.

[20] O. Zahran, H. Kasban, M. El-Kordy, and F. E. A. El-Samie, "Automatic weld defect identification from radiographic images," *NDT & E International*, vol. 57, pp. 26–35, 2013.

[21] T. Y. Lim, M. M. Ratnam, and M. A. Khalid, "Automatic classification of weld defects using simulated data and an mlp neural network," *Insight-Non-Destructive Testing and Condition Monitoring*, vol. 49, no. 3, pp. 154–159, 2007.

[22] J. Zapata, R. Vilar, and R. Ruiz, "Performance evaluation of an automatic inspection system of weld defects in radiographic images based on neuro-classifiers," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8812–8824, 2011.

[23] I. Valavanis and D. Kosmopoulos, "Multiclass defect detection and classification in weld radiographic images using geometric and texture features," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7606–7614, 2010.

[24] W. Hou, D. Zhang, Y. Wei, J. Guo, and X. Zhang, "Review on computer aided weld defect detection from radiography images," *Applied Sciences*, vol. 10, no. 5, p. 1878, 2020.

[25] M. Carrasco and D. Mery, "Segmentation of welding defects using a robust algorithm," *Materials Evaluation*, vol. 62, no. 11, pp. 1142–1147, 2004.

[26] D. Mery, "Automated detection of welding discontinuities without segmentation," *Materials Evaluation*, vol. 69, no. 6, pp. 656–663, 2011.

[27] M. Ben Gharsallah and E. Ben Braiek, "Weld inspection based on radiography image segmentation with level set active contour guided off-center saliency map," *Advances in Materials Science and Engineering*, vol. 2015, Article ID 871602, 10 pages, 2015.

[28] C. Ajmi, S. E. Ferchichi, and K. Laabidi, "New procedure for weld defect detection based-gabor filter," in *Proceedings of the 2018 International Conference on Advanced Systems and Electric Technologies (ICASET)*, pp. 11–16, Hammamet, Tunisia, March 2018.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Munich, Germany, October 2015.

[30] W. Hou, Y. Wei, Y. Jin, and C. Zhu, "Deep features based on a dcnn model for classifying imbalanced weld flaw types," *Measurement*, vol. 131, pp. 482–489, 2019.

[31] C. Ajmi, J. Zapata, S. Elferchichi, A. Zaafouri, and K. Laabidi, "Deep learning technology for weld defects classification based on transfer learning and activation features," *Advances in Materials Science and Engineering*, vol. 2020, Article ID 1574350, 16 pages, 2020.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[33] D. Mery, V. Riffo, U. Zscherpel et al., "GDXray: the database of X-ray images for nondestructive testing," *Journal of Nondestructive Evaluation*, vol. 34, no. 4, pp. 1–12, 2015.

[34] R. Miao, Z. Jiang, Q. Zhou et al., "Online inspection of narrow overlap weld quality using two-stage convolution neural network image recognition," *Machine Vision and Applications*, vol. 32, no. 1, pp. 1–14, 2021.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, https://arxiv.org/abs/1409.1556.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, https://arxiv.org/abs/1512.03385.

[37] Q. Wang, W. Jiao, P. Wang, and Y. Zhang, "A tutorial on deep learning-based data analytics in manufacturing through a welding case study," *Journal of Manufacturing Processes*, vol. 63, pp. 2–13, 2021.

[38] A. Paszke, S. Gross, F. Massa et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., Red Hook, NJ, USA, 2019.

[39] D. Mery and C. Arteta, "Automatic defect recognition in x-ray testing using computer vision," in *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1026–1035, Santa Rosa, CA, USA, March 2017.

[40] S. J. Oh, M. J. Jung, C. Lim, and S. C. Shin, "Automatic detection of welding defects using faster R-CNN," *Applied Sciences*, vol. 10, no. 23, pp. 1–10, 2020.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," vol. 28, pp. 91–99, 2015.

[42] R. Guo, H. Liu, G. Xie, and Y. Zhang, "Weld defect detection from imbalanced radiographic images based on contrast enhancement conditional generative adversarial network and transfer learning," *IEEE Sensors Journal*, 2021.

[43] L. Yang, H. Wang, B. Huo, F. Li, and Y. Liu, "An automatic welding defect location algorithm based on deep learning," *NDT & E International*, vol. 120, Article ID 102435, 2021.

[44] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, https://arxiv.org/abs/1411.1784.

[45] T. Gantala and K. Balasubramaniam, "Automated defect recognition for welds using simulation assisted tfm imaging with artificial intelligence," *Journal of Nondestructive Evaluation*, vol. 40, no. 1, pp. 1–24, 2021.

[46] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[47] Y. Yan, D. Liu, B. Gao, G. Y. Tian, and Z. C. Cai, "A deep learning-based ultrasonic pattern recognition method for inspecting girth weld cracking of gas pipeline," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 7997–8006, 2020.

[48] J. Hu, W. Xu, B. Gao et al., "Pattern deep region learning for crack detection in thermography diagnosis system," *Metals*, vol. 8, no. 8, 2018.

[49] American Petroleum Institute, *API 5L: Specification for Line Pipe*, American Petroleum Institute, Washington, NJ, USA, 2004.

[50] American Petroleum Institute, *API 1104: Standard for Welding of Pipelines and Related Facilities*, American Petroleum Institute, Washington, NJ, USA, 2001.

[51] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7036–7045, Long Beach, CA, USA, June 2019.

[52] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, Seattle, WA, USA, June 2020.

[53] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, Beach, CA, USA, June 2019.

[54] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[55] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on computer Vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[56] S. Naddaf-Sh, M.-M. Naddaf-Sh, A. R. Kashani, and H. Zargarzadeh, "An efficient and scalable deep learning approach for road damage detection," in *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pp. 5602–5608, Atlanta, GA, USA, December 2020.

[57] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[58] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, Salt Lake City, UT, USA, June 2018.

[59] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, https://arxiv.org/abs/1712.04621.

[60] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–123, Long Beach, CA, USA, June 2019.

[61] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proceedings of the European Conference on Computer Vision*, pp. 566–583, Glasgow, UK, August 2020.

[62] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing*, pp. 237–242, 2020.

[63] "Coco detection challenge (bounding box)," 2021, https://cocodataset.org/#detection-eval.

[64] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[65] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, October 2017.

[66] A. Howard, M. Sandler, G. Chu et al., "Searching for Mobilenetv3," 2019, https://arxiv.org/abs/1905.02244.

[67] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "Cspnet: a new backbone that can enhance learning capability of Cnn," 2019, https://arxiv.org/abs/1911.11929.

[68] G. Jocher, A. Stoken, J. Borovec et al., "Ultralytics/yolov5: v5.0-YOLOv5-P6 1280 models, AWS, Supervise.Ly and YouTube integrations," 2021, https://zenodo.org/record/4679653#.YTmqdrAzbIU.

[69] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019, https://github.com/facebookresearch/detectron2.

[70] 2020 Efficientdet (A Pytorch Implementation of Efficientdet).

[71] I. Loshchilov and F. Hutter, "Sgdr: stochastic gradient descent with warm restarts," 2016, https://arxiv.org/abs/1608.03983.

[72] 2021 Exponential Moving Average." https://www.tensorflow.org/api_docs/python/tf/train/ExponentialMovingAverage.

[73] 2021 Nvidia Apex Optimizers." https://nvidia.github.io/apex/optimizers.html.

[74] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2020, https://arxiv.org/abs/1606.08415.

[75] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, https://arxiv.org/abs/1710.05941.

[76] D. Misra, "Mish: a self regularized non-monotonic neural activation function," 2019, https://arxiv.org/abs/1908.08681.

[77] R. Wightman, "Pytorch image models," 2019, https://github.com/rwightman/pytorch-image-models.

[78] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," 2016, https://arxiv.org/abs/1506.02640.

[79] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of*

the IEEE International Conference on Computer Vision, pp. 618–626, Venice, Italy, October 2017.

[80] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, Las Vegas, NV, USA, June 2016.

Hindawi
Complexity
Volume 2022, Article ID 8213855, 10 pages
https://doi.org/10.1155/2022/8213855

Research Article

An Improved Multibranch Convolutional Neural Network with a Compensator for Crowd Counting

Zhiyun Zheng, Zhenhao Sun, Guanglei Zhu, Zhenfei Wang, and Junfeng Wang

School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

Correspondence should be addressed to Junfeng Wang; iewangjf@zzu.edu.cn

Received 28 May 2021; Revised 25 February 2022; Accepted 14 March 2022; Published 28 April 2022

Academic Editor: Jenq-Haur Wang

Image-based crowd counting has extremely important applications in public safety issues. Most of the previous studies focused on extremely dense crowds. However, as the number of webcams increases, a crowd with extremely high density can obtain less error by summing the images of multiple close-range webcams, but there are still some problems such as heavy occlusions and large-scale variation. To solve the above problems, this paper proposes a new type of multibranch neural network with a compensator, in which features are extracted through multibranch subnetworks of different scales. The weights between the branches are adjusted by the compensator, and the captured features are distinguished among different branches. To avoid learning nearly the same features in each branch and reducing the training deviation, the dataset is labeled with head scale, and the adaptive grading loss function is used to calculate the estimated loss of the subregions. The experimental results show that the accuracy of the network proposed in this paper is about 10% higher than that of the comparison network.

1. Introduction

In recent years, the crowd counting based on computer vision has been widely used in video surveillance, traffic control, security services, etc., which has attracted great attention from people. At present, the solution of crowd counting has gradually developed from target detection to display density distribution, and the total number is given by integrating the density map. Inspired by the great success of convolutional neural networks in computer vision tasks, such as object detection[1, 2], image segmentation [3, 4], and object tracking [5, 6], people began to pay attention to the application of convolutional neural networks method in crowd counting.

Due to the challenges of heavy occlusions, large-scale variation, perspective effect, and large density differences in crowd counting [7], crowd counting faces great difficulties, especially for the overconcentration of crowd distribution caused by perspective effect, which seriously affects the crowd distribution in the image. The number of people owned in a smaller area accounts for a larger proportion of the number of people owned by the entire picture. As shown in Figure 1, human head targets are mainly concentrated in a small area in the middle of the image, while the total number of human head targets in the larger area below the image is significantly smaller than that in the middle area of the image. It is obvious that, compared with the crowd counting error produced in the area below the image, the error produced by the crowd counting in the middle area of the image has a more significant impact on the final evaluation index of the whole image. Therefore, the current focus of reducing crowd counting error is mainly to improve the accuracy of crowd counting in extremely dense areas.

An area with a large population density in the image also means that the head target is extremely small. In an image with a large number of people, it is often only necessary to successfully capture the head target features with medium and below target scales to get a good counting result, but this also leads to the neglect of large-scale human head targets. Moreover, the number of features that can be obtained from a single head target with a very small scale is relatively small, and overfitting of such datasets will reduce the generalization ability of the model, making it difficult to be applied in actual scenarios.

Figure 1: Image of people with large density differences.

This is also one of the problems in current population counting.

With the increasing demand in reality, especially for semi-enclosed areas with high crowd density and danger, such as stadiums and theatres, the number of webcams gradually increases; in the meantime, the distance to the crowd is closer; then, the corresponding scale of nearby human head target also follows. By increasing the webcams, the distance between the webcams and the crowd is shortened, and thus the scene of great crowd density is transformed into the sum of the crowd count in multiple regions, making the result more accurate. However, the problem of population scale variation still exists. Even a close-up of the camera will cause an increase in the nearby human head target, leading to the greater variation in the human head target. Therefore, crowd counting with large-scale variation is of great value in real applications.

The head-scale information of the image is not marked in the head data annotation. Since the convolutional neural network relies heavily on the information provided by the dataset during the training process, the lack of scale annotation information will also affect the results of network learning. Therefore, this article adds grading labeling of head scales in the dataset to increase the missing head-scale information in the dataset; then, according to the scale information, the adaptive grading loss function is used. According to the head size information in the image blocks, the final loss is obtained by calculating the local loss and accumulating the loss of all blocks. Moreover, the loss of all blocks is accumulated, and then the final loss is obtained.

Aiming at the large-scale change of crowd counting, this paper proposes a multibranch convolutional neural network with a compensator. The multibranch structure can effectively solve the multiscale feature problem, and the compensator can perform weight compensation on the outputs of different branches. Different branches use convolution kernels of different scales. For different scale features, branches of different receptive fields can be captured correspondingly, and the performance at different branches is optimized at the same time. The network also uses the adaptive grading loss function and adaptively uses the asynchronous loss function according to the target scale. The head-scale grading label is manually added to the dataset, which preserves the head-scale information in the image to a certain extent. Finally, a comparative experiment is conducted on the dataset with a large variation in feature scale. Compared with the comparison network, the population counting accuracy of this network in the dataset is improved.

2. Related Work

Crowd counting is mainly divided into two categories, namely, detection-based methods and regression-based methods. Detection-based methods [8, 9] have good results when the crowd is sparse and the occlusion is not heavy. However, it is often difficult to obtain good crowd counting based on the detection in complex actual scenes, such as heavy occlusion and complex background. The regression-based method calculates the number of people by learning the mapping relationship between image features and density maps. However, this method ignores the spatial information in the image and cannot intuitively feel where the crowd gathers. Recently, with the great success of convolutional neural networks (CNNs), researchers in this field have focused on training CNNs-based models to generate high-quality density maps and thus improving counting performance [10, 11]. The practice has proved that the use of convolutional neural networks for crowd counting is very effective, but early models are affected by scale variation, which usually leads to a decrease in inaccuracy. For this reason, some scholars have proposed a series of multibranch neural networks to extract multiscale features, aiming to extract features of different scales by using convolution kernels of different scales to deal with the problem of large-scale variation. For example, the multicolumn convolutional neural network (MCNN) proposed by Zhang et al. [12] used multi-size filters to extract features with different sizes and finally integrated these features into the same density map. Similarly, Sam et al. [13] proposed the Switch-CNN which used a switch classifier to select the best classifier from a pool of density generators. Subsequently, in order to make the network have certain specialties, scholars began to optimize the network in a targeted manner, so that the network has the characteristic ability to solve some certain problems. For example, Sindagi and Patel [14] proposed a multilevel bottom-top and top-bottom fusion network (MBTTBF), which was carefully designed to combine multiple shallow and deep features. Chen et al. [15] proposed a scale pyramid network (SPN), which extracted multiscale features in parallel by using the dilated convolution of different dilation rates in a shared single-column CNN. A scale-based attention model was proposed to

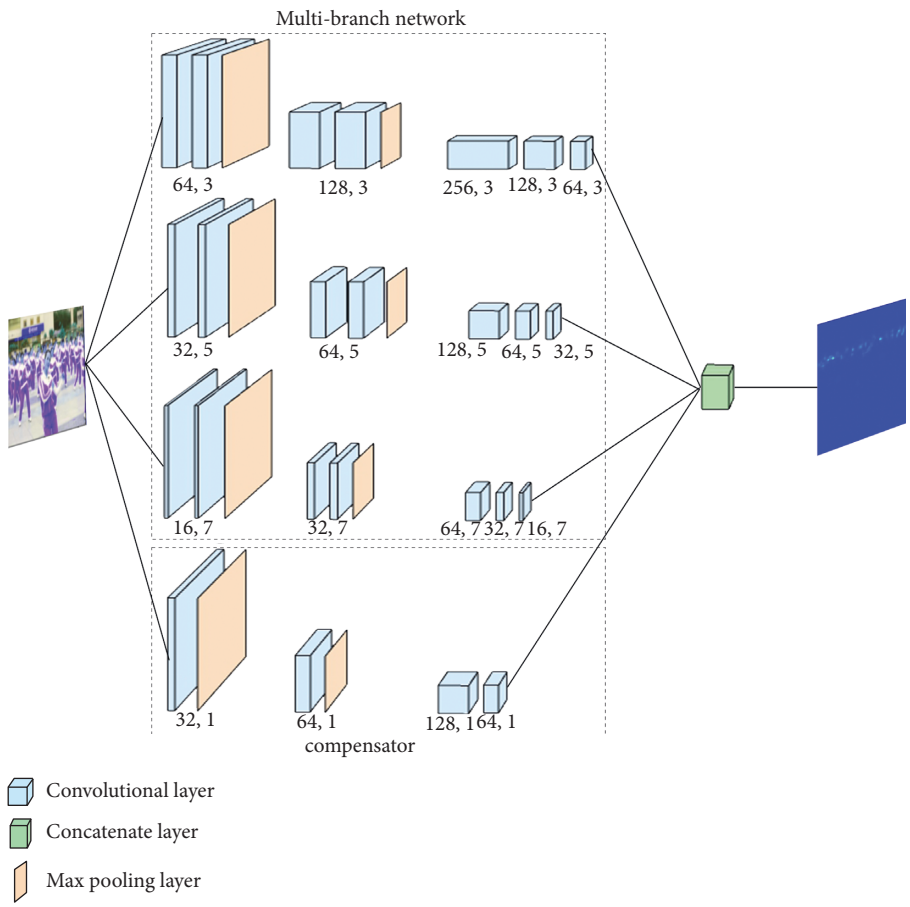


Figure 2: The overall architecture of the multibranch neural network with a compensator.

adaptively select the appropriate head size and shape [16, 17]. Wang [18] et al. proposed a combination of dilated convolutions with multiple dilation rates to process population characteristics of different scales using dilated convolutions to capture larger-scale targets, which has the advantage of fewer parameters.

Although these methods have made effective improvements, there are still some limitations:

(1) Multiple branches of exactly the same length will learn almost the same characteristics, which deviates from the original purpose of multibranch design.

(2) The dilated convolution itself has a strong ability to capture large-scale features, but it has a poor ability to capture small-scale features. Using dilated convolution requires a more complex network structure to ensure that the network can simultaneously capture features at different scales.

(3) The network structure is overly complex, and the network level is deep, so it is difficult to generalize well without a large enough dataset.

Aiming at the above problems, this paper proposes a multibranch neural network with a compensator.

(1) We used a multibranch network to solve multi-scale problems; at the same time, we increased the compensator module, adjusted the weights between the multibranch structures, and increased the degree of difference in the learned features of each branch.

(2) The dataset is added with scale level information, and the head features are divided into 4 levels according to the scale to facilitate the network to learn.

(3) In view of the large difference in target scales, a hierarchical loss function based on target scales was proposed.

In response to the above problems, this paper proposes a multibranch neural network with a compensator. The network adjusts the weights among the branches through the compensator and increases the differences in learning characteristics of each branch. The shallow structure ensures the generalization ability of the network.

3. The Proposed Method

3.1. Multibranch Convolutional Neural Network. In order to solve the problems of the huge variation in the scale of crowd counting and similar extraction features of each branch in the multibranch structure, a multibranch neural network with a compensator is proposed. The overall architecture is shown in Figure 2. The network consists of an

input, an output, a multibranch neural network, and a compensator.

The input part can receive input images of different pixels, but in order to prevent the image from being too large, the image larger than 224 pixels is reduced. Because this dataset needs to be reduced in a small proportion, it will not cause serious loss of precision due to the reduction, which will affect the crowd counting precision. The output part is connected to the output of the multibranch neural network; the compensator module and the outputs of different branches enter the output part together; then, the output part generates a density map showing the distribution of the crowd. The multibranch neural network part is a multiscale feature extraction module composed of three branches. A network which is excessively deep may weaken the generalization ability. However, the 3×3 convolution kernel, which has a better effect in extracting image features, is difficult to perform well when extracting large-scale features in a shallower network due to insufficient receptive fields. Based on this consideration, this paper adds branches of large-scale convolution kernels while using a shallower network. Finally, a network structure composed of three 7-layer branches is used to extract features of the image. The three branches in Figure 2 use 3×3, 5×5, and 7×7 convolution kernels from the top to the bottom. In order to optimize the training efficiency, the number of convolution kernels decreases by half each time as the scale of the convolution kernel increases.

By calculating the receptive field and comparing the receptive field with the image scale, it can be seen that the large-scale convolution kernel plays an important role in capturing large-scale objects. The receptive field size of the first branch is calculated as follows:

$$RF_i = (RF_{i-1} - 1)\text{Stride}_i + \text{KSize}_i, \tag{1}$$

where RF is the receptive field, i is the number of layers, Stride is the step size, and KSize is the size of the convolution kernel.

By bringing in the common influence of the convolution kernel and the pooling layer on the receptive field, the receptive field of the first branch above the network is 40×40. Similarly, the receptive fields of the second and third branches are calculated to be 76×76 and 112×112, respectively. As can be seen from Figure 3, the receptive field of 112×112 can already read a quarter of the input image with pixels of 224×224, which is enough to cope with scale changes in most cases. Even if there are features that exceed this scale, the number of them is less than 4, which has little impact on the counting results, but a larger number of parameters of the convolution kernel have a negative impact on counting accuracy, network training, network generalization, etc. Therefore, larger convolution kernels are not used.

4. Branch Weight Compensation

The bottom of Figure 2 is the compensator module, which compensates for the results of other branches through image features and optimizes the final result. After experimental tests, the network itself shows the ability to handle complex scenes; the excessively complex compensator module makes it difficult for the model to converge and to achieve good results. The simpler compensator module can actually improve the network's ability to extract simple features.

Figure 3: Effect image of receptive field.

The compensator module is implemented by using a single branch, which allows the network to adaptively learn the output weights of other branches by grasping the image input characteristics. After testing, the network will be used to learn feature information, even with a very small number of 3×3 convolution kernels, which defeats the original purpose of designing the compensator. Therefore, the compensator is only composed of a 1×1 filter and a max pooling layer, which is used to organize the feature information of the image and perform weight compensation on the first three branches, so as to optimize the weights of different branches under different scale features.

The three convolution operations ensure its nonlinear structure, and the two max pooling operations can first keep the same size as the output of other branches, that is, a quarter of the input image. At the same time, the receptive field of the 1×1 convolution kernel can be increased to 4×4, which helps the branch to better grasp the characteristics of the input image while keeping fewer parameters.

The outputs of the three branches of feature extraction are Y_i, Y_j, Y_k ($i \in \{1, 2, \ldots, 64\}$, $j \in \{1, 2, \ldots, 32\}$, $k \in \{1, 2, \ldots, 16\}$), ($i \in \{1, 2, \ldots, 64\}$, $j \in \{1, 2, \ldots, 32\}$, $k \in \{1, 2, \ldots, 16\}$); the output of the compensator is c_t ($t \in \{1, 2, \ldots, 128\}$). The weight when they perform the first convolution operation is ω_{mn} ($m \in \{1, 2, \ldots, 64\}$, $n \in \{1, 2, \ldots, 200\}$), $\sum \omega (c + Y_i + Y_j + Y_k)$. After the second convolution operation, the weight is δ_m, and the final weight compensation formula is calculated as follows:

$$M = \sum_{m=1}^{64} \delta_m \left[\sum \omega_{m,n} \left(c + Y_i + Y_j + Y_k \right) \right], \tag{2}$$

where i represents different branches, j represents the number of filters, Y represents the value of the branch output, and ω represents the weight corresponding to its Y.

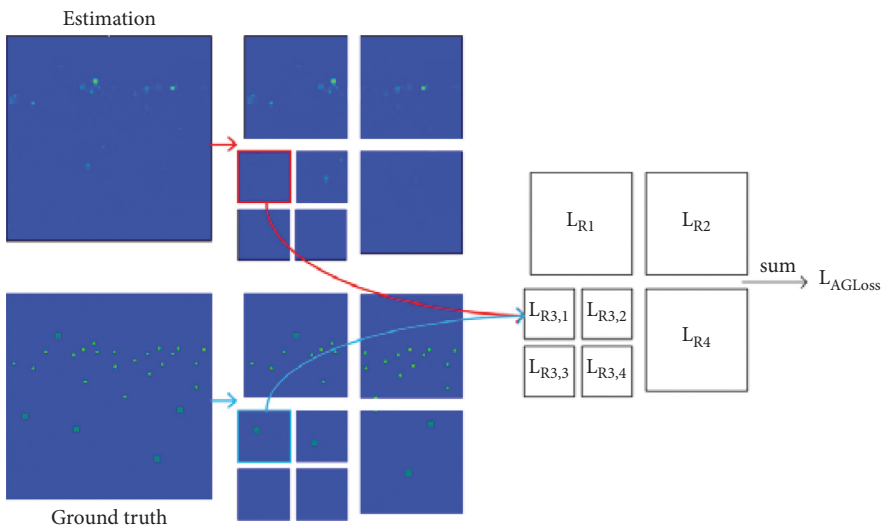


Figure 4: Demonstration of two-level adaptive classification loss.

By compensating for the weights, the output of each branch that had been originally fixed can have a certain choice space, which helps the multibranch structure to have more free choices when learning different features. To a certain extent, we avoid all branches trying to learn exactly the same characteristics, thus making differences between branches.

5. Adaptive Grading Loss Function

In the training phase, the previous CNN-based density estimation network usually uses the Euclidean distance between the entire estimated density map and the ground truth density map as the loss function [19], as shown in the following formula:

$$L(\theta) = \frac{1}{M}\sum_{k=1}^{M}\left\|D\left(X^k;\theta\right) - D^k\right\|_2^2, \tag{3}$$

where X^k is the k-th input image, D^k is its ground truth density map, θ is the parameter of the counting network, $D(X^k;\theta)$ is the estimated density map, and M is the size of the training set.

This loss function ignores the influence of feature information of different scales on the network training process. Since different scale features are often required to feed back different size steps, this loss function cannot meet the asynchronous requirements of backpropagation of different scale features, which weakens the learning ability of the network when training different scale features. In order to solve the above problems, the loss function is optimized by making full use of the dataset with scale information. For different crowd densities, an adaptive hierarchical loss function (AGLoss) [20] is proposed with reference to Adaptive Pyramid Loss (APLoss). AGLoss can adaptively divide the density map into subregions at different levels according to the real head scale. Then, the AGLoss computes the relative estimated loss for each part and adds it to get the final loss. The data is preprocessed hierarchically using the same structure as the 3×3 branch of the network so that subsequent networks use different loss functions.

Table 1: Specification of the employed dataset. Num: the number of images; max: the maximal crowd count; min: the minimal crowd count; ave: the average crowd count; total: the total number of labeled people.

Num	Resolution	Max	Min	Ave	Total
503	Different	78	2	25.9	13028

Table 2: Performance comparison of different methods on dataset.

Method	MAE	MSE
MSCNN [21]	12.6	368.9
MCNN [12]	8.8	350.8
DSA-CNN [22]	8.2	306.8
DCN [23]	8.1	311.3
Ours	7.3	302.4

Specifically, the AGLoss is calculated in the following way. First, the real population density map D^k is divided into a 2×2 first-level grid, and R_{i_1} is used to represent the subregions, where $i_1 \in \{1,2,3,4\}$. If the local population feature scale of the subregion R_{i_1} is greater than the given threshold T_{i_1}, it is divided into a 2×2 secondary subgrid. Figure 4 shows the two-stage AGLoss calculation. The human head feature scales in the secondary subgrids are all greater than T_{i_2}, then secondary subgrids are iteratively divided into 2×2 three-level subgrids until the feature scale of this area is less than T_{i_n}. Let $R_{i_1,\ldots,i_n}$ denote the in-layer subregion, $i_n \in \{1,2,3,4\}$. After the segmentation is completed, an uneven grading grid can be obtained. Apply the obtained adaptive grading grid to the estimated density map $D(X^k;\theta)$ and calculate the local loss of each subregion based on the following equation:

Figure 5: People with large-scale variation.

$$l^k_{R_{i_1,\ldots,i_n}} = \begin{cases} \dfrac{\left\| D_{R_{i_1,\ldots,i_n}}\left(X^k;\theta\right) - D^k_{R_{i_1,\ldots,i_n}} \right\|_2^2}{\left(1/\max\left(D^k_{R_{i_1,\ldots,i_n}}\right)\right)+1}, & \max\left(D^k_{R_{i_1,\ldots,i_n}}\right) < T_{i_n}, \\ \sum_{i_n=1}^{4} l^k_{R_{i_1,\ldots,i_n}}, & \text{otherwise.} \end{cases} \tag{4}$$

Finally, sum up all local losses to obtain the final AGLoss according to the following formula:

$$L_{\text{AGLoss}} = \frac{1}{M}\sum_{k=1}^{M}\sum_{i_n=1}^{4} l^k_{R_{i_n}}. \tag{5}$$

6. Case Study

6.1. Data Description. In order to obtain data of scenes with moderate density and large variation in the crowd scale, a dataset with large variation in head size and different shooting distances was selected from the public crowd counting competition dataset. The usual data preprocessing method is Gaussian blur processing for each head feature, while all heads use the same size Gaussian kernel processing, and all head features will be treated as the same scale feature during crowd counting feedback, which makes the image originally containing the head size information completely lost during backpropagation. However, it takes too much manpower to scale markings for all the heads precisely. In order to solve the above problems, human head features are scaled, so the human head is divided into four levels by the length of the pixels: greater than 112, greater than 56 and less than 112, greater than 28 and less than 56, and less than 28. Different scale blur processing was used for them.

The scenes contained in the images of the dataset are quite different, the distortion degree of the images is different, and the scale of the human head varies obviously. The details of the data are shown in Table 1. In this case, it is already very challenging to extract head features, and the total number of marked heads is small, which puts forward high requirements for the generalization ability of the network. The dataset is divided into the training set, the validation set, and the test set on a scale of 6 : 2:2.

6.2. Performance Assessment. In this paper, two metrics, namely, mean absolute error (MAE) and mean square error (MSE), are used to evaluate the performance of all considered methods.

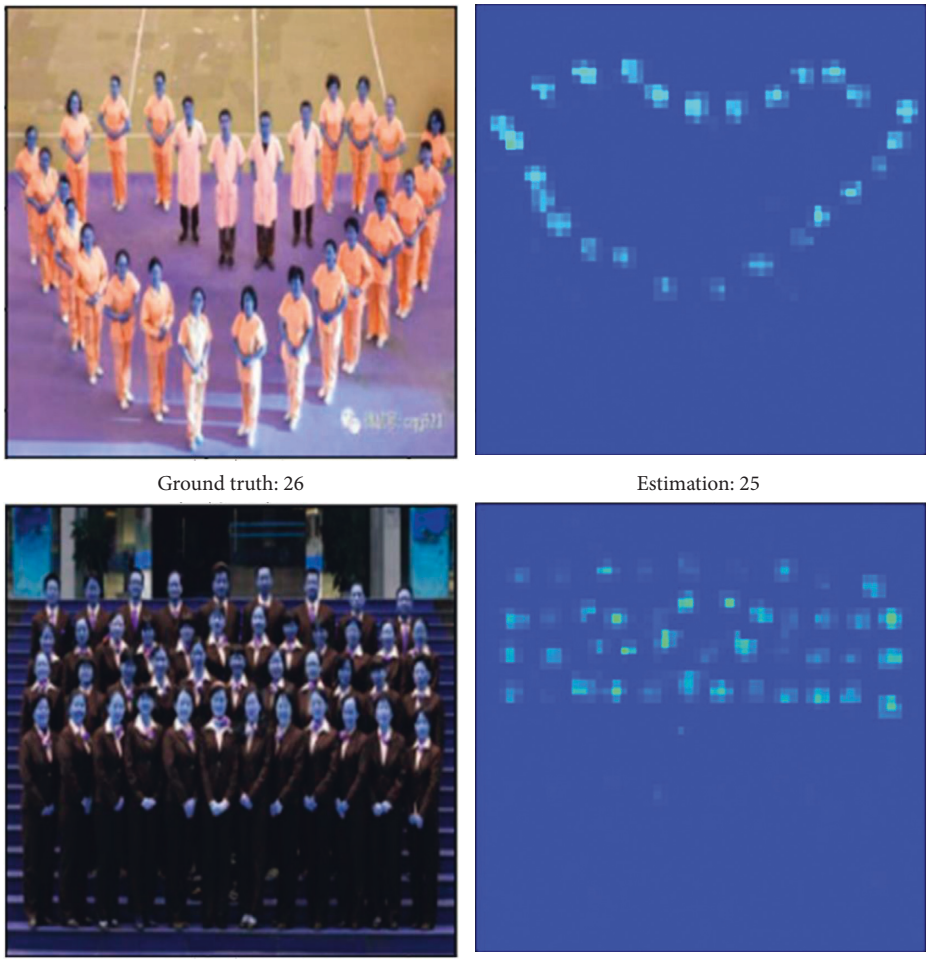


Figure 6: People with the same scale.

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|C_i - \widehat{C}_i\right|, \tag{6}$$

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}\left(C_i - \widehat{C}_i\right)^2, \tag{7}$$

where M is the number of test images, and C_i and $\widehat{C}_i$ are the actual number of people and the estimated number of the i-th image, respectively.

6.3. Experimental Results and Discussion. First, three branches of extracted features are pretrained, and other parts of the entire network are pretrained without changing the pretraining weights, so as to avoid the deviation between the parameters of the pretrained branches and those generated by the other parts. The direction of feature extraction and analysis changes. Finally, the entire pretrained network is trained again to obtain the final training model. The batch size used is 8, the learning rate is 0.0003, and the weight attenuation is 1e-5 (Adam optimizer).

MSCNN [21], MCNN [12], DSA-CNN [22], and DCN [23], were reproduced on the dataset, compared with the multibranch neural network with the compensator to evaluate the experimental results. MSCNN repeatedly passes the input image through a single convolution kernel of different scales and then merges the output results. The interactive use of convolution kernels of different scales can combine many situations of receptive fields to solve the problem of scale change. MCNN also adopts a multibranch structure. Different branches use a larger convolution sum firstly and then use three relatively small convolution kernels. The convolution kernel scales between different branches differ by 2. MCNN is a classic method, and the model is simple and efficient. DSA-CNN adds a 1×1 convolution kernel before the convolution kernels of different scales to form a DSAM, which is used repeatedly in the network to read multiscale features. DCN uses a dual-branch structure in the first half of the network: the first half uses a large-scale convolution kernel for large-scale feature extraction, and the second half outputs the results. The calculation results of different methods are shown in Table 2.

As shown in Table 2, the multibranch neural network with the compensator achieves an accuracy of 7.3 MAE. Because of the perspective effect, the proportion of data with smaller head size in the dataset is larger. Crowd counting

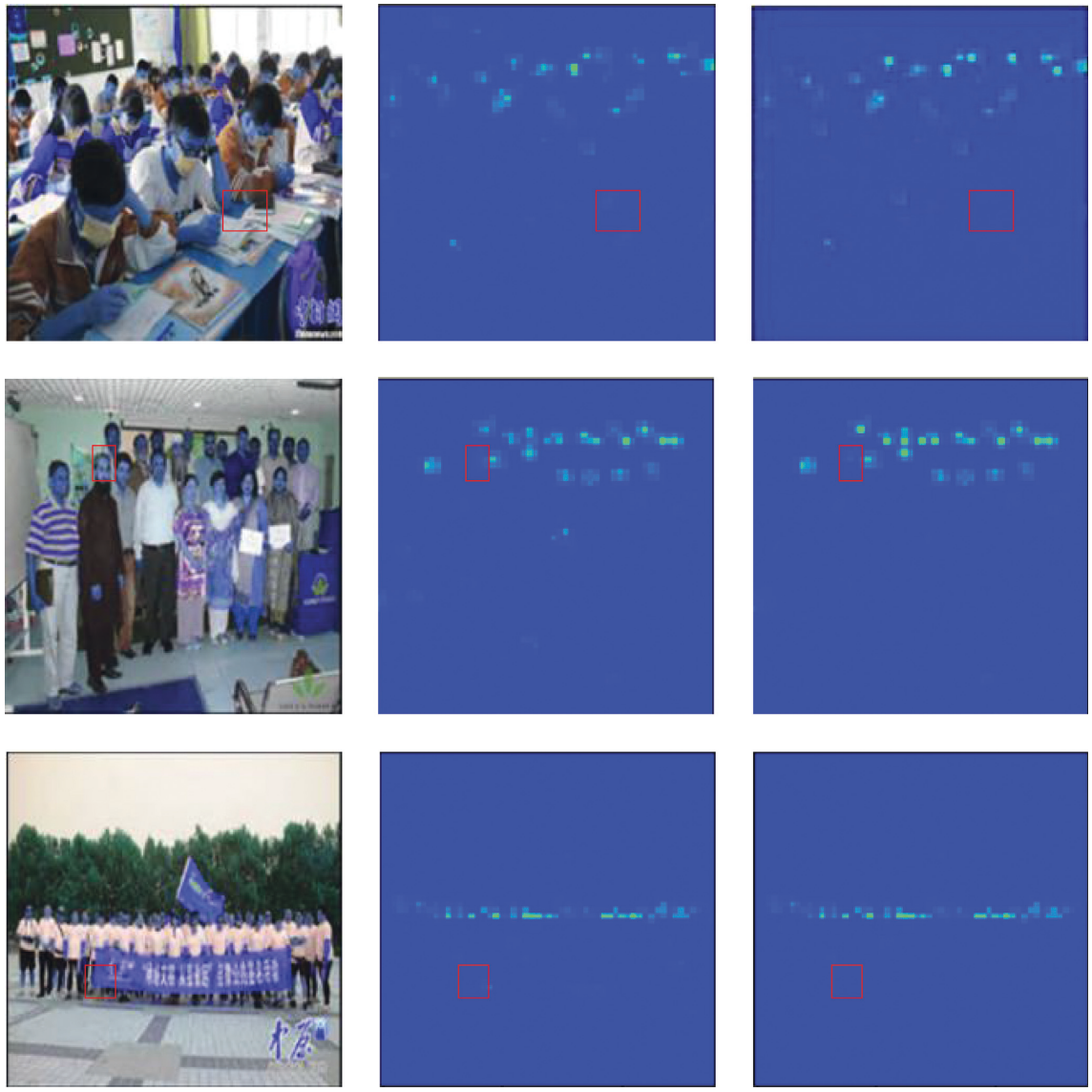


Figure 7: The effect of the compensator.

accuracy for small-scale features has a significant impact on the results. However, the dataset is aimed at people with different scale features, and there are also a certain number of features with larger scales. This requires the network to be able to count objects of different scales. Since MSCNN attaches equal importance to objects of all scales, its accuracy in counting small-scale objects with dense information is insufficient. MCNN is a classic method of crowd counting, and it also has better results for scale changes. The DSAM of DSA-CNN also has good results when counting crowds with large-scale variation. DCN also has excellent performance for two-branch networks of different scales. The network proposed in this paper performs crowd counting at different scales through different branches and pays more attention to small-scale targets with greater small information density in parameter allocation. The compensator module also has some optimization to the network. The hierarchical loss function uses different calculation methods for objects of different scales, which is more helpful for counting objects of different scales. For the detailed effect of each part of the network on the count, there is a detailed analysis later. The effect of counting images with large variation in feature scale is shown in Figure 5. The effect of counting images with similar feature scale sizes is shown in Figure 6. A multi-branch neural network with a compensator can successfully capture and count head features at different scales.

In order to analyze the effectiveness of the compensator and the adaptive grading function, we compare them with the networks without the compensator or the adaptive grading loss function and analyze the results. The MAE value of the network without the compensator module is 7.7, and that of the network with the compensation module is 7.3, which is relatively low. Through the comparison and the results displayed in Figure 7, it can be seen that the weight compensator has two main functions: one is to reduce the misjudgment of human head features as human heads, and

Table 3: AGLoss ablation of our ASNet on the dataset.

	1-level loss	2-level AGLoss	3-level AGLoss
MAE	7.9	7.3	7.5
MSE	306.3	302.4	305.8

Table 4: Branch comparison.

Configuration	MAE	MSE	Loss	RF	Filters	Parameters
3×3	7.6	313.6	0.00018	40	640	924225
5×5	8.1	328.4	0.00022	76	320	642849
7×7	14.2	417.6	0.00031	112	160	316177

the other is to improve the ability to recognize human head targets. Through these two aspects, the network has stronger results. The importance of the compensator is proved, which is helpful for the weight adjustment of the multibranch neural network. The adjustment of network weights also helps each branch of the multibranch neural network structure to better play different roles for targets of different scales.

The computational results of the proposed model with 2-level and 3-level AGLoss are presented in Table 3. It can be seen that AGLoss further improves the counting performance of the network with the compensator. The MAE of the network with the compensator and 2-level AGLoss reaches 7.3, which is better than the network with the compensator with MSE loss and 3-level AGLoss. The level 2 AGLoss has better generalization ability than the level 3 AGLoss. This may be caused by the over-grading of larger human head features, resulting in human head features being segmented in different regions. Therefore, the level 2 AGLoss was finally selected.

For further qualitative analysis, the capabilities of multiple branches are analyzed separately, which is also one of the reasons for the construction of branches. After using individual branches to make predictions and analyzing MAE, RMSE, and loss indicators, the results are shown in Table 4. Small-scale targets occupy a larger proportion in the dataset. The 3×3 branch is more effective than other branches. In addition, due to the large number of large-scale feature pixels, crowd counting only needs simple head information and does not require too much detailed information. Therefore, its information density is lower, and the use of fewer parameters of large-scale convolution kernels can also increase the efficiency of the network.

7. Conclusions

This paper proposed an improved multibranch convolution model for solving crowd counting in complex scenes. In the proposed model, multibranch structure was employed to capture head features of different scale, and an extra compensator was introduced to optimize weights of different branches according to different scale features. The feasibility of the proposed model was verified on a public crowd counting competition dataset. Experimental results showed that the proposed model accurately estimated the number of crowds with different head sizes and shooting distances. Meanwhile, compared with benchmarking methods, namely, MSCNN, MCNN, DSA-CNN, and DCN, the proposed model achieved the best evaluation performance. Furthermore, through the head-scale grading labeling, targets with different scales were optimized via adaptive grading loss function. Therefore, it is promising to utilize the proposed method to count the number of people in complex scenarios [24–27].

Data Availability

The image data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partially supported by National Key R&D Program of China (2018********01) and National Social Science Foundation Project (17BXW065).

References

[1] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[3] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, pp. 833–851, Springer Cham, New York, NY, USA, 2018.

[4] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Venice, Italy, October 2017.

[5] T. Zhang, C. Xu, and M. H. Yang, "Robust structural sparse tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 473–486, 2019.

[6] T. Zhang, C. Xu, and M. H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 365–378, 2019.

[7] Z. Wang, L. Wang, and C. Huang, "A Fast Abnormal Data Cleaning Algorithm for Performance Evaluation of Wind Turbine," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2020.

[8] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2913–2920, Miami, FL, USA, June 2009.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743–761, 2012.

[10] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1299–1302, New York, NY, USA, October 2015.

[11] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proceedings of the Computer Vision - ECCV 2016*, pp. 615–629, Amsterdam, Netherlands, October 2016.

[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 589–597, Las Vegas, NV, USA, June 2016.

[13] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4031–4039, Honolulu, HI, USA, July 2017.

[14] V. Sindagi and V. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1002–1012, Seoul, Republic of Korea, October 2019.

[15] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1941–1950, Waikoloa, HI, USA, January 2019.

[16] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, and X. Bai, "Learn to scale: generating multipolar normalized density maps for crowd counting," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8381–8389, Seoul, Republic of Korea, November 2019.

[17] M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang, "Crowd counting using scale-aware attention networks," in *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1280–1288, Waikoloa, HI, USA, January 2019.

[18] M. Wang, H. Cai, J. Zhou, and M. Gong, "Stochastic multi-scale Aggregation network for crowd counting," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020.

[19] Z. Wang, L. Wang, C. Huang, Z. Zhang, and X. Luo, "Soil-moisture-sensor-based automated soil water content cycle classification with a hybrid symbolic aggregate approximation algorithm," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 14003–14012, 2021.

[20] X. Jiang, L. Zhang, M. Xu et al., "Attention scaling for crowd counting," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4705–4714, Seattle, WA, USA, June 2020.

[21] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 465–469, Beijing, China, September 2017.

[22] J. Wu, W. Qu, H. Yu, Y. Zhou, and Z. Cui, "A novel crowd counting method via deep convolutional neural network," in *Proceedings of the 2019 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 162–167, Tianjin, China, August 2019.

[23] W. Zhang, Y. Wang, Y. Liu, and J. Zhu, "Deep convolution network for dense crowd counting," *IET Image Processing*, vol. 14, no. 4, pp. 621–627, 2020.

[24] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[25] E. Walach and L. Wolf, "Learning to Count with CNN Boosting," in *Computer Vision – ECCV 2016*, pp. 660–676, Springer Verlag Cham, New York, NY, USA, 2016.

[26] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1215–1219, Phoenix, AZ, USA, September 2016.

[27] C. Cong Zhang, H. Hongsheng Li, X. Wang, X. Xiaokang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–841, Boston, MA, USA, June 2015.

Hindawi
Complexity
Volume 2021, Article ID 8329487, 10 pages
https://doi.org/10.1155/2021/8329487

Research Article

Factors Affecting the Adoption of Blockchain Technology in the Complex Industrial Systems: Data Modeling

Yu Chengyue,[1] **M. Prabhu**,[2] **Mahendar Goli**,[3] **and Anoop Kumar Sahu**[4]

[1]*School of Economics and Management, Nanchang University, Nanchang 330031, China*
[2]*Department of Business Administration, College of Administration and Economics, Lebanese French University, Erbil, Kurdistan, Iraq*
[3]*School of Management, Anurag University, Hyderabad, India*
[4]*Department of Mechanical Engineering, School of Studies in Engineering and Technology, Guru Ghasidas Vishwavidyalaya (A Central University), Bilaspur, Chhattisgarh, India*

Correspondence should be addressed to Mahendar Goli; mahendar.sm@gmail.com

Received 1 June 2021; Revised 21 November 2021; Accepted 6 December 2021; Published 24 December 2021

Academic Editor: Long Wang

Nowadays, Blockchain Technology (BCT) is contributing toward addressing the challenges of complex industrial systems (CISs). The BCT reduces the complexity of cash data storage as well as retrieval system of finance, marketing, supply chain, inventory, and other departments. The objective of the present study is to investigate the factors, which affect the intention of professionals to adapt the BCT in the CISs by using an extension of the technology acceptance model. To fulfill the research objective, a theoretical research model is constituted by multiple hypotheses (H1–H6), i.e., perceived usefulness, perceived ease of use, perceived innovativeness, knowledge, risk, and trust after conducting the relevant literature survey in the context of BCT. Next, each hypothesis is tested by exploring the survey data of a sample of 287 professionals of different BCT user's companies such as retailing, e-commerce, manufacturing, and construction. Survey data is analyzed by executing the structural equation modeling with AMOS software. The factors and latent constructs loadings, reliability, convergent, discriminant, model fit-measurement, structural model, and the path analysis are conducted. The results reveal that the H1, H2, and H4–H6 dropped the positive impact and effect on professionals' intention to use the BCT in CISs. But, H3 has no effect for enhancing the intention of professionals to use BCT.

1. Introduction

Disruptions in technology brought phenomenal changes in the function of CISs. The supply chain (SC) plays a critical role in enhancing the effectiveness as well as productivity of CISs. Firms using emerging and radical technologies in performing operations can outperform their competitors. Blockchain technology (BCT) is one such emerging technology, which brings a competitive advantage and enables the CISs to function smoothly. BCT is implemented for transactional databases, carried out between the consensuses of equal independent parties [1]. BCT is a "digital, immutable, distributed ledger that chronologically records transactions in near real time" making the processes simpler, efficient, transparent, and secured [2]. BCT differs from the existing technologies in four ways: nonlocalization (decentralization), security, auditability, and smart execution [3], and it has diverse applications in the CISs. Despite many advantages of BCT in different practices of CISs, the adoption of blockchain technology is limited [4]. BCT is still in its infancy, which has to overcome the many barriers—behavioral, organizational, technological, or policy for its adoption in the CISs as per [5, 6].

A few research studies are organized on the BCT acceptance models in different industries, where most of the published research documents dealt with security concerns of BCT. It is found that insufficient research documents are published for auditing attention of professionals in the adaptation of the BCT in CISs. Therefore, to address the above concerns, the current study focussed on examining

the factors, which can influence the attention of professionals in the adoption of BCT in today's CISs. Bag et al. [7], Grzegorczyk [8], Hsiao [9], Dalmarco et al. [10], Kouaib and Almulhim [11], Shi and Wang [12], Patanakul and Rufo-McCarron [13], Sahu et al. [14–16], and Sahu et al. [17] probed that multivariable analysis helps to examine and improve the performance of BCT. The technology acceptance model posited by Davis [18] is one of the widely adopted theories to describe an individual's behavior toward new technologies. It is found as one of the most appropriate models to examine an individual's desire and readiness toward the usage of a technology [19]. Two key antecedents, namely, perceived usefulness and perceived ease of use, explained the users' behavior toward BCT as per Patanakul and Rufo-McCarron [13].

It is seen that the factors such as perceived usefulness [18], perceived ease of use, [18, 20], innovativeness [21], trust, motivation [22], and risk [23] are found as the most significant factors to investigate the behaviors of professionals toward adapting the BCT in CISs. It is also observed that the technology acceptance model (TAM) can be extended with knowledge or expert data [24]. Blockchain technology (BCT) is relatively a new technology, and its acceptance is influenced by factors such as ease of use, usefulness, and risk. The technology acceptance model is the suitable framework to understand users' acceptance of BCT in CISs. The current researches focussed on investigating the effects of factors, motivating the professionals to adapt the BCT in ICSs by using the technology acceptance model [25–27].

The aim of the present study is to evaluate the influence of factors and its hypothesis to gain the intention of professionals to adopt the BCT in the CISs. To examine the outcomes, the authors plan to use structural equation modeling with AMOS software. A list of contributions toward framing the aims of the study is depicted as follows:

(i) To frame a new theoretical research model based on the reference of the technology acceptance model.

(ii) To construct the new theoretical research model by factors and its hypothesis via conducting extant literature survey in the area of BCT adoption in CISs.

(iii) To conduct reliabe, convergent, and discriminant tests for assessing the validity of the model.

(iv) To assess the model fitness for measurement and also assess the structure of the model.

(v) To conduct the path analysis to conclude the influence of factors and its hypothesis to gain the intention of professionals to adopt the BCT in the CISs.

The research study is organized into the following sections. Section 2 (theoretical background, research model, and hypotheses development) includes the following subsections: growth of BCT usage (Section 2.1), theoretical foundation-technology acceptance model (Section 2.2), theoretical research model (Section 2.3), and hypotheses development (Section 2.4). The research method (Section 3) is introduced with its associated subsections, i.e., measures (Section 3.1) and materials (Section 3.2). Section 4 deals with data analysis and model fit test, where reliability and validity assessment (Section 4.1) includes the reliability test (Section 4.1.1), convergent validity (Section 4.1.2), and discriminant validity (Section 4.1.3). Model tests include the model fit-measurement model (Section 4.2), model fit-structural model (Section 4.3), and path analysis results (Section 4.4). At last, discussion (Section 5), implications and contribution of the study (Section 6), and conclusion and future research directions (Section 7) are introduced.

2. Theoretical Background, Research Model, and Hypotheses Development

2.1. Growth of Blockchain Technology (BCT) Usage. BCT has a profound impact on business operations. The BCT is a distributed and highly secure platform, ledger, or database of values—everything from money, assets, stocks, bonds, intellectual property, and deeds, to music, art, and even votes [28, 29]. BCT is a "digital, immutable, distributed ledger that chronologically records transactions in near real time" making the processes simpler, efficient, transparent, and secured [2]. It is a potential technology applied in diverse industries for improved operational efficiency. The BCT-based applications cover the supply chain, finance, e-commerce transactions, product traceability, user credits, financial services, trust systems, new energy, etc. [30]. For instance, BCT has its applications in the areas of tourism for managing ticket booking and loyalty programs [31], data privacy, security and sharing in healthcare [32], and financial services [33]. It is vital for organizations to adopt BCT for improved efficiency and performance in complex industrial systems.

2.2. Theoretical Foundation-Technology Acceptance Model. The technology acceptance model proposed by [18] provides the conceptual framework for the research on the adoption of BCT. TAM serves as a model to understand the user behavior toward the acceptance of new technologies and information systems. It is based on the premise that individuals use certain technologies to derive benefits from the usage. According to TAM, usage attitude is based on two major predictors: perceived ease of use and perceived usefulness [18]. Perceived ease of use is "the extent to which use of the technology is thought to be easy and effortless" [34] whereas perceived usefulness is the "degree to which use of the technology is thought to be useful and helpful" [34]. The technology acceptance model has been widely validated in the context of mobile shopping [35], social media usage [36], and in-store technologies [37]. Though TAM is a suitable framework to adopt new technologies, the adoption of blockchain is critical at the organizational level. BCT is relatively a new technology and its adoption is influenced by certain factors. Therefore, the present study extends TAM with knowledge [24], innovativeness [21], trust, motivation [22], and risk [23] for a better explanation and understanding of users' behavior.

2.3. Theoretical Research Model. The current research focussed on investigating the factors, affecting the attentions or behaviors of industrial professionals toward adopting the BCT in CISs. To attain the same, the research study proposes a theoretical research model constituted by hypotheses H1–H6 such as perceived innovativeness, knowledge, risk, and trust. The model is built by using the foundation of the technology acceptance model proposed by Davis [18]. The theoretical research model is depicted in Figure 1 where hypotheses H1–H6 are displayed.

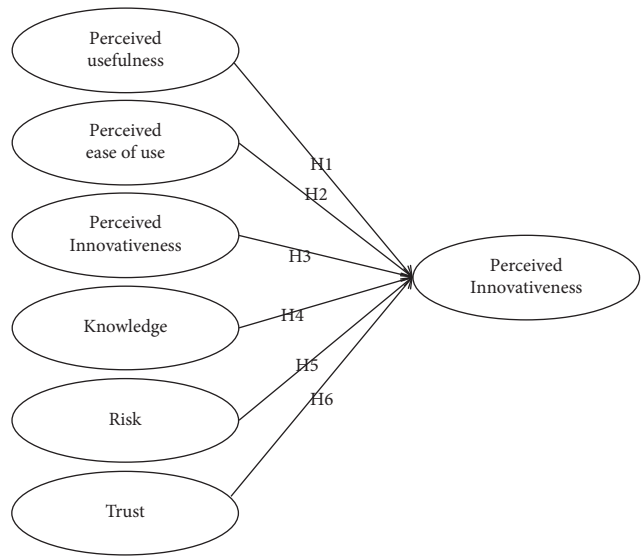


Figure 1: Theoretical research model.

2.4. Hypotheses Development

2.4.1. Perceived Usefulness. Perceived usefulness (PU) is found as a key influencer in determining the acceptance of technology by an individual. PU is conceptualized as "the subjective perspective of users about the specific merit application of system/technology that may either increase or decrease the job performance of users" [18]. Users tend to investigate new technology to ascertain if it will augment his/her job or activity performance. This investigation helps to develop a perception of the technology with respect to performance enhancement. They will continue to use the application if and only if there is no dissonance between perception and experience. Researchers established a significant relationship between the perceived usefulness and behavioral intention to use technology in the case of online shopping [21], e-government learning [38], mobile banking [22], online banking [39], and hotel tablet applications [40]. The positive relationship between perceived usefulness and intention to adopt BCT is confirmed by Kamble et al. [41] and Nuryyev et al. [42]. Therefore, we hypothesize the following:

H1: Perceived usefulness has a positive effect on the intention to use BCT.

2.4.2. Perceived Ease of Use. Perceived ease of use is one of the key exogenous constructs proposed in the technology acceptance model, which influences the user acceptance of a specific information system/technology. It is defined as "the degree to which the specific technology will be free from physical or psychological effort" [18, 20]. In the present research, perceived ease of use to the extent that users are free from physical or psychological effort to use a specific technology is considered. Previous research showed a significant relationship between perceived ease of use and user intention to use technologies in mobile banking [22], online banking [39], hotel tablet applications [40], and agricultural technology [43]. In the case of BCT, a significant relationship between perceived ease of use and intention to adopt is recently audited by Nuryyev et.al. [42]. Therefore, we hypothesize the following:

H2: Perceived ease of use has a positive effect on the intention to use BCT.

2.4.3. Perceived Innovativeness. Innovativeness is a personality trait that indicates an individual's intention to try new things [44] and a desire to be different [45]. Innovativeness is a key determinant to adopt emerging technologies. Research studies proved a positive relationship between users' perceived innovativeness and behavioral intention toward cloud classrooms [46], remote mobile payment services [47], mobile Internet [48], and mobile diet apps [49]. Nuryyev et al. [42] endorsed the relationship between perceived innovativeness and intention to use BCT. Therefore, we hypothesize the following:

H3: Perceived innovativeness has a positive effect on the intention to use BCT.

2.4.4. Knowledge. Knowledge is defined as "awareness, consciousness or familiarity gained by experience or learning" [50]. Users' knowledge of the product is a prerequisite for understanding and using it. In the context of technology, knowledge refers to the expertise and skills, gained to understand the usage of a specific technology. Research studies asserted a significant positive relationship between knowledge and intention in the case of renewable energy [51] and website usage [24]. It is imperative to have know-how knowledge of BCT to use it at optimum level. Knauer and Mann [52] found a positive relationship between knowledge and behavioral intention to use BCT. Therefore, we hypothesize the following:

H4: Knowledge has a positive effect on the intention to use BCT.

2.4.5. Risk. Mandrik and Bao [53] defined risk as "feelings of uncertainty or anxiety about the behaviour and the seriousness or importance of the possible negative outcomes of

that behaviour". Further, it is described as "a user's belief in the potential uncertain negative outcomes of using a product" [54]. Risk is one of the major reasons individuals avoid adopting new technology and it reduces the adoption intention (Chen [55]). Research revealed a negative association between individuals' perceived risk and adoption of e-commerce [56], mobile banking services [57], and remote mobile payment services [47]. Researchers found a negative relationship between risk and intention in the case of BCT [58, 59]. Therefore, we hypothesize the following:

H5: Risk has a negative effect on the intention to use BCT.

2.4.6. Trust. Trust is conceptualized as "existing when one party has confidence in the exchange partner's reliability and integrity" [60]. Trust is a key determinant of consumer behavior [61]; hence, it is crucial to develop user trust in a product, service, or technology. The concept of trust has gained importance to predict an individual's behavioral intention [62]. A number of studies indicated a positive relationship between trust and user intention regarding online travel purchase [63], e-commerce [64], mobile financial services [14], and remote mobile payment services [47]. In the case of BCT, trust exhibits a positive relationship with the behavioral intention [59, 65]. Therefore, we hypothesize the following:

H6: Trust has a positive effect on the intention to use BCT.

3. Research Method

3.1. Measures. The research has the aim to examine the factors that influence the behavior or attention of professionals about the adoption of BCT for addressing the supply chain operations of CISs. The measurement items for the current research are adapted from the previous researches. These items are modified to suit the context of BCT. The measurement items for both perceived usefulness and perceived ease of use are adapted from Childers et al. [66]. Perceived innovativeness's items are taken from Agarwal and Prasad [67]. Three items of knowledge factor are extracted from Golnaz et al. [15]. The measuring statements for the risk factor are obtained from Lu et al. [68]. Trust is measured with three items, which are drafted from Suh and Han [69]. Finally, behavioral intention is examined based on the technology acceptance model by Davis [18].

All the measurement items are examined by using a five-point Likert scale ranging from 1-strongly disagree to 5-strongly agree. A pilot test with 30 respondents is conducted to test validity. Industry experts using BCT to address the supply chain operations are consulted to ensure the content validity of the measurement instrument.

3.2. Materials. Supply chain managers/professionals using BCT in various industries such as retailing, e-commerce, and manufacturing are contacted via a web-based survey engine or instrument for collecting data. Respondents are consulted based on purposive sampling mode from different parts of India. Online questionnaires are shared with 700 employees, out of which 315 responses are returned/recorded. 28 questionnaires are incomplete. Finally, 287 usable samples were received with a 41 percent response rate for data analysis.

4. Data Analysis

4.1. Reliability and Validity Assessment

4.1.1. Reliability Test. Reliability of the constructs/factors is assessed based on Cronbach's alpha and composite reliability (CR) test, which used the surveyed sample data of latent constructs, shown in Table 1. Internal consistency of the constructs is measured using Cronbach's alpha, which should be beyond 0.70 as suggested by Hair et al. [16]. From Table 2, it is noted that Cronbach's alpha and composite reliability of the constructs exceed the cut-off value, which ensures the reliability of the constructs/factors.

4.1.2. Convergent Validity. Convergent validity is conducted by using the average variance extracted (AVE) test on the same surveyed sample data of latent constructs. For all the latent constructs, AVE values are above the threshold value, i.e., 0.60 as referred by Hair et al. [16]. From Table 2, AVE values calculated for all the constructs are more than 0.60 and thus well confirmed the convergent validity of constructs/factors.

4.1.3. Discriminant Validity. It is evaluated based on the shared variance between the factors [70]. Discriminant validity is confirmed if the square root of the average variance extracted is more than the correlation between the factor/construct and other factors/constructs. From Table 3, the diagonal values (square root of AVE) are more than the off-diagonal values (correlations between the factor/construct and other factors/constructs are confirmed). Hence, we can assure the discriminant validity of the constructs/factors.

4.2. Model Fit-Measurement Model. The measurement model is fit or not, which is tested after the latent constructs met the criteria for reliability and validity (convergent and discriminant validity) assessment. Structural equation modeling (SEM) tool under AMOS software is employed to test the fitness of indices. Model fit was examined by using Chi-Square/D*f* (CMID/d*f*), Root Mean Square Residual (RMR), Root Mean Square of Error Approximation (RMSEA), Comparative Fit Index (CFI), Normed Fit Index (NFI), Tucker-Lewis Index (TLI), and Goodness of Fit (GIF).

From Table 4, all the fit indices meet the criteria suggested by Bentler [71], Hu and Bentler [72], Hair et al. [16], Kim and Sundar [73], and Henseler et al. [74], which ensures that measurement model is fit. Table 5 has shown the measurement statements and sources of the model.

Table 1: Industry and work experience of the employees.

Work experience and industry ($n = 287$)	Frequency	Percentage
Work experience in the organization		
Less than one year	7	2.44
1-2 years	18	6.27
2–5 years	107	37.28
5–10 years	121	42.16
Above 10 years	34	11.85
Industry		
Construction	16	5.57
Manufacturing	95	33.10
Retailing	47	16.38
E-commence	65	22.65
Transport and storage	19	6.62
Information and communication services	38	13.24
Others	7	2.44

Table 2: Reliability and average variance extracted.

Factor	Cronbach's alpha	Composite reliability	Average variance extracted
Perceived usefulness	0.907	0.909	0.770
Perceived ease of use	0.933	0.934	0.826
Perceived innovativeness	0.860	0.864	0.680
Knowledge	0.923	0.926	0.808
Risk	0.918	0.919	0.791
Trust	0.918	0.848	0.741
Intention to use	0.907	0.909	0.770

Table 3: Intercorrelation matrix.

	PU	TR	PI	PEO	KN	RI	ITU
PU	0.877						
TR	−0.100	0.861					
PI	−0.072	0.366	0.824				
PEO	−0.030	0.293	0.258	0.908			
KN	−0.055	0.101	0.462	0.132	0.898		
RI	0.010	−0.036	0.289	0.034	0.481	0.889	
ITU	0.073	0.466	0.264	0.443	0.152	−0.058	0.877

Note. PU: perceived usefulness, TR: trust, PI: perceived innovativeness, PEO: perceived ease of use, KN: knowledge, RI: risk, ITU: intention to use.

Table 4: Measurement model.

	CMID/df	RMR	RMSEA	CFI	NFI	TLI	GFI
Cut-off value	<3	<0.5	<0.08	>0.9	>0.9	>0.9	>0.8
Actual value	1.317	0.025	0.033	0.989	0.957	0.986	0.935

4.3. Model Fit-Structural Model. After confirming the fitness of the measurement model, the model structure is assessed for its fitness by using the structural model. The test is conducted for the same CMID/df, RMR, RMSEA, CFI, NFI, TLI, and GFI indices. Table 6 shows that all the values are well above the threshold values referred by Bentler [71], Hu and Bentler [72], Hair et al. [16], Kim and Sundar [73], and Henseler et al. [74]. Hence, the structural model is found fit.

4.4. Path Analysis Results. It is observed from Table 7 that H1-perceived usefulness ($\beta = 0.133$, $p = 0.017$) and H2-perceived ease of use ($\beta = 0.309$, p=***) have the most significant influence on the intention of professionals to adopt BCT, whereas H3-perceived innovativeness has no influence on intention. H4-knowledge ($\beta = 0.139$, $p = 0.024$) and H6-trust ($\beta = 0.323$, p=***) exhibited positive influence, and H5-risk ($\beta = 0.16$, $p = 0.021$) proved to be a negative influencer on the intention to use BCT. Except for H3, all hypotheses, i.e., H1, H2, H4, H5, and H6, are accepted.

As discussed, Table 1 dealt with industries as well as the work experience of the employees, which are invited to address the survey questionnaires. 287 samples are received for data analysis. Table 8 depicted the factors and latent constructs loadings analysis, which confirmed the relevancy of factors/main construct and latent constructs. Table 2 showed the reliability analysis, which confirmed the reliability of factors. Table 3 confirmed the intercorrelation between the factors/constructs. Table 4 depicted the discriminant validity and confirmed the variances between the factors. Next, model fit-measurement and structural model are analyzed, where, in Table 4, the model fit-measurement confirmed the validity of the hypothesis and, in Table 6, confirmed the structure of the theoretical research model. Eventually, Table 7 exhibited the positive and negative

Table 5: Measurement statements and sources.

Factors/constructs	Latent constructs	Latent constructs (items) loading
Perceived usefulness	Usage of BCT improves the productivity. BCT is useful. BCT improves the effectiveness.	Childers et.al. [66]
Perceived ease of use	It is easy to understand BCT. It is easy to use BCT. Use of BCT does not require a lot of mental effort.	Childers et al. [66]
Perceived innovativeness	I like to experiment with BCT. In general, I would not hesitate to try out BCT. I would look for ways to experiment with BCT.	Agarwal and Prasad [67]
Knowledge	I understand BCT. I have sufficient knowledge about BCT. I have enough knowledge about BCT.	Golnaz et al. [15]
Risk	I do not feel very safe using BCT. I am worried about using BCT. I do not feel secure using BCT.	Lu et al. [68]
Trust	BCT is trustworthy. I trust in the benefits of BCT. I trust BCT.	Suh and Han [69]
Intention to use	I plan to use BCT in the future. I intend to use BCT in the future. I predict I will use BCT in the future.	Davis [18]

Table 6: Structural model.

	CMID/d*f*	RMR	RMSEA	CFI	NFI	TLI	GFI
Cut-off value	<3	<0.5	<0.08	>0.9	>0.9	>0.9	>0.8
Actual value	2.289	0.106	0.067	0.952	0.918	0.945	0.873

Table 7: Hypotheses testing.

Hypothesis	Estimate	SE	CR	*p* value	Result
H1: Perceived usefulness ⟶ intention to use	0.133	0.056	2.382	0.017	Accepted
H2: Perceived ease of use ⟶ intention to use	0.309	0.053	5.889	***	Accepted
H3: Perceived innovativeness ⟶ intention to use	0.048	0.056	0.859	0.391	Not accepted
H4: Knowledge ⟶ intention to use	0.139	0.062	2.257	0.024	Accepted
H5: Risk ⟶ intention to use	−0.16	0.069	−2.306	0.021	Accepted
H6: Trust ⟶ intention to use	0.323	0.049	6.565	***	Accepted

Table 8: Factor-latent constructs loadings.

Factors/constructs	Latent constructs	Latent constructs (items) loading
Perceived usefulness	Usage of BCT improves productivity. BCT is useful. BCT improves the effectiveness.	0.933 0.931 0.882
Perceived ease of use	It is easy to understand BCT. It is easy to use BCT. Use of BCT does not require a lot of mental effort.	0.897 0.919 0.917
Perceived innovativeness	I like to experiment with BCT. In general, I would not hesitate to try out BCT. I would look for ways to experiment with BCT.	0.817 0.858 0.837
Knowledge	I understand BCT. I have sufficient knowledge about BCT. I have enough knowledge about BCT.	0.899 0.891 0.857

Table 8: Continued.

Factors/constructs	Latent constructs	Latent constructs (items) loading
Risk	I do not feel very safe using BCT.	0.900
	I am worried about using BCT.	0.906
	I do not feel secure using BCT.	0.890
Trust	BCT is trustworthy.	0.901
	I trust in the benefits of BCT.	0.841
	I trust BCT.	0.915
Intention to use	I plan to use BCT in the future.	0.850
	I intend to use BCT in the future.	0.874
	I predict I will use BCT in the future.	0.894

attention of professionals/employees to adopt the BCT in addressing the supply chain operations of CISs.

5. Discussion

The research work investigated the subjective perception of employees of various firms against six critical factors/constructs, affecting the intention to use to adopt the BCT in addressing the supply chain operations of CISs. A theoretical research model based on the technology acceptance model is framed and tested using empirical data. The factors/constructs studied in the research work are perceived usefulness, perceived ease of use, perceived innovativeness, knowledge, risk, trust, and intention to use. The findings of the research indicate that perceived usefulness has shown a significant positive effect on the intention to use BCT (H1), in line with the research by Sharma [22] and Kamble et al. [41]. This shows that the usefulness of BCT is requisite to adopt the BCT. The perceived ease of use is positively related to intention to use (H2), consistent with the research studies by Kim [40] and Nuryyev et al. [42]. The perceived innovativeness has no significant effect on the intention to use BCT (H3). The results may be due to the small size of the sample and responses are drawn from various industries. Next, a positive relationship between knowledge toward the BCT and the intention to use it (H4) is found, which supports the results of the research performed by Bang et al. [51] and Knauer and Mann [52]. Risk is negatively related to the intention to use BCT (H5), similar to the results of Slade et al. [47] and Guych et al. [58]. Trust has a positive effect to draw intention to use (H6). It is evident from the hypothesis results in Table 7 that all the hypotheses (H1, H2, H4, H5, and H6) are accepted except H3.

6. Implications and Contribution of the Study

The research advances the literature in the field of BCT adoption for taking care of supply chain operations of CISs. The research tried to bridge the gap between the BCT and the adoption of the technology acceptance model. As a part of the contribution, a theoretical research model is constituted by multiple hypotheses (H1–H6), i.e., perceived usefulness, perceived ease of use, perceived innovativeness, knowledge, risk, and trust after conducting the relevant literature survey in the context of BCT. Next, each hypothesis is tested by exploring the survey data of a sample of 287 professionals of different BCT user's companies. To test each hypothesis to use BCT, as discussed from Tables 1–4 and Tables 6–8, factors as latent constructs loadings, reliability, convergent, discriminant validity, model fit-measurement, model fit-structural model, and path analyses are conducted to audit the positive and negative attention of professionals/employees to adopt the BCT in addressing the supply chain operations of CISs in today's industry 4.0. As a part of implications, the managers can adopt the presented work to investigate the positive and negative attention of employees of an individual or specific firm toward adopting the BCT and other advanced technologies, i.e., PayPal, Google Pay, Paytm, etc.

7. Conclusion and Future Research Directions

The conducted study proposed a theoretical research model constituted by multiple hypotheses (H1–H6), i.e., perceived usefulness, perceived ease of use, perceived innovativeness, knowledge, risk, and trust. The factors and latent constructs loadings, reliability, convergent, discriminant, model fit-measurement, and structural model are conducted. Eventually, the path analysis tested the constructs/factors, namely, perceived usefulness, perceived ease of use, perceived innovativeness, knowledge, risk, trust, and intention to use BCT with using the foundation on technology acceptance model. Perceived ease of use and perceived usefulness were found to be key predictors for the adoption of BCT in addressing the supply chain operations of CISs [75]. The study established the positive effect of H1, H2, and H4–H6 on professionals' intention to use the BCT in CISs, while H3 has no effect for enhancing the intention of professionals to use BCT.

The research has certain caveats that could be considered for future research. The research was conducted in India, the geographical limitation may affect the ability of the research, further studies may incorporate in other countries, and cultural differences could be tested. The research was cross-sectional and quantitative, and qualitative studies may produce better insights. The research is confined to a few select industries using blockchain technology in complex industrial systems. The research did not include attitude from the technology acceptance model; future research may include cost, hedonic value, and attitude as predictors of behavioral intention. Further research may test the moderating role of user experience in the adoption of blockchain

technology. The research focussed only on blockchain technology, and future research may integrate the Internet of things with blockchain technology.

Data Availability

The data used to support the findings of this study are available in Table 7.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

References

[1] T. Aste, P. Tasca, and T. Di Matteo, "Blockchain technologies: the foreseeable impact on society and industry," *Computer*, vol. 50, no. 9, pp. 18–28, 2017.

[2] T. Shah and S. Jani, *Applications of Blockchain Technology in Banking & Finance*, ParulCUniversity, Vadodara, India, 2018.

[3] J. Baker and J. Steiner, *Blockchain: The Solution for Transparency in Product Supply Chains*, Provenance, London, UK, 2015.

[4] H. Min, "Blockchain technology for enhancing supply chain resilience," *Business Horizons*, vol. 62, no. 1, pp. 35–45, 2019.

[5] M. Crosby, P. Pattanayak, S. Verma, and V. Kalyanaraman, "Blockchain technology: beyond bitcoin," *Applied Innovation*, vol. 2, no. 6-10, p. 71, 2016.

[6] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, "Where is current research on blockchain technology?-a systematic review," *PLoS One*, vol. 11, no. 10, Article ID e0163477, 2016.

[7] S. Bag, D. A. Viktorovich, A. K. Sahu, and A. K. Sahu, "Barriers to adoption of blockchain technology in green supply chain management," *Journal of Global Operations and Strategic Sourcing*, vol. 14, no. 1, 2020.

[8] T. Grzegorczyk, "Managing intellectual property: strategies for patent holders," *The Journal of High Technology Management Research*, vol. 31, no. 1, 2020.

[9] H. M. Hsiao, "Mobile payment services as a facilitator of value co-creation: a conceptual framework," *The Journal of High Technology Management Research*, vol. 30, no. 2, 2019.

[10] G. Dalmarco, F. R. Ramalho, A. C. Barros, and A. L. Soares, Providing industry 4.0 technologies: the case of a production technology cluster," *The Journal of High Technology Management Research*, vol. 30, no. 2, Article ID 100355, 2019.

[11] A. Kouaib and A. Almulhim, "Earnings manipulations and board's diversity: the moderating role of audit," *The Journal of High Technology Management Research*, vol. 30, no. 2, Article ID 100356, 2019.

[12] Z. Shi and G. Wang, "Integration of big-data ERP and business analytics (BA)," *The Journal of High Technology Management Research*, vol. 29, no. 2, pp. 141–150, 2018.

[13] P. Patanakul and R. Rufo-McCarron, "Transitioning to agile software development: lessons learned from a government-contracted program," *The Journal of High Technology Management Research*, vol. 29, no. 2, pp. 181–192, 2018.

[14] H. Chemingui, "Resistance, motivations, trust and intention to use mobile financial services," *International Journal of Bank Marketing*, vol. 31, no. 7, 2013.

[15] R. Golnaz, M. Zainalabidin, S. Mad Nasir, and F. Eddie Chiew, "Non-Muslims awareness of Halal principles and related food products in Malaysia," *International Food Research Journal*, vol. 17, pp. 667–674, 2010.

[16] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis: A Global Perspective*, Pearson Prentice Hall, Upper Saddle River, NJ, 7th ed. edition, 2010.

[17] A. K. Sahu, N. K. Sahu, and A. K. Sahu, "Appraisements of material handling system in context of fiscal and environment extent: a comparative grey statistical analysis," *International Journal of Logistics Management*, vol. 28, no. 1, pp. 1–30, 2017.

[18] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[19] M. G. M. Johar and J. A. A. Awalluddin, "The role of technology acceptance model in explaining effect on e-commerce application system," *International Journal of Managing Information Technology*, vol. 3, no. 3, pp. 1–14, 2011.

[20] F. D. Davis, "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts," *International Journal of Man-Machine Studies*, vol. 38, no. 3, pp. 475–487, 1993.

[21] Z. Wei, M. Y. Lee, and H. Shen, "What drives consumers in China to buy clothing online? Application of the technology acceptance model," *Journal of Textiles and Fibrous Materials*, vol. 1, Article ID 2515221118756791, 2018.

[22] S. K. Sharma, "Integrating cognitive antecedents into TAM to explain mobile banking behavioral intention: a SEM-neural network modeling," *Information Systems Frontiers*, vol. 21, no. 4, pp. 815–827, 2019.

[23] J. Li, Q. Ma, A. H. Chan, and S. S. Man, "Health monitoring through wearable technologies for older adults: smart wearables acceptance model," *Applied Ergonomics*, vol. 75, pp. 162–169, 2019.

[24] D. Manika, D. Gregory-Smith, and S. Papagiannidis, "The influence of prior knowledge structures on website attitudes and behavioral intentions," *Computers in Human Behavior*, vol. 78, pp. 44–58, 2018.

[25] A. K. Sahu, A. K. Sahu, and N. K. Sahu, "Knowledge based decision support system for appraisement of sustainable partner under fuzzy cum non-fuzzy information, Kybernetes," *The international journal of cybernetics, systems and management sciences*, vol. 47, no. 6, pp. 1090–1121, 2018a.

[26] A. K. Sahu, N. K. Sahu, A. K. Sahu, and M. S. Rajput, "Grey-based scorecard model for opting fruit supply bazaar locality under advanced chain of macro-micro parameter," *British Food Journal*, vol. 120, no. 1, pp. 59–79, 2018b.

[27] N. K. Sahu, A. K. Sahu, A. K. Sahu, and A. K. Sahu, Cluster approach integrating weighted geometric aggregation operator to appraise industrial robot," *Kybernetes*, vol. 47, no. 3, pp. 487–524, 2018c.

[28] D. Tapscott and A. Tapscott, *Blockchain Revolution: How the Technology behind Bitcoin Is Changing Money, Business, and the World*, Penguin, New York, USA, 2016.

[29] M. Swan and P. De Filippi, "Toward a philosophy of blockchain: a symposium: introduction," *Metaphilosophy*, vol. 48, no. 5, pp. 603–619, 2017.

[30] X. Zhu and D. Wang, "Research on blockchain application for E-commerce, finance and energy," *IOP Conference Series: Earth and Environmental Science*, vol. 252, no. 4, Article ID 042126, 2019.

[31] M. Kizildag, T. Dogru, T. C. Zhang et al., "Blockchain: a paradigm shift in business practices," *International Journal of Contemporary Hospitality Management*, vol. 32, no. 3, pp. 953–975, 2019.

[32] M. A. Engelhardt, "Hitching healthcare to the chain: an introduction to blockchain technology in the healthcare sector," *Technology Innovation Management Review*, vol. 7, no. 10, 2017.

[33] K. Fanning and D. P. Centers, "Blockchain and its coming impact on financial services," *Journal of Corporate Accounting & Finance*, vol. 27, no. 5, pp. 53–57, 2016.

[34] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies," *Management Science*, vol. 46, no. 2, pp. 186–204, 2000.

[35] A. Shukla and S. K. Sharma, "Evaluating consumers' adoption of mobile technology for grocery shopping: an application of technology acceptance model," *Vision: The Journal of Business Perspective*, vol. 22, no. 2, pp. 185–198, 2018.

[36] R. Rauniar, G. Rawski, J. Yang, and B. Johnson, "Technology acceptance model (TAM) and social media usage: an empirical study on Facebook," *Journal of Enterprise Information Management*, vol. 27, no. 1, 2014.

[37] S. K. Roy, M. S. Balaji, A. Quazi, and M. Quaddus, "Predictors of customer acceptance of and resistance to smart technologies in the retail sector," *Journal of Retailing and Consumer Services*, vol. 42, pp. 147–160, 2018.

[38] S. H.-P. Shyu and J.-H. Huang, "Elucidating usage of e-government learning: a perspective of the extended technology acceptance model," *Government Information Quarterly*, vol. 28, no. 4, pp. 491–502, 2011.

[39] A. George and G. S. G. Kumar, "Antecedents of customer satisfaction in internet banking: technology acceptance model (TAM) redefined," *Global Business Review*, vol. 14, no. 4, pp. 627–638, 2013.

[40] J. S. Kim, "An extended technology acceptance model in behavioral intention toward hotel tablet apps with moderating effects of gender and age," *International Journal of Contemporary Hospitality Management*, vol. 28, no. 8, pp. 1535–1553, 2016.

[41] S. Kamble, A. Gunasekaran, and H. Arha, "Understanding the Blockchain technology adoption in supply chains-Indian context," *International Journal of Production Research*, vol. 57, no. 7, pp. 2009–2033, 2019.

[42] G. Nuryyev, Y.-P. Wang, J. Achyldurdyyeva et al., "Blockchain technology adoption behavior and sustainability of the business in tourism and hospitality SMEs: an empirical study," *Sustainability*, vol. 12, no. 3, p. 1256, 2020.

[43] P. Verma and N. Sinha, "Integrating perceived economic wellbeing to technology acceptance model: the case of mobile based agricultural extension service," *Technological Forecasting and Social Change*, vol. 126, pp. 207–216, 2018.

[44] H. T. Hurt, K. Joseph, and C. D. Cook, "Scales for the measurement of innovativeness," *Human Communication Research*, vol. 4, no. 1, pp. 58–65, 1977.

[45] E. C. Hirschman, "Innovativeness, novelty seeking, and consumer creativity," *Journal of Consumer Research*, vol. 7, no. 3, pp. 283–295, 1980.

[46] J. Cao, Y. Shang, Q. Mok, and I. K.-W. Lai, "The impact of personal innovativeness on the intention to use cloud classroom: an empirical study in China," *Communications in Computer and Information Science, Technology in Education: Pedagogical Innovations*, Springer, Singapore, pp. 179–188, 2019.

[47] E. L. Slade, Y. K. Dwivedi, N. C. Piercy, and M. D. Williams, "Modeling consumers' adoption intentions of remote mobile payments in the United Kingdom: extending UTAUT with innovativeness, risk, and trust," *Psychology and Marketing*, vol. 32, no. 8, pp. 860–873, 2015.

[48] A. A. Alalwan, A. M. Baabdullah, N. P. Rana, K. Tamilmani, and Y. K. Dwivedi, "Examining adoption of mobile internet in Saudi Arabia: extending TAM with perceived enjoyment, innovativeness and trust," *Technology in Society*, vol. 55, pp. 100–110, 2018.

[49] B. Okumus, F. Ali, A. Bilgihan, and A. B. Ozturk, "Psychological factors influencing customers' acceptance of smartphone diet apps when ordering food at restaurants," *International Journal of Hospitality Management*, vol. 72, pp. 67–77, 2018.

[50] A. A. Rahman, E. Asrarhaghighi, and S. Ab Rahman, "Consumers and Halal cosmetic products: knowledge, religiosity, attitude and intention," *Journal of Islamic Marketing*, vol. 6, no. 1, pp. 148–163, 2015.

[51] H.-K. Bang, A. E. Ellinger, J. Hadjimarcou, and P. A. Traichal, "Consumer concern, knowledge, belief, and attitude toward renewable energy: an application of the reasoned action theory," *Psychology and Marketing*, vol. 17, no. 6, pp. 449–468, 2000.

[52] F. Knauer and A. Mann, "What Is in it for Me? Identifying Drivers of Blockchain Acceptance Among German Consumers," *The Journal of The British Blockchain Association*, vol. 3, no. 1, pp. 1–16, 2019.

[53] C. A. Mandrik and Y. Bao, *Exploring the Concept and Measurement of General Risk Aversion*, ACR North American Advances, United States, 2005.

[54] H. H. Chang and S. W. Chen, "The impact of online store environment cues on purchase intention: trust and perceived risk as a mediator," *Online Information Review*, vol. 32, no. 6, pp. 818–841, 2008.

[55] L. D. Chen, "A model of consumer acceptance of mobile payment," *International Journal of Mobile Communications*, vol. 6, no. 1, pp. 32–52, 2008.

[56] P. Pavlou, "Consumer intentions to adopt electronic commerce-incorporating trust and risk in the technology acceptance model," *Digit 2001 Proceedings*, vol. 2, 2001.

[57] C. Chen, "Perceived risk, usage frequency of mobile banking services," *Managing Service Quality: International Journal*, vol. 23, no. 5, 2013.

[58] N. Guych, S. Anastasia, Y. Simon, and A. Jennet, *Factors Influencing the Intention to Use Cryptocurrency Payments: An Examination of Blockchain Economy*, UB University of Munich - Central Library, Munich, Germany, 2018.

[59] S. Salem, "A proposed adoption model for blockchain technology using the unified theory of acceptance and use of technology (UTAUT)," *Open International Journal of Informatics (OIJI)*, vol. 7, no. 2, pp. 75–84, 2019.

[60] R. M. Morgan and S. D. Hunt, "The commitment-trust theory of relationship marketing," *Journal of Marketing*, vol. 58, no. 3, pp. 20–38, 1994.

[61] P. H. Schurr and J. L. Ozanne, "Influences on exchange processes: buyers' preconceptions of a seller's trustworthiness and bargaining toughness," *Journal of Consumer Research*, vol. 11, no. 4, pp. 939–953, 1985.

[62] S. Chandra, S. C. Srivastava, and Y. L. Theng, "Evaluating the role of trust in consumer adoption of mobile payment systems: an empirical analysis," *Communications of the Association for Information Systems*, vol. 27, no. 1, p. 29, 2010.

[63] E. Bonsón Ponte, E. Carvajal-Trujillo, and T. Escobar-Rodríguez, "Influence of trust and perceived value on the intention to purchase travel online: integrating the effects of assurance on trust antecedents," *Tourism Management*, vol. 47, pp. 286–302, 2015.

[64] C. Liu, J. T. Marchewka, J. Lu, and C.-S. Yu, "Beyond concern-a privacy-trust-behavioral intention model of electronic commerce," *Information & Management*, vol. 42, no. 2, pp. 289–304, 2005.
[65] M. M. Queiroz and S. Fosso Wamba, "Blockchain adoption challenges in supply chain: an empirical investigation of the main drivers in India and the USA," *International Journal of Information Management*, vol. 46, pp. 70–82, 2019.
[66] T. L. Childers, C. L. Carr, J. Peck, and S. Carson, "Hedonic and utilitarian motivations for online retail shopping behavior," *Journal of Retailing*, vol. 77, no. 4, pp. 511–535, 2001.
[67] R. Agarwal and J. Prasad, "The antecedents and consequents of user perceptions in information technology adoption," *Decision Support Systems*, vol. 22, no. 1, pp. 15–29, 1998.
[68] Y. Lu, S. Yang, P. Y. K. Chau, and Y. Cao, "Dynamics between the trust transfer process and intention to use mobile payment services: a cross-environment perspective," *Information & Management*, vol. 48, no. 8, pp. 393–403, 2011.
[69] B. Suh and I. Han, "Effect of trust on customer acceptance of Internet banking," *Electronic Commerce Research and Applications*, vol. 1, no. 3-4, pp. 247–263, 2002.
[70] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981.
[71] P. M. Bentler, "Comparative fit indexes in structural models," *Psychological Bulletin*, vol. 107, no. 2, pp. 238–246, 1990.
[72] L. T. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, pp. 1–55, 1999.
[73] K. J. Kim and S. S. Sundar, "Does screen size matter for smartphones? Utilitarian and hedonic effects of screen size on smartphone adoption," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 7, pp. 466–473, 2014.
[74] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The use of partial least squares path modeling in international marketing," in *New Challenges to International Marketing*, Emerald Group Publishing Limited, Bingley, UK, 2009.
[75] V. Venkatesh, J. Y. Thong, and X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS Quarterly*, vol. 36, no. 1, pp. 157–178, 2012.

Hindawi
Complexity
Volume 2021, Article ID 2122655, 10 pages
https://doi.org/10.1155/2021/2122655

Research Article

A Prediction Method for the RUL of Equipment for Missing Data

Chen Wenbai,[1] **Liu Chang**,[2] **Chen Weizhao**,[1] **Liu Huixiang**,[1] **Chen Qili**,[1] **and Wu Peiliang**[3]

[1]*School of Automation, Beijing Information Science & Technology University, Beijing 100101, China*
[2]*National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China*
[3]*School of Information Science & Technology, Yanshan University, Qinhuangdao 066004, China*

Correspondence should be addressed to Chen Wenbai; chenwb03@126.com and Liu Chang; changliu22@outlook.com

Received 30 August 2021; Accepted 3 November 2021; Published 20 December 2021

Academic Editor: Long Wang

We present a prediction framework to estimate the remaining useful life (RUL) of equipment based on the generative adversarial imputation net (GAIN) and multiscale deep convolutional neural network and long short-term memory (MSDCNN-LSTM). The method we proposed addresses the problem of missing data caused by sensor failures in engineering applications. First, a binary matrix is used to adjust the proportion of "0" to simulate the number of missing data in the engineering environment. Then, the GAIN model is used to impute the missing data and approximate the true sample distribution. Finally, the MSDCNN-LSTM model is used for RUL prediction. Experiments are carried out on the commercial modular aero-propulsion system simulation (C-MAPSS) dataset to validate the proposed method. The prediction results show that the proposed method outperforms other methods when packet loss occurs, showing significant improvements in the root mean square error (RMSE) and the score function value.

1. Introduction

Prognosis and Health Management (PHM) aims to monitor, predict, and manage the health of the system through models and algorithms and is widely used in aviation, military equipment, industrial manufacturing, and other fields [1]. As one of the important research issues of PHM, remaining useful life prediction (RUL) can provide strategy support for establishing the best maintenance management for equipment. The data-driven method for RUL prediction is developed to analyze the equipment operation data through modeling to determine the remaining available time of equipment. Therefore, the quality of the data is directly related to the accuracy of the RUL prediction [2].

Precision equipment and multisensor fusion are widely used in the industrial field, and obtaining complete monitoring data is crucial to predicting the remaining useful life (RUL) of equipment. In engineering applications, various factors, such as failure of data storage, sensor damage, and mechanical failure, may lead to missing information during equipment information collection and storage [3]. Data packet loss is especially detrimental in complex and harsh working environments, such as aerospace and agricultural production environments [4]. The high cost and difficulty of obtaining equipment degradation data and the existence of information intervals between samples make RUL prediction challenging.

In 1987, Rubin [5] proposed that missing data mechanisms fall into three categories: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR). Handling missing data appropriately is particularly important to ensure the accuracy of missing data imputation [6]. Scholars have conducted numerous studies, and the methods can be roughly divided into three categories: ignoring data or deletion, imputation, and statistical models.

Deleting the missing items in the dataset is the simplest data processing method. Strike et al. [7] simulated three types of mechanisms for dealing with missing data and used different techniques for processing missing data, such as

listwise deletion, mean imputation, and eight types of hot-deck imputation. A detailed simulation study was carried out, and it was concluded that simple deletion was a suitable choice when the missing data volume was small. The imputation method fills in missing values. The most probable value is typically used for imputation, which causes less information loss than incomplete samples obtained by deleting all missing values. Commonly used methods include mean imputation, median imputation, mode imputation, and maximum likelihood estimation. Inspired by machine learning, prediction models were used to estimate missing values from the available information in the dataset [8]. Troyanskaya et al. [9] used the k-nearest neighbor (KNN) to estimate missing values in gene microarray data. The imputation effect was better than the imputation method based on singular value decomposition (SVD). Duan et al. [10] used a deep learning model with a denoising stacked autoencoder (DSAE) to estimate missing values in traffic data. This method proved effective for traffic data imputation and analysis. A statistical model was used to impute the missing values based on the linear or nonlinear relationship between the missing data and the observed data. Ni et al. [11] proposed an advanced calculation method based on a Bayesian network to learn from the raw data. A Markov chain Monte Carlo method was used for sampling based on the probability distribution learned by the Bayesian network. It imputes the missing data multiple times and makes statistical inferences about the results. Li et al. [12] proposed a systematic calculation method of traffic flow data based on probabilistic principal component analysis and historical data to estimate missing flow data. A statistical model was used to impute the missing values based on the prior knowledge of the data model, providing excellent results. However, the statistical model has shortcomings due to the incomplete dataset and incomplete prior knowledge. Machine learning has substantial application potential for data imputation. This study focuses on exploiting the use of existing data and machine learning algorithms to impute missing values.

We propose an RUL prediction framework based on data imputation to deal with missing sensor data in engineering applications. First, the missing data are simulated using various missing sample rates. Then, the generative adversarial imputation net (GAIN) model is used to impute the missing values and fill in the dataset. Finally, the proposed multiscale deep convolutional neural network and long short-term memory (MSDCNN-LSTM) prediction model is used to obtain the RUL value of the equipment. The proposed method is well suited for predicting the RUL of equipment if the sensor data are affected by data packet loss in engineering applications. The performance of the proposed method is demonstrated using the commercial modular aero-propulsion system simulation (C-MAPSS) dataset.

2. Related Work

In recent years, deep learning has powerful function mapping capabilities and data processing capabilities. To extract the complex characteristics inside the spectrum and predict the nicotine volume in tobacco, Jiang et al. [13] proposed a one-dimensional fully convolutional network (1D-FCN) model. Hu et al. [14] presented a deep neural network-based visual analysis approach to process videos to detect different augmentative and alternative communication users in practice sessions.

Deep learning has also been widely used in data-driven RUL prediction methods. Babu et al. [15] first tried to use convolutional neural network (CNN) to predict the RUL of the engine, which improved the ability to automatically extract multidimensional features. Then, Li et al. [16] improved the prediction accuracy by using the deep CNN (DCNN) structure and time window data processing. In order to make the CNN model learn more detailed features, Li et al. [17] proposed an algorithm of MSDCNN, that is, the DCNN with different convolution kernel sizes. In order to extract the time correlation features of condition monitoring data, Kong et al. [17] proposed a hybrid algorithm of CNN and long short-term memory (LSTM) to learn spatial and temporal features. Huang et al. [18] developed a novel deep convolutional neural network-bootstrap-based integrated prognostic approach for the remaining useful life (RUL) prediction of rolling bearing. Hu et al. [19] applied the long short-term memory (LSTM) model for RUL prediction of turbine engines and studied a parameter optimization method with Bayesian theory.

In this article, we use the RUL prediction model of MSDCNN-LSTM proposed by Liu et al. [20] to learn more detailed features in a high-dimensional space and predict RUL of aircraft engines. The hybrid MSDCNN-LSTM model consists of an MSDCNN submodel and an LSTM submodel. Among the MSDCNN-LSTM model, the MSDCNN is used to extract high-dimensional features from the input data by time window processing, and the LSTM performs time-series learning on the input data at the same time. Then, the feature map of MSDCNN and LSTM are added and flatten. Finally, the output is sent to a dense layer that represents the RUL output value. The structure chart of MSCNN-LSTM is shown in Figure 1.

3. Proposed Method

3.1. Missing Data Imputation Method Based on GAIN. In 2014, Lan Goodfellow et al. [21] first proposed the generative adversarial net (GAN), which generates data in an adversarial manner with generators and discriminators. The method attracted the attention of researchers and was verified theoretically and practical in engineering applications. The GAN has wide applicability in image, text, and audio processing and other fields.

We used the GAIN model [22] to generate time-series data with a similar distribution as the original for missing data imputation. The basic structure of the model is shown in Figure 2.

The generator is used to observe each part of the real data, and the missing data are imputed according to the observations. The vector $\overline{X}$ in the missing data imputation is expressed as follows:

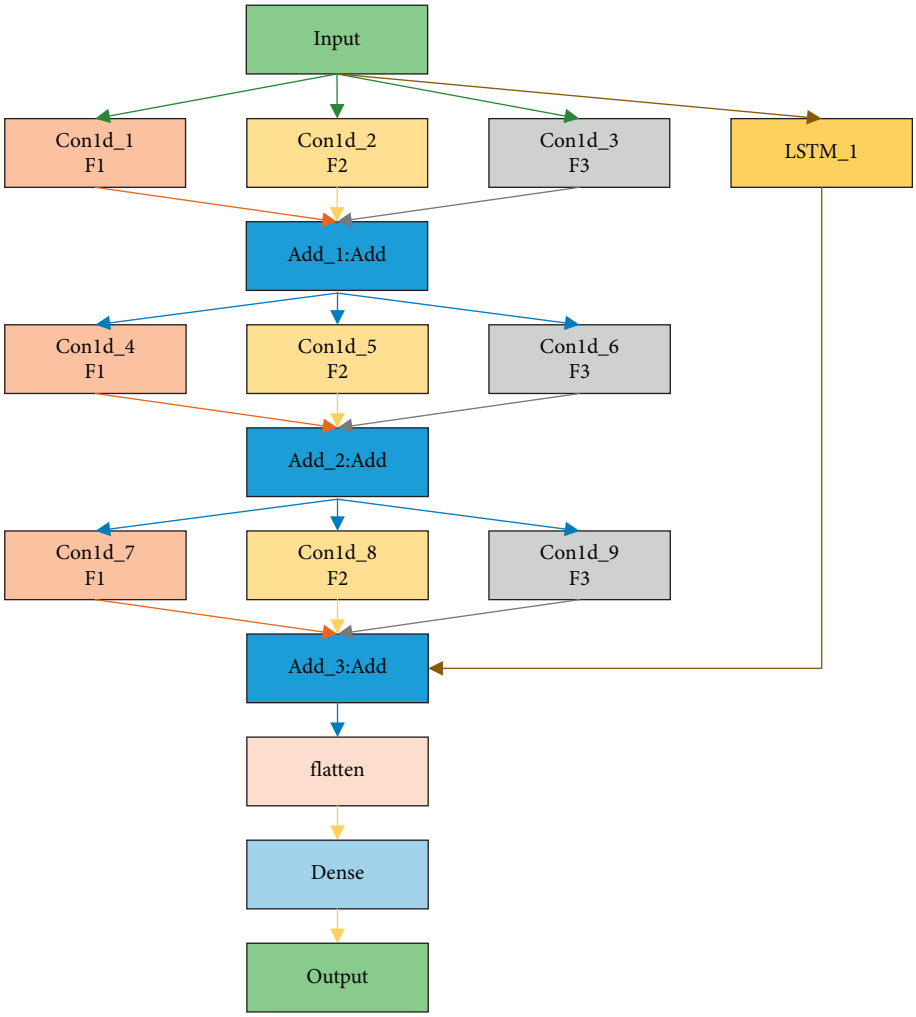


Figure 1: The structure chart of MSCNN-LSTM.

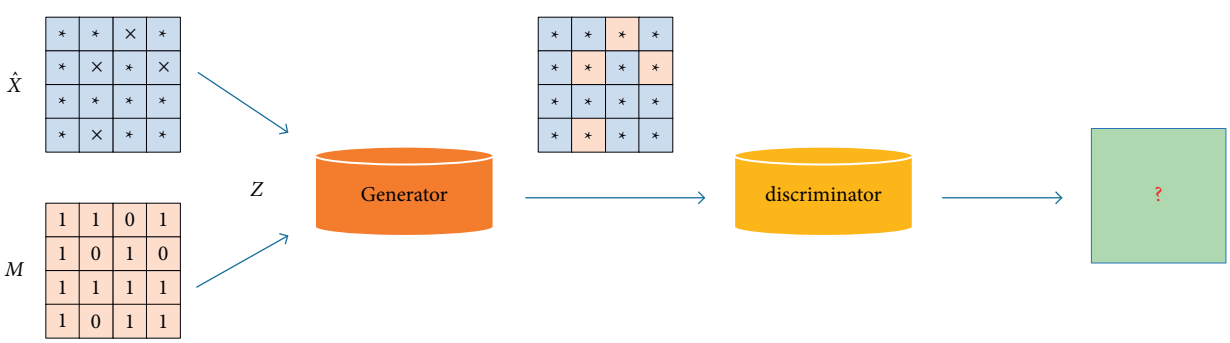


Figure 2: The basic structure of GAIN.

$$\overline{X} = G(\widehat{X}, M, (1-M)\odot Z), \tag{1}$$

where $\widehat{X}$ represents a small sample with missing data, M represents a binary matrix with the same size as $\widehat{X}$, Z represents noise, and $\odot$ represents the multiplication of the corresponding elements.

Finally, the generator outputs a complete vector $\widetilde{X}$ after imputation as follows:

$$\widetilde{X} = M\odot\widehat{X} + (1-M)\overline{X}. \tag{2}$$

Since some of the output of the generator is real and some is generated, the difference between the GAIN and the GAN is that the discriminator of GAIN does not determine the authenticity of the entire vector but detects the real and generated parts, i.e., it predicts the value of m in M. The model trains D by maximizing the probability of correctly

predicting M and trains G by minimizing the probability of correctly predicting M. The objective function is expressed as

$$\min_{G}\max_{D} V(G,D) = E_{\widetilde{X},M}\left[M^{T}\log D(\widetilde{X}) + (1-M)^{T}\log(1-D(\widetilde{X}))\right]. \tag{3}$$

The discriminator distinguishes the source of each part of the input data, and the obtained discriminant matrix is represented by $\widehat{M}$. The cross-entropy loss function is used to evaluate each element in M:

$$\mathscr{L}_{D}(m,\widehat{m}) = \left[\sum_{i=1}^{d} m_i\log(\widehat{m}_i) + (1-m_i)\log(1-\widehat{m}_i)\right]. \tag{4}$$

The loss function of the generator is defined as

$$\mathscr{L}_{G}(m,\widehat{m}) = \left[\sum_{i=1}^{d} 1 - m_i\log(\widehat{m}_i)\right] + \alpha\mathrm{MSE}. \tag{5}$$

3.2. RUL Prediction Framework Based on GAIN and MSDCNN-LSTM. An RUL prediction framework that combines the GAIN and MSDCNN-LSTM is designed; it consists of three parts: preprocessing the missing data, missing data imputation based on the GAIN model, and RUL prediction based on the MSDCNN-LSTM model, as shown in Figure 3.

First, data preprocessing is performed on the C-MAPSS dataset. The method described in Section 1 is used to construct sample data for the training set with different missing data rates. Subsequently, the GAIN network is used to impute the samples with different missing data rates, and the generator and the discriminator generate data in an adversarial manner to obtain a dataset close to the original one. Finally, the generated samples are used as the input of the MSDCNN-LSTM prediction model. The MSDCNN and LSTM models process the data simultaneously. The multiscale structure of the MSDCNN substantially improves the feature extraction capability. Convolution kernels of different sizes ($F1$, $F2$, and $F3$) are used to extract features from the input data, and the feature maps are spliced together and combined with the time series to obtain the prediction results of the LSTM. Continuous iteration is used to evaluate the trained model using two indicators (root mean square error (RMSE) and the score function), and the test set data are input to obtain the RUL prediction result.

The RUL prediction process based on the combination of GAIN and MSDCNN-LSTM is shown in Figure 4. First, the missing data are generated using missing data rates of 0.1, 0.2, 0.3, 0.4, and 0.5. Then, we use the GAIN network to impute the missing values of the samples. During the training of the GAIN network, we set the epoch to 1000 times, and the newly generated time-series dataset are standardized and processed by a time window. After setting the RUL labels of the training set and test set, the next stage is model training and system prediction. The parallel MSDCNN-LSTM hybrid model performs multiscale feature extraction on the time-series training set, and the parameters and weights in the model are updated using a minibatch of 512. When the early stopping conditions set by the system are met, the model training ends early. If the early stopping condition is not met, the minibatch training is continued until the maximum epoch. After the model is trained, we input the test set data to predict the RUL result of each engine.

4. Experimental and Results

4.1. Experimental Dataset and Settings. The experiments are conducted on a server computer configured with an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz and an NVIDIA GeForce TITAN XP. The C-MAPSS dataset is used to verify the proposed method. The C-MAPSS dataset is divided into four subsets (FD001, FD002, FD003, and FD004) according to the operating conditions and failure modes. Each dataset contains engine degradation data monitored by 21 sensors, as listed in Table 1. Each subset is divided into a training set and a test set. FD002 and FD004 have 6 operating conditions, and FD003 and FD004 have 2 failure modes.

The score function and the RMSE are used as evaluation indicators. The formula of the score function is [23]

$$\mathrm{Score} = \begin{cases} \sum_{i=1}^{N}\left(e^{-d_i/13} - 1\right), & d_i < 0, \\ \sum_{i=1}^{N}\left(e^{d_i/10} - 1\right), & d_i \geq 0. \end{cases} \tag{6}$$

The formula of the RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}, \tag{7}$$

where d_i represents the difference between the predicted value of RUL and the true value, $d_i = \mathrm{RUL}_{\mathrm{pre}} - \mathrm{RUL}_{\mathrm{act}}, i = 1, 2, \ldots, N$. When d_i is less than 0, the predicted value is less than the true value, and the result is referred to as an advanced prediction; otherwise, it is a lagging prediction.

The lower the value of the score function and RMSE, the better the predictive ability of the model. The RMSE is a symmetric function and provides the same result for an advanced prediction and lagging prediction. However, the score function is an asymmetric function and is more sensitive to lagging prediction. Because lagging prediction has more serious consequences, it results in stronger penalties than advanced prediction. Therefore, these indicators comprehensively measure the performance of the algorithm.

4.2. Simulating the Missing Data Rate. It is assumed that the original dataset is X, $X = [X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots, X_{im}] \in R^{m\times n}$, $i \in [1, n]$ and $j \in [1, m]$, where m is the number of sensors, n is the length of the time series, and x_{ij} is the measured value of the jth sensor corresponding to the ith period. Here, we define a binary matrix M,

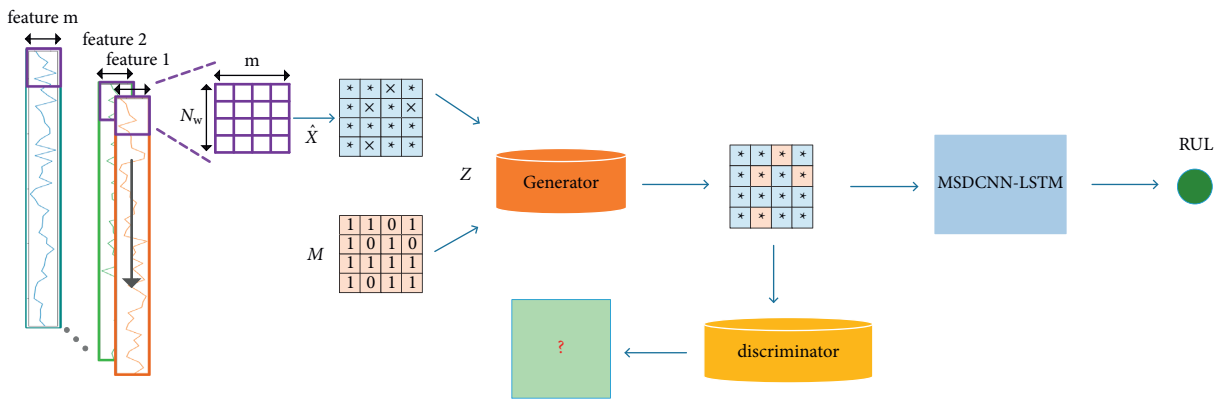


Figure 3: RUL prediction framework based on the GAIN and MSDCNN-LSTM.

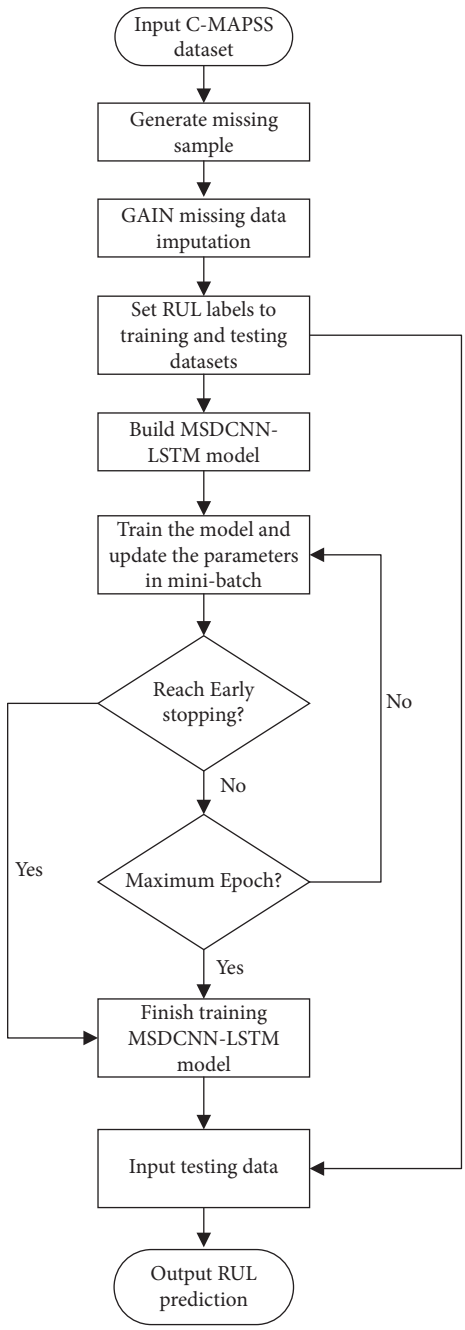


Figure 4: Flowchart of the RUL prediction.

Table 1: C-MAPSS subdatasets.

Subdatasets	FD001	FD002	FD003	FD004
Training engines	100	260	100	249
Test engines	100	259	100	248
Conditions	1	6	1	6
Failure modes	1	1	2	2

$M = [M_{i1}, M_{i2}, \ldots, M_{ij}, \ldots, M_{im}] \in R^{m\times n}$, $i \in [1, n]$ and $j \in [1, m]$, which has the same size as the original data X and consists of 0 and 1 values. The reconstructed missing data $\widehat{X}$ can be expressed as

$$\widehat{X} = \begin{cases} X_{i,j}, & M_{i,j} = 1, \\ \text{Nan}, & \text{other}, \end{cases} \tag{8}$$

where M represents the observed component of X. A value of 1 indicates the observed data, and a value of 0 represents the missing data $\widehat{X}$. Datasets with different missing sample rates can be created by changing 0 to another value.

4.3. Data Imputation Simulation Results and Analysis. In the 4 subsets of the C-MAPSS dataset, 7 sensor data with no changes were eliminated. Therefore, the sensor numbers used in this experiment are 2, 3, 4, 7, 8, 9, 11, 12, 13, 14, 15, 17, 20, and 21. The RMSE was used as an evaluation indicator to evaluate the imputation effect of GAIN.

Table 2 lists the imputation accuracy of GAIN for different missing data rates (0.1, 0.2, 0.3, 0.4, and 0.5) on the C-MAPSS dataset.

As the missing data rate increases, the RMSE values of the four subdatasets FD001, FD002, FD003, and FD004 increase, and the accuracy decreases. In the case of a high missing data rate, there is less sample information, and it is difficult to fill in the missing sample data. The imputation performance of GAIN is better for the FD002 and FD004 datasets with complex working conditions and a large sample size than for the FD001 and FD003 datasets with simple working conditions and a small sample size. Therefore, the imputation performance is better for a larger sample size, and the prediction accuracy decreases as the missing data rate increases.

Figure 5 shows the visualization results of GAIN after missing data imputation for a missing data rate of 0.5. The horizontal axis represents the operating cycle of the first engine, and the vertical axis is the result of the first sensor data after maximum-minimum standardization. The middle black rectangle represents the real data, and the red dots represent the results of GAIN after imputation. Although the effect of missing data is more serious when the missing data rate is high, the data after imputation based on GAIN fluctuates in a small range around the real data, and the overall distribution is consistent with the real data distribution.

Table 3 shows the influence of the loss function on the GAIN model performance. During data imputation, the loss function is particularly important for training the generator and discriminator models. After conducting experiments, we found that the model performance was best when the

Table 2: The RMSE value of GAIN for different missing data rates.

Missing data rates	FD001	FD002	FD003	FD004
0.1	0.1041	0.0217	0.0915	0.0210
0.2	0.1050	0.0223	0.0917	0.0200
0.3	0.1076	0.0257	0.0988	0.0231
0.4	0.1079	0.0419	0.0997	0.0525
0.5	0.1091	0.1011	0.1126	0.0569

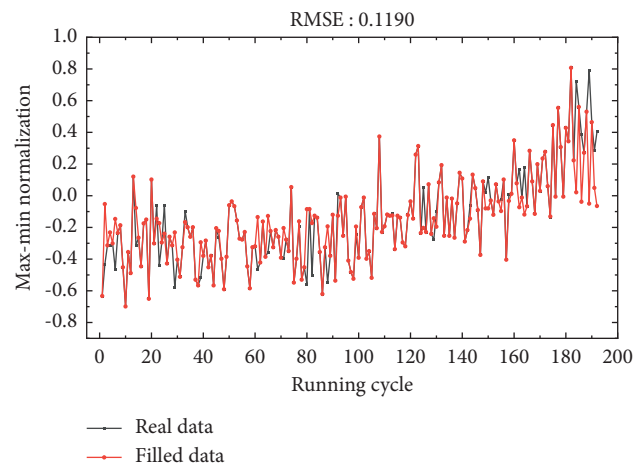


Figure 5: The result of GAIN imputation.

cross-entropy loss function and the mean square error loss function were used. We use the FD001 dataset with a missing data rate of 0.5 as an example to verify the impact of the loss function on the model performance and compare the simulation results obtained from different combinations of loss functions. It can be seen from Table 3 that the combinations of the two loss functions and the adjustment of the parameters ∂ significantly affect the results. The optimal RMSE value is obtained when the cross-entropy loss function is used for the discriminator, and the cross-entropy loss function + ∂ mean square are used for the generator. In the experiment, different combinations were used under the same conditions to verify the effect of the parameter ∂ in the loss function. The results are listed in Table 4.

Adding the parameter ∂ to the generator loss function improves the imputation accuracy, but $1/\partial$ and $(1-\partial)$ substantially increase the RMSE value. Therefore, the model provides the best performance when the coefficient of the RMSE in the generator loss function is ∂.

Figure 6 shows the results of the GAIN imputation and other methods. The GAIN imputation, mean imputation, median imputation, and mode imputation are compared using the FD001 dataset. The horizontal axis represents the missing data rate, and the vertical axis represents the RMSE. As the missing data rate increases, the RMSEs of the four methods show an upward trend, and the imputation performance decreases. The results for different missing data rates indicate that the mean value imputation results are more stable than the mode and median imputation methods. However, the GAIN achieves the smallest RMSE values for the different missing data rates, indicating that it outperforms the other methods.

Table 3: The influence of the loss function on model performance.

D_loss	G_loss	RMSE
Cross entropy	Cross entropy + ∂ RMSE	0.1190
Cross entropy + ∂ RMSE	Cross entropy + ∂ RMSE	0.1347
Cross entropy	Cross entropy	0.5065
Cross entropy + ∂ RMSE	Cross entropy	0.4692
Cross entropy	RMSE	0.1423

Table 4: The influence of the parameters in the loss function on model performance.

D_loss	G_loss	RMSE
Cross entropy	Cross entropy + ∂ RMSE	0.1190
Cross entropy	Cross entropy + ∂^2 RMSE	0.1227
Cross entropy	Cross entropy + $1/\partial$ RMSE	0.4900
Cross entropy	Cross entropy + $(1-\partial)$ RMSE	0.5664
Cross entropy	Cross entropy + RMSE	0.1245

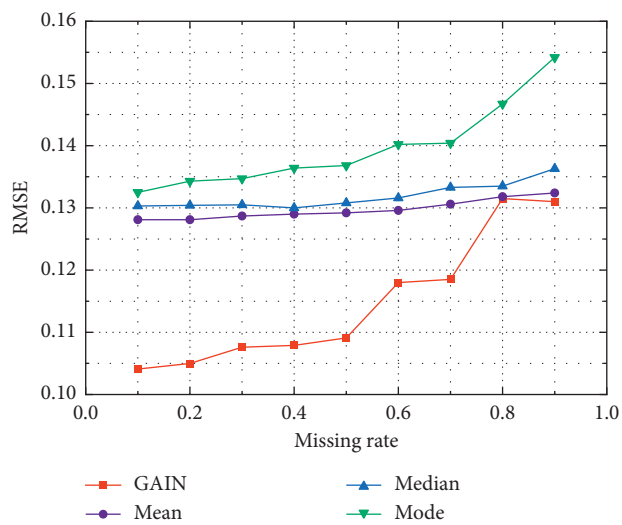


Figure 6: The results of different imputation methods for different missing data rates.

Figure 7 shows the RUL prediction results of all engine units after GAIN imputation on the C-MAPSS dataset when missing data rate is 0.1. The test engine is sorted by RUL from small to large to better observe the changes in prediction accuracy. The horizontal axis represents test engine unit, and the vertical axis represents the RUL. In the figure, the black dots represent the real RUL, and the red dots represent the model prediction results. It can be seen from Figure 7 that, at the initial stage of engine operation, the RUL value is relatively large and the prediction error is relatively large. When the engine runs for a long time or is about to fail, the degradation information is more obvious, and the predictive performance is significantly enhanced. The proposed framework reflects a good forecasting effect.

Table 5 shows the RUL prediction results with and without GAIN imputation for a missing data rate of 0.1. It is worth noting that the system automatically replaces missing data with 0 to ensure the smooth execution of the RUL prediction algorithm. Therefore, when the missing data rate is 0.1, the score function value cannot be obtained, and it causes difficulties for the subsequent RUL prediction, such as a substantial increase in the RMSE value. However, after the missing data are imputed by GAIN, the prediction results are significantly improved. The RMSE has increased by at least 80.16%, and the score function value has increased by at least 99.98%.

Table 6 shows the prediction results of the proposed GAIN method for missing data rates of 0.1, 0.2, 0.3, 0.4, and 0.5. As the missing data rate increases, the prediction accuracy of the 4 subdatasets decreases. The score function is more affected than the RMSE. When the missing rate is less than 0.4, the proposed RUL prediction framework based on data imputation can show better performance. When the missing data rate is higher than or equal to 0.4, too much data information is lost, resulting in low prediction performance. However, the prediction result of the proposed framework is much better than using no missing data imputation for a missing data rate of 0.1.

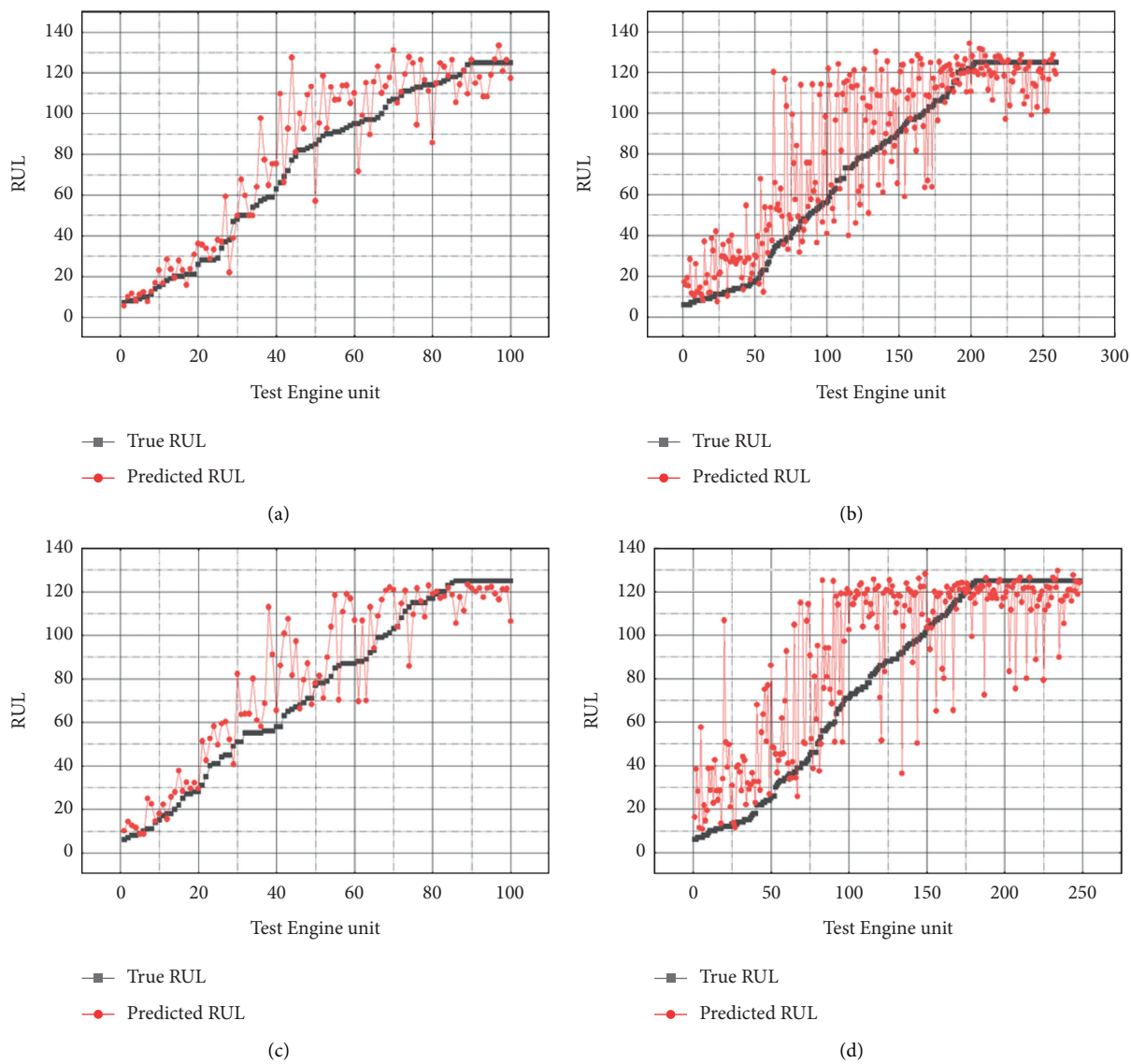


Figure 7: The results of RUL prediction after GAIN imputation when missing data rate is 0.1. (a) 100 test engine units on FD001. (b) 259 test engine units on FD002. (c) 100 test engine units on FD003. (d) 248 test engine units on FD004.

Table 5: The results of the RUL prediction with and without data imputation for a missing data rate of 0.1.

Method	FD001		FD002		FD003		FD004	
	RMSE	*Score*	*RMSE*	*Score*	*RMSE*	*Score*	*RMSE*	*Score*
GAIN imputation	14.48	274.62	18.73	2067.25	15.83	338.05	23.28	4750.18
No imputation	71.47	1620023.5	77.04	1815133.4	75.62	3656987.8	66.73	4859614.0

Table 6: The results of the RUL prediction with data imputation for different missing data rates.

Missing rate	FD001		FD002		FD003		FD004	
	RMSE	*Score*	*RMSE*	*Score*	*RMSE*	*Score*	*RMSE*	*Score*
0.1	14.48	274.62	18.73	2067.25	15.83	338.05	23.28	4750.18
0.2	15.74	361.07	23.62	3647.92	17.45	494.89	24.45	5744.81
0.3	18.17	506.06	23.74	3208.44	23.07	847.91	32.52	13376.09
0.4	19.61	521.18	43.19	36818.28	29.13	1531.91	32.07	42081.27
0.5	19.65	503.78	51.96	703764.25	26.44	1216.59	43.38	34483.86

5. Conclusions

This paper proposed a RUL prediction method based on the combination of GAIN and MSDCNN-LSTM. Experiments were carried out with a missing data rate of 0.1–0.5 to simulate data packet loss in industrial production. In the GAIN method, the generator interacts with the discriminator to impute the missing data and generate a sample dataset that is close to reality. This dataset is used as the input of the MSDCNN-LSTM prediction model. A comparison of the simulation results of GAIN and other methods indicated that the GAIN imputation method outperformed the mean, median, and mode imputation methods. The proposed prediction framework was compared with no data imputation when packet loss occurred and exhibited a significant improvement. The RUL prediction framework showed better prediction performance than other methods on the C-MAPSS dataset for different missing data rates.

Data Availability

The data used to support the findings of the study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (Grant no. 4202026), Qin Xin Talents Cultivation Program of Beijing Information Science and Technology University (QXTCP A202102), Beijing Postdoctoral Science Foundation (Grant no. ZZ-2019-65), Chaoyang District Postdoctoral Science Foundation (Grant no. 2019ZZ-45), the National Key R&D Program of China (Grant no. 2018YFB1308300), and National Natural Science Foundation of China (Grant no. 62103056).

References

[1] B. Bagheri, M. Rezapoor, and J. Lee, "A unified data security framework for federated prognostics and health management in smart manufacturing," *Manufacturing Letters*, vol. 24, pp. 136–139, 2020.

[2] S. F. L. T. M. Zhang and C. H. Hu, "Deep convolutional generative adversarial network based missing data generation method and its application in remaining useful life prediction," *Acta Aeronautica et Astronautica Sinica*, vol. 42, no. 6, (in Chinese), Article ID 625207, 2021.

[3] X. Hu, H. Zhang, D. Ma, and R. Wang, "Hierarchical pressure data recovery for pipeline network via generative adversarial networks," *IEEE Transactions on Automation Science and Engineering*, pp. 1–11, 2021.

[4] F. Qu, J. Liu, Y. Ma, D. Zang, and M. Fu, "A novel wind turbine data imputation method with multiple optimizations based on GANs," *Mechanical Systems and Signal Processing*, vol. 139, Article ID 106610, 2020.

[5] D. B. Rubin, "The calculation of posterior distributions by data augmentation: comment: A n sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 543–546, 1987.

[6] A. Ngueilbaye, H. Wang, D. A. Mahamat, and S. B. Junaidu, "Modulo 9 model-based learning for missing data imputation," *Applied Soft Computing*, vol. 103, Article ID 107167, 2021.

[7] K. Strike, K. El Emam, and N. Madhavji, "Software cost estimation with incomplete data," *IEEE Transactions on Software Engineering*, vol. 27, no. 10, pp. 890–908, 2001.

[8] J. M. Jerez, I. Molina, P. J. García-Laencina et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105–115, 2010.

[9] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[10] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 168–181, 2016.

[11] D. Ni and J. D. Leonard, "Markov chain Monte Carlo multiple imputation using bayesian networks for incomplete intelligent transportation systems data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1935, no. 1, pp. 57–67, 2005.

[12] L. Qu, L. Jianming, H. Li, and Y. Zhang, "PPCA-based missing data imputation for traffic flow volume: a systematical approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 512–522, 2009.

[13] D. Jiang, G. Hu, G. Qi, and N. Mazur, "A fully convolutional neural network-based regression approach for effective chemical composition analysis using near-infrared spectroscopy in cloud," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 1, pp. 74–82, 2021.

[14] G. Hu, S. H. K. Chen, and N. Mazur, "Deep neural network-based speaker-aware information logging for augmentative and alternative communication," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 2, pp. 138–143, 2021.

[15] G. Sateesh Babu, P. Zhao, and X.-L. Li, "Deep convolutional neural network based regression approach for estimation of remaining useful life," in *International conference on database systems for advanced applications*, pp. 214–228, Springer, Dallas, TX, USA, April 2016.

[16] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliability Engineering & System Safety*, vol. 172, pp. 1–11, 2018.

[17] H. Li, W. Zhao, Y. Zhang, and E. Zio, "Remaining useful life prediction using multi-scale deep convolutional neural network," *Applied Soft Computing*, vol. 89, Article ID 106113, 2020.

[18] C. G. Huang, H. Z. Huang, Y. F. Li, and P. Weiwen, "A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing," *Journal of Manufacturing Systems*, vol. 61, pp. 757–772, 2021.

[19] S. Hu, S. Zhang, and X. Xu, "RUL prediction by LSTM model with bayesian parameter optimization for turbine engines," *Journal of Physics: Conference Series*, IOP Publishing, vol. 1646, no. 1, , Article ID 012122, 2020.

[20] C. Liu and W. Chen, "A RUL prediction method of equipments based on MSDCNN-LSTM," *Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University*, vol. 39, no. 2, pp. 407–413, 2021.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., *Generative Adversarial Networks*, https://arxiv.org/abs/1406.2661, 2014.

[22] J. Yoon, J. Jordon, and M. Schaar, "Gain: missing data imputation using generative adversarial nets," in *International Conference on Machine Learning*, pp. 5689–5698, July 2018, https://arxiv.org/abs/1806.02920.

[23] X. S. Si, W. Wang, C. H. Hu, and M. Y. Chen, "A Wiener-process-based degradation model with a recursive filter algorithm for remaining useful life estimation," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 219–237, 2013.

Hindawi
Complexity
Volume 2021, Article ID 4679739, 11 pages
https://doi.org/10.1155/2021/4679739

Research Article

Storage Assignment Optimization in Robotic Mobile Fulfillment Systems

Ruiping Yuan,[1,2] **Juntao Li**,[1,2] **Wei Wang**,[1] **Jiangtao Dou**,[1] **and Luke Pan**[1]

[1]*School of Information, Beijing Wuzi University, Beijing 101149, China*
[2]*Beijing Key Laboratory of Intelligent Logistics Systems, Beijing 101149, China*

Correspondence should be addressed to Juntao Li; ljtletter@126.com

Received 20 August 2021; Revised 16 October 2021; Accepted 3 November 2021; Published 25 November 2021

Academic Editor: Long Wang

Robotic mobile fulfillment system (RMFS) is a new type of parts-to-picker order picking system, where robots carry inventory pods to stationary pickers. Because of the difference in working mode, traditional storage assignment methods are not suitable for this new kind of picking system. This paper studies the storage assignment optimization of RMFS, which is divided into products assignment stage and pods assignment stage. In the products assignment stage, a mathematical model maximizing the total correlation of products in the same pods is established to reduce the times of pod visits, and a scattered storage policy is adopted to reduce system congestion. A heuristic algorithm is designed to solve the model. In the pods assignment stage, a model is established minimizing the total picking distance of the mobile robots considering the turnover rate and the correlation of pods as well as the workload balance among picking corridors. A two-stage hybrid algorithm combining greedy algorithm and improved simulated annealing is designed to solve the model. Finally, a simulation experiment is carried out based on the historical order data of an e-commerce company. Results show that the storage assignment method proposed in the paper significantly improves the efficiency of order picking.

1. Introduction

E-commerce orders are generally large in quantity, small in batch, and unstable, so e-commerce order fulfillment can be quite challenging for warehouses. Robotic mobile fulfillment system (RMFS) is a new type of parts-to-picker order picking system, which was first brought to the market by Amazon in 2012. RMFS is particularly suited for e-commerce distribution centres that handle strong demand fluctuations and large assortments of small products, so it has become the development trend of e-commerce picking system in recent years [1]. In an RMFS, mobile robots are used to bring movable shelves, also known as inventory pods, to picking stations, where pickers take the required products off pods. After that, the robots bring the pods back to the storage area and continue to carry the next pods required. By eliminating the need for the pickers to walk and search the inventory, RMFS greatly improves the efficiency of order picking [2]. As a new picking mode, there are many optimization problems to be studied in RMFS, such as storage assignment, order batching [3], multirobot task allocation [4], and path planning [5]. As an important optimization direction of RMFS, storage assignment has a direct impact on the total travel time/distance of robots, hence the efficiency of order picking.

In a traditional picker-to-parts picking system, the scientific storage assignment method can shorten the walking distance, reduce the search time, and improve the efficiency of order picking [6]. Hausman et al. [7] were the first to study the storage assignment strategy of traditional picking systems, and the subsequent literature conducted more in-depth research from demand correlation [8], COI (cube-per-order index) [9], turnover rate [10], and demand and structure correlation [11].

Unlike the traditional picker-to-parts picking system, storage assignment in RMFS includes not only products assignment (to decide which product to assign on which pod) but also pods assignment (to decide where the pods are

put in the storage areas) [12–14]. Because of the different ways of working, the traditional storage allocation algorithm cannot be directly applied to RMFS.

In recent years, the storage allocation problem in RMFS has attracted scholars' attention. Xi et al. [15] proposed a collaborative optimization model on the products assignment and order batching problem of RMFS considering the relevance of products. Yuan et al. [16] studied the pods assignment in RMFS, considering the random, class-based, and turnover-based assignment strategies. The research shows that class-based storage with two or three classes can achieve most of the potential benefits, and these benefits increase with greater variation in the pod velocities. Then Roy et al. [17] and Weidinger and Boysen [18] studied the pods assignment and products assignment in RMFS using random assignment policy and scattered assignment policy, respectively. Based on this, Weidinger et al. [19] further studied the reassignment of pods as a special interval scheduling problem and designed an adaptive algorithm. The results show that the proposed adaptive assignment rules are better than traditional assignment strategies. Lamballais et al. [20] developed an analytical model to optimize the inventory allocation across the pods. It is found that spreading inventory units across multiple pods is a better allocation strategy for e-commerce order fulfillment.

From the literature review, we can find that though the research on storage assignment under traditional manual picking mode is very rich, the research on that of the new picking system RMFS has just begun with the following shortcomings: (1) most of the research only focuses on the products allocation or pods allocation, but few on the joint optimization of the two stages; (2) the correlation between products is considered to reduce the pod visits in previous literature, but the correlation between pods is ignored. If we assign the pods with a strong correlation close to each other, it can effectively shorten the travel distance of robots during pods switching; and (3) the optimization objective of most existing literature is to minimize the travel distance of robots but seldom consider system congestion caused by the unbalanced workload among picking corridors, which affects system efficiency even more than travel distance of robots. Thus, reducing system congestion can effectively improve the picking efficiency of RMFS.

This paper studies the storage assignment optimization of RMFS and a joint optimization model of products assignment and pods assignment is built to improve the overall picking efficiency. The main innovations of this paper are as follows: (1) in the modelling, on the basis of the correlation and turnover rate of products, the correlation between pods is also considered to shorten the travel distance of robots during pods switching; (2) scattered storage policy and workload balance among picking corridors are adopted to reduce system congestion. Scattering products in different pods can increase the choice of pod visits and balancing the workload can reduce the congestion of some corridors; and (3) effective hybrid algorithms, including heuristic algorithm, greedy algorithm, and improved simulated annealing algorithm, are designed to solve the models.

The rest of the paper is organized as follows. In Section 2, we describe the storage assignment problem in RMFS. In Section 3, we build the mathematical models of products assignment and pods assignment. In Section 4, we design hybrid algorithms to solve the models. In Section 5, we verify the proposed methods using the real data of an e-commerce company. We finally conclude our study and discuss future research opportunities in Section 6.

2. Problem Description

The typical layout of an RMFS is shown in Figure 1, which is adopted by Amazon Kiva Systems [2] and most RMFS providers. The picking system consists of pods, picking stations, mobile robots, picking corridors, a conveyor belt, and so on. The storage area is composed of neatly distributed pods, each of which can store different products. In order to improve order picking efficiency, multiple orders arrived in a certain period of time are usually combined into one batch for picking, which is called wave picking. Orders in one wave are allocated to picking stations, and the items to be picked in the orders of each picking station are merged to generate a picking list, which consists of many picking tasks. These picking tasks are assigned to robots according to some rules, and these robots cooperate to complete the tasks.

The products picking process is as follows: A mobile robot is sent to the pod containing the required products and transports them to the picking station, where a picker picks the products from the pod, and then the robot sends the pod back to the storage areas (that is called one pod visit). After putting the previous pod back, the robot travels to carry the next pod (that is called pods switching) and so forth until all of its tasks are finished.

The optimization objective of RMFS is to minimize the total travel distance of robots or the total picking time of orders. The storage assignment in RMFS refers to assigning products to the appropriate pods (products assignment) and pods to the appropriate location in the storage areas (pods assignment), which obviously has a direct impact on the total travel time/distance of robots, hence the efficiency of order picking.

In most studies, storage assignment optimization is usually achieved by putting the related products on the same pods to reduce pod visits or by putting the pods with a high turnover rate closer to the picking stations to reduce the travel distance of robots. However, the picking corridors are generally very narrow in RMFS to save storage space. If the products with a high turnover rate are all placed in the zone close to the picking stations, it is easy to cause road congestion, and the robots have to queue and wait, which will affect the picking efficiency more than the saved travel distance. Therefore, system congestion factors should be given enough consideration in the storage allocation of RMFS.

In this paper, we study the joint optimization of products assignment and pods assignment of RMFS considering the correlation, turnover rate of products and pods, as well as system congestion factors. The main ideas of optimization in each stage are discussed below.

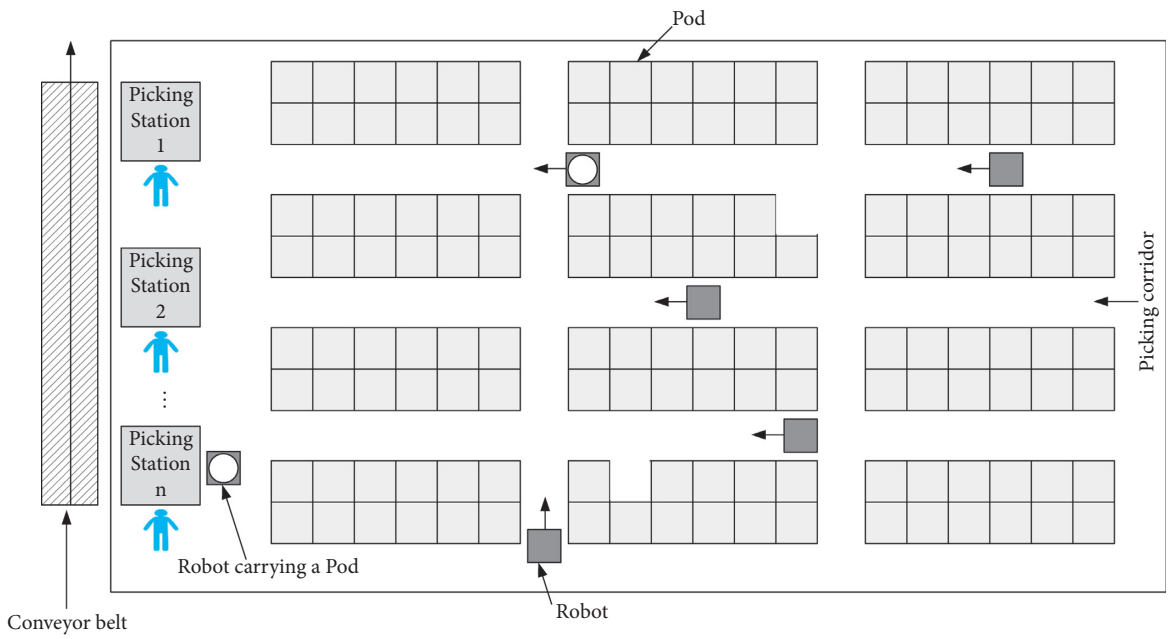


Figure 1: The typical layout of an RMFS.

2.1. The Products Assignment Stage. In the products assignment stage to decide which product to assign to which pod, we take the following measures to improve efficiency:

(1) Store the products with strong correlation in the same pods to reduce the times of pod visits. The stronger the correlation of products in one pod, the more likely the products are required by the same order so that more kinds of products can be picked out by one pod visit, which obviously can reduce the total number of pod visits to fulfill the orders.

(2) Spread the same kind of products across multiple pods other than just one pod to reduce the possibility of system congestion. When the products are scattered in different pods, it increases the choices of pod visits. When one pod is not available or it is in a congested area, there are still other pods can be chosen to fulfill the order. Thus, it can reduce system congestion and improve picking efficiency.

2.2. The Pods Assignment Stage. In the pods assignment stage, to decide where the pods to put in the storage areas, the following two aspects are considered:

(1) In order to shorten the travel distance of robots during pods switching, the correlation between pods is considered. After putting the previous pod back, a robot has to travel to carry the next pod. If the next pod is put close to the previous pod, it will shorten the travel distance of pods switching and reduce the total travel distance of all robots.

(2) While the pods with a high turnover rate are placed near picking stations, the workload balance among picking corridors is considered simultaneously. The pods that contain the best-selling products have a high turnover rate and visit frequency, so putting them close to the picking stations can reduce the robots' travel distance. However, it can also cause congestion in hot areas. In order to reduce the possibility of road congestion, we balance the workload among corridors by setting a maximum workload volume for each corridor.

3. Model Formulation

3.1. Assumptions and Parameter definition. The following assumptions, which are reasonable in reality, are listed for the mathematical model formulation:

(i) The pods are empty at the beginning.

(ii) The inventory of each kind of product is 4 times its average demand. Yuan et al. [21] have proved that when the inventory is 4 times the average demand, there is little possibility of shortage.

(iii) One kind of product can be stored in different pods.

(iv) The number of pods is sufficient to store all products.

(v) All pods have the same fixed layers (usually 6–8), and each layer can only store one kind of product in order to operate conveniently for the picker.

(vi) The quantity of products stored on one pod can satisfy the demand of the products in one order.

(vii) Only one pod can be allocated for each position in the storage area.

(viii) A pod can only visit one picking station at one time and then is returned to its original position.

3.2. Mathematical Model of Products Assignment. This section will formulate an optimization model for products assignment with the objective of maximizing the total correlation of products in the same pods to reduce the number of pod visits. Meanwhile, scattered storage strategy is adopted to avoid system congestion. Before describing the model, the following parameters are given:

i, j represent product type, $i, j = 1, 2, \ldots, P$, P is the total number of product types

n represents order, $n = 1, 2, \ldots, N$, N is the total number of orders

m represents pod, $m = 1, 2, \ldots, M$, M is the total number of pods

q represents the total number of layers on one pod

l represents the maximum number of items stored in each layer of a pod

a_i represents the average demand of product i

d_i represents the total number of storage layers needed of products i

$$d_i = \begin{cases} \dfrac{4a_i}{l}, & 4a_i \text{ can be divided by } l, \\ \left[\dfrac{4a_i}{l}\right] + 1, & \text{otherwise.} \end{cases} \tag{1}$$

$B_n = \{b_{1n}, b_{2n}, \ldots, b_{Pn}\}^T$ represents products set included in order n. If product i is included in order n, $b_{in} = 1$; otherwise, $b_{in} = 0$.

R represents products correlation matrix:

$$R = \begin{bmatrix} r_{11} & r_{21} & \cdots & r_{P1} \\ r_{21} & r_{22} & \cdots & r_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{P1} & r_{P2} & \ldots & r_{PP} \end{bmatrix}, \tag{2}$$

where r_{ij} is the correlation between products i and j. The more frequent the two kinds of products appear in the same order, the stronger their correlation is. The calculation method of the correlation between products is shown in equation (3). When $i \neq j$, the numerator is the number of orders both containing products i and j. The denominator is the number of orders either containing product i or j. When $i = j$, r_{ij} is the correlation of the same kind of product, in order to spread them across multiple pods, set $r_{ij} = 0$ in this case. r_{ij} can be obtained from historical orders data using the following equation:

$$r_{ij} = \begin{cases} \dfrac{\sum_{n=1}^{N} b_{in} b_{jn}}{\sum_{n=1}^{N} b_{in} + \sum_{n=1}^{N} b_{jn} - \sum_{n=1}^{N} b_{in} b_{jn}}, & i \neq j, \\ 0, & i = j. \end{cases} \tag{3}$$

x_{im} represents the decision variable and can be expressed as follows:

$$x_{im} = \begin{cases} 1, & \text{product } i \text{ is assigned to pod } m, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

y_{im} is an integer variable, which represents the number of storage layers product i occupying on pod m.

The objective is to maximize the total correlations between products on the same pods as follows:

$$\max z_1 = \sum_{i=1}^{P} \sum_{j>i}^{P} \sum_{m=1}^{M} \frac{r_{ij} x_{im} x_{jm}}{M}. \tag{5}$$

The constraints are presented as follows:

$$\sum_{i=1}^{P} y_{im} \leq q, \quad m = 1, 2, 3, \ldots, M, \tag{6}$$

$$\sum_{m=1}^{M} y_{im} = d_i, \quad i = 1, 2, 3, \ldots, P, \tag{7}$$

$$y_{im} \geq x_{im}, \quad i = 1, 2, 3, \ldots, P, m = 1, 2, 3, \ldots, M, \tag{8}$$

$$x_{im} = 0, 1, \quad i = 1, 2, 3, \ldots, P, m = 1, 2, 3, \ldots, M, \tag{9}$$

$$0 \leq y_{im} \leq q, \quad i = 1, 2, 3, \ldots, P, m = 1, 2, 3, \ldots, M. \tag{10}$$

Equation (6) stipulates that the number of storage layers assigned to the products is no more than the maximum layers of one pod. Equation (7) ensures that the total demand number of storage layers for each type of product is satisfied. Equations (8)–(10) are the basic constraints for the decision variables.

3.3. Mathematical Model of Pods Assignment. This section will formulate an optimization model for pods assignment with the objective of minimizing the total travel distance of all robots, considering the turnover rate of the pods, the correlation between the pods and the workload balance among picking corridors. Based on Section 3.2, some additional parameters are given.

t represents picking corridor, $t = 1, 2, \ldots, T$, T is the total number of picking corridors.

Each corridor contains K storage positions. p_{tk} represents the k-th storage position of corridor t, where only one pod can be placed, $k = 1, 2, \ldots, K$, $t = 1, 2, \ldots, T$.

w_{tk} represents the shortest average distance between p_{tk} and the picking stations.

$d_{kk'}$ represents the distance between the current position k and the next position k', where the next designated pod is located.

$H_m = \{h_{m1}, h_{m2}, \ldots, h_{mP}\}$ represents the relationship vector between pod m and the products.

$$h_{mi} = \begin{cases} 1, & \text{if pod } m \text{ contains product } i, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

H is the pod-product relation matrix, $H = [H_1, H_2, \ldots, H_m]^T$.

$C_m = \{c_{m1}, c_{m2}, \ldots, c_{mN}\}$ represents the relationship vector between pod m and the orders.

$$c_{mn} = \begin{cases} 1, & \text{if pick order } n \text{ need carry pod } m, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

C is the pod-order relation matrix, $C = [C_1, C_2, \ldots, C_M]^T$.

Decision variable:

$$p_{mtk} = \begin{cases} 1, & \text{if pod } m \text{ is allocated on position } k \text{ in corridor } t, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The objective function is to minimize the total picking distance :

$$\min z_2 = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{m=1}^{M} c_{mn} p_{mtk} (2w_{tk} + d_{kk'}). \tag{14}$$

$2w_{tk}$ is the distance a robot travels to carry the pod located at p_{tk} to the picking station and return the pod back. $d_{kk'}$ is the distance a robot travels to fetch the next pod.

The constraints are presented as follows:

$$\sum_{t=1}^{T} \sum_{k=1}^{K} p_{mtk} = 1, \quad m = 1, 2, \ldots, M, \tag{15}$$

$$\sum_{m=1}^{M} p_{mtk} \leq 1, \quad t = 1, 2, \cdots, T; k = 1, 2, \ldots, K, \tag{16}$$

$$\sum_{m=1}^{M} h_{mi} c_{mn} \geq b_{in}, \quad i = 1, 2, \ldots, P; \; n = 1, 2, \ldots, N, \tag{17}$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} c_{mn} p_{mtk} \leq \left[\frac{\sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn}}{Z} \right], \quad t = 1, 2, \ldots, T; \; Z = 1, 2, \ldots, T, \tag{18}$$

$$c_{mn} = 0, 1, \quad n = 1, 2, \ldots, N; m = 1, 2, \ldots, M, \tag{19}$$

$$p_{mtk} = 0, 1, \quad m = 1, 2, \ldots, M; t = 1, 2, \ldots, T; k = 1, 2, \ldots, K. \tag{20}$$

Equation (15) stipulates that each pod must be placed at a position in the storage area. Equation (16) denotes that each storage position can only be placed by one pod at most. Equation (17) ensures that each kind of product that is contained in the order can be picked up from the pods that are carried to the picking station. Equation (18) represents the workload is balanced among picking corridors by setting a threshold for the total visits times of pods placed in the same corridor. Z can be any integer between 1 and T. When $Z = 1$, the constraint condition is always satisfied, that is, the threshold control does not work. When $Z = T$, the threshold constraint control is the strongest; equations (19) and (20) are the basic constraints of decision variables.

4. Algorithm Design

The integer programming models built in the previous section are nonlinear and have been proved to be NP-hard problems. A large amount of data is also needed to solve the models, so it is difficult to get the optimal solution. Therefore, according to the characteristics of the problem, this paper designs effective hybrid algorithms to solve the optimization problem quickly.

4.1. Algorithm Design for Products Assignment Model. A heuristic algorithm is designed to solve the products assignment model. The steps are as follows:

(1) Step 1: Calculate the required storage layers of each product. According to the historical order data and equation (1), the average demand and the required storage layers of each product are obtained.

(2) Step 2: Calculate the products information matrix RA. According to the historical order data and equation (3), the correlation between two products is calculated, and the products correlation matrix R is obtained. The products information matrix RA is formed by adding the required number of storage layers of each product to the right side of R.

(3) Step 3: Find the maximum correlation value in the products information matrix RA and assign the two corresponding products to one pod. At the same time, reassign the correlation value to 0, update the required storage layers of the two products, and get the updated RA. In the updated RA, find the products most relevant to the above two products, assign the products to the same pod, and update RA. Repeat the process until all the layers of the pod are occupied and then update H. Notice that when the required storage layer of a product is reduced to 0, all the correlation values between the products and other products are reassigned to 0.

(4) Step 4: Recover the correlation information of the products in RA whose required storage layers are not 0. Repeat step 3 for other pods until the required layers of all products are 0.

The pseudo code of the heuristic Algorithm 1 is as follows:

4.2. Algorithmic Design for Pods Assignment Model. The pods assignment problem is similar to the knapsack problem, which can be solved by a simulated annealing algorithm. In order to adapt to the nature of the problem and improve the search efficiency, the simulated annealing algorithm is improved from three aspects: (1) a greedy algorithm based on pods correlation is used to generate the initial solution; (2) the principle of pods with high turnover rate placed near picking stations is used to generate the new solution; and (3) at the same time, the workload among corridors is balanced to reduce system congestion.

```
Input: Historical order data
Output: The pod-product relation matrix H
#Calculate products correlation matrix R
for (i = 1; i ≤ P; i++) do
    for (j = 1; j ≤ P; j++) do
        r (i, j) ← Calculate the correlation between products i and j
    end
end
#Calculate products information matrix RA
    for (i = 1; i ≤ P; i++) do
        RA (i, P+1) ← Calculate the required storage layers for product i
end
#Calculate which product to put on which pod
for (m = 1; m ≤ M; m++) do
    r←1
    product i, j← find (A = = max(max(RA))), then update RA
    r←3
    while r ≤ q do
        Find the most relevant product u in RA and update RA
        update H_m
        if RA (u, P+1) = = 0 then
            RA (u, :) ←0
            RA (:,u) ←0
        end
        if sum (RA (:, P+1)) = = 0 then
            end calculation
        end
        r←r + 1
    end
end
```

Algorithm 1: The heuristic algorithm of products assignment.

4.2.1. Data Preparation. Firstly, we prepare the data needed. According to the pod-product relation matrix H obtained in products assignment stage, the pod-order relation matrix C is constructed based on the principle of maximum set coverage. Then the pods correlation matrix E is constructed, and the pod turnover rate is calculated.

(1) Pod-Order Relation Matrix C. The maximum set coverage strategy is adopted to decide which pod to carry, that is, when one order needs to be picked, the robot carries the pods containing the most types of products in the order to fulfill it. c_{mn} is set to 0 at the beginning. For each order, repeat the following steps to update the pod-order relation matrix C.

Step 1: For the current order n, retrieve the types of unpicked products in the order.

Step 2: Search for the pod that contains the most types of the above products and choose the pod to serve the order.

Step 3: Check whether all the products in order n have been picked, if so, update $c_{1n}, c_{2n}, \ldots, c_{Mn}$ in the pod-order relation matrix C and change the corresponding data from 0 to 1. For example, if pod m is selected to serve order n, c_{mn} is updated to 1; otherwise, return to step 1.

(2) Pods Correlation Matrix E. The correlation between two pods relates with the proportion of the common orders they can fulfill, which is measured by cosine correlation as shown in equation (21). C_m is the relationship vector between pod m and the orders, and C_g is the relationship vector between pod g and the orders. $C_m \cdot C_g$ represents the number of the common orders pod m and g can fulfill, respectively, and $|C_m|$ and $|C_g|$ is the number of the orders pod m and g can fulfill, respectively.

$$e_{mg} = \cos\left(C_m, C_g\right) = \frac{C_m \cdot C_g}{\left|C_m\right| \cdot \left|C_g\right|}. \tag{21}$$

(3) Pod Turnover Rate Fre_m. The pod turnover rate Fre_m is calculated using the historical order information and pod-order relation matrix C.

$$\mathrm{Fre}_m = \frac{\sum_{n=1}^{N} c_{mn}}{\sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn}}. \tag{22}$$

4.2.2. Two-Stage Hybrid Algorithm. A two-stage hybrid algorithm combining greedy algorithm and simulated annealing is designed to solve the pods assignment model.

Stage I: according to the pods correlation matrix, a greedy algorithm is designed to generate the initial solution of pods assignment.

Stage II: the simulated annealing algorithm is improved to optimize the solution and get the final pods assignment results.

The steps of each stage are as follows:

(1) Stage I. Generate initial solution: The purpose of this stage is to generate the initial solution of the subsequent simulated annealing algorithm considering the correlation of pods. The specific steps of the algorithm are as follows:

Step 1: According to the pods correlation matrix E, select a pair of pods with the maximum correlation.

Step 2: Select a pair of positions with the minimum distance in the position distance matrix D. If there are more than one pair of positions with minimum distance, choose the pair closer to the picking stations.

Step 3: Check whether the total pod turnover rate in the corridor will exceed the threshold value after the two pods in step 1 are placed in the two positions in step 2. If not, put the pods there. Then update the pods correlation matrix and distance matrix by reassigning the values corresponding to the selected pods and positions to -1. Otherwise, select the next position pair with a smaller distance.

Step 4: Repeat steps 1 to 3 until each pod is assigned to a storage position and the initial solution is obtained.

(2) Stage II. Optimize the initial solution: In this stage, the simulated annealing algorithm is improved to optimize the initial solution considering pod turnover rate and the balance among corridors.

Step 1. Initialization: Generate initial solutions $P = P_0$ from the previous stage and set the initial temperature T_0, which is big enough. Set $T = T_0$ and determine the number of iterations, which is metropolis chain length L.

Step 2: For current temperature T, switch the positions of pods to get new solution P' according to pod turnover rate and workload balance among corridors. Select some pairs of pods with a higher turnover rate and switch their positions with the pods with a lower turnover rate but closer to the picking stations. The purpose is to move the pods with a high turnover rate closer to the picking stations to reduce travel distance. Then check whether the workload balance constraint (equation (18)) is still satisfied after the pods movement to avoid congestion. If it is satisfied, a new solution P' is generated; otherwise, select other pods to switch positions to get a new solution P'.

Step 3: Compute the increment of P': $\Delta f = f(P') - f(P)$. $f(P)$ is the cost function of P. If $\Delta f < 0$, accept P' as a new current solution, set $P = P'$, $l = l + 1$. If $\Delta f \geq 0$, calculate acceptance probability $\exp(-\Delta f/T)$ of P'. That is, generate a random number rand with uniform distribution in (0,1) intervals; if $\exp(-\Delta f/T) > \text{rand}$, P' is accepted as a new current solution and set $P = P'$ and $l = l + 1$. Check if l reach the maximum number of iteration, if so, go to step 4; otherwise, return to step 2.

Step 4: According to the attenuation function, set $T = T * \Delta T$ (ΔT is generally 0.95~0.99) to reduce temperature. Check if the minimum temperature $T_{\min}$ is reached, if so, output the current solution P as the best solution, and end the program; otherwise, return to step 2.

5. Simulation Analysis

The storage assignment models and algorithms proposed in this paper are verified by the data of an e-commerce company. The company sells 250 kinds of products online and uses RMFS for order picking. The products are stored on mobile pods, each of which is divided into 8 layers and the maximum capacity of each layer is 70 units. There are 10 mobile robots and 6 picking workstations in the warehouse. Suppose the travel speed of robots is 1 m/s and the picking time of one item is 5 s. We compare our proposed method with other methods in products assignment, pods assignment, and the overall picking efficiency.

5.1. Simulation Results of Products Assignment. The simulation program of the storage assignment method proposed in this paper is constructed using MATLAB. The optimal result for product assignment is obtained using the data of 2,000 history orders of the e-commerce company. The result for the products assignment is shown in Table 1.

We can see from Table 1 that one kind of product is assigned to different pods; for example, product 131 is allocated to pods 8 and 9, that is, when product 131 needs to be picked, there are two choices of pod visits, which can reduce system congestion and improve the efficiency of order picking.

The heuristic algorithm proposed in the paper is compared with the random assignment method and the Apriori algorithm. Random assignment is to store the products on the pods randomly without considering the correlation of products. Apriori algorithm was first proposed by Agrawal and Shafer [22] and has been widely used in storage assignments [23]. The basic idea is to count the frequency of multiple products in the same order and put the products with high frequency on the same pod. The methods were compared under different order size ($n = 500$, 1,000, 1,500, and 2,000) in the number of pod visits. The results are shown in Figure 2.

It can be found from Figure 2 that the times of pod visits of the heuristic algorithm are the smallest among the three methods under different order sizes, 32.7%~36.6% less than random algorithm and 16.9%~18.6% less than the Apriori algorithm. Besides, with the increase of order size, the growth rate of pod visits in the proposed heuristic algorithm slows down, which means that the bigger the order size is, the greater the possibility of the correlation between

Table 1: Products assignment result.

Pod number	Product number							
1	26	203	100	103	137	96	247	238
2	175	41	165	250	33	50	185	25
3	8	208	49	14	189	112	19	153
4	142	62	159	58	128	36	60	237
5	121	139	155	111	148	221	141	100
6	144	236	195	107	232	135	197	45
7	43	82	93	102	66	50	232	219
8	63	35	7	187	5	45	193	131
9	213	150	95	125	79	149	131	16
...								
164	202	180	187	34	1	144	99	250
165	161	190	55	208	118	7	—	—

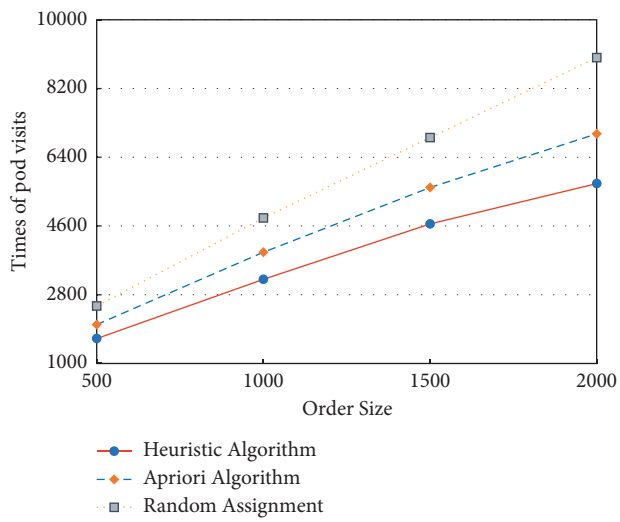


Figure 2: Times of pod visits using different algorithms.

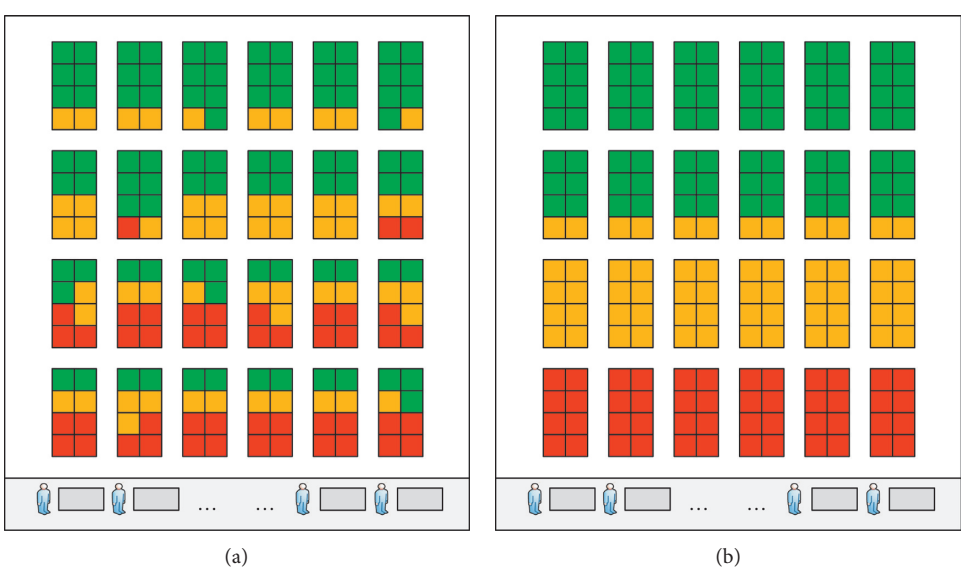


Figure 3: Comparison of workload balance using different pods assignment methods: (a) pods assignment using the proposed method and (b) pods assignment using ABC classification method.

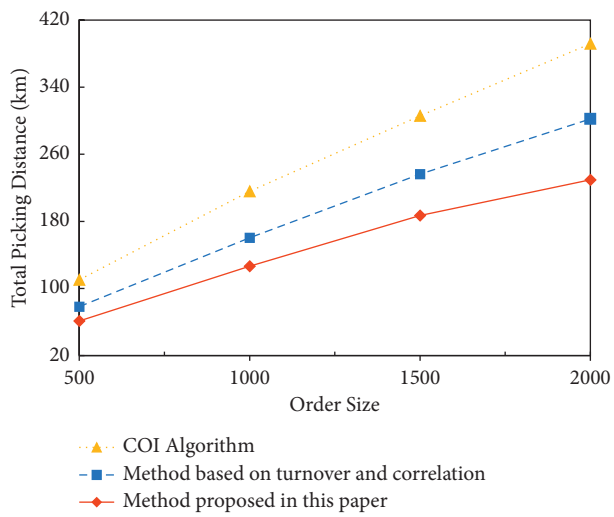


Figure 4: Total piking distance of different methods.

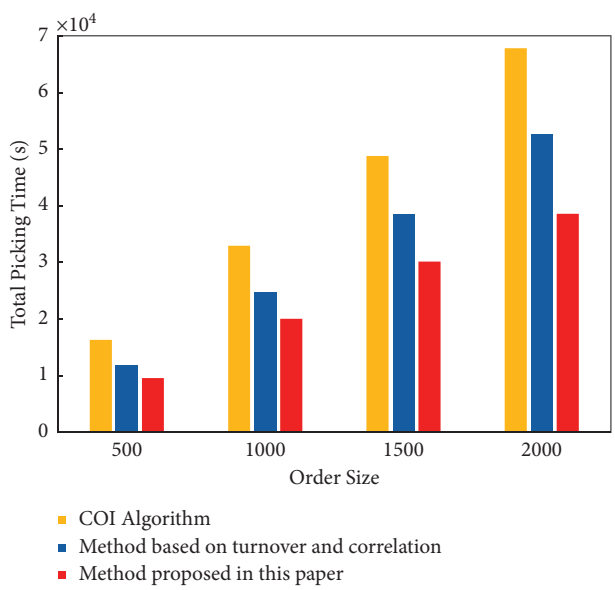


Figure 5: Total picking time of different methods.

products, so the advantage of the heuristic algorithm is more obvious. The result shows that maximizing the correlations of the products in one pod can effectively reduce the times of pod visits and improve the picking efficiency.

5.2. Simulation Results of Pods Assignment. In order to alleviate system congestion, the method of balancing the workload among corridors is adopted, and the pod assignment result is shown in Figure 3(a). The typical pod layout using the traditional ABC classification method [11] is shown in Figure 3(b). In the figures, the red ones represent pods with a high turnover rate, accounting for about 25% of the total number of pods; the yellow ones represent pods with a medium turnover rate, accounting for about 30% of the total number of pods; and the green ones represent pods with a low turnover rate, accounting for about 45% of the total number of pods.

We can see from Figure 3(b) that all the pods with the high turnover rate (the red ones) are assigned to the storage positions in the first row, which is closest to the picking stations, and most pods with the medium turnover rate (the yellow ones) are assigned to the storage positions in the second row. This will lead to a high total access rate of pods in this area, resulting in congestion of picking robots in the corridors nearby. In Figure 3(a), the red and yellow pods are properly scattered in different areas. We can see that the red pods account for only about 48% of the total positions in the first row. In this way, the workload of corridors can be effectively balanced, and system congestion can be alleviated.

5.3. Comparative Analysis of the Overall Picking Efficiency. In order to validate the efficiency of the storage assignment models and algorithm proposed in this paper, it is compared with other storage assignment methods commonly used

(COI algorithm in reference [24] and the algorithm just based on turnover and product correlation in reference [21]) from total picking distance and picking time.

The COI algorithm is also called the cube-per-order index algorithm, which considers the volume and the turnover rate of products [24]. The method used in reference [21] is mainly considered the correlation and turnover rate of products but does not consider the correlation between pods. The method proposed in this paper considers the correlation of products, the turnover rate, and the correlation of pods, as well as the workload balance of corridors. The experimental results of the total picking distance and picking time using the three methods are shown in Figures 4 and 5, respectively.

Figure 4 shows that the total picking distance using the method proposed in this paper is much shorter than that of the COI algorithm and the method, only considering the correlation and the turnover rate of products. And with the increase of order size, the growth rate of picking distance in our proposed method slows down, while the growth rate of other methods shows a linear growth.

Figure 5 shows that the total picking time using the storage assignment method proposed in this paper is much lower than that of the COI algorithm and methods based on correlation and turnover rate of products. And with the increase of order size, the gap of the picking time among methods increases.

Simulation results above show that the storage assignment method proposed in this paper can significantly improve the order picking efficiency by alleviating system congestion and reducing picking distance. Besides, with the increase of order size, the improvement of this method is more obvious.

6. Conclusions and Prospects

This paper studies the joint optimization of products assignment and pods assignment in RMFS. A two-stage mathematical model is established considering the correlation, turnover rate of products and pods, as well as system congestion factors. Hybrid algorithms, including heuristic algorithm, greedy algorithm, and improved simulated annealing algorithm, are designed to solve the models. Simulation experiments are carried out based on the historical data of an e-commerce company. The method proposed in this paper is verified by comparing with other commonly used storage assignment methods from times of pod visits, system congestion situation, total picking distance, and time. Experimental results show that the storage assignment method proposed in the paper significantly improves the efficiency of order picking.

The paper mainly solves the static storage assignment problem considering the characteristics of the operation process of RMFS. However, due to the complexity and dynamics of system scheduling, there are still some problems that need to be further studied. For example, in practical application, the characteristics of orders will change over time, so it is necessary to adjust the storage allocation dynamically according to the changes. Thus, the dynamic storage assignment optimization can further improve the picking efficiency and is a very promising research direction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The paper was funded by National Natural Science Foundation of China (72101033 and 71831001), Beijing Key Laboratory of Intelligent Logistics Systems (BZ0211), Canal Plan-Youth Top-notch Talent Project of Beijing Tongzhou District (YHQN2017014), Ningxia Science and Technology Key Research and Development Project (2018BEG03003), and Scientific Research Program of Beijing Municipal Commission of Education (KM202110037003).

References

[1] T. Lamballais, D. Roy, and M. B. M. De Koster, "Estimating performance in a robotic mobile fulfillment system," *European Journal of Operational Research*, vol. 256, no. 3, pp. 976–990, 2017.

[2] N. Boysen, R. De Koster, and F. Weidinger, "Warehousing in the E-commerce era: a survey," *European Journal of Operational Research*, vol. 277, no. 2, pp. 396–411, 2019.

[3] X. M. Luo, X. Y. Xia, and J. B. Li, "Study on the batch of warehouse orders of FMCG E-commerce," *Journal of Systems Science and Mathematical Sciences*, vol. 36, no. 6, pp. 847–859, 2016.

[4] R. P. Yuan, H. L. Wang, L. R. Sun, and J. T. Li, "Research on the task scheduling of part-to-picker order picking system based on logistics AGV," *Operations Research and Management Science*, vol. 27, no. 10, pp. 133–138, 2018.

[5] Z. Zheng, Q. Guo, J. Chen, and P. J. Yuan, "Collision-free route planning for multiple AGVs in automated warehouse based on collision classification," *IEEE Access*, vol. 6, pp. 2602-2603, 2018.

[6] R. de Koster, T. Le-Duc, and K. J. Roodbergen, "Design and control of warehouse order picking: a literature review," *European Journal of Operational Research*, vol. 182, no. 2, pp. 481–501, 2007.

[7] W. H. Hausman, L. B. Schwarz, and S. C. Graves, "Optimal storage assignment in automatic warehousing systems," *Management Science*, vol. 22, no. 6, pp. 629–638, 1976.

[8] Y. D. Li, "Model and algorithm for cartonization and slotting optimization simultaneously in wave-picking zone-based system," *Systems Engineering-Theory & Practice*, vol. 33, no. 5, pp. 1269–1276, 2013.

[9] F. Caron, G. Marchet, and A. Perego, "Routing policies and COI-based storage policies in picker-to-part systems," *International Journal of Production Research*, vol. 36, no. 3, pp. 713–732, 1998.

[10] J. B. Li, G. Y. Yang, and F. Chen, "Retail warehouse center storage location assignment research for E-commerce,"

Industrial Engineering and Management, vol. 18, no. 4, pp. 102–108, 2013.

[11] J. Li, M. Moghaddam, and S. Y. Nof, "Dynamic storage assignment with product affinity and ABC classification—a case study," *International Journal of Advanced Manufacturing Technology*, vol. 84, no. 9, pp. 1–16, 2016.

[12] X. B. Xu and Z. Q. Ma, "Robotic mobile fulfillment systems: state-of-the-art and prospects," *Acta Automatica Sinica*, vol. 46, no. 9, pp. 1–25, 2020.

[13] N. Boysen, D. Briskorn, and S. Emde, "Parts-to-picker based order processing in a rack-moving mobile robots environment," *European Journal of Operational Research*, vol. 262, no. 2, pp. 550–562, 2017.

[14] R. Yuan, H. Wang, and J. Li, "The pod assignment model and algorithm in robotic mobile fulfillment systems," in *Proceedings of the 2019 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pp. 99–103, Zhengzhou, China, November 2019.

[15] X Xi, C. Liu, and L. Miao, "Storage assignment and order batching problem in Kiva mobile fulfilment system," *Engineering Optimization*, vol. 50, no. 11, pp. 1941–1962, 2018.

[16] R. Yuan, S. C. Graves, and T. Cezik, "Velocity-based storage assignment in semi-automated storage systems," *Production and Operations Management*, vol. 28, no. 2, pp. 354–373, 2019.

[17] D. Roy, S. Nigam, R. De Koster, I. Adan, and J. Resing, "Robot-storage zone assignment strategies in mobile fulfillment systems," *Transportation Research Part E: Logistics and Transportation Review*, vol. 122, pp. 119–142, 2019.

[18] F. Weidinger and N. Boysen, "Scattered storage: how to distribute stock keeping units all around a mixed-shelves warehouse," *Transportation Science*, vol. 52, no. 6, pp. 1412–1427, 2018.

[19] F. Weidinger, N. Boysen, and D. Briskorn, "Storage assignment with rack-moving mobile robots in KIVA warehouses," *Transportation Science*, vol. 52, no. 6, pp. 1479–1495, 2018.

[20] T. Lamballais, D. Roy, and R. B. M. De Koster, "Inventory allocation in robotic mobile fulfillment systems," *IISE Transactions*, vol. 52, no. 1, pp. 1–17, 2020.

[21] R. Yuan, T. Cezik, and S. C. Graves, "Stowage decisions in multi-zone storage systems," *International Journal of Production Research*, vol. 56, no. 1-2, pp. 333–343, 2018.

[22] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 962–969, 1996.

[23] D. Ming-Huang Chiang, C.-P. Lin, and M.-C. Chen, "Data mining based storage assignment heuristics for travel distance reduction," *Expert Systems*, vol. 31, no. 1, pp. 81–90, 2014.

[24] M. Li and Y. Zhang, "A study of workload distribution and COI-based storage policies," *Industrial Engineering*, vol. 18, no. 1, pp. 37–41, 2015.

Hindawi
Complexity
Volume 2021, Article ID 4795396, 10 pages
https://doi.org/10.1155/2021/4795396

Research Article

Research on Surface Defect Detection of Rare-Earth Magnetic Materials Based on Improved SSD

Bin Zhang, Shuqi Fang, and Zhixi Li

School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China

Correspondence should be addressed to Shuqi Fang; fsq1611@sina.com

Received 2 September 2021; Accepted 26 October 2021; Published 13 November 2021

Academic Editor: Long Wang

In order to overcome the limitation of manual visual inspection of surface defects of rare-earth magnetic materials and increase production efficiency of traditional rare-earth enterprises, a detection method based on improved SSD (Single Shot Detector) is proposed. The SSD model is improved from two aspects for better performance in the detection of small defects. First of all, the multiscale receptive field module is embedded into the backbone network of the algorithm to improve the feature extraction ability of the model. Secondly, the interlayer feature fusion strategy of bidirectional feature pyramid in PANet (path aggregation network) is integrated into the model. In order to enhance the detection ability of the model, the high-level semantic information is strengthened by an efficient channel attention mechanism. The detection speed of the improved SSD algorithm is 55FPS, and the mAP (mean Average Precision) is up to 83.65%, which is 3.41% higher than of the original SSD algorithm, and the ability to identify small defects is significantly improved.

1. Introduction

Computer vision technology has very important application and theoretical significance in defect detection of rare-earth magnetic materials. At present, most magnetic material manufacturers in China use the traditional method of manual sorting to classify their products. It is not only time consuming and laborious, but also increasing the cost of products. At the same time, an overreliance on manual sorting raises a problem that cannot be ignored: operators in the sorting process will inevitably cause visual and physical fatigue. Lack of attention for a long time will directly lead to an increase in the rate of false and missed detection. With the development of computer technology and image processing technology, automatic detection based on image processing technology is an inevitable trend. The visual-based surface quality inspection method has very important research value, for example, the steel surface damage detection [1], the railway track defect detection [2], the wafer electron microscope image defect detection [3], and a wide range of applications in other fields [4, 5].

In previous studies, researchers usually use traditional machine learning methods to carry out a series of studies in the field of defect recognition. Chang et al. [6] proposed a defect detection method based on LVQ neural network. Aiming at recognizing the defects on the LED wafer, the geometric features and texture features are extracted by analyzing each ROI for detection. Yazdchi et al. [7] presented a texture segmentation technology based on multifractal dimensions to detect steel surface defects. Fourier analysis in the time domain is utilized to detach the defective region from the image and specify its position. The features of multifractal dimension, mean and variance of column variance, and maximum value of principal component vector are extracted as the input of the three-layer and multilayer perceptron classifier. Demetgul et al. introduced a KNN-based fabric defect classification algorithm, which uses wavelet transform, threshold, and morphology for image processing [8]. Jeon et al. proposed a method based on wavelet reconstruction to detect cracks at the corners of billets [9]. Sun et al. proposed a new single-shot target detection network with a mask prediction branch. They proposed an improved RFB module to increase the size of receptive field for better performance on small object detection [10]. Although these methods have achieved good results, they usually require explicit feature extraction, which leads to unsatisfactory generalization of detection methods.

In recent years, with the development of deep learning, its outstanding performance in the field of image processing has attracted people's attention. Deep learning avoids the manual extraction of features and realizes automatic feature extraction. Combined with CNN (Convolutional Neural Network), relevant researchers have carried out a series of research work in the field of defect detection of motor magnetic tiles [11–13]. Due to the various types of rare-earth magnets, there is currently little work on the automatic detection of rectangular rare-earth magnetic patches domestically and overseas. Therefore, this research combines machine vision and deep learning and other related theories and adopts SSD (Single Shot Detector) [14] as the prototype of defective target detection and improves it. The improved SSD algorithm has a better recognition effect for small defects, and the detection accuracy is significantly improved compared with the original SSD algorithm.

2. SSD Target Detection Algorithm Model

The industrial defect detection task requires high detection speed. Compared with two-stage model such as Faster R-CNN [15], the one-stage target detection algorithm is more suitable for actual production and practice. This research selects SSD as the basic network and improves it. The improved SSD can detect defects on the surface of the magnet more accurately and has a higher recognition rate for small targets. It provides a feasible method for the industrial scene where defect detection is needed.

2.1. SSD Algorithm Principle. SSD is a one-stage target detection algorithm that borrows ideas from the anchors mechanism in Faster R-CNN and uses multiscale feature map for detection. The backbone feature extraction network of the SSD algorithm is the classic VGG-16 [16] convolutional neural network, which has been modified. The SSD algorithm replaces fc6 and fc7 in the VGG-16 with a convolutional layer structure and, on this basis, adds four additional convolutional layers to obtain more feature maps for detection. The structure of the SSD network is shown in Figure 1.

The feature extraction network uses feature maps of different sizes output by the specific effective feature layers for prediction and outputs the location information of the target and the confidence of the category. Among them, the output feature map scale of the Conv4_3 layer structure is 38×38, the output feature map scale of the fc7 layer structure is 19×19, and the output feature map scale of the Conv6_2 layer structure is 10×10. The output sizes of Conv7_2, Conv8_2, and Conv9_2 layers are 5×5, 3×3, and 1×1, respectively. Six feature maps of different scales can detect targets of different sizes. The front feature map has a smaller receptive field and is mainly used to identify small targets, while the back feature map has a larger receptive field and is used to detect large targets. Each pixel of the feature map has several default boxes, as shown in Figure 2. The proportion between the size of the default box and the picture can be expressed as follows:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), \quad k \in [1, m]. \tag{1}$$

Among them, m is the number of valid feature map (excluding Conv4_3). $S_{\max}$ and $S_{\min}$ represent maximum and minimum proportions, with values of 0.9 and 0.2, respectively. The default box scale of the first effective feature layer is independently set to 0.1, and the scale is 30 on the 300×300 original image. The aspect ratio of the default box is set as $\alpha_r = \{1, 2, 3, 1/2, 1/3\}$; then, the width and height of the default box can be obtained by the following formula:

$$\begin{aligned} w_k &= s_k \sqrt{\alpha_r}, \\ h_k &= \frac{s_k}{\sqrt{\alpha_r}}. \end{aligned} \tag{2}$$

It is special when $\alpha_r = 1$. In addition to the default box with the scale of s_k, there is another scale of s_k', and its calculation formula is as follows:

$$s_k' = \sqrt{s_k s_{k+1}}. \tag{3}$$

Since the feature map output by Conv4_3 is relatively special, the default boxes with an aspect ratio of 3 and 1/3 are not used. From this, the number of default boxes for each pixel of the feature layer as the center is 4, and the remaining layers are 6. So the number of SSD default boxes is 8732. Due to the large number of default boxes obtained and the limited number of GT (Ground Truth) matched by IOU value, the SSD algorithm uses hard negative mining to sample negative samples. Those with large errors are taken as negative samples, while those with small errors are considered as positive samples. The final sample ratio of positive and negative ones is controlled at about 1 : 3.

The bounding box of SSD algorithm is fine-tuned by the default box, and its essence is a regression task. The offset value of each bounding box to the default box is inferred, and the final target position information is obtained through a transformation. The transformation process consists of two parts: encoding and decoding [14]. The position of the default box is represented by $d_b = (d_b^{cx}, d_b^{cy}, d_b^{w}, d_b^{h})$, which respectively correspond to the coordinate of the center point and the width and height of the default box. The position of the bounding box is expressed as $b_b = (d_b^{cx}, d_b^{cy}, d_b^{w}, d_b^{h})$, which respectively correspond to the coordinates of the center point and the width and height of the bounding box.

2.2. Problems with SSD Algorithm. Compared with Faster R-CNN and YOLO [17] algorithm, SSD algorithm has higher detection accuracy and detection speed, which benefits from its good design ideas. The SSD algorithm directly acts on the output information of the effective feature layer at different scales on the detection layer to generate the bounding box and the confidence of the detection target. The small target is detected by the shallow structure such as Conv4_3, and the large target is detected by the top effective feature layer such as Conv8_2.

The SSD algorithm uses a pyramidal sampling structure to express the semantic information of the image. The shallow

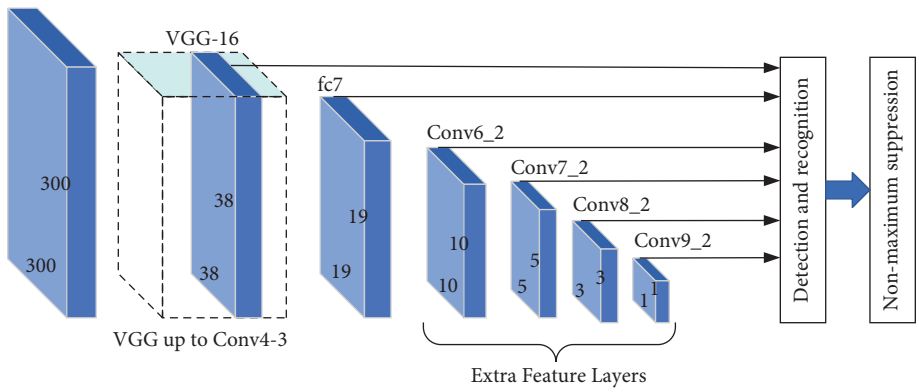


Figure 1: SSD algorithm structure.

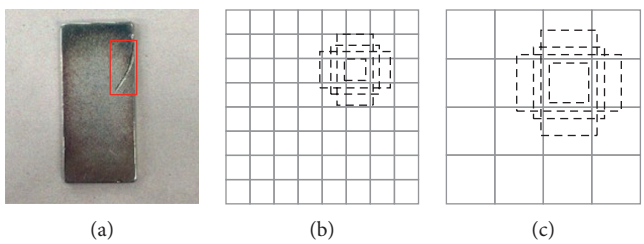


Figure 2: Comparison of the default box and the real box under different scale feature graphs. (a) Image with GT boxes. (b) 8×8 feature map. (c) 4×4 feature map.

information has high pixels and has a strong ability to locate targets. However, the small target features obtained by the low-level convolutional layer lack semantic information. The nonlinear degree of the features is insufficient, and the features of their scales are not merged, resulting in the loss of part of the detailed information of the model learning. This will not be able to effectively use the contextual information related to the model, which is not conducive to the detection of related targets on the image. These deficiencies will lead to the failure to identify small-size defects in the defect detection process of rare-earth magnetic materials.

After multiple convolutions, the high-level feature maps for detecting large targets have larger receptive fields and the learned information is more abstract, but its lower resolution may still cause missed targets. The high-level semantic information in the SSD algorithm has not been enhanced, and there is still room for improvement.

3. Defect Detection Method of Rare-Earth Magnetic Patch Based on Improved SSD

3.1. PANet Bidirectional Feature Pyramid Structure. FPN (Feature Pyramid Network) [18] is a network topology structure that gathers high-level features and low-level features. It is different from the pyramid hierarchical sampling structure in SSD algorithm, so that the information learned by the model not only retains the location information but also contains stronger semantic information. Its network topology is shown in Figure 3. According to the topological structure, FPN performs upsampling (such as bilinear interpolation and deconvolution) on the high-level feature map and predicts the low-level semantic information after horizontal superposition, so that the feature image has strong semantic information and achieves the effect of feature fusion.

Inspired by FPN, Shu Liu et al. proposed a path aggregation network called PANet; it was first applied to instance segmentation tasks and achieved excellent results [19]. Its framework is shown in Figure 4. It can be seen from the figure that PANet has improved the FPN structure, adopting the structure of the top-down and bottom-up bidirectional feature pyramid fusion and combining the downsampling on the basis of the original top-down horizontal superposition of FPN, adding the bottom-up path enhancement strategy. Its purpose is to shorten the information path and use the accurate positioning of the shallow level. This effectively improve the utilization rate of the underlying characteristics of the network.

3.2. Parallel Multiscale Convolutional Layer. At present, most of the algorithms with excellent detection results rely on backbone networks with strong feature extraction capabilities, such as the improved DSSD algorithm based on SSD [14]. This network replaces the original VGG-16 feature extraction network with a deeper network like ResNet101 and adds a DSSD network layer on this basis. These changes significantly improve the accuracy of the model, but require a lot of computational cost, and the detection speed is much lower than that of SSD. Some SSD series detection algorithms based on lightweight backbone networks, such as

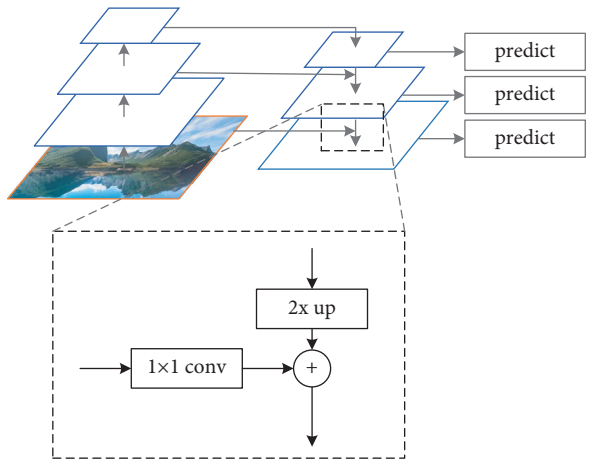


Figure 3: FPN structure.

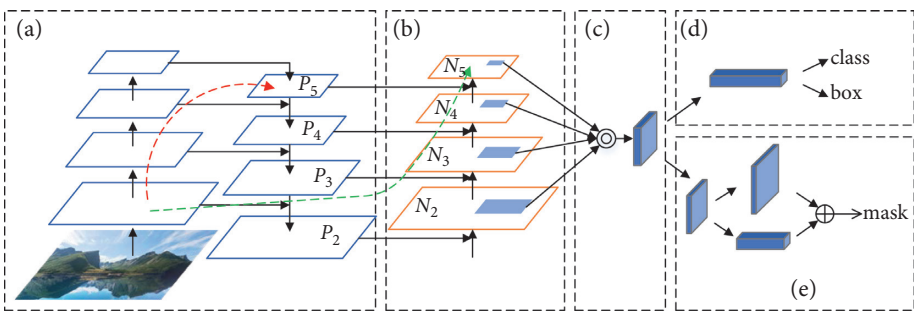


Figure 4: PANet structure.

MobileNet-SSD model [20], have relatively low detection accuracy despite their fast computing speed.

In order to avoid a large number of operations caused by deep convolutional neural network, Liu et al. [21] added a receptive field module to the top of VGG-16 network of SSD algorithm to obtain performance gains and control the computational cost within a controllable range. The structure of the receptive field module is similar to that of InceptionV4, and multiscale information is obtained by using parallel convolution kernels with different receptive fields. In the receptive field module, according to the size of the receptive field, each convolution branch performs a dilated convolution with different expansion coefficients to simulate the group receptive field mechanism in human vision. It also uses the residual learning idea and adds jump connections. This paper uses the receptive field module to enhance the VGG-16 backbone network and replaces the 5×5 convolutional layer in the receptive field module with two consecutively stacked 3×3 convolutional layers to achieve the same receptive field and reduce network parameters, as shown in Figure 5.

3.3. Efficient Channel Attention Mechanism. In the reasoning process of the SSD algorithm, the identification of medium and large targets is completed by the high-level feature map. Due to its low output resolution, the target may also be missed. In this paper, an efficient channel attention module ECA [22] is introduced to optimize the high-level and low-resolution semantic information graph output by SSD algorithm. The obtained information graph with low resolution and high-semantic features can better identify medium and large defect targets, so as to achieve the purpose of improving mAP (mean Average Precision).

The attention mechanism is to imitate the signal processing mechanism of the human brain. This strategy has good adaptability and enhancement for computer vision tasks. ECA structure also obtains the importance degree of different characteristic channels through supervised learning. The nonlinear full-connection layer in SENet [23] is improved to avoid the impact of dimensionality reduction on the attention of the learning channel and ensure the efficiency and computing effect of the network. This module can achieve cross-channel interaction without dimensionality reduction. That is, after global average pooling of the output feature graph, each channel interacts with its K neighbors through rapid one-dimensional convolution. The topology of ECA module is shown in Figure 6.

The correlation of feature channels can be expressed as follows:

$$\omega = \sigma\left(\mathrm{CID}_k\left(y\right)\right), \tag{4}$$

where CID represents one-dimensional convolution operation, k represents the coverage of cross-channel interaction, and its size is proportional to the number of channel dimensions. When the channel dimension C is constant, the value of k can be determined by the following formula:

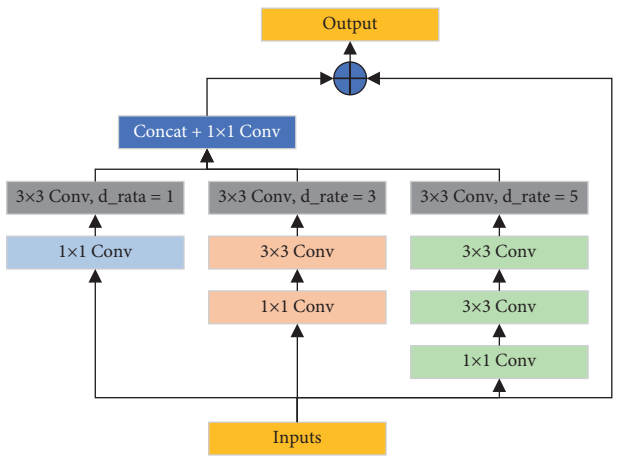


Figure 5: Structure of the multiscale receptive field layer.

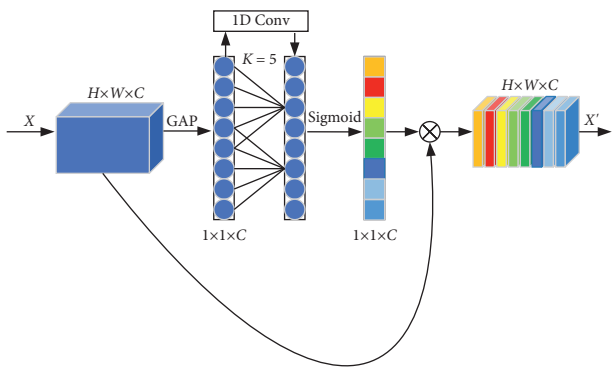


Figure 6: Structure diagram of the efficient channel attention module.

$$k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{\text{odd}}. \tag{5}$$

Here, odd is the closest odd number and γ and b are set to 2 and 1, respectively.

3.4. Establishment of the Improved SSD Network Model. In order to make the original SSD algorithm better solve the problem of small-size defect identification and improve recognition effect, the SSD algorithm is improved based on the relevant theoretical basis analyzed in the above section. The framework of the defect detection method for rare-earth magnetic materials based on the improved SSD model is shown in Figure 7.

It can be seen from the figure that the improved SSD algorithm replaces the Conv6 and Conv7 convolutional layers of the original backbone extraction network by embedding multiscale receptive field modules Block_1, Block_2, and Block_3, thereby improving the ability of the VGG-16 backbone network to extract features and not affecting the inference speed of the detection algorithm. Conv8 and Conv9 layers are not modified. The effective feature layer dimension information output by the embedded Block_1 receptive field module is $19 \times 19 \times 1024$, which has the same size as the output of the fc7 layer and can extract more effective feature information. Therefore, this study extracts the features of the convolutional layer to replace the output of the fc7 layer and uses the PANet-based bidirectional feature pyramid fusion with the Con4_3 and Con3_1 layer structure to make full use of the context information between different levels. The final output is 38×38 and 19×19 effective feature layers. The network structure of feature fusion between layers is shown in Figure 8. Among them, Block_1 adopts spatial pyramid pooling to increase the receptive field. After the interlayer fusion between the top-level feature layer and the low-level feature layer, alternate convolution of 1×1 and 3×3 convolutional layers is helpful to reduce the amount of parameters and extract very effective features. The deep effective feature layers Block_2, Block_3, Conv8_2, and Conv9_2 output four sizes of low-resolution and high-semantic information maps through the ECA channel attention mechanism network.

4. Rare-Earth Magnetic Patch Data Set and Experimental Analysis

4.1. Data Set. The image data sets used in this paper were all collected from Guangxi Jinyuan Rare Earth Co., Ltd. (Guangxi Nonferrous Metals Group). The data sets were manually classified with the assistance of sorting workers and photographed on site with industrial cameras. The focus of the research in this paper is four types of common defect data samples with category and location information, including unfilled corner, line mark, deformation, and crack, as shown in Figure 9. The defect targets are marked using LabeIImg labeling software before the experiment. A total of 1534 data samples were labeled and divided into training set, verification set, and test set in a ratio of 8 : 1 : 1. After image annotation, XML files including image number, category, location of target marker box, and other information will be obtained for model training.

4.2. Network Model Training and Experimental Parameter Setting. The loss function of the improved SSD algorithm consists of two parts, namely the classification confidence loss function (Conf) associated with the target category and the location loss function (Loc) associated with the bounding box regression, as shown in

$$L(x, c, l, g) = \frac{1}{N}\left(L_{\text{Conf}}(x, c)\right) + \alpha L_{\text{Loc}}(x, l, g). \tag{6}$$

Among them, x represents the matching result; c and l respectively represent the confidence of the classification and the position of the bounding box; α is the scale factor, which is used to adjust the ratio of the bit loss function to the confidence loss function; g is the true box label; N represents the number of default boxes that match the actual box.

The confidence loss function can be expressed as follows:

$$L_{\text{Conf}}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log\left(\hat{c}_i^{p}\right) - \sum_{i \in Neg} \log\left(\hat{c}_i^{0}\right). \tag{7}$$

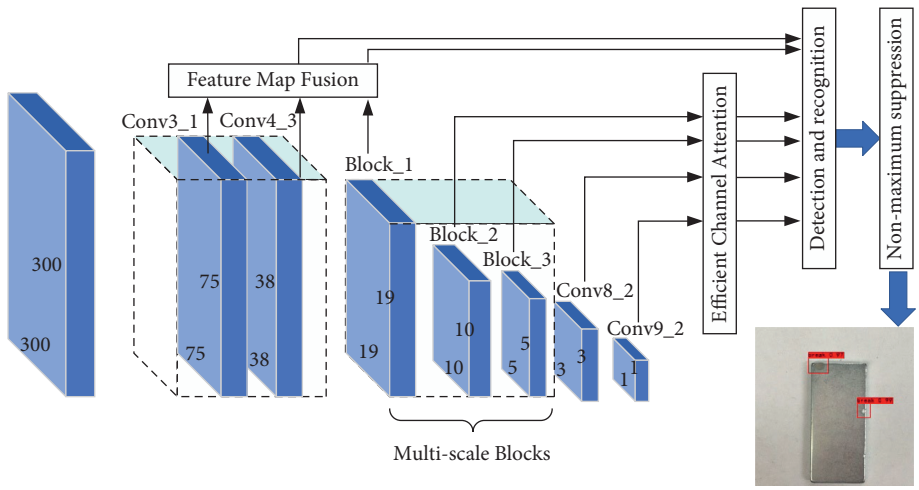


Figure 7: Improved SSD algorithm detection framework.

Among them, Pos represents the number of positive samples in the bounding box, Neg represents the number of negative samples in the bounding box, j represents the j-th true box, i represents the i-th bounding box, and x_{ij}^{p} represents whether it belongs to the p category. c_i^p represents the confidence of the corresponding category of the i-th bounding box, and $\hat{c}_i^p$ represents the confidence of the background, which is calculated by the following formula:

$$\hat{c}_i^p = \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)}. \tag{8}$$

The location loss function can be expressed as follows:

$$L_{\text{Loc}}(x, l, g) = \sum_{i \in Pos}^{N} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^{k} \text{smooth}_{L1}\left(l_i^m - g_j^m\right). \tag{9}$$

Among them, k represents the category and $\text{smooth}_{L1}(x)$ represents the smooth L_1 norm,

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1, \\ |x| - 0.5, & \text{otherwise}. \end{cases} \tag{10}$$

All experiments in this paper are based on the Ubuntu 18.04 operating system and the NVIDIA high-speed hardware computing platform with 1 × GeForce RTX 2080TI graphics card. The improved SSD target detection algorithm is implemented by the deep learning framework PyTorch 1.5.1 and python 3.6. The image data are all downsampled to 300 × 300 resolution as the input of the network. The parameters of this model are set as follows: the batch size of one training sample is 24, and the data set is iterated for 80,000 times in total. The stochastic gradient descent method is used as the network optimizer. The learning rate is initialized to 1E-3 and set to 1E-4 and 1E-5 respectively when the iteration reaches 60,000 and 70,000 times. Momentum is set to 0.9. In order to make the training process of the model more stable, the training of the model adopts the strategy of warm up learning rate automatic adjustment [24]. At the same time, the idea of transfer learning was introduced to load the weight parameters before the Conv4_3 layer which had been trained in the VGG-16 network to accelerate the training of the network. The curve of the loss function is shown in Figure 10.

4.3. Analysis of Experimental Results. In target detection, mAP index is usually used to evaluate the comprehensive detection performance of a model, and the calculation of map value is related to the special-recall carve for each detection category. The abscissa and ordinate of the curve represent the recall rate and accuracy, respectively, and the area enclosed below is the average precision (AP). mAP is calculated by adding the AP values of each category and averaging them. Different from the classification task, the IOU values of the bounding box and the real box are used as the boundary to define the positive and negative samples in target detection. Values greater than the IOU are called positive samples, values less than the IOU are called negative samples, and the IOU value is usually set to 0.5. The test results of the test set in the rare-earth magnetic patch data set and the performance parameters before and after improvement are shown in Table 1.

According to the test results, the improved SSD algorithm in this paper has a 3.41% improvement in mAP compared with the original version of the SSD algorithm. This is mainly because the algorithm in this paper improves the ability of defect feature extraction and adopts the strategy of feature fusion between layers. The full use of contextual semantic information makes the rate of missed detection lower for small defects such as unfilled corners, which makes up for the deficiency of SSD algorithm in small target recognition ability to a certain extent. In addition to the enhancement of the backbone network embedded receptive field module, this algorithm also introduces the ECA network module to the low-resolution effective feature map of the top layer to obtain more effective features, which makes the detection rate of nonobvious defect targets such as line marks higher. The experiment found that using MobileNetV2 [20] as the backbone feature extraction network to detect the data of rare-earth magnetic patches can make the network have a high degree of lightweight and a fast

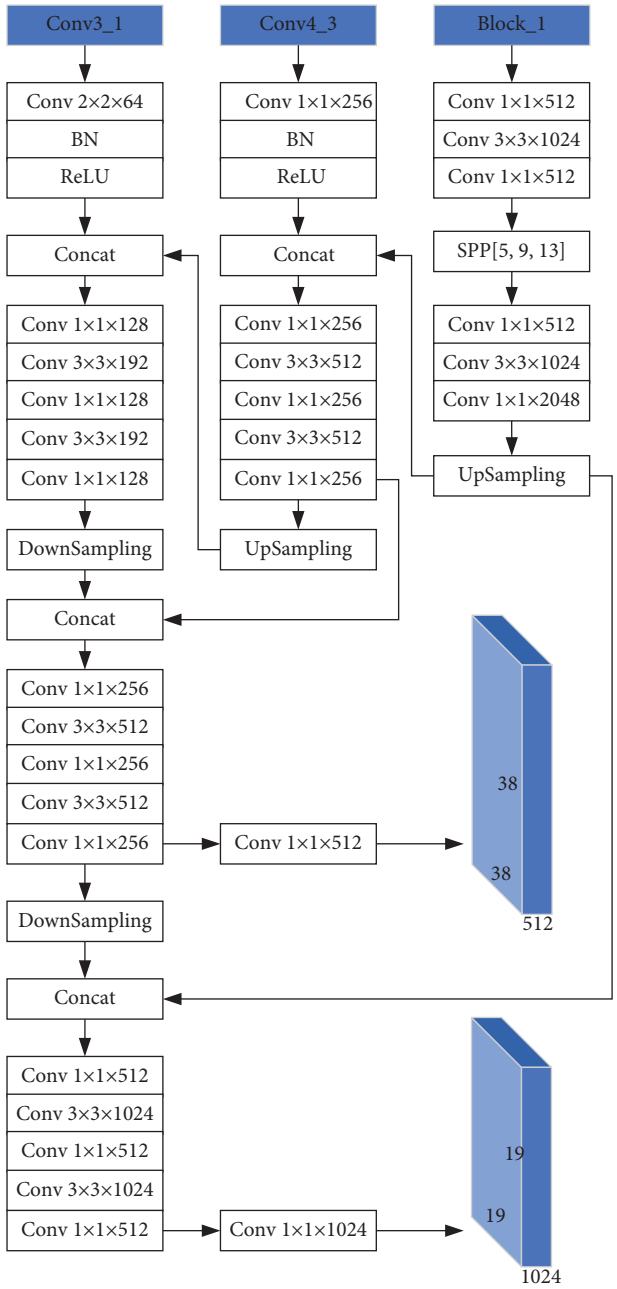


Figure 8: Structure of feature fusion network based on bidirectional FPN.

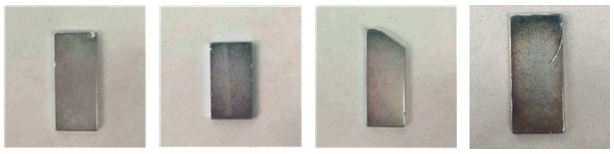


Figure 9: Common defect appearance of rare-earth magnetic patch.

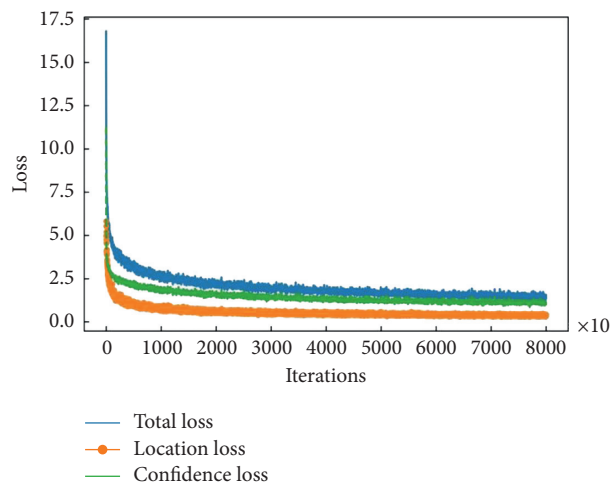


Figure 10: Loss function curve.

Table 1: Comparison of performance indexes between the improved SSD algorithm and other algorithms.

Algorithm	mAP/%	Detection accuracy (AP)/%			
		Unfilled corner	Crack	Deformation	Line mark
Faster R-CNN	77.48	69.86	85.11	63.51	91.45
YOLO-V3	81.88	86.58	84.63	66.25	90.06
MobileNetV2-SSD	73.83	62.66	79.74	67.50	85.43
SSD (original)	80.24	80.28	82.42	70.57	87.69
SSD (improvement)	**83.65**	86.37	85.31	70.89	92.02

detection speed, making it easy to deploy the device, but its detection accuracy is not reliable. Compared with the Faster R-CNN two-stage model that only uses the top layer to learn features, the improved algorithm in this paper has greater advantages. At the same time, compared with the more advanced one-stage model YOLO-V3 [24], defect recognition accuracy is slightly lower, but the comprehensive detection effect is better.

We perform ablation study to explore the effects of PANET bidirectional feature pyramid structure, parallel Multiscale Convolutional Layer, and efficient channel attention mechanism on detection accuracy. Here we investigate four models SSD, SSD + multiscale receptive field layer, SSD + multiscale receptive field layer + PANet, and SSD + multiscale receptive field layer + PANet + ECA. It can be seen from Table 2. After the enhancement of the backbone network embedded receptive field module and integration of the bidirectional feature pyramid in PANET for interlayer feature fusion, the accuracy of mAP was significantly improved to 81.37% and 83.24%. When combining the ECA module, we observed our best performance (83.65%).

In order to more intuitively evaluate the effect of SSD algorithm on defect detection before and after improvement, some experimental results are shown in Figures 11 and 12.

As shown in Figure 11, in addition to the improved SSD algorithm in this paper, the SSD algorithm with the lightweight MobileNetV2 convolutional neural network as the backbone feature extraction network and the original SSD algorithm can both locate and identify defects. The improved SSD algorithm in this paper is more accurate in the bounding box (the detection results of line marks and deformation are shown in the figure).

In actual manufacturing, unfilled corner type defects are very common. The size of such defects is different, and small-sized unfilled corner defects often appear. The improved SSD defect detection algorithm shows stronger recognition ability for this type of defect, the algorithm has a lower rate of missed detection, and the recognition rate is significantly improved. At the same time, its recognition ability is also improved for defects with larger size but nonobvious features, as shown in Figure 12.

This paper compares the detection speed of SSD before and after improvement, which is 108 frames per second and 55 frames per second, respectively, on GeForceRTX2080TI high-performance graphics card, as shown in Table 3. Because the original SSD algorithm integrates the bidirectional feature pyramid structure and channel attention mechanism and enhances the trunk extraction network, the number of parameters is more than the original algorithm. Compared with the original algorithm, the proposed algorithm integrates the bidirectional feature pyramid in PANET for interlayer feature fusion. The algorithm has a large amount of computation, which improves the overall mAP and increases the inference time of the model. In the application of industrial video analysis and processing, the speed of reading image data from the camera is about 25 frames per second. The improved SSD target detection algorithm in this paper meets the requirements of real-time online detection while maintaining good detection performance.

Table 2: Ablation study: mAPs of different modules.

Structure	mAP/%
SSD	80.24
SSD + multiscale receptive field layer	81.37
SSD + multiscale receptive field layer + PANet	83.24
SSD + multiscale receptive field layer + PANet + ECA	83.65

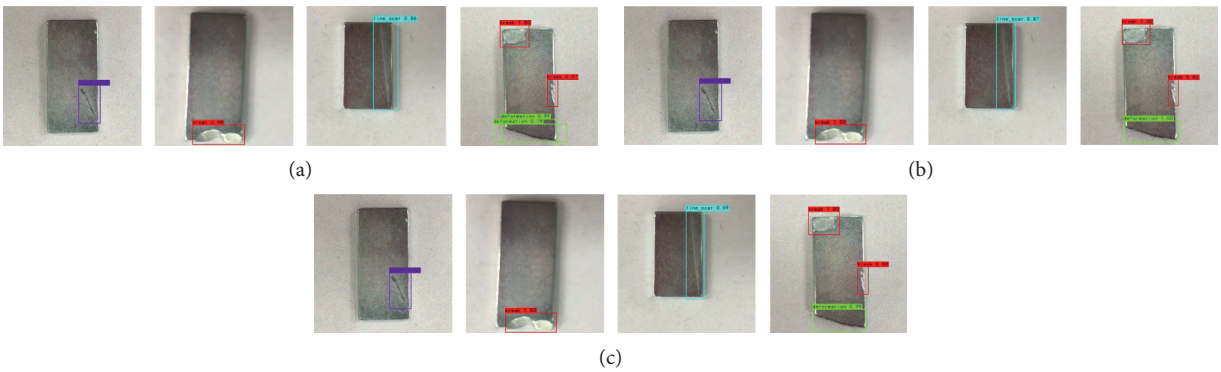


Figure 11: The detection effect of SSD algorithm on large-size defects before and after improvement. (a) Detection effect of MobileNetV2-SSD. (b) Detection effect of original SSD. (c) Detection effect of improved SSD.

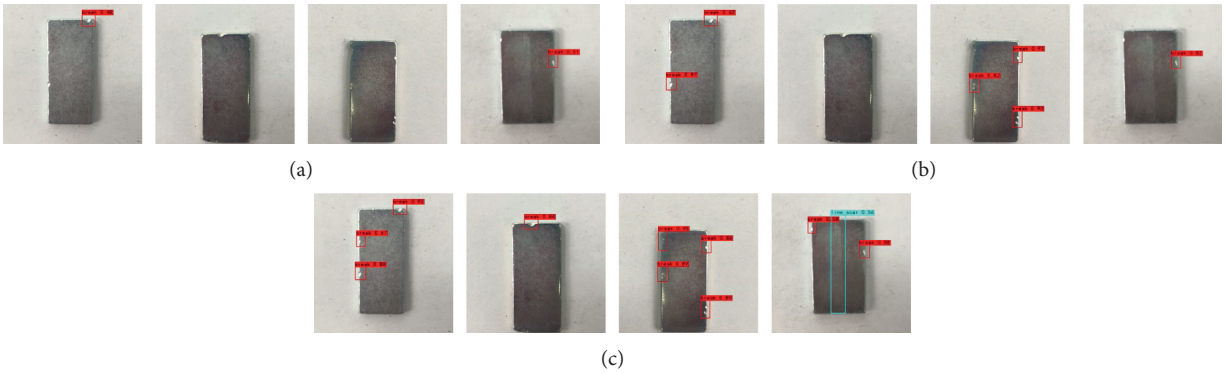


Figure 12: The detection effect of SSD algorithm on unfilled corners and line marks before and after improvement. (a) The detection effect of MobileNetV2-SSD. (b) The detection effect of original SSD. (c) The detection effect of improved SSD.

Table 3: Parameter comparison before and after improvement of SSD algorithm.

Algorithm	Parameter (M)	Computation (GMac)	mAP (%)	Speed (FPS)
SSD (original)	24.15	30.64	80.24	108
SSD (improvement)	71.09	57.63	83.65	55

5. Conclusion

This paper presented a defect detection method of rare-earth magnetic patch based on improved SSD. The SSD algorithm with a relatively balanced detection accuracy and inference speed was selected to detect magnetic patch defects and analyze the detection results. It was observed that SSD had a good identification effect for obvious defects with large size, but for defects with obscure features and small size, there was often a situation of missing inspection. In order to further improve the detection accuracy of SSD algorithm, this paper embedded multiscale receptive field module into the backbone network of SSD algorithm to improve the feature extraction ability of the model, using a bidirectional feature pyramid idea to integrate high-level features and low-level features, combined with efficient channel attention mechanism to enhance the detection ability of the network. Experiments have proved that, compared with the original algorithm, the improved algorithm in this paper has a significant improvement in the

recognition ability of small-size defects, with mAP reaching 83.65%. And the recognition ability of large-size defects with nonobvious features has also been enhanced. The detection speed can reach 55FPS on the experimental platform, which meets the requirements of online automatic detection.

Data Availability

The image data sets used in this paper were all collected from Guangxi Jinyuan Rare Earth Co., Ltd. (Guangxi Nonferrous Metals Group). The data sets were manually classified with the assistance of sorting workers and photographed on site with industrial cameras.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant no. 61762028 and Guangxi Automatic Testing Technology and Instrument Key Laboratory Foundation under grant no. PF19004P.

References

[1] N. Neogi, D. K. Mohanta, and P. K. Dutta, "Review of vision-based steel surface inspection systems," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 50, 2014.

[2] X. Tang and Y. Wang, "Visual inspection and classification algorithm of rail surface defect," *Computer Engineering*, vol. 39, no. 3, pp. 25–30, 2013.

[3] X. Fang and Z. Shi, "Wafer defect detection and classification algorithms based on convolutional neural network," *Computer Engineering*, vol. 44, no. 8, pp. 218–223, 2018.

[4] H. Ponce, C. Cevallos, R. Espinosa, and S. Gutiérrez, "Estimation of low nutrients in tomato crops throughthe analysis of leaf images using machine learning," *Journal of Artificial Intelligence and Technology*, vol. 1, pp. 131–137, 2021.

[5] D. Jiang, G. Hu, G. Qi, and N. Mazur, "A fully convolutional neural network-based regression approach for effective chemical composition analysis using near-infrared spectroscopy in cloud," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 1, pp. 74–82, 2021.

[6] C. Chang, C. Chang, C. Li, and M. Jeng, "Learning vector quantization neural networks for LED wafer defect inspection," in *Proceedings of the Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, Kumamoto, Japan, September 2007.

[7] M. Yazdchi, M. Yazdi, and A. G. Mahyari, "Steel surface defect detection using texture segmentation based on multifractal dimension," in *Proceedings of the 2009 International Conference on Digital Image Processing*, Bangkok, Thailand, March 2009.

[8] K. Yildiz, A. Buldu, and M. Demetgul, "A thermal-based defect classification method in textile fabrics with K-nearest neighbor algorithm," *Journal of Industrial Textiles*, vol. 45, no. 5, pp. 780–795, 2016.

[9] Y.-J. Jeon, D.-C. Choi, S. J. Lee, J. P. Yun, and S. W. Kim, "Defect detection for corner cracks in steel billets using a wavelet reconstruction method," *Journal of the Optical Society of America A*, vol. 31, no. 2, pp. 227–237, 2014.

[10] P. Sun, Y. Zhao, and S. Zhu, "An approach to improve SSD through mask prediction of multi-scale feature maps," *Pattern Analysis and Applications*, vol. 24, no. 9, 2021, republish.

[11] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *The Visual Computer*, vol. 36, no. 1, pp. 85–96, 2018.

[12] C. Liu, J. Zhang, and J. Lin, "Detection and identification of surface defects of magnetic tile based on neural network," *Surface Technology*, vol. 48, no. 8, pp. 330–339, 2019.

[13] Z. Wen and M. Zhou, "Recognition of blowholes and cracks on surface of magnetic tile based on deep learning," *Ordnance Material Science and Engineering*, vol. 43, no. 6, pp. 106–112, 2020.

[14] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[17] J. Redmon, S. Divvala, R. Girshick, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," 2016, https://arxiv.org/abs/1506.02640.

[18] T. Y. Lin, D. Piotr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2017, https://arxiv.org/abs/1612.03144.

[19] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," 2018, https://arxiv.org/abs/1803.01534.

[20] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a surface defect detection algorithm based on mobileNet-SSD," *Applied Sciences*, vol. 8, no. 9, p. 1678, 2018.

[21] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," *Computer Vision—ECCV 2018*, vol. 11215, pp. 404–419, 2018.

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: efficient channel attention for deep convolutional neural networks," 2020, https://arxiv.org/abs/1910.03151.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, https://arxiv.org/abs/1709.01507.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, https://arxiv.org/abs/1512.03385.

Hindawi
Complexity
Volume 2021, Article ID 3237342, 19 pages
https://doi.org/10.1155/2021/3237342

Research Article

A Job-Shop Scheduling Problem with Bidirectional Circular Precedence Constraints

Pisut Pongchairerks

Industrial Engineering Program, Faculty of Engineering, Thai-Nichi Institute of Technology, Bangkok 10250, Thailand

Correspondence should be addressed to Pisut Pongchairerks; pisut@tni.ac.th

Received 17 June 2021; Revised 25 July 2021; Accepted 14 August 2021; Published 9 November 2021

Academic Editor: Jenq-Haur Wang

This paper introduces a job-shop scheduling problem (JSP) with bidirectional circular precedence constraints, called BCJSP. In the problem, each job can be started from any operation and continued by its remaining operations in a circular precedence-relation chain via either a clockwise or counterclockwise direction. To solve BCJSP, this paper proposes a multilevel metaheuristic consisting of top-, middle-, and bottom-level algorithms. The top- and middle-level algorithms are population-based metaheuristics, while the bottom-level algorithm is a local search algorithm. The top-level algorithm basically controls a start operation and an operation-precedence-relation direction of each job, so that BCJSP becomes a JSP instance that is a subproblem of BCJSP. Moreover, the top-level algorithm can also be used to control input parameters of the middle-level algorithm, as an optional extra function. The middle-level algorithm controls input parameters of the bottom-level algorithm, and the bottom-level algorithm then solves the BCJSP's subproblem. The middle-level algorithm evolves the bottom-level algorithm's parameter values by using feedback from the bottom-level algorithm. Likewise, the top-level algorithm evolves the start operations, the operation-precedence-relation directions, and the middle-level algorithm's parameter values by using feedback from the middle-level algorithm. Performance of two variants of the multilevel metaheuristic (i.e., with and without the mentioned extra function) was evaluated on BCJSP instances modified from well-known JSP instances. The variant with the extra function performs significantly better in number than the other. The existing JSP-solving algorithms can also solve BCJSP; however, their results on BCJSP are clearly worse than those of the two variants of the multilevel metaheuristic.

1. Introduction

The job-shop scheduling problem (JSP) [1, 2] and the open-shop scheduling problem (OSP) [3, 4] are well-known in practical applications. They are also interesting academic topics since they are NP-hard problems [5]. They both involve scheduling jobs onto machines in order to minimize makespan, i.e., the schedule's length. Each job consists of a number of operations; each operation must be processed on a predetermined machine with a given processing time. Each machine cannot process more than one operation at a time, and it cannot be stopped during processing an operation. To complete each job in JSP, all of its operations must be processed in the sequence from the first to the last operations. This sequence is called an operation-precedence-relation chain. The operation-precedence-relation chain of each job is very strict and, thus, cannot be changed. In contrast to JSP, OSP has no operation-precedence-relation chains. This means all operations of each job in OSP can be processed in any orders. The job-shop scheduling problem with bidirectional circular precedence constraints (BCJSP) introduced in this paper is an intermediate problem between JSP and OSP. It is a generalized JSP where the operation-precedence-relation chain of each job is circular and bidirectional.

BCJSP has a wide range of real applications, e.g., health check-up service, automobile repair shop, and instrument calibration service. In fact, when taking layouts and distances into account, many OSP's applications become BCJSP's applications. The health check-up service, as a

BCJSP application, starts with multiple optional check-up programs offered to hospital customers. An optional check-up program consists of specific diagnoses, each of which is provided in a different room. For customer satisfaction, the best circular route of all diagnosis rooms has been predetermined by a hospital service manager for each check-up program. The manager may assign each customer to receive any diagnosis of his chosen program as the first diagnosis. However, the manager has to assign the customer to receive his next diagnosis in a nearest predetermined room, and so on. The customer finishes his check-up activities after successfully receiving all diagnoses of his chosen program. Notice that if the diagnosis rooms are very close to each other, the manager can then assign the customer to receive all diagnoses in any orders. Then, this application fits with OSP rather than BCJSP.

There are a number of JSP and OSP's variants recently presented in the literature, e.g., [6–11]. Some of them have some partially similar properties to the BCJSP's properties. For example, the extended resource-constrained project scheduling problem [12] allows processing its activities bidirectionally, but not circularly. The flexible job-shop scheduling problem (FJSP) has more flexibility than JSP and is also defined as a generalized form of JSP. However, the flexibility of FJSP [13, 14] is due to a number of selectable machines for each operation, while the flexibility of BCJSP is due to the bidirectional circular precedence relations of operations.

To solve BCJSP, this paper introduces a multilevel metaheuristic (called MUL) based on the adaptive parameter control concept [15]. MUL consists of the top-level algorithm (called TOP), the middle-level algorithm (called MID), and the bottom-level algorithm (called BOT). In the MUL's top level, TOP controls the start operation and the operation-precedence-relation direction of every job in BCJSP. TOP is also usable to control the MID's input parameters if requested. In the middle level, MID transfers the start operation and the operation-precedence-relation direction of every job given by TOP into BOT. However, the MID's main function is to control the BOT's input parameters. In the bottom level, BOT uses the start operations and the operation-precedence-relation directions to generate a JSP instance, which is a subproblem of the BCJSP instance. BOT then acts as a local search algorithm for solving the generated JSP instance. MID evolves the BOT's input-parameter values based on feedback from BOT, while TOP evolves the operation-precedence-relation chains and the MID's input-parameter values based on feedback from MID.

The BOT combined with MID is similar to the two-level metaheuristic developed by [16]. A major difference of the MID-BOT combination from the algorithm of [16] is in the solution-decoding procedure. Once the MID-BOT combination is combined with TOP, they all together have become MUL. MUL can be defined as an adaptive multistart iterated local search algorithm for solving BCJSP. There are two variants of MUL proposed in this paper, i.e., the base-specification MUL (called MUL-B) and the top-specification MUL (called MUL-T). The only difference between the two variants is in their TOP-MID relationships. In both MUL-B and MUL-T, their TOPs control the start operation and the operation-precedence-relation direction of every job. In only MUL-T, its TOP also controls the MID's input parameters. Performance of MUL-B and MUL-T was evaluated on the BCJSP instances, modified from the JSP instances of [17, 18, 19]. On the BCJSP instances, MUL-B's and MUL-T's results were compared with each other. Because the existing JSP-solving algorithms can be used to solve BCJSP, their results were also used in the performance comparisons. Note that the existing JSP-solving algorithms mean the algorithms developed for solving JSP in the literature, e.g., [1, 2, 16, 20, 21].

The remainder of this paper is divided into six sections. Section 2 provides an overview of the relevant publications of the research topic. Section 3 describes the job-shop scheduling problem with bidirectional circular precedence constraints (BCJSP). Section 4 presents the procedure of MUL, where the procedures of BOT, MID, and TOP are described in Sections 4.1–4.3, respectively. The differences between MUL-B and MUL-T, as the two variants of MUL, are also described in Section 4. Section 5 presents the experiment's results for MUL-B's and MUL-T's performance evaluations. Section 6 then discusses the experiment's results. Finally, Section 7 concludes the research's findings.

2. Related Works

Metaheuristics can be classified into two categories based on their numbers of solutions used in each iteration, i.e., single-point-based and population-based search algorithms [22]. As its name implies, a single-point-based search algorithm starts with a single solution. Then, it moves from its current solution to another solution repeatedly. Local search is a well-known type of single-point-based search algorithms. A local search algorithm improves its solution gradually within a local region of the solution space. Although a local search algorithm aims to find just a local optimal solution, the algorithm with a good initial solution occasionally finds a global optimal solution.

Iterated local search [23] is another well-known type of single-point-based search algorithms. It can be defined as a local search algorithm that can escape a local region of the solution space. During its exploration, an iterated local search algorithm uses a neighbor operator repeatedly to find a local optimal solution. After that, it tries to escape the current local region into another local region by using a perturbation operator. In general, an iterated local search algorithm starts with a single solution; however, some recent variants, e.g., [24, 25], have multistart properties.

There are three operators, i.e., swap, insert, and inverse, commonly used as a neighbor operator and a perturbation operator [26]. To define these three operators, let h and v be two different random integers from 1 to the number of members in a solution-representing permutation. The swap operator is to swap between the two members in the h-th and the$v$-th positions. The insert operator is to remove a member from the h-th position and then insert it back at the v-th position. The inverse operator is to inverse the sequence of

all members from the h-th to the$v$-th positions. Some iterated local search algorithms, e.g., [27, 28], use the swap operator or insert operator multiple times as their perturbation operators.

As mentioned, the population-based search algorithm category is the alternative of the single-point-based search algorithm category. A population-based search algorithm starts with a set (i.e., a population) of solutions instead of a single solution. At each iteration, a population-based search algorithm evolves its population by using the information from its previous iterations. Some population-based search algorithms were purposely developed for solving discrete optimization problems, such as genetic algorithm [29] and ant colony optimization [30]. In contrast, some others were intentionally developed for exploring in real-number search spaces, such as particle swarm optimization [31], differential evolution [32], and cuckoo search [33] algorithms.

A common drawback of most metaheuristics is that there is no single set of input-parameter values performing best for all problem's instances. However, several techniques can be applied for handling such a drawback. One of these techniques is adaptive parameter control [15], where an upper-level metaheuristic acts as a parameter controller for a lower-level metaheuristic. (Note that upper-level metaheuristic is commonly called *metaevolutionary algorithm* [15, 34].) In addition, an upper-level metaheuristic can be applied to control parameters of a being-considered problem for generating its simpler subproblems [35, 36]. For applying the adaptive parameter control technique, the metaheuristics usually have only two levels [4, 15, 16, 37–39]. However, for solving highly complicated problems, they may require more than two levels [35, 36].

A two-level metaheuristic of [37] was developed for solving JSP. It consists of the algorithms named UPLA and LOLA in its upper and lower levels, respectively. LOLA is a local search algorithm exploring in a solution space of parameterized-active schedules (hybrid schedules) [40–42], where each parameterized-active schedule is decoded from an operation-based permutation [29, 43]. UPLA is the population-based search algorithm intentionally developed for being a parameter controller. Its population consists of a number of value combinations of input parameters of LOLA. For updating an input-parameter-value combination, each input-parameter value is iteratively moved by a sum of two changeable opposite-direction vectors. The first vector's direction is toward the memorized best-found value, whereas the second vector's direction is away from it. The magnitudes of these two vectors are generated randomly between zeros and their given maximum values.

The two-level metaheuristic of [16] is a recent variant of [37]. In this variant, MUPLA and LOSAP are the upper-level and lower-level metaheuristics, respectively. LOSAP [16] is a local search algorithm exploring in a probabilistic-based hybrid neighborhood structure. To generate each neighbor solution, LOSAP randomly uses one of the two predetermined neighbor operators by a preassigned probability. (Other applications of randomly using one of two different operators can be found in [44, 45].) Note that while LOLA's solution space is a set of parameterized-active schedules, LOSAP's solution space is just a set of semiactive schedules. It means that the LOSAP's search ability is mainly based on its hybrid neighborhood structure, not based on a special solution space like LOLA. LOSAP has many optional operators proposed for being its perturbation and neighbor operators. In addition, LOSAP uses a different criterion from that of LOLA on accepting a new best-found solution.

MUPLA [16] is a population-based metaheuristic designed to be a parameter controller for LOSAP. Thus, its population consists of a number of value combinations of the LOSAP's input parameters. Each input-parameter-value combination contains specific values of the perturbation operator, the scheduling direction, the ordered pair of two neighbor operators, the probability of selecting a neighbor operator, and the start solution-representing permutation. A major change of MUPLA from UPLA is that each input-parameter-value combination in its population includes a specific start solution-representing permutation. Thus, the MUPLA combined with LOSAP acts as a multistart iterated local search algorithm, while the UPLA combined with LOLA is just an iterated local search algorithm.

3. Problem Definition

BCJSP is an intermediate problem between JSP and OSP; however, it can be explained simpler as a JSP's generalized variant. BCJSP aims to find a feasible schedule that minimizes makespan (i.e., a total length of the schedule). The problem comes with m given machines (i.e., M_1, M_2, . . ., M_m) and n given jobs (i.e., J_1, J_2, . . ., J_n). At the beginning (i.e., time 0), all jobs have already been arrived, and all machines have not yet been occupied. Each job J_i (where $i = 1, 2, \ldots, n$) consists of m operations (i.e., O_{i1}, O_{i2}, . . ., O_{im}). Each operation must be processed by a predetermined machine with a predetermined processing time. Each machine cannot process more than one operation at a time, and it cannot be stopped or paused during processing an operation. In other words, an operation preemption is not allowed. BCJSP differs from JSP in that an operation-precedence-relation chain of each job J_i (where $i = 1, 2, \ldots, n$) is circular and bidirectional, as shown in Figure 1.

Figure 1 shows the relationships of the operations O_{i1}, O_{i2}, . . ., O_{im} of the job J_i in BCJSP. For each job J_i, let the precedence relations of O_{i1}, O_{i2}, . . ., O_{im} be all together connected as a circular chain. It means that any operation from O_{i1} to O_{im} can be selected as the start operation of the job J_i (i.e., the operation processed first in the job J_i). Then, to complete the job J_i, the remaining operations in the circular chain must be processed one-by-one in either a clockwise or counterclockwise direction. In Figure 1, the operations connected together by green arrows present the circular operation-precedence-relation chain in clockwise direction, while those by blue arrows present the chain in counterclockwise direction.

A BCJSP instance can be divided into $(2m)^n$ JSP instances as all of its subproblems. Each subproblem is generated from the BCJSP instance by assigning a specific start operation and a specific operation-precedence-relation direction into every job. To generate a subproblem, let O_{ik} be

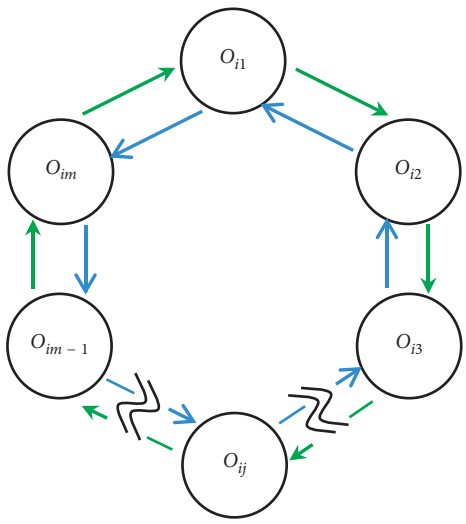


Figure 1: A diagram of the operation-precedence-relation chain of the job J_i in BCJSP.

the start operation of the job J_i, which is selected from any operation of O_{i1}, O_{i2}, ..., O_{im}. Then, let the operation-precedence-relation direction of the job J_i be selected from either clockwise or counterclockwise. If the clockwise direction is selected, let the operation-precedence-relation chain of the job J_i be $O_{ik} \Rightarrow O_{ik+1} \Rightarrow \ldots \Rightarrow O_{im} \Rightarrow O_{i1} \Rightarrow \ldots \Rightarrow O_{ik-1}$; otherwise, let it be $O_{ik} \Rightarrow O_{ik-1} \Rightarrow \ldots \Rightarrow O_{i1} \Rightarrow O_{im} \Rightarrow \ldots \Rightarrow O_{ik+1}$. In this paper, $D \Rightarrow E$ means that D must be finished before E can be started.

To clarify the above paragraph, consider a BCJSP instance that has three machines (i.e., M_1, M_2, M_3) and two jobs (i.e., J_1 and J_2). Each job J_i (where $i = 1$ and 2) then consists of three operations (i.e., O_{i1}, O_{i2}, and O_{i3}). To generate a subproblem, there are six options for the job J_1's operation-precedence-relation chain: (1) $O_{11} \Rightarrow O_{12} \Rightarrow O_{13}$, (2) $O_{12} \Rightarrow O_{13} \Rightarrow O_{11}$, (3) $O_{13} \Rightarrow O_{11} \Rightarrow O_{12}$, (4) $O_{13} \Rightarrow O_{12} \Rightarrow O_{11}$, (5) $O_{12} \Rightarrow O_{11} \Rightarrow O_{13}$, and (6) $O_{11} \Rightarrow O_{13} \Rightarrow O_{12}$. In addition, there are six options for the job J_2's operation-precedence-relation chain: (1) $O_{21} \Rightarrow O_{22} \Rightarrow O_{23}$, (2) $O_{22} \Rightarrow O_{23} \Rightarrow O_{21}$, (3) $O_{23} \Rightarrow O_{21} \Rightarrow O_{22}$, (4) $O_{23} \Rightarrow O_{22} \Rightarrow O_{21}$, (5) $O_{22} \Rightarrow O_{21} \Rightarrow O_{23}$, and (6) $O_{21} \Rightarrow O_{23} \Rightarrow O_{22}$. Of each job, the first three options are generated in clockwise direction, while the last three options are generated in counterclockwise direction. Based on the six options of each job, this BCJSP instance can be divided into 36 JSP instances as all of its subproblems.

BCJSP is a generalization of JSP and is also much more complex than JSP. Every single BCJSP instance can be divided into $(2m)^n$ JSP instances as all of its subproblems. Because JSP with $m = n = 3$ has been proven to be NP-hard [5], BCJSP with $m \geq 3$ and $n \geq 3$ thus belongs to a class of NP-hard problems. In the literature, no algorithms excepting MUL-B and MUL-T have been developed for BCJSP. Although the existing JSP-solving algorithms without modifications can be used to solve BCJSP, they may not perform well on BCJSP. This is because, with the same m and n, a solution space of BCJSP is much larger than that of JSP.

4. Methods

As mentioned, MUL represents the proposed multilevel metaheuristic for solving BCJSP. It consists of BOT, MID, and TOP algorithms in its bottom, middle, and top levels, respectively. BOT is a local search algorithm, modified from LOSAP [16], for solving subproblems of the BCJSP instance. Each subproblem is a JSP instance modified from the BCJSP instance by assigning a specific start operation and a specific operation-precedence-relation direction into every job. MID, as a variant of MUPLA [16], is a population-based metaheuristic for controlling BOT's input parameters. Another function of MID is to transfer the start operation and the operation-precedence-relation direction of every job from TOP into BOT. TOP is a population-based metaheuristic developed based on the framework of MUPLA [16]. TOP is used to control the start operation and the operation-precedence-relation direction of every job in the BCJSP instance. If requested, TOP can also control the MID's input parameters as an extra optional function.

In this paper, there are two variants of MUL, i.e., the base-specification MUL (MUL-B) and the top-specification MUL (MUL-T). MUL-B is the MUL whose TOP controls only the start operation and the operation-precedence-relation direction of every job in BCJSP. MUL-T is the MUL whose TOP controls the start operation and operation-precedence-relation direction of every job and also the MID's input parameters. The details of BOT, MID, and TOP are described in Sections 4.1–4.3, respectively.

4.1. BOT Algorithm. BOT is a local search algorithm for solving subproblems of the being-solved BCJSP instance; each subproblem is a JSP instance. BOT generates each subproblem from the BCJSP instance by assigning a specific start operation and a specific operation-precedence-relation direction into every job. Let A_i and B_i represent the start operation and the operation-precedence-relation direction, respectively, of the job J_i (where $i = 1, 2, \ldots, n$). For each job J_i, A_i can be any operation selected from O_{i1}, O_{i2}, ..., O_{im}; in addition, B_i can be either a clockwise or counterclockwise direction. In this paper, BOT receives A_i and B_i from TOP via MID.

To illustrate how to use A_i and B_i, assume $A_1 = O_{12}$, $A_2 = O_{23}$, $A_3 = O_{31}$, $B_1 =$ counterclockwise, $B_2 =$ clockwise, and $B_3 =$ counterclockwise be assigned into a 2-machine/3-job BCJSP instance. By assigning $A_1 = O_{12}$ and $B_1 =$ counterclockwise, the job J_1's operation-precedence-relation chain becomes $O_{12} \Rightarrow O_{11} \Rightarrow O_{13}$. By assigning $A_2 = O_{23}$ and $B_2 =$ clockwise, the job J_2's operation-precedence-relation chain becomes $O_{23} \Rightarrow O_{21} \Rightarrow O_{22}$. By assigning $A_3 = O_{31}$ and $B_3 =$ counterclockwise, the job J_3's operation-precedence-relation chain becomes $O_{31} \Rightarrow O_{33} \Rightarrow O_{32}$. As a result, a subproblem of the BCJSP instance in the form of JSP has successfully been generated.

After BOT has successfully generated a JSP instance (which is a subproblem of the BCJSP specified by A_i and B_i), BOT acts as a local search algorithm for solving the JSP instance. BOT uses operation-based permutations [29, 43] to represent semiactive schedules, where an operation-based permutation is a permutation with m repetitions of the numbers 1, 2, . . ., n. However, the transformation into a schedule for the BCJSP's subproblem differs from the transformation used by [29, 43] for the classical JSP. The difference is due to the specific order of operations in the operation-precedence-relation chain assigned by A_i and B_i. For example, on a 2-machine/3-job BCJSP instance, assume a given permutation be (2, 1, 2, 3, 1, 1, 3, 3, 2). In addition, assume $A_1 = O_{12}$, $A_2 = O_{23}$, $A_3 = O_{31}$, $B_1 =$ counterclockwise, $B_2 =$ clockwise, and $B_3 =$ counterclockwise. From the given permutation, BOT constructs a semiactive schedule in the order of O_{23}, O_{12}, O_{21}, O_{31}, O_{11}, O_{13}, O_{33}, O_{32}, and O_{22}. Notice that the schedule is not constructed in the order of O_{21}, O_{11}, O_{22}, O_{31}, O_{12}, O_{13}, O_{32}, O_{33}, and O_{23} like that used for the classical JSP.

BOT, as modified from LOSAP [16], improves its solutions by using PT and PN. Let PT represent the perturbation operator, and let $PN \equiv (PN_f, PN_s)$ represent the ordered pair of the first neighbor operator (PN_f) and the second neighbor operator (PN_s). BOT offers five options for PT, i.e., n-medium swap, n-large swap, n-medium inverse, n-large insert, and n-medium insert. In addition, BOT offers four options for $PN \equiv (PN_f, PN_s)$, i.e., (1-small inverse, 1-medium insert), (1-large swap, 1-large insert), (1-medium swap, 1-medium insert), and (1-small swap, 1-small insert).

In the names of the above-mentioned operators, the number in front of the hyphen sign indicates the number of repeated uses of the operator mentioned in back of the hyphen sign. For example, the n-medium inverse operator is to use the medium inverse operator n times on a permutation. In addition, the words *small*, *medium*, and *large* are used to restrict the value of v from h (note that the uses of h and v for operators are already reviewed in Section 2). Let h and v be two different random integers within $[1, mn]$, where mn is the number of all operations in the BCJSP instance. The words *small* and *medium* then provide additional limitations on generating v as follows: v must be generated within $[h-4, h+4]$ for *small*, while v must be generated within $[h-0.2mn, h+0.2mn]$ for *medium*. For *large*, there are no additional limitations.

The procedure of BOT is presented in Algorithm 1, and its flowchart is presented in Figure 2. The parameters and abbreviations used in Algorithm 1 and Figure 2 are defined as follows:

(i) Let m and n, respectively, be the number of all machines and the number of all jobs in BCJSP. Thus, mn is the number of all operations in BCJSP.

(ii) Let $A_i \in \{O_{i1}, O_{i2}, \ldots, O_{im}\}$ represent the start operation of the job J_i.

(iii) Let $B_i \in \{$clockwise, counterclockwise$\}$ represent the operation-precedence-relation direction of the job J_i.

(iv) Let PT and P stand for the perturbation operator and the start operation-based permutation, respectively.

(v) Let $PN \equiv (PN_f, PN_s)$ represent the ordered pair of the first neighbor operator (PN_f) and the second neighbor operator (PN_s).

(vi) Let PR be the probability of selecting the first neighbor operator (PN_f) of PN. Consequently, the probability of selecting the second neighbor operator (PN_s) is equal to unity minus PR.

(vii) Let P_0 (which is a permutation with m repetitions of the numbers 1, 2, . . ., n) stand for the current best-found operation-based permutation.

(viii) Let Π_0 be the permutation of mn operations decoded from P_0.

(ix) Let S_0 stand for the current best-found schedule decoded from Π_0. In addition, let $Makespan(S_0)$ represent the makespan of S_0.

(x) Let P_1 (which is a permutation with m repetitions of the numbers 1, 2, . . ., n) stand for the current neighbor operation-based permutation.

(xi) Let Π_1 be the permutation of mn operations decoded from P_1.

(xii) Let S_1 stand for the current neighbor schedule decoded from Π_1. In addition, let $Makespan(S_1)$ represent the makespan of S_1.

Although there are two proposed variants of MUL (i.e., MUL-B and MUL-T), the procedures of BOTs in MUL-B and MUL-T are both identical to Algorithm 1. The differences between MUL-B and MUL-T are in their MIDs and TOPs.

4.2. MID Algorithm. MID is a population-based metaheuristic modified from MUPLA [16]. It is a channel to transfer A_i and B_i (where $i = 1, 2, \ldots, n$) from TOP into BOT. However, a main function of MID is to be a parameter controller for BOT. At the t-th iteration, the MID's population contains N members, i.e., $C_1(t), C_2(t), \ldots, C_N(t)$. For $g = 1$ to N, let $C_g(t) \equiv (pt_g(t), pn_g(t), pr_g(t), p_g(t))$ represent a value combination of the BOT's input parameters PT, PN, PR, and P, respectively. Each of $pt_g(t+1)$, $pn_g(t+1)$, and $pr_g(t+1)$ is updated from its old value via two opposite-direction vectors. The first vector's direction is toward the memorized best-found value, whereas the second vector's direction is away from it. Differently, $p_g(t+1)$ is set to the final operation-based permutation returned from the BOT with $C_g(t)$-given input-parameter values.

The procedure of MID is presented in Algorithm 2, and its flowchart is presented in Figure 3. The following list presents the definitions of parameters and abbreviations used in Algorithm 2 and Figure 3. In addition, the transformation (i.e., decoding method) of each member of $C_g(t)$ is also given:

(i) Let n be the number of all jobs in BCJSP.

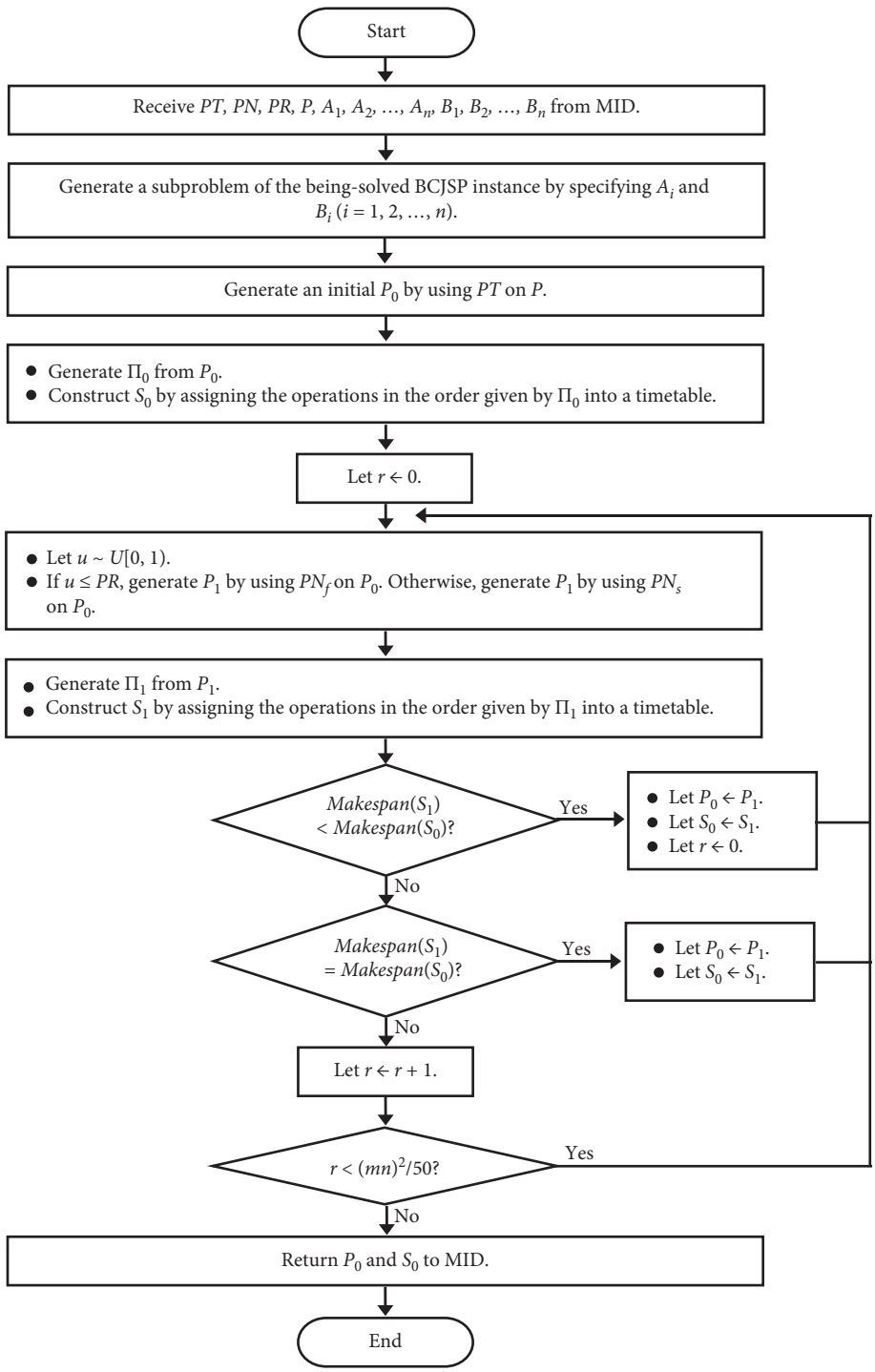


Figure 2: A flowchart of BOT.

(ii) Let N be the number of all members in the MID's population.

(iii) Let $C_g(t) \equiv (pt_g(t), pn_g(t), pr_g(t), p_g(t))$ represent the g-th member (where $g = 1, 2, \ldots, N$) in the MID's population at the t-th iteration. In addition, let $Score(C_g(t))$ stand for the performance score of $C_g(t)$. Note that the lower the performance score, the better the performance.

(iv) Let $pt_g(t) \in \mathbb{R}$ represent the perturbation operator (PT) of BOT. Equation (1) is used to transform $pt_g(t)$ into PT.

(v) Let $pn_g(t) \in \mathbb{R}$ represent the ordered pair of the first and second neighbor operators (PN) of BOT. Equation (2) is used to transform $pn_g(t)$ into PN.

(vi) Let $pr_g(t) \in \mathbb{R}$ represent the probability of selecting the first neighbor operator (PR) of BOT.

(1) Receive values of BOT's input parameters (i.e., PT, PN, PR, and P) from MID. In addition, receive A_i and B_i (where $i = 1, 2, \ldots, n$) from MID.
(2) To generate a subproblem of the BCJSP instance, let the start operation and the operation-precedence-relation direction be assigned to every job by using Steps 2.1 to 2.4.
  (2.1) Let $i \leftarrow 1$.
  (2.2) Let $O_{ik} \leftarrow A_i$.
  (2.3) If B_i = clockwise, let the job J_i's operation-precedence-relation chain be $O_{ik} \Rightarrow O_{ik+1} \Rightarrow \ldots \Rightarrow O_{im} \Rightarrow O_{i1} \Rightarrow O_{i2} \Rightarrow \ldots \Rightarrow O_{ik-1}$. Otherwise, let it be $O_{ik} \Rightarrow O_{ik-1} \Rightarrow \ldots \Rightarrow O_{i1} \Rightarrow O_{im} \Rightarrow O_{im-1} \Rightarrow \ldots \Rightarrow O_{ik+1}$.
  (2.4) If $i < n$, let $i \leftarrow i + 1$ and repeat from Step 2.2. Otherwise, go to Step 3.
(3) Generate an initial P_0 by using PT on P. Then, transform P_0 into S_0 by using Steps 3.1 and 3.2.
  (3.1) Generate Π_0 by changing the j-th repetition of the number i in P_0 into the operation listed in the j-th order of the job J_i's operation-precedence-relation chain. (Note that the job J_i's operation-precedence-relation chain is given in Step 2.)
  (3.2) Construct S_0 by assigning the operations in the order given by Π_0 (from left to right) into a timetable. In the timetable, each operation must be assigned to its predetermined machine at its earliest possible start time. (Note that the earliest possible start time of each operation is the maximum between the completion time of its immediate-predecessor operation in its job and the completion time of the current latest operation on its machine.)
(4) Find a local optimal schedule by using Steps 4.1 to 4.3.
  (4.1) Let $r \leftarrow 0$.
  (4.2) Randomly generate $u \sim U[0, 1)$. If $u \leq PR$, then generate P_1 by using PN_f on P_0; otherwise, generate P_1 by using PN_s on P_0. Then, transform P_1 into S_1 by using Steps 4.2.1 and 4.2.2.
    (4.2.1) Generate Π_1 by changing the j-th repetition of the number i in P_1 into the operation listed in the j-th order of the job J_i's operation-precedence-relation chain.
    (4.2.2) Construct S_1 by assigning the operations in the order given by Π_1 (from left to right) into a timetable. In the timetable, each operation must be assigned to its predetermined machine at its earliest possible start time.
  (4.3) Update P_0, S_0, and r by using Steps 4.3.1 to 4.3.3.
    (4.3.1) If $Makespan(S_1) < Makespan(S_0)$, let $P_0 \leftarrow P_1$ and $S_0 \leftarrow S_1$, and repeat from Step 4.1.
    (4.3.2) If $Makespan(S_1) = Makespan(S_0)$, let $P_0 \leftarrow P_1$ and $S_0 \leftarrow S_1$, and repeat from Step 4.2.
    (4.3.3) If $Makespan(S_1) > Makespan(S_0)$, let $r \leftarrow r + 1$. Then, repeat from Step 4.2 if $r < (mn)^2/50$; otherwise, go to Step 5.
(5) Return P_0 and S_0 as the final (best-found) operation-based permutation and the final (best-found) schedule, respectively, to MID.

Algorithm 1: The procedure of BOT.

In the transformation, let $PR \leftarrow pr_g(t)$ if $0 \leq pr_g(t) \leq 1$; in addition, let $PR \leftarrow 0$ if $pr_g(t) < 0$, and let $PR \leftarrow 1$ if $pr_g(t) > 1$.

(vii) Let $p_g(t)$ represent the start operation-based permutation (P) of BOT, and let it be a member of all possible operation-based permutations. In the transformation, let $P \leftarrow p_g(t)$.

(viii) For updating $pt_g(t+1)$, let y_{pt} and w_{pt} be the controlling weights of the maximum magnitudes of the first and second vectors, respectively. However, if $pt_g(t) = pt_{\text{best}}$, let w_{pt} be the controlling weights of the maximum magnitudes of both vectors.

(ix) For updating $pn_g(t+1)$, let y_{pn} and w_{pn} be the controlling weights of the maximum magnitudes of the first and second vectors, respectively. However, if $pn_g(t) = pn_{\text{best}}$, let w_{pn} be the controlling weights of the maximum magnitudes of both vectors.

(x) For updating $pr_g(t+1)$, let y_{pr} and w_{pr} be the controlling weights of the maximum magnitudes of the first and second vectors, respectively. However, if $pr_g(t) = pr_{\text{best}}$, let w_{pr} be the controlling weights of the maximum magnitudes of both vectors.

(xi) Let P_0 and S_0, respectively, stand for the final operation-based permutation and the final schedule returned from BOT. In addition, let $Makespan(S_0)$ stand for the makespan of S_0.

(xii) Let $C_{\text{best}} \equiv (pt_{\text{best}}, pn_{\text{best}}, pr_{\text{best}}, p_{\text{best}})$ and S_{best} stand for the best $C_g(t)$ and the best S_0, respectively, ever found by the population. In addition, let $Score(C_{\text{best}})$ stand for the performance score of C_{best}.

(1) Receive $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n, y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr} from TOP.
(2) Let $t \leftarrow 1$ and $Score(C_{\text{best}}) \leftarrow +\infty$.
(3) Generate $C_g(t)$ by randomly generating $pt_g(t)$, $pn_g(t)$, and $pr_g(t) \sim U[0,1)$ and randomly generating $p_g(t)$ from any possible operation-based permutation ($g = 1, 2, \ldots, N$).
(4) Evaluate $Score(C_g(t))$ and update $p_g(t+1)$, C_{best}, and S_{best} by using Steps 4.1 to 4.6.
(4.1) Let $g \leftarrow 1$.
(4.2) Transform $C_g(t)$ into the values of PT, PN, PR, and P of BOT.
(4.3) Execute BOT with the values of PT, PN, PR, and P (taken from Step 4.2) and the values of $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n$ (taken from Step 1). This is done for receiving P_0 and S_0 from BOT.
(4.4) Let $Score(C_g(t)) \leftarrow Makespan(S_0)$, and let $p_g(t+1) \leftarrow P_0$.
(4.5) If $Score(C_g(t)) \leq Score(C_{\text{best}})$, let $C_{\text{best}} \leftarrow C_g(t)$, $Score(C_{\text{best}}) \leftarrow Score(C_g(t))$, and $S_{\text{best}} \leftarrow S_0$.
(4.6) If $g < N$, let $g \leftarrow g+1$ and repeat from Step 4.2. Otherwise, go to Step 5.
(5) Update $pt_g(t+1)$, $pn_g(t+1)$, and $pr_g(t+1)$, where $g = 1, 2, \ldots, N$, by using Steps 5.1 to 5.3.
(5.1) Let $g \leftarrow 1$.
(5.2) Generate $pt_g(t+1)$, $pn_g(t+1)$, and $pr_g(t+1)$ by below three equations, respectively. Let u_1 and $u_2 \sim U[0,1)$.

$$pt_g(t+1) = \begin{cases} pt_g(t) + (0.02 + 0.01y_{pt})u_1 - (0.005 + 0.01w_{pt})u_2 & \text{if } pt_g(t) < pt_{\text{best}}, \\ pt_g(t) - (0.02 + 0.01y_{pt})u_1 + (0.005 + 0.01w_{pt})u_2 & \text{if } pt_g(t) > pt_{\text{best}}, \\ pt_g(t) + (0.005 + 0.01w_{pt})u_1 - (0.005 + 0.01w_{pt})u_2 & \text{if } pt_g(t) = pt_{\text{best}}. \end{cases}$$

$$pn_g(t+1) = \begin{cases} pn_g(t) + (0.02 + 0.01y_{pn})u_1 - (0.005 + 0.01w_{pn})u_2 & \text{if } pn_g(t) < pn_{\text{best}}, \\ pn_g(t) - (0.02 + 0.01y_{pn})u_1 + (0.005 + 0.01w_{pn})u_2 & \text{if } pn_g(t) > pn_{\text{best}}, \\ pn_g(t) + (0.005 + 0.01w_{pn})u_1 - (0.005 + 0.01w_{pn})u_2 & \text{if } pn_g(t) = pn_{\text{best}}. \end{cases}$$

$$pr_g(t+1) = \begin{cases} pr_g(t) + (0.02 + 0.01y_{pr})u_1 - (0.005 + 0.01w_{pr})u_2 & \text{if } pr_g(t) < pr_{\text{best}}, \\ pr_g(t) - (0.02 + 0.01y_{pr})u_1 + (0.005 + 0.01w_{pr})u_2 & \text{if } pr_g(t) > pr_{\text{best}}, \\ pr_g(t) + (0.005 + 0.01w_{pr})u_1 - (0.005 + 0.01w_{pr})u_2 & \text{if } pr_g(t) = pr_{\text{best}}. \end{cases}$$

(5.3) If $g < N$, let $g \leftarrow g+1$ and repeat from Step 5.2. Otherwise, go to Step 6.
(6) If the stopping criterion is not met, let $t \leftarrow t+1$ and repeat from Step 4. Otherwise, return S_{best} to TOP.

Algorithm 2: The procedure of MID.

$$PT = \begin{cases} n\text{-medium swap if } pt_g(t) < 0.20, \\ n\text{-large swap if } 0.20 \leq pt_g(t) < 0.40, \\ n\text{-medium inverse if } 0.40 \leq pt_g(t) < 0.60, \\ n\text{-large insert if } 0.60 \leq pt_g(t) < 0.80, \\ n\text{-medium insert if } pt_g(t) \geq 0.80, \end{cases} \tag{1}$$

$$PN = \begin{cases} (\text{1-small inverse, 1-medium insert}) & \text{if } pn_g(t) < 0.25, \\ (\text{1-large swap, 1-large insert}) & \text{if } 0.25 \leq pn_g(t) < 0.50, \\ (\text{1-medium swap, 1-medium insert}) & \text{if } 0.50 \leq pn_g(t) < 0.75, \\ (\text{1-small swap, 1-small insert}) & \text{if } pn_g(t) \geq 0.75. \end{cases} \tag{2}$$

In Algorithm 2, MID starts its procedure by receiving A_i, B_i (where $i = 1, 2, \ldots, n$), and its input-parameter values (i.e., $y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr}) from TOP. MID assigns $t \leftarrow 1$ and $Score(C_{\text{best}}) \leftarrow +\infty$; then, it generates $C_g(t)$ randomly. To solve a BCJSP subproblem specified by A_i and B_i, MID then starts a repeated loop by executing BOT N times. In the g-th execution (where $g = 1, 2, \ldots, N$), BOT is executed with $C_g(t)$-given parameter values to return P_0 and S_0; then, let $Score(C_g(t)) \leftarrow Makespan(S_0)$ and $p_g(t+1) \leftarrow P_0$. If MID finds any $C_g(t)$ better than or equal to C_{best}, then let this $C_g(t)$ and its corresponding S_0 become a new C_{best} and a new S_{best}, respectively. After that, MID completes $C_g(t+1)$ by using the two opposite-direction vectors to generate each of $pt_g(t+1)$, $pn_g(t+1)$, and $pr_g(t+1)$. If the stopping criterion is not met, MID assigns $t \leftarrow t+1$ and starts the repeated loop's next round.

As mentioned earlier, there are two variants of MUL, i.e., the base-specification MUL (MUL-B) and the top-specification MUL (MUL-T). The difference between MUL-B and MUL-T in their MIDs is given as follows. In MUL-T, the input parameters (i.e., $y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr}) of its MID are controlled by TOP. This means MID in MUL-T is identical to Algorithm 2. Differently, these input-parameter values of MID in MUL-B are constants. The procedure of MID in MUL-B is thus modified from Algorithm 2 by removing $y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr} from Step 1 and

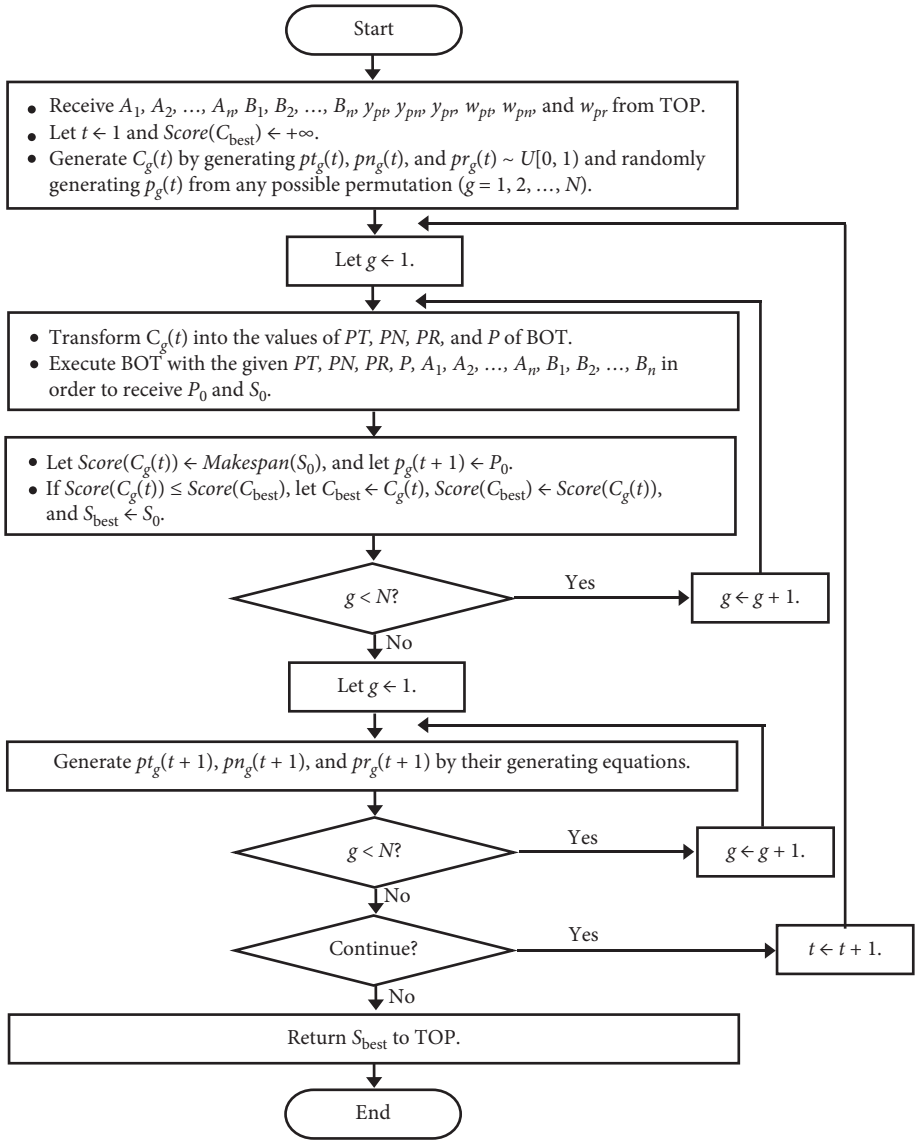


Figure 3: A flowchart of MID.

changing y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr} in Step 5.2 to constants. In this paper, the values of y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr} of MID in MUL-B all are set to 0.5, as mentioned again in Section 5.

4.3. TOP Algorithm. Like MID, TOP is developed based on the framework of MUPLA [16]. The main function of TOP is to control the start operation and the operation-precedence-relation direction of every job in the being-solved BCJSP instance. As previously mentioned, A_i and B_i represent the start operation and the operation-precedence-relation direction, respectively, of the job J_i (where $i = 1, 2, \ldots, n$, and n is the number of all jobs in BCJSP). After assigning A_i and B_i, the BCJSP instance has become a JSP instance that is a subproblem of the BCJSP instance. In addition to the main function, TOP can control the MID's input parameters, i.e., y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr}, as its optional extra function.

At the τ-th iteration, the TOP's population contains M members, i.e., $\zeta_1(\tau), \zeta_2(\tau), \ldots, \zeta_M(\tau)$. For $q = 1$ to M, let $\zeta_q(\tau) \equiv ((\alpha_{1q}(\tau), \alpha_{2q}(\tau), \ldots, \alpha_{nq}(\tau)), \beta_{1q}(\tau), \beta_{2q}(\tau), \ldots, \beta_{nq}(\tau), \gamma_{1q}(\tau), \gamma_{2q}(\tau), \ldots, \gamma_{6q}(\tau))$ represent a value combination of $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n, y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr}, respectively. Each member of $\zeta_q(\tau+1)$, such as $\alpha_{1q}(\tau+1)$, is usually updated from its old value via two opposite-direction vectors; however, it is occasionally regenerated by a reinitialization. For the two opposite-direction vectors, the first vector's direction is toward the memorized best-found value, whereas the second vector's direction is away from it.

The procedure of TOP is presented in Algorithm 3, and its flowchart is presented in Figure 4. The following list shows the definitions of parameters and abbreviations used in Algorithm 3 and Figure 4. In addition, the transformation (i.e., decoding method) of each member of $\zeta_q(\tau)$ is also given:

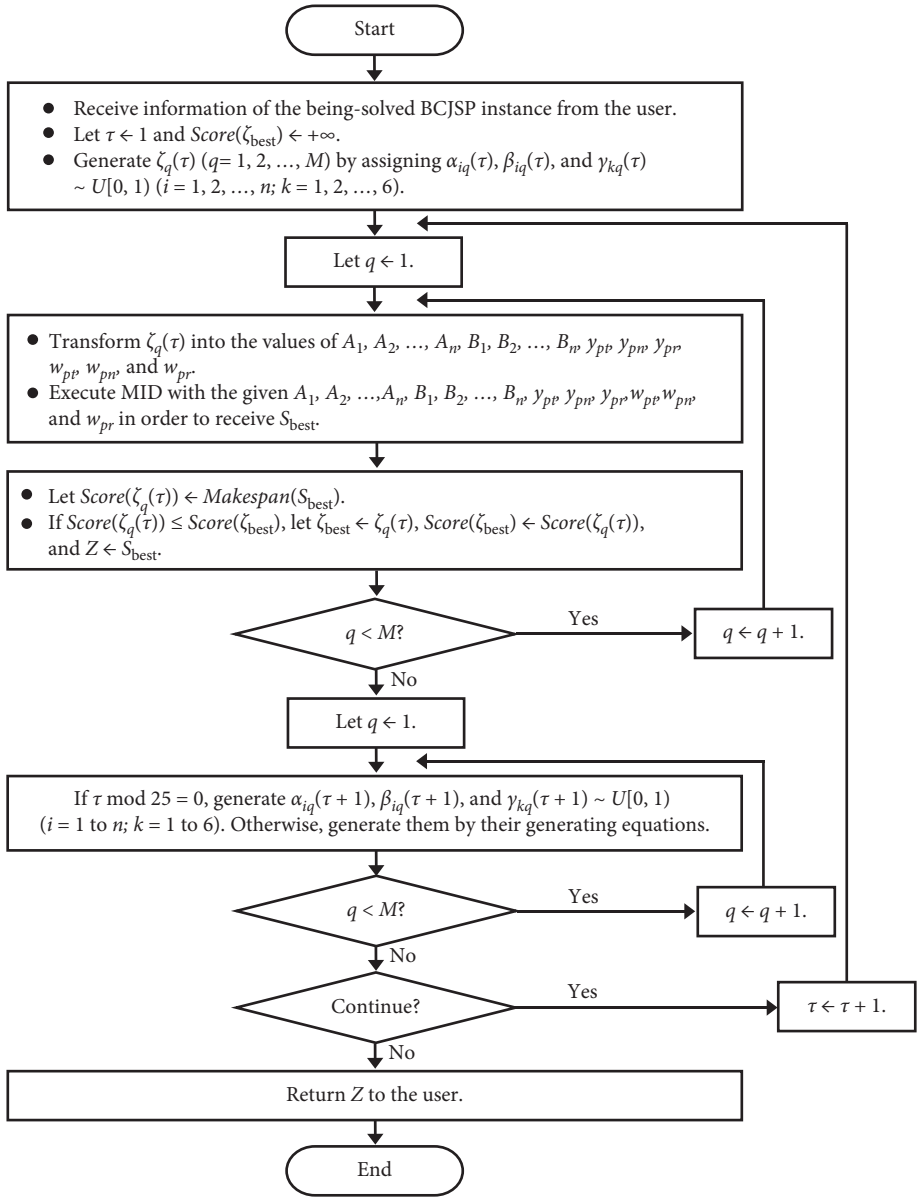


Figure 4: A flowchart of TOP.

(i) Let m and n, respectively, be the number of all machines and the number of all jobs in the being-solved BCJSP.

(ii) Let A_i and B_i, respectively, stand for the start operation and the operation-precedence-relation direction of the job J_i.

(iii) Let M be the number of all members in the TOP's population.

(iv) Let $\zeta_q(\tau) \equiv ((\alpha_{1q}(\tau), \alpha_{2q}(\tau), \ldots, \alpha_{nq}(\tau)), \beta_{1q}(\tau), \beta_{2q}(\tau), \ldots, \beta_{nq}(\tau), \gamma_{1q}(\tau), \gamma_{2q}(\tau), \ldots, \gamma_{6q}(\tau))$ represent the q-th member (where $q = 1, 2, \ldots, M$) in the TOP's population at the τ-th iteration. In addition, let $Score(\zeta_q(\tau))$ stand for the performance score of $\zeta_q(\tau)$. Note that the lower the performance score, the better the performance.

(v) Let $\alpha_{iq}(\tau) \in \mathbb{R}$ represent A_i (where $i = 1, 2, \ldots, n$). To transform $\alpha_{iq}(\tau)$ into A_i, let $k \leftarrow$ the integer part of $m\alpha_{iq}(\tau) + 1$. After that, reassign $k \leftarrow 1$ if $k < 1$, and reassign $k \leftarrow m$ if $k > m$. Finally, let $A_i \leftarrow O_{ik}$.

(vi) Let $\beta_{iq}(\tau) \in \mathbb{R}$ represent B_i (where $i = 1, 2, \ldots, n$). In the transformation, let $B_i \leftarrow$ clockwise if $\beta_{iq}(\tau) < 0.5$, and let $B_i \leftarrow$ counterclockwise otherwise.

(vii) Let $\gamma_{1q}(\tau)$, $\gamma_{2q}(\tau)$, $\gamma_{3q}(\tau)$, $\gamma_{4q}(\tau)$, $\gamma_{5q}(\tau)$, and $\gamma_{6q}(\tau) \in \mathbb{R}$ represent the MID's y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr}, respectively. In their transformations, let $y_{pt} \leftarrow \gamma_{1q}(\tau)$, $y_{pn} \leftarrow \gamma_{2q}(\tau)$, $y_{pr} \leftarrow \gamma_{3q}(\tau)$, $w_{pt} \leftarrow \gamma_{4q}(\tau)$, $w_{pn} \leftarrow \gamma_{5q}(\tau)$, and $w_{pr} \leftarrow \gamma_{6q}(\tau)$.

(viii) Let S_{best} and $Makespan(S_{best})$ stand for the best schedule returned from MID and its makespan, respectively.

(1) Receive information of the being-solved BCJSP instance from the user.
(2) Let $\tau \leftarrow 1$ and $Score(\zeta_{\text{best}}) \leftarrow +\infty$.
(3) Generate $\zeta_q(\tau)$, where $q = 1, 2, \ldots, M$, by randomly generating $\alpha_{iq}(\tau)$, $\beta_{iq}(\tau)$, and $\gamma_{kq}(\tau) \sim U[0, 1)$ ($i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, 6$).
(4) Evaluate $Score(\zeta_q(\tau))$ and update ζ_{best} and Z by using Steps 4.1 to 4.6.
(4.1) Let $q \leftarrow 1$.
(4.2) Transform $\zeta_q(\tau)$ into the values of $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n, y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr}.
(4.3) Execute MID with the values of $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n, y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}$, and w_{pr} taken from Step 4.2. This is done for receiving S_{best} from MID.
(4.4) Let $Score(\zeta_q(\tau)) \leftarrow Makespan(S_{\text{best}})$.
(4.5) If $Score(\zeta_q(\tau)) \leq Score(\zeta_{\text{best}})$, then let $\zeta_{\text{best}} \leftarrow \zeta_q(\tau)$, $Score(\zeta_{\text{best}}) \leftarrow Score(\zeta_q(\tau))$, and $Z \leftarrow S_{\text{best}}$.
(4.6) If $q < M$, then let $q \leftarrow q + 1$ and repeat from Step 4.2. Otherwise, go to Step 5.
(5) Update $\zeta_q(\tau + 1)$, where $q = 1, 2, \ldots, M$, by using Steps 5.1 to 5.3.
(5.1) Let $q \leftarrow 1$.
(5.2) If $\tau \bmod 25 = 0$, then randomly generate $\alpha_{iq}(\tau + 1)$, $\beta_{iq}(\tau + 1)$, and $\gamma_{kq}(\tau + 1) \sim U[0, 1)$, where $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, 6$. Otherwise, generate $\alpha_{iq}(\tau + 1)$, $\beta_{iq}(\tau + 1)$, and $\gamma_{kq}(\tau + 1)$ by below three equations, respectively ($i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, 6$). Let u_1 and $u_2 \sim U[0, 1)$.

$$\alpha_{iq}(\tau + 1) = \begin{cases} \alpha_{iq}(\tau) + 0.025u_1 - 0.01u_2 & \text{if } \alpha_{iq}(\tau) < \alpha_{i\,\text{best}}, \\ \alpha_{iq}(\tau) - 0.025u_1 + 0.01u_2 & \text{if } \alpha_{iq}(\tau) > \alpha_{i\,\text{best}}, \\ \alpha_{iq}(\tau) + 0.01u_1 - 0.01u_2 & \text{if } \alpha_{iq}(\tau) = \alpha_{i\,\text{best}}. \end{cases}$$

$$\beta_{iq}(\tau + 1) = \begin{cases} \beta_{iq}(\tau) + 0.025u_1 - 0.01u_2 & \text{if } \beta_{iq}(\tau) < \beta_{i\,\text{best}}, \\ \beta_{iq}(\tau) - 0.025u_1 + 0.01u_2 & \text{if } \beta_{iq}(\tau) > \beta_{i\,\text{best}}, \\ \beta_{iq}(\tau) + 0.01u_1 - 0.01u_2 & \text{if } \beta_{iq}(\tau) = \beta_{i\,\text{best}}. \end{cases}$$

$$\gamma_{kq}(\tau + 1) = \begin{cases} \gamma_{kq}(\tau) + 0.025u_1 - 0.01u_2 & \text{if } \gamma_{kq}(\tau) < \gamma_{k\,\text{best}}, \\ \gamma_{kq}(\tau) - 0.025u_1 + 0.01u_2 & \text{if } \gamma_{kq}(\tau) > \gamma_{k\,\text{best}}, \\ \gamma_{kq}(\tau) + 0.01u_1 - 0.01u_2 & \text{if } \gamma_{kq}(\tau) = \gamma_{k\,\text{best}}. \end{cases}$$

(5.3) If $q < M$, then let $q \leftarrow q + 1$ and repeat from Step 5.2. Otherwise, go to Step 6.
(6) If the stopping criterion is not met, then let $\tau \leftarrow \tau + 1$ and repeat from Step 4. Otherwise, return Z to the user.

Algorithm 3: The procedure of TOP.

(ix) Let $\zeta_{\text{best}} \equiv (\alpha_{1best}, \alpha_{2\text{best}}, \ldots, \alpha_{n\text{best}}, \beta_{1\text{best}}, \beta_{2\text{best}}, \ldots, \beta_{n\text{best}}, \gamma_{1\text{best}}, \gamma_{2\text{best}}, \ldots, \gamma_{6\text{best}})$ be the best $\zeta_q(\tau)$ ever found by the population. In addition, let $Score(\zeta_{\text{best}})$ stand for the performance score of ζ_{best}.

(x) Let Z represent the best schedule ever found by the TOP's population.

In Algorithm 3, TOP starts its procedure by receiving information of a BCJSP instance from the user. TOP assigns $\tau \leftarrow 1$ and $Score(\zeta_{\text{best}}) \leftarrow +\infty$; then, it generates $\zeta_q(\tau)$ randomly. After that, TOP starts a repeated loop by executing MID M times. In the q-th execution (where $q = 1, 2, \ldots, M$), MID is executed with $\zeta_q(\tau)$-given parameter values to return S_{best}; then, let $Score(\zeta_q(\tau)) \leftarrow Makespan(S_{\text{best}})$. If TOP finds any $\zeta_q(\tau)$ better than or equal to ζ_{best}, then let this $\zeta_q(\tau)$ and its corresponding S_{best} become a new ζ_{best} and a new Z, respectively. After that, TOP chooses to update each member of $\zeta_q(\tau + 1)$ by either the two opposite-direction vectors or the reinitialization. If the stopping criterion is not met, TOP assigns $\tau \leftarrow \tau + 1$ and starts the repeated loop's next round.

As mentioned, there are two variants of MUL, i.e., MUL-B and MUL-T. The difference between MUL-B and MUL-T in their TOPs is given as follows. In MUL-T, $\zeta_q(\tau)$ of its TOP's population consists of $\alpha_{iq}(\tau)$, $\beta_{iq}(\tau)$, and $\gamma_{kq}(\tau)$. This means TOP in MUL-T is identical to Algorithm 3. In MUL-B, $\zeta_q(\tau)$ of its TOP's population consists of only $\alpha_{iq}(\tau)$ and $\beta_{iq}(\tau)$. It means that $\zeta_q(\tau)$ of TOP in MUL-B is equivalent to $(\alpha_{1q}(\tau), \alpha_{2q}(\tau), \ldots, \alpha_{nq}(\tau), \beta_{1q}(\tau), \beta_{2q}(\tau), \ldots, \beta_{nq}(\tau))$. Thus, the procedure of TOP in MUL-B is modified from Algorithm 3 by removing $\gamma_{kq}(\tau)$ from Step 3; removing y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn} and w_{pr} from Steps 4.2 and 4.3; and removing $\gamma_{kq}(\tau + 1)$ and its generating equation from Step 5.2.

5. Results

The performance of the two proposed variants of MUL (i.e., MUL-B and MUL-T) was evaluated via an experiment on 53 BCJSP instances. These BCJSP instances were modified from the well-known JSP instances, i.e., FT06, FT10, and FT20 instances of [17]; LA01 to LA40 instances of [18]; and ORB01 to ORB10 instances of [19]. The modification of each instance was done by letting the operation-precedence-relation chains of all jobs be circular and bidirectional. In this paper, let BCFT06, BCFT10, and BCFT20 represent the BCJSP instances modified from FT06, FT10, and FT20, respectively. Let BCLA01 to BCLA40 represent the BCJSP instances modified from LA01 to LA40, respectively. Then, let BCORB01 to BCORB10 represent the BCJSP instances modified from ORB01 to ORB10, respectively. For each BCJSP instance, let its original JSP instance stand for the JSP instance later modified to become it. For example, FT06 is the original JSP instance of BCFT06.

Because of the BCJSP's novelty, no algorithms excepting MUL have intentionally been developed for BCJSP. Thus, to

evaluate the performance of MUL-B and MUL-T, their performance was compared with each other and with the performance of the existing JSP-solving algorithms. Without modifications, the existing JSP-solving algorithms can solve BCJSP as well. However, they hardly find good solutions for BCJSP. This is because, for each BCJSP instance, the solution space of its original JSP instance is just a subset of its whole solution space. Consequently, for each BCJSP instance, the JSP-solving algorithms cannot return better solutions than the optimal solution of its original JSP instance. This is the reason that the original JSP instance's optimal solution is used as the best possible result from all JSP-solving algorithms on each BCJSP instance.

The settings of MUL-B and MUL-T for the experiment are summarized as follows:

(i) MUL-B and MUL-T were coded in C# and executed on an Intel® Core™ i5-3320M CPU processor @ 2.60 GHz with RAM of 4 GB (3.87 GB usable).

(ii) In each of MUL-B and MUL-T, let the population of MID consist of two members (i.e., $N = 2$ in Algorithm 2), and let MID be stopped after the 10-th iteration (i.e., $t = 10$ is the maximum iteration in Algorithm 2).

(iii) In only MUL-B, the constant value of 0.5 was assigned to the MID's input parameters y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr}.

(iv) In each of MUL-B and MUL-T, let the population of TOP consist of two members (i.e., $M = 2$ in Algorithm 3), and let TOP be stopped if it could not find an improving solution (a better solution) within 100 consecutive iterations.

(v) Each of MUL-B and MUL-T was executed for three runs on each BCJSP instance with different random seed numbers.

Reasons for the above-mentioned parameter settings are given as follows. For limiting MUL-B's and MUL-T's computational time, the population sizes of their MIDs and TOPs were set to 2, the smallest possible size. With the same reason, their MIDs were set to stop after their 10-th iterations. However, for avoiding premature stops, their TOPs were set to stop after 100 consecutive iterations without finding a better solution. For performing as well as MUPLA [16], the parameters y_{pt}, y_{pn}, y_{pr}, w_{pt}, w_{pn}, and w_{pr} of MID in MUL-B were set to 0.5. This made the equations in Step 5.2 of Algorithm 2 become identical to those of MUPLA [16].

The experiment's results on the 53 BCJSP instances are presented in Table 1. Terms and abbreviations used in Table 1 and in its discussion are defined as follows:

(i) Let a solution and a solution value stand for a schedule and a schedule's makespan, respectively.

(ii) For each BCJSP instance, let its original JSP instance mean the JSP instance that was later modified to become it. For example, LA01 is the original JSP instance of BCLA01.

(iii) Let *Ins* column present the name of each BCJSP instance.

(iv) Let *JSP Opt* column present the optimal solution value of the original JSP instance of each BCJSP instance. The values in this column were given by published articles, e.g., [20, 21].

(v) For each of MUL-B and MUL-T, let *Best* and *Avg* stand for its best final solution value and its average final solution value, respectively, over three runs.

(vi) Let *UB* stand for the upper bound of the optimal solution value of each BCJSP instance. *UB* in this paper is a minimum among the values of *JSP Opt*, *Best* of MUL-B, and *Best* of MUL-T.

(vii) For each of MUL-B and MUL-T, let *%BD* stand for a percent deviation of its *Best* from *UB*, and let *% AD* stand for a percent deviation of its *Avg* from *UB*.

(viii) For each *JSP Opt*, let *%JD* stand for its percent deviation from *UB*.

(ix) For each of MUL-B and MUL-T, let *Avg Iters* column present the average number of iterations used until stopped on each instance. In parentheses, this column presents the average minimum number of iterations required to find the last improving solution. Note that the number of iterations used by each of MUL-B and MUL-T means the number of iterations used by its TOP.

(ix) For each of MUL-B and MUL-T, let *Avg time* column present an average computational time consumed until stopped on each instance in HH: MM: SS form. In parentheses, this column presents an average minimum computational time required to find the last improving solution.

Table 2 then summarizes the results shown in Table 1. In Table 2, the 53 BCJSP instances are classified into nine categories based on the numbers of machines (m) and the numbers of jobs (n). Terms and abbreviations used in Table 2 are defined as follows:

(i) *Category* column presents each category of the 53 BCJSP instances in the form $m \times n$, where m is the number of machines and n is the number of jobs.

(ii) *Members* column presents all instances that are members of each category. Then, *No. of Ins* column presents the number of all instances in each category.

(iii) Let *Avg %JD* stand for an average *%JD* over all instances in each category.

(iv) For each of MUL-B and MUL-T, let *Avg %BD* and *Avg %AD* stand for an average *%BD* and an average *%AD*, respectively, over all instances in each category.

(v) For each of MUL-B and MUL-T, let *Avg Iters* column present the average number of iterations used until stopped. In parentheses, this column presents the average minimum number of iterations required to

Table 1: Experiment's results.

Ins	UB	JSP Opt	%JD	MUL-B						MUL-T					
				Best	%BD	Avg	%AD	Avg iters	Avg time	Best	%BD	Avg	%AD	Avg iters	Avg time
BCFT06	47	55	17.02	48	2.13	49.3	4.89	178 (78)	0:00:10 (0:00:04)	47	0.00	48.7	3.62	217 (117)	0:00:12 (0:00:06)
BCFT10	739	930	25.85	780	5.55	786.0	6.36	150 (50)	0:07:21 (0:02:25)	739	0.00	756.0	2.30	183 (83)	0:08:56 (0:04:14)
BCFT20	1119	1165	4.11	1119	0.00	1119.0	0.00	101 (1)	0:08:18 (0:00:03)	1119	0.00	1119.0	0.00	101 (1)	0:08:15 (0:00:04)
BCLA01	666	666	0.00	666	0.00	666.0	0.00	112 (12)	0:00:29 (0:00:03)	666	0.00	666.0	0.00	112 (12)	0:00:30 (0:00:03)
BCLA02	635	655	3.15	635	0.00	635.0	0.00	114 (14)	0:00:30 (0:00:03)	635	0.00	635.0	0.00	103 (3)	0:00:27 (0:00: 00.4)
BCLA03	588	597	1.53	588	0.00	588.0	0.00	115 (15)	0:00:30 (0:00:04)	588	0.00	588.0	0.00	103 (3)	0:00:26 (0:00:01)
BCLA04	537	590	9.87	537	0.00	537.0	0.00	135 (35)	0:00:38 (0:00:10)	537	0.00	537.0	0.00	190 (90)	0:00:55 (0:00:25)
BCLA05	593	593	0.00	593	0.00	593.0	0.00	101 (1)	0:00:26 (0:00:00.3)	593	0.00	593.0	0.00	103 (3)	0:00:26 (0:00:01)
BCLA06	926	926	0.00	926	0.00	926.0	0.00	101 (1)	0:02:21 (0:00:01)	926	0.00	926.0	0.00	101 (1)	0:02:22 (0:00:01)
BCLA07	869	890	2.42	869	0.00	869.0	0.00	101 (1)	0:02:22 (0:00:01)	869	0.00	869.0	0.00	101 (1)	0:02:14 (0:00:01)
BCLA08	863	863	0.00	863	0.00	863.0	0.00	101 (1)	0:02:35 (0:00:01)	863	0.00	863.0	0.00	101 (1)	0:02:40 (0:00:01)
BCLA09	951	951	0.00	951	0.00	951.0	0.00	101 (1)	0:02:38 (0:00:02)	951	0.00	951.0	0.00	101 (1)	0:02:30 (0:00:01)
BCLA10	958	958	0.00	958	0.00	958.0	0.00	101 (1)	0:02:22 (0:00:01)	958	0.00	958.0	0.00	101 (1)	0:02:21 (0:00:01)
BCLA11	1222	1222	0.00	1222	0.00	1222.0	0.00	101 (1)	0:07:15 (0:00:05)	1222	0.00	1222.0	0.00	101 (1)	0:06:51 (0:00:03)
BCLA12	1039	1039	0.00	1039	0.00	1039.0	0.00	101 (1)	0:08:38 (0:00:06)	1039	0.00	1039.0	0.00	101 (1)	0:08:46 (0:00:03)
BCLA13	1150	1150	0.00	1150	0.00	1150.0	0.00	101 (1)	0:08:12 (0:00:05)	1150	0.00	1150.0	0.00	101 (1)	0:08:02 (0:00:05)
BCLA14	1292	1292	0.00	1292	0.00	1292.0	0.00	101 (1)	0:07:02 (0:00:03)	1292	0.00	1292.0	0.00	101 (1)	0:06:43 (0:00:04)
BCLA15	1207	1207	0.00	1207	0.00	1207.0	0.00	101 (1)	0:07:27 (0:00:05)	1207	0.00	1207.0	0.00	101 (1)	0:07:21 (0:00:03)
BCLA16	798	945	18.42	798	0.00	806.0	1.00	162 (62)	0:08:00 (0:03:11)	810	1.50	814.0	2.01	153 (53)	0:07:20 (0:02:37)
BCLA17	717	784	9.34	735	2.51	735.0	2.51	141 (41)	0:07:04 (0:02:00)	717	0.00	731.7	2.05	234 (134)	0:11:43 (0:06:47)
BCLA18	765	848	10.85	765	0.00	775.7	1.40	191 (91)	0:09:25 (0:04:20)	768	0.39	777.7	1.66	168 (68)	0:08:22 (0:03:23)
BCLA19	783	842	7.54	783	0.00	802.3	2.46	184 (84)	0:09:25 (0:04:21)	801	2.30	808.0	3.19	248 (148)	0:12:17 (0:07:18)
BCLA20	810	902	11.36	812	0.25	823.0	1.60	157 (57)	0:07:35 (0:02:48)	810	0.00	828.7	2.31	182 (82)	0:08:59 (0:04:01)
BCLA21	967	1046	8.17	981	1.45	985.7	1.93	212 (112)	1:05:44 (0:34:46)	967	0.00	979.7	1.31	181 (81)	0:55:07 (0:24:20)
BCLA22	900	927	3.00	911	1.22	925.3	2.81	172 (72)	0:54:31 (0:23:14)	900	0.00	905.0	0.56	150 (50)	0:47:03 (0:15:45)
BCLA23	1032	1032	0.00	1032	0.00	1032.0	0.00	141 (41)	0:43:54 (0:12:49)	1032	0.00	1032.0	0.00	141 (41)	0:42:37 (0:13:01)
BCLA24	932	935	0.32	938	0.64	948.3	1.75	180 (80)	0:55:37 (0:24:16)	932	0.00	950.3	1.96	174 (74)	0:54:07 (0:23:12)
BCLA25	907	977	7.72	918	1.21	935.0	3.09	226 (126)	1:09:46 (0:39:20)	907	0.00	927.0	2.21	253 (153)	1:17:00 (0:46:58)
BCLA26	1218	1218	0.00	1218	0.00	1218.0	0.00	146 (46)	2:35:48 (0:47:58)	1218	0.00	1218.0	0.00	119 (19)	2:09:40 (0:20:37)
BCLA27	1207	1235	2.32	1218	0.91	1222.0	1.24	162 (62)	2:58:45 (1:08:36)	1207	0.00	1222.7	1.30	162 (62)	3:02:29 (1:10:44)

Table 1: Continued.

Ins	UB	JSP Opt	%JD	MUL-B						MUL-T					
				Best	%BD	Avg	%AD	Avg iters	Avg time	Best	%BD	Avg	%AD	Avg iters	Avg time
BCLA28	1216	1216	0.00	1216	0.00	1217.7	0.14	223 (123)	4:03:14 (2:16:13)	1216	0.00	1216.0	0.00	177 (77)	3:13:08 (1:24:44)
BCLA29	1105	1152	4.25	1105	0.00	1105.0	0.00	152 (52)	2:46:47 (0:55:04)	1105	0.00	1115.0	0.90	188 (88)	3:26:20 (1:37:06)
BCLA30	1355	1355	0.00	1355	0.00	1355.0	0.00	101 (1)	1:43:32 (0:00:51)	1355	0.00	1355.0	0.00	101 (1)	1:39:45 (0:00:36)
BCLA31	1784	1784	0.00	1784	0.00	1784.0	0.00	101 (1)	7:50:40 (0:02:52)	1784	0.00	1784.0	0.00	101 (1)	8:55:01 (0:04:44)
BCLA32	1850	1850	0.00	1850	0.00	1850.0	0.00	102 (2)	10:21:25 (0:07:08)	1850	0.00	1850.0	0.00	102 (2)	10:30:37 (0:10:30)
BCLA33	1719	1719	0.00	1719	0.00	1719.0	0.00	101 (1)	9:13:34 (0:07:22)	1719	0.00	1719.0	0.00	101 (1)	9:35:45 (0:06:18)
BCLA34	1721	1721	0.00	1721	0.00	1721.0	0.00	101 (1)	10:45:44 (0:06:43)	1721	0.00	1721.0	0.00	101 (1)	11:06:51 (0:08:12)
BCLA35	1888	1888	0.00	1888	0.00	1888.0	0.00	101 (1)	7:31:40 (0:02:25)	1888	0.00	1888.0	0.00	101 (1)	8:07:04 (0:04:14)
BCLA36	1186	1268	6.91	1186	0.00	1210.7	2.08	196 (96)	3:59:35 (1:52:51)	1207	1.77	1218.0	2.70	177 (77)	3:40:50 (1:37:49)
BCLA37	1247	1397	12.03	1247	0.00	1271.3	1.95	220 (120)	4:36:09 (2:24:27)	1292	3.61	1295.3	3.87	171 (71)	3:38:52 (1:33:27)
BCLA38	1114	1196	7.36	1165	4.58	1173.0	5.30	198 (98)	4:11:10 (2:04:43)	1114	0.00	1145.0	2.78	195 (95)	4:15:04 (2:07:00)
BCLA39	1186	1233	3.96	1186	0.00	1196.0	0.84	195 (95)	4:06:29 (1:57:48)	1191	0.42	1192.7	0.56	214 (114)	4:26:56 (2:19:38)
BCLA40	1150	1222	6.26	1156	0.52	1182.3	2.81	166 (66)	3:24:58 (1:24:33)	1150	0.00	1179.3	2.55	145 (45)	2:58:02 (0:56:47)
BCORB01	789	1059	34.22	812	2.92	818.3	3.71	248 (148)	0:12:31 (0:07:30)	789	0.00	810.3	2.70	150 (50)	0:07:45 (0:02:35)
BCORB02	763	888	16.38	763	0.00	783.0	2.62	164 (64)	0:08:24 (0:03:19)	797	4.46	803.7	5.33	173 (73)	0:09:01 (0:03:47)
BCORB03	741	1005	35.63	741	0.00	764.3	3.14	192 (92)	0:09:29 (0:04:28)	773	4.32	780.7	5.36	333 (233)	0:16:43 (0:11:39)
BCORB04	831	1005	20.94	839	0.96	849.7	2.25	163 (63)	0:08:01 (0:03:07)	831	0.00	839.3	1.00	217 (117)	0:10:54 (0:05:48)
BCORB05	705	887	25.82	705	0.00	717.3	1.74	213 (113)	0:10:26 (0:05:42)	727	3.12	729.0	3.40	180 (80)	0:08:51 (0:03:56)
BCORB06	817	1010	23.62	822	0.61	835.3	2.24	136 (36)	0:06:40 (0:01:43)	817	0.00	832.3	1.87	187 (87)	0:09:27 (0:04:21)
BCORB07	353	397	12.46	356	0.85	358.7	1.61	201 (101)	0:09:45 (0:04:58)	353	0.00	358.0	1.42	155 (55)	0:07:35 (0:02:42)
BCORB08	671	899	33.98	675	0.60	684.3	1.98	208 (108)	0:10:14 (0:05:19)	671	0.00	683.3	1.83	186 (86)	0:09:19 (0:04:10)
BCORB09	772	934	20.98	773	0.13	792.0	2.59	154 (54)	0:07:50 (0:02:42)	772	0.00	789.3	2.24	179 (79)	0:09:15 (0:04:01)
BCORB10	780	944	21.03	812	4.10	825.7	5.86	193 (93)	0:09:30 (0:04:34)	780	0.00	788.7	1.12	266 (166)	0:13:24 (0:08:19)

find the last improving solution. Note that the number of iterations used by each of MUL-B and MUL-T means the number of iterations used by its TOP.

(vi) For each of MUL-B and MUL-T, let *Avg time* column present an average computational time consumed until stopped. In parentheses, this column presents an average minimum computational time required to find the last improving solution.

(vii) In the competition between MUL-B and MUL-T, let *W*, *D*, and *L* in *W-D-L* present the numbers of instances won, drawn, and lost, respectively. On each instance, MUL-B is judged to win MUL-T if the MUL-B's *%BD* is better than the MUL-T's *%BD*, and vice versa. In addition, they are judged to draw if their *%BDs* are equal.

Figure 5 presents the rates of solution improvements over iterations of MUL-B and MUL-T. In making the figure, *Avg-%AD-over-iteration* plots and *Avg-%BD-over-iteration* plots were plotted and compared with *Avg %JD*. Let *Avg %JD* be calculated from an average *%JD* over 53 instances. For each of MUL-B and MUL-T, let *Avg-%AD-over-iteration* plot and *Avg-%BD-over-iteration* plot present the average *%AD* over 53 instances

Table 2: A summary of experiment's results.

Category	Members	No. of Ins	Avg %JD	MUL-B Avg %BD	MUL-B W-D-L	MUL-B Avg %AD	MUL-B Avg iters	MUL-B Avg time	MUL-T Avg %BD	MUL-T W-D-L	MUL-T Avg %AD	MUL-T Avg iters	MUL-T Avg time
5×10	BCLA01–05	5	2.91	0.00	0-5-0	0.00	115 (15)	0:00:31 (0:00:04)	0.00	0-5-0	0.00	122 (22)	0:00:33 (0:00:06)
5×15	BCLA06–10	5	0.48	0.00	0-5-0	0.00	101 (1)	0:02:28 (0:00:01)	0.00	0-5-0	0.00	101 (1)	0:02:25 (0:00:01)
5×20	BCFT20, BCLA11–15	6	0.69	0.00	0-6-0	0.00	101 (1)	0:07:49 (0:00:05)	0.00	0-6-0	0.00	101 (1)	0:07:40 (0:00:04)
6×6	BCFT06	1	17.02	2.13	0-0-1	4.89	178 (78)	0:00:10 (0:00:04)	0.00	1-0-0	3.62	217 (117)	0:00:12 (0:00:06)
10×10	BCFT10, BCLA16–20, BCORB01–10	16	20.53	1.16	6-0-10	2.69	178 (78)	0:08:51 (0:03:54)	1.01	10-0-6	2.49	200 (100)	0:09:59 (0:04:45)
10×15	BCLA21–25	5	3.84	0.90	0-1-4	1.92	186 (86)	0:57:54 (0:26:53)	0.00	4-1-0	1.21	180 (80)	0:55:11 (0:24:39)
10×20	BCLA26–30	5	1.31	0.18	0-4-1	0.28	157 (57)	2:49:37 (1:01:44)	0.00	1-4-0	0.44	149 (49)	2:42:16 (0:54:45)
10×30	BCLA31–35	5	0.00	0.00	0-5-0	0.00	101 (1)	9:08:37 (0:05:18)	0.00	0-5-0	0.00	101 (1)	9:39:04 (0:06:48)
15×15	BCLA36–40	5	7.30	1.02	3-0-2	2.60	195 (95)	4:03:40 (1:56:52)	1.16	2-0-3	2.49	180 (80)	3:47:57 (1:42:56)
All	All 53 instances	53	8.09	0.59	9-26-18	1.36	149 (49)	1:40:03 (0:21:05)	0.41	18-26-9	1.21	154 (54)	1:40:49 (0:19:22)

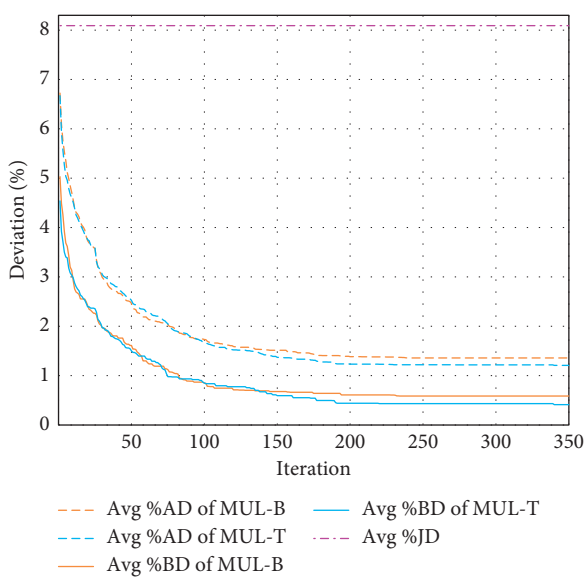


Figure 5: Avg %ADs and Avg %BDs over iterations of MUL-B and MUL-T compared with Avg %JD.

and the average *%BD* over 53 instances, respectively, from the first to the 350-th iterations. One obstacle of plotting each plot was that each algorithm was usually stopped before the 350-th iteration, as shown in Table 1. To deal with the obstacle, since the algorithms were stopped, the last values of *%AD* and *%BD* had been used in the plots until the 350-th iterations.

6. Discussion

6.1. Result Comparisons of MUL-B and MUL-T with Other Algorithms. As mentioned previously, no existing algorithms excepting MUL-B and MUL-T have been developed to solve BCJSP. This means no other existing BCJSP-solving algorithms could be used for comparison against MUL-B and MUL-T. It is the cause that, in this research, the MUL-B's and MUL-T's results were compared against *JSP Opt*. For each BCJSP instance, *JSP Opt* in Table 1 denotes the optimal solution value of its original JSP instance. (Note that, for example, FT06 is the original JSP instance of BCFT06.) Rather than comparison with many JSP-solving algorithms, let *JSP Opt* be a representative of the best possible result from all existing JSP-solving algorithms on BCJSP.

As shown in Table 1, MUL-B returns better solution values than *JSP Opt* on 32 instances, equivalent values on 20 instances, and worse values on one instance. This means MUL-B performs better than the best performance of the existing JSP-solving algorithms on 32 instances, has equivalent performance on 20 instances, and has worse performance on one instance. MUL-T returns better solution values than *JSP Opt* on 33 instances, equivalent values on 20 instances, and worse values on none of the instances. This means MUL-T performs better than the best performance of the existing JSP-solving algorithms on 33 instances, has equivalent performance on 20 instances, and has worse performance on none of the instances.

Note that, in Table 1, *%JD*, *%BD*, and *%AD* represent the percent deviations from *UB* of *JSP Opt*, *Best*, and *Avg*, respectively. One-sided paired t tests conclude that the population mean *%BDs* of MUL-B and MUL-T are both significantly better than the population mean *%JD* (with p values of 4×10^{-7} and 3×10^{-7}, respectively). Moreover, the population mean *%ADs* of MUL-B and MUL-T are also significantly better than the population mean *%JD* (with p values of 10×10^{-7} and 5×10^{-7}, respectively). As an

interpretation, the two variants of MUL outperform the existing JSP-solving algorithms on BCJSP.

A cause for the above-mentioned results is that MUL-B and MUL-T can both explore the whole solution space of the BCJSP instance, while the existing JSP-solving algorithms can explore only the original JSP instance's solution space. For each BCJSP instance, the solution space of its original JSP instance is just a small subset of its whole solution space. Then, a BCJSP instance usually has multiple better solutions than the optimal solution of its original JSP instance. Consequently, MUL-B and MUL-T both possibly find better solutions than the original JSP instance's optimal solution, whereas the existing JSP-solving algorithms cannot.

6.2. Result Comparisons between MUL-B and MUL-T. As shown in Tables 1 and 2, MUL-T performs slightly better than MUL-B in the average solution quality. Over 53 instances, *Avg %BD* of MUL-T (i.e., 0.41%) is slightly better than *Avg %BD* of MUL-B (i.e., 0.59%). Likewise, *Avg %AD* of MUL-T (i.e., 1.21%) is slightly better than *Avg %AD* of MUL-B (i.e., 1.36%). However, based on one-sided paired t tests, no sufficient evidence suggests that the population mean *%BD* of MUL-T is better than that of MUL-B (with p value of 0.24). In addition, no enough evidence suggests that the population mean *%AD* of MUL-T is better than that of MUL-B (with p value of 0.19).

By counting the instances won in Table 1, the outperformance of MUL-T over MUL-B can be detected more clearly. On each instance, let MUL-T be judged to win MUL-B if the MUL-T's *%BD* is lower than the MUL-B's *%BD*, and vice versa. Over 53 instances, MUL-T wins MUL-B on 18 instances, draws on 26 instances, and loses on nine instances. As noticed, the number of instances won by MUL-T (i.e., 18) is twice of the number of instances won by MUL-B (i.e., 9). Although a proportion of drawn instances (i.e., 26 out of total 53) is very high, most of the drawn instances are small or easy to solve (e.g., the instances in the 5×10, 5×15, 5×20, and 10×30 categories in Table 2). To determine whether the total number of all instances won by MUL-T is greater than that by MUL-B, a one-sided binomial test was conducted at a significance level of 0.1. Let a sample size be the number of instances not drawn from the total 53 instances (i.e., 27). The test's result concludes that, from all not-drawn instances, the number of all instances won by MUL-T is significantly greater than that by MUL-B (with p value of 0.061).

As mentioned previously, the main difference between MUL-B and MUL-T is in their TOPs. The MUL-B's and MUL-T's TOPs both control the start operations and the operation-precedence-relation directions in BCJSP. However, only TOP of MUL-T additionally controls the MID's input parameters as the extra function. The outperformance of MUL-T over MUL-B abovementioned indicates that this extra function can enhance the performance of MUL. This means the MID's performance can be enhanced by using TOP to control its input parameters. Once the MID's performance has been enhanced, MID can provide better input-parameter values for BOT. Consequently, BOT can find better solutions.

6.3. Number of Iterations and Computational Time Consumed. For MUL-B and MUL-T, the numbers of used iterations directly affect computational time spent. The more the number of iterations used, the longer the computational time consumed. Table 1 presents the average number of iterations and the average computational time of the three runs of each algorithm. For each instance, *Avg Iters* and *Avg time* columns present the average number of iterations and the average computational time, respectively, used until the stopping criterion is met. (Note that the stopping criterion was to stop when TOP could not find an improving solution within 100 consecutive iterations.) In parentheses, *Avg Iters* and *Avg time* columns show the average minimum number of iterations and the average minimum computational time, respectively, required to find the last improving solution. A summary of the data from Table 1 just-mentioned can be found in Table 2.

As mentioned above, the stopping criterion for each algorithm was to stop when its TOP could not find an improving solution within 100 consecutive iterations. The purpose of using this stopping criterion is to avoid a premature stop for each algorithm. However, it probably results in an unnecessary high consumption of the number of iterations and of computational time. For example, on BCLA34, MUL-B and MUL-T both find their best-found solutions at their first iterations; however, they have to proceed until the 101-st iterations. For computational time on BCLA34, the both algorithms find their best-found solutions within 10 minutes; however, they have to stop after around 11 hours. Such cases also happen on many other instances, not only BCLA34.

Because the given stopping criterion causes a very long computational time, the user should not wait until MUL-B and MUL-T meet their stopping criteria. The user can break off the algorithms' executions at any time to receive their current best-found solutions. To choose a breaking-off time, the user may use information from the values in parentheses of *Avg Iters* and *Avg time* columns in Table 1. *Avg Iters* column indicates that, on more than 50% of all instances, the algorithms find their last improving solutions within 55 iterations; on all instances excepting BCORB03, they find their last improving solutions within 170 iterations. *Avg time* column indicates that, on more than 50% of all instances, the algorithms find their last improving solutions within 4.5 minutes; on all instances, they find their last improving solutions within 145 minutes.

As suggested from *Avg Iters* column, MUL-B and MUL-T should be broken off during the 55-th to the 170-th iterations. In addition, as suggested from *Avg time* column, they should be broken off during the 4.5-th to the 145-th minutes. Thus, as a minimum requirement, the user should break off each algorithm at the 55-th iteration or the 4.5-th minute of computational time, whichever comes first. If an optimal or near-optimal solution is required, the user is suggested to break off each algorithm at the 170-th iteration

or the 145-th minute of computational time, whichever comes first. Moreover, the user may decide to break off each algorithm at any iteration index during the 55-th to the 170-th iterations or at any time during the 4.5-th to the 145-th minutes. The decision is made based on his trade-off between the computational time and the possibility of finding an improving solution. Note that the computational time in the above suggestions may be changed if the user's computer has a different specification from the computer used in this research.

6.4. Solution Improvement Rate over Iteration. Since started, *Avg-%AD-over-iteration* plots and *Avg-%BD-over-iteration* plots of MUL-B and MUL-T in Figure 5 have already been below *Avg %JD* (i.e., 8.09%). At their first iterations, *Avg % AD* of MUL-B, *Avg %AD* of MUL-T, *Avg %BD* of MUL-B, and *Avg %BD* of MUL-T are 6.72%, 6.67%, 5.03%, and 4.53%, respectively. Since then, the values of the four plots have still been reduced continuously. The four plots in the mentioned order finally become 1.36%, 1.21%, 0.59%, and 0.41%, respectively, at their final values. Obviously, all their final values are much lower than *Avg %JD*. This means MUL-B and MUL-T perform better than the JSP-solving algorithms on BCJSP because *Avg %JD* is the best possible result of the JSP-solving algorithms.

All given plots in Figure 5 behave similarly to each other in their patterns. These plots can be divided into three periods based on their similar patterns. The first period of each plot, where the value is reduced very quickly, is approximately started from the first iteration to the 55-th iteration. The second period, where the value is reduced gradually, is approximately started from the 55-th iteration to the 170-th iteration. The third period, where the value is reduced hardly, is approximately started from the 170-th iteration onwards. After the 170-the iteration, the plot has become more and more stable. The patterns of the plots emphasize that the user should break off each algorithm during the 55-th to the 170-th iterations, where the number of more iterations is a trade-off for the higher possibility of finding an improving solution.

7. Conclusions

MUL is the multilevel metaheuristic developed to solve the job-shop scheduling problem with bidirectional circular precedence constraints (BCJSP). MUL consists of TOP, MID, and BOT algorithms in its top, middle, and bottom levels, respectively. TOP is the population-based metaheuristic developed to control the start operation and the operation-precedence-relation direction of each job in BCJSP. If requested, TOP can also control the MID's input parameters. MID is the population-based metaheuristic developed to control the BOT's input parameters. BOT generates a subproblem of the BCJSP instance, in the form of JSP, by using the start operations and the operation-precedence-relation directions given by TOP. BOT then acts as a local search algorithm to solve the generated subproblem. The population in MID is evolved by the feedback from BOT, while the population in TOP is evolved by the feedback from MID. In this paper, there are two proposed variants of MUL, i.e., MUL-B and MUL-T. These two variants both use their TOPs to control the start operation and the operation-precedence-relation direction of each job. However, only MUL-T additionally uses its TOP to control the MID's input parameters. Because MUL-B and MUL-T were intentionally developed for BCJSP, they perform much better than the existing JSP-solving algorithms on BCJSP. When comparing the two MUL variants, MUL-T outperforms MUL-B slightly in the average solution quality and significantly in the number of instances won.

Data Availability

The data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The author acknowledges partial financial support for publication from the Thai-Nichi Institute of Technology, Thailand.

References

[1] E. Nowicki and C. Smutnicki, "A fast taboo search algorithm for the job shop problem," *Management Science*, vol. 42, no. 6, pp. 797–813, 1996.

[2] T. Yamada and R. Nakano, "Job-shop scheduling," in *Genetic Algorithms in Engineering Systems*, A. M. S. Zalzala and P. J. Fleming, Eds., pp. 134–160, The Institution of Electrical Engineers, London, UK, 1997.

[3] C. Blum, "Beam-ACO—hybridizing ant colony optimization with beam search: an application to open shop scheduling," *Computers & Operations Research*, vol. 32, no. 6, pp. 1565–1591, 2005.

[4] P. Pongchairerks and V. Kachitvichyanukul, "A two-level particle swarm optimisation algorithm for open-shop scheduling problem," *International Journal of Computing Science and Mathematics*, vol. 7, no. 6, pp. 575–585, 2016.

[5] P. Brucker, Y. N. Sotskov, and F. Werner, "Complexity of shop-scheduling problems with fixed number of jobs: a survey," *Mathematical Methods of Operations Research*, vol. 65, pp. 461–481, 2007.

[6] M. M. Ahmadian, M. Khatami, A. Salehipour, and T. C. E. Cheng, "Four decades of research on the open-shop scheduling problem to minimize the makespan," *European Journal of Operational Research*, vol. 295, no. 2, pp. 399–426, 2021.

[7] J. Zhang, G. Ding, Y. Zou, S. Qin, and J. Fu, "Review of job shop scheduling research and its new perspectives under industry 4.0," *Journal of Intelligent Manufacturing*, vol. 30, no. 4, pp. 1809–1830, 2019.

[8] B. Çaliş and S. Bulkan, "A research survey: review of AI solution strategies of job shop scheduling problem," *Journal of Intelligent Manufacturing*, vol. 26, no. 5, pp. 961–973, 2015.

[9] C. Lu, X. Li, L. Gao, W. Liao, and J. Yi, "An effective multi-objective discrete virus optimization algorithm for flexible job-shop scheduling problem with controllable processing times," *Computers & Industrial Engineering*, vol. 104, pp. 156–174, 2017.

[10] L. Yin, X. Li, L. Gao, C. Lu, and Z. Zhang, "A novel mathematical model and multi-objective method for the low-carbon flexible job shop scheduling problem," *Sustainable Computing: Informatics and Systems*, vol. 13, pp. 15–30, 2017.

[11] Z. Adak, M. Ö. A. Akan, and S. Bulkan, "Multiprocessor open shop problem: literature review and future directions," *Journal of Combinatorial Optimization*, vol. 40, pp. 547–569, 2020.

[12] J. Kuster, D. Jannach, and G. Friedrich, "Extending the RCPSP for modeling and solving disruption management problems," *Applied Intelligence*, vol. 31, no. 3, pp. 234–253, 2009.

[13] I. A. Chaudhry and A. A. Khan, "A research survey: review of flexible job shop scheduling techniques," *International Transactions in Operational Research*, vol. 23, no. 3, pp. 551–591, 2016.

[14] X. Li and L. Gao, "Review for flexible job shop scheduling," in *Engineering Applications of Computational Methods*, vol. 2, pp. 17–45, Springer, Berlin, Germany, 2020.

[15] Á. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 2, pp. 124–141, 1999.

[16] P. Pongchairerks, "An enhanced two-level metaheuristic algorithm with adaptive hybrid neighborhood structures for the job-shop scheduling problem," *Complexity*, vol. 2020, Article ID 3489209, 15 pages, 2020.

[17] H. Fisher and G. L. Thompson, "Probabilistic learning combinations of local job-shop scheduling rules," in *Industrial Scheduling*, J. F. Muth and G. L. Thompson, Eds., pp. 225–251, Prentice-Hall, Englewood, NJ, USA, 1963.

[18] S. Lawrence, *Resource Constrained Project Scheduling: An Experimental Investigation of Heuristic Scheduling Techniques (Supplement)*, Carnegie Mellon University, Pittsburgh, PA, USA, 1984.

[19] D. Applegate and W. Cook, "A computational study of the job-shop scheduling problem," *ORSA Journal on Computing*, vol. 3, no. 2, pp. 149–156, 1991.

[20] J. F. Gonçalves and M. G. C. Resende, "An extended akers graphical method with a biased random-key genetic algorithm for job-shop scheduling," *International Transactions in Operational Research*, vol. 21, no. 2, pp. 215–246, 2014.

[21] B. Peng, Z. Lü, and T. C. E. Cheng, "A tabu search/path relinking algorithm to solve the job shop scheduling problem," *Computers & Operations Research*, vol. 53, pp. 154–164, 2015.

[22] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization," *ACM Computing Surveys*, vol. 35, no. 3, pp. 268–308, 2003.

[23] H. R. Lourenço, O. C. Martin, and T. Stützle, "Iterated local search," in *International Series in Operations Research and Management Science*, vol. 57, pp. 321–354, Springer, Boston, MA, USA, 2003.

[24] S. Kande, C. Prins, L. Belgacem, and B. Redon, "Multi-start iterated local search for two-echelon distribution network for perishable products," in *Proceedings of the International Conference on Operations Research and Enterprise Systems*, pp. 294–303, Lisbon, Portugal, January 2015.

[25] M. Avci and S. Topaloglu, "A multi-start iterated local search algorithm for the generalized quadratic multiple knapsack problem," *Computers & Operations Research*, vol. 83, pp. 54–65, 2017.

[26] C.-W. Chiou and M.-C. Wu, "A GA-tabu algorithm for scheduling in-line steppers in low-yield scenarios," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11925–11933, 2009.

[27] T. Davidović, P. Hansen, and N. Mladenović, "Scheduling by VNS: experimental analysis," in *Proceedings of the Yugoslav Symposium on Operations Research*, pp. 319–322, Belgrade, Serbia, October 2001.

[28] M.-E. Marmion, F. Mascia, M. López-Ibáñez, and T. Stützle, "Automatic design of hybrid stochastic local search algorithms," *Hybrid Metaheuristics*, vol. 7919, pp. 144–158, 2013.

[29] M. Gen and R. Cheng, *Genetic Algorithms and Engineering Design*, John Wiley & Sons, New York, NY, USA, 1997.

[30] M. Dorigo and T. Stützle, *Ant Colony Optimization*, MIT Press, Cambridge, MA, USA, 2004.

[31] J. Kennedy, R. C. Eberhart, and Y. Shi, *Swarm Intelligence*, Morgan Kaufmann, Burlington, MA, USA, 2001.

[32] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[33] X. S. Yang and S. Deb, "Engineering optimisation by cuckoo search," *International Journal of Mathematical Modelling and Numerical Optimisation*, vol. 1, no. 4, pp. 330–343, 2010.

[34] E. K. Burke, M. Gendreau, M. Hyde et al., "Hyper-heuristic: a survey of the state of the art," *Journal of the Operational Research Society*, vol. 64, no. 12, pp. 1695–1724, 2013.

[35] X. Fu, F. T. S. Chan, B. Niu, N. S. H. Chung, and T. Qu, "A three-level particle swarm optimization with variable neighbourhood search algorithm for the production scheduling problem with mould maintenance," *Swarm and Evolutionary Computation*, vol. 50, Article ID 100572, 2019.

[36] A. Kattan and S. Fatima, "PSO as a meta-search for hyper-GA system to evolve optimal agendas for sequential multi-issue negotiation," in *Proceedings of the 2012 IEEE Congress on Evolutionary Computation*, Brisbane, Australia, June 2012.

[37] P. Pongchairerks, "A two-level metaheuristic algorithm for the job-shop scheduling problem," *Complexity*, vol. 2019, Article ID 8683472, 11 pages, 2019.

[38] J. Grefenstette, "Optimization of control parameters for genetic algorithms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 1, pp. 122–128, 1986.

[39] S.-J. Wu and P.-T. Chow, "Genetic algorithms for nonlinear mixed discrete-integer optimization problems via meta-genetic parameter optimization," *Engineering Optimization*, vol. 24, no. 2, pp. 137–159, 1995.

[40] C. Bierwirth and D. C. Mattfeld, "Production scheduling and rescheduling with genetic algorithms," *Evolutionary Computation*, vol. 7, no. 1, pp. 1–17, 1999.

[41] P. Pongchairerks, "Particle swarm optimization algorithms and their applications to scheduling problems," D. E. Dissertation, Asian Institute of Technology, Khlong Nueng, Thailand, 2008.

[42] P. Pongchairerks, "Particle swarm optimization algorithm applied to scheduling problems," *ScienceAsia*, vol. 35, no. 1, pp. 89–94, 2009.

[43] C. Bierwirth, "A generalized permutation approach to job shop scheduling with genetic algorithms," *Spectrum*, vol. 17, no. 2-3, pp. 87–92, 1995.

[44] Q. Luo, Y. Zhou, J. Xie, M. Ma, and L. Li, "Discrete bat algorithm for optimal problem of permutation flow shop

scheduling," *Science World Journal*, vol. 2014, Article ID 630280, 15 pages, 2014.

[45] M. N. Janardhanan, Z. Li, P. Nielsen, and Q. Tang, "Artificial bee colony algorithms for two-sided assembly line worker assignment and balancing problem," in *Advances in Intelligent Systems and Computing*, vol. 620, pp. 11–18, Springer, Cham, Germany, 2018.

Hindawi
Complexity
Volume 2021, Article ID 9914076, 8 pages
https://doi.org/10.1155/2021/9914076

Research Article

An Analytical Study of the External Environment of the Coevolution between Manufacturing and Logistics Based on the Logistic Model

Yunfei Zhou[1] and Li Yan[2]

[1]*School of Management, Xi'an University of Finance and Economics, Xi'an 710100, China*
[2]*School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710021, China*

Correspondence should be addressed to Li Yan; yanli_10702@163.com

Received 24 August 2021; Revised 12 October 2021; Accepted 20 October 2021; Published 3 November 2021

Academic Editor: Long Wang

This paper focuses on the external environmental capacity of the coevolution system of the manufacturing industry and logistics industry. This paper first constructs a dynamic model of the external environmental capacity of the coevolution system by using the logistic model and then simulates the effects of two factors: one factor is the institutional environment affected by the random interference factors of policy and the other factor is the industrial environment affected by the random interference factors of the industrial economy on the coevolution system. This paper discusses the cooperative mechanism of external random interference factors on system evolution, analyzes the nonlinear variation of external environmental factors with time, and gives the estimation method. Finally, we provide an example to prove our findings.

1. Introduction

In the historical process of the evolution of the social division of labor, the form of industrial organization presents an obvious evolutionary phenomenon. From the perspective of organizational ecology, the formation and evolution process of an organization is not the strategic choice and adaptation of decision-makers within the organization but the choice of the external environment. On the one hand, the impact of external environmental changes on organizational evolution will be recorded in the number of enterprises entering or exiting the population [1], which will change the population density, environmental capacity, and survival mode of the organization. On the other hand, environmental change produces industrial evolutionary action through the self-organization mechanism of the organization [2], and this evolutionary action is bound to have more or less impact on the environment. In other words, industrial evolution operates in a complex environment, and its essence has a very close interactive relationship with the environment.

The logistic differential equation has become the main tool for modeling physical, engineering, economic, and biological models [3, 4] because it provides more ability in estimating the natural behavior of the model, and it also provides higher degrees of freedom [5]. In addition, the logistic equation involves memory and genetic characteristics, which are essential to describe the behavior of the ecological model [6]. On the other hand, the logistic model can demonstrate the interaction relationship between two species since the earliest model was published by Verhulst [7] in 1838. May [8] added a linear term in the population environment capacity of the logistic growth equation. This linear term can be described as the density of the symbiotic population to the environmental capacity of the equation. The symbiotic function can be seen as a function to enlarge the population environment capacity. Lotka [9] and Vloterra [10] extended it to simulate the interaction between two populations. This mathematical model is called Lotka–Volterra (LV) model or predator-prey model. The LV model can be used to describe the dynamics of some real-

world models. For example, Zhou and Wang [11] used the LV model to calculate the symbiotic relationship and symbiotic coordination degree of an industrial economy and industrial ecology in 30 provinces in China. Meng et al. [12] used the LV model to explore the symbiotic stability of two types of enterprise population: central enterprise population and satellite digital enterprise population in the digital economic ecosystem. Mao et al. [13] took the predator as an online bank and the prey as the third online payment system. Using the LV model, they quantitatively analyzed and predicted the impact of commercial banks' online payment systems on the development of third-party online payment systems. Mohammed et al. [14] used the LV model to simulate the dynamic behavior of New Coronavirus. However, most of these studies are deterministic external environmental capacity. The model does not consider the external environmental capacity of the interaction between the two groups. The external environment of the LV model shows randomness and affects the ecosystem in many forms [15–17]. For example, the external environment to be considered in industrial evolution includes resources and energy supply, environmental protection, other population functions, economic development, market, and political system changes, and the changes in time show a nonlinear functional relationship. The existing research does not estimate the impact of external random interference factors on systems coevolution.

To solve this problem, this paper uses a logistic equation to build an external environment capacity model of coevolution between manufacturing and logistics, and then the mathematical statistics method is used to estimate the internal environmental capacity and external random interference factors. Finally, this paper validates the feasibility and effectiveness of the model based on the data collected from the manufacturing and logistics industry in Shaanxi Province.

2. Methodology

2.1. External Environmental Capacity Model of the Coevolution between Manufacturing and Logistics. The industrial activities of manufacturing and logistics can be likened to the self-organization behavior of the biopopulation in the ecosystem. In the process of evolution, various populations are constantly exchanging material, energy, and information with the external environment so that the structure of the population has evolved into an orderly and stable state, which shows the self-organization regularity of the great nature. The relationship between population density and environmental capacity can be described by the logistic growth equation, as shown in the following:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN\left(1 - \frac{N}{k}\right), \tag{1}$$

where N is the population density of time t. N refers to the individual quantity of each population unit space. It is the state variable describing the evolution process of the industrial system. r is called the natural growth rate of the population, which refers to the ratio of the natural increase in population to the total population, that is, the difference between the average birth rate and the average mortality rate of the population, which indicates that the population of individuals in the absence of inhibition of the maximum growth, reflecting the inherent characteristics of the species; K is the maximum capacity of the environment. It is the upper limit of population density. In other words, it represents the maximum population density that environmental resources can carry.

Hypothesis $N_1(t)$ and $N_2(t)$ are the population density of the manufacturing and logistics, respectively; r_1 and r_2 are the natural growth rate of the manufacturing population and the logistics population separately; and $K_1(t)$ and $K_2(t)$ are the separately maximum ambient capacity. The synergetic evolutionary kinetic model of two industrial populations is expressed as follows:

$$\begin{cases} f_1(N_1(t)) = \dfrac{\mathrm{d}N_1(t)}{\mathrm{d}t} = r_1 N_1(t)\left(1 - \dfrac{N_1(t)}{K_1(t)}\right), \\ f_2(N_2(t)) = \dfrac{\mathrm{d}N_2(t)}{\mathrm{d}t} = r_2 N_2(t)\left(1 - \dfrac{N_2(t)}{K_2(t)}\right). \end{cases} \tag{2}$$

The impact of external environmental changes on the internal population of the system can be simplified into two cases: one is the impact of the market on the efficiency of resource allocation so that the change of external random interference factors of environmental capacity affects the change of industrial population-scale and the other is to transfer the resource allocation of enterprises in the system, to change the scale of the existing industrial population from emerging industrial population to emerging industrial population [18]. On this basis, we continue to consider the synergy between manufacturing and logistics to expand the environmental capacity of the existing population system, to promote the change of population growth rate. We try to analyze the coevolution between the system and the external environment by extending the environmental capacity parameter in the logistic equation.

Hypothesis $\alpha_1 f_1(N_2)$ and $\alpha_2 f_2(N_1)$ separately represent increment values of the environmental capacity of S_1 and S_2, which results from the synergy between manufacturing population and logistics population separately. According to the logistic equation of the population growth of manufacturing and logistics, we have

$$\begin{cases} \dfrac{\mathrm{d}N_1(t)}{\mathrm{d}t} = r_1 N_1(t)\left[1 - \dfrac{N_1(t)}{K_1^0 + \alpha_1 f_1(N_2)}\right], \\ \dfrac{\mathrm{d}N_2(t)}{\mathrm{d}t} = r_2 N_2(t)\left[1 - \dfrac{N_2(t)}{K_2^0 + \alpha_2 f_2(N_1)}\right], \end{cases} \tag{3}$$

where k_1^0 and k_2^0, respectively, represent the size of environmental capacity with two population systems, i.e., $t = 0$ and $f = 0$; $K_1 = K_1^0$; $K_2 = K_2^0$.

So, we can obtain the environmental capacity model with a system based on (2), as follows:

$$\begin{cases} K_1(t) = K_1^0 + \alpha_1 f_1[N_2(t)] = f_1^k(N_2), \\ K_2(t) = K_2^0 + \alpha_2 f_2[N_1(t)] = f_2^k(N_1). \end{cases} \tag{4}$$

As the evolution process between manufacturing and logistics must be influenced by the random disturbance factors in the outside system, in this, we mainly consider two points, one is by the influence of the industrial system and the policy random disturbance factors due to the guidance of the system and policy forcing the rapid development of manufacturing and logistics, which promotes the change of internal organization structure and the transformation and upgrading of industrial structure; another is by the influence of the industrial economy random disturbance factor, which is reflected in the output value of the economy.

Based on this, we hypothesis that represents the external environmental change of system, and the change will cause the external environmental capacity K with the change in the system. I represents the institutional system environment random disturbance factor which on the external environmental capacity of the system coefficient of influence is β_1 and β_2, respectively. E represents the institutional economic environment random disturbance factor which on the external environmental capacity of the system coefficient of influence is γ_1 and γ_2, respectively. So, under the influence of the random disturbance factor of the exterior environment, the dynamics model of manufacturing and logistics should be amended to the following:

$$\begin{cases} \dfrac{\mathrm{d}N_1(t)}{\mathrm{d}t} = r_1 N_1(t)\left[1 - \dfrac{N_1(t)}{k_1^0 + \alpha_1 f_1(N_2) + \beta_1 f_1(I) + \gamma_1 f_1(E)}\right], \\ \dfrac{\mathrm{d}N_2(t)}{\mathrm{d}t} = r_2 N_2(t)\left[1 - \dfrac{N_2(t)}{k_2^0 + \alpha_2 f_2(N_1) + \beta_2 f_2(I) + \gamma_2 f_2(E)}\right], \end{cases} \tag{5}$$

where $I = f(t)$ and $E = f(t)$, respectively, represent the time function of industrial system environment and industrial economic environment with T-time change; $N_1(t)$ and $N_2(t)$ represent the population density of the manufacturing and logistics, respectively, and $N_1|_{t=0} = N_1^0$ and $N_2|_{t=0} = N_2^0$.

From (5), we can see the synergy between S_1 and S_2 and influence function about external random disturbance factors of the industrial system environment and the industrial economy environment, which affect the growth rate of the population by changing the K value of environment capacity. So, under the joint action of internal synergy force and external environment random disturbance factors, the dynamics model of external environment capacity should be amended again to

$$\begin{cases} k_1(t) = k_1^0 + \alpha_1 f_1(N_2) + \beta_1 f_1(I) + \gamma_1 f_1(E) = f_1^k(N_2, I, E), \\ k_2(t) = k_2^0 + \alpha_2 f_2(N_1) + \beta_2 f_2(I) + \gamma_2 f_2(E) = f_2^k(N_1, I, E). \end{cases} \tag{6}$$

2.2. Parameter Estimation of the Factors. For better parameter estimation, we choose any adjacent two years $[t_i, t_{i+1}]$ on the population growth curve of manufacturing and logistics, as the observation interval, and then the interval length is $\Delta t_{i+1} = t_{i+1} - t_i = 1$ in Figure 1.

It can be seen that the increment of the population density on the evolution curve $[t_i, t_{i+1}]$ is $\Delta N_1^{i+1} = N_1^{i+1} - N_1^i$, the average is $\mathrm{aver}N_1^{i+1} = N_1^i - N_1^{i+1}/2$, and there is $\mathrm{aver}N_1^{i+1} \in [\min N_1(t), \max N_1 | t_i \le t \le t_{i+1}]$.

So, the slope of a line with two endpoints on an arbitrary interval $[t_i, t_{i+1}]$ in Figure 1 curve is as follows: $\Delta N_1^{i+1}/\Delta t_{i+1} = \Delta N_1^{i+1}$.

At the same time, the curves in the interval $[t_i, t_{i+1}]$ are also the logistic curves of the environmental capacity which is k_i^{i+1}, and the slope of the curve is the growth rate of the manufacturing population density, that is, $dN_1(t)/dt = r_1 N_1(t)[1 - N_1(t)/k_1^{i+1}]$ when $t \in t_i, t_{i+1}$.

As the rate of change of the population density is not very large in the logistic curve of the interval $[t_i, t_{i+1}]$, it can be approximated that the slope of logistic curves is equal to the slope of the interval endpoint in the interval $[t_i, t_{i+1}]$.

So, approximately there is $\Delta y_1^{i+1} = r_1 N_1(t)\ [1 - N_1(t)/k_1^{i+1}] = r_1(t)\mathrm{aver}N_1^{i+1}[1 - \mathrm{aver}N_1^{i+1}/k_1^{i+1}]$.

The iterative formula for the environmental capacity of the manufacturing population can be obtained by the finishing arrangement, as follows:

$$k_1^{i+1} = \frac{r_1 \mathrm{aver}N_1^{i+1}}{r_1 - \Delta N_1^{i+1}/\mathrm{aver}N_1^{i+1}} = f_1^k(r_1). \tag{7}$$

As the environmental capacity of the manufacturing population density is greater than 0, i.e., $k_1^{i+1} > 0$, there is $r_1 - \Delta N_1^{i+1}/\mathrm{aver}N_1^{i+1} > 0$, that is, to satisfy

$$r_1 > \frac{\Delta N_1^{i+1}}{\mathrm{aver}N_1^{i+1}} \quad (i = 0, 1, 2, \ldots,). \tag{8}$$

Therefore, given one r_1 value is $\hat{r}_1$, we can obtain a set of estimates $\hat{k}_1^{i+1}$ $(i = 0,1,2, ..., n)$ between the partitions through (7), that is,

$$\hat{k}_1^{i+1} = \frac{\hat{r}_1 \mathrm{aver}N_1^{i+1}}{\hat{r}_1 - \Delta N_1^{i+1}/\mathrm{aver}N_1^{i+1}} = f_1^{\hat{k}}(\hat{r}_1). \tag{9}$$

Substituting the above estimate value $\hat{r}_1$ and $\hat{k}_1^{i+1}$, we can obtain a set of logistic estimates $\hat{N}_1^{i+1}$ $(i = 0,1,2, \ldots, n)$, that is:

$$\hat{N}_1^{i+1} = \frac{\hat{k}_1^{i+1}}{1 + \left(\hat{k}_1^{i+1} - N_1^i\right)/N_1^i e^{-\hat{r}_1}} = f_1^{\hat{N}}(\hat{r}_1, \hat{k}_1). \tag{10}$$

In the same vein, we can obtain a set of logistic estimates of the population density and environmental capacity of the logistics as follows:

$$\hat{N}_2^{i+1} = \frac{\hat{k}_2^{i+1}}{1 + \left(\hat{k}_2^{i+1} - N_1^i\right)/N_1^i e^{-\hat{r}_1}} = f_2^{\hat{N}}(\hat{r}_2, \hat{k}_2), \tag{11}$$

$$\hat{k}_2^{i+1} = \frac{\hat{r}_2 \mathrm{aver}N_2^{i+1}}{\hat{r}_2 - \Delta N_2^{i+1}/\mathrm{aver}N_2^{i+1}} = f_2^{\hat{k}}(\hat{r}_2). \tag{12}$$

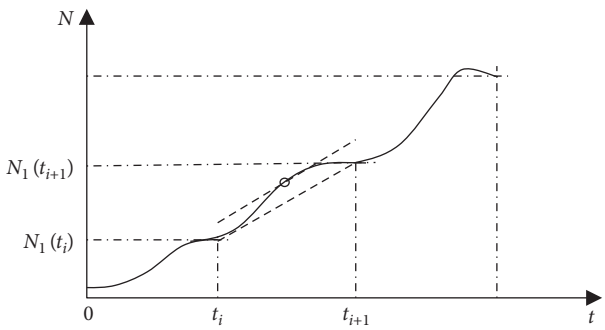


Figure 1: Population density evolution curve based on the logistic equation.

3. Examples Analysis

3.1. Data Selection. In order to verify the effectiveness and feasibility of the above models and methods, we select Shaanxi Province as the research object to study the estimation of the coevolution environmental capacity of the manufacturing industry and logistics industry. Because Shaanxi Province has not the value of manufacturing added statistics in the national economic accounting system and at the same time the logistics have not established a unified caliber to carry out added value statistics, in order to analyze the relationship of coevolution between the manufacturing and logistics of Shaanxi in a better way, in line with the principle of the availability of data and the consistency of statistical caliber, we use the industrial value-added (IVA) to measure the biomass of manufacturing to create the manufacturing population and use the transportation, warehousing, and postal services value-added to measure the biomass of logistics to create the logistics population, which will probe into the environmental capacity of the coevolution between manufacturing and logistics. Data originate from the "China Statistic Yearbook," "China Industrial Statistics Yearbook," "Shaanxi Statistic Yearbook," and "China Tertiary Industry Statistic Yearbook."

3.2. Numerical Estimation. For a given estimate, the variance of the estimated value of the manufacturing population density in each observation year is as follows:

$$d_j = s_j^2 = \sum_{i=0}^{n} \left(N_j^{i+1} - \widehat{N}_j^{i+1} \right)^2 = f_j^{s_j}\left(\widehat{r}_j\right) \quad (j = 1, 2). \quad (13)$$

We use the C++ program language to iterate over the infinite loop estimation $\widehat{r}_j$ towards the direction of variance reduction. When variance d_j cannot be reduced or it reaches the preset threshold, the corresponding estimate $\widehat{r}_j$ which is the natural growth rate r of the population can be calculated by substituting the natural growth rate into the Equation (8) to Equation (11) where the estimated value $\widehat{k}_j$ is the corresponding environmental capacity in each year.

According to the above methods, we can obtain the natural growth rate $r_1 = 0.2412$ and $r_2 = 0.1474$ and the environmental capacity estimates between manufacturing and logistics, as shown in Figure 2.

As shown in Figure 3, the environmental capacity changes with the increase of population density. These two variables show a trend of co-evolution. However, the external environmental capacity of the population shows unstable trends and shows mutations in some years node (such as 1987, 1993, and 2000), which indicates that the node is influenced by random disturbance. According to the situation of the industry system environment and economic environment of manufacturing and logistics in Shaanxi, we find the following fundamental reality:

(1) In 1987, this year is a new stage of China's economic development, which is transformed into the direction of marketization and privatization by the former wholly public-owned and planned. Affected by this, Shaanxi successively appeared machinery, building materials, textiles, printing and dyeing, papermaking, and other manufacturing industries, and the concept of logistics began to emerge, such as transport, warehousing, distribution, circulation processing, information processing, and other logistics links in the supply chain, which are gradually accepted by the enterprise.

(2) In 1993, the CPC "The third Plenary Session of the 14th CPC Central Committee" held which indicates that Chinese economy is transitioning from a planned economy to a market economy. All kinds of private enterprises began to appear and increase year by time; all kinds of industrial parks have been planning and constructed; the manufacturing market in Shaanxi starts booming; the manufacturing demand for the external service is increasing. As a result, we are moving towards a new developing era for logistics.

(3) In 2000, Chinese government launched "West Development Strategy of China." The strategy significantly accelerated economic development in Shaanxi. After China joined the WTO in 2002, there are more opportunities for China to collaborate with developed countries in the world, which drives the rapid growth of the heavy machinery, equipment, steel, molds and other raw materials. As a result, the automotive, metal smelting, general equipment, instrumentation engineering machinery, computer, communications, and electronic equipment and other manufacturing have also been developed rapidly. However, the benefits and service quality of logistics that underpin their external services are lagging.

Moreover, at the same time node, the sudden increase in manufacturing environment capacity also brought about a sudden increase in the environmental capacity of the logistics, but the logistics population did not grow, indicating that the node was negatively affected by external stochastic disturbance factors, such as the elimination of backward enterprises by economic development.

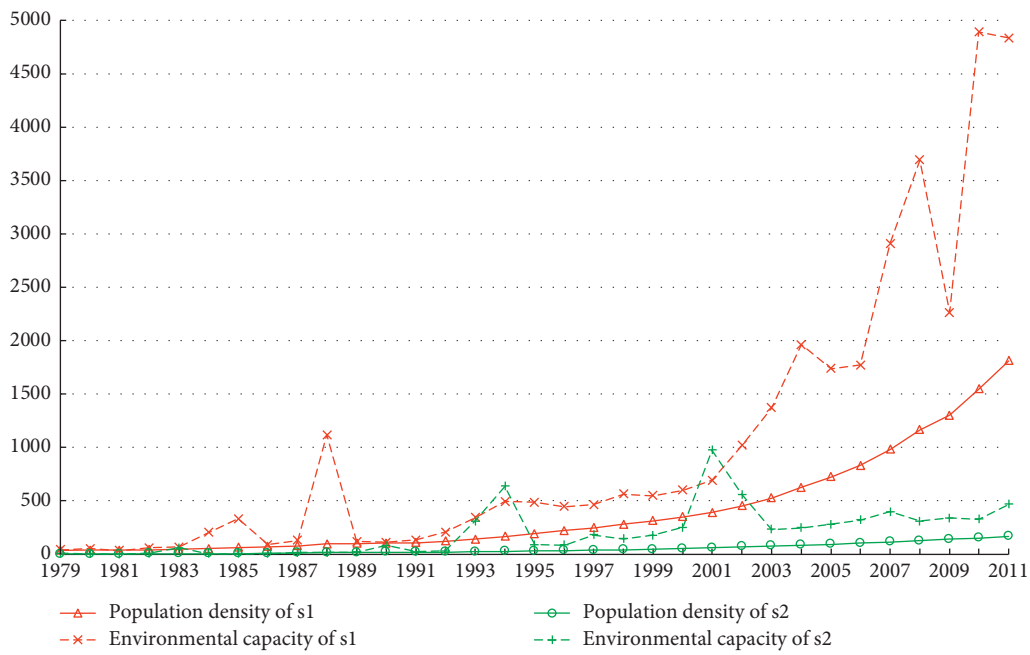


Figure 2: The relationship between S_1 and S_2.

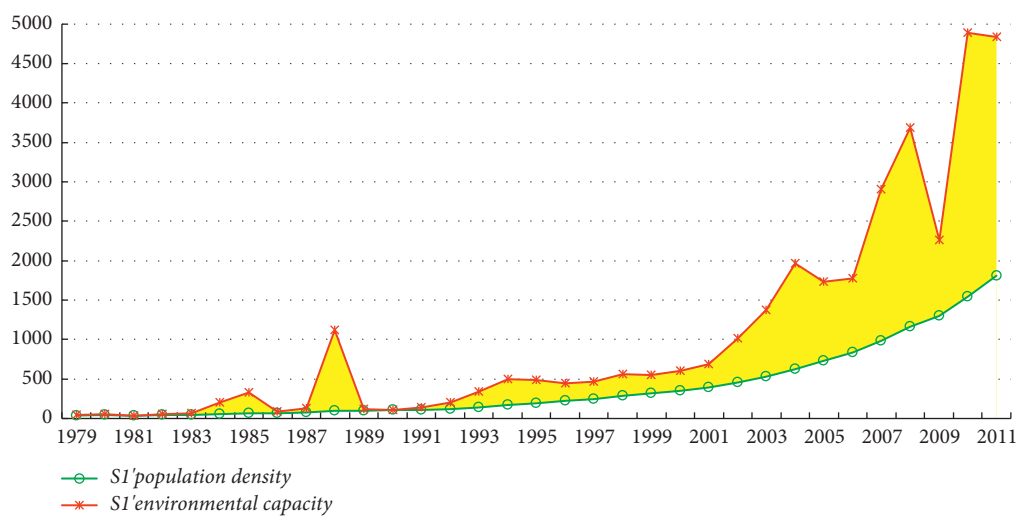


Figure 3: Relationship between K_1 and N_1.

3.3. External Environmental Capacity Estimation. Based on the above analysis, we can estimate the external environmental capacity model of coevolution between manufacturing and logistics in the following equation:

$$\begin{cases} k_1(t) = k_1^0 + \sum_{i=1}^{3} [\alpha_1(N_2)]^n + \sum_{i=1}^{3} [\beta_1(I)]^n + \sum_{i=1}^{2} [\gamma_1(E)]^n = f_1^k(N_2, I, E), \\ k_2(t) = k_2^0 + \sum_{i=1}^{3} [\alpha_2(N_1)]^n + \sum_{i=1}^{3} [\beta_2(I)]^n + \sum_{i=1}^{2} [\gamma_2(E)]^n = f_2^k(N_1, I, E). \end{cases} \tag{14}$$

We want to assure that the results of F-test, R^2, and t-test are significant. So we use progressive regression and tentative regression on $N_i(t)$, $i = 1, 2$, and k_i $(i = 1, 2)$, and get the following equations:

$$\begin{cases} K_1(t) = 25.531 - 10.877N_2(t) + 1.534\left(N_2(t)\right)^2 - 2.455\left(I(t)\right)^2 + 1.089\left(I(t)\right)^3, \\ K_2(t) = -47.844 - 0.786\left(N_1(t)\right)^2 + 12.487E(t), \end{cases} \tag{15}$$

where $t = 0, 1, 2, \ldots, T$ and $R^2 \geq 0.780$ ($F \geq 47.091$), which illustrated that the fitting effect of the equations on the sample points is good. The sig. <0.01 shows that the regression equation is well predicted.

We discuss the relationship between the population density and the environment capacity of manufacturing and logistics, respectively, as shown in Figures 3 and 4.

The figure clearly shows the trend of changes in the external environmental capacity and the population of time, and it can be seen that Shaanxi manufacturing and logistics in the current development space is very large. It is at the growing stage of the logistics curve, but the growth rate is different. To further discuss the growth rate of the two, the first derivative of (14) with respect to t, we have

$$\begin{cases} f'(K_1) = \dfrac{\mathrm{d}k_1}{\mathrm{d}t} = \dfrac{\mathrm{d}N_2}{\mathrm{d}t} + \dfrac{\mathrm{d}I}{\mathrm{d}t} = -10.877 + 3.068N_2(t) - 4.910I(t) + 3.267\left(I(t)\right)^2, \\ f'(K_2) = \dfrac{\mathrm{d}K_2}{\mathrm{d}t} = \dfrac{\mathrm{d}N_1}{\mathrm{d}t} + \dfrac{\mathrm{d}E}{\mathrm{d}t} = -1.572N_1(t) + 12.487. \end{cases} \tag{16}$$

The varying rate of ambient capacity varies between manufacturing and logistics as follows: the change rate of external environment capacity of manufacturing is a quadratic function, in which the synergy performance is a positive effect (3.068) from logistics, and the industry system environment performance is shown as a short-term positive effect, and there is a maximum value. However, the change rate of external environment capacity of logistics is characterized by a linear function, and synergy performance is a negative effect (−1.572), and the performance of the industrial system environment and the synergy of manufacturing is not obvious.

4. Results and Discussion

4.1. Manufacturing. The change of environment capacity of the manufacturing population is influenced by the synergy of its own enterprise's resource allocation, the synergy of logistics, and the industry system environment.

The K_1 equation embodies the complex process of manufacturing population evolution:

(1) The existing scale and market share of manufacturing enterprises provide huge environmental space for the growth of the manufacturing population. However, the synergy of the emerging logistics expands the environmental capacity of manufacturing and has formed a competitive relationship with it, which reduces the environmental capacity of manufacturing. From equation (16), we can find that the logistics population system has had a positive effect on the overall environmental capacity of the manufacturing population system from 1978 to 2011.

(2) The influence of the industrial system environment factor is to increase the policy to promote the development of manufacturing and expand its market space. However, due to the delay of policies on technology development, the expansion of the manufacturing market space has been suppressed. Therefore, the industrial system environment factor has a double effect. It can be found that from 1978 to 2011, industrial system environmental impact factor and the manufacturing population system environment capacity follow two quadratic curve trends. In other words, there is the maximum limit of environmental capacity.

4.2. Logistics. The change of environment capacity of logistics population is influenced by the cooperation of its own enterprise's allocation of resources, the synergy of manufacturing, and the industrial economic environment.

(1) The constant term in the K_2 equation represents the original environmental capacity of the logistics. The negative coefficient indicates that the environmental capacity of the logistics population evolution originates initially from the result of the manufacturing enterprise's allocation of resources. The manufacturing population system harms the overall environment capacity of the logistics population system during the years 1978–2011.

(2) The negative coefficient of $N_1(t)$ indicates that the synergy promoted by manufacturing to the logistics is not obvious, it originates from the inherent production and management model of the manufacturing, which does not create a broad market space for the logistics, and many logistics projects are constrained by the manufacturing, which makes the logistics development slowly.

The economic environment is a positive role, indicating the demand and contribution of the current economic development to logistics; from equation (16), we can find that the contribution of economic development along

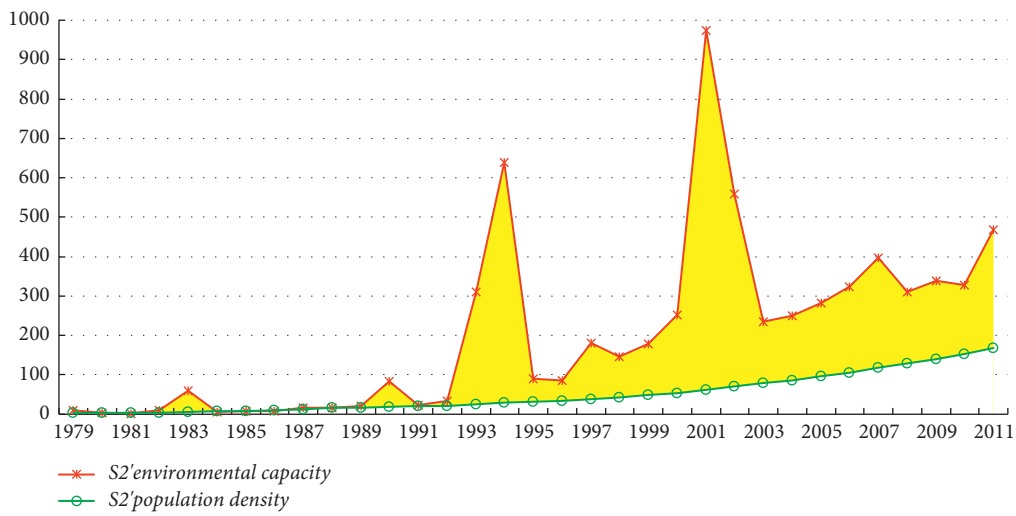


Figure 4: Relationship between K_2 and N_2.

1978–2011 to the environmental capacity of the logistics is positive.

5. Conclusions

The classical logistic equation illustrates the relationship of the individual population with time and environment, but the nonlinear function relation produced by the synergistic force between the two populations is not clear, and there is no exact expression. To solve this problem, we propose an improved logistic equation, to establish a system dynamics model of the coevolution between manufacturing and logistics and study the synergistic mechanism of external random interference factors on the coevolution system of manufacturing and logistics. The results show that the dual effects of policy system time lag and economic development demand affect the changes of manufacturing external environmental capacity and logistics ecosystem. This study reveals this internal mechanism. Through the analysis of external environmental capacity, it is further found that Shaanxi manufacturing and logistics are currently in the stage of commensalism and have not reached the ideal mature stage of mutualism. Our research extends the nonlinear environmental capacity of the LV model and its application fields. It is hoped that this study can provide enlightenment and help for follow-up research.

Data Availability

All underlying data that support the results can be found in "China Statistic Yearbook," "China Industrial Statistics Yearbook," "Shaanxi Statistic Yearbook," and "China Tertiary Industry Statistic Yearbook."

Conflicts of Interest

The authors declare that there are no conflicts of interest in the paper.

Acknowledgments

This work was supported by the Scientific Research Support Program of Xi'an University of Finance and Economics (Grant no. 20FCZD03), Ministry of Education Humanities and Social Sciences Fund Project (Grant no. 20YJAZH011), and Natural Science Basic Research Project of Shaanxi (Grant no. 2020JM-584). The authors gratefully acknowledge Professor Yunxiu Sai for his professional support at the "2018 International Joint Research Conference on Advanced Engineering Technology (JIAET 2018)," and he also provided valuable opinions on this study.

References

[1] P. A. Geroski and M. Mazzucato, "Modelling the dynamics of industry populations," *International Journal of Industrial Organization*, vol. 19, no. 7, pp. 1003–1022, 2001.

[2] X. Dong and S. G. Fisher, "Ecosystem spatial self-organization: free order for nothing?" *Ecological Complexity*, vol. 38, pp. 24–30, 2019.

[3] M. Javidi and N. Nyamoradi, "Dynamic analysis of a fractional order prey-predator interaction with harvesting," *Applied Mathematical Modelling*, vol. 37, no. 20-21, pp. 8946–8956, 2013.

[4] A. E. Matouk and I. khan, "Complex dynamics and control of a novel physical model using nonlocal fractional differential operator with singular kernel," *Journal of Advanced Research*, vol. 24, pp. 463–474, 2020.

[5] Y. Huang, F. Li, and J. Shi, "Stability of synchronized steady state solution of diffusive Lotka-Volterra predator-prey model," *Applied Mathematics Letters*, vol. 105, Article ID 106331, 2020.

[6] M. M. Amirian, I. N. Towers, Z. Jovanoski, and A. J. Irwin, "Memory and mutualism in species sustainability: a time-fractional Lotka-Volterra model with harvesting," *Heliyon*, vol. 6, no. 9, Article ID e04816, 2020.

[7] P. F. Verhulst, "Notice sur la loi que la population suit dans son accroissement," *Quetelet*, vol. 10, pp. 113–121, 1838.

[8] R. M. May, "Simple mathematical models with very complicated dynamics," *Nature*, vol. 261, no. 5560, pp. 459–467, 1976.

[9] A. Lotka, *Elements of Physical Biology*, Williams and Wilkins, Baltimore, MD, USA, 1925.

[10] V. Volterra, "Variazioni e fluttuazioni del numero di individui in specie animali conviventi," *Memoria della Reale Accademia Nazionale dei Lincei*, vol. 2, pp. 31–113, 1926.

[11] T. Zhou and W. Wang, "Study on coordination of provincial industrial eco-economic system based on Lotka-Volterra model," *Chinese Journal of Management Science*, vol. 22, no. s1, pp. 240–246, 2014.

[12] F. Meng, Z. Tian, and X. Yao, "Research on operation mechanism and evolution of digital economy ecosystem based on Lotka-Volterra model," *Journal of Hohai University (Philosophy and Social Sciences)*, vol. 22, no. 2, pp. 63–71, 2020.

[13] S. Mao, M. Zhu, X. Wang, and X. Xiao, "Grey-Lotka-Volterra model for the competition and cooperation between third-party online payment systems and online banking in China," *Applied Soft Computing*, vol. 95, Article ID 106501, 2020.

[14] W. W. Mohammed, E. S. Aly, A. E. Matouk, S. Albosaily, and E. M. Elabbasy, "An analytical study of the dynamic behavior of Lotka-Volterra based models of COVID-19," *Results in Physics*, vol. 26, Article ID 104432, 2021.

[15] C. Zhu and G. Yin, "On competitive Lotka-Volterra model in random environments," *Journal of Mathematical Analysis and Applications*, vol. 357, no. 1, pp. 154–170, 2009.

[16] D. H. Nguyen and G. Yin, "Coexistence and exclusion of stochastic competitive Lotka-Volterra models," *Journal of Differential Equations*, vol. 262, no. 3, pp. 1192–1225, 2017.

[17] L. Zu, D. Jiang, D. O'Regan, T. Hayat, and B. Ahmad, "Ergodic property of a Lotka-Volterra predator-prey model with white noise higher order perturbation under regime switching," *Applied Mathematics and Computation*, vol. 330, pp. 93–102, 2018.

[18] Y. Zhou, Y. Sai, and L. Yan, "Co-evolution mechanism of manufacturing and logistics population ecosystem driven by external environment capacity," in *Proceedings of the 2018 Joint International Advanced Engineering and Technology Research Conference (JIAET 2018)*, Xi'an China, May 2018.

Hindawi
Complexity
Volume 2021, Article ID 5624909, 7 pages
https://doi.org/10.1155/2021/5624909

Research Article

Cross-Model Transformer Method for Medical Image Synthesis

Zebin Hu,[1] **Hao Liu**,[1,2] **Zhendong Li**,[1,2] **and Zekuan Yu**[3]

[1]*School of Information Engineering, Ningxia University, Yinchuan 750021, China*
[2]*Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-founded by Ningxia Municipality and Ministry of Education, Yinchuan 750021, China*
[3]*Academy for Engineering and Technology, Fudan University, Shanghai 200433, China*

Correspondence should be addressed to Hao Liu; liuhao@nxu.edu.cn

Received 14 August 2021; Revised 25 September 2021; Accepted 7 October 2021; Published 25 October 2021

Academic Editor: Long Wang

Acquiring complementary information about tissue morphology from multimodal medical images is beneficial to clinical disease diagnosis, but it cannot be widely used due to the cost of scans. In such cases, medical image synthesis has become a popular area. Recently, generative adversarial network (GAN) models are applied to many medical image synthesis tasks and show prior performance, since they enable to capture structural details clearly. However, GAN still builds the main framework based on convolutional neural network (CNN) that exhibits a strong locality bias and spatial invariance through the use of shared weights across all positions. Therefore, the long-range dependencies have been destroyed in this processing. To address this issue, we introduce a double-scale deep learning method for cross-modal medical image synthesis. More specifically, the proposed method captures locality feature via local discriminator based on CNN and utilizes long-range dependencies to learning global feature through global discriminator based on transformer architecture. To evaluate the effectiveness of double-scale GAN, we conduct folds of experiments on the standard benchmark IXI dataset and experimental results demonstrate the effectiveness of our method.

1. Introduction

Magnetic resonance imaging (MRI) is a versatile and non-invasive imaging technique widely used in clinical applications. Tailored MRI pulse sequences enable to capture specific characteristics of the underlying anatomical information. For instance, T1-weighted brain images clearly depict the gray matter and white matter tissue, while T2-weighted images depict the fluid in the cortical tissue. Hence, acquiring complementary information about tissue morphology from multimodal images enables to improve accuracy and confidence in clinical diagnosis [1]. Unfortunately, acquiring multimodal MR imaging is often challenging due to numerous factors, such as uncooperative patients, limited availability of scanning time, and the expensive cost of prolonged exams [2, 3]. To address this issue, cross-modal medical image synthesis has been widely used, as it enables to synthesis unattained images in multimodal protocols from the subset of available images [4–7].

Currently, deep learning-based synthesis demonstrates more promising performance, which is compared with the traditional registration-based method [8, 9] and intensity-transformation-based methods [10, 11]. For the image synthesis task, convolutional neural network (CNN) architectures produce significant performance through minimizing pixelwise losses between synthetic and real images. However, pixelwise losses ignore high-level features in the training step. Since generative adversarial networks (GAN) were introduced by Goodfellow et al. [12], this problem was gradually solved by adversarial loss functions, which designed a training strategy between generator and discriminator networks based on the game theory. In this case, GAN enables to capture high-frequency texture information of medical images. Therefore, GAN-based methods surpass

many synthesis tasks based on traditional architectures [13, 14]. To be specific, the generator and discriminator networks of GAN deploy compact convolution filters, whereas CNNs are plugged with spatial locality on the entire images by the sliding window. This makes the long-range dependencies between distant regions lost [15].

Moreover, CNNs not only exhibit a strong locality bias but also a bias towards spatial invariance through the use of shared weights across all positions [16]. This prevents the networks from fully understanding the local region of the input image. To guide networks towards critical image regions, Zhao et al. [17] proposed the attention mechanisms that strengthen the features of important regions by learning the weight map and multiplying it on the feature map. However, conventional attention mechanisms still do not explicitly model long-range dependencies. Recently, transformer architectures have been applied to language tasks and are increasingly adopted in other areas such as segmentation tasks [18] and classification tasks [19]. In contrast to the predominant vision architecture, the emergent transformer architectures are integrated to learn complex relationships among its inputs, since it contains no built-in inductive prior on the locality of interactions such as sliding window. Hence, we consolidate transformer into our model due to capture more global information and make a comprehensive understanding of the input [16].

In this paper, we propose a double-scale deep learning method for cross-modal medical image synthesis. Motivated by the fact that low-level image structure and high-level feature is equally important to cross-modal medical image synthesis we integrate the ability of transformer to efficiently seek long-range interactions inside our model, which enables to capture global feature as complementary information for CNNs. To achieve this, we carefully design double-scale discriminator GAN which specifically consists of the transformer-based global discriminator and CNN-based local discriminator.

The main contributions of this paper are listed as follows. (1) We introduce a double-scale discriminator GAN for medical image synthesis. (2) The global discriminator of our model is designed on vision transformer that utilizes long-range dependencies between distant patches and captures global features.

2. Related Works

2.1. Medical Image Synthesis. Recently, GAN-based models have been successfully applied to kinds of tasks including data augmentation [20–22] and image synthesis tasks [23–25]. For example, Nie et al. [5] utilized MR images to synthesize computed tomography (CT) images with a context-aware GAN model; Wolterink et al. [7] utilized GAN to generate low-dose CT from routine-dose CT images. Nevertheless, as the traditional GAN has failed to meet the gradually higher application requirements, pix2pix [26] has recently begun to attract the attention of researchers, which utilizes paired data to enhance the pixel-to-pixel similarity between the real and the synthesized images, and then, Olut et al. [27] developed a CycleGAN-based method to synthesis MRA from T1-MRI and T2-MRI. These methods are unable to capture the features of critical image regions. Therefore, Zhao et al. [17] used a self-attention in the generator of GAN to enhance the feature of tumour and improve the performance of tumour detection. Isola et al. [26] used a patch-based discriminator to refine the extraction of features. However, these methods cannot solve the problem that the strong prior position information introduced by the sliding window in the convolution operation, which destroys the modelling of the distant dependence relationship, so that all the local information cannot be better captured.

2.2. The Transformer Architecture. The transformer architecture is designed to handle complicated interactions between inputs regardless of their relative position to one another through modelling interactions between its inputs solely through attention mechanism. Transformer is originally applied to language tasks, Floridi and Chiriatti [28] introduced GPT to use language modelling as its pretraining task. Recently, this method also can be used in computer vision. Esser et al. [16] proposed a VQGAN which represents images as a composition of perceptually rich image constituents and thereby overcomes the infeasible quadratic complexity when modelling images directly in pixel space. However, the codebook of VQGAN requires numerous datasets to fit, which is impractical in the medical image field. Meanwhile, the increased expressivity of transformers comes with quadratically increasing computational costs, because all pairwise interactions are taken into account. Finally, our method is based on a vision transformer which crops interactions between inputs based on nonoverlapping patch-level.

3. Approach

3.1. Overview of Our Method. The overview of double-scale GAN is illustrated in Figure 1. Our method is comprised of three main components: generator network, global discriminator network, and local discriminator network. In the remainder of this section, we explain the detailed composition of each network component and the loss functions.

3.2. Generator Network. The first component of our method is a deep encoder network that contains a series of convolutional layers to capture a hierarchy of localized features of source images. To learn a meaningful and effective high-level representation, we adopt an autoencoder structure as our main framework. In order to reduce the use of upsampling layer, deconvolution operation is used instead.

The detail of generator is illustrated in Figure 1. In the downsampling process, our method uses two convolutional layers of kernel size with 3 and stride with 2. In the upsampling process, our method uses two deconvolutional layers of kernel size with 3 and stride with 2. Besides, we also introduce instance normalization after each convolutional layer. After the instance normalization, the activation function ReLU is used in the encoder and decoder. For spatial and depth feature extraction, our method also adds 9 ResNet blocks between downsampling and upsampling.

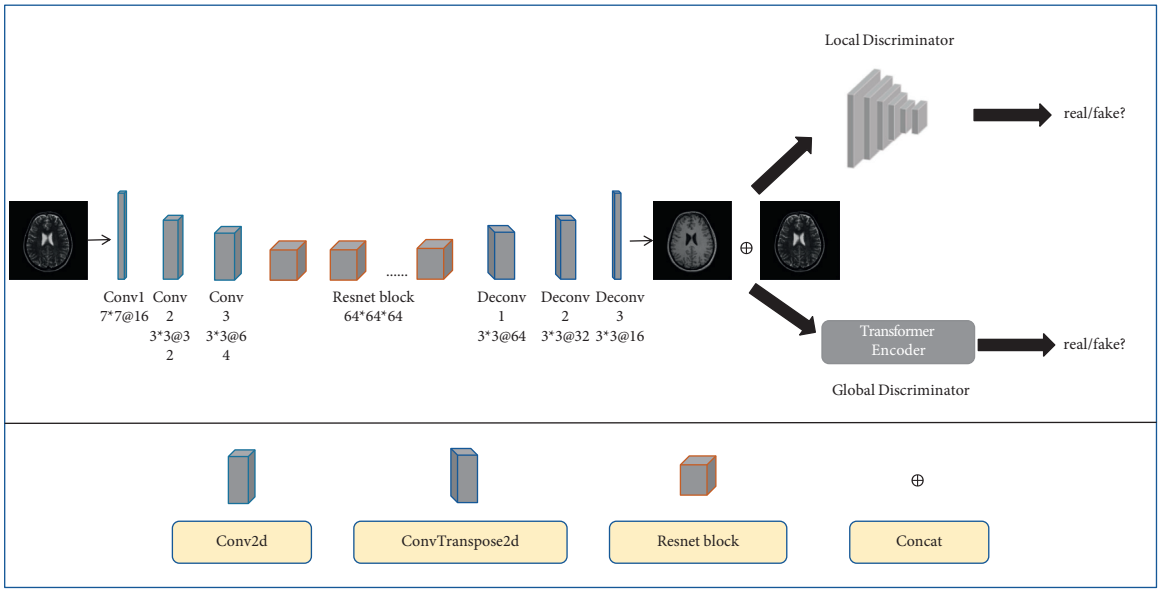


Figure 1: Schematic flow chart of the proposed algorithm for cross-modal medical image synthesis, which consists of generator, CNN-based local discriminator, and transformer-based global discriminator. The local discriminator guides the generator to learn structural representation with inductive bias. The global discriminator guides the generator to learn comprehensive features by utilizing long-range dependencies between patches of input image.

3.3. Local Discriminator Network. The local discriminator is based on a condition PatchGAN architecture [26]. It receives as input the concatenation of the source and target contrast images [29] and then obtains $30 * 30$ overlapped patches of $70 * 70$ size through sliding window for prediction to real or fake. Although this patch-based discriminant is more robust than the image-based discriminant in the extraction of local detail features, the overlapping patches it extracts destroy the long-range dependencies by introducing a strong prior position relationship, so as to have a comprehensive understanding of the input images.

3.4. Global Discriminator Network. In order to synthesize high-quality medical images, global and local features are equally important. Inspired by the DeblurGAN-v2 [30], we use a pure transformer method to replace convolutional network to capture long-range dependencies for a comprehensive understanding of the input image. The details of global discriminator network are depicted in Figure 2.

The input image is first split into $32 * 32$ nonoverlapping patches, in which kernel size is equal to stride:

$$P_1, P_2, \ldots, P_N = \text{split (input)}, \tag{1}$$

where P_i denotes the i-th patch of the input image; we set $N = 8^2$ to divide the input into 64 patches. Then, all patches are flattened to D dimension by a trainable linear projection. Similar to the class token in BERT [31], we also prepend a learnable embedding to the sequence of embedded patches. Position embeddings are added to the patch embeddings to retain positional information. Our method uses standard learnable 1D position embeddings because many studies have shown that using more advanced 2D-aware position embeddings not works [32], which can be therefore formulated as follows:

$$Z_0 = [x_{\text{class}}; P_1E; P_{2E}; \ldots; P_{\text{NE}}] + E_{\text{pos}}, \tag{2}$$

where Z_0 denotes the input of transformer encoder; E denotes embedding projection which maps patch image to vector; and E_{pos} denotes the learnable positional embedding that carries information about patch location.

The transformer encoder consists of two parts: multihead self-attention (MSA) and multilayer perceptrons (MLP). MSA enables to learn different levels of features benefit from multihead attention. In addition, layer norm (LN) is applied before every block, and residual connections after every block. At the end of these blocks, the output is taken by the classification head to output the real/fake prediction. The output of the l-th layer in the transformer encoder can be formulated as

$$Z'_l = \text{MSA}(\text{LN}(Z_{l-1})) + Z_{l-1}, \tag{3}$$

$$Z'_l = \text{MLP}(\text{LN}(Z'_l)) + Z'_l \quad , \tag{4}$$

where Z_{l-1} represents the feature extracted from the previous layer.

3.5. Loss Function. The first component of the loss function in our method is a pixelwise loss as inspired by the pix2pix architecture [26]:

$$L_1 = E_{x,y}|y - G(x)|_1, \tag{5}$$

where x denotes the source image and y denotes the target image.

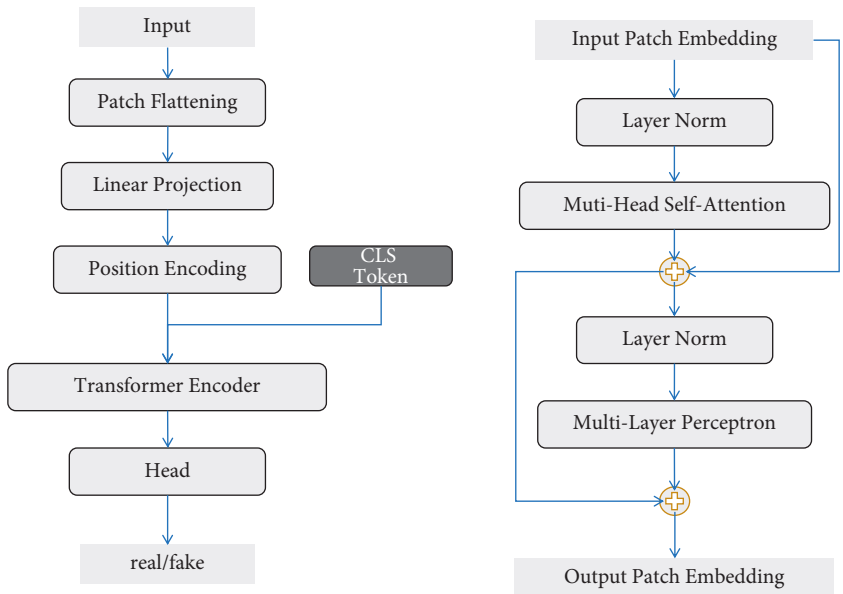


FIGURE 2: Detailed chart of global discriminator. The left side shows the overall computational flow of the global discriminator, and the right side shows the details of the transformer encoder on the left.

Unlike loss functions based on pixelwise differences, perceptual loss relies on differences in higher feature representations that are often extracted from networks pretrained for more generic tasks [33]. A commonly used network is VGGNet which trained on the ImageNet [34] dataset for object classification. Here, following [33], we extracted feature maps right before the second max-pooling operation of VGG16 pretrained on ImageNet:

$$L_{\text{per}} = E_{x,y}|V(y) - V(G(x))|_1, \tag{6}$$

where $V(\cdot)$ denotes pretrained VGG16.

The local discriminator is based on the conditional discriminator; its loss function can be formulated as

$$L_{\text{Local}}(G, D) = -E_{x,y}\left[(D(x, y) - 1)^2\right] - E_{x,z}\left[D(x, G(x, z))^2\right], \tag{7}$$

where z denotes the synthesis image from generator.

The global discriminator uses hinge loss to optimize the generator; hinge loss can be formulated as

$$\begin{aligned} L_{\text{Gocal}}(G, D) = &-E_{x,y}[\min(0, D(x, y) - 1)] \\ &- E_{x,z}[\min(0, -D(x, G(x, z)) \\ &- 1)] - \lambda_{\text{adv}} E_{x,y,z}[D(G(x, z), y)], \end{aligned} \tag{8}$$

By aggregating all the above losses, we can formulate our aggregate loss function as

$$L_{\text{aggregate}} = \lambda_{L_1} L_1 + \lambda_{\text{per}} L_{\text{per}} + \lambda_{\text{Local}} L_{\text{Local}} + \lambda_{\text{Global}} L_{\text{Global}}, \tag{9}$$

where λ_{L_1} denotes the weighing of the pixelwise loss; λ_{per} denotes the weighing of the perceptual loss; λ_{Local} denotes the weighing of the adversarial loss of local discriminator; and λ_{Global} denotes the weighing of the adversarial loss of global discriminator.

4. Experiments

In this section, we will first describe the information about the dataset used in our method and then introduce the implementation details of experiments. We present experimental results that compare with several state-of-the-art methods.

4.1. Dataset. The dataset used in the evaluation is provided by the IXI dataset. The experimental dataset we used totals 40 subjects, and each subject has corresponding T1-MRI and T2-MRI, where 30 subjects were used for training and 10 were used for testing. Acquisition parameters were as follows: T1-weighted images: TE = 4.603 ms, TR = 9.813 ms, and spatial resolution = $0.94 \times 0.94 \times 1.2\ \text{mm}^3$. T2-weighted images: TE = 100 ms, TR = 8178.34 ms, and spatial resolution = $0.94 \times 0.94 \times 1.2\ \text{mm}^3$. Since multicontrast images were unregistered, we use FSL [35] to register T1-MRI and T2-MRI. Finally, we use zero-padding to fill all images in axial cross-sections used in experiments to a consistent size of $256 * 256$.

4.2. Implementation Details. Our method is implemented in PyTorch. All methods were trained and tested on 1 NVIDIA Tesla V100 with 32 GB of memory for each GPU. In the stage of training of our method, we set the epoch as 100, learning rate as 0.0002, and batch size as 1 which causes the training time to increase to 5 hours. Model training was performed via the Adam optimizer with $\beta 1 = 0.5$ and $\beta 2 = 0.999$. In global discriminator, we use

Table 1: Comparisons of T2-weight MRI synthesis between our proposed method with different approaches of PSNR and SSIM (data in the table denote the average value and standard deviation of the test dataset).

Method	PSNR	SSIM
pix2pix	34.38 ± 0.84	0.775 ± 0.04
CycleGAN	34.75 ± 0.86	0.786 ± 0.03
PGAN (without global)	34.82 ± 0.98	0.892 ± 0.06
Ours (global and local)	34.91 ± 1.00	0.895 ± 0.07

Table 2: Comparisons of T1-weight MRI synthesis between our proposed method with different approaches of PSNR and SSIM (data in the table denote the average value and standard deviation of the test dataset).

Method	PSNR	SSIM
pix2pix	34.58 ± 0.84	0.758 ± 0.04
CycleGAN	34.73 ± 0.82	0.795 ± 0.04
PGAN (without global)	35.85 ± 1.09	0.887 ± 0.07
Ours (global and local)	35.34 ± 0.95	0.895 ± 0.07

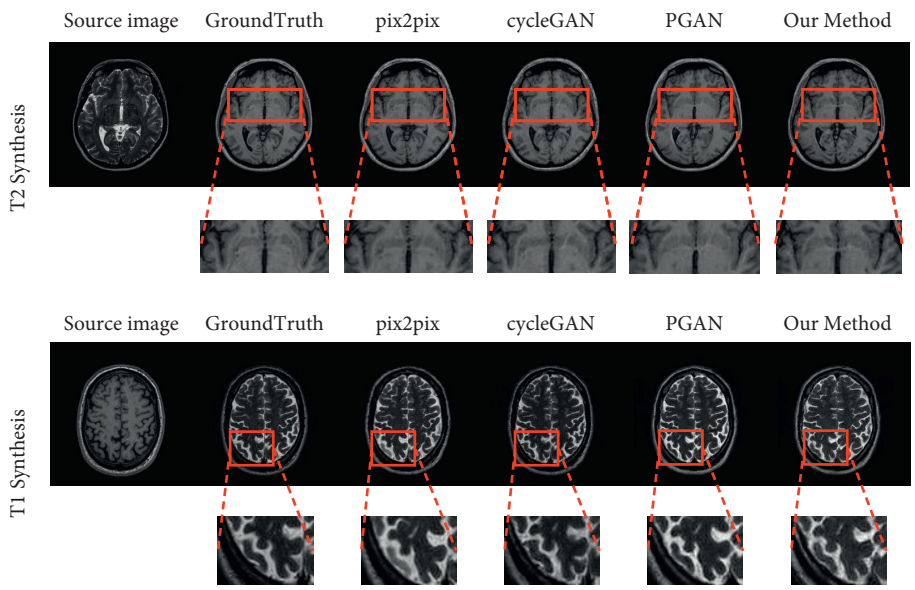


Figure 3: Synthesized images from all competing methods are shown along with the source images and the reference target image. Our method improves synthesis performance in regions that are depicted suboptimally in competing methods. Obviously, the composite images from our method have less noise and sharper tissue depiction.

multihead attention with 4 heads and set D as 64. In each multihead attention, we performed GeLu activation and set dropout as 0.1. Limited by the small size of the medical image dataset, we utilize pretrained model in global discriminator for object classification tasks on the ImageNet database. All weights were initialized using normal distribution with 0 mean and 0.02 std. We set the hyperparameter in the aggregate loss function as $\lambda_{L_1} = 1$, $\lambda_{\text{per}} = 1$, $\lambda_{\text{Local}} = 0.8$, and $\lambda_{\text{Global}} = 0.3$. For the fairness of the experiment, we designed 4-fold cross-validation by randomly sampling nonoverlapping training, validation, and testing sets in each fold.

4.3. Comparison Methods. To validate the effectiveness of the proposed synthesis method, we compare it with three state-of-the-art cross-modality synthesis methods:

(1) pix2pix [26]: this method is based on a convolutional GAN model and UNet backbone, which synthesizes the whole image by focusing on the pixelwise similarity.

(2) CycleGAN [27]: this method consists of two generators and two discriminators, which uses a cycle consistency loss to enable to train with unpaired data. In our comparison, we use the paired data to training this method and our method.

(3) PGAN [29]: this method is based on conditional GAN; its generator consists of a encoder, a decoder, and 9 ResNet blocks. Meanwhile, this method has shown superior performance in many cross-modal image synthesis tasks.

4.4. Results and Analysis. We employ two measurements to evaluate the synthesis performance of the proposed methods and our method in comparison: structural similarity index measurement (SSIM) and peak-signal-to-noise ratio (PSNR). The data in all tables are represented by the mean and standard deviation. Further details can be found in Tables 1 and 2.

To demonstrate the effectiveness of our double-scale discriminator method with regard to subjective quality, a demonstrated example is shown in Figure 3.

5. Conclusion

In this paper, we have proposed a double-scale discriminator GAN for cross-modal medical image synthesis. By compositing both CNN and transformer to design double-scale discriminator, our method has explicitly exploited the localization power of CNNs and the sensitivity of vision transformers to global context meanwhile. Experimental results have demonstrated the effectiveness of the proposed method. In the future, we will focus on the medical image generation method which integrated multiview and multimodal information through transformer, which solves the problem that 2D medical image generation cannot exploit 3D information and 3D medical image generation needs high computing power.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 61806104 and 62076142, in part by the West Light Talent Program of the Chinese Academy of Sciences under Grant XAB2018AW05, and in part by the Youth Science and Technology Talents Enrolment Projects of Ningxia under Grant TJGC2018028.

References

[1] B. J. Pichler, M. S. Judenhofer, and C. Pfannenberg, "Multimodal imaging approaches: pet/ct and pet/mri," *Molecular Imaging I*, vol. 185, pp. 109–132, 2008.

[2] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish Journal of Radiology*, vol. 80, pp. 93–106, 2015.

[3] B. B. Thukral, "Problems and preferences in pediatric imaging," *Indian Journal of Radiology and Imaging*, vol. 25, no. 4, p. 359, 2015.

[4] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 815–827, 2017.

[5] D. Nie, R. Trullo, J. Lian et al., "Medical image synthesis with context-aware generative adversarial networks," in *Proceedings of the International conference on medical image computing and computer-assisted intervention*, pp. 417–425, Springer, Quebec City, Quebec, Canada, September 2017.

[6] Y. Wang, L. Zhou, B. Yu et al., "3d auto-context-based locality adaptive multi-modality gans for pet synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 6, pp. 1328–1339, 2018.

[7] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Generative adversarial networks for noise reduction in low-dose ct," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[8] N. Burgos, M. J. Cardoso, K. Thielemans et al., "Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2332–2341, 2014.

[9] J. Lee, A. Carass, A. Jog, C. Zhao, and J. L. Prince, "Multi-atlas based ct synthesis from conventional mri with patch-based refifinement for mri-based radiotherapy planning," in *Proceedings of the Medical Imaging 2017: Image Processing*, vol. 10133, International Society for Optics and Photonics, Orlando, Florida, US, February 2017.

[10] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Mr image synthesis by contrast learning on neighborhood ensembles," *Medical Image Analysis*, vol. 24, no. 1, pp. 63–76, 2015.

[11] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical Image Analysis*, vol. 35, pp. 475–488, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, "Mustgan: multi-stream generative adversarial networks for mr image synthesis," *Medical Image Analysis*, vol. 70, Article ID 101944, 2021.

[14] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "Collagan: collaborative gan for missing image data imputation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, Long Beach, CA, USA, June 2019.

[15] T. Roughgarden, "Algorithmic game theory," *Communications of the ACM*, vol. 53, no. 7, pp. 78–86, 2010.

[16] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12 873–912 883, Nashville, TN, USA, June 2021.

[17] J. Zhao, D. Li, Z. Kassam et al., "Tripartite-gan: synthesizing liver contrast-enhanced mri to improve tumor detection," *Medical Image Analysis*, vol. 63, Article ID 101667, 2020.

[18] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: multimodal brain tumor segmentation using transformer," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 109–119, Springer, Strasbourg, France, October 2021.

[19] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3286–3295, Seoul, Korea, October 2019,.

[20] C. Bermudez, A. J. Plassard, L. T. Davis, A. T. Newton, S. M. Resnick, and B. A. Landman, "Learning implicit brain mri manifolds with deep learning," in *Proceedings of the Medical Imaging 2018: Image Processing*, vol. 10574, International Society for Optics and Photonics, Houston, Texas, US, February 2018.

[21] Z. Xu, C. Qi, and G. Xu, "Semi-supervised attention-guided cyclegan for data augmentation on medical images," in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 563–568, IEEE, San Diego, CA, USA, November 2019.

[22] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina, "Biomedical data augmentation using generative adversarial neural networks," in *Proceedings of the International conference on artificial neural networks*, pp. 626–634, Springer, Alghero, Italy, September 2017.

[23] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[24] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine-gan: semantic segmentation of multiple spinal structures," *Medical Image Analysis*, vol. 50, pp. 23–35, 2018.

[25] H. Zhao, H. Li, S. Maurer-Stroh, and L. Cheng, "Synthesizing retinal and neuronal images with generative adversarial nets," *Medical Image Analysis*, vol. 49, pp. 14–26, 2018.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, Honolulu, HI, USA, July 2017.

[27] S. Olut, Y. H. Sahin, U. Demir, and G. Unal, "Generative adversarial training for MRA image synthesis using multi-contrast MRI," in *Proceedings of the International workshop on predictive intelligence in medicine*, pp. 147–154, Springer, Granada, Spain, September 2018.

[28] L. Floridi and M. Chiriatti, "GPT-3: its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[29] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[30] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887, Seoul, Korea, October 2019.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the NAACLHLT*, Minneapolis, USA, June 2019.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16×16 words: transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2020.

[33] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European conference on computer vision*, pp. 694–711, Springer, Amsterdam, Netherlands, October 2016.

[34] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

Hindawi
Complexity
Volume 2021, Article ID 8086088, 12 pages
https://doi.org/10.1155/2021/8086088

Research Article

Fuzzy Wavelet Neural Network with the Improved Levenberg–Marquardt Algorithm for the AC Servo System

Run-Min Hou, Di-Fen Shi, Qiang Gao, and Yuan-Long Hou

School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Correspondence should be addressed to Run-Min Hou; 187189579@qq.com

Received 13 August 2021; Revised 13 September 2021; Accepted 11 October 2021; Published 23 October 2021

Academic Editor: Long Wang

In this study, a fuzzy wavelet neural network with the improved Levenberg–Marquardt algorithm (FWNN-LM) is proposed to conquer nonlinearity and uncertain disturbance problems in the AC servo system. First of all, use the particle swarm optimization algorithm based on Levenberg–Marquardt (LM) to optimize parameters in the FWNN controller. Second, the potentiality of fuzzy rules (PFR) method is developed to optimize the structure of the FWNN by error reduction ratio (ERR). Furthermore, stability of FWNN-LM is proved by the Lyapunov method. Finally, simulation and prototype test results show that this method can improve the accuracy and robustness of the system in presence of load disturbances and parameter perturbations.

1. Introduction

In recent years, varies studies show that the AC servo system exhibits good dynamical property [1, 2], but the stability still needs to improve. For the AC servo system, the dynamic mathematical model is a complex system with characteristics of large load, which can lead to nonlinearity and uncertain disturbance. In practical applications, an AC servo system performance may be affected due to unmodeled dynamics changed [3, 4].

After referring to many references, a lot of studies have shown the neural network is an important component of a complicated nonlinear system control policy under the circumstance of the lack of full model details [5–8]. The most prominent advantage of the neural network is approximate capability, and it can approximate function with any precision. However, it is hard to avoid local minimization for the BP neural network. If we use a sigmoid function as the stimulation function, it also causes slow convergence speed. Moreover, it cannot realize the mapping rules in time [9]. Fuzzy logic has become a hot topic research of neural networks in many studies. Dong et al. [10, 11] provide theoretical basis to modelling and controlling the nonlinear system. Consider there are many uncertainties existing in the fuzzy control process. Wang [12] provides a fuzzy neural network along with utilization to improve system robustness without accurate control; however, the parametric learning algorithm is presupposed for the topology of fuzzy systems.

As an alternative, multiple research studies concentrate on the use of the wavelet neural network (WNN) [13–16]. Compared with the usual sigmoid function neural network, the wavelet function possesses a better learning capacity in aspects of system identification. In recent years, Zekri et al. [17–19] studied the combination of wavelet theory and the fuzzy neural network (FNN). In the FWNN, fuzzy rules are corresponding to the sub-WNN, respectively, and the wavelet and fuzzy sets parameters learning can improve the FWNN approximation accuracy [20–23]. However, the main drawback of the WNN is that due to its feed-forward network structure, its application area is limited to static issues. The Levenberg–Marquardt (LM) method has a remarkable characteristic of local learning and a fast convergence performance at the same time [24]. However, the LM algorithm increases memory demands with the method of calculating some problems that come from the error function with the Jacobian matrix [25]. Moreover, another disadvantage is that the LM algorithm is still a local optimization method.

The particle swarm algorithm (PSO) is a global optimization algorithm, through collaboration and competition

between individuals to find the optimal solution, and the particle swarm optimization search process is started from the entire group, with the implicit parallel search features to improve the performance of the algorithm [22]. However, the PSO algorithm has some disadvantages such as slow convergence speed.

Based on the above analysis, in this study, an adaptive fuzzy wavelet neural network controller with LM is proposed to control the rotor position of the AC servo system for tracing reference trajectory with robustness. In the proposed control structure, the FWNN is a controller, and the LMPSO algorithm is employed for the online training of all weights of the FWNN. Moreover, potentiality of fuzzy rules (PFR) with using error reduction ratio (ERR) is developed to adjust the parameters and organize the structure of the FWNN. The stability of the system can be proved by using Lyapunov theory [26]. Finally, studies demonstrate promising results of a prototype AC servo system that can verify the feasibility and effectiveness by using the proposed algorithm.

The contents of this study can be listed as follows: the second section analyzes the servo system. After briefly introducing the FWNN in the third section, the following section four develops the FWNN-LM which has been proposed at great length. Afterwards, the convergence of the algorithm is analyzed in section five. And then, the simulation outcomes are discussed in section six. Last, a conclusion has been mentioned in the last section.

2. AC Servo System Analysis

The AC servo system control structure is shown in Figure 1.

In the stationary (d-q) frame of reference, the mathematical models of the permanent magnet synchronous motor can be expressed as follows:

$$\begin{cases} \dot{i}_d = -\dfrac{R}{L_d} i_d + \dfrac{L_q}{L_d} p i_q \omega_r + \dfrac{u_d}{L_d}, \\ \dot{i}_q = -\dfrac{R}{L_q} i_q + \dfrac{L_d}{L_q} p i_d \omega_r - \dfrac{\psi_f}{L_q} p \omega_r + \dfrac{u_q}{L_q}, \\ \dot{\omega}_r = \dfrac{1}{J}\left(T_e - T_L - B\omega_r\right), \end{cases} \tag{1}$$

where i_d, i_q, u_d, u_q, and L_d, L_q represent the electric currents, voltages, and inductance coefficient of the motor d and q axes, respectively; R is the motor stator resistor (Ohm), ψ_f represents the motor permanent magnet flux, p represents the motor pair of poles, J represents the motor inertia constant, B represents the viscous friction coefficient, ω_r stands for the motor angular velocity, T_e stands for motor electromagnetic torque, and T_L stands for the load torque.

The system is applied to a three-closed-loop control system. It uses the magnetic field-oriented control technology to complete the motor position and achieve high performance. Additionally, the simplification of the motor control system uses the $i_d = 0$ vector control approach. When $i_d = 0$, the motor mechanical equation can be expressed as

$$J\dot{\omega}_b + B\omega_b + T_L = T_e, \tag{2}$$

where ω_b is the mechanical angular velocity, and T_e can be written as

$$T_e = \frac{3}{2} p \psi_f i_q = K_t i_q, \tag{3}$$

where K_t is a moment constant that needs to be adjusted.

Generally, compared with the mechanical time constant, the motor current time constant has a much smaller numerical value; thus, the delay time of the current responding can be neglected. The state variables can be set as $x_1 = \theta$ and $x_2 = \dot{\theta}$; substitute equation (2) into equation (3), and the AC servo system can be rewritten as

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -\dfrac{B}{J} x_2(t) + \dfrac{K_t}{J} i_q(t) + \left(-\dfrac{1}{J} T_L\right), \end{cases} \tag{4}$$

where $-(B/J)$, K_t/J, and $-(1/J)T_L$ represent the nonlinear dynamic equations and the external disturbance, respectively.

3. Fuzzy Wavelet Neural Network

3.1. Wavelet Neural Network Structure. The structure of wavelet neural network is shown in Figure 2. As Figure 2 illustrates, K is the master nodes of the input layer, the hidden layer number is M, ω_{km} is the connection weighing between node k of the input layer and node m of the hidden layer, ω_m is the connection weighing between node m and the output layer, b_m is the translation parameter of wavelet function, and a_m is the scale variable of the wavelet function. The output can be written as [21]

$$\mathbf{L}(t) = \sum_{m=1}^{M} \boldsymbol{\omega}_m \boldsymbol{\psi}_m(\mathrm{net}_m), \tag{5}$$

where $\mathrm{net}_m = \sum_{k=1}^{K} \omega_{km} x_k - b_m/a_m$, $(m = 1, 2, \ldots, M)$. Choose the Morlet wavelet function as the generating function $\psi(x) = \cos(1.75x)\exp(-(1/2)x^2)$.

3.2. Fuzzy Wavelet Neural Network Structure. In the fuzzy wavelet network, each fuzzy rule corresponds to a given wavelet scale values of the wavelet neural network [27]. In order to describe FWNN-LMPSO clearly, a simple structure of the FWNN is shown in Figure 3.

The N_F fuzzy IF-THEN rules can be expressed as follows:

R_n: if x_1 is A_{1n},x_2 is A_{2n}, ..., and x_m is A_{mn}, then

$$\mathbf{y}_n = \mathbf{L}_n = \sum_{m=1}^{N_{w,n}} \boldsymbol{\omega}_{m,n} \boldsymbol{\psi}_{m,n}(x), \tag{6}$$

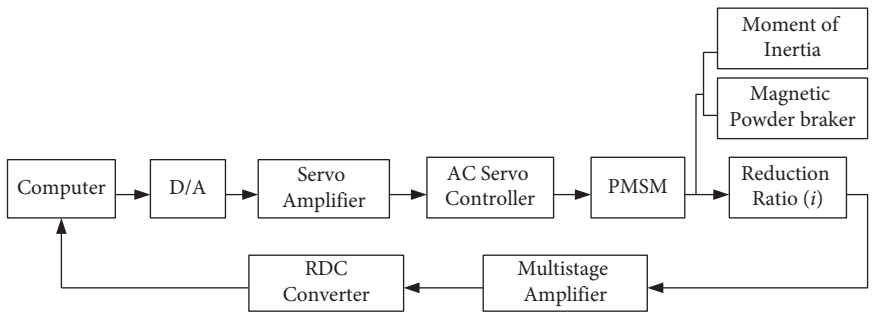


Figure 1: The structure of the AC servo system.

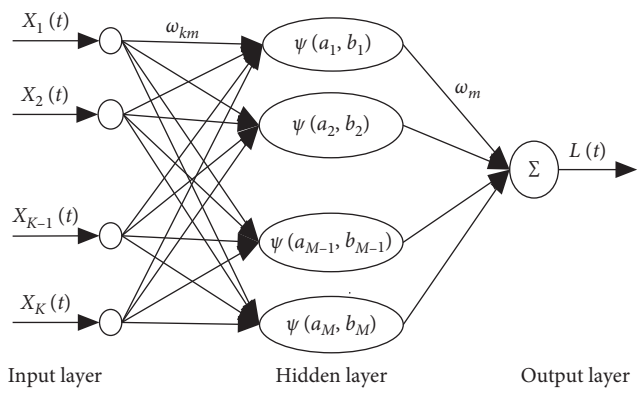


Figure 2: The structure of the wavelet neural network.

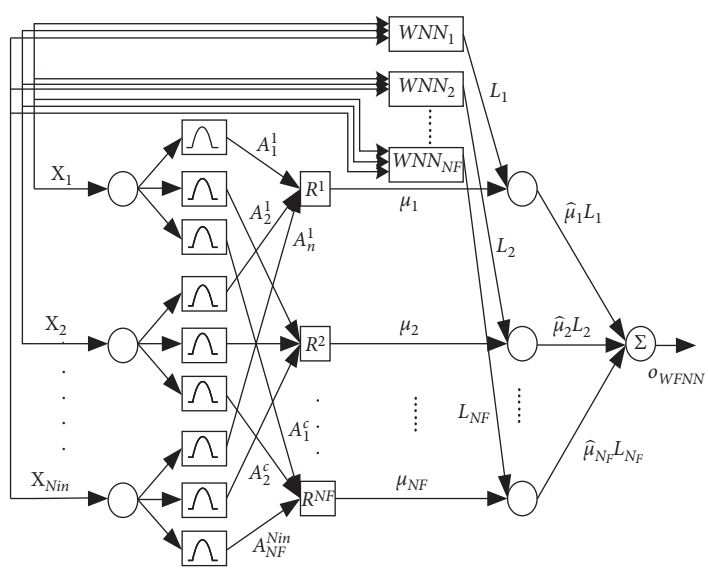


Figure 3: The structure of the FWNN.

where R_n is the fuzzy rule $(1 \le n \le N_F)$; A_{mn} is the membership function for the fuzzy set of Gaussian function, which can be expressed as

$$\mu_{A_{mn}}(x_m) = \exp^{\left(-(x_m - c_{mn})^2/\sigma_{mn}^2\right)}, \tag{7}$$

where x_m is the input of $m = 1{:}\ N_{in}$, N_{in} represents the number of input neurons; $n = 1{:}\ N_F$. The canter c_{mn} and width σ_{mn} can be used to define as a subordinate function.

The output of the entire FWNN structure by using product rules and defuzzification is shown as

$$\mathbf{O}_{\mathrm{FWNN}}(k) = \sum_{n=1}^{N_F} \hat{\mu}_n(x)\mathbf{L}_n, \tag{8}$$

where $\hat{\mu}_n(x) = (\mu_n(x)/\sum_{n=1}^{N_F}\mu_n(x))$, $\mathbf{L}_n = \sum_{j=1}^{N_{W(n)}}\omega_j\psi_j$, and $\mu_n(x) = \prod_m \mu_{A_{mn}}(x_m)$.

The output mean square error of the online learning is

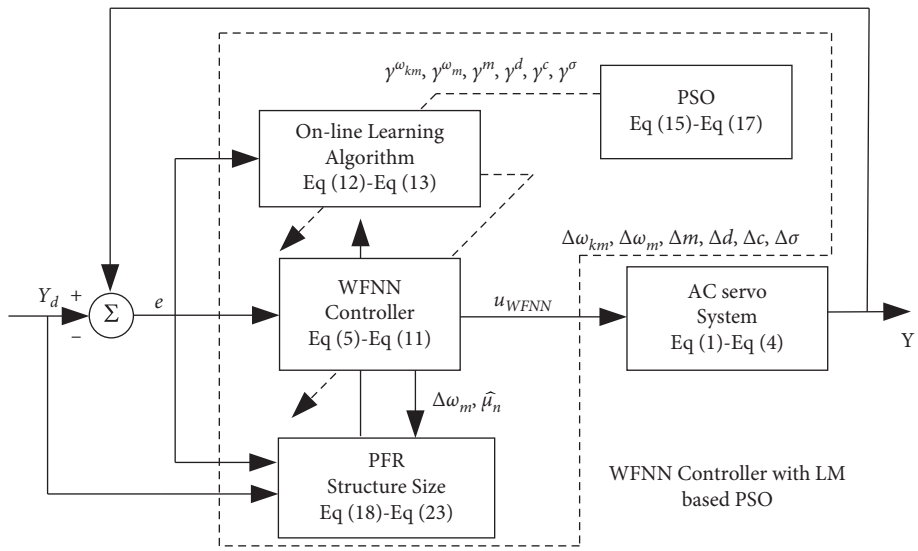


Figure 4: FWNN with LM-based PSO.

$$E = \frac{1}{2}\left[\left(O_d(k) - O_{\text{FWNN}}(k)\right)^2\right], \tag{9}$$

where O_d is the expected output of the training data. According to the descent algorithm, the FWNN parameters adjustment formulas are shown as

$$\omega_{km}^* = \omega_{km} - \gamma^1 \frac{\partial E}{\partial \omega_{km}};\ \omega_m^* = \omega_m - \gamma^2 \frac{\partial E}{\partial \omega_m};\ a_m^* = a_m - \gamma^3 \frac{\partial E}{\partial a_m},$$
$$b_m^* = b_m - \gamma^4 \frac{\partial E}{\partial b_m};\ c_{mn}^* = c_{mn} - \gamma^5 \frac{\partial E}{\partial c_{mn}};\ \sigma_{mn}^* = \sigma_{mn} - \gamma^6 \frac{\partial E}{\partial \sigma_{mn}}, \tag{10}$$

where γ^l $(l = 1{:}\ 6)$ is the learning rate, and the arguments of the FWNN controller can be expressed as

$$\boldsymbol{\gamma}^l = \left[\gamma^1, \ldots, \gamma^2, \ldots, \gamma^l\right] = \left[\gamma^{\omega_{km}}, \gamma^{\omega_m}, \gamma^m, \gamma^d, \gamma^c, \gamma^\sigma\right]. \tag{11}$$

4. FWNN with LM Algorithm-Based PSO

In the neural network, a sigmoid function is used as the activation function of the BP neural network, which leads to the result that the BP neural network is easy to get into a local minimum, slow convergence speed. Thus, to improve the performance of FWNN, the training process is required to adjust both the structure size and the parameters. An LMPSO method is used for adjusting the parameters; and a PFR is developed to design the structure of FWNNs. In the following, the LMPSO method and PFR method are described in detail.

4.1. LM Algorithm-Modified FWNN. The LM algorithm is an approximate Newton algorithm, which proves that the LM-based BPNN algorithm converges quick and accurate performance [28]. In this study, the total mean square error of $\mathbf{P}$ is given as

$$\mathbf{P} = \frac{1}{2}\sum_{p=1}^{P} e_p^2. \tag{12}$$

The LM algorithm is written as

$$\Delta\mathbf{h} = \left[\mathbf{J}^T(h)J(h) + \lambda I\right]^{-1}\mathbf{J}^T(h)\bar{\mathbf{e}}(h), \tag{13}$$

where the Jacobian matrix $\mathbf{J}(h)$is

$$\mathbf{J}(h) = \begin{bmatrix} \frac{\partial e_1(h)}{\partial h_1} & \frac{\partial e_2(h)}{\partial h_2} & \cdots & \frac{\partial e_1(h)}{\partial h_I} \\ \frac{\partial e_1(h)}{\partial h_1} & \frac{\partial e_2(h)}{\partial h_2} & \cdots & \frac{\partial e_2(h)}{\partial h_I} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial e_P(h)}{\partial h_1} & \frac{\partial e_P(h)}{\partial h_2} & \cdots & \frac{\partial e_P(h)}{\partial h_I} \end{bmatrix},$$
$$\bar{\mathbf{e}}(h) = \left[e_1, e_2, \ldots, e_P\right]^T. \tag{14}$$

The structure of the fuzzy wavelet neural network (FWNN) based on the LMPSO controller is shown in Figure 4.

4.2. LM Algorithm-Based PSO. PSO is a population-based heuristic global optimization technique. In this algorithm, the population is called a swarm, and the trajectory of each particle in the search space is adjusted by dynamically altering its velocity, according to its own flying experience and swarm experience in the search space. In the PSO algorithm, a group of particles represent a candidate solution. The velocity and position updating formulas of the PSO are illustrated as

$$\begin{cases} v_i(k+1) = w(k)v_i(k) + c_1 r_1 \left(P_i(k) - x_i(k)\right) + c_2 r_2 \left(P_g(k) - x_i(k)\right), \\ x_i(k+1) = x_i(k) + v_i(k+1), \end{cases} \tag{15}$$

where $v_i(k)$ represents the current rate of particle i^{th} during the iteration k; $x_i(k)$ represents the current position of the ith particle; $P_i(k)$ is on behalf of the optimum position of the ith particle previously appeared; $P_g(k)$ denotes the best previous position among all the particles; c_1 and c_2 are on behalf of the acceleration factors; r_1 and r_2 uniformed random number in the interval $[0, 1]$; $w$stands for the inertia weight in the interval $[0.4, 0.9]$. An appropriate fitness function to calculate the appropriate value is

$$F(t) = 0.4e_{\mathrm{MAE}} + 0.6e_{\max}, \tag{16}$$

where $F(t)$ is the fitness value; e_{MAE} represents the mean absolute error, $e_{\max}$ is the maximum absolute error, and connected with LM-based FWNN, there is

$$e_{\mathrm{MAE}} = \frac{\sum_{x=1}^{X} |e_x|}{X}; \; e_{\max} = \max(|\overline{e}(h)|); \; x_i = h. \tag{17}$$

In this study, the PSO-LM algorithm is used to make adjustments of FWNN which can make e_{MAE} and $e_{\max}$ more appropriate in actual conditions. A few particulars about the PSO procedure are shown as follows:

(1) Initialize the PSO parameters

(2) Determine $P_g(k)$ and $P_i(k)$

(3) Refresh the particle speed as well as the position by taking advantage of equation (29)

(4) Get the current fitness value FT and update$P_i(k)$, $P_g(k)$; if $\mathrm{FT}_i < P_i(k)$, $\mathrm{FT}_i < P_g(k)$.

(5) If$i < N$, set $i = i + 1$ and then go to step (3); otherwise, proceed to the next step.

(6) If $k <$ maxgen, set $k = k + 1$ and then go to step (2) or output $P_g(k)$ to LM-based FWNN.

4.3. Potentiality of Fuzzy Rules (PFR). The PFR values can be used to calculate the potentiality of fuzzy rules and extract the contributions of the normalized neuron. The FWNN model is expressed as follows:

$$\mathbf{y}_d(t) = \mathbf{W}(t)\mathbf{\Phi}(t) + \mathbf{e}(t), \tag{18}$$

where $\mathbf{W}(t) = [\mathbf{w}(t-k-1), \mathbf{w}(t-k+2), \ldots, \mathbf{w}(t)]^{\mathrm{T}}$ is the weight between the output layer and normalized layer, and $\Phi(t)$ is given by

$$\mathbf{\Phi}(t) = [\widehat{\mu}(t-k+1), \widehat{\mu}(t-k+2), \ldots, \widehat{\mu}(t)]. \tag{19}$$

The matrix $\Phi(t)$ can be transformed into a set of orthogonal basis vectors by QR decomposition as

$$\mathbf{\Phi}^{\mathrm{T}} = \mathbf{Q}(t)\mathbf{R}(t), \tag{20}$$

where $R(t)$ is an upper triangular matrix, and $\mathbf{Q}(t) = [q_1(t), q_1(t), \ldots, q_n(t)]$ have the same dimension as $\Phi(t)$. Then, the ERR is given by [29]

$$\mathrm{error}(t) = \frac{\left(\mathbf{y}_d(t)\mathbf{q}_l^{\mathrm{T}}(t)\right)^2}{\mathbf{q}_l^{\mathrm{T}}(t)\mathbf{q}_l(t)\mathbf{y}_d(t)\mathbf{y}_d^{\mathrm{T}}(t)}, \quad l = 1, 2, \ldots, N_F. \tag{21}$$

The PFR value of the l_{th} normalized neuron can be expressed as follows:

$$\mathrm{PFR}_l(t) = \frac{R_l(t)}{\sum_{l=1}^{N_F} R_l(t)}, \quad l = 1, 2, \ldots, N_F, \tag{22}$$

where $\mathrm{PFR}_l(t) \in (0, 1)$ is the potentiality of fuzzy rule in the lth normalized neuron, and

$$R_l(t+1) = \eta R_l(t) + \widehat{\mu}(t-\rho+1)\mathrm{error}(t), \quad \rho = k, k-1, \ldots, 1, \tag{23}$$

where $0 < \eta < 1$ is a constant.

5. Stability Analysis

Lyapunov function is used to assess the system stability, and it can be defined as

$$V(k) = \frac{1}{2}e^2(k), \tag{24}$$

where $e(k) = (y_d(k) - y(k)).y_d$ is the desired output, and $y(k)$ is the actual output.

$$\Delta V(k) = \frac{1}{2}\left(e^2(k+1) - e^2(k)\right), \tag{25}$$

where $e(k+1) = e(k) + \Delta e(k)$. Using Taylor's formula, $\Delta e(k)$ can be given as

$$\Delta e(k) = \sum_{n=1}^{N_F} \left\{ \left[\frac{\partial e(k)}{\partial \mathbf{P}_n^l(k)} \right]^T \Delta \mathbf{P}_n^l(k) \right\}, \tag{26}$$

where

$$\begin{cases} \dfrac{\partial e(k)}{\partial P_n^l(k)} = -A\dfrac{\mu_n(x)}{\sum_{h=1}^{N_F}\mu_h(x)}\dfrac{\partial L_n}{\partial P_n^l(k)}, \\ \Delta P_n^l(k) = \gamma^l Ae(k)\times\left(\dfrac{\mu_n(x)}{\sum_{h=1}^{N_F}\mu_h(x)}\right)\dfrac{\partial L_n}{\partial P_n^l(k)}, \\ \mathbf{P}_n = \left[P_n^1,\ldots,P_n^2,\ldots,P_n^l\right] = \left[\omega_{Km,(n)},\omega_{m,(n)}m_{m,(n)},d_{m,(n)},c_{m,(n)},\sigma_{m,(n)}\right], \\ n = 1{:}\ N_F, l = 1{:}\ 6, j = 1{:}\ m, k = 1{:}\ N_{\text{in}}, \end{cases} \tag{27}$$

where $A = (\partial y(k)/\partial u(k))$. Then, $\Delta e(k)$ can be written as

$$\Delta e(k) = \gamma^l A^2 e_1(k)\frac{1}{\left(\sum_{h=1}^{N_F}\boldsymbol{\mu}_h(x)\right)^2}\sum_{n=1}^{N_F}\boldsymbol{\mu}_n^2(x)\left\|\frac{\partial \mathbf{L}_n}{\partial \mathbf{P}_n^l(k)}\right\|^2. \tag{28}$$

Substituting equation (28) into equation (25), $\Delta V(k)$ can be rewritten as

$$\Delta V(k) = -\frac{1}{\left(\sum_{h=1}^{N_F}\boldsymbol{\mu}_h(x)\right)^2}\times\sum_{n=1}^{N_F}\left(\boldsymbol{\mu}_n^2(x)\left\|\frac{\partial \mathbf{L}_n}{\partial \mathbf{P}_n^l(k)}\right\|\right)^2 Ae_1(k)\lambda, \tag{29}$$

where

$$\lambda = \gamma^l\left[1-\frac{1}{2}\frac{\gamma^l}{\left(\sum_{h=1}^{N_F}\boldsymbol{\mu}_h(x)\right)^2}\times\sum_{n=1}^{N_F}\left(\boldsymbol{\mu}_n^2(x)\left\|\frac{\partial \mathbf{L}_n}{\partial \mathbf{P}_n^l(k)}\right\|^2\right)\right]. \tag{30}$$

If $\mu_n(x)\le 1$, that is,

$$\lambda \ge \gamma^l\cdot\left[1-\frac{1}{2}\frac{\gamma^l}{\left(\sum_{h=1}^{N_F}\boldsymbol{\mu}_h(x)\right)^2}\times\sum_{n=1}^{N_F}\left(\left\{\max\left(\boldsymbol{\mu}_n^2(x)\right)\right\}^2\left\{\max\left\|\frac{\partial \mathbf{L}_n}{\partial \mathbf{P}_n^l(k)}\right\|\right\}^2\right)\right]. \tag{31}$$

Thus, according to Lyapunov stability theory, when $\lambda > 0$, $\Delta V(k) < 0$, and the stability of the system will be guaranteed.

6. Simulation Test and Discussion

In order to test and verify the effectiveness of the FWNN-LMPSO control, it will be compared with FWNN-LM.

6.1. Simulation Experiment. In addition, all simulation programs are conducted in Matlab/Simulink, and a clock rate of 2.6 GHz and 4 GB of RAM on a PC running in a Microsoft 7.0 environment are selected. The main parameters of the AC servo system are given in Table 1. Figures 5–9 show the simulation results.

As shown in Figure 5, the moment of inertia changes from the initial value to 1.5 times. FWNN-LM generates an overshoot; it takes 4.15 s to reach the stable condition. Using FWNN-LMPSO control, the system responds quickly, and only needs 1.6 s to reach the steady state without overshoot.

Figure 6 shows step response when a 360 nm disturbance added at 3 s.

As Figure 6 shows, when the load added, it gets more deviation results on account of the response of the algorithm of FWNN-LM control. It also has a 5.15° delay in tracking the reference position. However, when using the FWNN-LMIPSO control algorithm, the offset can decrease to 1.25°. It costs 0.35 s to reach the target position. Above all, the system can perform better in the aspect of load disturbance suppressing.

Experimental result of tracking step signal with random disturbances is shown in Figure 7. It can be seen from Figure 7 that when adding random disturbance to the response signal, there is no offset occurring by using FWNN-LMPSO control. Moreover, random disturbance is also added in sinusoidal tracking experiment. The maximum error of FWNN-LM and FWNN-LMPSO is 0.089° and 0.057°, respectively.

The sinusoidal tracking error curves with a frequency of 1.67 rad/s and amplitude of 30 degree is shown in Figure 8.

In Figure 9, the number of FWNN-LM iterations is about 220 steps, the training error is 0.128, and the training error of FWNN-LMPSO is 0.035 when the number of iterations is about 95 steps. Therefore, the convergence rate of FWNN-LMPSO is better than the FWNN-LM method.

6.2. Semiphysical Experiment. The semiphysical experiment platform structure is shown in Figure 10. A step response with FWNN-LM control and FWNN-LMPSO control are

Table 1: The main parameters of the AC servo system.

System parameters	S	Value	Unit
Converted to the motor output shaft moment of inertia	J	5239	$Kg.m^2$
Converted to unbalanced torque and friction torque of the motor output shaft	T_L	9.32×10^3	$N \cdot m$
Electromagnetic torque coefficient	K_t	0.195	$N \cdot m/A$
Viscous friction coefficient	B	1.43×10^{-4}	$N \cdot m/(rad.s^{-1})$
Reduction ratio	i	1039	

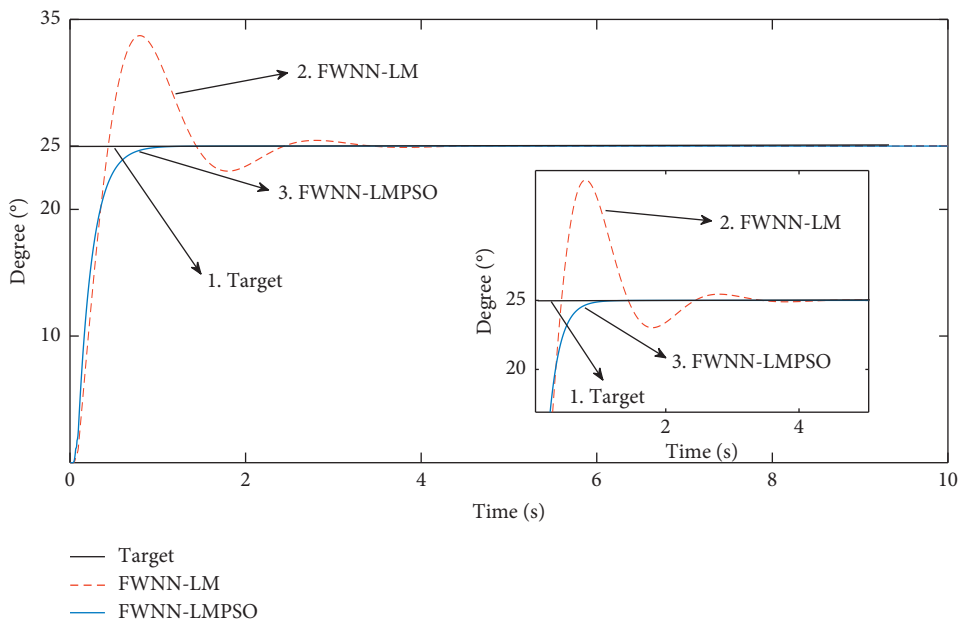


Figure 5: The moment of inertia changes the step response curve.

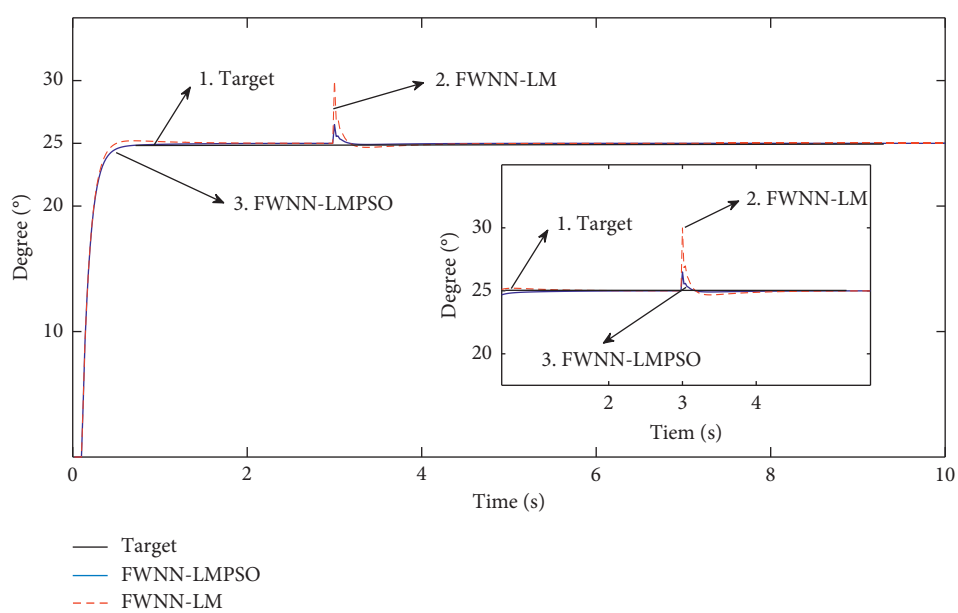


Figure 6: Step response curve of load disturbance.

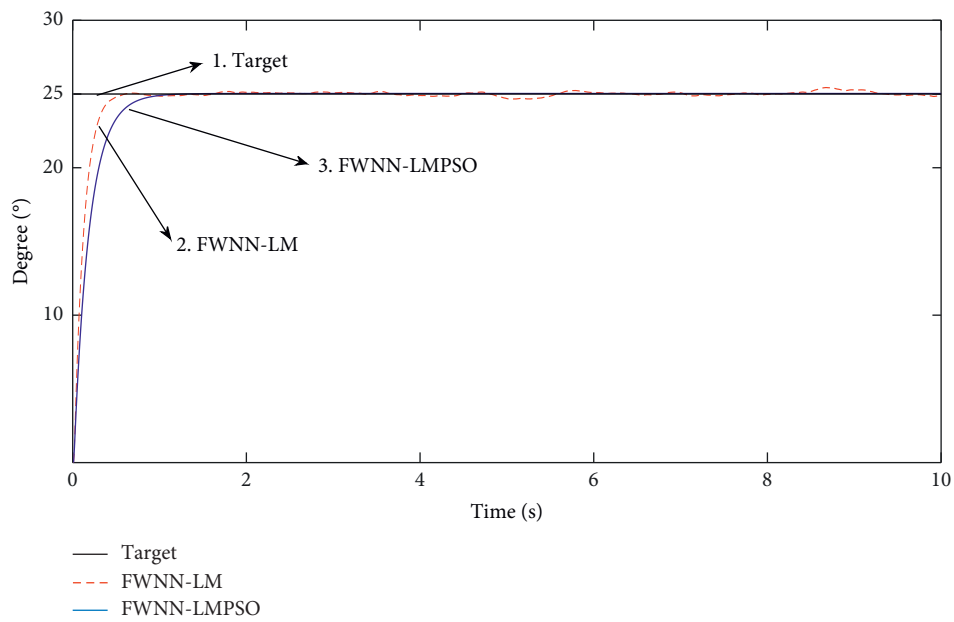


Figure 7: System dynamic response curves.

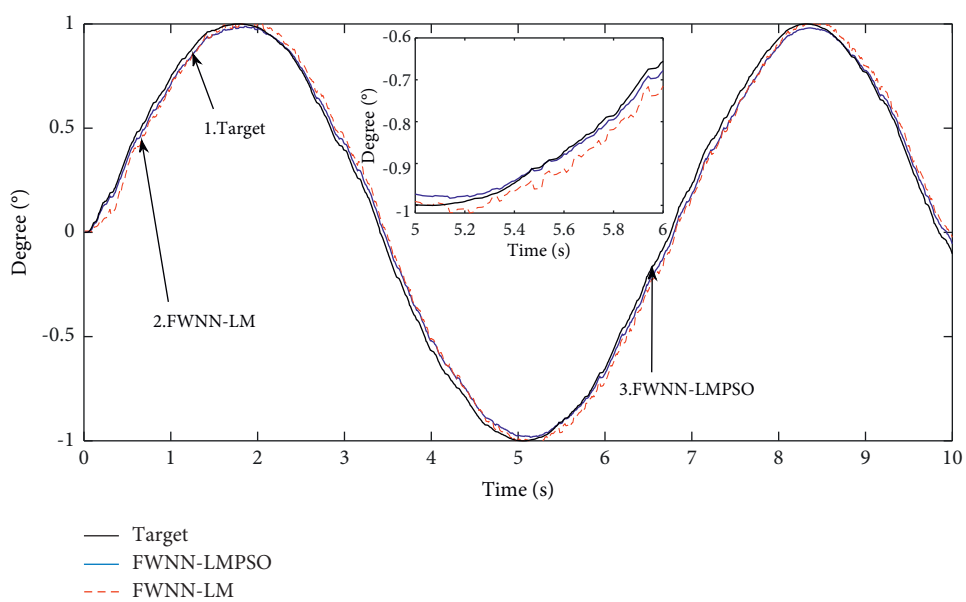


Figure 8: Sinusoidal tracking error curves with a neural network controller.

conducted on this semiphysical experiment platform to test the system performance.

According to Figures 11–14, when the system is under maximum load, its steady state time takes 1.41 s for the FWNN-LM control and the maximum steady state error is 2.63°; however, for FWNN-LMPSO control, the system required steady state time is 1.35 s, and maximum steady state error is 0.749°.

Compared with FWNN-LM control, the FWNN-LMSPSO control has better dynamic and steady state performance. In addition, it performs well in improving system antidisturbance performance.

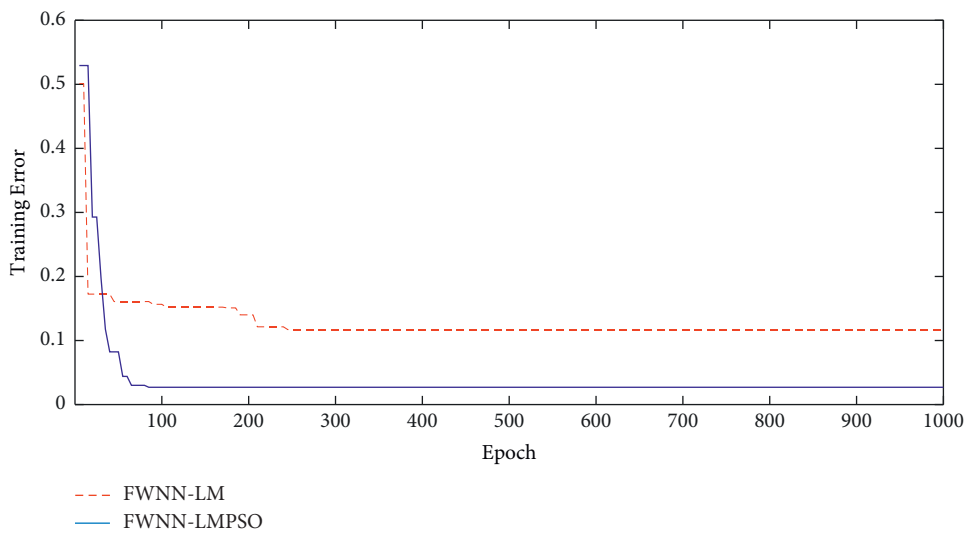


Figure 9: RMSE values in the training process.

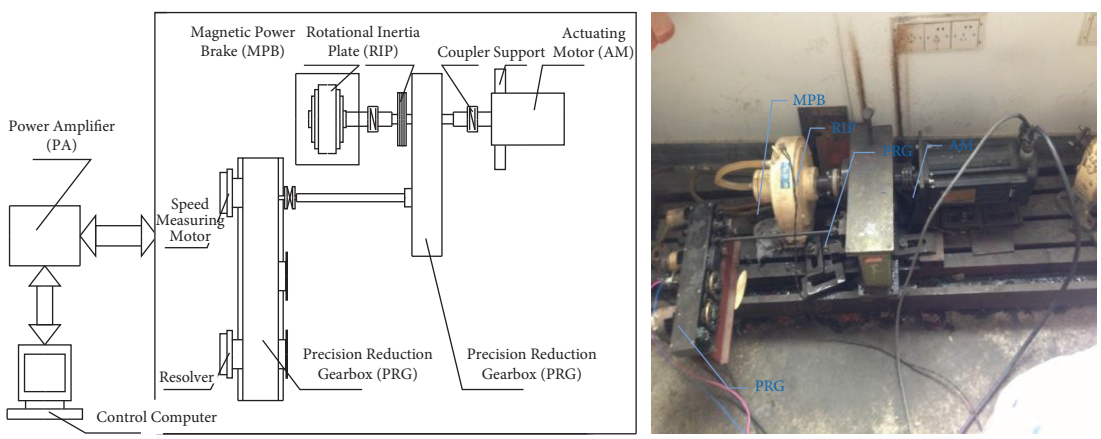


Figure 10: Semiphysical platform structure.

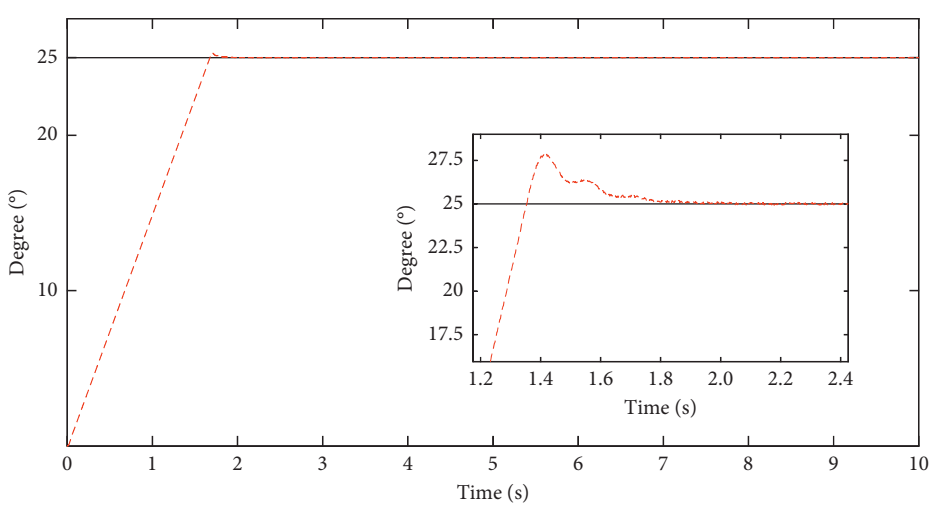


Figure 11: FWNN-LM step response curves with maximum load.

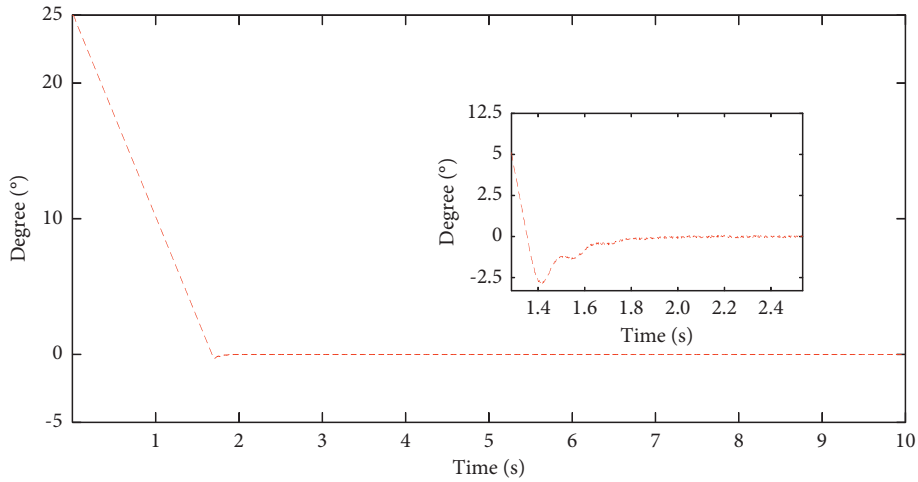


Figure 12: FWNN-LM error curves with maximum load.

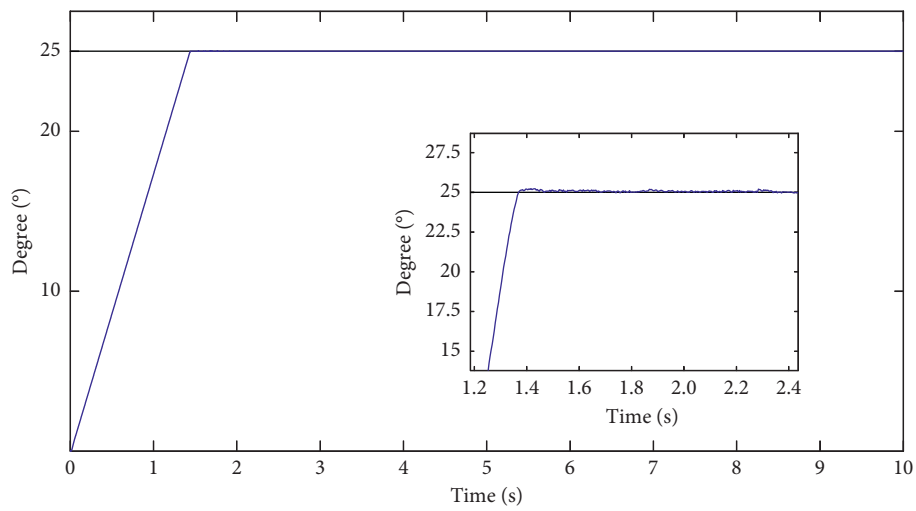


Figure 13: FWNN-LMPSO step response curves with maximum load.

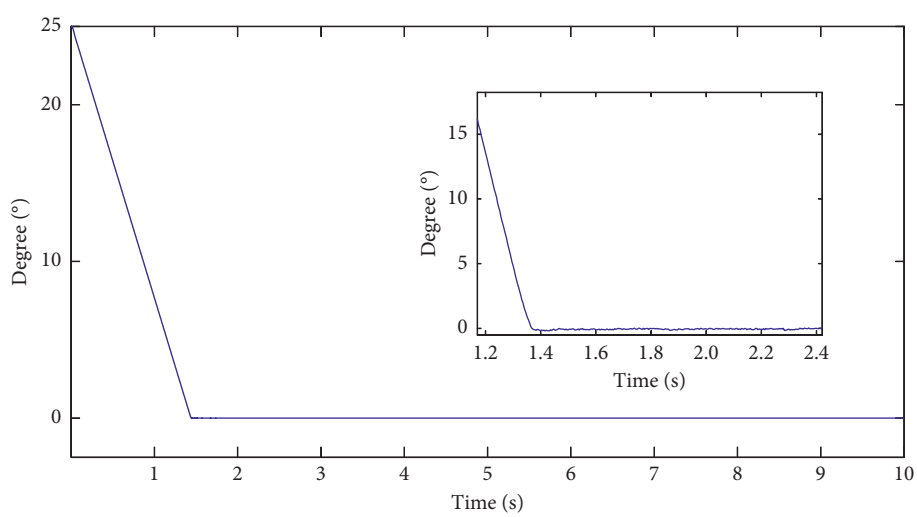


Figure 14: FWNN-LMPSO error curves with maximum load.

7. Conclusions

This study offers a new fuzzy wavelet neural network method in the AC servo system. Compared with the FWNN-LM controller, the proposed FWNN-LMPSO controller can be designed more accurately, more meaningful, and simpler. The main advantages of the existing method based on FWNN-LMPSO are as follows: first, in the FWNN-LMPSO based on the PFR method, fuzzy rules can be added to and removed from the structure learning method with the method of using the ERR value. Second, the LM algorithm improves the control accuracy through the adjustment of parameters, and the PSO learning algorithm is used to improve the learning speed. Lyapunov theory is also introduced to analysis system stability. Last, experimental results show the method has strong robustness and better dynamic performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (51805264).

References

[1] S.-M. Yang and K.-W. Lin, "Automatic control loop tuning for permanent-magnet AC servo motor drives," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 3, pp. 1499–1506, 2016.

[2] L. Liu, T. Gao, Y.-J. Liu, S. Tong, C. L. P. Chen, and L. Ma, "Time-varying IBLFs-based adaptive control of uncertain nonlinear systems with full state constraints," *Automatica*, vol. 129, no. 2021, Article ID 109595, 2021.

[3] S. Zhigang Liu and Z. Liu, "Adaptive speed control for permanent-magnet synchronous motor system with variations of load inertia," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 8, pp. 3050–3059, 2009.

[4] K. Jezernik and M. Rodic, "High precision motion control of servo drives," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 10, pp. 3810–3816, 2009.

[5] M. Sarma and K. K. Sarma, "An ANN based approach to recognize initial phonemes of spoken words of assamese language," *Applied Soft Computing*, vol. 13, no. 5, pp. 2281–2291, 2013.

[6] L. Liu, X. Li, Y.-J. Liu, and S. Tong, "Neural network based adaptive event trigger control for a class of electromagnetic suspension systems," *Control Engineering Practice*, vol. 106, no. 2021, Article ID 104675, 2021.

[7] O. Khayat, J. Razjouyan, F. N. Rahatabad, and F. Nowshiravan Rahatabad, "A fast learnt fuzzy neural network for huge scale discreted at a function approximation and prediction," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 4, pp. 693–701, 2013.

[8] J. Du, Y. Yang, D. Wang, and C. Guo, "A robust adaptive neural networks controller for maritime dynamic positioning system," *Neurocomputing*, vol. 110, no. 13, pp. 128–136, 2013.

[9] H. G. Han, Z. L. Lin, and J. F. Qiao, "Modeling of nonlinear systems using the self-organizing fuzzy neural network with adaptive gradient algorithm," *Neurocomputing*, vol. 266, pp. 1447–1459, 2017.

[10] J. Dong, Y. Wang, and G. H. Yang, "Output feedback fuzzy controller design with local nonlinear feedback laws for discrete-time nonlinear systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 6, pp. 1447–1459, 2010.

[11] H. K. Lam and J. Lauber, "Membership function dependent stability analysis of fuzzy model based control systems using fuzzy Lyapunov functions," *Information Sciene*, vol. 232, pp. 253–266, 2013.

[12] J. Wang and J. Xiao, "Constructing fuzzy wavelet network modeling," in *Proceedings of the 2005 International Conference on Intelligent Computing*, pp. 169–172, Hefei, China, 2005.

[13] P. P. Preseren and B. Stopar, "Wavelet neural network employment for continuous GNSS or bit function construction: application for the assisted-GNSS principle," *Applied Soft Computing*, vol. 13, no. 5, pp. 2526–2536, 2013.

[14] Y. Bodyanskiy, A. Dolotov, and O. Vynokurova, "Evolving spiking wavelet neuro fuzzy self-learning system," *Applied Soft Computing*, vol. 14, no. 1, pp. 252–258, 2014.

[15] D. Bayram and S. Seker, "Wavelet based neuro detector for low frequencies of vibration signals in electric motors," *Applied Soft Computing*, vol. 13, no. 5, pp. 2683–2691, 2013.

[16] M. K. Shahriari, F. Sheikholeslam, and M. Zekri, "Design of adaptive fuzzy wavelet neural sliding mode controller for uncertain nonlinear systems," *ISA Transactions*, vol. 52, no. 3, pp. 342–350, 2013.

[17] R. Cheng, P. Yan, and Y. Bai, "A novel approach to fuzzy wavelet neural network modeling and optimization," *Electrical Power & Energy Systems*, vol. 64, pp. 671–679, 2015.

[18] A. Ebadat, N. Noroozi, A. A. Safavi, and S. M. Mousavi, "New fuzzy wavelet network for modeling and control: the modeling approach," *Communications in Nonlinear Science and Numerical Simulation*, vol. 16, pp. 3385–3396, 2011.

[19] M. Davanipoor, M. Zekri, and F. Sheikholeslam, "Fuzzy wavelet neural network with an accelerated hybrid learning algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 3, pp. 463–470, 2012.

[20] R. Chandra, "Adaptive problem decomposition in cooperative coevolution of recurrent networks for time series prediction," *Neurocomputing*, vol. 86, no. 4, pp. 1–8, 2013.

[21] R. M. Hou, W. Li, Q. Gao, Y. Hou, and C. Wang, "Indirect adaptive fuzzy wavelet neural network with self-recurrent consequent part for AC servo system," *ISA Transactions*, vol. 70, pp. 298–307, 2017.

[22] J. B. Oliveira, J. Moura, P. B. Oliveira, and H. Freire, "A swarm intelligence-based tuning method for the sliding mode generalized predictive control," *ISA Transactions*, vol. 53, pp. 1501–1515, 2014.

[23] S. Sahin and M. A. Cavuslu, "FPGA implementation of wavelet neural network training with PSO/iPSO," *Journal of Circuits, Systems, and Computers*, vol. 27, no. 6, pp. 15–30, 2018.

[24] X. G. Fu, S. H. Li, M. Fairbank, D. C. Wunsch, and E. Alonso, "Training recurrent neural networks with the Levenberg-Marquardt algorithm for optimal control of agrid-connected converter," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 6, no. 9, pp. 1900–1912, 2015.

[25] A. M. Rather, A. Agarwal, and V. N. Sastry, "Recurrent neural network and a hybrid model for prediction of stock returns," *Expert Systems with Applications*, vol. 42, pp. 3234–3241, 2015.

[26] L. Liu, Y. J. Liu, A. Chen, S. Tong, and C. L. P. Chen, "Intergal Barrier Lyapunov function-based adaptive control for switched nonlinear systems," *Science China Information Sciences*, vol. 63, pp. 132203–132211, 2020.

[27] B. Z. Xia, D. Cui, S. Zhen et al., "State of charge estimate of lithium-ion batteries using optimized Levenberg-Marquardt wavelet neural netwok," *Energy*, vol. 153, no. 15, pp. 694–705, 2018.

[28] H. G. Han, Y. N. Guo, and J. F. Qiao, "Nonlinear system modeling using a self-organizing recurrent radial basis function neural network," *Applied Soft Computing*, 2017.

[29] H. G. Han, Y. N. Guo, and J. F. Qiao, "Nonlinear system modeling using a self-organizing recurrent radial basis function neural network," *Applied Soft Computing*, vol. 71, pp. 1105–1116, 2018.

Hindawi
Complexity
Volume 2021, Article ID 4551067, 28 pages
https://doi.org/10.1155/2021/4551067

Research Article

Identifying Major Research Areas and Minor Research Themes of Android Malware Analysis and Detection Field Using LSA

Deepak Thakur,[1] **Jaiteg Singh**,[1] **Gaurav Dhiman**,[2] **Mohammad Shabaz**,[1,3] **and Tanya Gera**[1]

[1]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*
[2]*Government Bikram College of Commerce, Patiala, Punjab, India*
[3]*Arba Minch University, Arba Minch, Ethiopia*

Correspondence should be addressed to Jaiteg Singh; jaiteg.singh@chitkara.edu.in and Mohammad Shabaz; mohammad.shabaz@amu.edu.et

Received 12 August 2021; Revised 25 August 2021; Accepted 28 August 2021; Published 7 September 2021

Academic Editor: Long Wang

Contemporary technologies have ensured the availability of high-quality research data shared over the Internet. This has resulted in a tremendous availability of research literature, which keeps evolving itself. Thus, identification of core research areas and trends in such ever-evolving literature is not only challenging but interesting too. An empirical overview of contemporary machine learning methods, which have the potential to expedite evidence synthesis within research literature, has been explained. This manuscript proposes Simulating Expert comprehension for Analyzing Research trends (SEAR) framework, which can perform subjective and quantitative investigation over enormous literature. TRENDMINER is the use case designed exclusively for the SEAR framework. TRENDMINER uncovered the intellectual structure of a corpus of 444 abstracts of research articles (published during 2010–2019) on Android malware analysis and detection. The study concludes with the identification of three core research areas, twenty-seven research trends. The study also suggests the potential future research directions.

1. Introduction

Data are ubiquitous, whether they are on blogs, social media platforms, discussion forums, reviews, literature, or research studies. Extracting information out of such multidimensional data is not only important but is challenging too. There is a paradigm shift in knowledge transfer among different subareas of the research held. Manual systematic reviews [1] or semiautomated [2–4] are two methods that can be employed for systematic reviews. Manual reviews are more critical and can be biased [5]. The selection of focus area, attribute selections, and interpretation entirely depends on the expertise of the reviewer. Elaborating present trends and forecasting future directions from the existing literature is not only challenging but also time-consuming for systematic manual reviews. In contrast, semiautomated methods are more generic in finding the trends [6]. Deployment of machine learning techniques within semiautomated review methods can facilitate researchers to gain a dynamic review of any literature of choice. This manuscript offers an empirical overview of contemporary machine learning methods, which have the potential to expedite evidence synthesis within research literature using Simulating Expert comprehension for Analyzing Research trends (SEAR) framework. SEAR deploys human-like intelligence to manage knowledge and information effectively. The framework leverages information modeling techniques to simulate how humans read, understand, interpret the meaning of words, and map the semantic relationship in text. The proposed SEAR framework has been deployed as TRENDMINER. As a use case, a corpus pertaining to Android security was used. During the last decade, pieces of malware are propagating at a tremendously high rate using persistent and sophisticated techniques [7]. This situation has led researchers to devise various analyses, detection, and mitigation methods, resulting in building a substantial body of literature. Continuous ongoing research augmentation of the

Android platform and malware has resulted in humongous literature. This research literature has offered numerous research prospects and has promulgated contemporary challenges within the domain. To the best of our knowledge, there is no literature investigating those challenges and research directions using semiautomated machine learning-based methods. Unlike previous works, this study is far beyond any generic study on mobile attack vectors or defense [8–11]. Instead, it oriented around emerging research trends and also suggested future directions using quantitative semiautomatic approaches. With respect to the technique being employed and dataset chosen, this study intends to answer the following research questions as framed by the research community [12]:

RQ1: can the proposed framework uncover leading researchers within a research domain?

RQ2: are those frameworks robust enough to determine the most investigated research areas?

RQ3: would the proposed framework reveal how the focus of topics within each core research area has changed over time?

RQ4: can it unfold the future directions within the research domain of choice?

Numerous topic modeling techniques such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA), and Correlated Topic Modeling (CTM) are compared and summarized in Table 1. LSA has found to be appropriate for this work as it was successfully deployed by various researchers to analyze research trends in domains such as Volunteered Geographic Information [13], Building Information Modeling [6], Supply Chain Management [14], and OpenStreetMap [5]. Several studies have demonstrated the validity of LSA in constructing a framework that leverages semantic-driven analysis to recognize and infer information from the content. Semantic-driven analysis understands the text structure, words, and the topic discussed in the document [15–29]. LSA is dependably effective in data recovery and question streamlining. It recognizes a whole of the settings where a word could show up and figures out how to set up a typical factor to address basic ideas. Examination in brain science proposes that LSA mirrors the human brain to sift through semantics from the content.

The authors in [30] proposed a method called word2vec-based LSA as a new topic modeling technique to study the trend analysis in blockchain technology. Their proposed methodology was composed of neural network-based word embedding and spherical K-means clustering. They also discussed the downside of traditional methods such as bibliometric and frequency-based analysis. They also compared their results with PLSA. In their findings, PLSA is not successful in capturing the context of the document whereas their proposed methodology was able to capture the context on real data. The authors in [31] reviewed various theoretical aspects of LSA and spatial models. They discussed various characteristics and properties empowering LSA as a suitable topic modeling technique. They also revealed some limitations and misunderstandings related to LSA. They argue that LSA has traveled a lot in providing good results as compared to other models. As a future scope, they mentioned that the fusion of different models tends to produce a coherent ecosystem. The authors in [32] performed text mining using LSA and nonnegative matrix factorization (NNMF). They discussed the strengths of LSA to process the highly sparse term-document matrix with less computation overhead. They discussed the stability of results and clustering performance while deploying LSA in their methodology. They also integrated K-means for cluster formation in their proposed methodology. In [33], the authors utilized LSA as an application to determine the memory reconstruction. LSA was applied to test that sleep reduces the semantic coherence of memory recall. In [34], the authors attempted to deploy kernel matrix estimation using LSA to increase the sharpness of the blurred image. The authors in [35] defined the applicability of LSA in determining problems in aerospace science. The authors in [36] utilized the LSA to extract the features across different knowledge domains such as information systems and operations management. In [37], the authors studied the impact of technology-enhanced learning in higher education. The topics were discovered and analyzed from the corpus related to technology-enhanced learning. The authors in [38] proposed a new taxonomy and future research directions in industry 4.0 using LSA. Various research themes related to the field were discovered and discussed.

Android security is an interesting area to explore. Malware authors tend to plant malicious code matrices inside legitimate applications to unlock their unscrupulous motives. A continuing thread of malware proliferation had let the research community perform various studies related to Android malware detection and analysis techniques. The traditional methods such as bibliometric analysis or frequency-based analysis focus on the quantitative analysis but not on the qualitative analysis [30]. These approaches are highly effort-demanding and time-consuming to perform trend analysis. The authors need to perform a full-text investigation to study the trends in the Android security field [39–41]. These approaches did not reveal the insights of the literature as they consider limited databases with limited time frames. The topic modeling techniques such as Latent Semantic Analysis (LSA) had confirmed their usefulness in determining comprehensive and detailed trend analysis. Studies in [42–45] have witnessed the use of topic modeling to identify research trends to a great extent and shown advantages over traditional methods. Table 1 shows the comparison of LSA with other topic modeling techniques. LSA focuses on revealing the diverse topics that emerged during the given timeline and provides a quantitative and qualitative evaluation. The results produced by LSA help the practitioner to pursue various potential research opportunities. LSA is used on top of this matrix to drastically reduce the vector size and capture latent topics in the corpus, while being able to infer relationships between relevant terms and respective documents, without any loss of context.

The remainder of the paper has been arranged as follows: Section 2 depicts the brief introduction to the SEAR framework. Materials and methods are discussed in Section

Table 1: Comparison of topic modeling techniques.

Technique name	Characteristics	Limitations	Area
Latent Semantic Analysis	Using SVD features, LSA can perform dimensionality reduction of TF-IDF. LSA works on synonyms of words.	Expert help is always required for labeling the topics. The interpretation of loading values sometimes becomes cumbersome.	(i) Spam filtering (ii) Automation in essay grading (iii) Topic identification.
Probabilistic Latent Semantic Analysis	Topics can be easily represented through multinomial random variables. Ability to partially handle polysemy.	Unable to perform document level modeling.	(i) Automation in essay grading (ii) Automation in question recommendation.
Latent Dirichlet Allocation	Provides multinomial distribution across words and Dirichlet distribution over topics. Capable of handling long-length documents.	Cannot predict relations among topics.	(i) Automatic labeling (ii) Emotion topic (iii) Sentiment summarization.
Correlated Topic Model	Uses logistic normal distribution for topic clustering. Produces topic graphs also.	Complex computation is involved in its processing. Too many generic words may lead to inefficiency.	(i) Query classification (ii) Topic identification (iii) Image retrieval.

3. Section 4 discusses the research questions and examines potential future research directions. Section 5 examines the outline of the proposed solution as an implication of future examination while Section 6 discusses the limitation of the investigation. Conclusions and findings are discussed in Section 7. Section 8 discusses the practical implications and future avenues of the research.

2. Proposed SEAR Framework

The proposed SEAR framework operates in the sequence as given in Figure 1.

Step 1: this step involves data gathering methods, creation of repository and XML parser, and conversion of documents to text files.

Step 2: this step involves data preprocessing of the corpus. Stop words and punctuations should be removed from the dataset, and it should be normalized before performing any text mining task.

Step 3: this step implements the TF-IDF and SVD technique, discussed in the further sections.

Step 4: this step involves the identification of core research areas and research trends. It also focuses on the mapping of research trends with the research area.

The SEAR framework utilizes a semantic analysis technique called LSA. It is a well-established algorithm to convert unstructured raw textual data into organized information objects and further analyze these objects to recognize patterns for the revelation of learning [2, 46, 47]. It employs a systematic and comprehensive approach to uncover the research trends in a vast literature dataset [3, 21, 24, 25, 48–52]. This study aims to map the semantic relationship between documents and terms in a large corpus to reveal the varied contextual latent classes using LSA.

The steps in applying LSA to the Android security corpus is identical to previously reported studies [3, 51, 53–57]. The following sections discuss the detailed procedure of this study.

3. The Use Case of SEAR Framework: TRENDMINER

TRENDMINER is the use case of the SEAR framework which takes text documents as an input, as shown in Figures 2 and 3. The dataset of 444 abstracts is considered sufficiently large enough for performing text mining, as explained in [3]. Python 3.7 programming language was used to perform all the experimentation. Table 2 shows the software versions used in our work. The machine used for the experimentation was configured with Intel Core i5 6200U with 2.4 GHz and 8 GB RAM. Once the literature dataset on Android security is successfully uploaded on TRENDMINER, it is further fed to Latent Semantic Analysis (LSA), which is a backbone of TRENDMINER. LSA is a text data mining and natural language processing technique used to retrieve and query a massive corpus of literature [51, 56, 58]. As a scientific and measurable strategy, LSA is utilized to recognize the latent concepts inside the textual data at the semantic level [59–63].

3.1. Step 1: Data Acquisition. This section reveals the keywords, search strategy, and selection criteria used for preparing the large corpus. Reputed databases were used for the collection of research articles on Android security. Inclusion and exclusion criteria were applied to refine the searching results to get relevant research articles. The repository was made to achieve standard uniformity across the research articles.

3.1.1. Task A: Dataset Preparation. The first task was to prepare the literature dataset for TRENDMINER. The approach followed for collecting the literature dataset is primarily focused upon the structure of Android applications, the probable vulnerabilities within existing application development along the methods adopted for malware identification and mitigation. The strategy adopted for searching and selecting literature is defined by 3C's Formula, depicted in Figure 4:

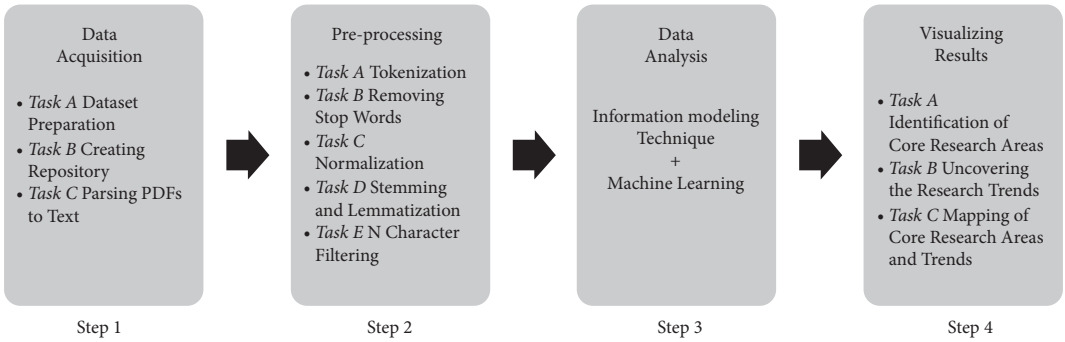


Figure 1: Sequence diagram of SEAR framework.

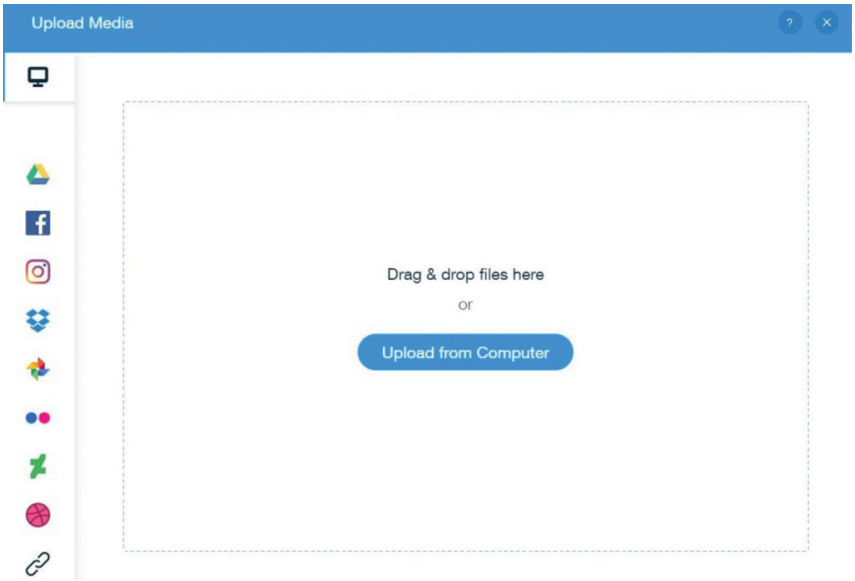


Figure 2: Uploading interface of TRENDMINER.

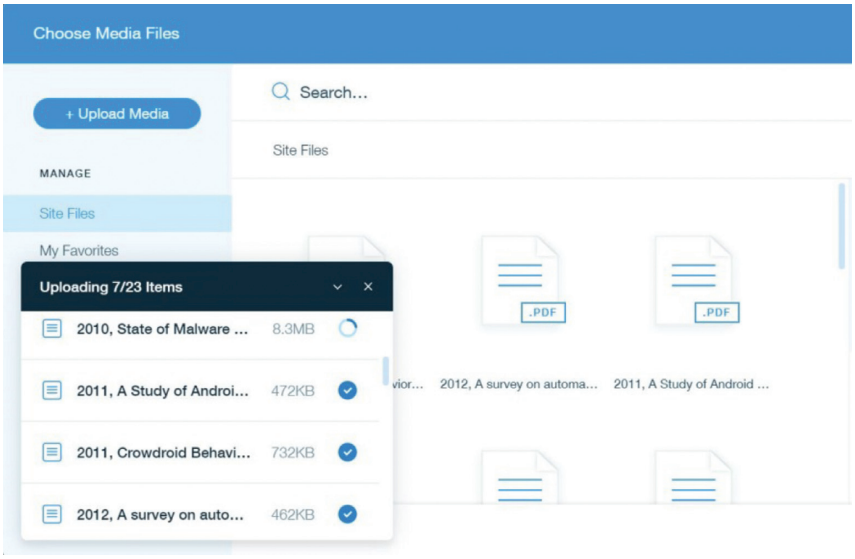


Figure 3: Files getting uploaded on TRENDMINER.

Table 2: Software specifications.

Library	Version	Implementation in TRENDMINER	Open source
PDFMiner	≥20140328	Used in data acquisition (parsing PDFs to text)	Yes
NLTK	≥3.4	Preprocessing (all tasks)	Yes
Scikit-Learn	≥0.20rc1	Data analysis	Yes

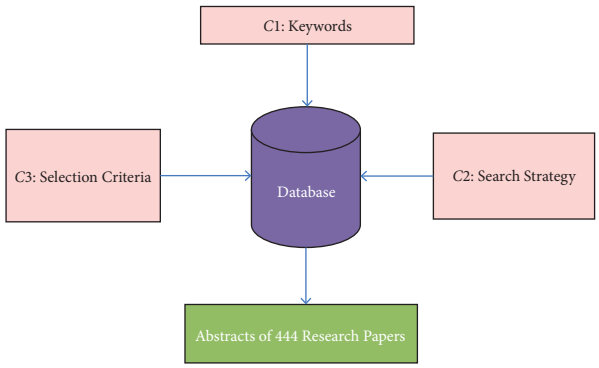


Figure 4: Dataset preparation using 3C's Formula.

(1) Component 1: Keywords. The articles were selected using keywords such as "malware," "vulnerability," "security," "privacy," "monitoring," "application," "smartphone," "android," "virus," "static," "dynamic," "detection," and "data flow."

(2) Component 2: Search Strategy. The TRENDMINER considered reputed research from prominent databases such as IEEE Xplore, ACM Computing Library, Science Direct, Springer, and Google Scholar which was queried to collect high-quality papers on Android malware analysis and detection techniques. Scopus indexed articles from prominent databases were duly included while searching the literature. Figure 5 illustrates the proportion of Scopus indexed articles in our corpus.

(3) Component 3: Selection Criteria. Raw results from the databases mentioned above were refined based on the Android operating system. Papers on operating systems such as Symbian and iOS were discarded.

3.1.2. Task B: Creating a Repository for TRENDMINER. Mendeley, a tool from Elsevier [64], has been used to build the literature database. It provides a systematic way to retrieve the authors, years, and abstracts of all the research papers indexed into its file system and also to export all of them as citations and XML tree structures. Parsing of resultant XML tree structure was one of the significant challenges during this study. A consistent naming convention for the whole literature dataset was necessary. Renaming the articles using particular objects common to all research documents will have a significant impact on their future ease.

A module in TRENDMINER was developed, known as XML Parser. The purposefully generated XML corpus was further parsed to a more structured format, i.e., comma-separated values (CSVs). Figure 6 shows the generic conversion process flow.

The exported files consist of metadata information such as authors, year of publication, and publishers. The following observations were made during the prelim analysis of the corpus. Based on the number of occurrences in the dataset, the top researchers with the most publications on Android security during the period 2010–2019 were calculated and are presented in Figure 7.

Figure 8 shows the top fifteen journals publishing articles related to Android security. Figure 7 interprets that the top authors were Wang, Xiaofeng, and Jiang, Xuxian, with 13 publications, with Zhou, Yajin closely following on 12. The graph obtained was from the analysis performed on the dataset chosen, as described above. Figure 8 identifies Computers and Security (Elsevier) and IEEE, as the top publishers, publishing research in the Android malware and security field. NDSS, Springer, and ACM closely follow them.

3.1.3. Task C: Parsing the PDF Documents to Text. Conversion of pdf to text was subsequently performed to make the dataset input ready, compatible with TRENDMINER. Various tool options for the conversion process are available, namely, PDFMiner, Tika, and Textract. PDFMiner [65] was opted in the experimental study because of the following significant benefits:

(i) PDFMiner can obtain the exact location of text on a page along with information such as fonts or a number of lines.

(ii) It facilitates the conversion of PDF files into other text formats (such as HTML).

(iii) It provides accurate results even under extreme conditions such as parsing large corpus.

3.2. Step 2: Preprocessing the Text Files. After the successful conversion to text files, the next step was to employ preprocessing procedures. The preprocessing module in TRENDMINER helps to gain quality information out of the text by applying appropriate preprocessing techniques. For any text mining algorithm, the preprocessing of the collected dataset is an essential step [66, 67]. This involves the expulsion of names, numbers, abbreviations, slang, acronyms, punctuation, and N characters as recommended in [3].

Preprocessing of corpus involves the execution of the following procedure, developed in Python platform using NLTK package. NLTK is Natural Language Toolkit [68].

3.2.1. Task A (Tokenization). In this step, large chunk of text was tokenized into sentences, then sentences into words.

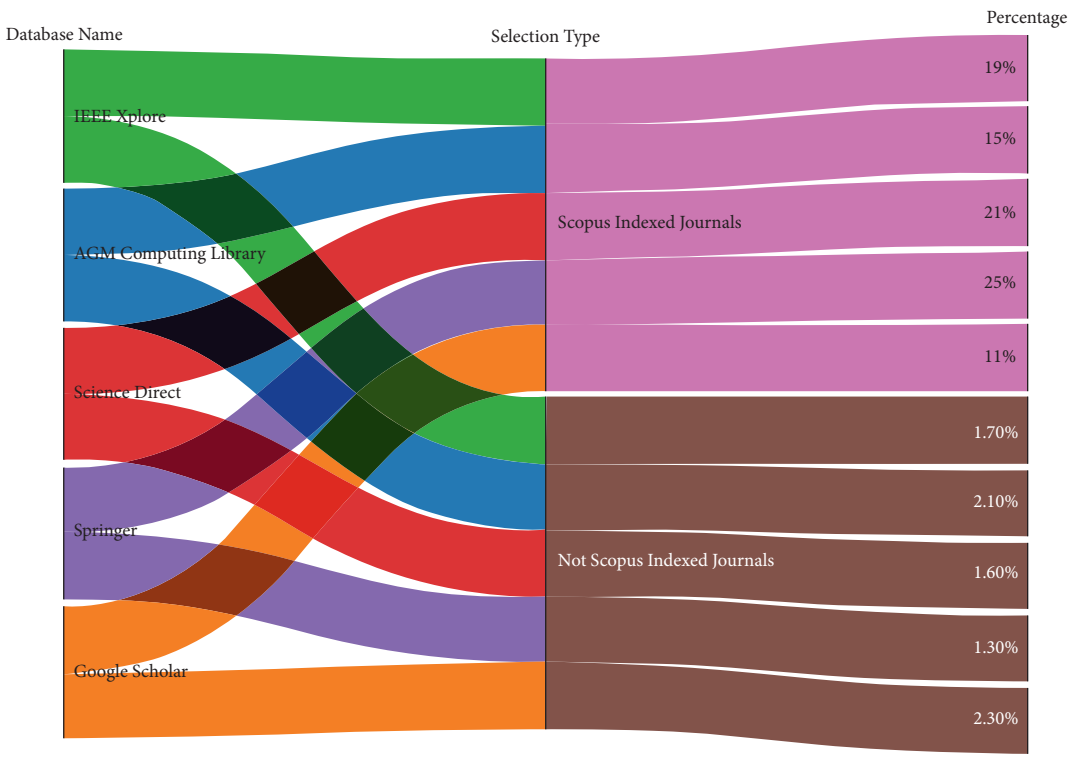


Figure 5: Distribution of Scopus indexed articles among the databases.

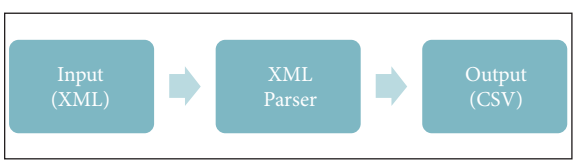


Figure 6: Parsing an XML to CSV.

3.2.2. Task B (Removing Stop Words). Stop words using NLTK's support and common words (sample, benign, learning, malware detection, malware, detection, training, layer, channel, attacker, password, market, call, warning, algorithm, installation, detector, socket, etc.) were removed.

3.2.3. Task C (Normalization). Normalization is applied over the words to introduce uniformity and maintain consistency among the text documents. The task of normalization is composed of several subtasks such as removing punctuation from the text, changing overall content to a similar case either uppercase or lowercase, and converting numbers to words. Normalization helps to keep all words on equivalent balance to allow smooth processing of the textual data.

3.2.4. Task D (Stemming and Lemmatizing). For further processing of documents, the dictionary size has to be reduced and should be populated with unique words. Stemming and lemmatizing are the techniques that are performed over the words to reduce the inflection. The idea is to reduce the words to the common root form. In stemming, base form is known as stem while in the case of lemmatizing, it is known as a lemma. Stems might not be actual or real words, but on the other hand, lemmas are the actual language words. These two techniques help in achieving faster processing of text documents.

3.2.5. Task E (Character Filtering). All words less than length 4 were omitted [3].

It is to be noted that the initial dataset contained 60,184 tokens which represents the length of the vocabulary in the entirety of the corpus. Before the dataset is fed to other computational steps, it has to be nonredundant and free from any kind of noise. After applying appropriate preprocessing procedures as discussed previously, the word list was retained with 1944 tokens. In this study, 444 documents and the resulted wordlist of the 1944 tokens represent columns and rows, respectively. A term frequency is created where each term maps to a count of occurrence in each document. Furthermore, this matrix is transformed into a weighted matrix using the TF-IDF weighting scheme.

3.3. Step 3: Data Analysis Using Information Modeling and Machine Learning Techniques. This work makes use of the information modeling technique to expedite the data analysis process over the corpus. With the conjunction of information modeling and machine learning techniques, human interpretable topics can be extracted from a document corpus. Machine learning approaches enhance the ability of information modeling techniques by allowing researchers to intelligently extract and manage the crucial

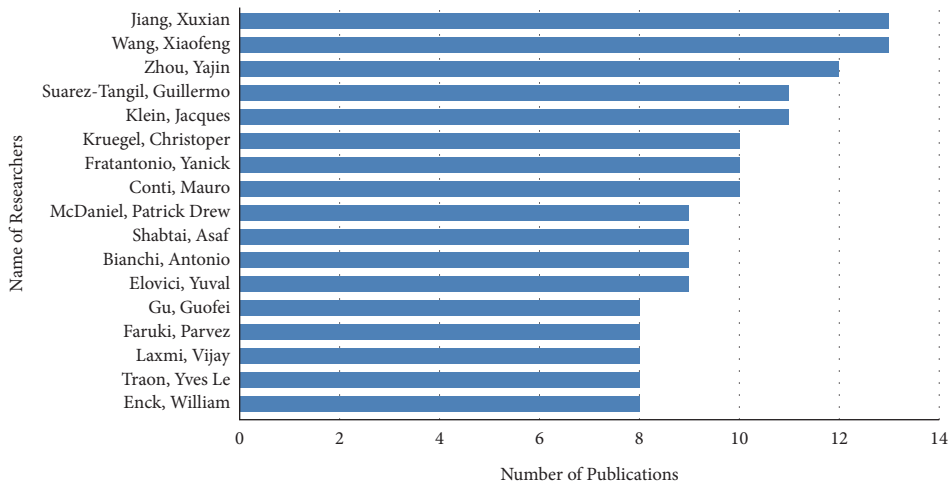


Figure 7: Top researchers in Android security research.

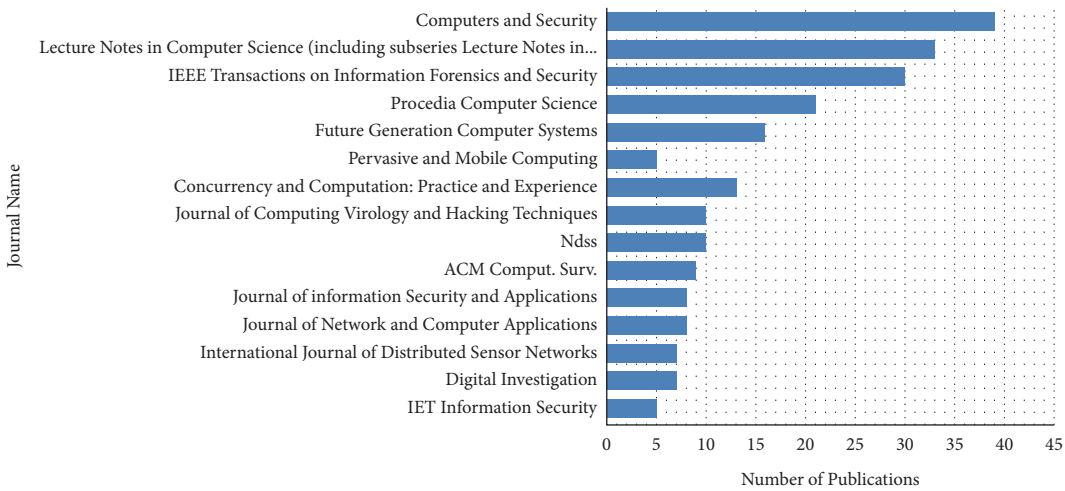


Figure 8: Top journals focusing on Android security field.

information to make smart decisions. Deploying Latent Semantic Analysis (LSA) as the information modeling technique can automatically identify topics and unveil hidden patterns in the vast corpus of data. LSA uses the matrix method called Singular Value Decomposition (SVD) to construct a low-rank approximation from extensive matrix data. SVD is the major strength of the LSA and one of the basic machine learning algorithms. It reduces the dimensions of the data without losing a significant amount of information. The main idea is to apply LSA on a document set and unsupervised machine learning approach on a reduced dimension set to group similar documents according to their topic areas. *K*-means, which is the unsupervised machine learning approach, fitted in the LSA model to uncover the latent structure of the corpus.

3.3.1. Task A: From Documents to Matrices—TF-IDF (Term Frequency Inverse Document Frequency). In this study, a mapping needs to be investigated from the documents to the latent topics that they all relate to. For that, the most important words were to be identified which can later lead to the latent topic discovery. The TRENDMINER leverages the essence of the technique, called Term Frequency Inverse Document Frequency (TF-IDF). There are other weighting methods are available for the analysis. The most common weighting schemes are TF-IDF and log-entropy. As per the study in [3], a potential weakness of log-entropy was discovered and it proved to be biased towards high-frequency terms in the dataset. For instance, log-entropy produces a better result with article titles or documents with a short text. TF-IDF performs better in discovering the patterns in large semantic spaces of larger groups of terms. Motivated by this finding, we utilized the TF-IDF technique as the weighting method in the study.

The Latent Semantic Analysis (LSA) topic model algorithm requires a document-term matrix as the main input. TF-IDF helped in maintaining a document-term matrix that described the frequency of terms that occur in a collection of documents. The documents and words in a matrix correspond to columns and rows, respectively. TF-IDF has widely been into usage for better topic analysis [3, 69, 70]. The

resulting document-term matrix of the example stated in the previous example is presented in Table 3.

(1) TF (Term Frequency). It processes the standardized Term Frequency (TF), which is determined as the frequency a term shows up in a report, separated by the complete number of terms in that record, refer to equation (1). TF matrix is shown in Table 4:

$$\mathrm{TF}(t,d) = \frac{\text{number of occurences of term } t \text{ appears in document } d}{\text{total number of terms in the document}}. \tag{1}$$

(2) IDF (Inverse Document Frequency). It estimates how significant a term is. IDF is processed as the logarithm of the quantity of records in the corpus isolated by the quantity of reports where the particular term shows up. Nonetheless, it is realized that specific terms, for example, "is," "of," and "that" or space explicit words, may seem a great deal of times however have little significance. In this way, there is a need to overload the continuous terms while increasing the uncommon ones, by figuring condition 2. IDF grid is introduced in Table 5:

$$\mathrm{IDF}(t,d) = \log \frac{\text{total number of documents}}{\text{number of documents with term } t \text{ in it}}. \tag{2}$$

The below equation (3) presents the TF-IDF scores:

$$w_{t,d} = \mathrm{TF}_{t,d} \times \mathrm{Log}\frac{N}{df_t}. \tag{3}$$

In equation (3), t means the terms, d signifies each record, and N indicates the complete number of reports. Consider Table 6, which addresses the report term lattice with TF-IDF scores for the recently expressed model. A term will have a huge weight when it much of the time happens across the archive yet inconsistently across the corpus. The word malware may show up frequently in an archive, but since it is probable reasonably entirely expected in the remainder of the corpus. To reveal the connection between the words and records and catch the latent themes inside the Android security dataset, dimensionality reduction must be performed, as examined in the following area.

3.3.2. Task B: Learning Latent Relationships between Documents Using SVD (LSA). Utilizing SVD, two sets of loading matrices were produced as the output of LSA. One is a document-to-topic matrix and the other one is a term-to-topic matrix. The topic solutions are the number of research themes in the literature dataset. High term or document loading in the matrix cell discloses the fact that a specific term or document is more inclined towards a particular topic solution. The researcher can adjust the detail level of a number of topic solutions for identifying research areas and trends. Smaller values of topic solution represent common research core areas, and higher values of topic solution represent principal research trends [51].

Truncated SVD is a framework variable-based math method that breaks down the TF-IDF lattice into a result of three grids: U, Σ, and V. The SVD disintegration is shown in

$$A = U \times \Sigma \times V^T. \tag{4}$$

Here, A addresses the TF-IDF lattice, U addresses the document-to-topic framework portraying relationship between documents attached to different concepts, V addresses the term-to-topic depicting relationship among concepts and terms, and Σ is composed of nonnegative numbers.

Suppose d is the number of records, t is the number of terms in the documents, and k is considered as the hyperparameter demonstrating the quantity of points to be separated from the corpus. A_k is the low-rank estimate of matrix A and can be delivered utilizing shortened SVD as continues in

$$A_k = U_k \times \Sigma_k \times V_k^T, \tag{5}$$

where U_k is the document-to-topic matrix $(d \times k)$, V_k is a term-to-topic matrix $(t \times k)$, and Σ_k is the topic-to-topic matrix $(k \times k)$. Table 6 shows the changed term frequencies subsequent to applying TF-IDF. SVD procedure must be applied to the TF-IDF matrix introduced in Table 6.

Tables 7 and 8 contain the factor loading values that are arbitrarily positive and negative. The set of terms and documents need to be mapped with the latent topics. To interpret the meaning of the loading values, the technique known as varimax rotation was applied on terms and document loading matrices. The varimax rotation helps to uncover the best correlation of terms with the latent topics. The rotation magnifies the association of terms and documents to the latent topics. Furthermore, a threshold value needs to be selected to discover the significant terms as discussed in [3, 5]. Empirical probability distribution was utilized to select the threshold values for different factor solutions. The loading values are transformed into a vector and sorted in descending order, thereby defining the threshold as retaining $1/n$ of the loadings, where n is the factor solution as explained in [5, 6]. For each factor solution, loading values are grouped by considering their absolute values to unveil latent topics. As an application of LSA followed by an unsupervised machine learning approach, discussed further, it will help to identify topic solutions.

TRENDMINER is used to identify the core research areas and significant research trends in Android security, and an optimal value for k topic solutions has to be determined. Choosing an optimal value for k is always a challenge; because the more the number of dimensions k chosen, the more will be the risk of induction of noise in the data [58, 71]. However, at the same time, selecting a smaller value of k will lead to losing important semantics. It is a good practice to include a bigger k, as an approach to deduce more trends or classify many trends into a single category [72]. A k-iterative process has been applied to uncover the core research areas and their subclassification of related trends. SVD provides the matrix of singular values that are defined as the square root of the eigenvalues. These values provide

Table 3: Document-term matrix describing frequency of terms.

Terms	Doc1	Doc2	Doc3	Doc4	Doc5
Access	0	0	0	0	1
Applic	1	1	1	1	1
Calendar	0	1	0	0	0
Connect	0	0	1	0	0
Contact	0	1	0	0	0
Daili	0	0	0	1	0
Data	0	1	0	1	0
Devic	2	0	0	0	0
Exact	0	0	1	0	0
Find	0	0	1	0	0
Identifi	1	0	0	0	0
Like	0	1	0	0	0
List	0	1	0	0	0
Locat	0	0	1	0	0
Malwar	1	1	0	1	0
Messag	0	0	0	0	1
Misus	0	1	0	0	1
Network	0	0	1	0	0
Number	0	1	0	0	0
Phone	0	1	0	0	0
Read	1	0	0	0	0
Record	0	0	0	1	0
Send	0	0	0	1	0
Server	0	0	0	1	0
Tower	0	0	1	0	0
Track	1	0	1	0	0
Uniqu	1	0	0	0	0
Usag	0	0	0	1	0
User	1	1	1	0	0
Various	0	0	0	1	0
Wifi	0	0	1	0	0

Table 4: Term frequency scores for each document.

Documents	Term frequency scores
Doc1	{"Malwar": 0.1111111111111111, "applic": 0.1111111111111111, "read": 0.1111111111111111, "uniqu": 0.1111111111111111, "devic": 0.2222222222222222, "identifi": 0.1111111111111111, "track": 0.1111111111111111, "user": 0.1111111111111111}
Doc2	{"Malwar": 0.09090909090909091, "applic": 0.09090909090909091, "misus": 0.09090909090909091, "user": 0.09090909090909091, "data": 0.09090909090909091, "like": 0.09090909090909091, "phone": 0.09090909090909091, "number": 0.09090909090909091, "contact": 0.09090909090909091, "list": 0.09090909090909091, "calendar": 0.09090909090909091}
Doc3	{"Applic": 0.1, "track": 0.1, "exact": 0.1, "locat": 0.1, "user": 0.1, "find": 0.1, "wifi": 0.1, "network": 0.1, "tower": 0.1, "connect": 0.1}
Doc4	{"Various": 0.1111111111111111, "malwar": 0.1111111111111111, "applic": 0.1111111111111111, "record": 0.1111111111111111, "daili": 0.1111111111111111, "usag": 0.1111111111111111, "data": 0.1111111111111111, "send": 0.1111111111111111, "server": 0.1111111111111111}
Doc5	{"Applic": 0.25, "access": 0.25, "messag": 0.25, "misus": 0.25}

the concept strength and are arranged in descending order. The k singular values are selected using a scree plot as depicted in Figure 9. As illustrated in the study [24], a high level of topics must be chosen using an empirical approach that involves multiple trials of LSA. The number of factors in individual trials ranged from 2 to 10. After reviewing high-loading terms/documents for each factor solution, experts decided to set three as core high-level research areas. It should be noted that it also depends upon the semantic space chosen for the experimentation.

Furthermore, based on the expert opinions and scree plot analysis [14, 73], dimensionalities of 27 topics were found to be significant elbow point detected through an iterative log-likelihood ratio test on eigenvalues [74]. The optimal number of twenty-seven topic solutions can be considered optimal for depicting the research trends in Android security of a large corpus; in addition, three topic solutions were considered to describe the core research areas. Topic clustering, topic labeling, and detailed analysis have been discussed in further sections.

Table 5: Inverse document frequency score for each term.

Terms	IDF score
Access	2.098612
Applic	1.000000
Calendar	2.098612
Connect	2.098612
Contact	2.098612
Daili	2.098612
Data	1.693147
Devic	2.098612
Exact	2.098612
Find	2.098612
Identifi	2.098612
Like	2.098612
List	2.098612
Locat	2.098612
Malwar	1.405465
Messag	2.098612
Misus	1.693147
Network	2.098612
Number	2.098612
Phone	2.098612
Read	2.098612
Record	2.098612
Send	2.098612
Server	2.098612
Tower	2.098612
Track	1.693147
Uniqu	2.098612
Usag	2.098612
User	1.405465
Various	2.098612
Wifi	2.098612

3.3.3. Task C: Topic Clustering. As stated in [3], clustering and factor analysis are the two analytic steps that are involved in post-LSA procedures. The authors discussed the main considerations that would let practitioners/decision-makers/researchers deploy these analytic steps as per their requirements. They focused on the fact that LSA has been used for clustering and factor analysis purposes. Based on the semantic space created in this study, the domain experts decided to pursue the clustering technique. The clustering approach was implemented through the *K*-means algorithm. Machine learning can be employed on top of results obtained after the application of Latent Semantic Analysis to significantly reduce the manual effort by a domain expert in determining the document to its closest topic. *K*-means is an unsupervised machine learning technique generally used when there are no labels of the data points and it learns them based on their relative positions in a vector space. The centroid feature weights may be used to identify the nature of the cluster while defining the groups, which may be used to label new data [75, 76]. *K*-means is easy to implement and can process extremely large samples [77]. Usually, the inputs into *K*-means are passed through a dimensionality reduction algorithm. LSA and *K*-means are applied in a linear combination for the interpretation of the results to find similar documents and their associations with the terms contained in the textual corpus [78–80], which is done to recommend research papers corresponding to a particular topic label. The interpretation of the results obtained is domain-specific. For instance, if data points were research articles on Android security in extensive literature, *K*-means will segregate the entire documents into k subgroups. The research trends in the domain of Android security which are a part of each subgroup or cluster have some common features, which are used for further analysis. The number of clusters was chosen to be three, with the selection done iteratively. It is to be noted that the choice of too few clusters may not reveal the actual underlying relationships, while too many clusters may account for noise, which would not be useful for any further analysis on the outputs obtained. The output, in the form of a multidimensional array, is composed of titles for all documents labeled with the respective cluster numbers. Taking the dot product of the components obtained from LSA with the cluster centroids, the results obtained are sorted to show only the top topics corresponding to each cluster, which require sensible topic labeling as discussed in the next section.

3.3.4. Task D: Topic Labeling. The term-to-topic and document-to-topic matrices consist of significant values to uncover topics. Each cell in both matrices represents the loading values which were later sorted in descending order. The results obtained from previous steps of TRENDMINER become the input for successful topic labeling. High-loading terms and documents were examined together and sensible labels were given against three and twenty-seven topic solutions, as shown in Figures 10 and 11. We have implemented the Delphi method [81] to perform the topic labeling process. The graphical representation of the Delphi method is also shown in Figure 12. Topic labeling is a collective intelligence task that involves the most reliable opinions of a group of experts. The Delphi method is an iterative method that worked under controlled monitoring and feedback mechanisms to build robust consensus.

3.4. Step 4: Results and Findings. As a result, three topic solutions present the major core research areas, as shown in Figures 13 and 14 along with the sensible topic label. Each topic solution is denoted as *Tm.n* where m denotes topic solution whereas n denotes an nth factor of the m topic solution. For instance, T27.3 illustrates the third factor of twenty-seven topic solutions. The graphical representations plotted for all point arrangements likewise give count about the publication distribution for every topic solution during three unique periods inside 2010–2019, as shown in Figures 13 and 15. The distribution check related with every subject arrangement addresses the significance of the comparing research region inside that theme arrangement. Furthermore, to reveal the examination patterns and future scope in the field of Android security, 27 point arrangements were found as depicted in Figures 15(a) and 15(b). The semantic relationship between 27 theme arrangements and three core research areas assists with recognizing research patterns inside each center exploration area of Android security, as depicted in Figures 10 and 11.

Table 6: Transformed term frequencies after TF-IDF generation.

Terms	Doc1	Doc2	Doc3	Doc4	Doc5
Access	0.000000	0.000000	0.000000	0.000000	0.589463
Applic	0.160859	0.164157	0.165134	0.176043	0.280882
Calendar	0.000000	0.344502	0.000000	0.000000	0.000000
Connect	0.000000	0.000000	0.346553	0.000000	0.000000
Contact	0.000000	0.344502	0.000000	0.000000	0.000000
Daili	0.000000	0.000000	0.000000	0.369447	0.000000
Data	0.000000	0.277942	0.000000	0.298067	0.000000
Devic	0.675160	0.000000	0.000000	0.000000	0.000000
Exact	0.000000	0.000000	0.346553	0.000000	0.000000
Find	0.000000	0.000000	0.346553	0.000000	0.000000
Identifi	0.337580	0.000000	0.000000	0.000000	0.000000
Like	0.000000	0.344502	0.000000	0.000000	0.000000
List	0.000000	0.344502	0.000000	0.000000	0.000000
Locat	0.000000	0.000000	0.346553	0.000000	0.000000
Malwar	0.226081	0.230717	0.000000	0.247423	0.000000
Messag	0.000000	0.000000	0.000000	0.000000	0.589463
Misus	0.000000	0.277942	0.000000	0.000000	0.475575
Network	0.000000	0.000000	0.346553	0.000000	0.000000
Number	0.000000	0.344502	0.000000	0.000000	0.000000
Phone	0.000000	0.344502	0.000000	0.000000	0.000000
Read	0.337580	0.000000	0.000000	0.000000	0.000000
Record	0.000000	0.000000	0.000000	0.369447	0.000000
Send	0.000000	0.000000	0.000000	0.369447	0.000000
Server	0.000000	0.000000	0.000000	0.369447	0.000000
Tower	0.000000	0.000000	0.346553	0.000000	0.000000
Track	0.272357	0.000000	0.279596	0.000000	0.000000
Uniqu	0.337580	0.000000	0.000000	0.000000	0.000000
Usag	0.000000	0.000000	0.000000	0.369447	0.000000
User	0.226081	0.230717	0.232090	0.000000	0.000000
Various	0.000000	0.000000	0.000000	0.369447	0.000000
Wifi	0.000000	0.000000	0.346553	0.000000	0.000000

Table 7: Term-loading with five latent topics.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Access	0.198118	−0.244194	−0.399978	0.007002	−0.317144
Applic	0.348381	−0.034393	−0.080514	−0.037085	−0.120670
Calendar	0.166381	−0.098917	−0.004802	0.024460	0.299730
Connect	0.107482	0.219059	−0.075849	−0.241415	0.008454
Contact	0.166381	−0.098917	−0.004802	0.024460	0.299730
Daily	0.128417	−0.108420	0.265857	−0.127129	−0.161347
Data	0.237841	−0.167279	0.210618	−0.082833	0.111647
Locat	0.107482	0.219059	−0.075849	−0.241415	0.008454
Malwar	0.284974	−0.031505	0.205508	0.104866	0.037136
Messag	0.198118	−0.244194	−0.399978	0.007002	−0.317144
Misus	0.294076	−0.276820	−0.326574	0.025384	−0.014050
Network	0.107482	0.219059	−0.075849	−0.241415	0.008454
Number	0.166381	−0.098917	−0.004802	0.024460	0.299730
Phone	0.166381	−0.098917	−0.004802	0.024460	0.299730
Read	0.130719	0.160295	0.045805	0.259253	−0.082933
Record	0.128417	−0.108420	0.265857	−0.127129	−0.161347

3.4.1. Task A: Identification of Core Research Areas in Android Security. Core research areas shown in Figure 13 were discovered as three topic solutions that focused on "Application Structure Analysis" (T3.1), "Static Level Monitoring" (T3.2), and "Automatic Malware Analysis" (T3.3). The word cloud for three topic solutions is shown in Figure 14. These articles emphasized imperative techniques to analyze, detect, and assess Android malware.

The outcomes showed that various high-stacking distributions joined to one exploration region, i.e., "Static Level Monitoring" (T3.2) in the three theme arrangements. Static investigation is the most used examination strategy for

Table 8: Document loading with five latent topics.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Doc1	0.542964	0.677204	0.434886	0.487392	0.471276
Doc2	0.491949	−0.297480	0.654891	−0.304044	−0.429196
Doc3	0.129873	−0.013342	−0.209489	0.688773	−0.649469
Doc4	0.640698	0.592341	−0.581166	−0.287077	0.919112
Doc5	−0.189251	0.670234	0.187921	−0.336431	−0.414465

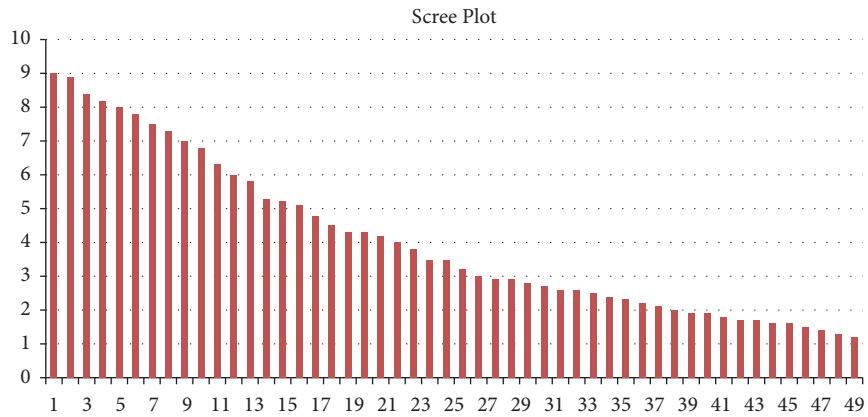


Figure 9: Scree plot.

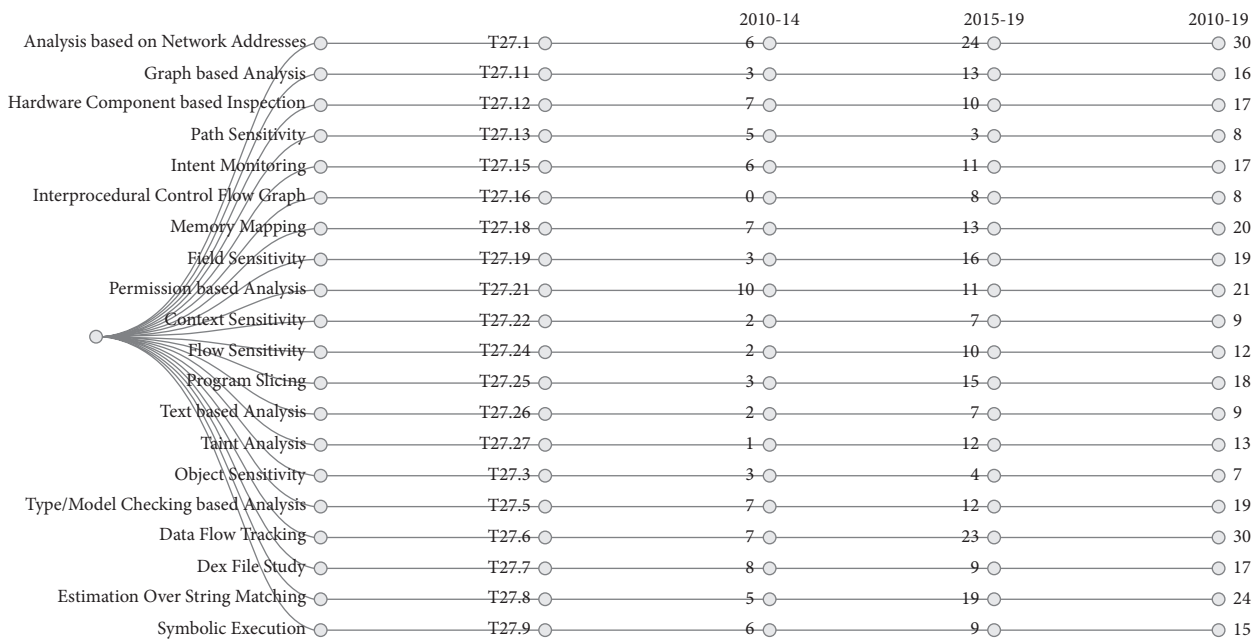


Figure 10: Mapping of core research area 3.2 and trends.

malware investigation; thus, it is obvious that "Static Level Monitoring" (T3.2) stayed in the moving exploration region over time 2010–2019. Results likewise showed that "Automatic Malware Analysis" (T3.3) additionally turned into a moving exploration region during the year 2015–2019. However, "Application Structure Analysis" (T3.1) had less impact on the set of papers collected in this study.

In the corpus, approaches dependent on Static Level Monitoring (T3.2) are the most well-known techniques (about 74%) utilized by the scientists to catch the security dangers in an Android system. Automatic Malware Analysis (T3.3) is around 20%, and Application Structure Analysis (T3.1) is 6%. First static analysis technique was introduced in 2009 [82] and in the year 2010, and dynamic analysis

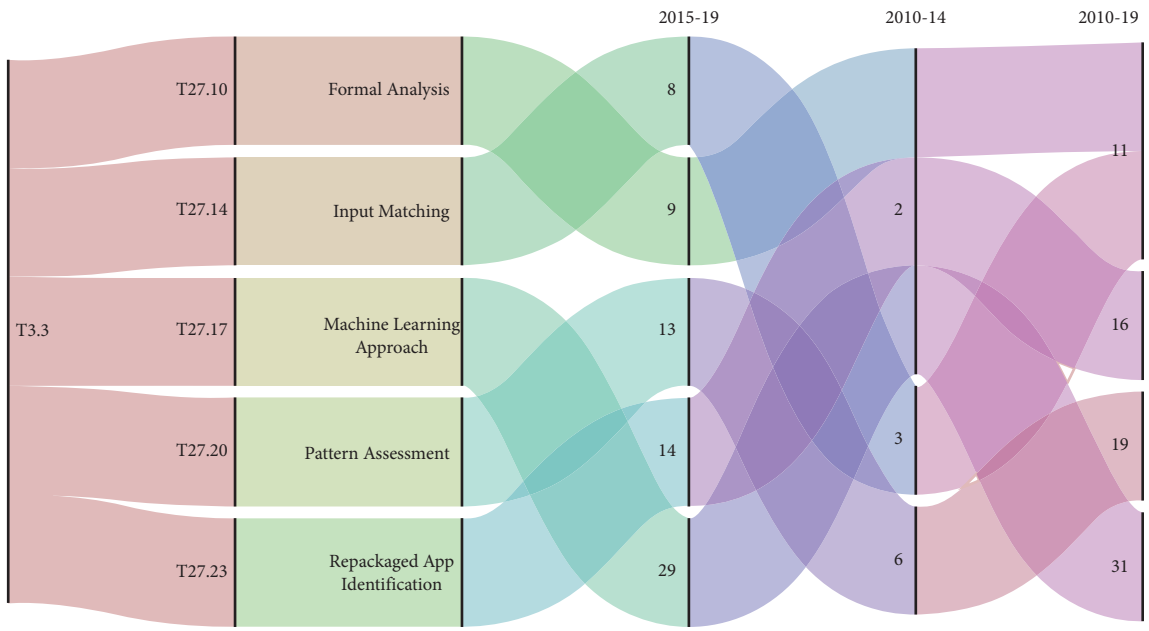


Figure 11: Mapping of core research area 3.3 and trends.

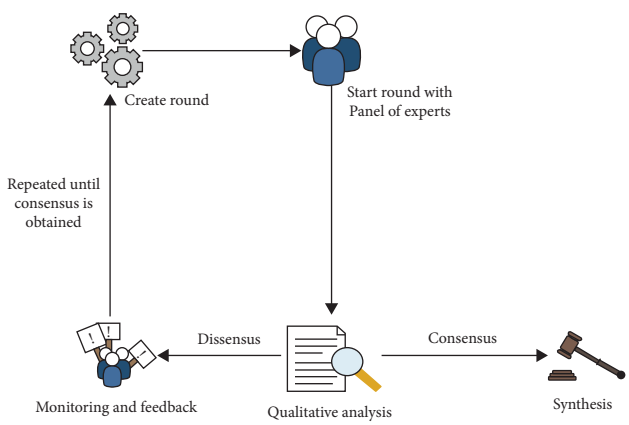


Figure 12: Working flow of the Delphi method.

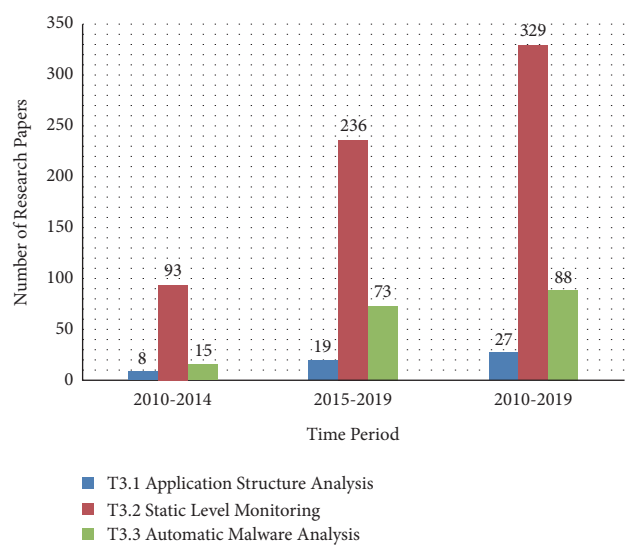


Figure 13: Publication count for three-factor solution during three different time periods.

technique was first explored by the researchers [83, 84]. The former investigated the data flows in applications that violate the security policies stored in an application's configurations. The latter identified the data leakage from sensitive sources of an application. Notwithstanding static and dynamic methodologies, there exist a couple of hybrid approaches that take advantage of the upsides of investigation such as static and dynamic. These techniques typically first apply static investigation to identify potential security threats in an Android system and after that perform dynamic procedures to enhance their accuracy by dispensing with the false alerts. For instance, in [85], the authors first used the static investigation to distinguish possibly vulnerable applications.

3.4.2. Task B: Identification of Android Security Research Trends and Task C: Core Research Areas and Trend Mapping. The TRENDMINER uncovered 27 subject core research trends as displayed in Figures 15(a) and 15(b). Figures 10 and 11 show the relationship of core areas with the research themes. The relationship is performed dependent on similarity scores. Documents were clustered into a lesser number of topic solutions as a start, while the higher value was chosen later. The points comparing to the last were to some degree identified with the previous and were checked utilizing similitude scores. The likeness scores were determined because of string coordinating, with the string similitudes indicating the closeness of the low and high upsides of theme arrangements. This was done to verify the understanding that the result while choosing a lower value of topic solutions would correspond somewhat to having chosen a comparatively higher value. The likeness scores present a reasonable connection between the core areas and their connected

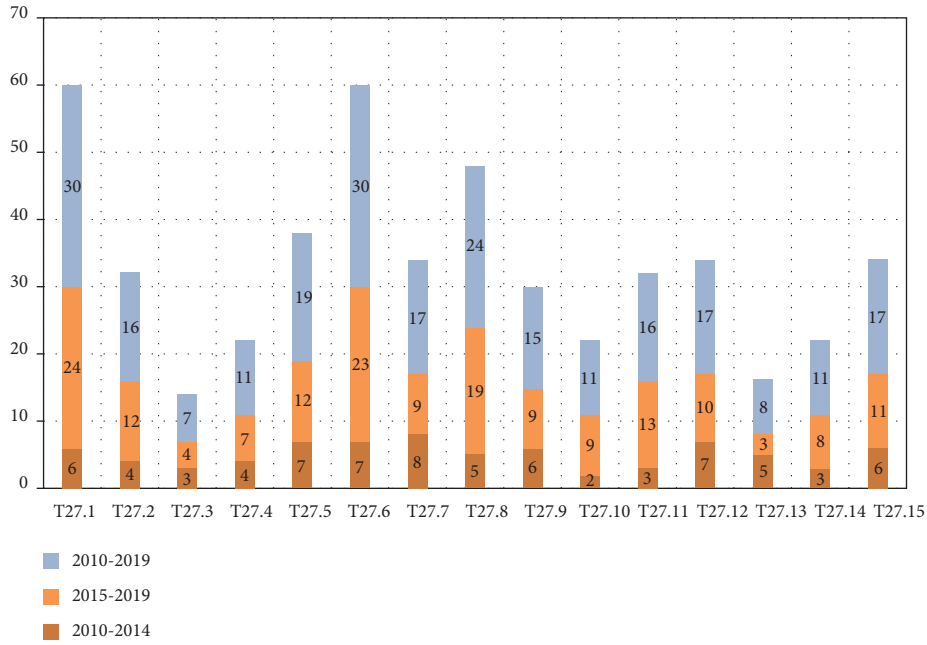
Figure 14: Word clouds generated by TRENDMINER for three topic solution (a)–(c). (a) Word cloud of topic solution 3.1. (b) Word cloud of topic solution 3.2. (c) Word cloud of topic solution 3.3.

patterns, which likewise approves the techniques created to show their semantic association.

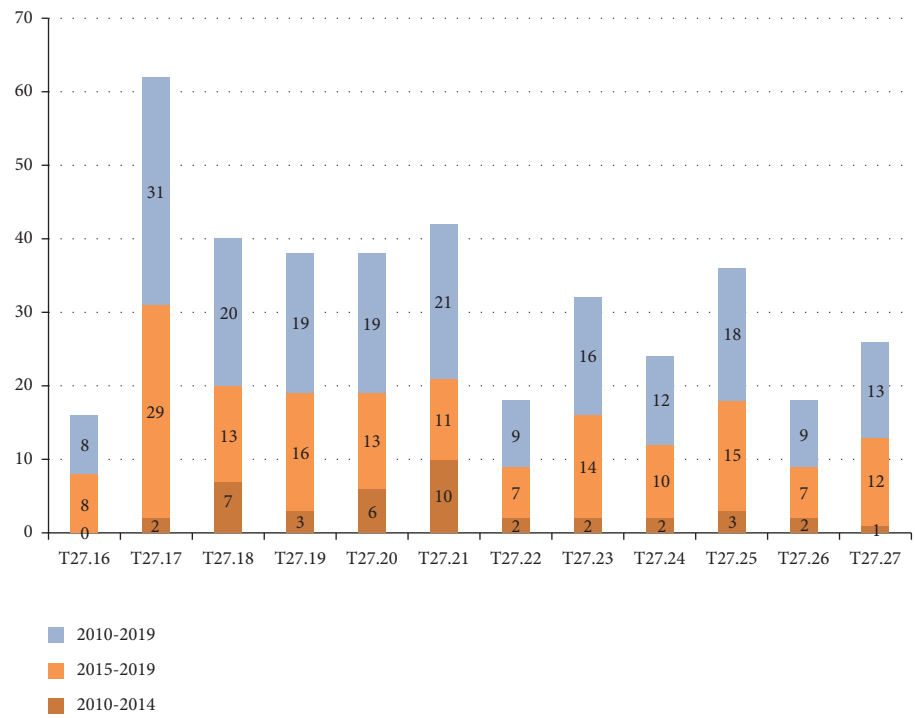
(1) Application Structure Analysis (T3.1). The trends Metadata-Based Study (T27.4) and App Level Features (T27.2) revealed the utilization of metadata. This pattern was found in the system named WHYPER [86], the researchers get to the permissions mentioned by applications' developers and utilized natural language processing (NLP) algorithms to search for sentences in application description that legitimizes the requirement for the mentioned permissions. Similarly, in another work, the study on metadata was accelerated by accounting additional information such as a number of application's screenshots, price, category, title, developer ID, website, and promotional videos. Furthermore, the analysis of application metadata was performed using machine learning algorithms. The trend Application Level features (27.2) unfolds the usage of CPU and memory usage to track malicious applications. In the project named MADAM, running processes, CPU utilization, memory state, Wi-Fi, and Bluetooth of the device were considered to train the k-nearest neighbor algorithm for effective detection [87].

(2) Static Level Monitoring (T3.2). It is the most investigated research area. Figure 10 demonstrates that out of twenty-seven topic solutions, twenty research trends such as Intent Monitoring (T27.15), Type and Model Checking-Based Analysis (T27.5), Memory Mapping (T27.18), Symbolic Execution (T27.9), Interprocedural Control Flow Graph (T27.16), Analysis Based on Network Addresses (T27.1), Program Slicing (T27.25), Context Sensitivity (T27.22), Text-Based Analysis (T27.26), Field Sensitivity (T27.19), Graph-Based Analysis (T27.11), Permission-Based Analysis (T27.21), Data Flow Tracking (T27.6), Dex File Study (T27.7), Object Sensitivity (T27.3), Flow Sensitivity (T27.24), Taint Analysis (T27.27), Hardware Component-Based Inspection (T27.12), Estimation over String Matching (T27.8), and Path Sensitivity (T27.13) mapped to T3.2.

In the topic solution Permission-Based Analysis (T27.21), authorizations played indispensable component for examination of vindictive applications, as most actions require explicit assents remembering the ultimate objective to be accomplished [88]. Permissions are declared in the manifest file and therefore, easy to obtain. Numerous systems, developed in studies [86, 89, 90], use static examination to evaluate the risks of the Android consent system and individual applications.



(a)



(b)

Figure 15: Twenty-seven factor solution during three different time periods (a) and (b).

Another significant research trend emerged as Analysis Based on Network Addresses (T27.1), focused on network addresses. Malware authors make use of network addresses to build communication with command and control (C&C) worker to send the client's classified information. Analysts discovered IP addresses as one of the key static components for investigation [91–93].

Another examination pattern that arose in this space is the Dex record study (T27.7), which played a vital role in understanding the dex files, which are usually cumbersome to interpret by humans. To recognize malevolent code sections, scientists first decompile the dex code into more possible organizations such as gathering, Smali, Dalvik bytecode, source code, container, Jimple, or Java bytecode [94]. This trend can be further relate to numerous articles

and tools deployed by researchers for successful translation such as dexdump [95], Pegasus [96], ded [97], SAAF [98], PScout [89], AppSealer [99], ded/DARE [100], dedexer [90], dex2jar [101], and FlowDroid [102].

The core research area discovered interesting research trends such as Data Flow Tracking (27.6), Interprocedural Control Flow Graph (27.16), and Graph-Based Analysis (27.11). All emerged trends relate to an interesting and pivotal branch in the field of static security mechanisms to identify commandeering vulnerabilities in the Android ecosystem. Data Flow Tracking (T27.6), which deals with tracing out the flow of sensitive information from the device to outside entities at the time application execution [103–107], came out as important and consistent topic. Information stream examination and control stream investigation help in understanding the hazardous usefulness such as protection spillage and communication administrations abuse [95, 108, 109] by tracking the flow of information across different points of execution.

Bytecode control-flow graph investigation recognizes all possible ways that an application can take while it is executed. These deduced trends helped in fostering advance investigation, by creating control flow bytecode graph (CFG) for intraprocedural analysis or between procedural investigation (crossing across various strategies). Creators in [110] formalized the Dalvik bytecode to play out the control stream investigation-based semantic marks to recognize malware applications. The studies [89, 95, 96, 102, 104, 108, 111] leverage the trends Data Flow Tracking (27.6), Interprocedural Control Flow Graph (27.16), and Graph-Based Analysis (27.11).

The trend of Intent Monitoring (T27.15) relates to the concept that intents declared in the application's manifest file are capable enough to leak the data to C&C servers. Intents are the objects which are used to move from one activity to another by making use of widgets in an Android application. Starting an activity, starting a service, and delivering a broadcast are the three fundamental use cases of intents, helps in establishing the communication between components in several ways. This trend was found in popular studies [91, 112]. The former employed numerous machine learning algorithms such as K-means, k-nearest neighbor, and naïve Bayes to analyze the intents, permissions, components, and APIs that were extracted from the manifest file. The latter employed support vector machines to detect malware and achieve a detection rate of 94%. Another trend Hardware Component-Based Inspection (T27.12) reflects the analysis of hardware components listed in an application for static investigation. Researchers in [91] made use of the components declared in the manifest file for analysis. This can be compelling as malicious applications with a specific end-goal demand all the hardware, e.g., camera, GPS, and microphone.

Estimation over String Matching (T27.8) is found as another significant trend in this area, which uncovered the analysis over various strings available in an Android application. Work done by researchers in [113] expressed that it is one of the broadly utilized strategies for recognizing the malware through analyzing the strings, accessible in the Android files. Scientists utilized the Vector Space Model (VSM) [114] and addressed the strings as vectors in a multidimensional space. Besides, scientists utilized distance estimates such as Manhattan distance, Euclidean distance, and Cosine similarity to learn irregularity of the data. The researchers assessed the outcomes over 666 samples of Android applications and accomplished 83.51% accuracy in their tests.

(3) Automatic Malware Analysis (T3.3). Figure 11 demonstrates research trends under T3.3. This core research explored the research trends Pattern Assessment (T27.20), Input Matching (T27.14), Repackaged App Identification (T27.23), Formal Analysis (T27.10), and Machine Learning Approach (T27.17) which were related to automation in identifying Android malware. To gather a predefined set of application features, researchers focus first to analyze application statically or dynamically. Furthermore, build a detection model capable of distinguishing malware and benign applications based on the training dataset. The trend proved as well explored and promising as researchers used numerous combination of different features such as API call sequences, permission request, package information, hardware components, application categories, and network activity to build detection models, as reported in studies [91, 115–118]. Another exploration pattern that arose was Repackaged App Identification (T27.23). Many articles such as [119] related to this trend were published in recent years. DroidMoss [88], Droidsim [120], DNADroid [121], ViewDroid [122], ResDroid [123], and AnDarwin [124] have witnessed to tame the problem of repackaging.

The trend Pattern Assessment (T27.20) uncovered the fact that an attacker can deduce sensitive information of the user by accessing the behavioral pattern of shared resources. The impact of this trend has been seen in a variety of articles [125–129] where side channel communication was compromised to infer confidential input patterns such as PIN, password, or screen taps.

4. Discussion and Potential Future Directions

This section determines that the results obtained from TRENDMINER can be used to answer the research questions stated in Section 1.

4.1. RQ1: Can the Proposed Framework Uncover Leading Researchers within a Research Domain? Figures 7 and 8 present the top journals and leading researchers in the Android security field. Some of the top journal lists include Computer and Security, IEEE Transaction on Information Forensic and Security, Future Generation Computer System, Journal of Information Security and Applications, and Journal of Networks and Computer Applications. Suarez Tangil has a major contribution in the research community who has framed a variety of antimalware techniques such as Alterdroid [130], Dendroid [131], and Droidsieve [132]. A fully automated malware identification mechanism with an appreciable accuracy of 82.93% has been framed by Wang

et al. [133]. Enck et al., who proposed a project named Taintdroid [83], are top leading researchers in this field. He had developed an effective model for tracking sensitive information leakage in third-party applications. On top of this, many other dynamic analysis tools such as Andrubis [134] and Droidbox [135] were deployed. He was first to perform on-device malware assessment in which authors defined a set of rules to identify dangerous permissions granted before installing the application, by the security service known as Kirin [136]. To detect kernel-level attacks, Yan and Yin presented a project named Droidscope [137], which is a unique method of dynamic analysis by keeping its process out of emulator and was able to achieve promising results. Faruki et al. [138] proposed a methodology called Androsimilar which produces marks by extricating measurably powerful components, to identify noxious Android applications. Proposed strategy was powerful against code jumbling and repackaging methods that will in general engender concealed variations of known malware by avoiding AV signatures.

4.2. RQ2: Are Those Frameworks Robust Enough to Determine the Most Investigated Research Areas? The consequences of the examination showed that Static Level Monitoring (T3.2) had been end up being the most generally researched point in Android malware investigation and location. The strategies utilized under Static Level Monitoring (T3.2) analyses the code without running the application on an Android emulator or gadget. The upside of static investigation is that the expense of calculation is low, less dreary, and low asset use. Figure 16 reveals that most of the trends tend to fall in the topic solution Static Level Monitoring (T3.2). Out of the 20 research trends, eleven such trends have shown a significant rise in time frame 2 (2015–2019) than time frame 1 (2010–2014). The rate of change varies from 0.82% to 4.01%. Nine such trends showed a downfall in time frame 2. Examination under this work uncovered that studies identified with static level observing significantly center around network addresses, information stream, control stream, string coordinating, consents, dex documents, setting, and purposes.

Static level monitoring emerged as an important technique to accomplish various security concerns such as detecting private data leaks, detecting component hijacking or intent injection, building frameworks for intercomponent vulnerabilities and content provider-based vulnerabilities, dangerous permissions used by malicious applications, energy consumption concerns by Android applications, comparing Android applications for clone detection, automatic testing by generating test cases, and checking the correctness of the Android application through code verification. On further investigation, it was found that there are various tools available for static monitoring, such as Soot, Dex2jar, Dexdump, Dedexer, Ded, Dare, and WALA. Soot is the most adopted support tool for static monitoring, and Jimple is the widely used intermediate representation (IR) format for the further analysis of Android applications. The trend line in Figure 16 illustrates that specific research trends orient towards sensitivities. Sensitivities maximize the precision and recall of static monitoring. The research trends Field Sensitivity (T27.19), Context Sensitivity (T27.22), and Flow Sensitivity (T27.24) are primarily taken into account by the Android research community. Other research trends, such as Path Sensitivity (T27.13) and Object Sensitivity (T27.3), have not gained much attention from the researchers. The trend line also revealed that the trend Taint Analysis (T27.27) widely used in data tracking emerged as the most applied technique in static monitoring.

4.3. RQ3: Would the Proposed Framework Reveal How the Focus of Topics within Each Core Research Area Has Changed over Time? In this study, two time frames 2010–2014 and 2015–2019 were used to maintain valid interpretations and comparisons among the topics. Table 9 shows the focus of major topics within each core research area changed over time. It depicts the paradigm shift from the time window from 2010–2014 to 2015–2019. The study made the following observations:

(a) Machine learning approaches are demonstrated to be compelling among other serious methodologies in the location of Android malware. These methodologies are all around investigated and promising during the period 2015–2019.

(b) Detection of piggybacked applications utilized delicate chart investigation/information followed by use of AI calculations, during the period 2015–2019.

(c) Permissions have discovered quite possibly the most utilized static elements to identify Android malware application. The trend was popular during 2010–2019. Some specific permissions are declared in the manifest file to activate certain events in an Android ecosystem.

(d) Static analysis is largely performed by the researchers to address security and privacy issues, due to its ease of implementation. However, static analysis is vulnerable to stealthy techniques such as encryption and native code, resulting in a downfall in the usage of pure static solutions. Nevertheless, it is still popular in the research community.

(e) Taint analysis is a widely applied technique in publications. It is the kind of information flow analysis in which objects are tainted and tracked using data flow analysis.

(f) During 2015–2019, one research trend emerged as "Analysis Based on Network Addresses" (T27.1), focused on network addresses. Malware authors make use of network addresses to build communication with the command and control (C&C) server to send the user's personal confidential data. Researchers found network addresses as one of the key static features for analysis.

(g) The trend "Text-Based Analysis" (T27.26) relies on extracting the critical phrases and keywords, for example, sensitive APIs and permission for the

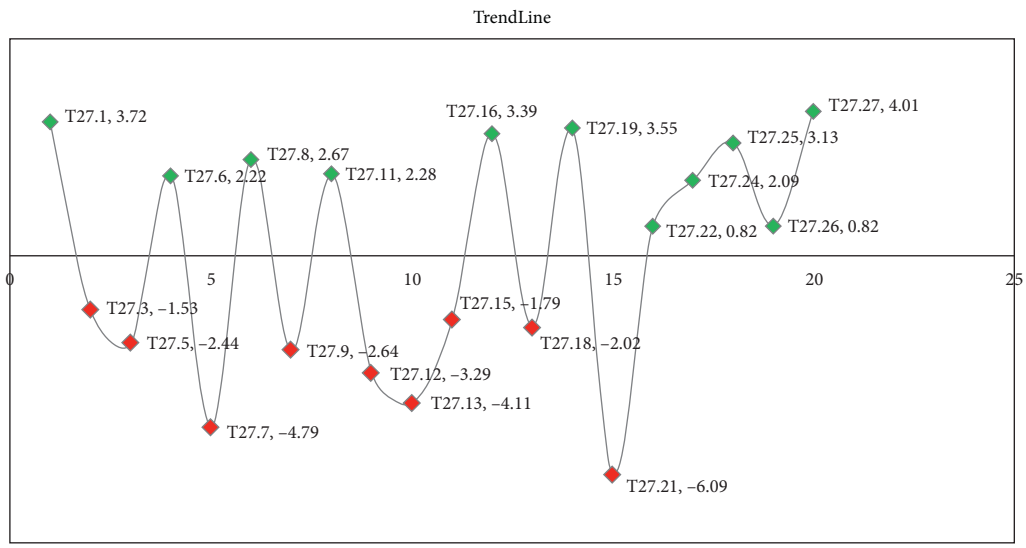


Figure 16: Impact of research trends during time frame 2015–2019.

Table 9: Focus of topics changed over time from 2010 to 2019.

Topic no.	Label	2010–2014	Impact in time frame 1 (%)	2015–2019	Impact in time frame 2 (%)	+/−
T27.1	Analysis Based on Network Addresses	6	6.45	24	10.17	+3.72
T27.3	Object Sensitivity	3	3.23	4	1.69	−1.53
T27.5	Type and Model Checking-Based Analysis	7	7.53	12	5.08	−2.44
T27.6	Data Flow Tracking	7	7.53	23	9.75	+2.22
T27.7	Dex File Study	8	8.60	9	3.81	−4.79
T27.8	Estimation over String Matching	5	5.38	19	8.05	+2.67
T27.9	Symbolic Execution	6	6.45	9	3.81	−2.64
T27.11	Graph-Based Analysis	3	3.23	13	5.51	+2.28
T27.12	Hardware Component-Based Inspection	7	7.53	10	4.24	−3.29
T27.13	Path Sensitivity	5	5.38	3	1.27	−4.11
T27.15	Intent Monitoring	6	6.45	11	4.66	−1.79
T27.16	Interprocedural Control Flow Graph	0	0.00	8	3.39	+3.39
T27.18	Memory Mapping	7	7.53	13	5.51	−2.02
T27.19	Field Sensitivity	3	3.23	16	6.78	+3.55
T27.21	Permission-Based Analysis	10	10.75	11	4.66	−6.09
T27.22	Context Sensitivity	2	2.15	7	2.97	+0.82
T27.24	Flow Sensitivity	2	2.15	10	4.24	+2.09
T27.25	Program Slicing	3	3.23	15	6.36	+3.13
T27.26	Text-Based Analysis	2	2.15	7	2.97	+0.82
T27.27	Taint Analysis	1	1.08	12	5.08	+4.01
T27.10	Formal Analysis	2	13.33	9	12.33	−1.00
T27.14	Input Matching	3	20	8	10.96	−9.04
T27.17	Machine Learning Approach	2	13.33	29	39.73	+26.39
T27.20	Pattern Assessment	6	40	13	17.81	−22.19
T27.23	Repackaged App Identification	2	13.33	14	19.18	+5.84
T27.2	App Level Features	4	50	12	63.16	13.16
T27.4	Metadata-Based Study	4	50	7	36.84	−13.16

analysis. This trend became popular during the time frame of 2015–2019.

(h) The trend "Symbolic Execution" (T27.9) showed a downfall in the time frame 2015–2019. It deals with generating all possible program inputs to explore all conditional branches inside the path. This process could be time-consuming and hence became less popular among the research communities during 2015–2019.

(i) Another research trend that emerged was "Repackaged App Identification" (T27.23). Repackaging is one of the popular techniques being employed by malware authors to generate fraudulent repackaged applications. Many articles related to this trend were published during the time frame 2015–2019.

(j) The trend "Metadata-Based Study" (T27.2) uncovered the utilization of metadata to identify and dissect Android malware applications. Metadata involves required authorizations, depiction, form, last refreshed, rating, number of establishments, and engineer data. This pattern encounters a defeat during 2015–2019.

(k) Table 9 revealed that the trend "Program Slicing" (T27.25) had gained momentum during 2015–2019. The trend "Program Slicing" (T27.25) specifies the technique by focusing on selected aspects of semantics for simplifying the programs. Slicing avoids those parts of the program that may not have caused the malicious behavior, instead focus attention on only those parts of programs that may contain malicious behavior. This technique tends to reduce the set of program behavior and hence became trending during 2015–2019.

(l) The trend "Field Sensitivity" (27.19) appears to be the most considered among all the sensitivities, depicted in Table 9. It may be due to the reason since Android apps are written in Java, an Object-Oriented language where object fields are pervasively used to hold data. Research trends such as "Context Sensitivity" (T27.22) and "Flow Sensitivity" (T27.24) are also largely taken into account. The least considered sensitivity is "Path Sensitivity" (T27.13) and Object Sensitivity (T27.3); probably, it is because of the scalability issues that it raises.

(m) The trends "Type and Model Checking-Based Analysis" (T27.5) showed a sudden fall during 2015–2019. When an Android application is developed for some task, it is common to define a certain set of properties that the application must satisfy. Model checking helps to ensure that the given system has met given specification or correctness properties. Type checking ensures that the given program is type-safe by keeping the possibility of type errors (e.g., applying integer operations on float numbers) to a minimum.

(n) Another research trend that emerged was the "Dex File Study" (T27.7), which played a vital role in understanding the dex files was popular during the time frame 2010–2014. Dex code is usually cumbersome to interpret by humans and therefore shows a downfall during 2015–2019.

(o) In the research trend "Permission-Based Analysis" (T27.21), permissions are declared in the manifest file and, therefore, easy to obtain, and that could be the reason for its popularity among researchers during 2010–2014. However, examining only permissions is not useful in detecting malicious applications. Therefore, this trend experiences a downfall during 2015–2019.

(p) Interesting research trends such as "Data Flow Tracking" (27.6), "Interprocedural Control Flow Graph" (27.16), and "Graph-Based Analysis" (27.11) are the data structures for the analysis. Data flow analysis and control flow analysis help in understanding unsafe functionality such as privacy spillage and telephony services misuse by tracking the flow of information across different points of execution. The advantage is that bytecode control-flow graph investigation recognizes all possible ways that an application can take while it is executed and hence popular during 2015–2019.

(q) The trend of "Intent Monitoring" (T27.15) relates to the concept that intents declared in the application's manifest file are capable enough to leak the data to C&C servers. Intents are the objects which are used to move from one activity to another by making use of widgets in an Android application. Starting an activity, starting a service, and delivering a broadcast are the three fundamental use cases of intents, which helps in establishing the communication between components in several ways. It was more popular in the time frame 2010–2014 than the time frame 2015–2019.

(r) Another trend, "Hardware Component-Based Inspection" (T27.12), reflects the analysis of hardware components listed in an application for static investigation. It can be compelling as malicious applications with a specific end-goal demand all the hardware, e.g., camera, GPS, and microphone. This trend gradually decreases in the time frame of 2015–2019.

(s) Another significant trend, "Estimation over String Matching" (T27.8), uncovered the analysis over various strings available in an Android application. Its impact is slightly more during 2015–2019.

(t) The trend "Application Level Features" (27.4) unfolds the usage of CPU and memory usage to track malicious applications. It remains trending during 2015–2019.

4.4. RQ4: Can It Unfold the Future Directions within the Research Domain of Choice? Many obstacles are set forth by Android malware, which needs to be carefully addressed after being thoroughly observed. Based on the results of

TRENDMINER, with no doubt, it is evident that Android security has gotten a ton of consideration in recently published literature. Perhaps, it is mainly due to the ubiquity of Android as a famous operating system in the community. Significant patterns are observed in the previous decade, as reflected by the aftereffects of this writing survey. Hence, based on the results of the TRENDMINER, a few recommendations are made that are discussed as follows:

(a) Mapping of API usage with permissions to achieve more fine-grained results: API calls are used to communicate and transfer sensitive information over the network. Malware families such as Fakeinst, Opfake, and Smsreg make use of API calls such as sendSMS() and readSMS(), which implies that collected information may be sent by SMS. There is an urgent need to deeply analyze the API calling patterns and what permissions these APIs demand [139].

(b) Complications in static analysis: static analysis techniques are incapable when applications are made using camouflage techniques [39, 139–143]. Static analysis also leads to a large number of false positives [7, 144].

(c) Evolution of intelligent malware: applications tend to use techniques such as rooting, antidebugging, code obfuscation, and kernel-level features to dodge the detection process [145, 146]. Despite this, most of the approaches still implement emulators. Limited efforts are made to curtail remote triggering. It enhances the stealthiness of malware by allowing malware authors to trigger and execute malware whenever they want [147].

(d) Development of nonintuitive features for robust malware analysis and detection: static and dynamic features need to be explored to the next level to characterize the behavior of an application [146] better. Attackers repackage the legitimate app to insert the malicious snippet and distribute it via stores [88].

(e) Need of automation in malware classification: development of semisupervised approaches to detect the malicious applications [146, 148] and faster detection and classification of malware families is required [141]. Also, the features and characteristics of a family that can be used to classify malware to a particular family have been less discussed among the research communities [7].

(f) Hindering the effectiveness of dynamic analysis: computation time and resource constraints are the major reasons for the hindered performance of dynamic analysis [7, 39, 140, 143]. To ensure that an application had triggered all its malicious behavior (all execution paths traversed) during dynamic analysis is a matter of concern [141, 142, 144].

(g) Limited availability of datasets: limited availability of ransomware datasets and lack of understanding of smart tactics limits the efficacy of detection mechanisms [149]. Generally, researchers download the samples from VirusTotal [150].

(h) Low-precision prediction mechanisms: the biggest challenge being faced by researchers is the high rate of false alarms in predicting ransomware. Most of the present techniques produce a large amount of false positive and false negative alarms, which affects the accuracy of detection mechanisms. There is a need for a cutting edge methodology to produce fewer false alarms [149].

The study uncovered that methodologies of examining malware incorporate static examination and dynamic investigation or perhaps a blend of both. The static examination essentially centers around dismantling the code, trailed by manual examination to look for the pernicious examples in the code. On the other hand, dynamic investigation executes the code in the virtual platform and breaks down its execution follow to notice the noxious conduct of an application. The static examination helps follow unique and full execution ways; subsequently, it gives total code inclusion; however, at last it experiences code obscurity. The application must be decoded first to perform static investigation. The issues of obstinate intricacy ruin the examination. Dynamic examination is more productive and need not bother with the executable to be unloaded or unscrambled. The dubious application is checked in a controlled arrangements. This cycle is time and asset devouring. It additionally raises adaptability issues. Besides, some malevolent conduct may be unseen on the grounds that the environment does not fulfill the setting off conditions. Besides, malware creators utilize mechanization innovation to produce a colossal measure of new malware variations, accordingly representing a major test to malware experts. The current situation with the-workmanship requests the combination of existing crude strategies with valuable methods to accomplish a powerful arrangement. The yield of TRENDMINER proposes that strengthening strategies ought to be utilized to supplement the arrangement of quickly developing Android malware families. Beneficial methods can end up being viable in deciding strange current vindictive conduct or security weaknesses. In view of the assortment of information got by this investigation, a plan for designing a cutting edge environment has been imagined for the characterization of Android malware families, as examined in the next section.

5. Towards Engineering a Visualization-Based Solution

Malware is developing quickly which is a result of the ability of malware creators to change little pieces of the first source code to produce new malware variations. A malware variation can be imagined as a grayscale image. A picture can catch even little changes. Thus, in the current work, a perception structure is proposed to decrease the impact of obscurity by changing the malware's noninstinctive components into unique finger impression images followed by the arrangement of Android malware families. The proposed methodology which is known as the SWAYAM (Stop WAY for Android Malware) system is shown in Figure 17.

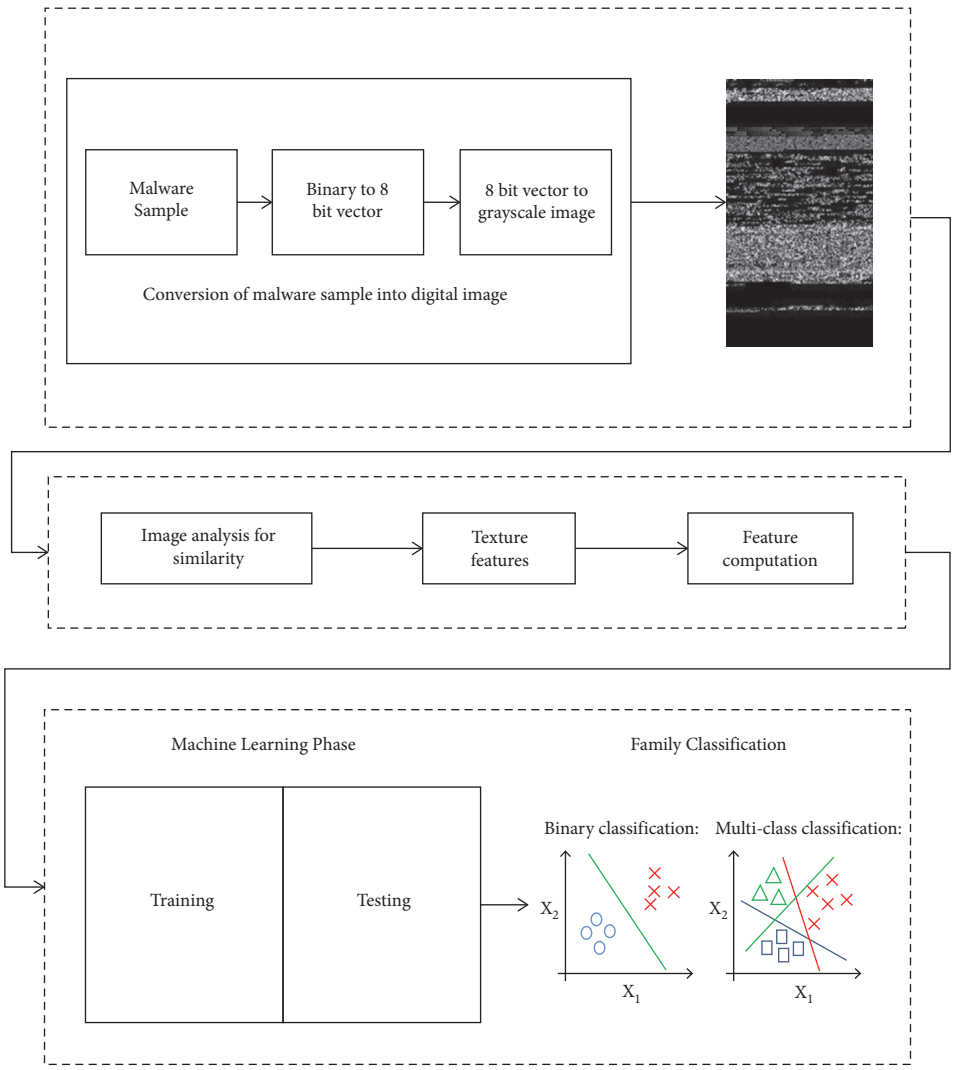


Figure 17: The proposed SWAYAM (Stop WAY for Android Malware) system.

5.1. Module I. This module deals with converting the malware samples into digital images. The malware binaries are first converted into 8-bit vectors and then converted into grayscale images. The overall structure of grayscale images is composed of various sections. Each section has a fixed width, but height is varied according to the file size. In a nutshell, malware samples tend to be represented as images and there is a strong propensity that malware variants from the same family form similar and visual implications [151]. On the other hand, malware samples from different families show dissimilar structural and visual implications.

5.2. Module II. Once the images are converted into digital images, the next step is to extract the features out of the images. Features play a vital role in classifying malware samples to a particular family. Various image descriptors such as Global Image deScripTors (GIST), Gray Level Co-occurrence Matrix-based (GLCM), and Local Binary Pattern (LBP) are available to extract the features from the images and thus formed a feature vector. Texture patterns, intensity, color patterns, and frequencies in images constitute the image features of the samples. Euclidean distance or standard deviation can be used to measure the distance in feature space [152].

5.3. Module III. Further machine learning algorithms or neural networks are employed over feature vectors to identify the family of a sample. For instance, in the KNN approach, a sample is classified to family f1 if it has k-nearest neighbors belonging to family f1. It is to be noted that many solutions leveraging machine learning and big data techniques are appearing to develop malware detection models [153–155]. Computer vision techniques have been becoming popular among the research communities to detect and classify malware applications [156, 157].

6. Limitation of the Study

This study encountered a few issues that may have arisen during the collection of the literature dataset on Android

security. It depends upon certain factors, for example, the type of queries and sources used while preparing the literature dataset. To discover the appropriate publications, the articles were selected using "malware" OR "vulnerability" OR "security" OR "privacy" OR "monitoring" OR "application" OR "smartphone" OR "android" OR "virus" OR "static" OR "dynamic" OR "detection" OR "data flow" as search keywords. The prominent databases which were leftover during the automated search were also browsed to get the influenced publication in the area. Relevant papers were filtered using inclusion and exclusion criteria on the search results to limit the purpose of the current study. Nonetheless, it may be possible that a few significant publications may have been left during the process.

TRENDMINER is backed by the goodness of the Latent Semantic Analysis (LSA) technique. LSA being an unsupervised way of uncovering synonyms improves the vector space model. However, the number of topic solutions cannot be decided statistically. To alleviate this situation, the value for an optimal number of topic solutions was decided after having intensive discussions with an expert. Ultimately, this work deduced that the process of topic labeling was purely based on human judgment, which may lead to subjective bias as well.

There might be impediments identified with the speculation of the outcomes. A stepwise procedure was followed to infer the core research areas and research trends. The procedure included literature collection, preprocessing of the dataset, generation of TF-IDF matrix, truncated SVD, and topic labeling. Every step in the algorithm tends to influence the results. For instance, the outcomes will be influenced if the dataset used in this study is modified to a composition of only titles or full-length articles.

Having done LSA representation of some documents, a new document cannot be just added to this collection. A new document, hence, can only be added incrementally. It fails to capture the elements of the new documents added. Hence, the performance of LSA degrades on the addition of new documents, allowing recomputation.

7. Conclusion

One of the key inspirations of the work was that the conventional manual literature reviews are often not ready to exploit huge literature because of human obstructions in time and insight. Hence, this study proposed another literature review method to deal with this challenge. This study unveiled a framework called the SEAR framework, which can perform subjective and quantitative investigation over enormous literature. It is an adaptable and versatile framework to draw information-driven investigation and conceptualize the advancement of inclining research measurements in any field of literature. The SEAR framework utilizes the linear combination of information modeling technique, i.e., LSA followed by the K-means clustering algorithm, which enables connections and groupings to be recognized that are usually missed by manual techniques constituting human interpretations. Machine learning techniques have reduced the manual effort to a great extent in determining the document to its closest topic.

TRENDMINER is designed as the use case of the SEAR framework. To exhibit the utility and use of TRENDMINER, a wide body of literature on the Android security field was utilized as the contextual investigation. The framework takes the contribution of 444 abstracts of research articles distributed during the period 2010–2019. This study identifies three core research areas and twenty-seven research trends as outcomes. Results demonstrated that specific research patterns have stayed reliable over the examined time frame. Taxonomy and future research directions in the field of Android security have been provided in this study. Time trend plots for each factor solution have been discussed. Some research trends have developed while a couple has likewise declined. TRENDMINER amplifies the utility and commitment by proposing potential future research directions in developmental research to mitigate human predispositions. This study also stresses answering the research questions framed with respect to the technique being employed and the dataset chosen. This paper additionally exhibited general suggestions to help new researchers to comprehend the idea of Android security research and assess their regions of interest for their latent capacity research alongside the related research pattern.

This examination additionally sets up an objective and observational establishment for future directions about the structure and analytical decomposition of Android security research. The particular research area and trends uncovered in this work can engage future research dimensions, which can be utilized by the research scientists and industry. Furthermore, researchers can pick at least one research area and make another investigation with the equivalent or another approach. Nonetheless, other factual factor investigation strategies can apply to this exploration. For future work, the researchers can apply a similar technique to a different comparable dataset to see the proclivity and decent variety of core research areas and trends inside related articles. To increase the application areas of this research, the SEAR framework can be enhanced by building a dynamic query system on the same or different corpus by applying deep learning models.

8. Practical Implications and Future Research Directions

This manuscript exhibits a panoramic view of the Android security field. The study has certain interesting practical implications. First, the research areas and trends uncovered in this work can engage future research dimensions, which can be utilized by the new research scientists and industry. The analysis obtained from the study can assist them to understand the diversity and depth of the Android security field. Second, the academic universities can enhance their teaching content and students' motivation by revising the curriculum to focus more on research activities related to the Android security field.

Third, perspectives drawn from the research will help the editors of the esteemed journals to plan the special sessions on Android malware research topics such as static analysis of Android applications, security and privacy for IoT and

multimedia devices, application-focused threats, new frontiers in Android malware analysis and detection, cryptojacking, component-based Android malware analysis, deep learning for Android malware classification, deep learning for digital forensics, and cybersecurity. There are avenues for future research which are discussed as follows.

8.1. Ranking Permissions for Android Malware Analysis and Detection. Using too many features for Android malware analysis and detection is a cumbersome task. Permissions as a special feature of the Android ecosystem are present in the manifest.xml file of the Android file structure. Permissions are needed to perform the application-sensitive operations. They are embedded in the manifest.xml file in the form of text. They play a vital role in detecting the suspicious application running on an Android device. Some permissions which malware authors use to exploit the sensitive information from the device are *access_coarse_location*, *access_ne_location*, *access_network_state*, *access_wi _state*, *battery_stats*, *answer_phone_calls*, *bind_carrier_messaging _service*, *read_contacts*, *read_call_log*, *read_phone_state*, *read_external_storage*, *read_sms*, *record_audio*, *request_install_packages*, *read_calendar*, *bluetooth_privileged*, *read_history_bookmarks*, and many more. The most important permissions in the malicious dataset can be identified using a technique called Term Frequency Inverse Document Frequency (TF-IDF) which can later lead to the discovery of malicious applications. It would help in maintaining an application-permission matrix that would describe the frequency of permissions that occur in the collection of malicious applications. TF-IDF assigns the permission value to each permission and calculates the sensitive value of each application by utilizing its weighing formula as discussed in this work. Furthermore, machine learning algorithms may be deployed to perform the detection or classification of Android malware applications.

8.2. Crowdsourced User Reviews at Application Stores. The suspicious application can also be identified by evaluating the user reviews at the application stores. The feedbacks of the users are vital as they tend to write reviews about the particular application based on their real-time usage and experience. The security firms cannot ignore the reviews whether they are positive or negative. The user reviews are expressed for various purposes such as functionality, UI (user interface)/design, battery consumption report, and other security issues of an application. Furthermore, the security issues in the application are broadly classified into four categories: malware code injected into the application for monetary benefits, spamming, information leakage, and use of overprivileged permissions in the application. Latent Semantic Analysis can be applied to crowdsourced user reviews to discover security-related issues of the application. At the initial step, relevant reviews can be filtered out from the noisy crowdsourced reviews by applying the preprocessing techniques as employed in this manuscript. The relevant terms in the reviews may be then mapped with Android API documentation to form the clusters based on the components addressed in the review.

Assume the user review for the cricket game application, "Whenever I open this CRC League application, it automatically clicks my photograph and also deducts one dollar from my account. I also received the message that says Thank you for subscribing to IOIO service." After reading this review, one undoubtedly thinks that this is a malicious application. There may be hundreds of reviews related to this context. The data-driven analysis here can understand the text structure, words, and the topic discussed in the review. This review reflects that this application accesses camera, sends the SMS, and deducts the amount from the user account. One may think that a cricket game can never be made for performing these types of sensitive operations. This scenario only depicts the security issue of an application. Therefore, the semantics of the review can be discovered to flag these applications as suspicious using LSA.

8.3. Preserving the Proprietary Rights of the Android Developers. Repackaging is an open issue in the Android malware detection and analysis field. Using this technique, malware authors first download the legitimate application from the application stores and then extract all files and folders of the application. After the extraction process, they inject the malicious code or segment into the application and upload the same on other application stores. They also entice users to download that malicious application by performing social engineering activities. Innocent users not aware of this fact get trapped and download the malicious version of the legitimate application. In this way, the malware penetrates the phone and their device gets compromised. Repackaging thus opens the other dimensions for the malware authors to generate malicious clone or plagiarized versions of the legitimate applications. In a nutshell, the proprietary rights of developers are widely exploited and abused among malware authors to create clone Android malware variants of legitimate applications. Furthermore, they also deploy the evasion technique to dodge the detection process. In this scenario, LSA can be used to infer the semantics from the corpus of source code files. The degree of similarity can be measured by comparing the code segments of the source code files.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research work was self-funded.

References

[1] A. White and K. Schmidt, "Systematic literature reviews," *Complementary Therapies in Medicine*, vol. 13, no. 1, pp. 54–60, 2005.

[2] D. Delen and M. D. Crossland, "Seeding the survey and analysis of research literature with text mining," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1707–1720, 2008.

[3] N. Evangelopoulos, X. Zhang, and V. R. Prybutok, "Latent semantic analysis: five methodological recommendations," *European Journal of Information Systems*, vol. 21, no. 1, pp. 70–86, 2012.

[4] S. Lee, J. Song, and Y. Kim, "An empirical comparison of four text mining methods," *Journal of Computer Information Systems*, vol. 51, no. 1, pp. 1–10, 2010.

[5] S. Sehra, J. Singh, and H. Rai, "Using latent semantic analysis to identify research trends in openstreetmap," *ISPRS International Journal of Geo-Information*, vol. 6, no. 7, p. 195, 2017.

[6] M. Yalcinkaya and V. Singh, "Patterns and trends in building information modeling (BIM) research: a latent semantic analysis," *Automation in Construction*, vol. 59, pp. 68–80, 2015.

[7] N. Xie, X. Wang, W. Wang, and J. Liu, "Fingerprinting android malware families," *Frontiers of Computer Science*, vol. 13, no. 3, pp. 637–646, 2019.

[8] M. Becher, F. C. Freiling, J. Hoffmann, T. Holz, S. Uellenbeck, and C. Wolf, "Mobile security catching up? Revealing the nuts and bolts of the security of mobile devices," in *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, IEEE, Oakland, CA, USA, 2011.

[9] W. Enck, "Defending users against smartphone apps: techniques and future directions," in *Proceedings of the 2011 International Conference on Information Systems Security*, 2011.

[10] P. Faruki, A. Bharmal, V. Laxmi et al., "Android security: a survey of issues, malware penetration, and defenses," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 998–1022, 2014.

[11] G. Suarez-Tangil, J. E. Tapiador, P. Peris-Lopez, and A. Ribagorda, "Evolution, detection and analysis of malware for smart devices," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 961–987, 2013.

[12] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, 2009.

[13] L. See, P. Mooney, G. Foody et al., "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information," *ISPRS International Journal of Geo-Information*, vol. 5, no. 5, p. 55, 2016.

[14] A. Kundu, V. Jain, S. Kumar, and C. Chandra, "A journey from normative to behavioral operations in supply chain management: a review using latent semantic analysis," *Expert Systems with Applications*, vol. 42, no. 2, pp. 796–809, 2015.

[15] E. Altszyler, S. Ribeiro, M. Sigman, and D. Fernández Slezak, "The interpretation of dream meaning: resolving ambiguity using latent semantic analysis in a small corpus of text," *Consciousness and Cognition*, vol. 56, pp. 178–187, 2017.

[16] A. Balahur, R. Mihalcea, and A. Montoyo, "Computational approaches to subjectivity and sentiment analysis: present and envisaged methods and applications," *Computer Speech & Language*, vol. 28, no. 1, pp. 1–6, 2014.

[17] J. N. De Boer, A. E. Voppel, M. J. H. Begemann, H. G. Schnack, F. Wijnen, and I. E. C. Sommer, "Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 93, pp. 85–92, 2018.

[18] O. B. Driss, S. Mellouli, and Z. Trabelsi, "From citizens to government policy-makers: social media data analysis," *Government Information Quarterly*, vol. 36, pp. 560–570, 2019.

[19] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems with Applications*, vol. 57, pp. 1–11, 2016.

[20] G. Gao, Y.-S. Liu, P. Lin, M. Wang, M. Gu, and J.-H. Yong, "Bimtag: concept-based automatic semantic annotation of online BIM product resources," *Advanced Engineering Informatics*, vol. 31, pp. 48–61, 2017.

[21] J. Guan, A. S. Manikas, and L. H. Boyd, "The international journal of production research at 55: a content-driven review and analysis," *International Journal of Production Research*, vol. 57, no. 15-16, pp. 4654–4666, 2019.

[22] P. D. Hutchison, R. J. Daigle, and B. George, "Application of latent semantic analysis in AIS academic research," *International Journal of Accounting Information Systems*, vol. 31, pp. 83–96, 2018.

[23] H. Kim, H. Lee, and J. Seo, "A reliable FAQ retrieval system using a query log classification technique based on latent semantic analysis," *Information Processing & Management*, vol. 43, no. 2, pp. 420–430, 2007.

[24] S. S. Kulkarni, U. M. Apte, and N. E. Evangelopoulos, "The use of latent semantic analysis in operations management research," *Decision Sciences*, vol. 45, no. 5, pp. 971–994, 2014.

[25] X. Lin, Y. Li, and X. Wang, "Social commerce research: definition, research themes and the trends," *International Journal of Information Management*, vol. 37, no. 3, pp. 190–201, 2017.

[26] O. Müller, T. Schmiedel, E. Gorbacheva, and J. Vom Brocke, "Towards a typology of business process management professionals: identifying patterns of competences through latent semantic analysis," *Enterprise Information Systems*, vol. 10, no. 1, pp. 50–80, 2016.

[27] G. Pilato and E. D'Avanzo, "Data-driven social mood analysis through the conceptualization of emotional fingerprints," *Procedia Computer Science*, vol. 123, pp. 360–365, 2018.

[28] Y. Tonta and H. R. Darvish, "Diffusion of latent semantic analysis as a research tool: a social network analysis approach," *Journal of Informetrics*, vol. 4, no. 2, pp. 166–174, 2010.

[29] C.-P. Wei, C. C. Yang, and C.-M. Lin, "A latent semantic indexing-based approach to multilingual document clustering," *Decision Support Systems*, vol. 45, no. 3, pp. 606–620, 2008.

[30] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V- LSA) for topic modeling: a study on blockchain technology trend analysis," *Expert Systems with Applications*, vol. 152, no. 113, p. 401, 2020.

[31] G. Jorge-Botana, R. Olmos, and J. M. Luzón, "Bridging the theoretical gap between semantic representation models without the pressure of a ranking: some lessons learnt from LSA," *Cognitive Processing*, vol. 21, no. 1, pp. 1–21, 2020.

[32] A. Hassani, A. Iranmanesh, and N. Mansouri, "Text mining using nonnegative matrix factorization and latent semantic analysis," *Neural Computing and Applications*, Springer, Berlin, Germany, 2021.

[33] X. Ren and M. N. Coutanche, "Sleep reduces the semantic coherence of memory recall: an application of latent semantic analysis to investigate memory reconstruction," *Psychonomic Bulletin & Review*, vol. 28, pp. 1336–1343, 2021.

[34] S. Gowthami and R. Harikumar, "Conventional neural network for blind image blur correction using latent semantics," *Soft Computing*, vol. 24, pp. 15223–15237, 2020.

[35] F. Sastre, A. Velazquez, L. S. de Leon, J. Montanes, and J. Rodrigo, "Method to solve redundant inverse problems based on a latent semantic analysis approach. application to an aerojet engine," *Aerospace Science and Technology*, vol. 102, Article ID 105854, 2020.

[36] N. Evangelopoulos and S. Y. Amirkiaee, "Extracting LSA topics as features for text classifiers across different knowledge domains," *Quality & Quantity*, vol. 54, no. 1, pp. 249–261, 2020.

[37] C. Shen and J. Ho, "Technology-enhanced learning in higher education: a bibliometric analysis with latent semantic approach," *Computers in Human Behavior*, vol. 104, Article ID 106177, 2020.

[38] A. A. Wagire, A. Rathore, and R. Jain, "Analysis and synthesis of industry 4.0 research landscape," *Journal of Manufacturing Technology Management*, vol. 31, 2019.

[39] M. F. A. Razak, N. B. Anuar, R. Salleh, and A. Firdaus, "The rise of "malware": bibliometric analysis of malware study," *Journal of Network and Computer Applications*, vol. 75, pp. 58–76, 2016.

[40] K. Liu, S. Xu, G. Xu, M. Zhang, D. Sun, and H. Liu, "A review of android malware detection approaches based on machine learning," *IEEE Access*, vol. 8, pp. 124579–124607, 2020.

[41] S. R. T. Mat, M. F. Ab Razak, M. N. M. Kahar, J. M. Arif, S. Mohamad, and A. Firdaus, "Towards a systematic description of the field using bibliometric analysis: malware evolution," *Scientometrics*, vol. 126, no. 3, pp. 2013–2055, 2021.

[42] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on big data in marketing: a text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1–7, 2018.

[43] S. Kavvadias, G. Drosatos, and E. Kaldoudi, "Supporting topic modeling and trends analysis in biomedical literature," *Journal of Biomedical Informatics*, vol. 110, Article ID 103574, 2020.

[44] M. Mustak, J. Salminen, L. Plé, and J. Wirtz, "Artificial intelligence in marketing: topic modeling, scientometric analysis, and research agenda," *Journal of Business Research*, vol. 124, pp. 389–404, 2021.

[45] J. Rumbut, H. Fang, and H. Wang, "Topic modeling for systematic review of visual analytics in incomplete longitudinal behavioral trial data," *Smart Health*, vol. 18, Article ID 100142, 2020.

[46] J. An, K. Kim, L. Mortara, and S. Lee, "Deriving technology intelligence from patents: preposition-based semantic analysis," *Journal of Informetrics*, vol. 12, no. 1, pp. 217–236, 2018.

[47] J. L. Hurtado, A. Agarwal, and X. Zhu, "Topic discovery and future trend forecasting for texts," *Journal of Big Data*, vol. 3, no. 1, p. 7, 2016.

[48] N. E. Evangelopoulos, "Latent semantic analysis," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 4, no. 6, pp. 683–692, 2013.

[49] K. R. Larsen and D. E. Monarchi, "A mathematical approach to categorization and labeling of qualitative data: the latent categorization method," *Sociological Methodology*, vol. 34, no. 1, pp. 349–392, 2004.

[50] J. F. López-Quintero, J. M. Cueva Lovelle, R. González Crespo, and V. García-Díaz, "A personal knowledge management metamodel based on semantic analysis and social information," *Soft Computing*, vol. 22, no. 6, pp. 1845–1854, 2018.

[51] A. Sidorova, N. Evangelopoulos, J. S. Valacich, and T. Ramakrishnan, "Uncovering the intellectual core of the information systems discipline," *MIS Quarterly*, vol. 32, no. 3, pp. 467–482, 2008.

[52] H. Tao, J. Li, T. Luo, and C. Wang, "Research on topics trends based on weighted k-means," in *Proceedings of the 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2017.

[53] J. Chen, W. Wei, C. Guo, L. Tang, and L. Sun, "Textual analysis and visualization of research trends in data mining for electronic health records," *Health Policy and Technology*, vol. 6, no. 4, pp. 389–400, 2017.

[54] S. Goyal, M. Ahuja, and J. Guan, "Information systems research themes: a seventeen-year data-driven temporal analysis," *Communications of the Association for Information Systems*, vol. 43, no. 1, p. 23, 2018.

[55] H. Jiang, M. Qiang, and P. Lin, "A topic modeling based bibliometric exploration of hydropower research," *Renewable and Sustainable Energy Reviews*, vol. 57, pp. 226–237, 2016.

[56] M. Kamber and J. Pei, *Data Mining*, Morgan Kaufmann, Burlington, MA, USA, 2006.

[57] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[58] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[59] M. M. Hossain, V. Prybutok, and N. Evangelopoulos, "Causal latent semantic analysis (CLSA): an illustration," *International Business Research*, vol. 4, no. 2, p. 38, 2011.

[60] P. Kherwa and P. Bansal, "Latent semantic analysis: an approach to understand semantic of text," in *Proceedings of the 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, 2017.

[61] C. S. Kim, S. J. Choi, and K. Y. Kwahk, "Investigation of research trends in information systems domain using topic modeling and time series regression analysis," *Journal of Digital Contents Society*, vol. 18, no. 6, pp. 1143–1150, 2017.

[62] S. K. Sehra, Y. S. Brar, N. Kaur, and S. S. Sehra, "Research patterns and trends in software effort estimation," *Information and Software Technology*, vol. 91, pp. 1–21, 2017.

[63] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 49–66, 2017.

[64] Elsevier, "Mendeley desktop," 2020, https://www.mendeley.com/download-desktop/.

[65] Y. Shinyama, *PDF Miner*, 2004.

[66] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, Cambridge, UK, 2007.

[67] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[68] Python, "Natural language toolkit (Nltk)," 2020, https://www.nltk.org/.

[69] J. H. Paik, "A novel TF-IDF weighting scheme for effective ranking," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013.

[70] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[71] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, 2004.

[72] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing applications," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[73] A. A. Wagire, A. P. S. Rathore, and R. Jain, "Exploration of pillars of industry 4.0 using latent semantic analysis technique," in *Intelligent Manufacturing and Energy Sustainability*, pp. 711–719, Springer, Berlin, Germany, 2020.

[74] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 918–930, 2006.

[75] R. Jingbiao and Y. Shaohong, "Research and improvement of clustering algorithm in data mining," in *Proceedings of the 2010 2nd International Conference on Signal Processing Systems*, vol. 1, 2010.

[76] M. Srinivas and C. K. Mohan, "Efficient clustering approach using incremental and hierarchical clustering methods," in *Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010.

[77] S. Singh and A. Yadav, "Study of k-means and enhanced k-means clustering algorithm," *International Journal of Advanced Research in Computer Science*, vol. 4, no. 10, 2013.

[78] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143–175, 2001.

[79] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[80] A. Rangrej, S. Kulkarni, and A. V. Tendulkar, "Comparative study of clustering techniques for short text documents," in *Proceedings of the 20th International Conference Companion on World Wide Web*, 2011.

[81] H. A. Linstone and M. Turoff, *The Delphi Method*, Addison-Wesley, Boston, MA, USA, 1975.

[82] A. P. Fuchs, A. Chaudhuri, and J. S. Foster, "SCanDroid: automated security certification of android application," Technical report, University of Maryland, College Park, MD, USA, 2009.

[83] W. Enck, P. Gilbert, S. Han et al., "Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Transactions on Computer Systems (TOCS)*, vol. 32, no. 2, p. 5, 2014.

[84] P. Faruki, V. Kumar, B. Ammar, M. S. Gaur, V. Laxmi, and M. Conti, "Platform neutral sandbox for analyzing malware and resource hogger apps," in *Proceedings of the 2014 International Conference on Security and Privacy in Communication Networks*, Springer, Beijing, China, 2014.

[85] D. Sounthiraraj, J. Sahs, G. Greenwood, Z. Lin, and L. Khan, "SMV-hunter: large scale, automated detection of SSL/TLS man-in-the-middle vulnerabilities in android apps," in *Proceedings of the 21st Annual Network and Distributed System Security Symposium*, San Diego, CA, USA, 2014.

[86] R. Pandita, X. Xiao, W. Yang, W. Enck, and T. Xie, "WHYPER: towards automating risk assessment of mobile applications," in *Proceedings of the 22nd USENIX Security Symposium*, Washington, DC, USA, 2013.

[87] G. Dini, F. Martinelli, A. Saracino, and D. Sgandurra, "MADAM: a multi-level anomaly detector for android malware," in *Proceedings of the 2012 International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security*, 2012.

[88] W. Zhou, Y. Zhou, X. Jiang, and P. Ning, "Detecting repackaged smartphone applications in third-party android marketplaces," in *Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy*, 2012.

[89] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, "PScout: analyzing the android permission specification," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, ACM, Raleigh, NC, USA, 2012.

[90] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, 2011.

[91] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "DREBIN: effective and explainable detection of android malware in your pocket," in *Proceedings of the 2014 Network and Distributed System Security Symposium*, vol. 14, San Diego, CA, USA, 2014.

[92] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for IoT security based on learning techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019.

[93] Z. Luoshi, N. Yan, W. Xiao, W. Zhaoguo, and X. Yibo, "A3: automatic analysis of android malware," in *Proceedings of the 1st International Workshop on Cloud Computing and Information Security*, University of Western Australia, Perth, Australia, 2013.

[94] K. Tam, A. Feizollah, N. B. Anuar, R. Salleh, and L. Cavallaro, "The evolution of android malware and android analysis techniques," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, p. 76, 2017.

[95] J. Kim, Y. Yoon, K. Yi, J. Shin, and S. Center, "Scandal: static analyzer for detecting privacy leaks in android applications," *MoST*, vol. 12, no. 110, p. 1, 2012.

[96] K. Z. Chen, N. M. Johnson, V. D'Silva et al., "Contextual policy enforcement in android applications with permission event graphs," in *Proceedings of the 2013 Network & Distributed System Security Symposium*, San Diego, CA, USA, 2013.

[97] A. Desnos and G. Gueguen, "Android: from reversing to decompilation," in *Proceedings of the 2011 Black Hat*, pp. 77–101, Abu Dhabi, UAE, 2011.

[98] J. Hoffmann, M. Ussath, T. Holz, and M. Spreitzenbarth, "Slicing droids: program slicing for smali code," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ACM, Coimbra, Portugal, 2013.

[99] M. Zhang and H. Yin, "Appsealer: automatic generation of vulnerability-specific patches for preventing component hijacking attacks in android applications," in *Proceedings of the 2014 NDSS*, San Diego, CA, USA, 2014.

[100] W. Enck, D. Octeau, P. D. McDaniel, and S. Chaudhuri, "A study of android application security," in *Proceedings of the*

2011 USENIX Security Symposium, vol. 2, San Francisco, CA, USA, 2011.

[101] C. Gibler, J. Crussell, J. Erickson, and H. Chen, "AndroidLeaks: automatically detecting potential privacy leaks in android applications on a large scale," in *Proceedings of the 2012 International Conference on Trust and Trustworthy Computing*, 2012.

[102] S. Arzt, S. Rasthofer, C. Fritz et al., "FlowDroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps," in *ACM Sigplan Notices*, vol. 49, pp. 259–269, ACM, New York, NY, USA, 2014.

[103] Y. Feng, S. Anand, I. Dillig, and A. Aiken, "Apposcopy: semantics-based detection of android malware through static analysis," in *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 2014.

[104] L. Li, A. Bartel, T. F. Bissyandé et al., "ICCTA: detecting inter-component privacy leaks in android apps," in *Proceedings of the 37th International Conference on Software Engineering*, 2015.

[105] M. Sun, T. Wei, and J. Lui, "TaintART: a practical multi-level information-flow tracking system for android runtime," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, Vienna, Austria, 2016.

[106] F. Wei, S. Roy, X. Ou, and Robby, "Amandroid: a precise and general inter- component data flow analysis framework for security vetting of android apps," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ACM, Scottsdale, AZ, USA, 2014.

[107] S. Wu, P. Wang, X. Li, and Y. Zhang, "Effective detection of android malware based on the usage of data flow apis and machine learning," *Information and Software Technology*, vol. 75, pp. 17–25, 2016.

[108] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang, "Riskranker: scalable and accurate zero-day android malware detection," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ACM, Windermere, UK, 2012.

[109] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang, "Hey, you, get off of my market: detecting malicious apps in official and alternative android markets," in *Proceedings of the 2012 NDSS*, vol. 25, San Diego, CA, USA, 2012.

[110] H. S. Karlsen, E. R. Wognsen, M. C. Olesen, and R. R. Hansen, "Study, formalisa-tion, and analysis of dalvik bytecode," in *Proceedings of the 2012 7th Workshop on Bytecode Semantics, Verification, Analysis and Transformation*, Tallinn, Estonia, 2012.

[111] P. Faruki, V. Laxmi, M. S. Gaur, and P. Vinod, "Mining control flow graph as API call-grams to detect portable executable malware," in *Proceedings of the 5th International Conference on Security of Information and Networks*, ACM, Jaipur, India, 2012.

[112] D. J. Wu, C. H. Mao, T. E. Wei, H. M. Lee, and K. P. Wu, "Droidmat: android malware detection through manifest and API calls tracing," in *Proceedings of the 2012 7th Asia Joint Conference on Information Security*, 2012.

[113] B. Sanz, I. Santos, X. Ugarte-Pedrero, C. Laorden, J. Nieves, and P. G. Bringas, "Anomaly detection using string analysis for android malware detection," in *Proceedings of the 2014 International Joint Conference SOCO13-CISIS13-ICEUTE13*, 2014.

[114] R. Baeza-Yates, B. Ribeiro-Neto, and B. d. A. N. Ribeiro, *Modern Information Retrieval*, Vol. 463, ACM Press, New York, NY, USA, 1999.

[115] Y. Aafer, W. Du, and H. Yin, "DroidAPIMiner: mining api-level features for robust malware detection in android," in *Proceedings of the 2013 International Conference on Security and Privacy in Communication Systems*, 2013.

[116] V. Avdiienko, K. Kuznetsov, A. Gorla et al., "Mining apps for abnormal usage of sensitive data," in *Proceedings of the 37th International Conference on Software Engineering*, 2015.

[117] J. Dave, S. Saharan, P. Faruki, V. Laxmi, and M. S. Gaur, "Secure random encryption for deduplicated storage," in *Proceedings of the 2017 International Conference on Information Systems Security*, 2017.

[118] H. Peng, C. Gates, B. Sarma et al., "Using probabilistic generative models for ranking risks of android apps," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 2012.

[119] Y. Zhou and X. Jiang, "Dissecting android malware: characterization and evolution," in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, 2012.

[120] X. Sun, Y. Zhongyang, Z. Xin, B. Mao, and L. Xie, "Detecting code reuse in android applications using component-based control flow graph," in *Proceedings of the 2014 IFIP International Information Security Conference*, 2014.

[121] J. Crussell, C. Gibler, and H. Chen, "Attack of the clones: detecting cloned applications on android markets," in *Proceedings of the 2012 European Symposium on Research in Computer Security*, 2012.

[122] F. Zhang, H. Huang, S. Zhu, D. Wu, and P. Liu, "ViewDroid: towards obfuscation-resilient mobile application repackaging detection," in *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks*, ACM, Oxford, UK, 2014.

[123] Y. Shao, X. Luo, C. Qian, P. Zhu, and L. Zhang, "Towards a scalable resource-driven approach for detecting repackaged android applications," in *Proceedings of the 30th Annual Computer Security Applications Conference*, 2014.

[124] J. Crussell, C. Gibler, and H. Chen, "AnDarwin: scalable detection of semantically similar android applications," in *Proceedings of the European Symposium on Research in Computer Security*, 2013.

[125] A. Al-Haiqi, M. Ismail, and R. Nordin, "On the best sensor for keystrokes inference attack on android," *Procedia Technology*, vol. 8, pp. 947–953, 2013.

[126] L. Deshotels, "Inaudible sound as a covert channel in mobile devices," in *Proceedings of the 8th USENIX Workshop on Offensive Technologies*, San Diego, CA, USA, 2014.

[127] E. Miluzzo, A. Varshavsky, S. Balakrishnan, and R. R. Choudhury, "Tapprints: your finger taps have fingerprints," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ACM, Low Wood Bay Lake District, UK, 2012.

[128] R. Schlegel, A. Kapadia, and X. Wang, "Soundcomber: a stealthy and context-aware sound Trojan for smartphones," in *Proceedings of the 2011 NDSS*, vol. 11, San Diego, CA, USA, 2011.

[129] N. Xu, F. Zhang, Y. Luo, W. Jia, D. Xuan, and J. Teng, "Stealthy video capturer: a new video-based spyware in 3G smartphones," in *Proceedings of the 2nd ACM Conference on Wireless Network Security*, ACM, Zurich, Switzerland, 2009.

[130] G. Suarez-Tangil, J. E. Tapiador, F. Lombardi, and R. Di Pietro, "Alterdroid: differential fault analysis of obfuscated

smartphone malware," *IEEE Transactions on Mobile Computing*, vol. 15, no. 4, pp. 789–802, 2015.

[131] G. Suarez-Tangil, J. E. Tapiador, P. Peris-Lopez, and J. Blasco, "Dendroid: a text mining approach to analyzing and classifying code structures in android malware families," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1104–1117, 2014.

[132] G. Suarez-Tangil, S. K. Dash, M. Ahmadi, J. Kinder, G. Giacinto, and L. Cavallaro, "DroidSieve: fast and accurate classification of obfuscated android malware," in *Proceedings of the 7th ACM on Conference on Data and Application Security and Privacy*, ACM, Scottsdale, AZ, USA, 2017.

[133] W. Wang, Y. Li, X. Wang, J. Liu, and X. Zhang, "Detecting android malicious apps and categorizing benign apps with ensemble of classifiers," *Future Generation Computer Systems*, vol. 78, pp. 987–994, 2018.

[134] M. Lindorfer, M. Neugschwandtner, L. Weichselbaum, Y. Fratantonio, V. Van Der Veen, and C. Platzer, "Andrubis 1,000,000 apps later: a view on current android malware behaviors," in *Proceedings of the 2014 3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, 2014.

[135] A. Desnos and P. Lantz, "Droidbox: an android application sandbox for dynamic analysis," Technical report, Lund University, Lund, Sweden, 2011.

[136] W. Enck, M. Ongtang, and P. McDaniel, "On lightweight mobile phone application certification," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, 2009.

[137] L. K. Yan and H. Yin, "Droidscope: seamlessly reconstructing the OS and dalvik semantic views for dynamic android malware analysis," in *Proceedings of the 21st USENIX Security Symposium*, Bellevue, WA, USA, 2012.

[138] P. Faruki, V. Laxmi, A. Bharmal, M. S. Gaur, and V. Ganmoor, "Androsimilar: robust signature for detecting variants of android malware," *Journal of Information Security and Applications*, vol. 22, pp. 66–80, 2015.

[139] A. T. Kabakus and I. A. Dogru, "An in-depth analysis of android malware using hybrid techniques," *Digital Investigation*, vol. 24, pp. 25–33, 2018.

[140] J. Fu, J. Xue, Y. Wang, Z. Liu, and C. Shan, "Malware visualization for fine- grained classification," *IEEE Access*, vol. 6, pp. 14510–14523, 2018.

[141] S. Ni, Q. Qian, and R. Zhang, "Malware identification using visualization images and deep learning," *Computers & Security*, vol. 77, pp. 871–885, 2018.

[142] S. Sun, X. Fu, H. Ruan, X. Du, B. Luo, and M. Guizani, "Real-time behavior analysis and identification for android application," *IEEE Access*, vol. 6, pp. 38041–38051, 2018.

[143] J. Yan, Y. Qi, and Q. Rao, "LSTM-based hierarchical denoising network for android malware detection," *Security and Communication Networks*, vol. 2018, Article ID 5249190, 18 pages, 2018.

[144] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, and H. Ye, "Significant permission identification for machine-learning-based android malware detection," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3216–3225, 2018.

[145] P. Faruki, H. Fereidooni, V. Laxmi, M. Conti, and M. Gaur, "Android code protection via obfuscation techniques: past, present and future directions," 2016, https://arxiv.org/abs/1611.10231.

[146] W. Wang, Z. Gao, M. Zhao, Y. Li, J. Liu, and X. Zhang, "Droidensemble: detecting android malicious applications with ensemble of string and structural static features," *IEEE Access*, vol. 6, pp. 31798–31807, 2018.

[147] S. Hyun, J. Cho, G. Cho, and H. Kim, "Design and analysis of push notification- based malware on android," *Security and Communication Networks*, vol. 2018, Article ID 8510256, 12 pages, 2018.

[148] Y. Liu, K. Guo, X. Huang, Z. Zhou, and Y. Zhang, "Detecting android malwares with high-efficient hybrid analyzing methods," *Mobile Information Systems*, vol. 2018, Article ID 1649703, 12 pages, 2018.

[149] Symantec, "Internet security threat report 2018," 2018, https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf.

[150] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, "Ransomware threat success factors, taxonomy, and countermeasures: a survey and research directions," *Computers & Security*, vol. 74, pp. 144–166, 2018.

[151] L. Nataraj, S. Karthikeyan, G. Jacob, and B. Manjunath, "Malware images: visualization and automatic classification," in *Proceedings of the 8th Inter- National Symposium on Visualization for Cyber Security*, ACM, Pittsburgh, PA, USA, 2011.

[152] L. Nataraj, D. Kirat, B. Manjunath, and G. Vigna, "SARVAM: search and retrieval of malware," in *Proceedings of the Annual Computer Security Conference (ACSAC) Workshop on Next Generation Malware Attacks and Defense (NGMAD)*, University of Western Australia, Los Angeles, CA, USA, 2013.

[153] K. Bakour and H. M. Ünver, "VisDroid: android malware classification based on local and global image features, bag of visual words and machine learning techniques," *Neural Computing and Applications*, vol. 33, no. 8, pp. 3133–3153, 2021.

[154] A. Mahindru and A. L. Sangal, "MLDroid-framework for android malware detection using machine learning techniques," *Neural Computing and Applications*, vol. 33, no. 10, pp. 5183–5240, 2021.

[155] Y. Zhao, L. Li, H. Wang et al., "On the impact of sample duplication in machine-learning-based android malware detection," *ACM Transactions on Software Engineering and Methodology*, vol. 30, no. 3, pp. 1–38, 2021.

[156] J. Singh, D. Thakur, F. Ali, T. Gera, and K. S. Kwak, "Deep feature extraction and classification of android malware images," *Sensors*, vol. 20, no. 24, p. 7013, 2020.

[157] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, "IMCFN: image-based malware classification using fine-tuned convolutional neural network architecture," *Computer Networks*, vol. 171, Article ID 107138, 2020.

Hindawi
Complexity
Volume 2021, Article ID 9567524, 12 pages
https://doi.org/10.1155/2021/9567524

Research Article

Assessing the Impact of Virtual Standby Systems in Failure Propagation for Complex Wastewater Treatment Processes

Fredy Kristjanpoller,[1] **Pablo Viveros**,[1] **Nicolás Cárdenas**,[1] **and Rodrigo Pascual**[2]

[1]*Department of Industrial Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile*
[2]*Mechanical Engineering Department, Universidad de Chile, Santiago, Chile*

Correspondence should be addressed to Fredy Kristjanpoller; fredy.kristjanpoller@usm.cl

Received 25 May 2021; Revised 24 July 2021; Accepted 8 August 2021; Published 19 August 2021

Academic Editor: Jenq-Haur Wang

This article proposes an original probabilistic modelling methodology named Virtual Standby (VSB), which enables a practical simulation, analysis, and evaluation of the impact on availability and reliability achieved by potential buffering policies on the performance of complex production systems. Virtual Standby (VSB) corresponds to a design and operational characteristic where some machines under a failure scenario are capable to provide for a limited time, continuity to the subsystems downstream before suffering delay which is currently not considered when assessing availability. This feature plays a relevant role on the propagation of the effect of a failure; indeed, it could prevent the propagation by guaranteeing the isolation time needed to recover from its failure, controlling and reducing the production losses downstream. A case study of the preliminary treatment process of a wastewater treatment facility (WWTF) is developed bearing in mind the systemic behaviour in the event of a failure and the specific features of each equipment. VSB is a big advantage for the representation of this complex processes because, among other things, it considers the impact of buffering policies on the perceived availability of the system. This model allows determining different production levels, with a better and easier fitting of the reliability, availability, and production forecast of the process. Finally, the comparison between the VSB simulation results with traditional procedures that do not consider the operational continuity under a failure scenario confirms the strength and precision of the proposal for complex systems.

1. Introduction

The performance of a system is the result of the synergic work of different sets of machines and individual machines adding to the overall performance. Each individual or set of machines is bounded by a set of constraints inherent to each machine or set of machines. Some of these are maintainability and maintenance requirements, reliability, nominal capacity, maintenance plan, operational limitations, layout of the system, and complexity degree.

The combination of all these aspects may create production bottlenecks [1, 2] and delays; hence, they must be corrected in a manner that is effective and accurate [3, 4]. Therefore, a combined analysis of reliability and productivity must be performed to allow optimal use of resources and achieve the required production goals [5, 6].

The traditional reliability analysis of complex systems is usually based on a logical and probabilistic modelling approach, which contributes to improve the key performance indicators (KPIs) of production systems [7, 8]. Nowadays, it is possible to find in the literature many alternatives available for reliability analysis of complex systems [9, 10]. The systematic studies are usually developed considering techniques and methodologies as Reliability Block Diagrams (RBDs) [11, 12], Fault Trees (FTs) [13], Reliability Graphics (RGs) [14], Petri Nets (PNs) [15], and Monte Carlo simulation [16–18] among others. More recently, other techniques have emerged such as Multistate Systems [19], Graph Topology [20], and fuzzy approaches [3] which have allowed to reveal subjacent connections rising from the process dynamic. Another approach would be to implement specially designed algorithms to assess availability and

reliability, such as computing the Equivalent Availability (EA) index that makes use of the shared load between pieces of equipment working under lower loads than their nominal capacity allowing the use of different combinations of equipment to achieve the availability goal [21]. In different scenarios, these techniques must be adapted or extended to account for the particularities of the system, especially for large, complex, and dynamic systems. Such is the case for classic RBD which must be adapted in order to measure effect of WIP or inventory buffering on the performance and availability of the system [4] (other techniques exist, for example, to adapt these types of analysis to demand fluctuations [22]). This is where the methodology developed in this paper fits.

Buffering policies allow machines, under any failure scenario, to provide continuity for a limited time to the production subsystem downstream [23, 24]. The effect and propagation under planned or unplanned stoppages and delays could be total or partially guaranteed, controlling and reducing the production losses depending on the time needed to recover, proper operating conditions (time to repair), and the required capacity to avoid material starvation.

The primary concern of this paper's proposed methodology (VSB) is to ease the process of building probabilistic models to simulate and analyse real production scenarios (wastewater treatment process in this case) involving different buffering policy opportunities [25, 26]. An initial approach for this method has been already developed in a case study for a mining process, which proved the potential for further research [27].

VSB is used within existing Monte Carlo simulation models which will be implemented in an especially designed environment for the case study that can estimate a set of expected performance indicators of a complex system and its equipment with which is possible to estimate statistical variability.

Alternatives to the VSB methodology to model reliability of a complex system, which currently exist in the literature, are as follows:

(i) *Traditional RBD Methodology* [11, 12]. This is a very useful and well-known method; nevertheless, this modelling does not allow to include the differential time effect due to the elements only having two states, and thus failure propagation is immediate.

(ii) *Markov Chain* [5]. In this case, it is only possible to model using constant or discrete-time evolution failure rate, restricting the assessment of the operational reality and complicating production and availability analyses. In general, this procedure does not reach enough detail in the results.

(iii) *Traditional RBD Methodology Using the Universal Generating Function for Data Analysis.* [19] This methodology combines classical RBD with a more accurate data analysis, which translates in better data fitting for failure rate and density functions because it considers the differences raised from the operation of multifunctional systems.

(iv) Finally, the operational continuity could be evaluated through the analysis or simulation of a buffer configuration [4], but considering the characteristics of this methodology, it would be necessary to incorporate and evaluate new variables, not currently contained in the problem under study, such as isolation time, upstream and downstream capacity, availability, nominal throughput, and physical buffer capacity. Even more, the model will have greater complexity if the operational continuity is provided by more than one element, implicating the generation of n buffers for each case and the incorporation of buffer model variables [4] without efficient resource utilization and possible loss of study focus.

This research claims that the development of VSB as a very specific methodology to model these specific buffering situations in production systems along with the use of Monte Carlo simulation provides an excellent and very practical tool to measure and assess the impact of buffering options on both the reliability and availability of complex production systems. These tools may help the analyst to focus on the study of specific modelling variables and therefore help solve problems in an effective and efficient way.

Table 1 shows a comparative analysis between the abovementioned methodologies related to their capacity to model operational continuity after a failure event or delay. This table exposes the differentiating strengths of VSB over the rest of the methodologies. It is necessary to emphasize the capacity of VSB to get valid results using relatively few information and with a moderate analysis effort.

Wastewater treatment is one of the several contexts in which the limitations of industrial processes play a critical role, because of the high impact of failure consequences, not just for the process but for human health and the environment also.

Water is the main responsible for life on the planet Earth and is one of the most important, if not the most important resources for any human settlement in the world. According to a press release from the UN in 2010 where they coined the term "sick water" [28], they address the need of transforming wastewater from a real hazard to health and the environment into a quality and useful resource that is a must for the 21st century in which water crisis is a fact as it is for Africa where it is forecasted that around 3 billion people will live in areas with water scarcity. In this context, they state that "improved sanitation and wastewater management are central to poverty reduction and improved human health" [28].

Since it is clear that sick water crisis is a highly critical problem for humanity to guarantee clean water access for people, the aim of this paper is to improve the assessment of availability and reliability in wastewater treatment processes through a novel method for modelling complex production lines using Virtual Standby.

Table 1: Comparison between methodologies for temporal operational continuity modelling.

Model	Evaluation aspects					
	Temporal continuity	Failure rate	Flexibility	Number of variables required	Modelling size	Construction and analysis effort
RBD [6, 7]	No	Variable	Medium	Medium	Medium	Low
Markov [3]	Yes	Constant	Medium	Medium	Medium	High
Buffer [13]	Yes	Variable	High	High	Big	High
VSB	Yes	Variable	High	Medium	Medium	Medium

2. Objective

The main goal of this research is to propose a novel modelling procedure for industrial processes accountable for failure propagation wherein buffering WIP is possible using probabilistic-based simulations of Virtual Standby backups for units performing specific tasks to minimize workflow interruptions.

According to the goal, this article is organized as follows: first, the problem statement and application of the proposed methodology are exposed in detail. After which, the analysis process is developed and abridged following the proposed methodology, and then an assessment is performed on the analysed data of reliability and maintainability analyses. Finally, a case study is developed, modelled, and solved concluding with some important remarks.

2.1. Problem Statement and VSB Proposal Methodology. As it was expressed in the Introduction section, in a manufacturing industrial process under specific conditions, the failure of one or more elements might not generate a system detention immediately; this capability depends on the system's ability to provide production during a limited time interval after failure, such may be the case for downstream work in the process, for example. This effect could be considered as a buffer [4], but the main variables of each situation are very different. In buffer modelling, the throughput capacity is a key variable to calculate what the starvation level should be for the proper isolation time. The buffer is a physical asset, with a specific capacity and of course with a required investment and maintenance cost and as such it should be considered when assessing availability; therefore, this is where this VSB becomes relevant because it will potentially improve the overall availability of a process reflecting the importance of buffering policies when analysing availability and reliability. In the VSB model, capacity is explained by two factors: a random variable (after failure capacity) and its relationship with the repair time (repair function). There is no relation with bottlenecks (upstream or downstream) or the starvation level. The main principles of VSB methodology are as follows:

(i) To model and represent the VSB scenario, a "virtual" backup must be created bounded by specific parameters for modelling failure and repair times which starts working at the time of failure of the primary equipment. Both primary and "virtual" backup equipment are necessary to model VSB scenario.

(ii) The VSB scenario must be applied only in machines where the above explained operational continuity effect exists. It is a very specific condition, so it is necessarily a deep process analysis to validate the VSB scenario inclusion.

(iii) As a preliminary criterion when modelling, the operating time of the "virtual" backup equipment i ($\mathrm{OT}^B_{i,j}$) should start at time $t = \mathrm{TTF}_i$, along with intervention j. The consecutive time to repair of the virtual backup i ($\mathrm{TTR}^B_{i,j}$) which is also the effective time to repair perceived by the system must be equal to the time to repair of the primary equipment at intervention j ($\mathrm{TTR}_{i,j}$) less than the operating time of "virtual" backup equipment i ($\mathrm{OT}^B_{i,j}$). The rules for the algorithm are expressed in the following equations:

$$\mathrm{OT}^B_{i,j} = \begin{cases} f^{\mathrm{vsb}}_{i,j}(t), & \mathrm{TTR}_{i,j} \geq f^{\mathrm{vsb}}_{i,j}(t), \\ \mathrm{TTR}_{i,j}, & \text{otherwise}, \end{cases} \tag{1}$$

$$\mathrm{TTR}^B_{i,j} = \mathrm{TTR}_{i,j} - \mathrm{OT}^B_{i,j}, \tag{2}$$

where $\mathrm{OT}^B_{i,j}$ is the operational backup time of equipment i during intervention j; $f^{\mathrm{vsb}}_{i,j}(t)$ is the distribution function of autonomy time of equipment i at intervention j; $\mathrm{TTR}_{i,j}$ is the time to repair of equipment i at intervention j; and $\mathrm{TTR}^B_{i,j}$ is the time to repair of the virtual backup equipment i at intervention j.

It is a conservative scenario because with this condition we make sure that after any intervention of the primary system, both assets are restored at the same time with perfect conditions (perfect renewal). This criterion will be graphical and numerically explained next.

Figure 1 represents both cases, with and without VSB scenario. The "Not VSB scenario" shows that any intervention of any single equipment i will affect directly to the operational time. In the second case, VSB can be modelled as a standby system, including "virtual" backup equipment i. The timeline for each equipment and system is depicted in Figure 2; it is possible to observe the effect of VSB which rises real operating time ($\mathrm{OT}_{i,j}$) to the effective operating time of the equipment ($\mathrm{OT}^E_{i,j}$) and reducing the real time to repair of the ($\mathrm{TTR}_{i,j}$) into the effective time to repair ($\mathrm{TTR}^E_{i,j}$). Each operating time increase for the system (Equipment i + Backup i) is equal to the operating time defined for the backup equipment ($\mathrm{OT}^B_{i,j}$). This logic also applies for the time to repair of each equipment, which is equal to the real

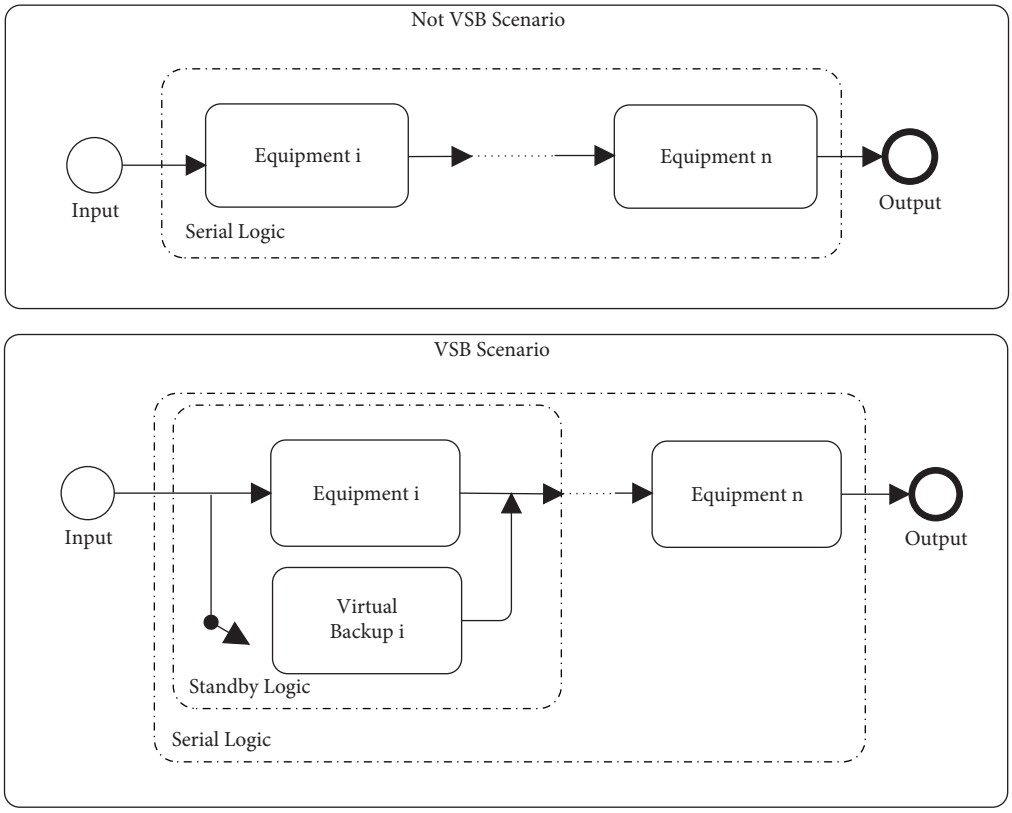


Figure 1: VSB logic representation.

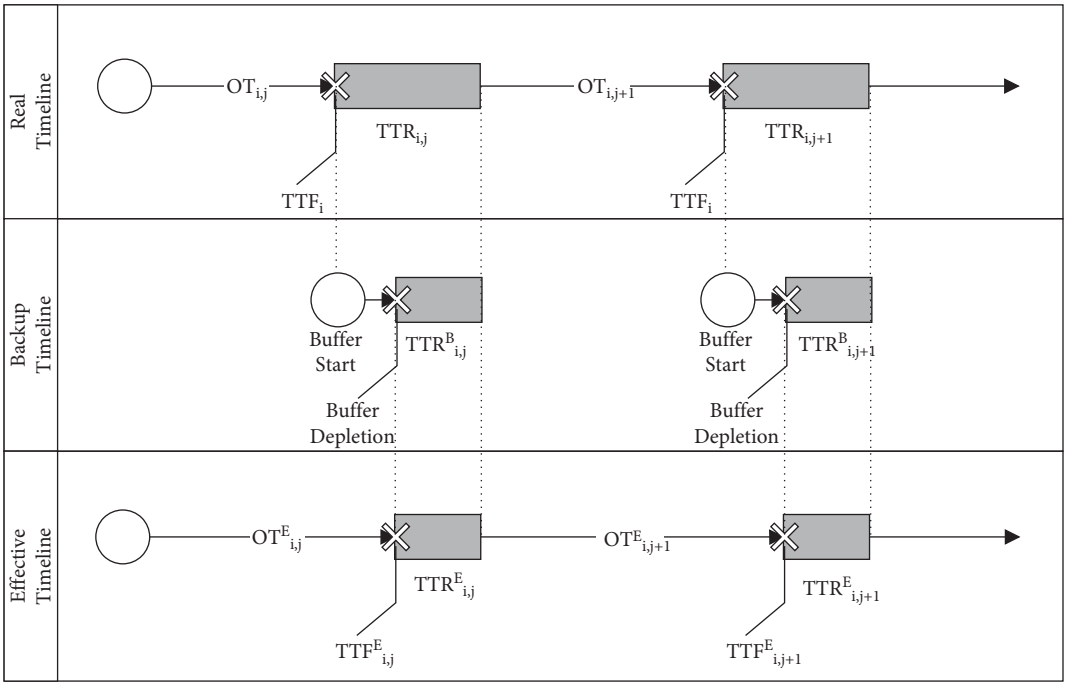


Figure 2: VSB effect on operating time for equipment *i*.

time to repair of the primary equipment and less the operating time of the "virtual" backup equipment. In terms of formulation, it is expressed through the following equations:

$$\mathrm{OT}_{i,j}^{E} = \mathrm{OT}_{i,j} + \mathrm{OT}_{i,j}^{B}, \tag{3}$$

$$\mathrm{TTR}_{i,j}^{E} = \mathrm{TTR}_{i,j} - \mathrm{OT}_{i,j}^{B} = \mathrm{TTR}_{i,j}^{B}, \tag{4}$$

where the sum of each effective operating time of equipment i at the time of intervention j ($\mathrm{OT}_{i,j}^{E}$) defines the effective operating time of the system (OT^{E}), i.e.,

$$\mathrm{OT}^{E} = \sum_{i\in I}\sum_{j\in J} \mathrm{OT}_{i,j}^{E}. \tag{5}$$

Likewise, effective time to repair of the system (TTR^{E}) is defined as follows:

$$\mathrm{TTR}^{E} = \sum_{i\in I}\sum_{j\in J} \mathrm{TTR}_{i,j}^{E}. \tag{6}$$

Thus, to introduce VSB impact on production performance evaluation, the simulation model must account for two scenarios: first, a scenario in which to measure the immediate effect of failure or detention and second a scenario in which a VSB is incorporated. This approach allows for the analysis to be more accurate.

As it was indicated at the beginning of this article, the motivation for this study is to develop an integral, flexible, and probabilistic methodology to model the behaviour and impact of buffering policies in complex systems; the following analysis will study historical statistical data regarding time to repair (TTR), operating time (OT), and its relation with reliability and delays due to maintainability.

Figure 3 describes the main stages of this proposal. Later on, methodology will be explained step by step to ease understanding through a case study.

As shown in Figure 3, VSB methodology is a framework which involves modelling the whole system from the beginning, recognizing the effects of failure on the whole process and the existence of VSB type buffer conditions. Fault Tree Diagrams can be performed to understand the operating logic of the system. The following is the parameterization of the operation and maintenance data of the involved equipment to perform the simulations using graphical models that follow the VSB logic (considering a virtual machine in standby). Finally, the interpretation of the simulation results is made. This interpretation is made in terms of reliability and maintainability indicators.

3. Case Study

As it was mentioned, wastewater treatment or sick water treatment is a critical problem to be addressed by every human settlement; therefore, in this context, it is important to find new and better ways to optimize the said process. The inherent nature of the process to cumulate WIP along the workflow is that buffering WIP is available at several stages of the wastewater treatment process, which often is not considered when assessing operational continuity; for this reason, using VSB will potentially improve the availability and reliability analysis.

Most WWTF workflows consist of two stages: a primary and a secondary stage, and there are also many different settings for these two stages. For the purpose of this paper, a primary stage will be considered where wastewater collected from the city through the sewage system flows into the facility which is immediately screened, usually using metal screens to dispose big elements that wastewater may contain, then it flows through a grit chamber to dispose medium size element, and finally it goes into a primary settling tank to clarify it where suspended solids are collected through settling; this collected material is called "primary sludge." Secondary treatment starts with aeration using blowers connected to aeration basins, then the wastewater flow goes into a secondary clarifier where the sludge is collected again, this time is called "activated sludge" because of the previous aeration process, and finally, before the treated water is released to the environment, it undergoes a decontamination process using UV light for modern processes or chlorine for older processes.

As for most industrial processes, failure is a constant threat randomly waiting to arise and the wastewater treatment process is not an exception. On the contrary, since this process involves working with human activity residues, the raw materials for the process have a wide range of possibilities, meaning that it is impossible for the operator of this process to control which residues will arrive to the plant. In this context, all systems of this process are exposed to different and unpredictable types of material damaging the equipment and therefore producing failures along the process and deeply affecting reliability levels; more specifically, when equipment fails because of the aforementioned hazardous materials, the process downstream will normally continue for a measurable period of time. This time frame is not considered in the classical analysis and therefore is not included when assessing availability or reliability in most (if not all) cases.

This paper presents and analyses a case study developed in the preliminary treatment stage (Figure 4) of a wastewater treatment facility (WWTF). The main goal of this stage is to protect the facility from clogs, jams, or materials that may render excessive wear of the machinery [29]. These are the first stages for most, if not for all, wastewater treatment processes, and its importance relies on the capability for removing undesired objects from the raw wastewater that, apart from being dangerous for the machinery, they take valuable space from the process.

A brief description of the process is as follows. An average of 8 MGD of wastewater flows to the plant; this influent from the plant first undergoes a fine screening process using metal bar screens after which wastewater is stored into two 2,400 ft^3 tanks; and then grit and scum is collected using a 2-grit teacup system of 8 MGD capacity (each). Wastewater is then collected into a 3955 ft^3 tank from which is pumped through a 150 hp, 10 MGD 4-pump system for preliminary treatment, which occurs in two $50\,\mathrm{ft} \times 50\,\mathrm{ft}$ clarifiers.

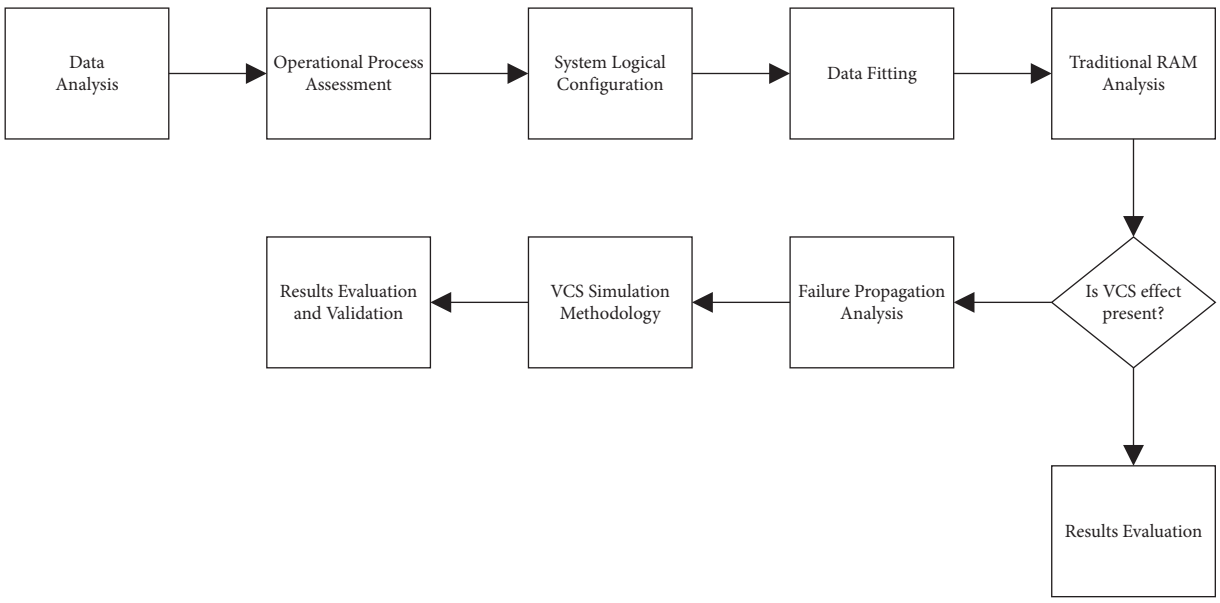


Figure 3: VSB simulation methodology scheme.

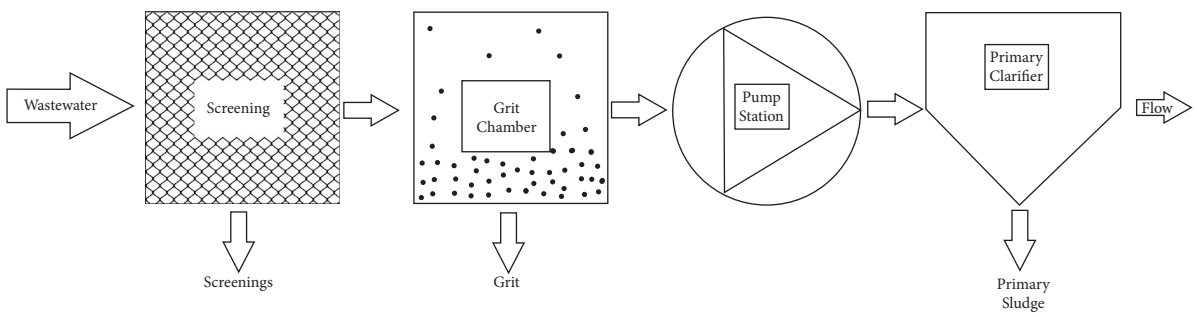


Figure 4: Process diagram for the preliminary treatment process.

Most important features of the t preliminary treatment process shown in Figure 4 are listed in Table 2.

4. Modelling the System

The relationship between subsystems operation under the same process (functional dependency) arises when asking "what if . . .?" This translates into a necessity to track any effect produced by a planned or random state change of a subsystem or equipment embedded in the system. Then, the effect on functioning and workload capacity over the system and its components must be studied and analysed. Usually and for the purpose of this proposal, two possible states are considered: degradation (normal established functioning) and nondegradation (failure state, preventive intervention, or operational detention) [30].

For the case study, four machines from a subprocess of the WWTF are set in serial. Therefore, if one of the pieces of equipment fails, the whole system fails. Accordingly, it was identified that to consider a VSB process for bar screens or grit chamber when they fail should be most beneficial for the expected results. In the case of failure of one or both of the mentioned equipment, the process downstream will continue to work properly for approximately. This feature is comparable with machines with the capacity of accumulating WIP during regular operation. This capacity is estimated of supplying around 30 minutes of downstream operation. Regarding the historical data analysis of this supplying capability, the simulation model will consider a discrete uniform distribution ($f_{i,j}^{\mathrm{vsb}}(t)$) between 26 and 30 minutes.

As an approximation, the VSB scenario is equivalent to add a standby system [31]; this standby is a redundancy method that involves having one system as a backup for another identical primary system. The standby system is required only upon failure of the primary system. This configuration is constrained by random variables, perfect repair, instantaneous, and perfect switch, mindful that the lifetime of the backup is equal to the defined time for the VSB.

Continuing with the wastewater treatment process, the following Fault Tree diagrams were developed (Figures 5 and 6) to support the understanding and representation of VSB logic.

Under the purpose of reducing the amount of analysis and not sacrificing the outcome quality, it is considered that

Table 2: Preliminary treatment process information.

Equipment	ID	Basic function
Bar screen	BS_001	Large solids removal
Grit chamber	GRIT_001	Removal of heavy inorganic solids such as grit, sand, and gravel, among others
Pump station	PUMP_001	Transport wastewater from grit chamber to primary clarifier
Primary clarifier	CLAR_001	Removal of settleable organic solids

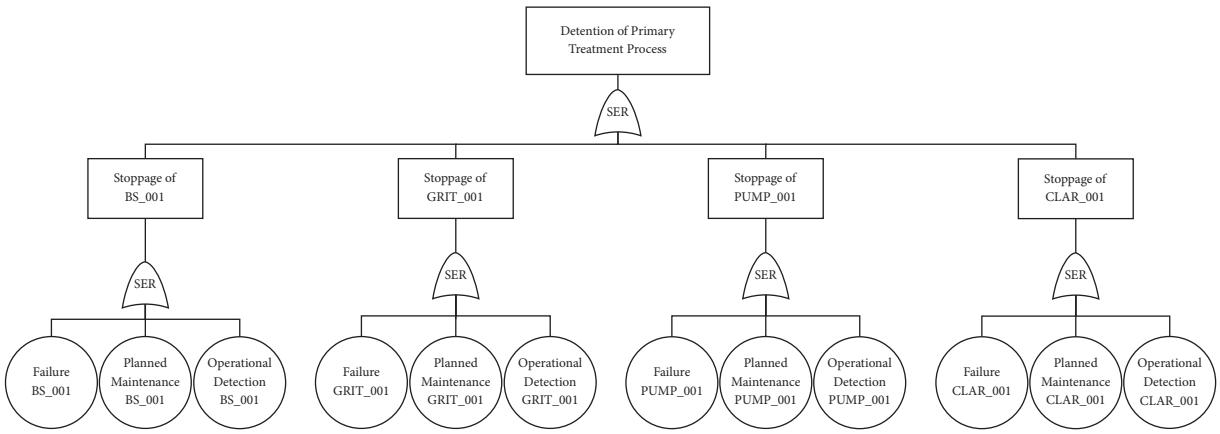


Figure 5: FT representation of the primary treatment process-immediate effect.

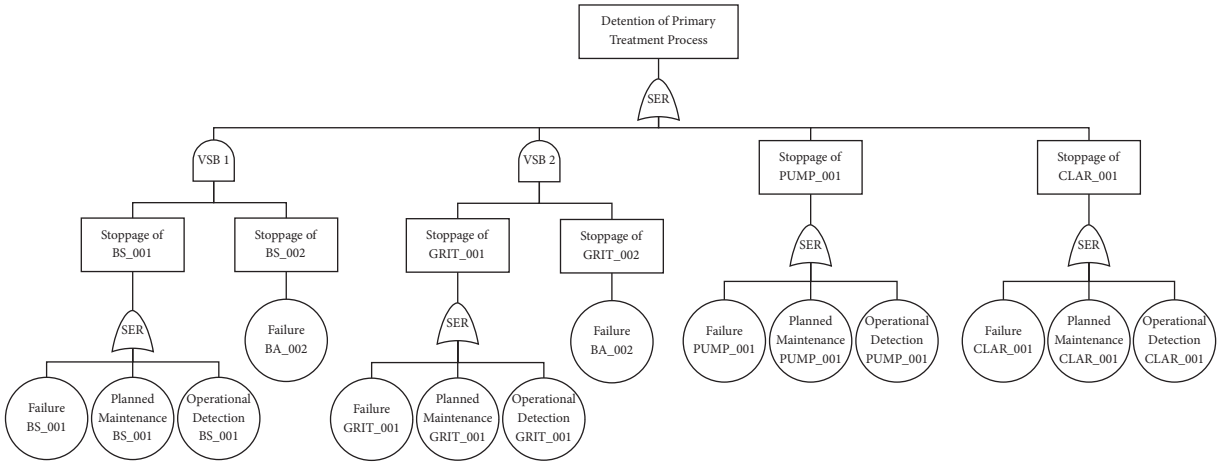


Figure 6: FT representation of the primary treatment process-VSB logic approximation.

the simulation will not account for operational or planned detentions, as it is graphically represented by the FT diagrams (Figures 5 and 6).

4.1. Data Parameterization. Usually, when describing failure behaviour or repair processes, it is necessary to define a probability distribution to model said features. Hence, several statistical distributions have been assessed and parameters are estimated using, a specially designed environment for the analysis.

Table 3 shows most important parameters and KPI regarding reliability and maintainability.

4.2. Simulation Methodology Application. As it was mentioned, before modelling the system and performing a simulation, the model has to consider all specific features of the system regarding operational conditions and all constraints that may exist due to the real physical relation between components. These selected features are listed in Table 2. Details of constraints regarding logical and functional dependency can be found in section: Modelling the System. The simulation must include an average production rate, which is equivalent for all equipment based on the serial relationship presented. Each piece of equipment must be able to produce at the required rate by the process, and this being totally or partially as demanded by the system.

Table 3: Reliability and maintainability information.

Equipment	Operating time parameterization				Time to repair parameterization			
	Best fit distribution	Parameter 1	Parameter 2	MTTFi	Best fit distribution	Parameter 1	Parameter 2	MTTRi
BS_001	Weibull	$\alpha = 45.88$	$\beta = 1.22$	42.98	Normal	$\mu = 1.4$	$\sigma = 1.17$	1.40
GRIT_001	Weibull	$\alpha = 68.86$	$\beta = 1.18$	65.06	Normal	$\mu = 2.4$	$\sigma = 1.21$	2.40
PUMP_001	Weibull	$\alpha = 28.75$	$\beta = 1.08$	27.91	Normal	$\mu = 1.2$	$\sigma = 0.78$	1.20
CLAR_001	Weibull	$\alpha = 98.42$	$\beta = 1.14$	93.91	Normal	$\mu = 3.2$	$\sigma = 1.58$	3.20

For the case study, the production rate considered is 8 MGD, and it assumes that the influent is equivalent to the daily output demanded by the process or the daily rate of effluent. This means that in a classical analysis, the whole system will stop for lack of influent or for capacity problems when critical equipment fails upstream.

The graphical models (based for the simulation) developed are presented and analysed next.

4.3. Considerations for the Simulation Model. Processing systems depend in part on the established operating logic. In general, the continuous simulators, or discrete that includes continuous control and monitoring variables, develop the estimation of indicators and identification of states through monitoring at certain intervals of time. In most cases, said procedure is slightly more efficient compared with methods that focus on the state change of components in the system where monitoring and consultation are performed when something in the system changes state, either a random or a planned condition. For this, a continued evaluation of the state of each element of the system is not needed since for the interest of this proposal, it is important to analyse the impact on operational time, availability, and reliability by comparing the behaviour of the system with and without VSB. Hence, for simulation purposes, the statistical environment designed in this paper is based on discrete-time event occurrence data allowing the impact of functional dependencies to be visible.

It is possible to establish the principal components to develop a modelling task such as tree of components representing the hierarchical structure of the systems and the flow chart.

4.4. Implementing VSB Simulation Methodology and Analysis. As it was described, this proposal considers a traditional scenario in contrast with the VSB scenario (i.e., immediate effect scenario vs. VSB scenario) as it can be observed in Figures 7 and 8.

For both scenarios, data inputs about the characteristics of each piece of equipment considered in the simulation are required (see Table 2). Furthermore, for VSB scenario, it is considered that repair interventions are independent, and that the standby equipment (VSB) starts working at the exact moment of failure of the primary equipment (bar screens and grit chamber in this case). This is usually known as standby [31].

As was explained before, the parameters of life degradation for the analysed equipment are modelled through a discrete uniform distribution, $(f_{i,j}^{\text{vsb}}(t))$ [27–31] min.

In the simulation model, the VSB machine must provide downstream systems the same autonomy level provided by the primary machine to survive after a failure. Thus, the expected operating time of the virtual machine cannot exceed, in equivalent terms, the autonomy level of the primary machine. Accordingly, the operating time of the virtual machine in the case study will be modelled by a uniform distribution $(f_{i,j}^{\text{vsb}}(t))$ [27–31] min.

For the virtual machine approximation (standby), it must be met that the virtual machine must be in perfect reliable condition every time that the primary machine starts operating (after an intervention). Then, the time to repair (TTR) of the virtual machine must be less or equal than the difference between the TTR of the primary machine and the equivalent time of autonomy for the virtual machine. With this, the virtual machine operates, and it is maintained while the primary machine is been restored. In the best case, when the primary machine is repaired in less time than the autonomy equivalent time, the system assumes TTR equal to 0.

It is important to highlight that the mean time to failure (MTTF) for the virtual machine will be directly dependent on the uniform distribution considered. If the MTTF of the virtual machine is compared with the MTTR of the virtual machine, most of the time the MTTF will be shorter than MTTR.

5. Simulation Results

Considering the elements that compose the system and the redundancy configurations, a horizon of 365 days of operation was selected (approximately 8,760 hours under normal conditions) rendering 100,000 replications of said horizon. This is mostly to assess a representative sample with which generate more accurate indicators and histograms. It is also important to highlight that some machines have very short autonomy times (e.g., Primary Clarifier); therefore, when analysing the time horizon in cases where the system is unable to provide influent to the aforementioned pieces of equipment, this autonomy time will become significant.

5.1. Analysis and Results 1: Immediate Effect Scenario. The performance indicators to measure are availability, operation time, mean time to failure (MTTF), mean time to repair (MTTR), and the total effluent produced by the system. The outcome for the scenario with immediate effect is compared with traditional statistical analysis (RBD). The expected indicators of simulation approach to RBD are quite different because in the simulation model, the failure propagation is direct, and in RBD approach, the assets are

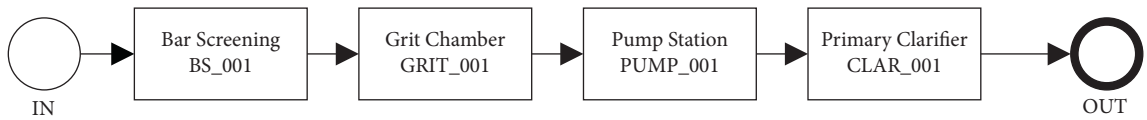


Figure 7: Graphical representation of modelling-immediate effect scenario.

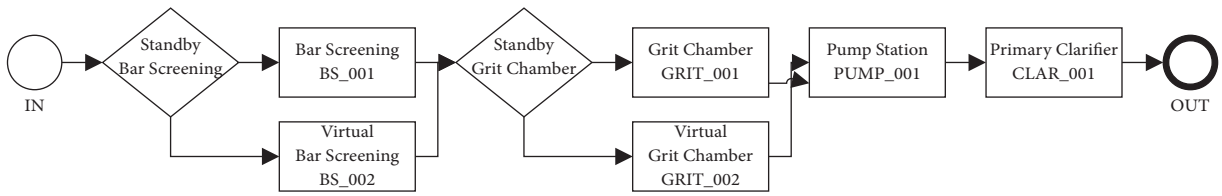


Figure 8: Graphical representation of modelling-VSB scenario.

modelled in an independent way. The specific results are shown in Table 4.

According to the results of simulation plus the requirements for VSB, BS_001 and GRIT_001 would be the incumbent equipment based on the availability indicator (96.68% and 96.19% respectably) and their buffering capabilities. In this scenario, the expected mean percentual availability of the system is 85.91%, in which during this available time, 88.98% corresponds to the system actually working. Since the logical configuration is in series, any change of state of a machine or set of machines will induce a state change on the overall system.

For that last reason, it will be important to identify which pieces of equipment require reliability improvements in order to decrease frequency of system failures; in this case, both bar screen (BS_001) and grit chamber (GRIT_001) were identified. This means increasing the mean time to failure, 45.88 and 68.86, respectively. When analysing and comparing between the simulation and RBD results, it is possible to verify a specific deviation. As it was commented before, the difference originates from the assumption of independent machine behaviour in the RBD model, while in the simulation model, the effect of individual failure propagation is incorporated. Indeed, in the simulation model, whenever failure occurs, the operating time of all working machines will stop (wear stops because of failure propagation, and of course the reliability is not improved). During that time, the machine that failed is maintained; furthermore, for maintenance, actions shorter than the buffer spam failure will not spread on the system, and it will continue working normally. This feature is essential to conclude the incompatibility with RBD modelling, and that the VSB proposal is an interesting alternative to address the problem.

5.2. Analysis and Results 2: VSB Scenario. When introducing the Virtual Standby effect, a more realistic availability estimation is obtained. Table 5 shows the main results for availability, operation time, expected production, and maintainability.

Again, bar screens and grit chamber are the incumbent equipment because of their availability and buffering capabilities. The pump system and clarifier have increased their operating time thanks to the VSB because, as mentioned before, it considers buffers acting as actual working pieces of equipment and not just an accumulation of inventory as they have been considered until now. For the VSB scenario, the mean percentage availability for the system is 86.62%, in which during this available time 89.72% corresponds to the system actually working. This last indicator is most important since it allows for the scenarios to be compared. It is also possible to observe that the mean time to repair (1.72 hours) and frequency of failure (11.13 hours of functioning) have improved, which is reflected on the fact that produced effluent increased (+0.13 million gallons) along with availability (+0.71%).

As a particular case, it was considered that the amount of time that takes to repair the primary equipment is the same as the time horizon used for the calculation of the percentage mean availability for virtual equipment (BS_002 and GRIT_002). In other words, said percentage indicates the relative amount of time where the virtual equipment operates supporting the primary equipment, which in this case 26.97% for the bar screen and 21.15% for the grit chamber.

6. Discussing Simulation Results and VSB Methodology Advantages

Comparing the results presented in Tables 4 and 5, it is possible to determine the effect the incumbent machines under a failure scenario are capable to provide, by themselves and for a limited time, granting continuity to the subsystem downstream.

To understand the differences between the simulation models (with and without VSB) is relevant to analyse key indicators such as the outcome of the VSB simulation for MTTR (1.72 hours) and frequency of failure (11.13 hours of uninterrupted work) that are higher than the values obtained in the immediate effect scenario (10.31 and 1.67, respectively), supporting the increased production (+0.13 million gallons) and availability (+0.71%) results.

The VSB simulation model generates improvement in the reliability of the process (some specific detentions have no effect on the overall system). From the maintainability

Table 4: Simulation results-immediate effect scenario and comparison with RBD.

Equipment/system	Performance indicators					
	Mean % availability	Mean % oper. time	Mean processing WW (MGD)	MTTF (hours)	MTTR (hours)	RBD availability (%)
Primary treatment	85.91	88.98	7.95	10.31	1.67	86.60
BS_001	96.68	88.98	7.95	45.88	1.4	96.85
GRIT_001	96.19	88.98	7.95	68.86	2.4	96.44
PUMP_001	95.76	88.98	7.95	28.75	1.2	95.88
CLAR_002	96.47	88.98	7.95	98.42	3.2	96.70

Table 5: Simulation results-VSB scenario.

Equipment/system	Performance indicators				
	Mean % availability	Mean % oper. time	Mean processed WW (MGD)	MTTF (hours)	MTTR (hours)
Primary treatment process	**86.62**	**89.72**	**8.08**	**11.13**	**1.72**
Standby bar screening[1]	**97.24**	**89.72**	**8.08**	**93.44**	**3.53**
BS_001	96.68	89.72	7.97	45.88	1.4
BS_002[2]	26.97	89.72	0.11	0.48	1.3
Standby grit chamber[1]	**96.43**	**89.72**	**8.08**	**76.63**	**3.29**
GRIT_001	96.19	89.72	7.99	68.86	2.4
GRIT_002	21.15	89.72	0.09	0.59	2.2
PUMP_001	95.76	89.72	8.08	28.75	1.2
CLAR_002	96.47	89.72	8.08	98.42	3.2

[1]Two new subsystems (standby configuration) are recognized, representing the integration of the main and virtual equipment. The creation of these new subsystems is needed according to evaluation of the resilience operational impact over the indicators of interest. [2]This virtual equipment approaches the impact of resilience condition on the main system and subsystems.

point of view, the real downtime will be reduced or compensated according to the isolation time generated by the upstream system. Therefore, it is important to study in detail each process to understand and find improvement opportunities.

When analysing the simulation results, it is possible to understand the strength of VSB proposal for complex systems. First, when a simulation model is built, the RBD model is relegated because of the supposition of machine independence that avoids the inclusion of the operation continuity effect. When the immediate effect scenario is compared with the results of VSB simulation, the positive impact of the proposal is evident, considering the precision addressed for reliability, maintainability, and availability indicators. Summarizing, it is possible to evidence the following VSB model advantages:

(i) It incorporates dependencies between the machines of a process

(ii) It evaluates the effect of the machines under a failure scenario or subject to delays, being capable to provide continuity to the subsystem downstream

(iii) It adjusts the operational capacity to a specific process condition

(iv) It has the flexibility to include buffering effects or self-autonomy, without complex modelling

(v) In processes with low reliability level, the VSB model will have a high impact because virtual machines will be required with a higher frequency and the operational continuity downstream will be activated when each failure occurs

(vi) In processes where the operational continuity effect is presented in many machines, the VSB model will be more efficient than other methodologies, considering limitations of the representation that RBD and Markov Chains have and the complexity of traditional buffer inventory level modelling, as explained in the Introduction section

(vii) When including VSB proposal, the model will be more accurate in reliability and maintainability assessment, mainly due to the representation of more realistic operational conditions associated to the operational continuity and repair processes

7. Conclusions

Performance analysis must be an integral part of engineering and reliability assessment and operational management, controlling operating plants or evaluating newly designed projects, especially for complex systems. Simulation is a widely used method to estimate indicators such as performance on early stages of development, especially when features such as physical dependency, maintainability, and reliability among others can be embedded in the model.

Evidently the most important outcome of this paper is the validation of the proposal methodology introducing the

VSB effect to improve accuracy when modelling an industrial facility and the development of a case study of a wastewater treatment process (primary treatment).

The obtained indicators show that when using VSB, computed availability increases a 0.71% and consequently so does the produced effluent by the plant. They also evidence critical equipment or possible bottlenecks due to maintainability and reliability issues. These results are detailed in the Simulation Results section. As a summary, the results of the modelling allow the following:

(i) Forecast performance of each equipment, subsystem, and overall wastewater treatment system

(ii) To evidence the equipment with the poorest performance

(iii) Track relevant incumbents on the outcome of performance, especially for reliability and maintainability

(iv) Acknowledge the risk level (probability) for decision making processes

(v) Evaluate the results for the scenarios, and determine the expected effect of VSB operational restriction

Concluding, this proposal has developed an innovative probabilistic methodology to simulate, analyse, and evaluate quantitatively the Virtual Standby (VSB) impact on production performance. A case study in a wastewater treatment line was developed, and the model has allowed to determine different production levels based on VSB impact. It also encouraged the use of this model on the early stages of any project (design stage) to promote highly efficient investments and future productivity.

Data Availability

The data used to support the findings of this study are provided in the Supplementary Materials.

Disclosure

The research work was partially performed within the context of PhD research work of Pablo Viveros and Fredy Kristjanpoller at University of Seville. The research work was performed within the context of UTFSM Project-PI_LIR_2020_5.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Supplementary Materials

This section includes the Event of Failure data used to support this study. (*Supplementary Materials*)

References

[1] E. Goldratt, *The Goal: A Process of Ongoing Improvement*, North River Press, Great Barrington, MA, USA, 1992.

[2] C. Roser, M. Nakano, and M. Tanaka, *A Practical Bottleneck Detection Method*, The Winter Simulation Conference, Arlington, TX, USA, 2001.

[3] L. Hu, W. Huang, G. Wang, and R. Tian, "Redundancy optimization of an uncertain parallel-series system with warm standby elements," *Complexity*, vol. 2018, Article ID 3154360, 10 pages, 2018.

[4] M. Macchi, F. Kristjanpoller, A. Arata, M. Garetti, and L. Fumagalli, "Introducing buffer inventories in the RBD analysis of production systems," *Reliability Engineering & System Safety*, vol. 104, pp. 84–95, 2012.

[5] J. Buzacott and J. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[6] D. Huang and R. Billinton, "Impacts of repair state residence time distributions in an electric power generating capacity adequacy assessment," *Proceedings of the Institution of Mechanical Engineers-Part O: Journal of Risk and Reliability*, vol. 221, pp. 297–305, 2007.

[7] P. Viveros, E. Zio, A. Arata, and F. Kristjanpoller, "Integrated system reliability and productive capacity analysis of a production line. A case study for a Chilean mining process," *Proceedings of the Institution of Mechanical Engineers-Part O: Journal of Risk and Reliability*, vol. 226, pp. 305–317, 2012.

[8] A. Jeang and C. Hun, "Process parameters determination for precision manufacturing," *Quality and Reliability Engineering International*, vol. 16, pp. 33–44, 2000.

[9] K. Das, R. Lashkari, and S. Sengupt, "Reliability consideration in the design and analysis of cellular manufacturing systems," *International Journal of Production Economics*, vol. 105, pp. 243–262, 2007.

[10] A. Christou, "Monte Carlo reliability model for microwave monolithic integrated circuits," *Quality and Reliability Engineering International*, vol. 24, pp. 315–329, 2007.

[11] E. Zio and N. Pedroni, "Building confidence in the reliability assessment of thermal-hydraulic passive systems," *Reliability Engineering & System Safety*, vol. 94, pp. 268–281, 2009.

[12] E. Zio, L. Podofillini, and V. Zille, "A combination of Monte Carlo simulation and cellular automata for computing the availability of complex network systems," *Reliability Engineering & System Safety*, vol. 91, pp. 181–190, 2006.

[13] M. Marseguerra and E. Zio, *Basics of the Monte Carlo Method with Application to System Reliability*, LiLoLe-Verlag GmbH, Hagen, Germany, 2002.

[14] A. Crespo, A. Sánchez, and L. Benoit, "Monte Carlo based assessment of system availability. A case study for cogeneration plants," *Reliability Engineering & System Safety*, vol. 88, pp. 273–289, 2005.

[15] E. Zio and N. Pedroni, "Reliability estimation by advanced Monte Carlo simulation," in *Simulation Methods for Reliability and Availability of Complex Systems*Springer, Berlin, Germany, 2010.

[16] M. Metropolis and S. Ulam, "The montecarlo method," *Journal of the American Statistical Association*, vol. 44, pp. 335–341, 1949.

[17] I. Sobol, *A Primer for the Monte Carlo Method*, CRC Press, Boca Raton, FL, USA, 1994.

[18] J. Vargas, J. Koppe, and S. Pérez, "Monte Carlo simulation as a tool for tunneling planning," *Tunnelling and Underground Space Technology*, vol. 40, pp. 203–209, 2014.

[19] M. López-Campos, F. Kristjanpoller, P. Viveros, and R. Pascual, "Reliability assessment methodology for massive manufacturing using multi-function equipment," *Complexity*, vol. 2018, Article ID 3236986, 8 pages, 2018.

[20] S. Lin, Y. Wang, and L. Jia, "System reliability assessment based on failure propagation processes," *Complexity*, vol. 2018, Article ID 9502953, 19 pages, 2018.

[21] F. Kristjanpoller, P. Viveros, E. Zio, R. Pascual, and O. Aranda, "Equivalent availability index for the performance measurement of haul truck fleets," *Maintenance and Reliability*, vol. 22, no. 4, pp. 583–591, 2020.

[22] C. Hsu and H. Li, "Reliability evaluation and adjustment of supply chain network design with demand fluctuations," *International Journal of Production Economics*, vol. 132, pp. 141–145, 2011.

[23] R. Meller and D. Kim, "The impact of preventive maintenance on system cost and buffer size," *European Journal of Operational Research*, vol. 95, pp. 577–591, 1996.

[24] F. Bernabei, R. Ferretti, M. Listanti, and G. Zingrillo, "A methodology for buffer design in ATM switches," *European Transactions on Telecommunications*, vol. 2, pp. 367–379, 1991.

[25] Y. Lin, "System reliability of a stochastic-flow network through two minimal paths under time threshold," *International Journal of Production Economics*, vol. 124, pp. 382–387, 2010.

[26] J. Sun, L. Xi, S. Du, and B. Ju, "Reliability modeling and analysis of serial-parallel hybrid multioperational manufacturing system considering dimensional quality, tool degradation and system configuration," *International Journal of Production Economics*, vol. 114, pp. 149–164, 2008.

[27] P. Viveros, A. Crespo, F. Kristjanpoller et al., "Probabilistic performance assessment for crushing system. A case study for a mining process," in *Proceedings of the PSAM 12–Probabilistic Safety Assessment and Management*, Honolulu, HI, USA, June 2014.

[28] United Nations Environment Programme, *Sick Water: The Central Role of Wastewater Management in Sustainable Development-A Rapid Response Assessment*, United Nations Environment Programme, Nairobi, Kenya, 2010, https://wedocs.unep.org/handle/20.500.11822/9156.

[29] F. R. Spellman, *Handbook of Water and Wastewater Treatment Plant Operations*, CRC Press, Boca Raton, FL, USA, 4th edition, 2020.

[30] M. Gorjian, M. Lin, M. Murthy, Y. Prasad, and S. Yong, *Engineering Asset Lifecycle Management. A Review on Degradation Models in Reliability Analysis*, Springer, Berlin, Germany, 2010.

[31] A. Birolini, *Quality and Reliability of Technical Systems*, Springer, Berlin, Germany, 1994.

Hindawi
Complexity
Volume 2021, Article ID 8235108, 10 pages
https://doi.org/10.1155/2021/8235108

Research Article

A Mountain Summit Recognition Method Based on Improved Faster R-CNN

Yueping Kong, Yun Wang, Song Guo, and Jiajing Wang

School of Information and Control, Xi'an University of Architecture and Technology, Xi'an 710055, China

Correspondence should be addressed to Yueping Kong; kongyp@xauat.edu.cn

Received 20 May 2021; Revised 21 June 2021; Accepted 4 August 2021; Published 12 August 2021

Academic Editor: Long Wang

Mountain summits are vital topographic feature points, which are essential for understanding landform processes and their impacts on the environment and ecosystem. Traditional summit detection methods operate on handcrafted features extracted from digital elevation model (DEM) data and apply parametric detection algorithms to locate mountain summits. However, these methods may no longer be effective to achieve desirable recognition results in small summits and suffer from the objective criterion lacking problem. Thus, to address these problems, we propose an improved Faster region-convolutional neural network (R-CNN) to accurately detect the mountain summits from DEM data. Based on Faster R-CNN, the improved network adopts a residual convolution block to replace the traditional part and adds a feature pyramid network (FPN) to fuse the features with adjacent layers to better address the mountain summit detection task. The residual convolution is employed to capture the deep correlation between visual and physical morphological features. The FPN is utilized to integrate the location and semantic information in the extracted feature maps to effectively represent the mountain summit area. The experimental results demonstrate that the proposed network could achieve the highest recall and precision without manually designed summit features and accurately identify small summits.

1. Introduction

Mountain summits are essential topographic feature points that are widely utilized in military and nonmilitary domains, such as biodiversity assessment [1], landslide risk analysis [2], and glacier and snow-covered [3] summit analysis. The summit is the area with the maximum elevation from sea level area. Summits are usually located in a complex and giant topographic system, with complex structural and functional differences [4]. Automating the summit detection process will greatly advance and enrich our geospatial knowledge; thus, it is valuable to study effective methods for the automatic detection of mountain summits.

In the literature, there are two main streams of methodologies, heuristic-based methods and data-driven methods, which have been extensively discussed in summit detections. A common trait of the heuristic-based methods is that they rely on features selected by the algorithm designer and landform recognition rules that depend on parameters configured by the user. In a prior work [5], a fuzzy set theory was applied to terrain analysis, which computes the fuzzy membership of each digital elevation model (DEM) pixel to six different morphometric classes, Pass, Pit, Plane, Ridge, Channel, and Peak, which are obtained through the evaluation at multiple scales. In another study [6], a multiscale and multisemantic method, which combines landform attributes and the surrounding environment to compute the membership value of each grid around the mountain summit, was proposed to detect mountain summits. The author considers the mountain to be a fuzzy entity with various attributes, such as topographic relief, average slope, and relative altitude. In another work [7], an accurate summit detection method based on morphological analysis was presented to detect summits, in which the author concluded that the summit should be located in a nonflat area and should be the highest point with respect to the eight adjacent grids around it. Moreover, the author further illustrated that different summits should be

separated from a certain level horizontal and vertical distance. Thus, a 3×3 sliding window is applied on the DEM to find the highest point in a local area and regard it as a summit candidate. Then, the relative distance between neighbors of the candidate is analyzed to more accurately locate the summit. Although the methods above can well address the false detection and missing detection problem of mountain summit detection tasks, most of these methods require manually designed features, and, thus, their representation abilities are limited. Furthermore, it is nontrivial and cumbersome to manually select parameters especially when multiple parameters are involved.

With the rapid development of remote sensing technology, high-resolution DEM data have become easily accessible, which is characterized by complex backgrounds, diverse feature structures, and rich details. Easily accessible high-resolution DEM data have provided an incredible opportunity to study summit detection from a data-driven perspective. The reported data-driven methods can be classified into two categories: machine learning methods and deep learning (DL) methods. Recent surveys of the applications of DL in remote sensing can be found in areas such as scene classification [8], object detection [9, 10], land use, and land cover analysis [11]. In a prior work [12], 446 recorded landslides and landslide-related conditioning factors were acquired, stored, and analyzed through remote sensing and geographic information system technologies. Then, the landslide susceptibility of Ningdu County was predicted using supervised machine learning models (support vector machine and chi-squared automatic interaction detection models) and unsupervised machine learning models (K-means and Kohonen models) based on 11 conditioning factors. In another study [13], three machine learning models, boosted regression tree (BRT), classification and regression tree (CART), and random forest (RF), were compared to produce groundwater spring potential maps.

Recently, DL has been widely used in various computer vision applications, where it can automatically conduct feature selection from data samples. Convolutional neural networks (CNNs) have been successfully used to perform object detection and image recognition [14–16], and CNNs contain a series of mathematical operations, such as convolution, pooling, and thresholding, to automatically learn the target features from low-level semantics. Due to its strong ability to capture the spatial correlation and the more advanced mechanism of feature extraction, CNNs, as well as DL technology, are hot topics in geography. In a prior work [17], a DL method was developed to detect terrain features, including craters, which combines the Faster region CNN (R-CNN) model with a ZF-net architecture to recognize some common cases, such as multiple separated but very close craters and very small craters. In another study [18], a DL approach was proposed for automatic terrain feature identification from remote sensing images, which extends the Faster-RCNN architecture with deep CNNs and adopts ensemble learning to detect nine different types of terrain features. Torres et al. [19] proposed an automatic summit recognition method based on DL. This method regards the summit recognition task as a classification problem and performs well compared to traditional methods. However, the sliding window makes the network only focus on local features, which ignores the summit's overall shape and spatial structure. With appropriately selected network structures, DL methods provide flexible options for better addressing various scenarios of terrain feature identification. However, the study of DL methods in terrain feature identification is still in its infancy, and further exploration is needed to discover its full potential.

In this paper, we focus on how to apply a DL model to summit detection to achieve high accuracy without manually designed features. A mountain summit recognition approach based on the Faster R-CNN framework is proposed for more effective mountain summit detection. The proposed approach borrows ideas from residual convolution [20] and feature pyramid network (FPN) [21] to automatically extract the feature of the mountain summit and directly output the summit's location in an end-to-end manner without setting parameters.

The main contributions of this paper can be summarized as follows: (1) We formalize summit detection as an image processing task to train the DL model with DEM data and locate the boundary coordinates of the summit. (2) We propose an advanced method for identifying summits from DEM data, which uses the residual structure to improve the convolutional layer and merges features of different levels. (3) We created a new summit detection data set, including its location and boundaries, to build the proposed model. (4) A computational experiment demonstrated that the proposed method could outperform the benchmarks, especially in terms of detecting small mountain tops and pseudosummits.

2. Methodology

According to the spatial, scale, and controlling area characteristics, the mountain summit can be classified into several types. Each type of summit has distinctive geological properties that are not easy to represent in a single model. Faster R-CNN is the most representative CNN for object detection. The region proposal network (RPN) is presented for efficient and accurate region proposal generation. It is possible to use a very deep network to improve the overall object detection accuracy by sharing convolutional features with a downstream detection network. However, some limitations of Faster R-CNN, such as poor feature extraction ability and inefficient feature utilization mechanism, result in the tendency to miss small summits. Therefore, the hierarchical structure of FPN is applied in the proposed method to integrate features of different scales to more accurately locate and identify summits. The overall improved Faster R-CNN is illustrated in the schematic diagram shown in Figure 1.

As shown in Figure 1, the improved Faster R-CNN consists of four parts components: a feature extractor, an FPN, an RPN, and a summit classifier. The DEM is fed to the feature extractor to shrink its size and increase the number of channels through a sequence of stacked convolution layers. The output of the feature extractor is a series of feature maps that represent the summit from different

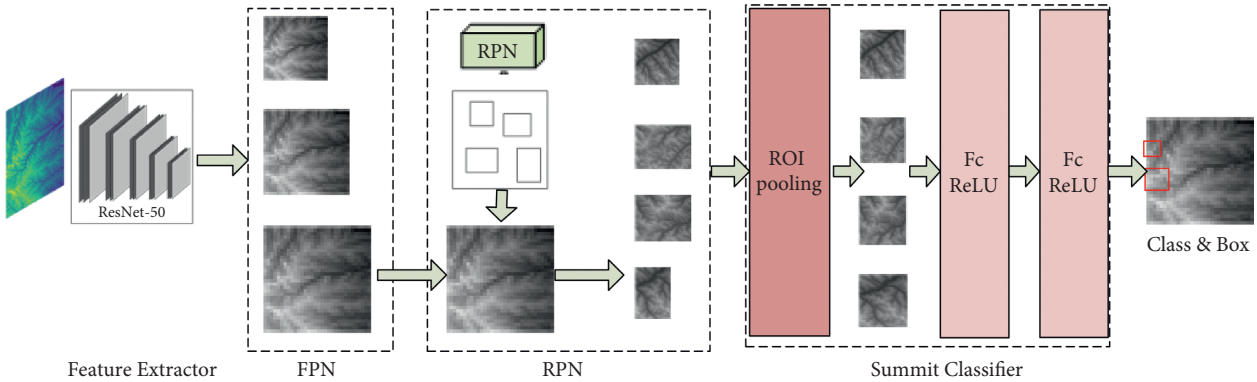


Figure 1: The framework of improved faster R-CNN.

perspectives learned from the data. Next, these feature maps are sent to the FPN to generate several fused feature maps containing the summit's semantic information and location information. Finally, the RPN generates the location of the summit through the fused feature maps. Meanwhile, the parts of the fused feature maps are sent to the summit classifier to determine whether it is a summit.

The performance of a neural network increases with the depth of the network layers. However, neural network models with many layers are subject to problems during training, including gradient vanishing and gradient exploding. The ResNet model effectively addresses these problems by introducing a deep residual framework. In this work, we evaluated the performance of the ResNet-50 and ResNet-101 architectures. ResNet-101 achieved only a 0.1% accuracy improvement, while the computational cost increased significantly. This is because the summits in the DEM are relatively small and their features may no longer be identified in those deeper network levels. We chose ResNet-50 as the feature extractor in the improved Faster R-CNN framework to balance accuracy and computational complexity.

Faster R-CNN only uses RPN to perform region suggestion operations in the last convolutional layer, while the semantic information displayed by small targets in the high-level features is very limited. It is not easy to obtain more comprehensive information to predict the summit location. Figure 2 shows the feature maps extracted by ResNet-50. We can see that shallow features identify edges by comparing the brightness of adjacent pixels, while deeper features can find a specific set of contours and corners to detect the entire part of the summit and finally identify the summit in the image. However, as the number of layers in the network increases, the semantic information in feature maps becomes increasingly prominent, and the location information is gradually blurred. To find all the possible summit-like regions for subsequent inferring, the FPN structure fuses semantic and location information of the mountain summit so that the features at each scale have wealthy semantic information. The improved structure is depicted in Figure 3.

We combine the feature from pyramid levels 4 (P4), 3 (P3), and 2 (P2) to generate the finest feature map. Since the summit is so small that it cannot be retained at this level, the output from the fifth convolutional layer (C5) is excluded for proposal detection. Afterward, the feature maps P2, P3, and P4 are used as an input of the RPN. Based on the location regression layer of the RPN, a regional suggestion box is generated to determine the possible locations of the mountain summit, and the classification layer of the RPN determines the probability of the existence of the summit area in the box. In Faster R-CNN, three anchor boxes of different scales and aspect ratios are predefined manually according to the PASCAL VOC data set. These anchor boxes are used as the reference bounding boxes for the algorithm to predict the target position for the first time. It should be noted that, if we use the default anchor box, the convergence speed of the bounding box regression slows down during the training process of Faster R-CNN. Moreover, once an error occurs in the RPN, it is difficult for the summit classifier to correct because they share some features between them. Therefore, considering the size of the summit areas in the SUMMIT-DEM data set, k-means clustering [22] is used to adjust the size of the anchor box in the proposed network. Table 1 shows the anchor box information of the proposed and previous methods.

3. Experimental Data

Deep learning is much more potent than traditional approaches due to its ability to learn high-level and abstract features from data. Therefore, a large amount of data is needed. Although many large databases, such as VOC [23] and COCO [24], are available in object detection, few publicly available data sets for terrain elements detection are based on optical images.

We use the DEM data marked by NASA [25] to build the samples data set of summits area, named SUMMIT-DEM. The DEM avoids the influence of the illumination and viewing angle on experimental results, but it lacks many details representing the summit areas, such as morphology, orientation, and contrast. We render DEM data into different modes through different visualization technologies to enable the network to represent mountain summit area features better.

Firstly, different elevation values are assigned to different gray scales to achieve three-dimensional terrain expression

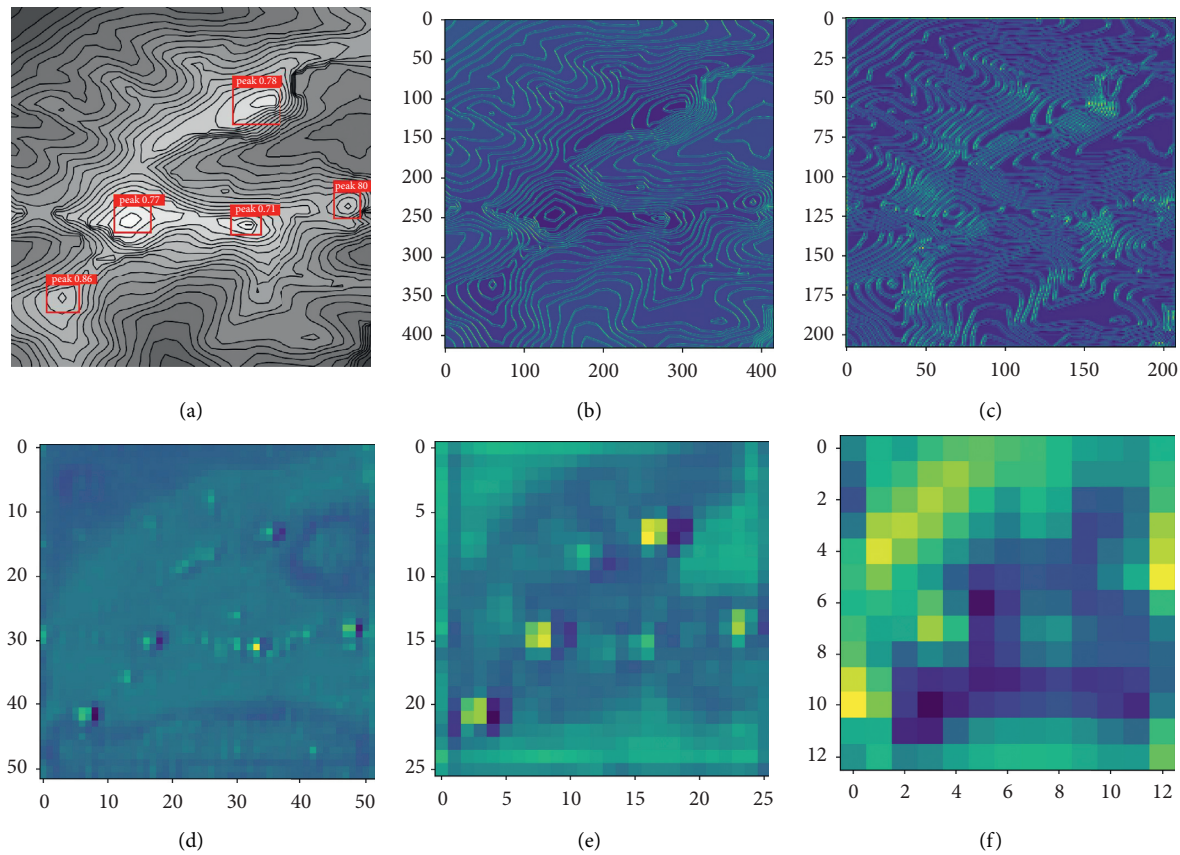


Figure 2: Feature maps from the shallow to deep layers of the input image.

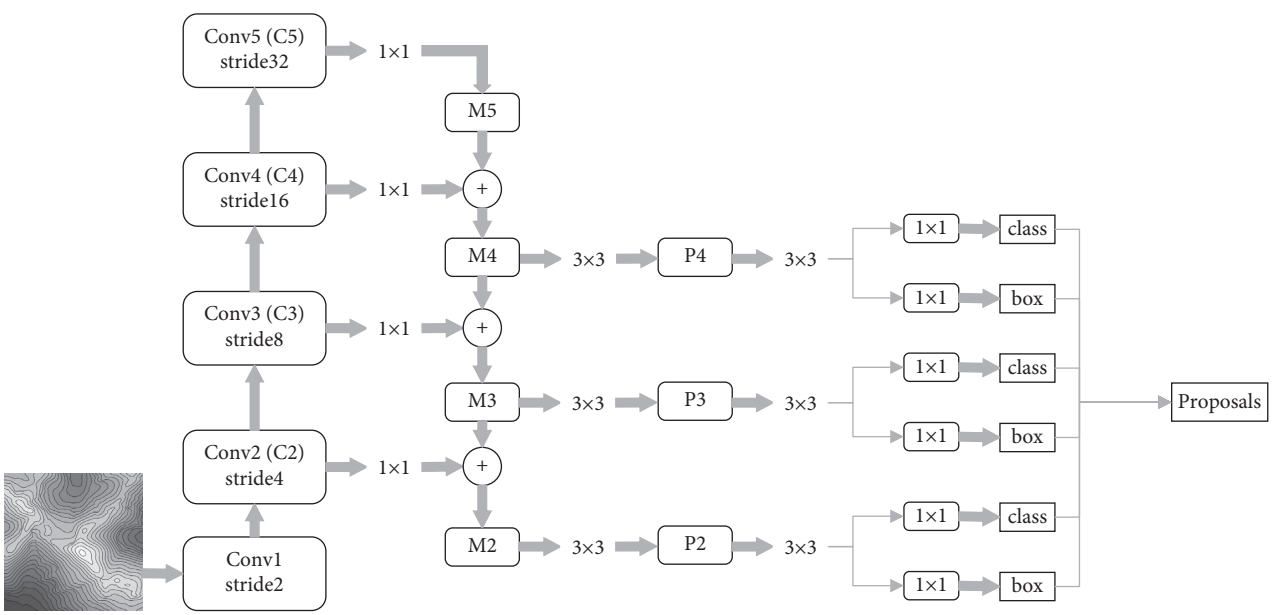


Figure 3: Feature pyramid module structure of the improved faster R-CNN.

on a two-dimensional plane through tonal differences. The range of elevation value is $[H_{\min}, H_{\max}]$, and the corresponding gray range is $[G_{\min}, G_{\max}]$. Then, for any elevation, the corresponding gray value G_i can be calculated through equation (1). After that, the gray value is normalized to between 0 and 1, which can reduce part of the noise in the input data without changing the relative elevation between elements. The converted image is shown in Figure 4(a). Then, the contour lines are generated by a serial of elevation intervals from the DEM, which can scientifically reflect the

Table 1: Anchor information of the proposed RPN and the original RPN.

Method	Feature	Size	Ratio
Original RPN*	C5	128, 256, 512	0.5, 1, 2
Our RPN	P2	42	0.66, 0.58, 0.59
	P3	102	0.66, 0.58, 0.59
	P4	302	0.66, 0.58, 0.59

*The original RPN represents the RPNs of faster R-CNN.

primary geomorphological forms and changes such as ground elevation, mountain body, slope, slope shape, and mountain strike, as shown in Figure 4(b). Finally, to satisfy the input of the network, two kinds of data are superimposed to form the sample shown in Figure 4(c). Figure 4(d) shows a sample of the summit area after annotation.

By this way, we have made the sampling data set SUMMIT-DEM, which includes a total of 1000 images and 3,345 samples of the mountain summit area. The data set was divided into training, verification, and testing set with a ratio of 7 : 2 : 1.

$$G_i = G_{\min} + \frac{G_{\max} - G_{\min}}{H_{\max} - H_{\min}} \times \left(H_i - H_{\min}\right). \tag{1}$$

4. Experiments and Discussion

Two experiments were conducted to evaluate the advantages of the improved Faster R-CNN. Ablation experiments were first conducted on three improved modules (feature extractor, FPN, and anchor box size) to find the contributions of each module. Then, different heuristic-based and DL methods, including Faster R-CNN [26], YOLOv3 [27], SSD [28], and Landserf Peak Classification (LPC) [29, 30], were compared to validate the effect of improved Faster R-CNN. The selection of the methods considered their relevance and heterogeneity along with the availability of the source code or a tool supporting their execution.

4.1. Evaluation Metrics and Parameter Selection. The precision, recall, F1 score, and average precision (AP) were selected to evaluate the performance of the improved Faster R-CNN. These metrics are defined as follows:

$$\begin{aligned} P &= \frac{T_p}{\left(T_p + F_p\right)}, \\ R &= \frac{T_p}{\left(T_p + F_n\right)}, \\ F_1 &= \frac{2 \times P \times R}{(P + R)}, \\ AP &= \frac{1}{11} \sum_{R \in \{0, 0.1, \ldots, 1\}} P(R), \end{aligned} \tag{2}$$

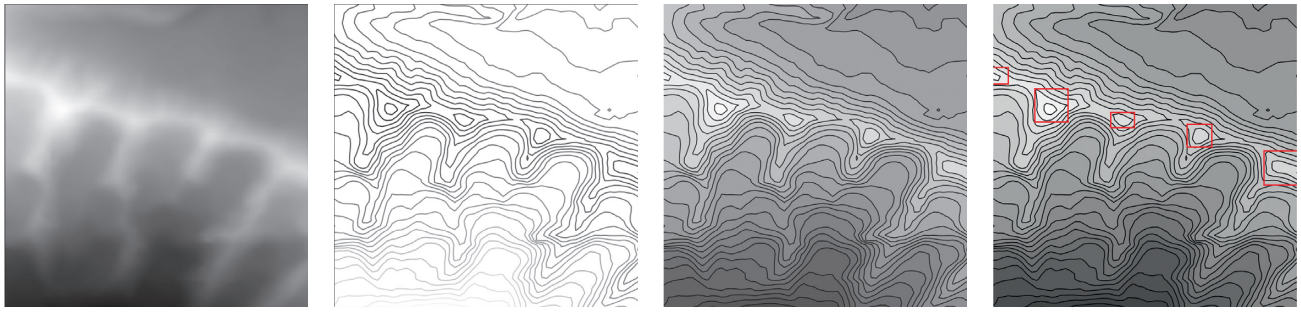
where T_p, F_p, and F_n denote the true positive, false positive, and false negative rates, respectively. Let P be the precision and R be the recall, while F_1 balances P and R. Besides, AP summarizes the shape of the precision and recall curves to avoid the problem that the threshold is difficult to evaluate the effect of the model absolutely, and it is defined as the mean precision at a set of eleven equally spaced recall levels $[0, 0.1, \ldots, 1]$.

Each method was executed with different parameters, the values were sampled from the parameter space, and all the resulting parameter combinations were tested. For the traditional method, LPC took the DEM and two parameters as input. The two parameters were as follows: (1) The minimum height that a point must have had to be considered as a candidate summit. For this parameter, we tested values from 400 m to 6,000 m with a step size of 100 m because these two values were the lowest and the highest elevations of the territory under evaluation. (2) The minimum distance that was the local maxima in a region. We tested values from 900 m to 30 m with a step size of 15 m. This yielded 3,363 configurations. Each configuration made the algorithm run independently once, for a total of 3363 runs. Deep learning models have only one parameter: the probability threshold value to determine if a point is a summit. We tested a value range from 0.01 to 1 with a step of 0.01, yielding 100 configurations. The DL algorithm only ran once, choosing different thresholds to obtain different precision and recall.

4.2. Ablation Experiment. We analyzed the contributions of each module in our method, namely, ResNet-50, feature fusion, and size of anchor box to the overall performance. The experimental results are given in Table 2.

By comparing the results of Faster R-CNN, the replacement of ResNet-50 brought performance improvements on the AP, with a margin of 1.98%. This means that ResNet-50 replaced VGG16 as a feature extractor that can better represent the summit feature. The scale and aspect ratio of the anchor boxes were adjusted in Improved 2, and then the AP increased to 92.97%. By comparing Improved 3 and Faster R-CNN, the addition of the FPN and the adjustment of the anchor box brought performance improvements on the AP, with a margin of 2.82%. This validates the effectiveness of our FPN and anchor box adjustment strategy. Finally, the proposed method (Row 6 in Table 2) was tested, and its AP reached 94.49%. The effectiveness of the three improvements, including the replacement of ResNet-50, the addition of FPN, and the adjustment of the anchor box, was consistently demonstrated.

The feature maps on the SUMMIT-DEM data set are shown in Figure 5. The input images, detection results, and features of the last two layers of Faster R-CNN and improved

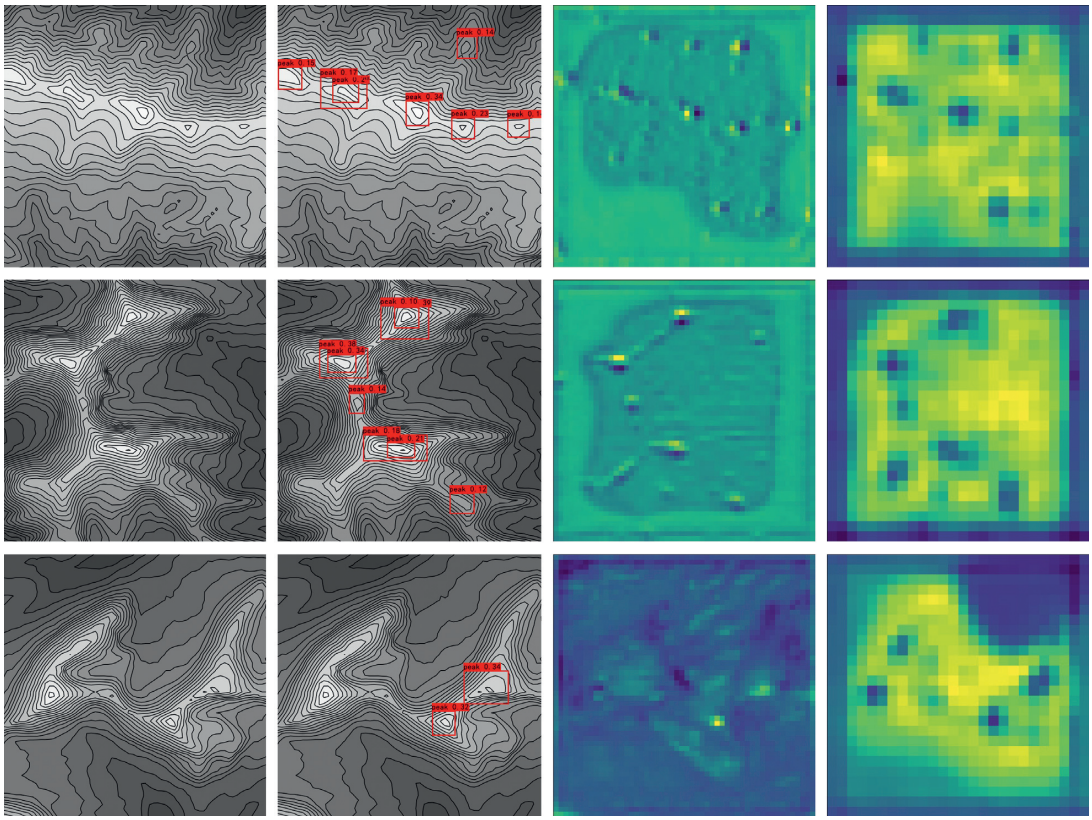


(a) (b) (c) (d)

Figure 4: The sample construction process of mountain summits area. (a) Grayscale image based on elevation data. (b) Contour line image. (c) Superimposed images of (a) and (b). (d) Ground-truth images. The red box is the label of the mountain summit area.

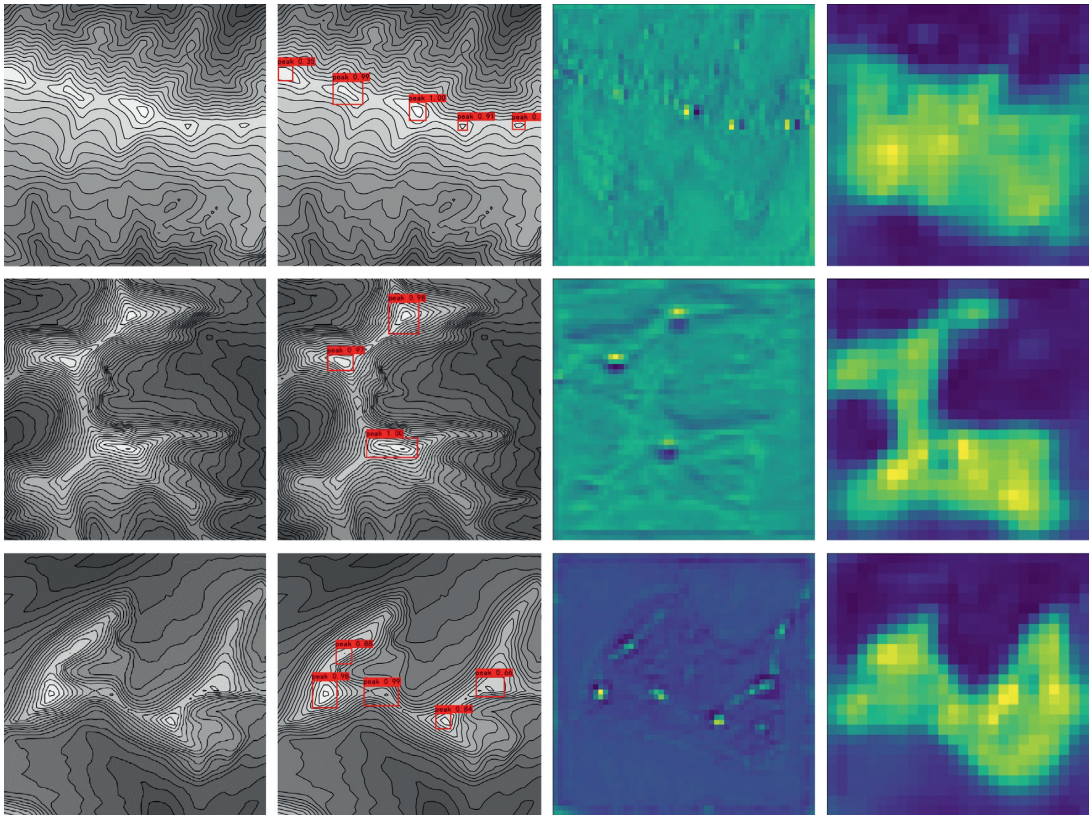
Table 2: Results of the ablation experiments.

Methods	Feature extractor	Feature fusion	Anchor size	AP (%)
Faster R-CNN	VGG16	RPN	(128, 256, 512) {0.5, 1, 2}	90.55
Improved 1	**ResNet-50**	RPN	(128, 256, 512) {0.5, 1, 2}	92.53
Improved 2	VGG16	RPN	**(42, 102, 302) {0.66, 0.58, 0.59}**	92.97
Improved 3	VGG16	**FPN + RPN**	**(42, 102, 302) {0.66, 0.58, 0.59}**	93.37
Ours	**ResNet-50**	**FPN + RPN**	**(42, 102, 302) {0.66, 0.58, 0.59}**	**94.49**



(a)

Figure 5: Continued.



(b)

Figure 5: The feature maps of Faster R-CNN and the improved Faster R-CNN. (a) From left to right, the first to fourth items are the input images, detection results, feature maps of C4, and feature maps of C5. (b) From left to right, the first to fourth items are the input images, detection results, feature maps of P3, and feature maps of P2.

Faster R-CNN are shown in Figure 5. In Column 3, the filter could learn the summit locations in the form of blobs. As expected, the improved Faster R-CNN had more information at its disposal to distinguish false summits and find small summits because it preserved the complete location and semantic features via the FPN. Furthermore, the improved Faster R-CNN could exploit the "context" of a location. In Column 4, the improved Faster R-CNN preserved the correlations among the different locations comprised in the adjacent points, that is, the network paid more attention to the surrounding mountains, which affected the computation of the summit locations and, ultimately, the accuracy.

4.3. Evaluation of Traditional and DL Methods. We compared the proposed method with different methods, including Faster R-CNN (FR), YOLOv3, SSD, and LPC on the SUMMIT-DEM data set.

Figure 6 shows the summit detection results of the different methods. As shown, the proposed model was the closest to the ground truth in various summit categories, including minor, submajor, and major. More importantly, the proposed model could well identify the pseudosummits and find the small summits, demonstrating the effectiveness of the proposed improved Faster R-CNN.

Table 3 reports the recall, precision, F1 score, and AP of the proposed method compared with the other methods. The improved Faster R-CNN achieved excellent results on all the evaluation metrics. The SSD method had the worst performance because of its limited ability to extract shallow features and a hierarchical prediction mechanism, which made the predicted feature maps have a low utilization rate. In particular, comparing Faster R-CNN (Column 3) and the improved Faster R-CNN (Column 2), the recall improved with a margin of 6.45%, which means more summits were discovered. These results clearly illustrate the superior performance and robustness of the improved Faster R-CNN.

The precision-recall curves are shown in Figure 7. The PR curves of the improved Faster R-CNN, represented by the straight blue lines, consistently outperformed all the other methods. Compared with DL methods, the heuristic-based method, represented by the straight purple lines, was more sensitive to parameter changes. At point precision = 0.5 and recall = 0.5, small parameter changes

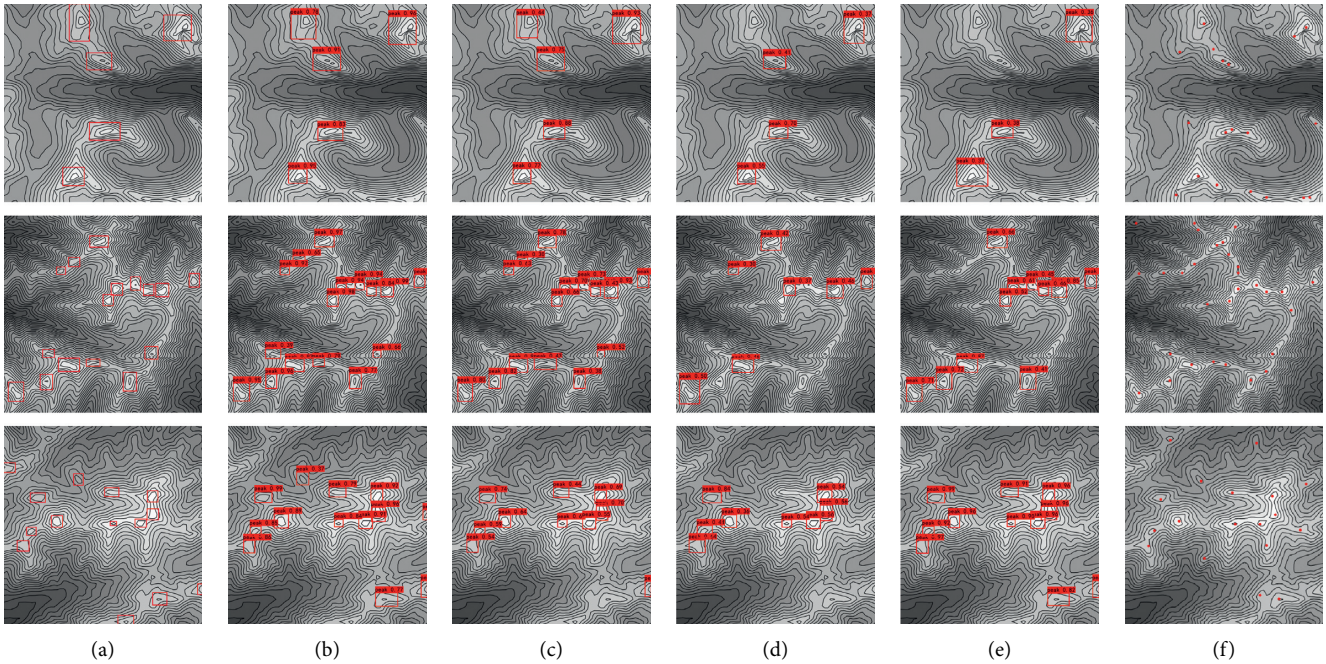


Figure 6: Mountain summit area recognition results of different networks. (a) Ground-truth images. The red boxes denote the ground-truth mountain summits area. (b) The recognition results of the proposed improved Faster R-CNN. (c) The recognition results of YOLOv3. (d) The recognition results of SSD. (e) The recognition results of the network were designed by prior work [26]. (f) The recognition results of LPC.

Table 3: The recognition results of different network models on the test data set.

Measures	Improved FR	FR-based	YOLOv3	SSD	LPC
Recall	**86.63%**	80.18%	81.89%	66.41%	53.79%
Precision	**93.98%**	91.55%	91.16%	90.57%	77.58%
F1 score	**0.90**	0.85	0.86	0.77	0.64
AP	**94.49%**	90.53%	90.69%	85.59%	59.62%

*The boldface in the table is the best result of the evaluation index.

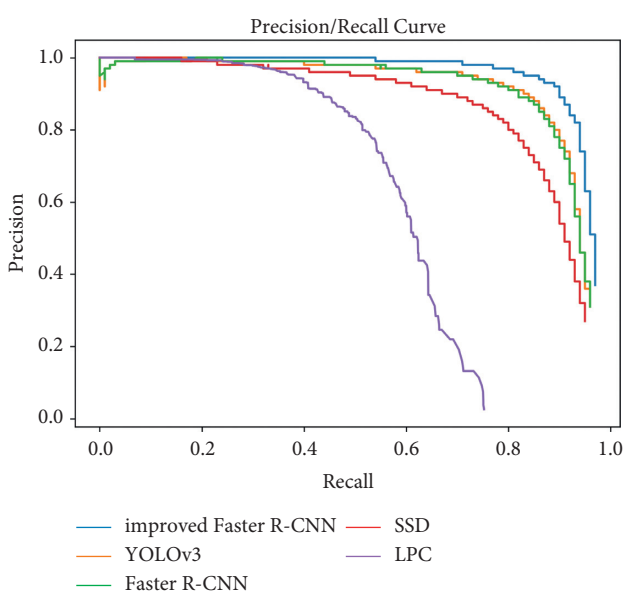


Figure 7: The P-R curve of each algorithm.

could lead to very different results in precision and recall. These results convincingly demonstrate the effectiveness of the proposed model.

5. Conclusions

In this paper, a novel DL method, improved Faster R-CNN, was proposed for summit detection without manually designed summit features. In the improved Faster R-CNN, ResNet-50 is used as the feature extractor to obtain better features of the summit, and the hierarchical structure of the FPN is applied to integrate features of various scales. Benefiting from these two improved modules, efficient information communication across multiple layers is conducted, reducing the information loss during RPN anchor box generation, which leads to more accurate summit detection results. In experimental studies, SUMMIT-DEM data set was used to study the performance of the improved Faster R-CNN. Experiments were conducted in different popular DL and heuristic-based methods, demonstrating the effectiveness and robustness of the improved Faster R-CNN.

Our future work will pursue several directions: (1) Topographic elements are often symbiotic, such as when the saddle is between two summits and the ridge is the connection between the summits; thus, we will improve the model by applying other elements and some methods [31] to process a set of objects simultaneously through interaction between their appearance feature and topology, which could allow modeling of their relations. (2) Multi-information fusion is also a problem worthy of attention. The DEM avoids the influence of the illumination and viewing angle on the experimental results, but it loses many detailed features, such as color, texture, and contrast. We will try to use our method to experiment on remote sensing images in the future and find a method to combine DEM features and remote sensing image features to recognize topographic elements. (3) Due to the conventional nature of cartography, which often only contains prominent mountains for morphological, historical, and cultural reasons, a data set may omit many locations with summit-like characteristics. Therefore, some of the output classified as false positives may indeed be true positives under a complete ground truth. We will apply semisupervised learning [32] to improve the quality of summit data sets, specifically, combining labeled and unlabeled data to change the learning behavior of the network.

Data Availability

The processed data used to support the findings of this study have not been made available because the data also forms part of an ongoing study.

Conflicts of Interest

There are no potential competing interests in our paper.

Acknowledgments

This work was supported by the Foundation of State Key Laboratory of Geo-Information Engineering (Grant no. SKLGIE2018-Z-4-1), the Natural Science Foundation of Shaanxi Province (Grant no. 2019JM-183), and the National Key R&D Program of China (Grant no. 2019YFD1100901).

References

[1] C. Körner, W. Jetz, J. Paulsen, D. Payne, K. Rudmann-Maurer, and E. M. Spehn, "A global inventory of mountains for biogeographical applications," *Alpine Botany*, vol. 127, no. 1, pp. 1–15, 2017.

[2] C. De Jong and T. Barth, "Challenges in hydrology of mountain ski resorts under changing climatic and human pressures," in *Proceedings of the ESA Surface Water Storage and Runoff: Modeling, In-Situ Data and Remote Sensing*, Karlsruhe, Germany, September 2008.

[3] R. Fedorov, A. Camerada, P. Fraternali, and M. Tagliasacchi, "Estimating, ESA Proceedings, 2008 snow cover from publicly available images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1187–1200, 2016.

[4] S. Yamada, "Mountain ordering: a method for classifying mountains based on their morphometry," *Earth Surface Processes and Landforms*, vol. 24, no. 7, pp. 653–660, 1999.

[5] P. Fisher and J. Wood, "What is a mountain? Or the Englishman who went up a boolean geographical concept but realised it was fuzzy," *Geography*, vol. 83, no. 3, pp. 247–256, 1998.

[6] Y. Deng and J. P. Wilson, "Multi-scale and multi-criteria mapping of mountain peaks as fuzzy entities," *International Journal of Geographical Information Science*, vol. 22, no. 2, pp. 205–218, 2008.

[7] T. Podobnikar, "Mountains'peaks determination supported with shapes analysis," *Geographia Technica*, vol. 5, pp. 111–119, 2010.

[8] R. Pires de Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: transfer learning analysis," *Remote Sensing*, vol. 12, no. 1, p. 86, 2020.

[9] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 294–308, 2020.

[10] Z. Huang, B. Sui, J. Wen et al., "An intelligent ship image/video detection and classification method with improved regressive deep convolutional neural network," *Complexity*, vol. 2020, Article ID 1520872, 11 pages, 2020.

[11] X. Zhang, L. Han, L. Han, and L. Zhu, "How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery?" *Remote Sensing*, vol. 12, no. 3, p. 417, 2020.

[12] Z. Chang, Z. Du, F. Zhang et al., "Landslide susceptibility prediction based on remote sensing images and GIS: comparisons of supervised and unsupervised machine learning models," *Remote Sensing*, vol. 12, no. 3, p. 502, 2020.

[13] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, "GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran[J]," *Environmental Monitoring and Assessment*, vol. 188, no. 1, pp. 1–27, 2016.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] J. Gu, Z. Wang, J. Kuen et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[16] H. Sang, L. Xiang, S. Chen et al., "Image recognition based on multiscale pooling deep convolution neural networks," *Complexity*, vol. 2020, Article ID 618031, 13 pages, 2020.

[17] W. Li, B. Zhou, C. Y. Hsu et al., "Recognizing terrain features on terrestrial surface using a deep learning model: an example with crater detection," in *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, pp. 33–36, Los Angeles Area, CA, USA, 2017.

[18] W. Li and C.-Y. Hsu, "Automated terrain feature identification from remote sensing imagery: a deep learning approach," *International Journal of Geographical Information Science*, vol. 34, no. 4, pp. 637–660, 2020.

[19] R. N. Torres, P. Fraternali, F. Milani et al., "A deep learning model for identifying mountain summits in digital elevation model data," in *Proceedings of the 2018 IEEE first international conference on artificial intelligence and knowledge engineering (AIKE)*, pp. 212–217, IEEE, Laguna Hills, CA, USA, September 2018.

[20] K. He, X. Zhang, S. Ren et al., "Deep Residual Learning for Image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[21] T. Y. Lin, P. Dollár, R. Girshick et al., "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, pp. 281–297, Los Angeles, CA, USA, June 1967.

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[24] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Computer Vision-ECCV 2014*, pp. 740–755, Prague, Czech Republic, May 2014.

[25] T. G. Farr and M. Kobrick, "Shuttle radar topography mission produces a wealth of data," *Eos, Transactions American Geophysical Union*, vol. 81, no. 48, pp. 583–585, 2000.

[26] S. Ren, K. He, R. Girshick et al., "Faster R-CNN: towards real-time object detection with region proposal networks," 2015, http://arxiv.org/abs/1506.01497.

[27] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[28] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multiBox detector," in *European Conference on Computer Vision-ECCV 2016*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[29] J. Wood, "The geomorphological characterisation of digital elevation models," Thesis, University of Leicester, Leicester, England, 1996.

[30] J. Wood, "Chapter 14 geomorphometry in LandSerf," *Developments in Soil Science*, vol. 33, pp. 333–349, 2009.

[31] H. Hu, J. Gu, Z. Zhang et al., "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, Salt Lake City, UT, USA, June 2018.

[32] A. Tarvainen and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," 2017, http://arxiv.org/abs/1703.01780.

Hindawi
Complexity
Volume 2021, Article ID 9997783, 11 pages
https://doi.org/10.1155/2021/9997783

Research Article

Three Survival-Related Genes of Esophageal Squamous Cell Carcinoma Identified by Weighted Gene Coexpression Network Analysis

Di Lu,[1] He Wang,[2] Xuanzhen Wu,[1] Jianxue Zhai,[1] Xiguang Liu,[1] Xiaoying Dong,[1] Siyang Feng,[1] Xiaoshun Shi,[1] Jianjun Jiang,[1] Zhizhi Wang,[1] Zhiming Chen,[1] Shuhua Zhao,[3] Jinhua Zhong,[3] Gang Xiong,[1] Hua Wu,[1] Haofei Wang,[1] and Kaican Cai[1]

[1]*Department of Thoracic Surgery, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China*
[2]*Department of Thoracic Surgery, Peking University Shenzhen Hospital, Shenzhen 518000, China*
[3]*Department of Biological Information Research, HaploX Biotechnology, Shenzhen 518000, China*

Correspondence should be addressed to Kaican Cai; doc_cai@163.com

Received 23 March 2021; Accepted 24 July 2021; Published 9 August 2021

Academic Editor: Chao Huang

Background. The aim of this study was to identify novel biomarkers associated with esophageal squamous cell carcinoma (ESCC) prognosis. *Methods*. 81 ESCC samples collected from The Cancer Genome Atlas (TCGA) were used as the training set, and 179 ESCC samples collected from the Gene Expression Omnibus database (GEO) were used as the validation set. The protein-coding genes of 25 samples from patients who completed the follow-up in TCGA were analyzed to construct a coexpression network by weighted gene coexpression network analysis (WGCNA). Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analyses were performed for the selected genes. The least absolute shrinkage and selection operator (LASSO) Cox regression model was constructed to analyze survival-related genes, and an optimal prognostic model was developed as well as evaluated by Kaplan–Meier and ROC curves. *Results*. In this study, a module containing 43 protein-coding genes and strongly related to overall survival (OS) was identified through WGCNA. These genes were significantly enriched in retina homeostasis, antimicrobial humoral response, and epithelial cell differentiation. Besides, through the LASSO regression model, 3 genes (PDLIM2, DNASE1L3, and KRT81) significantly related to ESCC survival were screened and an optimal prognostic 3-gene risk prediction model was constructed. ESCC patients with low and high OS in both sets could be successfully discriminated by calculating a risk score with the linear combination of the expression level of each gene multiplied by the LASSO coefficient. *Conclusions*. Our study identified three novel biomarkers that have potential in the prognosis prediction of ESCC.

1. Introduction

Esophageal cancer (EC) is a highly aggressive malignancy and one of the leading causes of cancer-related death worldwide [1]. There are two histologic subtypes of EC: esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC). Among them, ESCC is the main type of EC, accounting for about 90% [2]. Currently, the main treatments for ESCC include chemotherapy, radiation, and surgery [3]. Despite significant advances in the treatment of ESCC in recent years, the overall 5-year survival rate for ESCC patients is still less than 25%, and the prognosis is still poor, with metastasis and recurrence frequently occurring [4]. In addition, due to late diagnosis and the lack of effective targeted therapy, most patients are diagnosed at an advanced stage [5]. Therefore, it is urgent to explore new therapeutic methods and new therapeutic targets.

Although many genes and mechanisms have been shown to be closely related to the occurrence and development of ESCC, the overall genes and regulation of ESCC

are still unclear. In recent years, with the development of high-throughput technology, bioinformatics has been increasingly applied to explore diseases, molecular mechanisms, and find biomarkers for diseases [6]. The Cancer Genome Atlas (TCGA) contains genomic and clinical information on many types of cancers, which is publicly available [7]. For example, Shergalis et al. performed bioinformatics analysis using TCGA and found that 20 genes were overexpressed and associated with poor survival outcomes in GBM patients [8]. Weighted gene coexpression network analysis (WGCNA) is one of the important methods to understand gene function and gene association from whole genome expression. It can be used to detect coexpression modules of highly related genes and modules of interest related to clinical features, providing good insights for predicting the function of coexpressed genes and discovering genes that play a key role in human diseases [9]. The least absolute shrinkage and selection operator (LASSO) is a penalty regression method that can be used to analyze gene expression profiles. In addition, due to the high dimension and collinearity of the Lasso Cox regression model, it can be combined with WGCNA for biomarker identification [10]. Therefore, a better understanding and application of the above methods are more conducive to the identification of ESCC-related genes and to explore their potential clinical roles and molecular mechanisms to understand the occurrence and development of ESCC. It has been applied to gene-network constructions and survival model identification for many kinds of cancer like lung adenocarcinoma [11]. In this study, RNA-Seq data from TCGA and microarray data from the Gene Expression Omnibus (GEO) database of ESCC were downloaded. Then, we conducted a WGCNA-based analysis in order to identify key modules and hub genes associated with ESCC pathogenesis. Additionally, the potential functions of genes in this identified module were analyzed by gene ontology (GO) and pathway-enrichment analyses. More importantly, a 3-gene risk prediction model was constructed using Cox and LASSO regression models, which could help us better predict ESCC prognosis.

2. Methods

2.1. Gene Expression Data and Clinical Data. Gene expression data and clinical data for patients with ESCC were obtained from TCGA (https://cancergenome.nih.gov/) and the GEO data repository (http://www.ncbi.nlm.nih.gov/geo/) on September 20, 2018, and included 173 samples and 358 samples, respectively. Gene expression levels from TCGA were measured by RNA sequencing, denoted by fragments per kilobase of transcript per million mapped reads (FPKM): FPKM = $10^9 \times$ number of reads mapped to the gene/(number of reads mapped to all protein-coding genes $\times$ length of the gene in base pairs). Gene expression profiles from GEO were assessed by the Agilent-038314 CBC *Homo sapiens* microarray V2.0 and were analyzed using the GeneSpring software V11.5 (Agilent). Clinical information, including pathologic TNM stage and follow-up information, was also collected. This research flow chart is shown in Figure 1.

2.2. WGCNA Network Construction and Module Detection. A weighted gene coexpression network was constructed using the WGCNA package in *R*. The adjacency coefficient (a_{ij}) was calculated by the absolute value of Pearson's correlation coefficient of genes i and j to the power of β. $a_{ij} = |\mathrm{cor}(x_i, x_j)|^{\beta}$, where x_i is the series of expression values for gene i. The lowest power β is chosen when the scale-free topology fit index curve flattens out upon reaching a high value. In addition to considering the connection between two correlated genes, WGCNA also takes into account associated genes, and the topological overlaps (T_{ij}) are calculated from a_{ij} as follows, to compose a topological overlap matrix (TOM), as a similarity evaluation reflecting relevancy and overlap between genes:

$$Tij = \begin{cases} \dfrac{lij + aij}{\min\{ki, kj\} + 1 - aij}, & i \neq j, \\ 1, & i = j, \end{cases}$$
$$lij = \sum_{u \neq i,j} aiuauj, \tag{1}$$
$$ki = \sum_{u \neq i} aiu,$$

where u represents the common genes linking genes i and j together and T_{ij} takes account of the overlap between neighboring genes of genes i and j. TOM is subtracted from one and converted into a topological overlap dissimilarity matrix referred to as the corresponding dissimilarity of TOM (dissTOM). A hierarchical clustering tree (dendrogram) of genes is then created based on dissTOM. Finally, modules of highly correlated and coexpressed genes are created via a Dynamic Tree Cut algorithm.

2.3. Relating Modules to External Clinical Traits and Identifying Hub Genes. Correlations between modules and clinical traits, including pathologic stage and survival time, were estimated by Spearman's correlation tests. Significantly correlated modules were visualized using Cytoscape 3.7.1. Genes with multiple links were defined as hub genes.

2.4. Gene Ontology and Pathway-Enrichment Analysis. The potential biological functions and signaling pathways of the genes in the selected modules were analyzed by enrichment GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in Metscape (http://metascape.org).

2.5. LASSO Cox Regression Model Construction. LASSO Cox regression models were constructed using the glmnet package in R. By utilizing several hub genes from the selected modules, the function returned a series of values of λ and models. The coefficients of most original genes were penalized to zero in line with increasing values of the tuning parameter λ. λ was chosen when the partial likelihood deviance reached its lowest. A suitable model was then chosen

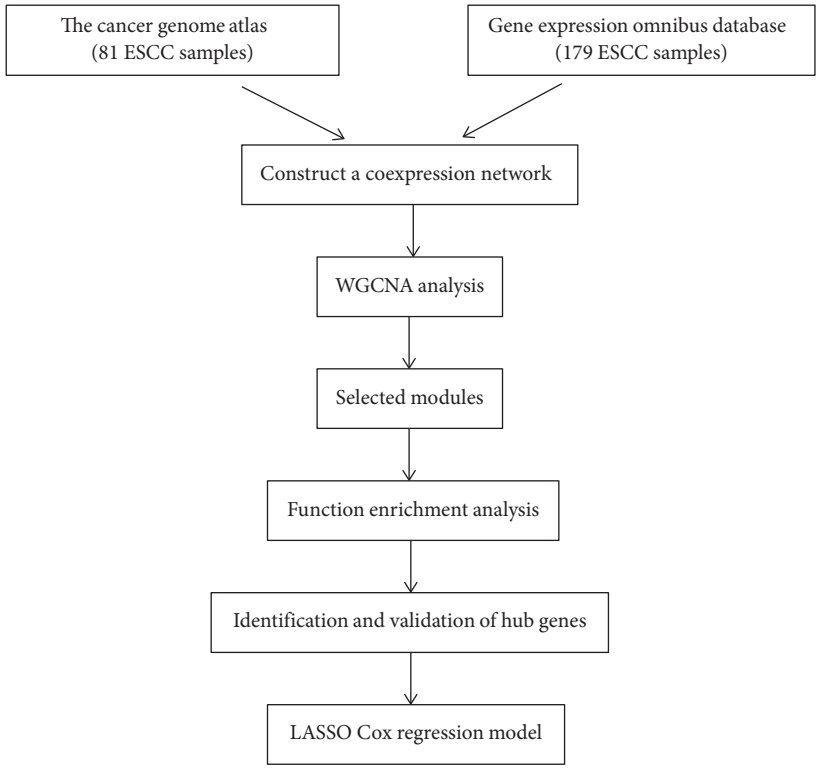


Figure 1: This flow chart of this study.

by 10-fold cross-validation using the function cv.glmnet. Using the function lambda.min, the remaining genes with nonzero LASSO coefficients were obtained. The risk score for each ESCC patient was calculated by a linear combination of the FPKM of each gene (G_k) multiplied by the LASSO coefficient (c_k):

$$\text{risk score} = \sum_{k=1}^{n} G_k \times c_k. \quad (2)$$

We followed the methods and routes of Wang et al. [11] to accomplish the gene network and survival model construction.

2.6. Statistical Analysis. Statistical analyses were conducted using SPSS (version 20.0). Receiver operating characteristic (ROC) curves were drawn, and the areas under the curves (AUC) were calculated to predict 1-year survival. The cut-off risk score was decided when the Youden index (sensitivity plus specificity minus 1) for the ROC curve was highest. The samples were then divided into high-risk and low-risk groups according to the cut-off. Survival was compared between the high- and low-risk groups using univariate and multivariable Cox regression analyses which were used to calculate the Hazard ratios (HRs). The independent significance of different factors was tested by a multivariate Cox regression analysis using backward selection to remove nonsignificant variables from the analysis. The P value threshold was 0.10 ($P > 0.10$).

3. Results

3.1. Data Preprocessing. Samples from TCGA were regarded as a training set and used for network and LASSO Cox regression model construction. A total of 173 samples from patients diagnosed between the ages of 28 and 90 years and classified as stage IA–stage IV were collected from TCGA, and 81 samples with both gene expression and clinical information were used for subsequent analysis. 25 samples from patients who completed the follow-up were subjected to sample clustering, and no outlier samples were removed before network construction (Figure 2(a)). The threshold for average gene expression value was set as 1. Protein-coding genes with average expression values less than the threshold value in all samples were excluded. A total of 25 samples including 12,439 protein-coding genes and clinical information were obtained for WGCNA.

Samples from GEO were regarded as a validation set and used for model verification. The microarray platform provided only a 60-base sequence for each probe, and the maximal exact matches (MEM) algorithm of the Burrows–Wheeler Alignment (BWA) Tool was used for sequence alignment. The parameters used were –t 6 and –k 19, and genome hg19 was used for reference. It was required that a transcript completely covered the genome position compared with the probe. 45,099 symbols were retained as protein-coding genes, and then levels of each probe were subjected to exponential transformation. One hundred and seventy-nine samples of paired normal tissues were removed, and the final validation set comprised 179 ESCC samples.

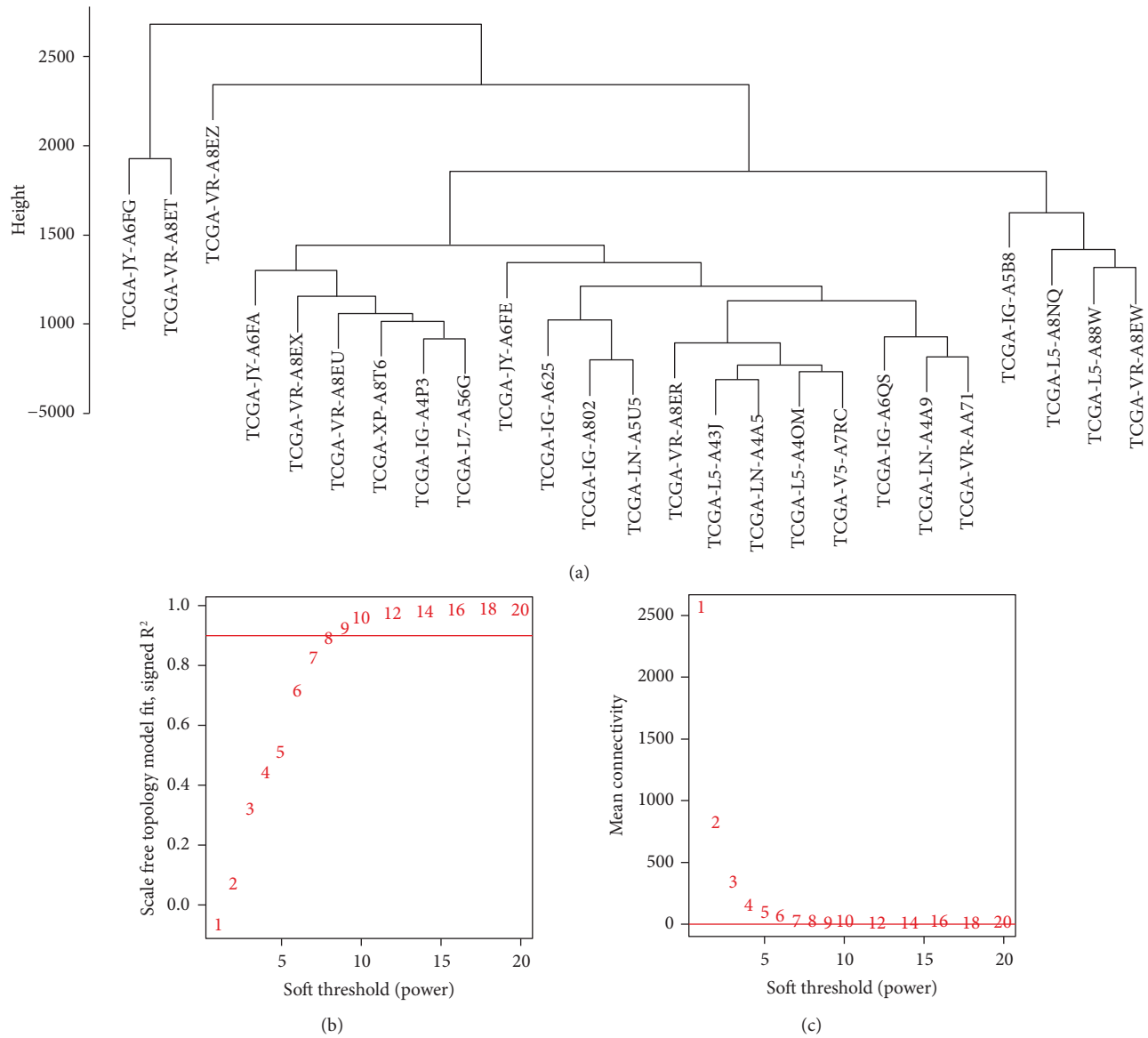
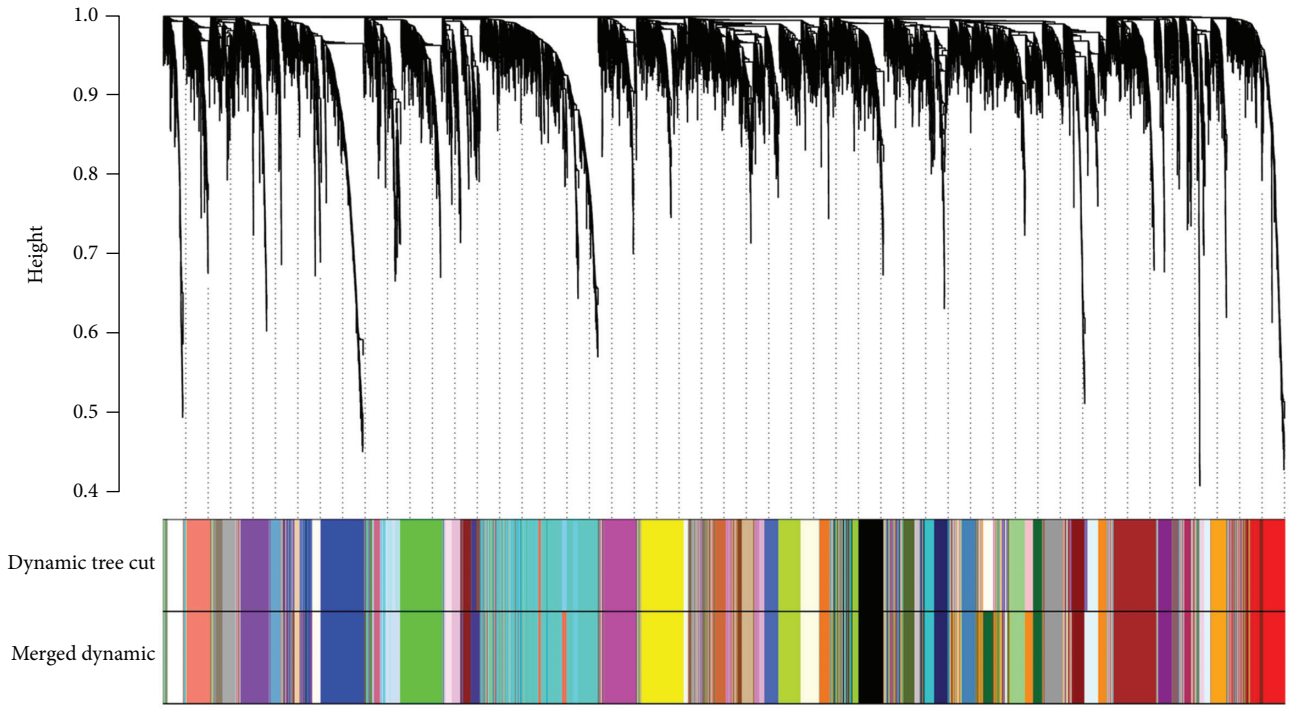


Figure 2: Sample clustering and determination of soft-threshold power in the WGCNA. (a) Sample clustering. Clustering dendrogram of samples based on their Euclidean distances. No outlier samples were removed. (b) β decision. Scale-free topology fit R^2 and series of soft thresholds. The red line indicates an R_2 value of 0.90. (c) Mean connectivity and series of soft thresholds. The red line indicates a mean connectivity value of 0.

3.2. Weighted Gene Coexpression Network of ESCC. As presented in Figure 2(b), the soft threshold is higher with the elevated R2, indicating that the network closely approaches to scale-free distribution. In this study, the soft thresholding power β was set as 8 and the scale-free topology fit index curve flattened out at 0.90 (Figure 2(c)). Then, cluster dendrogram was constructed and dynamic tree cut was performed (Figure 3(a)). Specifically, the constructed weighted gene coexpression network included 55 modules, including 36–875 genes. 36 genes that were not successfully integrated into any other modules were integrated into the gray module and were omitted in downstream analysis.

3.3. Identifying Modules with Clinical Significance. The correlations between each module and clinical traits including pathologic TNM stage and survival were calculated. Among them, 6 modules were positively correlated pathologic TNM stage, whereas 2 modules are negatively correlated with pathological TNM staging. Besides, the dark seagreen4 module and thistle2 module are positively correlated with survival time, suggesting that these two modules may be involved in ESCC tumorigenesis (Figure 3(b)).

3.4. Functional Characterization of Genes in the Selected Module. To explore the biological function of the identified genes in the dark seagreen4 and thistle2 modules, GO term and KEGG pathway-enrichment analyses were performed on 43 and 106 genes in the two modules, respectively. Among the genes in the dark seagreen4 module, retina homeostasis, antimicrobial humoral response, and epithelial cell differentiation were the most significantly enriched



(a)

	Thistle2	Dark seagreen4	Maroon	White	Orange red4	Light cyan1	Orange	Red	Light pink4	Dark magenta	Dark slate blue	Brown4	Plum2	Magenta	Violet	Sienna3	Green yellow	Tan	Saddle brown	Royal blue	Dark grey	Bisque4	Purple	Lavender blush3	Dark orange2	Light green	Dark orange	Dark green
Tumor stage	−0.19 (0.40)	0.04 (0.80)	0.16 (0.50)	−0.03 (0.90)	−0.29 (0.20)	−0.30 (0.10)	−0.40 (0.04)	−0.36 (0.08)	−0.16 (0.40)	−0.37 (0.07)	0.33 (0.10)	0.11 (0.60)	0.51 (0.01)	0.30 (0.10)	0.50 (0.01)	0.55 (0.00)	0.40 (0.05)	0.54 (0.01)	0.62 (0.00)	0.40 (0.05)	−0.22 (0.30)	−0.06 (0.80)	0.38 (0.06)	0.26 (0.20)	0.26 (0.20)	0.04 (0.80)	−0.01 (1.00)	−0.05 (0.80)
Overall survival	0.47 (0.02)	0.48 (0.01)	0.33 (0.10)	0.08 (0.70)	0.28 (0.20)	0.25 (0.20)	0.15 (0.50)	0.18 (0.40)	0.07 (0.70)	0.11 (0.60)	−0.21 (0.30)	−0.04 (0.90)	−0.09 (0.70)	−0.22 (0.30)	−0.03 (0.90)	−0.23 (0.30)	−0.10 (0.60)	−0.06 (0.80)	−0.13 (0.50)	−0.03 (0.90)	−0.17 (0.40)	−0.01 (1.00)	0.21 (0.30)	0.12 (0.60)	0.15 (0.50)	0.01 (0.60)	0.15 (0.50)	0.05 (0.80)

	Midnight blue	Oliver green	Light cyan	Dark red	Steel blue	Light yellow	Grey60	Navajo white2	Yellow	Dark turquoise	Yellow green	Black	Honeydew1	Salmon	Pale violet red	Light steel blue1	Green	Thistle1	Plum1	Sky blue3	Floral white	Blue	Cyan	Coral1	Turquoise	Brown	Grey
Tumor stage	0.10 (0.60)	0.07 (0.80)	−0.04 (0.90)	0.14 (0.50)	0.23 (0.30)	0.38 (0.06)	0.29 (0.20)	0.02 (0.90)	0.32 (0.10)	0.46 (0.02)	0.13 (0.50)	0.20 (0.30)	0.21 (0.30)	0.30 (0.10)	−0.06 (0.80)	0.12 (0.60)	−0.44 (0.03)	0.00 (1.00)	−0.11 (0.60)	−0.20 (0.30)	−0.15 (0.50)	−0.14 (0.50)	0.08 (0.70)	−0.05 (0.80)	0.10 (0.60)	−0.09 (0.70)	0.01 (1.00)
Overall survival	0.10 (0.60)	0.10 (0.70)	−0.03 (0.90)	0.20 (0.30)	0.11 (0.60)	0.01 (1.00)	−0.02 (0.90)	−0.01 (1.00)	0.22 (0.30)	−0.04 (0.90)	0.36 (0.08)	0.12 (0.60)	0.40 (0.05)	0.15 (0.50)	0.31 (0.10)	−0.01 (0.70)	−0.18 (0.40)	−0.32 (0.10)	0.03 (0.90)	−0.07 (0.70)	−0.07 (0.70)	0.08 (0.70)	−0.24 (0.20)	−0.09 (0.70)	−0.26 (0.20)	−0.08 (0.70)	0.32 (0.10)

−1 −0.5 0 0.5 1

(b)

Figure 3: WGCN and their module-trait associations identified by the WGCNA. (a) Weighted gene coexpression network of ESCC identified 55 modules. A dendrogram was produced based on the WGCNA package in *R* by average linkage hierarchical clustering of 12,439 protein-coding genes. (b) Module-trait associations. Each column represents a module eigengene, and each row represents a clinical trait. Each cell contains the correlation coefficient (first line) and *P* value (in parentheses). A *P* value < 0.05 using Spearman's correlation test was considered statistically significant. ESCC: esophageal squamous cell carcinoma; ME: module eigengene; TNM: tumor-node-metastasis; WGCN: weighted gene coexpression network; WGCNA: weighted gene coexpression network analysis.

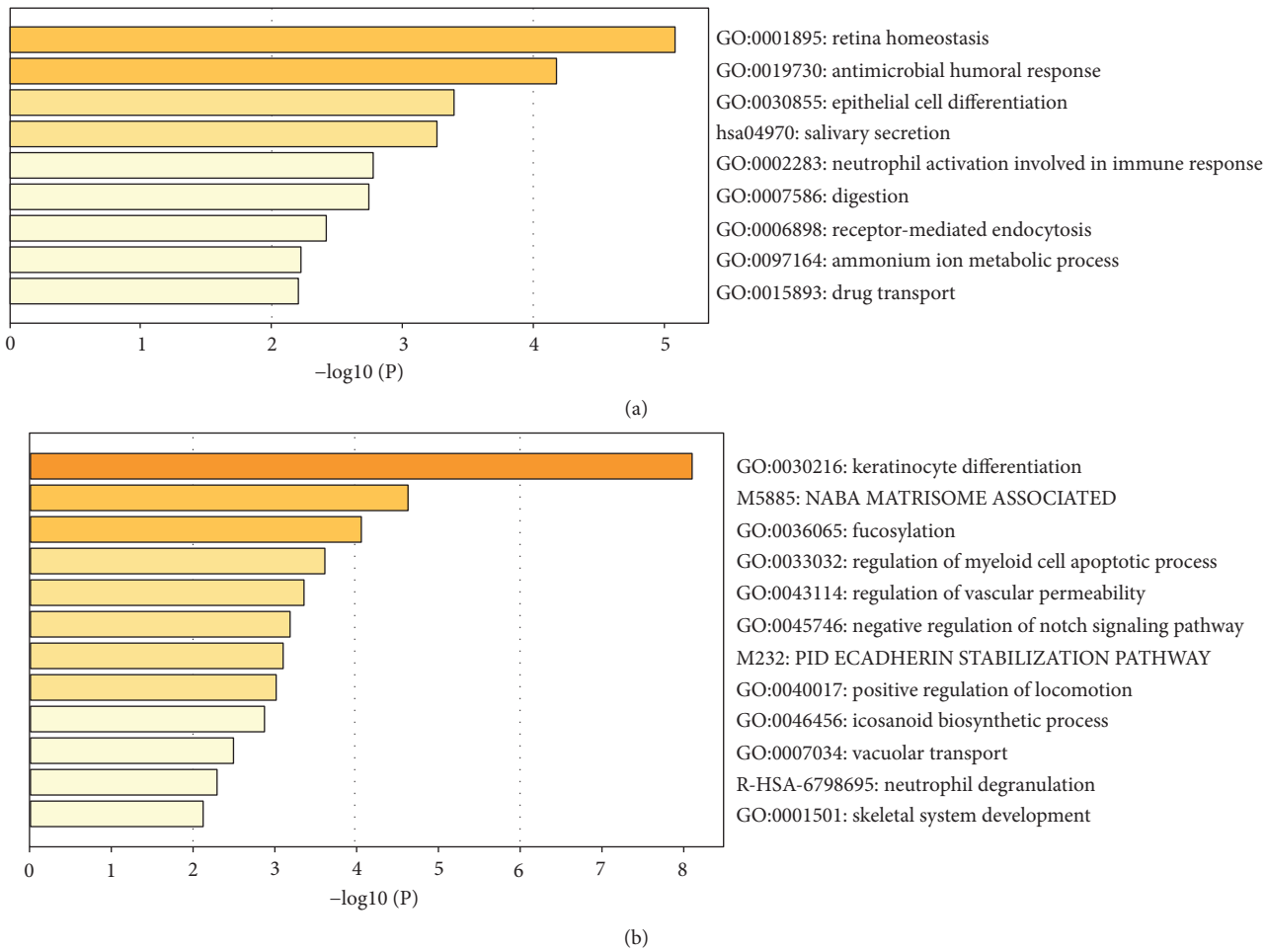


Figure 4: Top clusters in the dark seagreen4 module and the thistle2 module: (a) heatmap of enrichment clusters in the dark seagreen4 module; (b) heatmap of enrichment clusters in the thistle2 module.

(Figure 4(a)). While keratinocyte differentiation, NABA MATRISOME ASSOCIATED, and fucosylation were the most significantly enriched among the genes in the thistle2 module (Figure 4(b)).

Further, we analyzed the dark seagreen4 module with the strongest correlation with survival time by Cytoscape 3.7 to screen for hub genes. As shown in Figure 5, the top 20 genes with the most links were defined as hub genes in the dark seagreen4 module, including *AZGP1*, *CRISP3*, *KRT13*, *PDLIM2*, *PIGR*, *BPIFB1*, *BPIFB2*, *CLIC3*, *DNASE1L3*, *ENDOU*, *KRT81*, *MUC5B*, *PRR4*, *SCGB3A1*, *TFF3*, *SPINK7*, *ASPG*, *KRT6C*, *AQP5*, and *PCP4*.

3.5. Prognostic Signature Construction via LASSO Cox Regression Model Using the Training Set. The LASSO Cox regression model was constructed using the glmnet package in *R* by utilizing several hub genes in the dark-seagreen4 modules. Three genes (*PDLIM2*, *DNASE1L3*, and *KRT81*) with nonzero coefficients at the selected λ were obtained (Figures 6(a) and 6(b)). Based on the genes with nonzero coefficients, the risk score of every patient was calculated according to the linear combination of the expression of each gene multiplied by the LASSO coefficients: $(-0.0529 \times \text{PDLIM2}) + (0.0045 \times \text{DNASE1L3}) + (-0.0021 \times \text{KRT81})$.

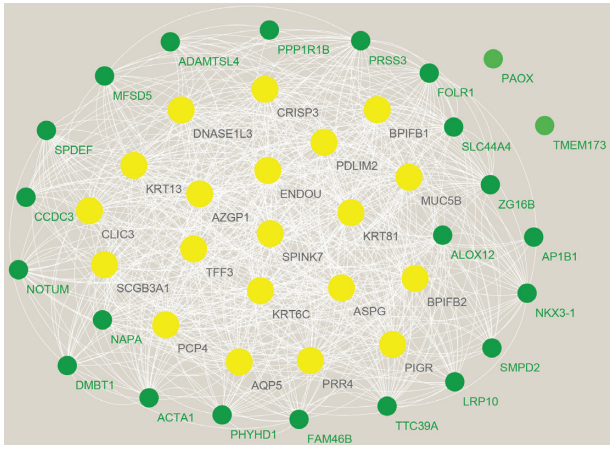


Figure 5: Coexpression network of the dark seagreen4 module. There are 43 connected genes in the dark seagreen4 module. Nodes are genes, and lines represent their connections. Twenty yellow nodes are the hub genes of the network.

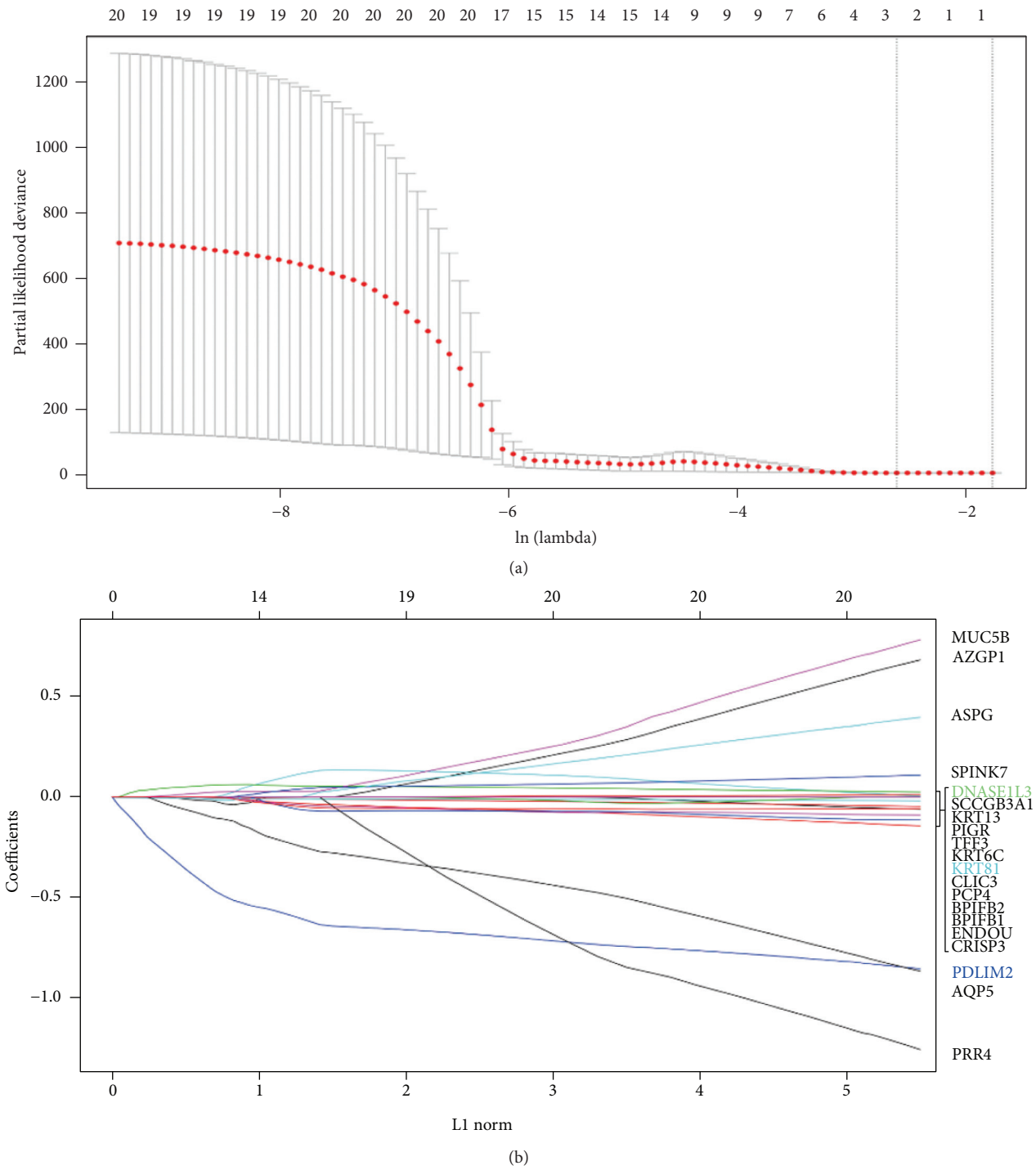


Figure 6: LASSO Cox regression model construction. (a) λ selection by 10-fold cross-validation. Continuous upright lines are partial likelihood deviance ± standard error (SE); imaginary lines are depicted at the optimal values by minimum criteria (lambda.min, left vertical dotted line) and 1−SE criteria (lambda.1se, right vertical dotted line). The partial likelihood deviance with changing of log (λ) is plotted. The value 0.073 is chosen for λ by 10-fold cross-validation with the minimum criteria. (b) Process of LASSO Cox model fitting. Each curve represents a gene. The trend of each coefficient against the L1-norm is plotted when λ changes. L1-norm is the total absolute of nonzero coefficients. LASSO: least absolute shrinkage and selection operator; L1-norm: L1 regularization, the total absolute of nonzero coefficients; SE: standard error.

3.6. Survival Analysis. ROC curve was plotted to assess the 1-year survival, and results shown in Figure 7 reveal that the cut-off risk score was decided as −0.136. As shown in the Kaplan–Meier curve in Figure 8(a), patients in the high-risk group had worse overall survival (OS) than those in the low-risk group. The mean OS was 26.3 months (95% CI,

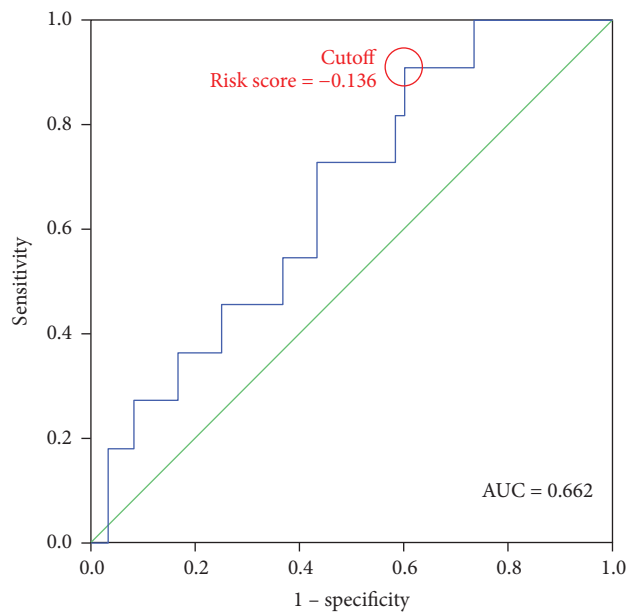


Figure 7: Cut-off risk score decision. The ROC curve is drawn, and AUC is calculated. The AUC is 0.662 (95% CI, 0.504–0.821) to predict 1-year survival using the training set. The cut-off risk score is decided as −0.136. AUC: area under the curve; ROC: receiver operating characteristic.

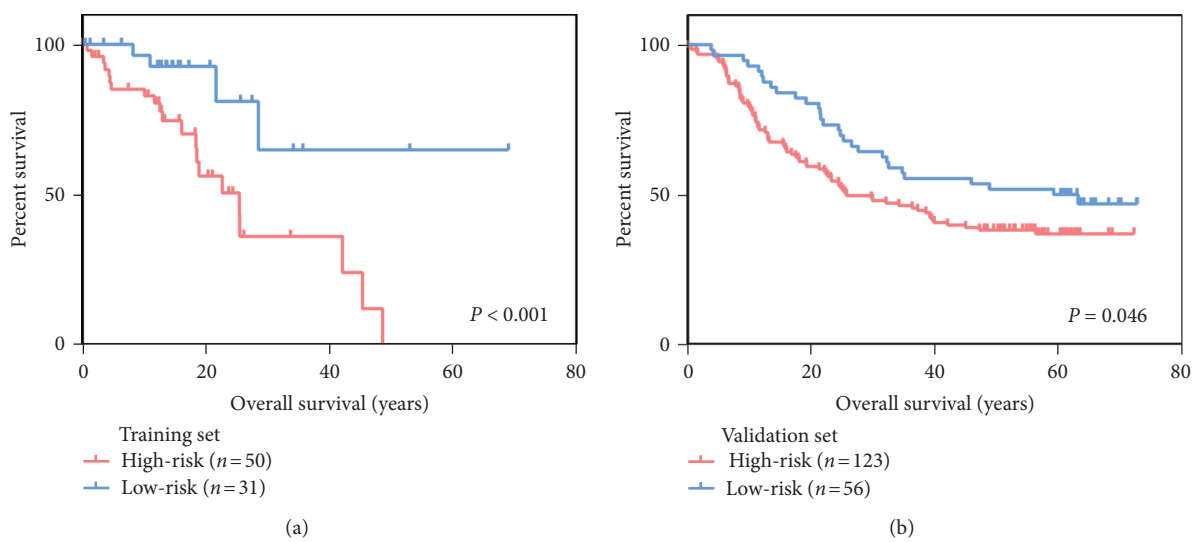


Figure 8: LASSO Cox regression model validation. (a) Survival comparison between the high- and low-risk groups of the training set using Kaplan–Meier analysis and log-rank tests. (b) Survival comparison between the high- and low-risk groups of the validation set using Kaplan–Meier analysis and log-rank tests. LASSO: least absolute shrinkage and selection operator.

20.1–32.5) in the high-risk group and 54.0 months (95% CI, 38.8–69.2) in the low-risk group ($P < 0.001$). Similarly, in the validation set, the mean OS was 38.0 months (95% CI, 33.1–43.0) in the high-risk group and 47.8 months (95% CI, 41.0–54.6) in the low-risk group ($P = 0.046$, Figure 8(b)).

3.7. Prognostic Factors Analysis. The risk score was then verified as an independent prognostic factor for OS, and univariate and multivariate analyses of potential prognostic factors in both training and validation set for OS were performed. In both sets, the risk score was correlated with OS in univariable analyses. After multivariable analyses adjustment by clinicopathological variables, the risk score and TNM stage remained prognostic factors for OS in the training set (Table 1) and the risk score remained the exclusive independent predictive factor for OS in the validation set (Table 2). ESCC patients with a high-risk score had poorer OS (training set: HR 5.319, 95% CI: 1.576–17.951, and $P = 0.007$; validation set: HR 1.767, 95% CI: 1.112–2.807, and $P = 0.016$). Clinical and pathological TNM staging plays an important role in predicting the prognosis of patients with esophageal cancer. Compared with the TNM stage, except for stage IV patients in the training set (HR 10.372,

Table 1: Univariate and multivariable Cox regression analysis of prognosis factors in the training set for OS.

Variable	Univariate analysis			Multivariable analysis		
	HR	95% CI	*P* value	HR	95% CI	*P* value
OS						
Risk score (high vs. low)	5.319**	1.576–17.951	0.007	6.815***	1.840–25.248	0.004
TNM stage						
(Stage II vs. I)	1.035	0.221–4.843	0.965	1.758	0.326–9.485	0.512
(Stage III vs. I)	1.927	0.404–9.188	0.410	3.254	0.581–18.216	0.179
(Stage IV vs. I)	3.929	0.626–24.640	0.144	10.372*	1.345–79.977	0.025
Age (≥60 vs. <60 years)	1.321	0.585–2.983	0.503			

Abbreviation. CI: confidence interval; HR: hazard ratio; OS: overall survival; TNM: tumor-node-metastasis; $^{*}P < 0.05$; $^{**}P < 0.01$; and $^{***}P < 0.01$.

Table 2: Univariate and multivariable Cox regression analysis of prognosis factors in the validation set for OS.

Variable	Univariate analysis			Multivariable analysis		
	HR	95% CI	*P* value	HR	95% CI	*P* value
OS						
Risk score (high vs. low)	1.543*	1.004–2.372	0.048	1.767*	1.112–2.807	0.016
TNM stage						
(Stage II vs. I)	1.260	0.500–3.174	0.625	1.362	0.540–3.437	0.513
(Stage III vs. I)	1.327	0.531–3.317	0.545	1.762	0.690–4.500	0.236
Age (≥60 vs. <60 years)	0.911	0.623–1.335	0.634			

Abbreviation. CI: confidence interval; HR: hazard ratio; OS: overall survival; TNM: tumor-node-metastasis; and $^{*}P < 0.05$.

95% CI: 1.345–79.977 $P = 0.025$), stage II patients and stage III patients did not have poorer OS than stage I patients. Unexpectedly, age had little effect on ESCC.

4. Discussion

With the development of microarray and RNA sequencing technology, a new era of biological big data is coming [12]. Data mining strategies can be used to explore key biological phenotypes related to high-dimensional data sets and can also realize the characterization of human cancer, the identification and definition of important genes in the tumorigenesis process, and the diagnosis of prognostic characteristics [13–15]. In this study, 10 gene coexpression modules were identified by WGCNA analysis of coexpression gene networks in TCGA and GSE53625 datasets. Of these modules, dark seagreen4 and thistle2 module are closely related to the survival time of ESCC, especially dark seagreen4. Therefore, it is speculated that the genes contained in module 4 are the key regulatory factors of ESCC, so we conducted further analysis. GO and KEGG enrichment analyses were used to evaluate the potential functional role of gene modules in ESCC [16]. It was observed that the genes in dark seagreen 4 were mainly involved in antimicrobial humoral response, epithelial cell differentiation, and neutrophil activation involved in immune response, all of which were involved in the cancer development. More importantly, based on the LASSO Cox regression model analysis of the TCGA data set, we screened out 3 genes that can predict overall survival, namely, PDLIM2, DNASE1L3, and KRT81 and constructed the best prognostic 3-gene risk prediction model.

PDZ and LIM domain containing protein 2 (PDLIM2), encoded by the *PDLIM2* gene, is a member of the actin-associated LIM family of proteins, which plays an important role in cytoskeletal organization and cell differentiation and is related to tumorigenesis [17]. Previous studies have shown that PDLIM2 plays a biological role as a tumor suppressor or a tumor promoter in different malignancies. For example, in human castration-resistant prostate cancer (CRPC)-like cells, PDLIM2 was significantly upregulated, while the inhibition of PDLIM2 can reduce the malignant phenotype in CRPC-like cells [18]. In colon cancer, PDLIM2 was significantly downregulated and involved in cancer regulation by inhibiting NF-κB activation [19]. These studies confirmed the tumor specificity of PDLIM2.

DNASE1L3 encodes an enzyme called deoxyribonuclease gamma. Monogenic lupus can be caused by nucleic acid degradation and repair defects because DNA that has not been properly cleared can serve as an immunogen library to drive the response of T cells and B cells [20]. Interestingly, the mutation in DNASE1L3 abolished the functional activity of the nuclease and caused defective DNA degradation, resulting in complete permeability of systemic lupus erythematosus [21]. DNase1L3, a genetically engineered human recombinant DNase, was described as attaining safeguarding stem cell-based regenerative therapy against iatrogenic cancerogenesis [22]. Malecki et al. found that the targeted expression of recombinant DNASE1, DNASE1L3, DNASE2, and DFFB can completely eradicate ovarian cancer cells in vitro, while healthy cells are not affected [23].

KRT81 encodes Hb-1, a type of keratin, which is expressed in all epithelial cell types. According to reports, it plays an important role in maintaining cell integrity, protein synthesis, and intracellular signal transduction [24,25]. Xie et al. confirmed that KRT81 rs3660GG type is an independent prognostic marker in non-Hodgkin's lymphoma [26]. Campayo et al. revealed that SNP-KRT81 plays an important role in the recurrence of nonsmall cell lung cancer

[27]. In addition, KRT81 also proved to be a novel and promising marker for squamous cell lung cancer [27].

Recent studies constructed prognostic models for ESCC based on gene expression data, with the aim of improving its early diagnosis and personalized treatment. Song et al. integrated the analysis of ESCC microarray data from GDS3838 and TCGA-ESCC and confirmed the prognostic value of PDLIM2 expression, which was independently associated with longer OS in ESCC patients [28]. In the current study, the risk score tends to be larger with lower expression of PDLIM2, and ESCC patients with a high-risk score had significantly poorer OS. As two independent studies, these similar findings suggested the accuracy and reliability of the analysis process. Alaei et al. constructed a large coexpression network of coding and noncoding genes using the WGCNA method and found important functional modules. Kaplan–Meier estimators and log-rank test statistics were then used to identify the majority of selected protein-coding and noncoding genes associated with poor prognosis in ESCC [29]. In this study, LASSO Cox regression model analysis was applied to construct a prognostic 3-gene risk prediction model. After the optimal cut-off point was determined by ROC analysis, the patients were divided into high-risk group and low-risk group. The results showed that the high-risk group had a lower percentage of survival, while the low-risk group had a higher percentage of survival. In addition, the verification result of risk score as an independent prognostic factor of OS showed that whether it is in the training set or the validation set, the risk score is its prognostic factor.

Zhan et al. developed a 3-gene signature by an ESCC-specific protein-protein interaction (PPI) network involving LOXL2 and actin-related proteins [30]. Sun et al. identified a three-gene prognostic signature based on the expression of GASC1 and 6 GASC1-targeted genes [31]. The signature was verified as an independent prognostic factor of ESCC. Guo et al. developed an immunogenomic risk score for predicting survival outcomes among esophageal cancer (EC) patients by utilizing immune-related genes [32]. Fei et al. constructed two prognostic risk score models of two EC sub-types, respectively [33]. The risk scores were verified as prognostic factors for EC OS and a classifier for the evaluation of groups with different risks. Since the underlying biological mechanisms of most ESCC biomarkers remains unclear, screening through candidate or pathway-based strategies rather than systematic screening results in limited prediction effect [34].

However, our prognostic model had some limitations. Firstly, we did not obtain other clinical information that might have influenced OS, such as primary health problems and follow-up treatment. Secondly, gene expression profiles obtained from GSE53625 were assessed by Agilent-038314 CBC *Homo sapiens* microarray V2.0 and have already been analyzed, so we did not start with the raw data. Thirdly, further validation in an independent set such as with RT-qPCR validation is required to confirm the diagnostic value of our model. Fourthly, we are unsure whether the risk score is feasible for metastatic ESCC because these samples were from primary tumors, the initial site of the cancer. The risk score was obtained from the dark seagreen4 module, which showed the strongest correlation with survival time. Further explorations are needed to detect markers from other modules.

5. Conclusion

In this study, a survival-related risk score for ESCC was identified using the WGCNA and a LASSO Cox regression model to explore a new molecular characterization of ESCC associated with prognosis, and we produced a risk model to aid its diagnosis and management. Our results suggested that three novel markers could effectively be of diagnostic and therapeutic value for the management of ESCC.

Data Availability

The data underlying this article are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Di Lu, He Wang, and Xuanzhen Wu contributed to this study equally.

Acknowledgments

The authors would like to thank the 27th European Conference on General Thoracic Surgery for the presentation of the abstract of this work. This work was supported by the Science and Technology Planning Project of Guangdong Province (No. 2017B020226005) and Start-Up Foundation for Scientific Research of Southern Medical University (Nos. PY2018N030 and PY2018N028).

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] J. R. Siewert and K. Ott, "Are squamous and adenocarcinomas of the esophagus the same disease?" *Seminars in Radiation Oncology*, vol. 17, no. 1, pp. 38–44, 2007.

[3] A. Pennathur, M. K. Gibson, B. A. Jobe, and J. D. Luketich, "Oesophageal carcinoma," *The Lancet*, vol. 381, no. 9864, pp. 400–412, 2013.

[4] J. Hu, R. Li, H. Miao, and Z. Wen, "Identification of key genes for esophageal squamous cell carcinoma via integrated bioinformatics analysis and experimental confirmation," *Journal of Thoracic Disease*, vol. 12, no. 6, pp. 3188–3199, 2020.

[5] H. Zeng, R. Zheng, S. Zhang et al., "Esophageal cancer statistics in C hina, 2011: estimates based on 177 cancer registries," *Thoracic Cancer*, vol. 7, no. 2, pp. 232–237, 2016.

[6] T. Can, "Introduction to bioinformatics," *miRNomics: MicroRNA Biology and Computational Analysis*, vol. 1107, pp. 51–71, 2014.

[7] J. N. Weinstein, E. A. Collisson, E. A. Collisson et al., "The cancer genome Atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[8] A. Shergalis, A. Bankhead 3rd, U. Luesakul, N. Muangsin, and N. Neamati, "Current challenges and opportunities in treating glioblastoma," *Pharmacological Reviews*, vol. 70, no. 3, pp. 412–445, 2018.

[9] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, "Comparing statistical methods for constructing large scale gene networks," *PloS One*, vol. 7, no. 1, Article ID e29348, 2012.

[10] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[11] H. Wang, D. Lu, X Liu et al., "Survival-related risk score of lung adenocarcinoma identified by weight gene co-expression network analysis," *Oncology Letters*, vol. 18, pp. 4441–4448, 2019.

[12] H. Zhao, S. Zhang, S. Shao, and H. Fang, "Identification of a prognostic 3-gene risk prediction model for thyroid cancer," *Frontiers in Endocrinology*, vol. 11, p. 510, 2020.

[13] A. Hébrant, G. Dom, M. Dewaele et al., "mRNA expression in papillary and anaplastic thyroid carcinoma: molecular anatomy of a killing switch," *PloS One*, vol. 7, no. 10, Article ID e37807, 2012.

[14] H. Zhang, X. Zhao, M. Wang, and W. Ji, "Key modules and hub genes identified by coexpression network analysis for revealing novel biomarkers for larynx squamous cell carcinoma," *Journal of Cellular Biochemistry*, vol. 120, no. 12, pp. 19832–19840, 2019.

[15] Y. Zou and L. Jing, "Identification of key modules and prognostic markers in adrenocortical carcinoma by weighted gene co-expression network analysis," *Oncology Letters*, vol. 18, pp. 3673–3681, 2019.

[16] N. Chen, G. Zhang, J. Fu, and Q. Wu, "Identification of key modules and hub genes involved in esophageal squamous cell carcinoma tumorigenesis," *Using WCGNA*, vol. 27, Article ID 1073274820978817, 2020.

[17] Z. Qu, J. Fu, P. Yan, J. Hu, S.-Y. Cheng, and G. Xiao, "Epigenetic repression of PDZ-LIM domain-containing protein 2," *Journal of Biological Chemistry*, vol. 285, no. 16, pp. 11786–11792, 2010.

[18] M. Kang, K.-H. Lee, H. S. Lee et al., "PDLIM2suppression efficiently reduces tumor growth and invasiveness of human castration-resistant prostate cancer-like cells," *The Prostate*, vol. 76, no. 3, pp. 273–285, 2016.

[19] B. Y. Oh, J. Cho, H. K. Hong et al., "Exome and transcriptome sequencing identifies loss of PDLIM2 in metastatic colorectal cancers," *Cancer Management and Research*, vol. 9, pp. 581–589, 2017.

[20] P. Costa-Reis and K. E. Sullivan, "Monogenic lupus: it's all new!," *Current Opinion in Immunology*, vol. 49, pp. 87–95, 2017.

[21] A. Bodaño, J. Amarelo, A. González, J. J. Gómez-Reino, and C. Conde, "NovelDNASE Imutations related to systemic lupus erythematosus," *Arthritis and Rheumatism*, vol. 50, no. 12, pp. 4070-4071, 2004.

[22] M. Malecki, C. LaVanne, D. Alhambra, C. Dodivenaka, S. Nagel, and R. Malecki, "Safeguarding stem cell-based regenerative therapy against iatrogenic cancerogenesis: transgenic expression of DNASE1, DNASE1L3, DNASE2, DFFB controlled by POLA1 promoter in proliferating and directed differentiation resisting human autologous pluripotent induced stem cells leads to their death," *Journal of Stem Cell Research and Therapy*, vol. Suppl 9, 2013.

[23] M. Malecki, J. Dahlke, M. Haig, L. Wohlwend, and R. Malecki, "Eradication of human ovarian cancer cells by transgenic expression of recombinant DNASE1, DNASE1L3, DNASE2, and DFFB controlled by EGFR promoter: novel strategy for targeted therapy of cancer," *Journal of Genetic Syndromes & Gene Therapy*, vol. 4, p. 152, 2013.

[24] P. A. Coulombe and M. B. Omary, ""Hard" and "soft" principles defining the structure, function and regulation of keratin intermediate filaments," *Current Opinion in Cell Biology*, vol. 14, no. 1, pp. 110–122, 2002.

[25] V. Karantza, "Keratins in health and cancer: more than mere epithelial cell markers," *Oncogene*, vol. 30, no. 2, pp. 127–138, 2011.

[26] Y. Xie, L. Diao, L. Zhang, C. Liu, Z. Xu, and S. Liu, "A miR-SNP of the KRT81 gene is associated with the prognosis of non-Hodgkin's lymphoma," *Gene*, vol. 539, no. 2, pp. 198–202, 2014.

[27] M. Campayo, A. Navarro, N. Viñolas et al., "A dual role for KRT81: a miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma," *PloS One*, vol. 6, no. 7, Article ID e22509, 2011.

[28] G. Song, J. Xu, L. He et al., "Systematic profiling identifies PDLIM2 as a novel prognostic predictor for oesophageal squamous cell carcinoma (ESCC)," *Journal of Cellular and Molecular Medicine*, vol. 23, no. 8, pp. 5751–5761, 2019.

[29] S. Alaei, B. Sadeghi, A. Najafi, and A. Masoudi-Nejad, "LncRNA and mRNA integration network reconstruction reveals novel key regulators in esophageal squamous-cell carcinoma," *Genomics*, vol. 111, no. 1, pp. 76–89, 2019.

[30] X.-H. Zhan, J.-W. Jiao, H.-F. Zhang et al., "A three-gene signature from protein-protein interaction network of LOXL2 - and actin-related proteins for esophageal squamous cell carcinoma prognosis," *Cancer Medicine*, vol. 6, no. 7, pp. 1707–1719, 2017.

[31] L. L. Sun, J. Y. Wu, Z. Y Wu et al., "A three-gene signature and clinical outcome in esophageal squamous cell carcinoma," *International Journal of Cancer*, vol. 136, pp. E569–E577, 2015.

[32] X. Guo, Y. Wang, H. Zhang et al., "Identification of the prognostic value of immune-related genes in esophageal cancer," *Frontiers in Genetics*, vol. 11, p. 989, 2020.

[33] Z. Fei, R. Xie, Z. Chen et al., "Establishment of a novel risk score system of immune genes associated with prognosis in esophageal carcinoma," *Frontiers in Oncology*, vol. 11, p. 625271, 2021.

[34] Y. Yu, Z. Li, C. Huang et al., "Integrated analysis of genomic and transcriptomic profiles identified a prognostic immunohistochemistry panel for esophageal squamous cell cancer," *Cancer Medicine*, vol. 9, no. 2, pp. 575–585, 2020.

Hindawi
Complexity
Volume 2021, Article ID 9475754, 13 pages
https://doi.org/10.1155/2021/9475754

Research Article

A Loss Reduction Optimization Method for Distribution Network Based on Combined Power Loss Reduction Strategy

Jihua Xie, Chang Chen, and Huan Long

School of Electrical Engineering, Southeast University, Nanjing 210098, China

Correspondence should be addressed to Huan Long; hlong@seu.edu.cn

Received 11 June 2021; Accepted 13 July 2021; Published 23 July 2021

Academic Editor: Long Wang

Power loss reflects the effective utilization rate of energy and the management level of power grids. In this paper, we propose a combined power loss reduction strategy optimization framework to improve the power loss reduction effect in a distribution network. The weak points of the distribution network are analyzed based on power flow calculation. The corresponding power loss reduction strategies are generated considering the following three aspects: replacing distribution lines, distribution transformers, and reactive power compensation. A combined power loss reduction strategy optimization model considering the comprehensive benefits of power loss reduction is established. A method for solving the optimization model based on the cost-benefit ratio is also proposed. Experiments based on the dataset from Tianjin show that the proposed loss reduction optimization method can effectively reduce power loss and formulate a reasonable loss reduction modification scheme in the distribution network.

1. Introduction

Power loss rate is an essential comprehensive index to measure the technical management and operation management levels of power supply enterprises. Since the power loss of the distribution network occupies a considerable proportion in the whole power system, the loss reduction modification of the distribution network has always been the critical work for power supply enterprises to improve their economic operation [1–4]. Thus, loss reduction optimization for the distribution network is a vital problem for power supply enterprises.

Loss reduction strategies of a distribution network can be mainly divided into management and technical strategies. Since the management strategies are primarily related to human factors, the primary task of power supply enterprises is to optimize the power loss management system and standardize the power loss management process [5–8]. Thus, the technical strategies of loss reduction are mainly taken into consideration in this paper.

The current research work on the loss reduction of the distribution network has been studied from many aspects. In [9], various loss reduction technical strategies of the distribution network were comprehensively summarized from two aspects of power equipment configuration and grid system operation. In [10], an evolutionary programming-based technique was proposed to optimize the placement of distributed generation units energized by wind and solar energy in a radial distribution system. In [11], based on considering the stochastic nature of distributed generation, a comprehensive optimization model for the simultaneous allocation of capacitor banks and distributed generation was proposed, and a hybrid algorithm based on Tabu search and genetic algorithms was also proposed to solve the model. In [12], on the basis of considering the uncertainty of distributed generation, electric vehicles, and other loads, Latin hypercube sampling was employed to generate random variables, and a bilayer optimization model was constructed. The improved harmony search algorithm was used to realize the dynamic reconfiguration of the distribution network. In [13], the multiobjective distribution network reconfiguration model considering distributed power generation and load uncertainty was proposed, which could optimize multiple important goals of the distribution network and effectively reduce the power loss of the distribution network. In [14], combining a microscopic analysis and the macro

statistics of the distribution network, an energy saving modification investment planning model, constrained by the investment and weighting factors, was developed to evaluate the energy saving. In [15], the Bat algorithm was used to solve the problem of reactive power source optimization for bus voltage deviation index minimization by the optimal placement of a number of capacitor banks in the network buses. In [16], optimal D-STATCOM placement and size was determined based on the index vector method for radial distribution networks under a reconfigured network to reduce the power loss. In [17], a multiobjective evolutionary algorithm based on a fuzzy decision-making method was proposed to reduce the power loss and improve the reliability of the radial distribution system.

Although there are currently a large number of references on loss reduction strategies, they mainly focus on the theoretical elaboration of different loss reduction strategies. Previous studies are mostly focused on calculating the power saving amount of various specific strategies such as reactive power compensation and power equipment replacement and analyzing the effect of energy saving and loss reduction of different strategies [18]. However, there is little research on the selection method of loss reduction modification scheme based on the combination of multiple loss reduction strategies, leading to the possibility that the loss reduction effect may not be the optimal situation.

In order to solve the abovementioned problems, a novel loss reduction optimization method for the distribution network is proposed in this paper based on the combined power loss reduction strategy that is divided into three stages: weak point analysis of power loss, generation of loss reduction strategy, and combined loss reduction strategy optimization. The weak point analysis of power loss of the distribution network is first carried out based on power flow calculation. The corresponding power loss reduction strategies are then generated considering three aspects: replacing distribution lines, distribution transformers, and reactive power compensation. A combined power loss reduction strategy optimization model considering the comprehensive benefits of power loss reduction is established. A method for solving the optimization model based on the cost-benefit ratio is proposed. The dataset from a power supply company in Tianjin is utilized to validate the proposed methodology.

2. Methodology

2.1. The Structure of the Proposed Algorithm. Although there are many loss reduction technical strategies in the current distribution network, there is little research on loss reduction optimization based on a combination of multiple types of loss reduction strategies. The current loss reduction strategies are relatively simple and lack pertinence. Thus, a framework of combined loss reduction strategy optimization in the distribution network is proposed in this paper, as shown in Figure 1, which is mainly divided into three stages: weak point analysis of power loss, generation of loss reduction strategy, and combined loss reduction strategy optimization.

In the stage of weak point analysis of power loss, considering that the load of distribution transformers is constantly changing during the operation of the distribution network, in order to make the results of loss reduction analysis more consistent with the actual situation, the clustering algorithm is employed to generate typical load curves for all distribution transformers to establish a typical loss reduction scenario. Based on the statistical analysis of the loss operation data through the power flow calculation, the weak points of the power loss of the distribution network feeder can be identified, including the severely aged branches, the distribution transformers with low power factor, and the branches with excessive power loss. In the stage of generation of loss reduction strategy, based on the results of the weak point analysis, the corresponding loss reduction strategies are generated for each loss reduction object (distribution transformer, distribution line, etc.), considering the three aspects of replacing distribution lines, distribution transformers, and reactive power compensation. The commonly used energy saving and loss reduction strategies are depicted in Figure 2. Finally, in the stage of combined loss reduction strategy optimization, to take into account the loss reduction effect and economy, a combined loss reduction strategy optimization model for the distribution network is established. A novel method for solving the above optimization model based on the cost-benefit ratio is proposed to optimize the solution process and formulate a reasonable combined strategy for loss reduction modification scheme for the distribution network.

2.2. Combined Loss Reduction Strategy Optimization Model. The loss reduction modification scheme of the distribution network is composed of different types of loss reduction strategies. Each type of loss reduction strategy has a variety of specific implementation situations for choice. When formulating a loss reduction modification scheme, it is necessary to consider the loss reduction effect of the distribution network feeder after the loss reduction modification and to analyze the economy of the loss reduction modification.

2.2.1. Objective Function. This paper mainly generates loss reduction strategies from distribution transformer, distribution line, and reactive power compensation of distribution network. Thus, the cost of power loss, the replacement cost of distribution lines, the replacement cost of distribution transformers, and the cost of reactive power compensation are needed to be considered. To optimize the comprehensive benefits of loss reduction, the objective function of the combined loss reduction strategy optimization model is established as shown in the following equation:

$$\min C = C_{\text{loss}}^{i} + C_{\text{vc}}^{i} + C_{l}^{i} + C_{t}^{i}, \tag{1}$$

where C represents the sum of costs involved in the loss reduction modification scheme with multiple strategies combination; C_{loss}^{i} represents the power loss cost of the i-th loss reduction modification scheme; C_{vc}^{i} denotes the reactive

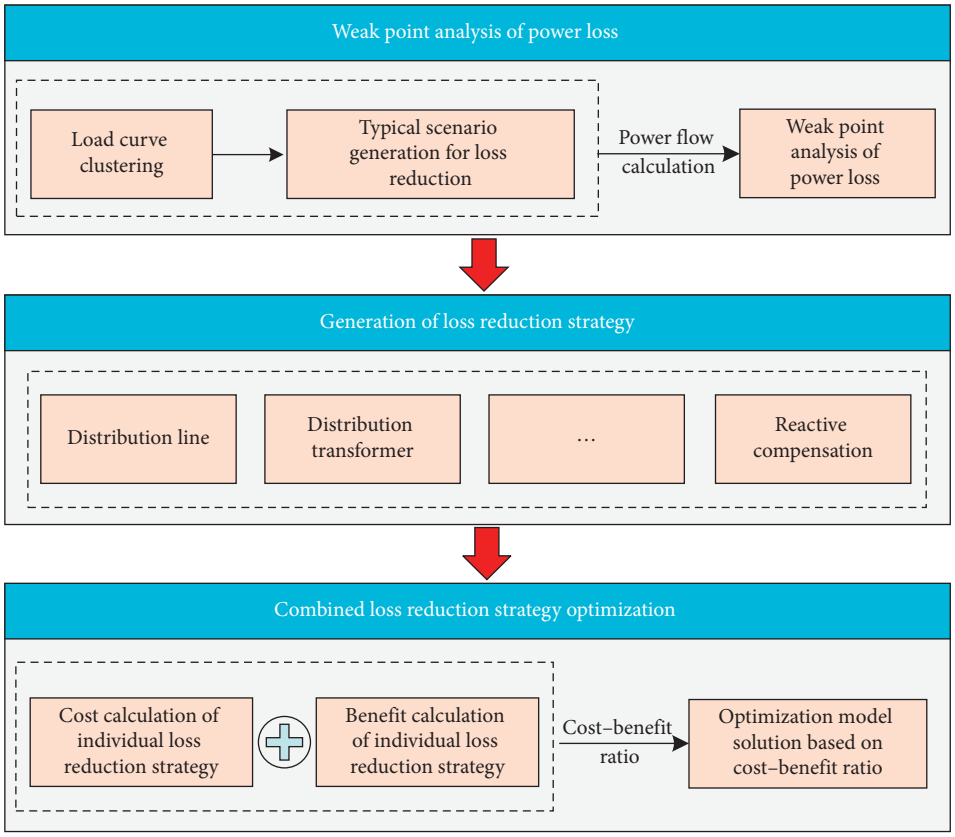


Figure 1: The proposed framework of combined loss reduction strategy optimization.

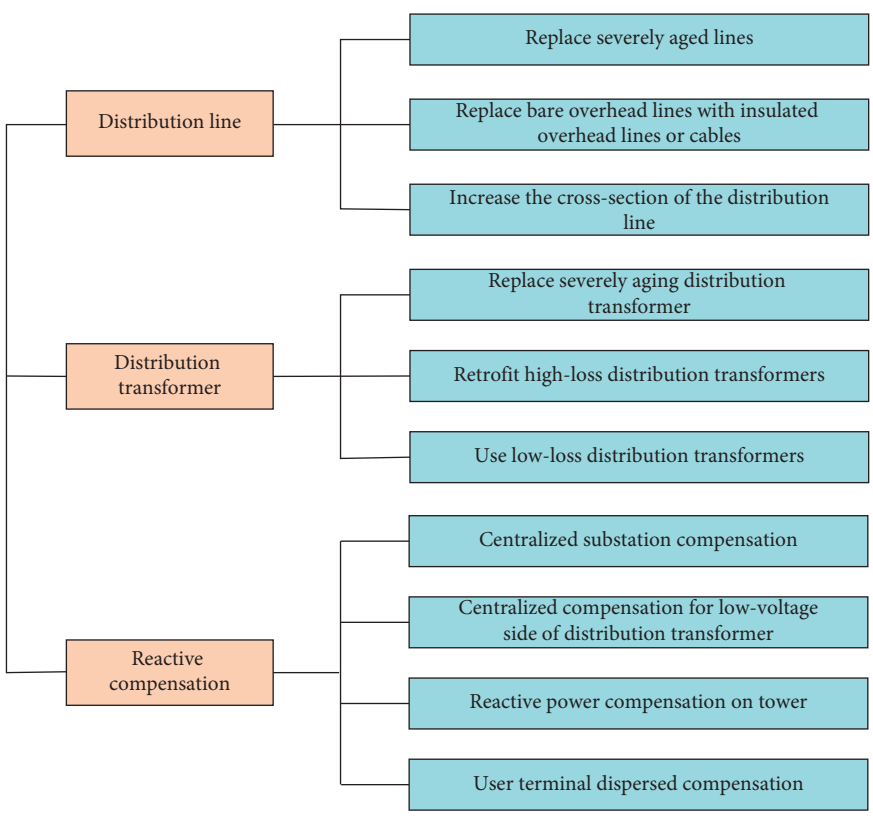


Figure 2: Commonly used energy saving and loss reduction strategies of the distribution network.

power compensation cost of the i-th loss reduction modification scheme; C_l^i stands for the replacement cost of the distribution line of the i-th loss reduction modification scheme; and C_t^i denotes the replacement cost of the distribution transformer of the i-th loss reduction modification scheme.

(1) Power Loss Cost. Power loss cost refers to the loss of electricity sale cost caused by power loss, which can be calculated by multiplying the power loss quantity with the corresponding electricity price, as shown in the following equation:

$$C_{\text{loss}}^i = \sum_{j=1}^{24} P_{\text{loss}}^{i,j} \cdot \tau_e \cdot n_{\text{LT}}, \tag{2}$$

where $P_{\text{loss}}^{i,j}$ represents the power loss at the j-th hour of the i-th loss reduction modification scheme; τ_e represents the electricity price; and n_{LT} represents the time frame considered for power loss cost, set as 365 in this paper.

(2) Reactive Power Compensation Cost. In this paper, the total compensated reactive power capacity of the distribution network is used to estimate its reactive power compensation cost, as shown in the following equation:

$$C_{\text{vc}}^i = \sum_{j=1}^{24} \sum_{k_{\text{vc}}=1}^{N_{\text{vc}}^i} Q_{k_{\text{vc}}}^{i,j} \cdot \tau_{\text{vc}} \cdot n_{\text{LT}}, \tag{3}$$

where N_{vc}^i represents the number of reactive power compensation points of the i-th loss reduction modification scheme; $Q_{k_{\text{vc}}}^{i,j}$ denotes the reactive power compensation capacity at the j-th hour at the k-th point of the i-th loss reduction modification scheme; and τ_{vc} represents the unit construction cost for reactive power compensation.

(3) Replacement Cost of the Distribution Line. The replacement cost of the distribution line is related to the length and type of the line, which is calculated using the following equation:

$$C_l^i = \sum_{k_l=1}^{N_l^i} L_{k_l}^i \cdot \tau_{l,k_l}^i, \tag{4}$$

where N_l^i represents the number of distribution lines that need to be replaced in the i-th loss reduction modification scheme; $L_{k_l}^i$ represents the length of the k_l-th distribution line to be replaced in the i-th loss reduction modification scheme; and $\tau_{l,kl}^i$ represents the unit construction cost of the k_l-th distribution line to be replaced in the i-th loss reduction modification scheme.

(4) Replacement Cost of the Distribution Transformer. The replacement cost of the distribution transformer is related to the type and capacity of the distribution transformer, which is described in the following equation:

$$C_t^i = \sum_{k_t=1}^{N_t^i} \tau_{t,k_t}^i, \tag{5}$$

where N_t^i represents the number of distribution transformers that need to be replaced in the i-th loss reduction modification scheme and $\tau_{t,kt}^i$ represents the unit construction cost of the k_t-th distribution transformer to be replaced in the i-th loss reduction modification scheme.

2.2.2. Constraints

(1) Power Loss Rate Constraint. Based on the development goals of the electric power development plan, power supply enterprises usually set the target value of the power loss rate after loss reduction modification, expressed in equation (6). For example, in the "13th Five-Year Plan for Electric Power Development (2016–2020)," the target value of the power loss rate is 6.5%.

$$P_{\text{loss}}\% = \frac{P_{\text{sup}} - P_{\text{sales}}}{P_{\text{sup}}} \times 100\% < \eta, \tag{6}$$

where $P_{\text{loss}}\%$ represents the power loss rate of the distribution network feeder after loss reduction modification; P_{sup} represents the power supply; P_{sales} represents the power sale quantity; and η represents the target value of the power loss rate after loss reduction modification of the distribution network feeder.

(2) Power Flow Constraint

$$\begin{cases} P_i = U_i \sum_{j \in i} U_j \left(G_{ij} \cos \delta_{ij} + B_{ij} \sin \delta_{ij} \right), \\ Q_i = U_i \sum_{j \in i} U_j \left(G_{ij} \sin \delta_{ij} - B_{ij} \cos \delta_{ij} \right), \end{cases} \tag{7}$$

where P_i represents the active power injected into the bus i; Q_i represents the reactive power injected to the bus i; U_i represents the voltage of bus i; δ_{ij} denotes the phasor between bus i and j; G_{ij} denotes the conductance between bus i and j; B_{ij} represents the susceptance between bus i and j; and $G_{ii} + jB_{ii}$ represents the self-admittance of bus i.

(3) Branch Transmission Capacity Constraint. The actual transmission capacity of the branch usually cannot exceed the maximum transmission capacity of the branch. In order to make the current operate within the normal range, the branch transmission capacity constraint is expressed in the following equation:

$$0 \le I_{ij} \le I_{\max}, \tag{8}$$

where $I_{\max}$ is the maximum allowable flow carrying capacity of the branch.

(4) Node Voltage Constraint. In order to make the node voltage operate within the normal range, the node voltage constraint is expressed as shown in the following equation:

$$U_{\min} \le U_i \le U_{\max}, \tag{9}$$

where $U_{\min}$ and $U_{\max}$ are the minimum and maximum values of the node voltage, respectively.

(5) Reactive Power Compensation Capacity Constraint. The constraint of reactive power compensation capacity is shown in the following equation:

$$Q_{i,\min} \le Q_i \le Q_{i,\max}, \tag{10}$$

where $Q_{i,\ \min}$ and $Q_{i,\ \max}$ represent the minimum and maximum values of reactive power compensation capacity at bus i, respectively.

2.2.3. Solution Method Based on Cost-Benefit Ratio. The purpose of solving the combined loss reduction strategy optimization model is to optimize the set of multiple loss reduction modification schemes composed of different loss reduction strategies for all loss reduction objects (distribution lines, distribution transformers, etc.), formulating a loss reduction modification scheme with the best comprehensive benefit of loss reduction, considering both the effect of loss reduction and the economy of loss reduction. The current research generally solves the combined loss reduction strategy optimization model through the enumeration method. For the alternative loss reduction modification schemes that meet the constraints, the objective function values are directly compared to determine the final loss reduction modification scheme with the best comprehensive benefit of loss reduction.

However, the number of alternative loss reduction modification schemes is closely related to the number of loss reduction objects determined by power loss weak point analysis results and the number of corresponding loss reduction strategies. Thus, there may be a huge number of alternative loss reduction modification schemes, which will inevitably lead to a large amount of calculation, resulting in low solution efficiency of the optimization model. A solution method based on the cost-benefit ratio is proposed in this paper to solve the above problem.

In this paper, the cost-benefit ratio, μ_{LR}, represents the ratio of the cost of loss reduction, C_{LR}, to the benefit of loss reduction, B_{LR}, as shown in equation (11). C_{LR} consists of the replacement cost of distribution lines, the replacement cost of distribution transformers, and the cost of reactive power compensation, described in equation (12). B_{LR} is the cost corresponding to the loss reduction electricity after the loss reduction modification in equation (13). It can be seen that when C_{LR} is lower and B_{LR} is higher, the corresponding μ_{LR} is smaller, which means that the corresponding loss reduction strategy should be selected.

$$\mu_{\mathrm{LR}} = \frac{C_{\mathrm{LR}}}{B_{\mathrm{LR}}}, \tag{11}$$

$$C_{\mathrm{LR}} = C_{\mathrm{vc}} + C_l + C_t, \tag{12}$$

$$B_{\mathrm{LR}} = C_{\mathrm{loss1}} - C_{\mathrm{loss2}}, \tag{13}$$

where C_{loss1} represents the power loss cost of the distribution network feeder before the loss reduction modification and C_{loss2} denotes the power loss cost after the loss reduction modification.

The solution process of the combined loss reduction strategy optimization model based on the cost-benefit ratio is shown in Figure 3. The specific steps are described as follows:

Step 1: based on the loss reduction strategies generated by the results of power loss weak point analysis, the loss reduction strategies that meet the constraints (power flow constraint, branch transmission capacity constraint, node voltage constraint, and reactive power compensation capacity constraint) are selected through power flow calculation. Then, the loss reduction cost, loss reduction benefit, and cost-benefit ratio when each loss reduction strategy is implemented separately are calculated.

Step 2: according to the order of cost-benefit ratio, the loss reduction strategies with the lowest cost-benefit ratio of each loss reduction object are selected and determined as the individual optimal loss reduction strategy corresponding to the loss reduction object.

Step 3: based on the individual optimal loss reduction strategies determined in Step 2, a set of alternative strategies for the loss reduction modification scheme are constructed according to the order of the cost-benefit ratio from low to high.

Step 4: the number of loss reduction strategies in the loss reduction modification scheme, r, is set to 1.

Step 5: the first r alternative loss reduction strategies are combined to construct a loss reduction modification scheme, and a loss reduction cost-benefit analysis is conducted based on the power flow calculation.

Step 6: if the termination condition is not met, then $r = r + 1$, continue to Step 5; if the termination condition is met, the current loss reduction modification scheme is determined as the final distribution network loss reduction modification scheme. The termination condition in this paper is that $P_{\mathrm{loss}}\% < \eta$.

It is worth noting that if the loss reduction modification scheme combines all the alternative strategies and cannot be less than the target power loss rate, it is necessary to regenerate loss reduction strategies with better loss reduction effects based on the power loss weak points.

3. Case Study

In this section, the data used in the experiment are first described. The loss reduction result of the selected feeder of the distribution network is displayed in the remaining sections.

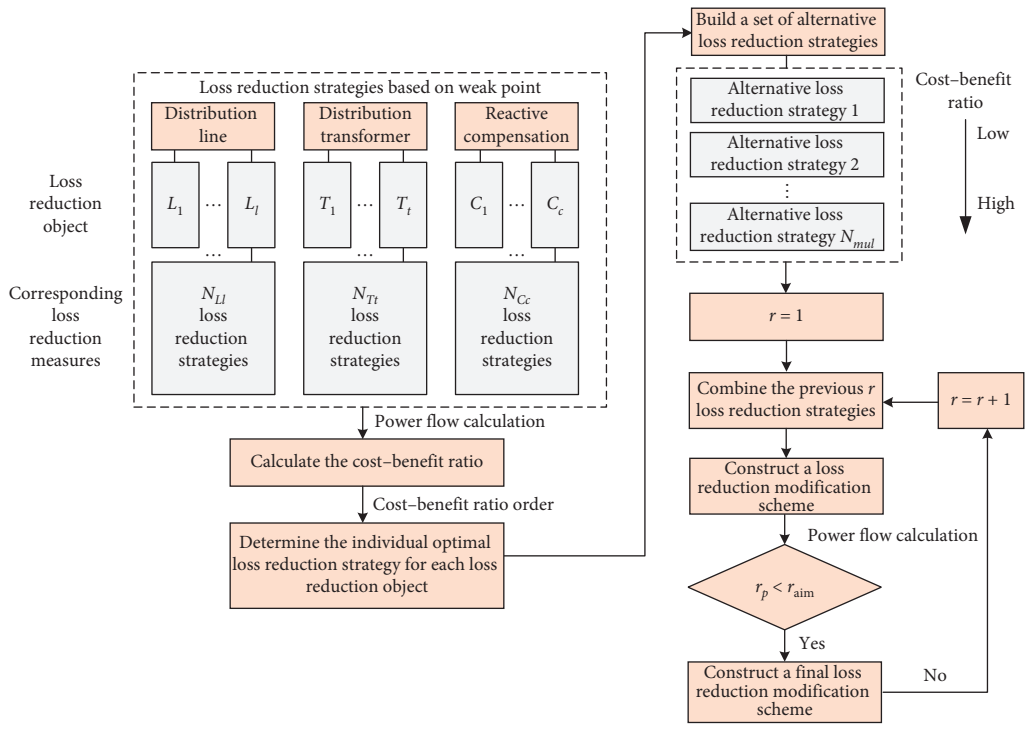


Figure 3: The solution process of the combined loss reduction strategy optimization model based on the cost-benefit ratio.

3.1. Dataset. The dataset utilized in this paper was collected from the Tianjin Electric Power Company in China. A 10 kV feeder of the distribution network is selected to conduct loss reduction, the topology of which is shown in Figure 4. The data, covering from January 1, 2019, to December 31, 2019, include the topology parameters, the parameters of power equipment, and load.

Compared with normal feeders, the feeders in the distribution network that need loss reduction generally have a higher power loss rate with a part of aged power equipment. Thus, in order to simulate the aging situation, the relevant parameters of the distribution lines, distribution transformers, and loads are modified to a certain extent. The specific modification is shown in Table 1. The parameter modification method of the aged transformer and the aged lines is to change their resistance parameters. In this paper, their resistance parameters are increased to 1.04~1.14 times of the original values [19].

3.2. Typical Scenario Generation for Loss Reduction. In order to generate a typical loss reduction scenario, the daily load curves sampled every 15 minutes of each distribution transformer in Figure 4 are clustered based on K-means [20,21]. Taking transformer $T1$ as an example, the load clustering results of transformer is shown in Figure 5. Figure 6 shows the center curves of the three clusters with the largest number of samples of part of distribution transformers. The cluster center of the cluster with the largest number of samples for each distribution transformer is taken as the typical load curve of each distribution transformer in a typical loss reduction scenario.

3.3. Generation of Loss Reduction Strategy. Based on the typical loss reduction scenario, the result of the power flow calculation is that the power loss rate of the feeder in Figure 4 is 6.4%, and its average power factor is 0.9 when no loss reduction strategies are selected in this typical scenario. The power factor of each public transformer is shown in Figure 7.

In order to generate targeted loss reduction strategies, according to the power flow calculation results and the power equipment parameters set above, the loss reduction strategies for each loss reduction object are proposed from the three aspects of distribution lines, transformers, and reactive power compensation. The details of the specific loss reduction strategies presented in this section are shown in Tables 2–4, the specific new type numbers of which are shown in Tables S1–S3 in Supplementary Material, respectively. The main ideas for generating loss reduction strategies in this paper are as follows:

(1) Distribution line: for severely aged lines or bare overhead conductors, replace them with new conductors or build new overhead insulated conductors/cables with the same cross-section according to actual conditions. The conductors can also be expanded accordingly considering the development requirements of the load;

(2) Distribution transformers: replace S7 series and other high-loss old transformers with S11, S13, or amorphous alloy-type distribution transformers;

(3) Reactive power compensation: select the distribution transformer with a power factor less than 0.85 to select reactive power compensation strategies,

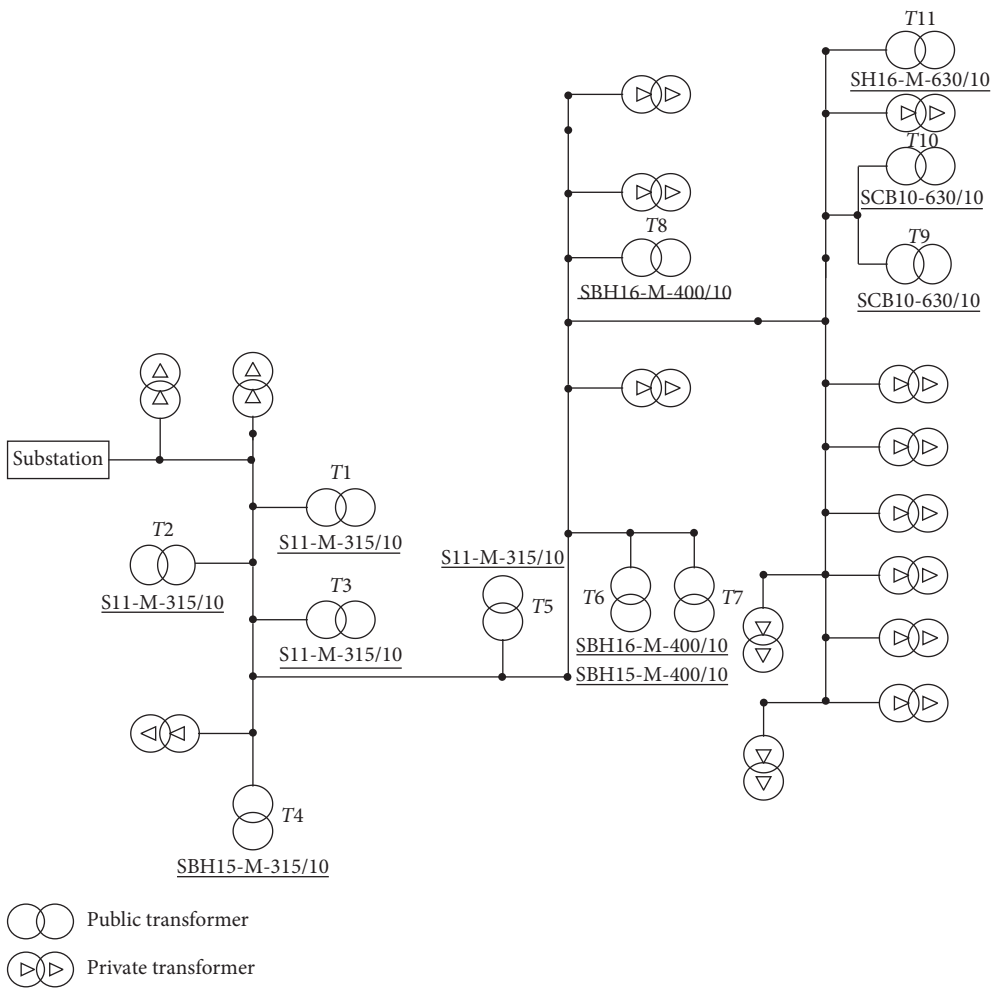


Figure 4: The selected feeder topology for loss reduction.

Table 1: Parameter modification details of transformers and lines.

Equipment	Head node of branch	Tail node of branch	Parameter modification details
Transformer	13	14	Change the equipment type to type 11
	16	17	
	18	19	Change the equipment type to type 11 and set a certain degree of aging
	27	28	
	51	52	
	51	53	Change the equipment type to type 7
	55	56	
Line	12	13	Set a certain degree of aging
	18	20	
	21	22	
	22	23	
	26	27	
	34	35	
	35	37	
	41	42	

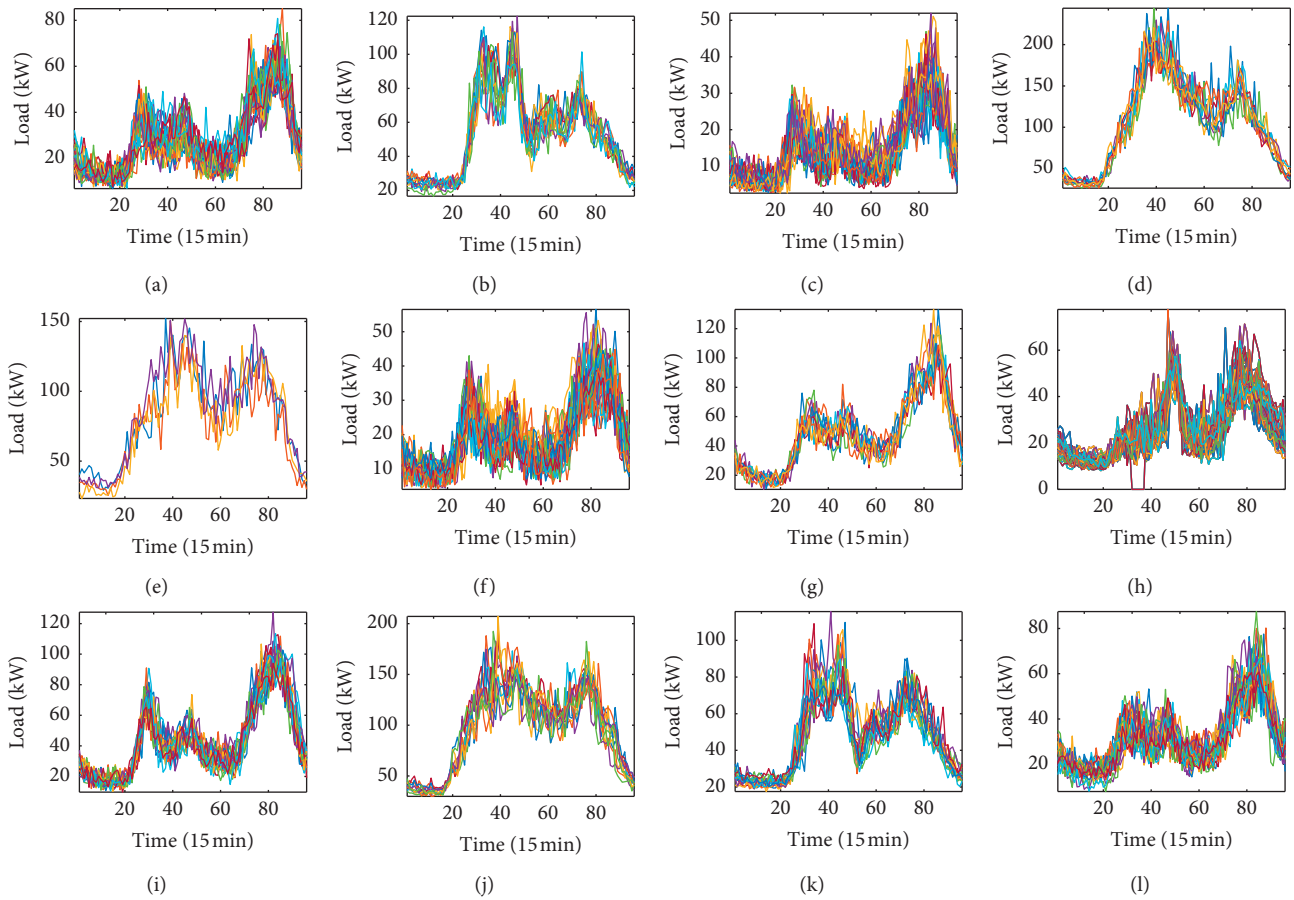


Figure 5: Load clustering results of transformer $T1$.

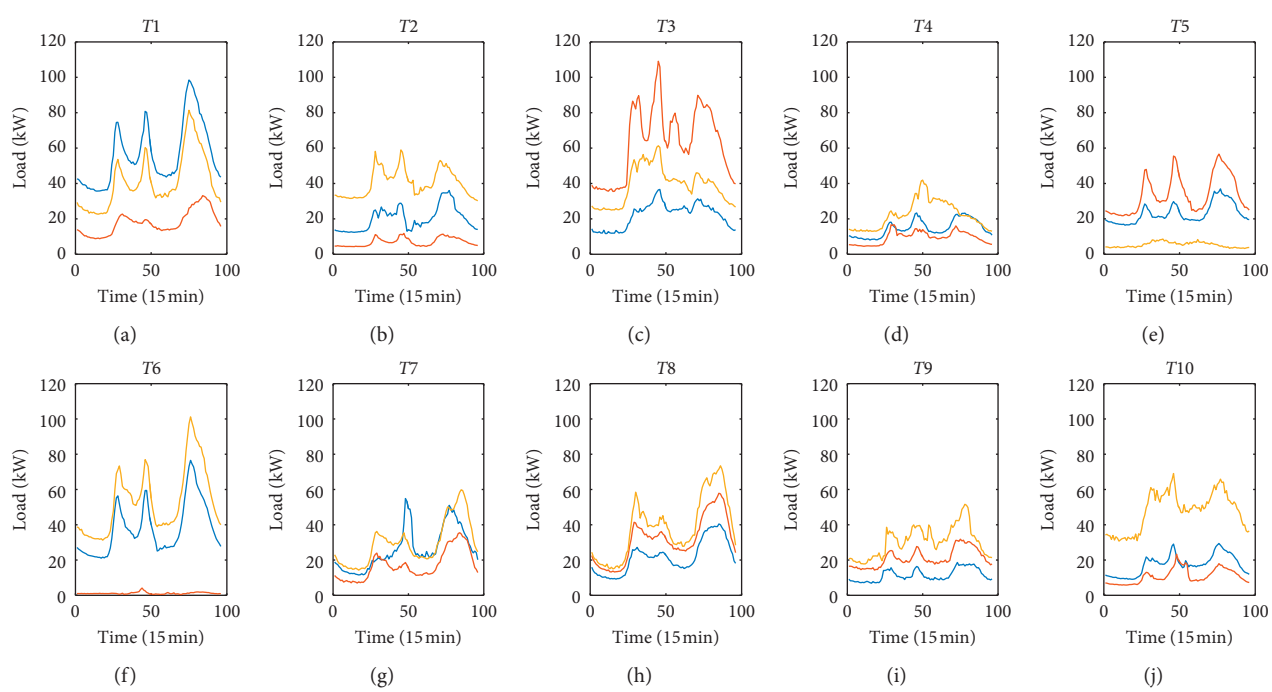


Figure 6: The center curves of the three clusters with the largest number of samples of part of distribution transformers.

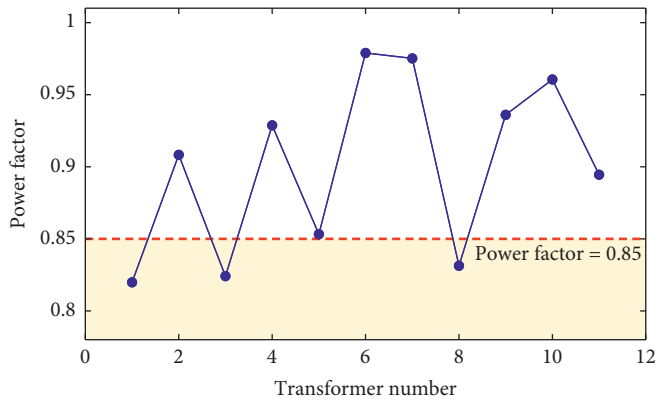


Figure 7: The power factor of each public transformer.

Table 2: Individual loss reduction strategies (distribution line).

Strategy number	Head node of branch	Tail node of branch	New type number	Strategy number	Head node of branch	Tail node of branch	New type number
1	12	13	9	18	34	35	11
2	12	13	10	19	34	35	17
3	12	13	17	20	35	37	10
4	12	13	6	21	35	37	11
5	18	20	10	22	35	37	17
6	18	20	11	23	41	42	10
7	18	20	17	24	41	42	11
8	21	22	10	25	41	42	17
9	21	22	11	26	30	31	10
10	21	22	17	27	30	31	11
11	22	23	10	28	30	31	17
12	22	23	11	29	31	32	10
13	22	23	17	30	31	32	11
14	26	27	10	31	31	32	17
15	26	27	11	32	42	43	10
16	26	27	17	33	42	43	11
17	34	35	10	34	42	43	17

Table 3: Individual loss reduction strategies (distribution transformer).

Strategy number	Head node of branch	Tail node of branch	New type number	Strategy number	Head node of branch	Tail node of branch	New type number
1	18	19	11	5	51	53	12
2	27	28	11	6	51	53	20
3	51	52	12	7	55	56	12
4	51	52	20	8	55	56	20

Table 4: Individual loss reduction strategies (reactive power compensation).

Strategy number	Load compensation node	Target power factor	Strategy number	Load compensation node	Target power factor
1	24	0.9	16	24	0.95
2	19	0.9	17	19	0.95
3	14	0.9	18	14	0.95
4	24	0.91	19	24	0.96
5	19	0.91	20	19	0.96
6	14	0.91	21	14	0.96
7	24	0.92	22	24	0.97
8	19	0.92	23	19	0.97

Table 4: Continued.

Strategy number	Load compensation node	Target power factor	Strategy number	Load compensation node	Target power factor
9	14	0.92	24	14	0.97
10	24	0.93	25	24	0.98
11	19	0.93	26	19	0.98
12	14	0.93	27	14	0.98
13	24	0.94	28	24	0.99
14	19	0.94	29	19	0.99
15	14	0.94	30	14	0.99

Table 5: Unit construction cost of related power equipment.

Equipment	Type	Unit construction cost	Unit
Line	YJV22-3 ∗ 50	11.5	$\times 10^4$¥/km
	YJV22-3 ∗ 70	40	$\times 10^4$¥/km
	YJV22-3 ∗ 120	60	$\times 10^4$¥/km
	JKLYJ-70	16	$\times 10^4$¥/km
	JkLYJ-120	24	$\times 10^4$¥/km
Transformer	S11-315	9	$\times 10^4$¥
	S11-630	13	$\times 10^4$¥
	S13-630	15	$\times 10^4$¥
Reactive compensation	—	80	¥/kVar

Table 6: Cost-benefit calculation results of individual loss reduction strategies (distribution transformers).

Strategy number	Loss reduction rate (%)	Loss reduction benefit ($\times 10^4$¥)	Loss reduction cost ($\times 10^4$¥)	Cost-benefit ratio
1	0.028	144.2885	9	0.062
2	0.006	144.2496	9	0.062
3	0.006	144.2499	13	0.090
4	0.006	144.2499	15	0.104
5	0.020	144.2742	13	0.090
6	0.020	144.2742	15	0.104
7	0.006	144.2498	13	0.090
8	0.006	144.2498	15	0.104

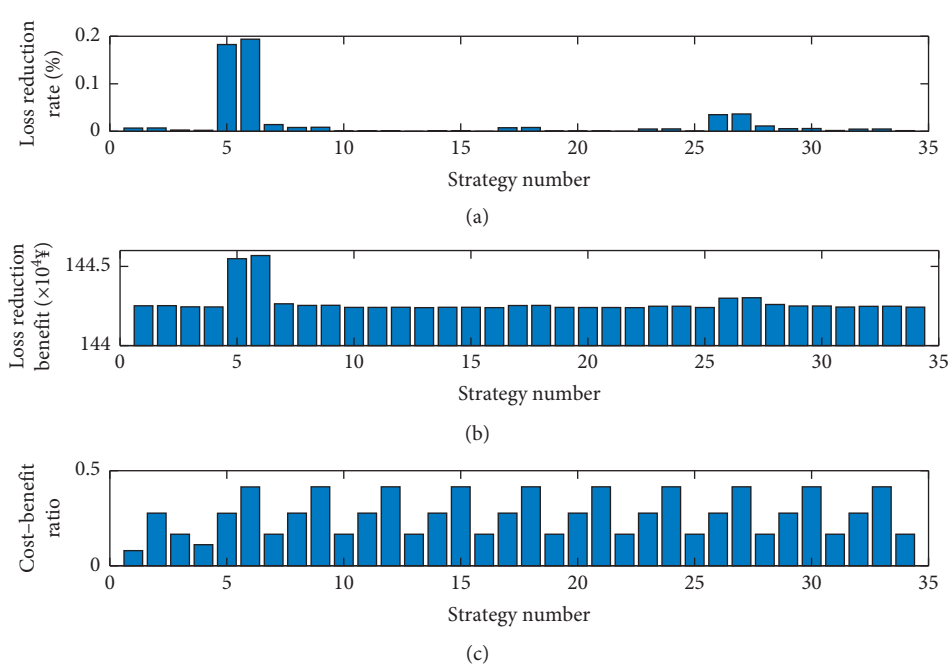


Figure 8: Cost-benefit calculation results of individual loss reduction strategies (distribution lines).

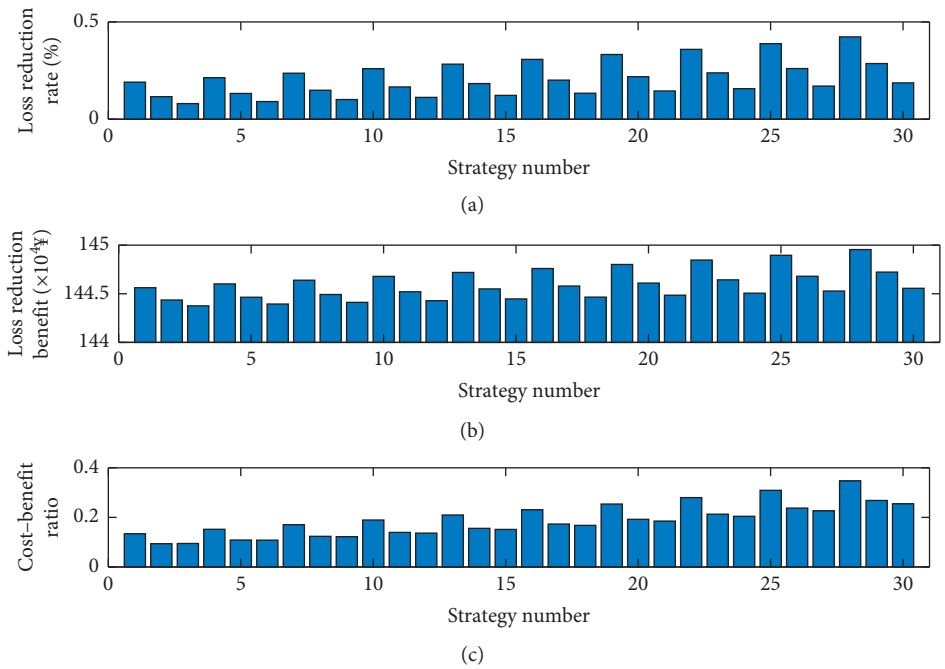


Figure 9: Cost-benefit calculation results of individual loss reduction strategies (reactive power compensation).

Table 7: Details of the individual optimal loss reduction strategies after sorting.

Number	Equipment	Corresponding strategy number	Number	Equipment	Corresponding strategy number
1	Transformer	1	11	Line	28
2	Transformer	2	12	Line	31
3	Line	1	13	Line	34
4	Transformer	5	14	Line	10
5	Transformer	3	15	Line	19
6	Transformer	7	16	Line	25
7	Reactive power compensation	2	17	Line	13
8	Reactive power compensation	3	18	Line	16
9	Reactive power compensation	1	19	Line	22
10	Line	7			

making its power factor reach the target value 0.9~0.99.

3.4. Cost-Benefit Analysis of Individual Loss Reduction Strategy. Table 5 lists the unit construction cost of power equipment related to the loss reduction strategy selected in this paper.

From Tables 2 to 5, in the typical loss reduction scenario, the loss reduction rate, loss reduction benefit, and the corresponding cost-benefit ratio of each individual loss reduction strategy compared to the case where no loss reduction strategies can be calculated. The calculation results are shown in Table 6 and Figures 8 and 9.

3.5. Combined Loss Reduction Strategy Optimization. Based on the cost-benefit calculation results of each individual loss reduction strategy, the cost-benefit ratio order can be obtained by sorting the cost-benefit ratio from low to high, and the individual optimal loss reduction strategies for each loss reduction object can be determined. Table 7 and Figure 10 show the details of the individual optimal loss reduction strategies after sorting and the corresponding loss reduction rate, loss reduction benefit, and cost-benefit ratio, respectively.

The cost-benefit ratio of replacing the transformer type or reactive power compensation is lower. Thus, their comprehensive loss reduction benefit after considering the investment cost and the extent of loss reduction is better. This is because the distribution feeder has a relatively high cable rate. Most of the bare overhead conductors to be renovated are basically not on the main trunk. Therefore, when an individual loss reduction strategy is selected, the corresponding loss reduction rate is relatively small.

According to the sequence of the individual optimal loss reduction strategies listed in Table 7, the alternative loss reduction strategy is combined in sequence starting from the first one. The power flow calculation is used to determine whether the target power loss rate value 6% is met. Figure 11 shows the result of the power loss rate after a combination of

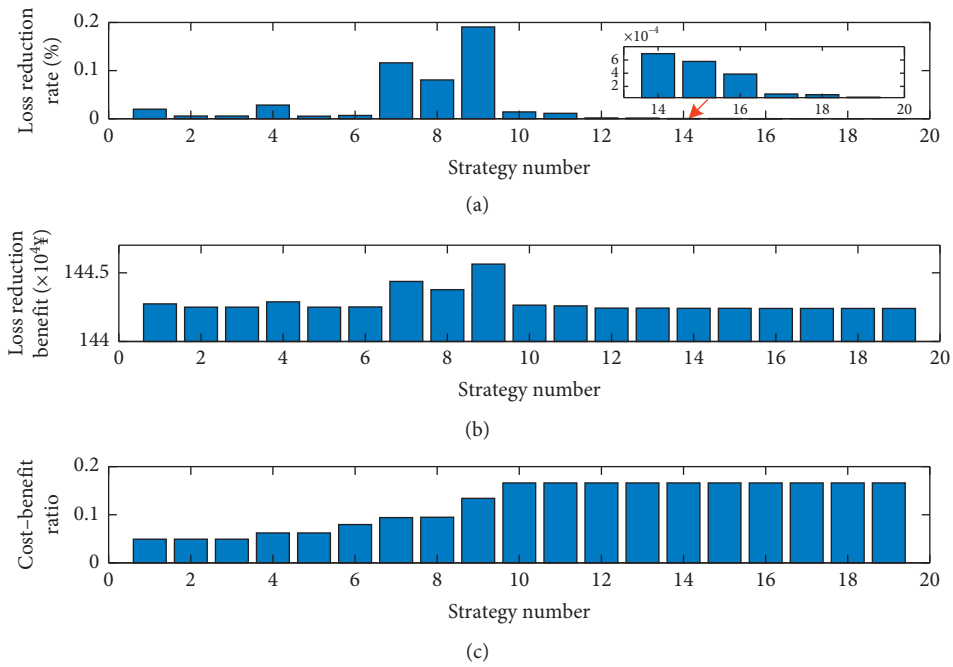


Figure 10: The cost-benefit calculation result of the individual optimal loss reduction strategy based on the cost-benefit ratio ranking.

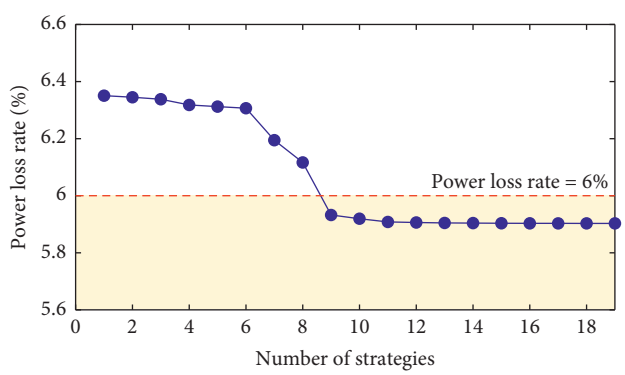


Figure 11: The result of the power loss rate after a combination of multiple strategies.

multiple strategies. It can be seen that when the top 9 kinds of loss reduction strategies are combined, the power loss rate of the feeder is already less than 6% at this time. After the loss reduction, the average power factor is 0.93, the loss reduction benefit is 1.4499 million yuan, and the loss reduction cost is 1.152 million yuan.

4. Conclusions

In this paper, a framework of combined power loss reduction strategy optimization is proposed to improve the power loss reduction effect in the distribution network, containing three stages: weak point analysis of power loss, generation of loss reduction strategy, and combined loss reduction strategy optimization.

Experiments were conducted using the dataset from the Tianjin Electric Power Company in China. Based on the power flow calculation, the analysis result of power loss weak points was obtained. To achieve the purpose of targeted loss reduction, the corresponding power loss reduction strategies were generated considering three aspects of replacing distribution lines, distribution transformers, and reactive power compensation. The corresponding power loss reduction strategies were generated considering three aspects: replacing distribution lines, distribution transformers, and reactive power compensation. A combined power loss reduction strategy optimization model considering the comprehensive benefits of power loss reduction was established. In order to solve the problem that the enumeration methods were generally used in most of the existing research to solve the above model, which caused a low efficiency of power loss reduction, a method for solving the optimization model based on the cost-benefit ratio was proposed. The result of the case study suggested that the proposed loss reduction optimization method could effectively formulate a reasonable loss modification scheme in the distribution network.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 51807023 and Natural Science Foundation of Jiangsu Province under Grant BK20180382.

Supplementary Materials

Table S1: line types. Table S2: transformer types. Table S3: branch information of the 10 kV feeder in the actual distribution network. (*Supplementary Materials*)

References

[1] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1023–1032, 2018.

[2] J. Jun Wang, Z. Jiangping Yu, and J. Yu, "Practical calculation method of theoretical line loss in 10 kV power distribution network based on Dscada," in *Proceedings of the 2008 China International Conference on Electricity Distribution*, Guangzhou, China, December 2008.

[3] C. M. P. Dos Santos, "Determination of electric power losses in distribution systems," in *Proceedings of the 2006 IEEE/PES Transmission & Distribution Conference and Exposition: Latin America*, Caracas, Venezuela, August 2006.

[4] S. Zhang, X. Dong, Y. Xing, and Y. Wang, "Analysis of influencing factors of transmission line loss based on GBDT algorithm," in *Proceedings of the 2019 International Conference on Communications, Information System and Computer Engineering (CISCE)*, Haikou, China, July 2019.

[5] J. F. Manirakiza and A. O. Ekwue, "Technical losses reduction strategies in a transmission network," in *Proceedings of the 2019 IEEE Africon*, Accra, Ghana, September 2019.

[6] M. Kundu, S. Jadhav, and K. Bagdia, "Technical loss reduction through active repair of distribution transformers: results from the field," in *Proceedings of the 2017 7th International Conference on Power Systems (ICPS)*, Pune, India, December 2017.

[7] D.-S. He, W. Lin, and Z.-Q. Liang, "The Energy efficiency diagnosis research of regional power grid loss reduction," in *Proceedings of the 2014 China International Conference on Electricity Distribution (CICED)*, Shenzhen, Chinadoi, September 2014.

[8] M. T. Au, T. M. Anthony, and M. Mohamad, "Strategies in technical loss reduction and it's impact on harmonic performance of distribution network," in *in Proceedings of the 2009 IEEE Bucharest PowerTech*, Bucharest, Romania, June 2009.

[9] L. Ying, M. Liu, L. Deng et al., "A comprehensive review of the loss reduction in distribution network," *Power System Protection and Control*, vol. 45, no. 19, pp. 162–169, 2017.

[10] D. K. Khatod, V. Pant, and J. Sharma, "Evolutionary programming based optimal placement of renewable distributed generators," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 683–695, 2013.

[11] B. R. Pereira, G. R. M. Martins da Costa, J. Contreras, and J. R. S. Mantovani, "Optimal distributed generation and reactive power allocation in electrical distribution systems," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 975–984, 2016.

[12] L. Xie, Z. Tang, X. Huang et al., "Bi-layer dynamic reconfiguration of a distribution network considering the uncertainty of distributed generation and electric vehicles," *Power System Protection and Control*, vol. 48, no. 10, pp. 1–11, 2020.

[13] X. Wang, Z. Wei, G. Sun et al., "Multi-objective distribution network reconfiguration considering uncertainties of distributed generation and load," *Electric Power Automation Equipment*, vol. 36, no. 6, pp. 116–121, 2016.

[14] Y. J. Zhang, X. T. Zhang, Q. H. Li, L. Ran, and Z. X. Cai, "Gray theory based energy saving potential evaluation and planning for distribution networks," *International Journal of Electrical Power & Energy Systems*, vol. 57, pp. 298–303, 2014.

[15] B. C. Neagu, O. Ivanov, and G. Georgescu, "Reactive power compensation in distribution networks using the bat algorithm," in *Prceedings of the 2016 International Conference and Exposition on Electrical and Power Engineering (EPE)*, Iasi, Romania, October 2016.

[16] A. R. Gupta and A. Kumar, "Energy saving using D-STATCOM placement in radial distribution system under reconfigured network," *Energy Procedia*, vol. 90, pp. 124–136, 2016.

[17] S. A. Nowdeh, I. F. Davoudkhani, M. J. H. Moghaddam et al., "Fuzzy multi-objective placement of renewable energy sources in distribution system with objective of loss reduction and reliability improvement using a novel hybrid method," *Applied Soft Computing*, vol. 77, pp. 761–779, 2019.

[18] W. Huang, J. Jiang, W. Chen et al., "Study on differentiated energy saving and loss reduction countermeasures for medium-voltage and low-voltage distribution network," *Power Capacitor & Reactive Power Compensation*, vol. 41, no. 5, pp. 0164–0170, 2020.

[19] L. Ding, "Research on the influence of aging and high resistance grounding fault on 10 kV line," *Jiangxi Electric Power*, vol. 44, no. 8, pp. 35–38, 2020.

[20] J. A. H. A. Wong, "Algorithm as 136: a K-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100–108, 1979.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR, USA, August 1996.

Hindawi
Complexity
Volume 2021, Article ID 8199013, 9 pages
https://doi.org/10.1155/2021/8199013

Research Article

A Defect Detection Method for the Surface of Metal Materials Based on an Adaptive Ultrasound Pulse Excitation Device and Infrared Thermal Imaging Technology

Yibo Ai,[1,2] Yingjie Zhang,[1] Xingzhao Cao,[1] and Weidong Zhang[1,2]

[1]*National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China*
[2]*Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai 519080, China*

Correspondence should be addressed to Weidong Zhang; zwd@ustb.edu.cn

Received 13 May 2021; Revised 15 June 2021; Accepted 26 June 2021; Published 10 July 2021

Academic Editor: Chao Huang

Ultrasonic excitation has been widely used in the detection of microcracks on metal surfaces, but there are problems such as poor excitation effect of ultrasonic pulse, long time to reach the best excitation, and difficult to find microcracks. In this paper, an adaptive ultrasonic pulse excitation device and infrared thermal imaging technology have been combined, as well as their control method, to solve the problem. The adaptive ultrasonic pulse excitation device adds intelligent modules to realize automatic adjustment of detection parameters, which can quickly obtain reliable excitation; the multidegree-of-freedom base realizes the three-dimensional direction change of the ultrasonic gun to adapt to different excitation occasions. When the appropriate ultrasonic excitation makes microcracks in the resonance state, the microcracks can be frictionated, which produce heat rise with the temperature. Then, the microcrack defect can be detected by the infrared thermal instrument through the different surface temperatures with imaging recognition method. Our detection experiments of the titanium alloy plates and the aluminum alloy profiles of marine engineering show that the method can get reliable detection parameters in a short time and measure the crack length effectively. It can be used in many aspects such as crack detection in mechanical structures or complex equipment operating conditions and industrial production processes.

1. Introduction

Metal components are widely used in marine engineering, aerospace, vehicle engineering, and many other fields, and cracks and other defects are unavoidable in the process of production and operation. Therefore, the detection of cracks in metal parts of large-scale equipment has also put forward higher and higher requirements. Surface cracks are a common form of failure in metal components. If the cracks cannot be detected in time, the cracks will gradually expand, which will lead to the failure of components, which will lay a serious hidden danger for the safe operation of the equipment [1]. However, in engineering practice, microcracks on the metal surfaces are difficult to be directly observed by human eyes. Therefore, nondestructive testing methods such as ultrasound, radiation, penetration, infrared, and eddy current are often used for regular testing [2–6]. Among them, ultrasonic testing is suitable for various types of workpieces, which can detect surface cracks, internal cracks, and quantify cracks [7–11]. It is currently the most important method for crack detection and quantification [12].

Infrared thermal wave nondestructive detection is a specialized technology that uses the principle of infrared radiation to inspect and measure the surface of equipment or materials and other objects. It has the advantages of fast, large observation area, intuitive, accurate, noncontact, and other conventional detection technologies. It is suitable for field applications, online in-service testing, and so on [13–15]. Ultrasonic infrared thermal wave nondestructive detection technology uses the characteristics of ultrasound to transfer energy in the form of waves and inject it into the sample. If there are defects such as cracks, debonding, and

tomography in the sample, friction, hysteresis, and thermoelastic effects will cause the injected energy to accumulate here and convert it into internal energy, which will increase the temperature. The changes in the temperature field will be captured by the infrared camera being inspected to identify defects. The commonly used ultrasonic infrared thermal imaging system consists of four parts: an ultrasonic excitation device, an infrared thermal imager, a computer microprocessor, and a measured object, as shown in Figure 1.

Scholars at home and abroad have conducted in-depth research on the use of ultrasonic infrared thermal wave technology to detect material surface cracks. Jia [16] built a nondestructive testing platform for rail tread cracks with infrared heat waves when inspecting rail foot cracks. According to the distribution of rail surface temperature information, the location of internal defects in the rail was determined. Jing et al., [17] used ultrasonic thermal wave imaging technology to detect cracks in locomotive knuckle parts. Experiments show that this technology is not sensitive to surface shape, rust, dust, and pollution and has special applications in the detection of defects such as cracks. Tang Schwarz et al. [18] simulated the heating phenomenon of austenitic stainless-steel weld cracks under ultrasonic excitation through the finite element analysis software Ansys and qualitatively analyzed the impact of ultrasonic excitation amplitude, frequency, direction, and weld reinforcement on the cracks. The authors in [19] proposed a new type of nondestructive testing method for composite structures of aluminum-carbon fiber composites. The interface between the three materials was characterized by ultrasonic and thermal imaging generated by an electromagnetic ultrasonic transducer. By detecting the artificial integration defects in the inner layer of carbon fiber composite materials, the quality of the interface bonding between the carbon fiber cloth and the thermoplastic layer is characterized.

The above analysis shows that infrared thermal wave as a new type of nondestructive testing technology has been widely favored by scholars, but there are still some problems to be solved when it is applied to the detection of metal cracks in engineering. The effect of ultrasonic infrared detection of cracks mainly depends on the three aspects of ultrasonic excitation energy and heat generation mechanism of crack action, parameters in the detection process, and signal processing of infrared heat maps. There is no scientific conclusion about the heat generation mechanism in the current research, and the mainstream friction heat generation mechanism still cannot perfectly explain all heat generated in the crack area [20–22]. The detection parameters include the parameters of the workpiece and the detection conditions of the ultrasonic infrared system. A large number of studies and experiments have shown that the detection parameters in the ultrasonic infrared nondestructive testing process have a significant impact on the detection effect, such as excitation frequency [23], excitation force [23], excitation time [24], and the difference between the horn and the tested part changes in the pretightening force [25] which will affect the transmission of ultrasound, affect the effect of ultrasound excitation, and cause large fluctuations in the test results. The signal processing of the infrared heat map is to process the infrared heat map obtained from the experiment to obtain intuitive and accurate crack information. In addition, the requirements for detection speed in metal crack defect detection and industrial production are increasingly high [26–29].

To improve the accuracy of crack defect detection, this paper starts researching from two aspects: optimal detection parameter selection and infrared heat map processing. An adaptive ultrasonic excitation device is designed to emit ultrasonic waves to the metal profiles so that its internal defects absorb and couple ultrasonic energy and produce local temperature imbalance areas. And the infrared imager is used to collect thermal images and crack information. In this way, it can diagnose the cracks in the production or use of metal profiles like titanium and aluminum of marine engineering equipment and solve the hidden safety problems caused by surface microcracks.

2. Design of Ultrasonic Infrared Detection System

The ultrasonic infrared detection system used in our research is shown in Figure 2, which mainly contains three major parts, which are an ultrasonic excitation system, an infrared thermal image detection system, and a data processing system.

2.1. Ultrasonic Excitation System. The ultrasonic excitation system is the adaptive ultrasonic excitation device proposed in this paper. At present, there are two main types of ultrasonic excitation devices such as one is the vertical ultrasound excitation device and the other is the handheld ultrasound excitation device. The vertical ultrasonic excitation device has a large structure, high power, stable and reliable excitation effect, and high efficiency and is easy to operate automatically, but it costs a lot. The handheld ultrasonic excitation device is small in structure, easy to move, can adapt to the ultrasonic excitation of workpieces in multiple locations, and has low testing cost. However, it has small power, poor excitation effect, low detection efficiency, and limited application. Therefore, it is of great practical value to research and design an ultrasonic excitation device with small structure, easy operation, and good excitation effect to achieve the best excitation effect. The adaptive ultrasonic excitation device proposed in this study can overcome the abovementioned problems, and the multidegree-of-freedom base of the ultrasonic gun enables the device to adapt to different operating environments and meet the needs of actual use.

This paper designs an adaptive ultrasonic pulse excitation device, and its structure and toolchain are shown in Figure 3. The device includes an ultrasonic controller and a multidegree-of-freedom base of the ultrasonic gun. The ultrasonic controller includes an ultrasonic generator and an ultrasonic transducer. Among them, the ultrasonic generator adopts the stepless frequency modulation ultrasonic generator, the frequency range is 13 kHz~75 kHz, which can be modulated by 0.1 kHz, and the ultrasonic transducer

frequency is 20 kHz. The ultrasonic generator is connected to the ultrasonic transducer through a line, and the working parameters of the ultrasonic generator are adjusted through adaptive control to quickly obtain better ultrasonic excitation effects. The multidegree-of-freedom base of the ultrasonic gun can adapt to different operating environments so that the device can meet the actual needs. The device adopts the principle of information feedback, which can ensure that reliable experimental parameters (excitation frequency, excitation power, excitation time, compression force between the horn and the tested time, incentive position, and other parameters) are obtained in a short time.

The ultrasonic controller of the adaptive pulse excitation device includes the control panel and the internal functional modules of the ultrasonic controller. The ultrasonic controller control panel includes the power switch, the auto/manual mode switch button, and the ultrasonic controller. The control panel includes pulse excitation frequency adjustment knob and its display screen, ultrasonic pulse excitation power adjustment knob and its display screen, ultrasonic pulse excitation time adjustment knob and its display screen, and the compression force adjustment knob between the ultrasonic gun horn and the test piece and its display screen; the internal functional module of the ultrasonic controller includes four major modules, which are parameter manual control module, parameter automatic control module, parameter feedback module, and data access module. The ultrasonic controller powers the circuit board through the power switch, selects the auto/manual mode, adjusts the experimental parameters according to the corresponding mode, and starts the work of each module; among them, each parameter adjustment knob and display screen is connected to the parameter manual control module.

The multidegree-of-freedom base of the ultrasonic gun of the adaptive pulse excitation device includes moving wheels, chassis, plane rotating frame and its driving mechanism, height rising frame and its driving mechanism, angle tilt frame and its auxiliary mechanism; ultrasonic gun card tool and base control panel are mainly used to fix the direction and position of the ultrasonic gun. Under the multidegree-of-freedom base chassis of the ultrasonic gun, there are moving wheels. The front surface of the chassis has a base control panel. The base control panel is connected to three driving mechanisms: plane rotation, height rise, and angle elevation. The plane rotating frame and its driving mechanism are placed on the chassis. The rotating frame and its driving mechanism are equipped with a height raising frame and its driving mechanism. The upper part of the height raising frame and its driving mechanism is the angle pitch frame and its auxiliary mechanism and ultrasonic gun fixture; the ultrasonic gun fixture includes a servo motor, lead screw, and threaded clamping cylinder. The servo motor can drive the lead screw to expand and contract the threaded clamping cylinder that is screwed to the screw. A force sensor is placed in front of the ultrasonic gun to transmit the pressure measurement value to the ultrasonic controller. The force sensor adopts the S-type load cell MIK-LCS1, with a range of 0–200 kg and an accuracy of 0.03% FS.

2.2. Infrared Thermal Image Detection System. The infrared thermal image detection system mainly includes an infrared thermal imager, its auxiliary equipment, and data processing system.

The infrared thermal imager uses FLUKE Ti10 handheld thermal imager with a resolution of 640×480, 1.3 million pixels, and a thermal sensitivity (NETD) of 200 mK at 30°C. After ultrasonic excitation, the surface temperature of the surface crack will be higher than that of other locations because of the friction. The temperature change at the crack defect is greater, which leads to the uneven distribution of the temperature field on the surface of the alloy profile. The infrared imager is used for detection, and the temperature of different areas can be displayed intuitively.

The data processing system analyzes, processes, and evaluates the collected image data. Users can interact with other systems through the data processing system and can also perform image processing, data management, and other functions through related software operations. The software includes functions such as image reading and accessing, heat image grayscale processing, crack contour extraction, and crack length measurement. The software interface is shown in Figure 4. The image processing mainly includes operations such as image background difference, image segmentation, feature extraction, defect recognition, and defect reconstruction on the collected infrared heat map. The processing flow is shown in Figure 5.

3. Control Method of Adaptive Ultrasonic Pulse Excitation Device

This paper proposes an adaptive ultrasonic pulse excitation device control method. For the internal function modules of ultrasonic control, including parameter manual control module, parameter automatic control module, parameter feedback module, and data storage module, there are two control methods: manual control mode and automatic control mode.

3.1. Manual Control Mode of Adaptive Ultrasonic Pulse Excitation Device. Manual control mode includes the ultrasonic controller parameter manual control module, parameter feedback module, and data access module. The frequency adjustment knob, power adjustment knob, time adjustment knob, and compression force adjustment knob between the horn and the test piece on the ultrasonic controller panel are manually adjusted, and the module reads the set parameters and then passes through the module; each submodule output (including frequency control submodule, power control submodule, time control submodule, and compression force submodule between the horn and the tested sample) would be transferred as high-frequency alternating current and signals that meet the set parameters, and it is transmitted to the ultrasonic gun and its multidegree-of-freedom base so that the ultrasonic gun can emit ultrasonic waves with set frequency, power, and excitation time. At the same time, the parameter feedback module (as shown in Figure 6) also reads the set parameters

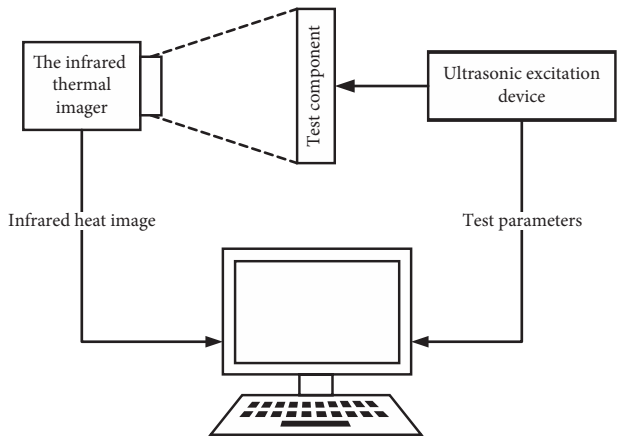


Figure 1: Composition of the ultrasound excitation infrared thermal imaging system.

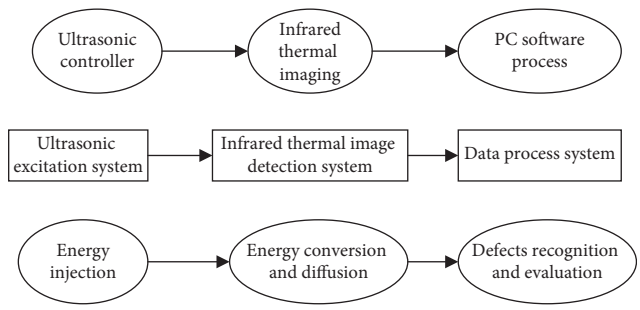


Figure 2: Relationship diagram of ultrasonic and infrared nondestructive testing equipment.

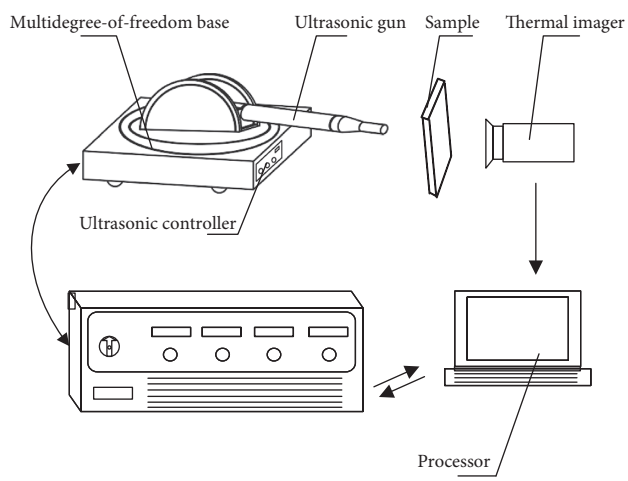


Figure 3: Schematic diagram of the adaptive ultrasonic pulse excitation device and its working chain.

and receives the processed excitation data obtained by the infrared thermal imager from the microprocessor. After comprehensive analysis and evaluation of the data, the data are stored in the data access module.

The data access module in Figure 7 mainly consists of two parts: the corresponding experimental conditions, including the relevant parameters of ultrasonic excitation, and the data information for subsequent infrared image processing.

3.2. Automatic Control Mode of Adaptive Ultrasonic Pulse Excitation Device. The automatic control mode includes parameter automatic control module, parameter manual control module, parameter feedback module, and data storage module. The flow chart of automatic control mode is shown in Figure 8.

The working steps of automatic control mode are as follows:

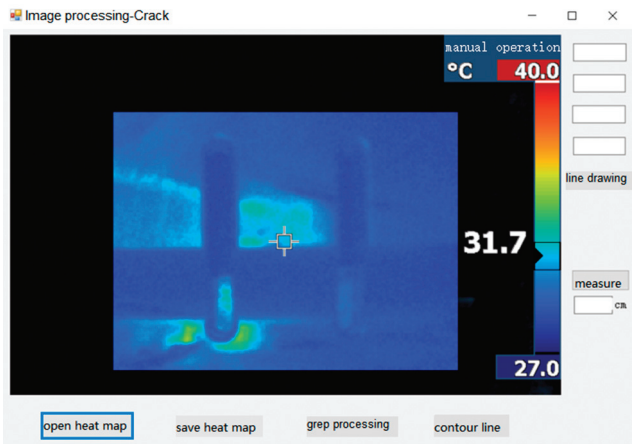


Figure 4: Data processing software interface.

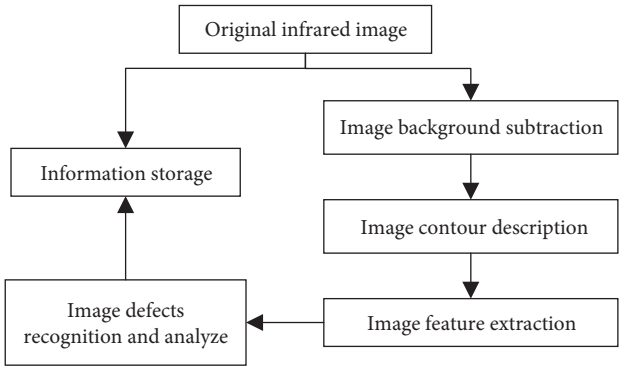


Figure 5: Thermal image data processing flow chart.

(1) The automatic control mode is selected under the ultrasonic controller panel, and the parameters feedback module reads a set of existing optimal working parameters from the data access module and transmits them as the initial working parameters to the parameters automatic module.

(2) After the data conversion of the parameter automatic control module, the signal is transferred to the circuit conversion module of the parameter manual control module, adjusted by the submodules in the parameter manual control module (including frequency control submodule, power control submodule, time control submodule, and compression force submodule between the horn and the tested sample). The output high-frequency alternating current and signals that meet the set parameters will be transmitted to the ultrasonic gun and its multidegree-of-freedom base.

(3) The parameter feedback module receives the excitation data and judges whether the range of the best excitation effect has been reached after analysis and evaluation. If the range of the best excitation effect is reached, further excitation is stopped and the data are written into the data storage. If the range of the best excitation effect is not reached, the modified experimental parameters will be given and passed to the parameter automatic control module.

(4) Repeat steps (1) to (3) until the best excitation effect is achieved; when the requirements are met, the parameters are stored in the data access module and stop working.

Combining manual control mode and automatic control mode to realize ultrasonic infrared nondestructive testing is an important innovation of this article. It can effectively solve the shortcomings of the traditional ultrasonic pulse excitation device with poor excitation effect and long time to achieve the best excitation effect. At the same time, there are many newly designed ultrasonic guns. The degree-of-freedom base can meet the needs of use in complex environments.

4. Experiments and Results Discussion

To verify the effectiveness of the adaptive ultrasonic excitation device and its control method proposed in this paper, a crack detection experiment on the alloy was carried out. The proposed method is a nondestructive crack detection method for metal surfaces, and there is no strictly special size of the sample and crack size. The experiment uses the ultrasonic infrared nondestructive testing system in Figure 2.

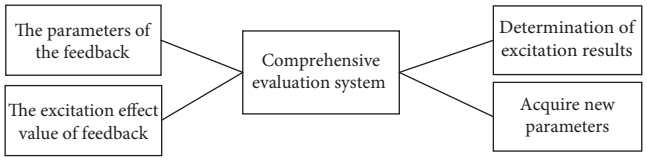


Figure 6: Schematic diagram of the working principle of parameter feedback module.

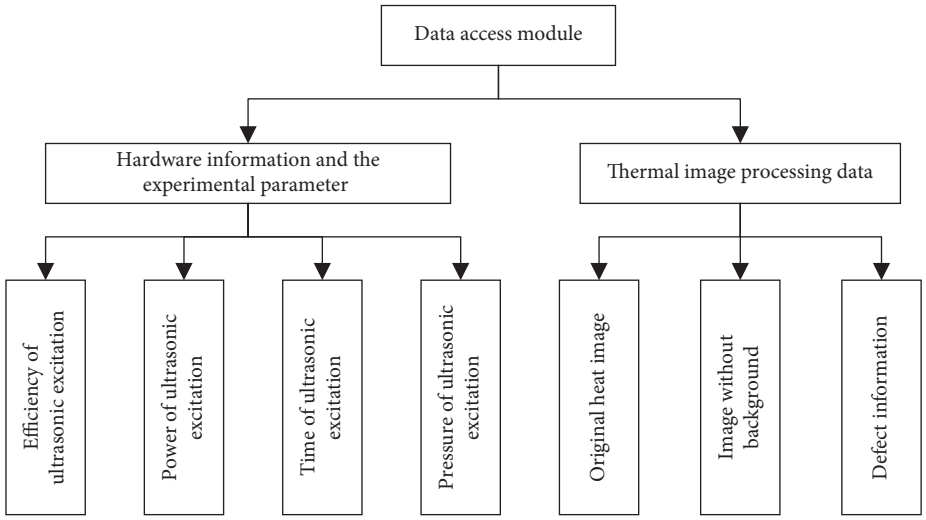


Figure 7: Schematic diagram of data access module.

The experiment is to detect the crack of U-groove specimen after the fatigue test. A sample is shown in Figure 9. The excitation gun is contacting the sample on one edge of the sample because the edge is flat and is easy to tough the gun. And for the placement, it is easy to detect the cracks by the infrared thermal imager at a positive perspective.

The ultrasonic excitation system is the active energy source for the crack detection experiment of metal material profile. The multidegree-of-freedom base of the ultrasonic gun is moved, and the excitation parameters are tested accordingly. The parameter feedback module reads a set of existing optimal working parameters from the data access module as the initial working parameters and passes them to the parameter automatic control module. Then, the data of the parameter automatic control module will be conversed and be transferred to the circuit conversion module of the parameter manual control module. After the control of each submodule in the parameter manual control module, it outputs high-frequency alternating current and signals that meet the set parameters and transmits them to the ultrasonic gun and multidegree freedom base. Then, the parameter feedback module receives, analyzes, and evaluates the excitation data. It is judged whether the range of the best excitation effect has been reached. The excitation will be stopped when the data access module reaches the range of the best excitation effect. The modified experimental parameters will be given and passed to the parameter automatic control module. After reaching the requirement, the parameter will be written into the data access module and stop working.

The multidegree-of-freedom base of the ultrasonic gun is moved to a suitable position and fixed. The orientation of the three rotating frames is adjusted according to the actual situation, the pressure control signal from the ultrasonic controller is received to control the rotation angular displacement of the servo motor, and then the threaded cartridge is pushed by the screw to move forward to obtain the set pressing force. The pressing force can be measured by a force sensor placed in front of the ultrasonic gun fixture. When the pressing force is not suitable, the force sensor transmits the measured value to the ultrasonic controller for observation and adjustment.

Parts of the images collected by the experiment are shown in Figure 10. The grayscale images of the excitation heat maps and the measured crack length after processing by the data processing system are shown in Figure 11. The crack can be observed more clearly, and its length can be measured through the following equation:

$$l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_1)^2}. \quad (1)$$

Finally, a database is established to record the experimental data, including the original experimental conditions (relevant parameters of ultrasonic excitation and so on), the original heat maps, and the final thermal images defect identification information.

It can be seen that the proposed adaptive ultrasonic infrared crack defect detection system can identify defects in real time. The heat produced by the cracks' friction after receiving the ultrasonic excitation can be captured by the infrared thermal imager. A software has been developed to

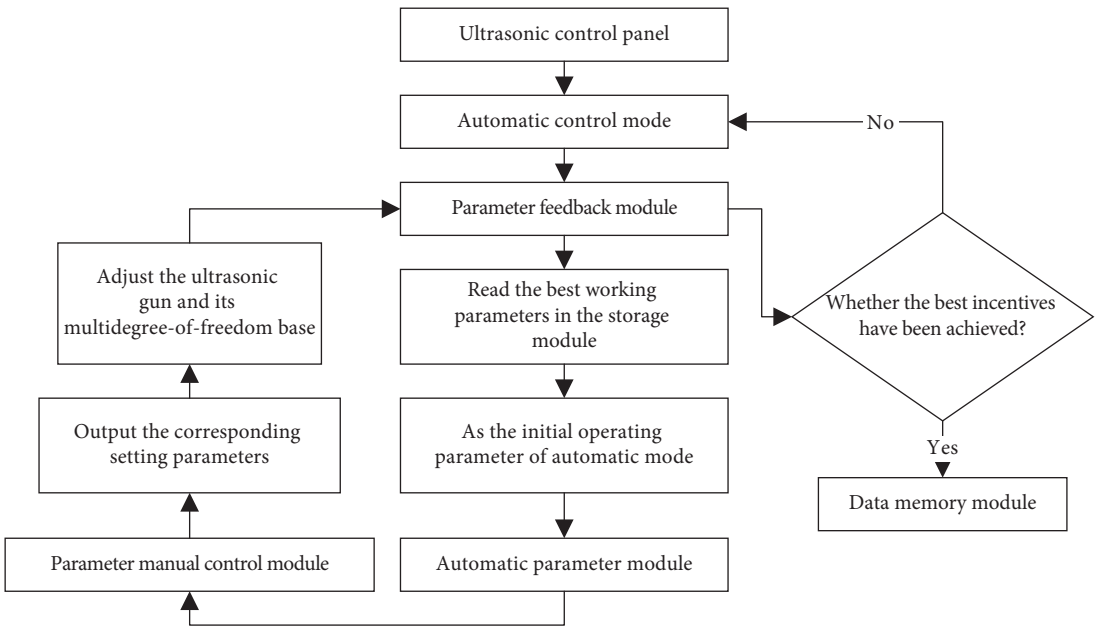


Figure 8: Flow chart of automatic control mode.

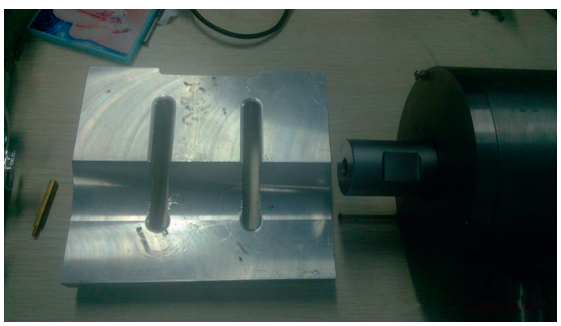
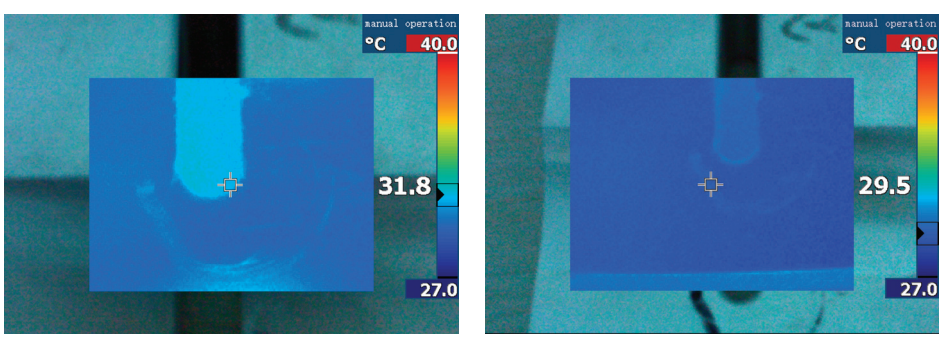
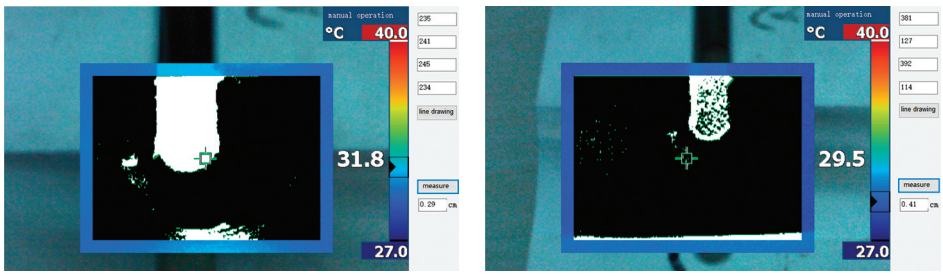


Figure 9: The sample.



(a) (b)

Figure 10: Part of experimental acquisition images.



(a) (b)

Figure 11: The processed images and the crack length.

recognize the defect from the thermal map and provide image process operation choice. Therefore, the defect can be detected, and its size can be actually calculated easily.

5. Conclusions

This paper designs an adaptive ultrasonic excitation device and its control method. The ultrasonic controller based on the original parameter manual control module adds a parameter feedback module, parameter automatic control module, and data access module. The main function of the parameter feedback module is to collect the current equipment working parameters and evaluate and judge the working parameters for the parameters setting of the next step when the equipment is running; the parameter automatic control module is corresponded to the parameter manual control module. With the assistance of the parameter feedback module and data access module, it can automatically correct the working parameters of the equipment and quickly obtain the actual best optimal working parameters; the data access module is mainly to output the best optimal working parameter. It can store the optimal working parameter and write to the parameter feedback module.

Compared with the existing ultrasonic excitation device, this device has the advantages of small device structure, convenient operation, and quicker obtaining the best excitation effect. At the same time, the multidegree-of-freedom base of the ultrasonic gun can adapt to different operating environments and meet the needs of actual use. Our quantitative detection experiments on the surface cracks of metal sheets prove that this technology can detect contact interface defects effectively such as fatigue cracks in composite materials. It can well solve the problem of identifying defects in parts of the metal profile and plates in service environments. And it can also be applied widely in industrial production.

Data Availability

The relevant data in this article were obtained by the author's personal investigation and research and can be made available upon request.

Additional Points

Featured Application. This paper provides a nondestructive detection method for microcracks on the surface of metal materials. Ultrasound excitation will be applied to the surface of the metal material, and the infrared thermal instrument will be used to detect the defect through the improved adaptive infrared imaging recognition method.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yibo Ai and Weidong Zhang conceived and designed the study. Yibo Ai and Xingzhao Cao designed the device and developed the software. Yingjie Zhang gathered the data. Yibo Ai and Yingjie Zhang wrote the paper. All authors read and approved the manuscript. Yibo Ai and Yingjie Zhang contributed equally to this work.

Acknowledgments

This work was supported by the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. 311021013), the Fundamental Research Funds for the Central Universities of China (Nos. FRF-GF-20-24B and FRF-MP-19-014), and the 111 Project (No. B12012).

References

[1] D. Wang, *Fracture Mechanics*, Harbin Institute of Technology Press, Harbin, China, 1989.

[2] K. C. Kim, J. Y. Kwon, and N. W. Kang, "A novel forced-resonance microwave method to detect surface cracks in metal," *Ieice Electronics Express*, vol. 13, no. 17, Article ID 20160715, 2016.

[3] Q. Tang, C. Bu, Y. Liu et al., "Computer simulation of metal surface micro-crack inspection using pulsed laser thermography," *International Journal of Multimedia & Ubiquitous Engineering*, vol. 3, no. 11, pp. 249–256, 2016.

[4] S.-H. Yang, K.-B. Kim, and J.-S. Kang, "Detection of surface crack in film-coated metals using an open-ended coaxial line sensor and dual microwave frequencies," *NDT & E International*, vol. 54, no. 54, pp. 91–95, 2013.

[5] C. Xu, X. Gong, W. Zhang, and G. Chen, "An investigation on eddy current pulsed thermography to detect surface cracks on the tungsten carbide matrix of polycrystalline diamond compact bit," *Applied Sciences*, vol. 7, no. 4, p. 429, 2017.

[6] W. Zhu, Z. Liu, D. Jiao et al., "Eddy current thermography with adaptive carrier algorithm for non-destructive testing of debonding defects in thermal barrier coatings," *Journal of Nondestructive Evaluation*, vol. 3137 pages, 2018.

[7] X. Kou, C. Pei, and Z. Chen, "Fully noncontact inspection of closed surface crack with nonlinear laser ultrasonic testing method," *Ultrasonics*, vol. 114, Article ID 106426, 2021.

[8] L. Feng and X. Qian, "Enhanced sizing for surface cracks in welded tubular joints using ultrasonic phased array and image processing," *NDT & E International*, vol. 116, Article ID 102334, 2020.

[9] Z. A. Wei, A. Sq, L. A. Li et al., "An quantitative inspection method for internal defects based on laser ultrasonic technology," *Optik*, vol. 216, Article ID 164873, 2020.

[10] A. Messager, A. Junet, T. Palin-Luc et al., "In situ synchrotron ultrasonic fatigue testing device for 3D characterisation of internal crack initiation and growth," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 3, no. 43, pp. 558–567, 2020.

[11] Y. Zhu, F. Li, and W. Bao, "Fatigue crack detection under the vibration condition based on ultrasonic guided waves," *Structural Health Monitoring*, vol. 3, no. 20, pp. 931–941, 2021.

[12] Z. X. Zhou, "Overview of NDT methods for mechanical cracks," *Journal of Mechanical & Electrical Engineering*, vol. 10, no. 34, 2017.

[13] G. Dua, V. Arora, and R. Mulaveesala, "Defect detection capabilities of pulse compression based infrared non-destructive testing and evaluation," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7940–7947, 2020.

[14] Z. Qu, P. Jiang, and W. Zhang, "Development and application of infrared thermography non-destructive testing techniques," *Sensors*, vol. 14, no. 20, p. 3851, 2020.

[15] F. Flora, M. Boccaccio, G. Fierro et al., "Non-destructive thermography-based system for damage localisation and characterisation during induction welding of thermoplastic composites," *Thermosense: Thermal Infrared Applications XLII*, vol. 11409, Article ID 114090I, 2020.

[16] W. Jia, "Experiment into nondestructive testing of rail foot cracks using infrared thermal waves," *Infrared Technology*, vol. 2, no. 42, pp. 163–167, 2020.

[17] F. Jing, L. Peng, H. Jiang, L. Chen, and W. Yibing, "Crack detection of locomotive hook tongue based on ultrasonic thermography," *Infrared Technology*, vol. 2, no. 42, pp. 158–162, 2020.

[18] M. Schwarz, M. Schwarz, S. Herter et al., "Nondestructive testing of a complex aluminium-CFRP hybrid structure with EMAT and thermography," *Journal of Nondestructive Evaluation*, vol. 38, no. 1, 35 pages, 2019.

[19] S. D. Holland, L. Koester, J. Vaddi et al., "VibroSim: a hybrid computational/empirical model of vibrothermography non-destructive evaluation," *Review of Progress in Quantitative Nondestructive Evaluation: Incorporating the European-American Workshop on Reliability of NDE*, vol. 1706, pp. 249–276, Article ID 100008, 2016.

[20] J. Vaddi, S. D. Holland, and R. Reusser, "Transducer degradation and high amplitude behavior of broadband piezoelectric stack transducer for vibrothermography," *Review of Progress in Quantitative Nondestructive Evaluation*, vol. 31, no. 1430, pp. 552–558, 2012.

[21] Lesthaeghe, "Evaluation of some parameters influencing vibrothermographic crack heating," Dissertations & These-Gradworks, Iowa State University, Ames, IA, USA, 2015.

[22] F. Ma and X. Guo, "Modeling and analysisi of vibrotherography for the detection of microcracks," *Nondestructive Testing*, vol. 9, no. 37, pp. 6–10, 2015.

[23] C. Zhang, A. Song, F. Fuzhou et al., "Study on optimization methods of ultrasonic infrared thermography detection conditions," *Infrared and Laser Engineering*, vol. 2, no. 45, pp. 77–84, 2016.

[24] X. Han and Y. Song, "Study the effect of engagement force of ultrasound transducer on crack detectability in sonic IR imaging," *AIP Conference Proceedings*, vol. 1, no. 1511, pp. 532–538, 2013.

[25] F. Fuzhou, C. Zhang, Q. Min et al., "Heating characteristics of metal plate crack in sonic IR imaging," *Infrared and Laser Engineering*, vol. 5, no. 44, pp. 1456–1461, 2015.

[26] J. P. Yun, W. C. Shin, G. Koo, M. S. Kim, C. Lee, and S. J. Lee, "Automated defect inspection system for metal surfaces based on deep learning and data augmentation," *Journal of Manufacturing Systems*, vol. 55, pp. 317–324, 2020.

[27] H. Gao, X. Qin, R. J. D. Barroso et al., "Collaborative learning-based industrial IoT API recommendation for software-defined devices: the implicit knowledge discovery perspective," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 99, pp. 1–11, 2020.

[28] H. Gao, W. Huang, and Y. Duan, "The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments," *ACM Transactions on Internet Technology*, vol. 21, no. 6, 2021.

[29] L. Kuang, T. Gong, S. OuYang, H. Gao, and S. Deng, "Offloading decision methods for multiple users with structured tasks in edge computing for smart cities," *Future Generation Computer Systems*, vol. 105, pp. 717–729, 2020.

Hindawi
Complexity
Volume 2021, Article ID 6325578, 20 pages
https://doi.org/10.1155/2021/6325578

Research Article

Urban Road Network Emergency: An Integrative Vulnerability Identification Method

Huaikun Xiang

School of Automotive & Transportation Engineering, Shenzhen Polytechnic, Shenzhen 518055, China

Correspondence should be addressed to Huaikun Xiang; xianghuaikun@szpt.edu.cn

Received 2 May 2021; Accepted 30 May 2021; Published 11 June 2021

Academic Editor: Long Wang

The vulnerability of an urban road network is affected by many factors, such as internal road network layout, network structure strength, and external destructive events, which have great uncertainty and complexity. Thus, there is still no unified and definite vulnerability analysis scheme available to cities. This paper proposes an integrative vulnerability identification method for urban road networks, which mainly relates to the vulnerability connotation and characteristics analysis of urban road networks during emergency, and vulnerability comprehensive evaluation indices design based on urban road network connectivity, traffic efficiency and performance, and an empirical study on a vulnerability identification method of an urban road network. In the empirical case, a real road network and traffic operation data were used from Science and Technology Park of Shenzhen City, China. In the context of one certain emergency scenario, the stated preference survey method and maximum likelihood method are used to solve the road users' random travel choice behavior parameters; subsequently, based on the traffic equilibrium distribution prediction, the traffic vulnerability identification methods of the road network in this region were verified before and after the emergency. The method presented here not only considers the impact of network topology changes on road network traffic function during emergency but also considers the impact of dynamic changes in road network traffic demand on vulnerability; therefore, it is closer to the actual distribution of urban road network traffic vulnerability.

1. Introduction

The urban road traffic system is called the "lifeline" system of a city, which plays a key role in the daily commuting, logistics, and business travel of the city as discussed by various sources [1–3]. However, all kinds of traffic accidents, road failures (such as landslides or flash floods), or bad weather (such as rainstorms, fog, or blizzards) often create significantly adverse effects on the normal operation of urban transportation network. Under the impact of various emergencies, the vulnerability of urban road network (VURN) is constantly exposed. Some vulnerable links in the network not only lead to the delay of network travel time but also cause cascading failure or an avalanche effect of the whole road network due to interruption and congestion. The American Road and Transportation Builders Association points out that nearly 75% of freight transportation in the United States is carried out through vulnerable links (such as highways and traffic bottlenecks), which cause approximately 243 million hours of delay for truck drivers every year [4, 5]. In this context, the VURN is regarded as an important research topic [6].

The VURN's main research purpose is to obtain a method that can quickly, accurately, and reliably identify those vulnerable links in the road network, to provide assistance to reduce the losses caused by various emergencies on vulnerable roads [7, 8], and fundamentally provide decision support for improving the risk resistance ability of urban road networks. Usually, vulnerability analysis of the road network is carried out on road topology network models based on network theory. Generally, the method of gradually moving some nodes or connecting edges in the road network is adopted to calculate and compare the changes of connectivity or traffic efficiency indexes of the whole road network before and after the change of road network structure; accordingly, the nodes or connecting

edges which calculation results vary greatly are defined as the vulnerable parts of the road network. Usually, the performance parameters of complex topological network are selected as the connectivity index or traffic efficiency index of the road network, such as the degree, betweenness centrality, community structure, and cluster coefficient [9–17]. Based on these indicators to identify the VURN, although the results can identify the vulnerable part of the network to a certain extent, there are obvious defects in the practical application of the identification results. The main problem is that it cannot reflect the VURN and change the state of vulnerability [11]. Research shows that the public network with statistical significance may show completely different network characteristics [18].

Due to carrying various complex traffic and transportation activities, urban road networks will be impacted to varying degrees in addition to the topological structure changes of the road network after an emergency, which introduces difficulties to the vulnerability identification of the whole road network. The key problem is how to accurately, effectively, and reliably define the quantitative relationship between various traffic activities and the VURN. In this case, scholars have proposed a series of vulnerability evaluation indicators to reflect the traffic changes of a road network, such as the travel time based on the network robustness index [19, 20]. Based on the analysis of the connotation and characteristics of the VURN, this paper proposes an integrative identification index which can comprehensively reflect the changes of urban road network vulnerability and designs the corresponding identification model and algorithm. The main contributions are as follows:

(1) A three-dimensional model for vulnerability analysis of an urban road network under emergency is established. The model can better explain the concept of the VURN, which mainly involves three variables, namely, the process variable of emergency, the traffic loss variable of the road network caused by the emergency, and the probability variable of the emergency.

(2) An integrative vulnerability identification index system of an urban road network is established. The VURN is affected by many factors, not only in the connectivity and stability of the network topology, but also in the effectiveness and reliability of the network service function, which are related to the process of emergency. Therefore, based on the comprehensive identification index system of the emergency occurrence process, the city can more accurately identify the VURN.

(3) The traffic demand is included in the vulnerability identification process. The comprehensive vulnerability identification model is constructed before and after the emergency, and the corresponding efficient algorithm is designed to identify the VURN.

The vulnerability identification method of an urban road network based on a comprehensive identification index system proposed here can reflect the vulnerability of a road network more comprehensively, so that it can identify those key locations and sections more accurately. The remainder of the paper is organized as follows: a conceptual and characteristic framework for vulnerability analysis is proposed. An integrative vulnerability identification method of the VURN during emergency is discussed, which mainly includes the design of the vulnerability index system, the improvement of the network traffic impedance function, and network traffic assignment modeling and model solving algorithm design. The proposed vulnerability identification method based on actual road network traffic survey data is verified and analyzed. The results of urban road network vulnerability discrimination under different index combinations are analyzed.

2. Related Work

Early studies of vulnerability emerged in the field of engineering and were then extended to the social economic and political institutional level by social scientists. In research related to road network disruption, efforts were made to define the vulnerability of road network [6, 21, 22]. Since then, several studies have attempted to provide methods for evaluating network performance in terms of vulnerability or robustness [20, 23–27]. In fact, the vulnerability assessment method of urban road networks is closely related to the vulnerability index, and this is related to the understanding of the concept of vulnerability. Therefore, the cognition of the concept directly determines the design of an urban road network vulnerability assessment method. The following literature review focuses on the relevant research results on urban road network vulnerability in recent years.

To date, there is no definite and unified definition of the concept of road network vulnerability [28]. Berdica [6] discussed the concept and considered that the vulnerability of a road network is sensitive to the time when the service capacity has declined sharply. Holmgren [29] believes that vulnerability should comprise the sensitivity of road networks to disaster risk. Husdal [30] considers that "vulnerability describes the nonoperability of the networks under varying strenuous conditions (i.e., the susceptibility to fail to function)". Taylor et al. [31] believe that if a small number of road sections in the network fail or significantly degrade the accessibility of nodes, then the nodes of the network are fragile; and if the failure of a certain road section will likely reduce the accessibility of the whole or some nodes of the network, then this road section is considered critical.

Erath et al. [32] studied the VURN in Switzerland. They defined the vulnerability of road network as the product of the probability of road failure under dangerous conditions and the sum of direct and indirect consequences caused by interruption. Jenelius et al. [33] believe that road network vulnerability can be defined as the study of the potential degradation of the road transport system and its impact on society; hence, risk theory can be used to measure the vulnerability of a road network. According to Yin et al. [15], the issues of vulnerability are "which areas in the road transportation system are easy to interrupt" and "which connections are the most critical to the operation of the

whole system." Therefore, the vulnerability of a road network can be regarded as the integration of the possibility of road section failure and the potential impact of failure. Yang et al. [34] believe that the vulnerability of a road traffic network is that some units (including nodes and road sections) in the network are damaged or have failed due to the disruption of emergency situations, through the interaction between road network units and the continuous influence of the outside world. This type of loss or failure transfers among other units in varying degrees and finally contributes to the loss measurement of the whole road traffic network.

Emergency can be natural or man-made damage events, or small-scale events, such as explosion or fire in buildings, or large-scale damage events, such as an earthquake, radiation accident, bombing, or dangerous weather conditions in cities or areas [35]. The International Strategy for Disaster Reduction (UNISDR) defined a disaster as the negative effects of hazards on vulnerable socioeconomic systems, where vulnerability limits the coping capability to the impact of the hazard [36]. Emergencies are characterized by strong randomness, wide coverage, and a significant negative diffusion effect [37]. Once emergencies occur, it will be in a state of rapid diffusion initially. Once this state affects the urban transportation activities, its impact on urban road network traffic begins [28, 38, 39]. After a period of development and change, the impact on road traffic will reach a maximum, mainly manifested in the effect on road traffic infrastructure, the interruption of traffic flow operation, and the impact of travel demand fluctuation. When the emergency is finally under control, the impact ends. From this point forward, road traffic will enter the process of gradual recovery. For different road traffic networks, the recovery time varies. A short recovery time indicates that the road traffic network is more robust, and conversely, it indicates that the network is more vulnerable [40].

The connotation of road traffic network vulnerability should include the following characteristics: (1) vulnerability is an inherent attribute existed recessively when there is an emergency and is present in a road traffic network. (2) When encountering emergency, the inherent vulnerability of a road traffic network emerges. If the road traffic network is more prone to emergency situations, the inherent vulnerability is greater. (3) In examining the impact of emergency, if the loss degree of the road traffic network is more serious, it indicates that the inherent vulnerability is greater and that the road network is more vulnerable to external interference factors. (4) Once the emergency and losses are under control, if the service level of the road traffic network is conducive to effective restoration under external repair, the impact of the vulnerability is considered low; conversely, the impact is categorized as high. These characteristics of road traffic network vulnerability are shown in the concept map of a three-dimensional model as shown in Figure 1.

Figure 1 shows a three-dimensional coordinate system composed of probability, time, and loss, where t_s is the start time of the emergency, t_m is the time when the impact of the event reaches the maximum, and t_e is the impact end time. The time period from t_s to t_m is called the influence period, in which the emergency will have a significant impact on the road traffic infrastructure, traffic operation, and the choice behavior of residential travel. The impact of the infrastructure and traffic operation mode has the most direct effect, and the response is also the fastest and easy to be spread, which is the most active and easily changed element in the whole system. Due to the transmission of information takes time, there is no way to quickly inform road users of a potential road block. This contributes to further unnecessary congestion in the emergency zone.

The period from t_m to t_e is called the recovery interval. Owing to the superposition of various influences, the recovery of a road traffic network needs a period. $P_{\max}$ is the maximum probability of an emergency, and $l_{\max}$ is the maximum loss caused by an emergency. Therefore, we define the vulnerability of a road traffic network as follows: the vulnerability of a road traffic network refers to the comprehensive embodiment of the probability of a road traffic network being affected, the severity of loss consequences, and the difficulty of recovery in an emergency encounter.

3. Materials and Methods

3.1. Index Design. The VURN is the result of a variety of complex, different, and uncertain factors, such as road conditions, network structure, traffic flow state, rescue point setting, resource scheduling, and repair strategy. When a serious traffic accident occurs, some units in the road network will be damaged or even invalid due to the situation. Simultaneously, the urban transportation department takes measures to ensure smooth continuity of the road network and conducts emergency rescue work to address the incident. If the road network is continuously affected by emergency, multiple types of damage can occur and spread impacting the road network. The transmission of road network vulnerability between road network units is triggered with the function failure of road network units [38], resulting in the so-called "avalanche effect," which creates significant difficulties for emergency rescue. This continues until the road network function returns to normal and the network is stabilized. Figure 2 expresses the road traffic flow distribution of this process as a schematic diagram.

In Figure 2, t_s represents the starting time of the event, and the number of vehicles on the road is $x(t_s)$. After the incident, the road is interrupted, as it is not possible to share the information with every vehicle on the road in a timely manner, and because the vehicles entering the section have no choice, the inflow rate of vehicles in the accident section exceeds the outflow rate, until the accident is managed; that is, entering time point t_m, the number of vehicles on the road is $x(t_m)$. Starting from t_m, the outflow rate of vehicles on the road will be slightly higher than the inflow rate until the original capacity of the road is fully restored. This is time t_e, and the number of vehicles on the road is $x(t_e)$. Under ideal conditions, $x(t_s)$ is approximately equal to $x(t_e)$.

The VURN reflects the degree to which the network and network units are vulnerable to various interference factors. The more sensitive the network is to interference factors, the more vulnerable the parts are to the impact [28]. According

to the previous analysis, after the road network suffers from emergency, the road network structure will be affected to varying degrees, and the traffic operation will be disturbed, which will be reflected in the loss degree of the whole network [41, 42], specifically in the changes of network connectivity [43], operation timeliness [44], service level, etc. Therefore, based on the influence degree of interference factors, this study designs a comprehensive identification index system of urban road network vulnerability under emergency conditions.

Depending on the severity of the situation, once an emergency has occurred, the topology of the original road network may significantly change, and this manifests in the node failure caused by interruption to traffic and sudden congestion at the intersection, thus affecting the connectivity of the road network.

3.1.1. Connectivity of Road Network and Its Change Measure in Case of Emergency. In this study, the road network connectivity G is defined as the ratio of the actual number of edges to the maximum number of edges in the network. For a network node, the greater the change of connectivity before and after failure, the more significant the impact of the node is on the robustness of the road network, and the same is true for the connected edges of the network. The calculation formula of network connectivity is as follows:

$$G = \frac{D}{3V_d - 6}, \tag{1}$$

where D is the number of edges in the network and V_d ($V_d > 3$) is the number of nodes in the network. When the point fails, the actual number of edges D' of the road network after the failure is counted, and the corresponding connectivity G' and the change ΔG of the road network connectivity before and after the failure are calculated:

$$\Delta G = \left|G - G'\right| = \left|\frac{D}{3V_d - 6} - \frac{D'}{3\left(V_d - 1\right) - 6}\right|. \tag{2}$$

3.1.2. Network Efficiency and Its Change Measure during Emergency. In the event of emergency, the sudden congestion of the road section leads to the failure of a link, which affects the efficiency of the network traffic. In this paper, the mean value of the edge betweenness efficiency of the topological network is defined as the network efficiency under emergency. Research shows that the greater the network efficiency change rate before and after failure, the more significant the impact of the road section on road network robustness [41]. The calculation formula of network efficiency E is as follows:

$$E = \frac{\sum_{k_1,k_2 \in N, k_1 \neq k_2}\left(1/l_{k_1k_2}\right)}{n(n-1)}, \quad (n > 1), \tag{3}$$

where N is the node set, n is the total number of road network nodes, and $l_{k_1k_2}$ and $l'_{k_1k_2}$ are the shortest paths between the connection nodes k_1 and k_2 before and after the failure. When an edge fails, count and calculate the actual number of edge betweenness and the corresponding network efficiency after the failure and calculate the change of network efficiency ΔE before and after the edge failure according to formula (3):

$$\Delta E = \left|\frac{\sum_{k_1,k_2 \in N, k_1 \neq k_2}\left(1/l_{k_1k_2}\right)}{n(n-1)} - \frac{\sum_{k_1,k_2 \in N, k_1 \neq k_2}\left(1/l'_{k_1k_2}\right)}{n(n-1)}\right|. \tag{4}$$

3.1.3. Road Network Traffic Performance and Its Change Measure during Emergency. Considering that the urban road network is facing the impact of emergency, besides the change of the structure performance, the traffic operation of the road network will also be directly affected. Therefore, this study designs a comprehensive vulnerability measurement index which is more consistent with the actual situation of urban road traffic. Let C_i denote the overall loss of the road network after unit i in the road network is impeded by an emergency, also known as path impedance, I denotes the number of road network units, K denotes the number of failure units in this case, k denotes the kth failure unit (road section or intersection) in the road network, and $t(x)_k$ denotes the loss of unit k under traffic operation condition x, and it is as follows:

$$C_i = \sum_{k=1}^{K} t(x)_k, \quad k = 1, 2, \ldots, K, \; i = 1, 2, \ldots, I. \tag{5}$$

In the formula, $t(x)_k$ is related to the emergency rescue technology R_k at k and the failure loss of the upstream failure unit. For the specific calculation method, see the theoretical derivation in "Traffic Analysis during Emergency" in Section 3.2. Here, the final loss of road network users is selected as the measure. As the unit loss propagates in turn, $k-1$ represents the nearest upstream failure unit k, so $t(x)_k = f\left(t(x)_{k-1}, R_k\right)$ is obtained. The final vulnerability of the road network is caused by the sudden impedance of the road network unit to measure the maximum road network loss of the project:

$$L = \max C_i, \quad i = 1, 2, \ldots, I. \tag{6}$$

In the above formula, L is the indicator of traffic VURN. $\Delta L = |L - L'|$ is the change of total loss before and after the emergency.

3.1.4. Integrated Identification Index of Urban Road Network Traffic Vulnerability during Emergency. From the previous analysis of urban road traffic network vulnerability, we can observe that the VURN is mainly characterized by its own structure and traffic function. Therefore, based on a summary of existing research results, this study selects three key index parameters to construct the traffic vulnerability identification index of the urban road network, which include the connectivity index, the traffic efficiency index, and the traffic performance index. However, these three indicators only judge the VURN from different aspects. In order to judge the VURN more accurately and comprehensively, it

is necessary to develop a more comprehensive identification index based on these three vulnerability indicators.

First, the three subindicators are standardized without dimension, so that the three subindicators of vulnerability after data standardization are $\{\Delta G', \Delta E', \Delta L' | \in (0,1)\}$. On this basis, based on the weighted method, the comprehensive identification index of vulnerability v under an emergency is constructed as follows:

$$v = \varepsilon \Delta G' + \varnothing \Delta E' + \gamma \Delta L'. \tag{7}$$

In the above formula, ε, $\varnothing$, and γ represent the weight factors of the road network connectivity index, traffic efficiency index, and urban road network traffic performance index, respectively, whereε, $\varnothing$, and γ are less than 1, and $\varepsilon + \varnothing + \gamma = 1$. Because it is difficult to determine the correlation between the three vulnerability subevaluation indexes objectively and quantitatively, this study uses the classic Delphi method [45] which involves collecting industry expert scores to determine the weight factor. Simultaneously, according to the Delphi method to determine the comprehensive level of urban road network traffic vulnerability under emergency situations, it is divided into five levels, including very low (Level 1), low (Level 2), medium (Level 3), high (Level 4), and very high (Level 5). The specific scale is shown in Table 1.

3.2. Traffic Analysis during Emergency

3.2.1. Road Network Impedance during Emergency

(1) Section Impedance. When an emergency occurs, if different types of travelers (a total of D) have uncertain estimation bias for the travel time of the road section in the traffic network, the estimated cost T_a^d of the class d travelers for the road section a can be expressed as follows:

$$T_a^d = t_a(x_a) + \varepsilon_a^d, \quad \forall a, d, \tag{8}$$

where x_a is the flow of segment a and $t_a(x_a)$ is the cost determined by segment a. $t_a(x_a)$ adopts the road resistance calculation formula which reflects the road traffic operation characteristics under emergency situations derived from the Bureau of Public Road (BPR) function of the United States:

$$t_a = t_a^0\left[1 + \boldsymbol{\alpha}(y_a)^{\beta}\right],$$
$$y_a = \begin{cases} \dfrac{x_a}{c_a}, & y_a \in [0,1], \\ \dfrac{(2c_a - x_a),}{c_a} & y_a \in [1,2]. \end{cases} \tag{9}$$

In the above formula, t_a^0 is the driving time of the road section between two intersections when the traffic volume is 0, also known as zero flow cost; c_a is the actual traffic capacity (vehicle/h) of road section a; and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the undetermined parameters of the model, with the general values of 0.15 and 4, respectively [45], which can also be calibrated according to the measured data [46]. The relationship between impedance and traffic flow is increasing monotonically, which reflects the traffic congestion effect under emergency situations. ε_a^d is the random deviation of class d travelers to Section a, and its mathematical expectation is $E[\varepsilon_a^d] = 0$. Therefore, the estimated cost T_a^d of the link impedance is a random variable, and $E[T_a^d] = t_a$, $\forall a, d$.

(2) Path Impedance. In the urban road network $G(V, E)$, there is an effective path set K_{rs} between all the OD (origin destination) pairs rs, and then the cost estimate $C_{d,k}^{rs}$ of the d class traveler for the kth path in the road network is

$$C_{d,k}^{rs} = \sum_a T_a^d \boldsymbol{\delta}_{a,k}^{rs}, \quad \forall r, s, d; k \in K_{rs}, \tag{10}$$

where $\boldsymbol{\delta}_{a,k}^{rs}$ is the correlation coefficient between link a and path k. If link a is on the kth path between rs, then $\boldsymbol{\delta}_{a,k}^{rs} = 1$; otherwise, $\boldsymbol{\delta}_{a,k}^{rs} = 0$. Similarly, according to the accumulation of random variables, $C_{d,k}^{rs}$ is also a random variable, and $E[C_{d,k}^{rs}] = c_{d,k}^{rs}$, that is,

$$c_{d,k}^{rs} = \sum_a t_a \boldsymbol{\delta}_{a,k}^{rs}, \quad \forall r, s, d; k \in K_{rs}, \tag{11}$$

where $c_{d,k}^{rs}$ is the mathematical expectation of the random variable $C_{d,k}^{rs}$. From the above analysis, the path impedance $C_{d,k}^{rs}$ not only reflects the road traffic operation state of each road section but also has the monotonous change characteristics, which can ensure the calculation needs of the unique solution of the stochastic user equilibrium assignment model.

3.2.2. Travel Choice during Emergency. In case of emergency, each traveler will make their own choice on the travel path and travel mode according to their own specific conditions. Each path has an equal probability of being selected [47]. According to the Wardrop traffic equilibrium principle, the probability $P_{d,k}^{rs}$ of the class d traveler selecting the route k among rs is the probability that the estimated cost (road impedance) is the probability that the estimated cost (road impedance) is the least on all possible paths between the origin and destination (OD) pairs, namely,

$$P_{d,k}^{rs} = P\left(c_{d,k}^{rs} \le c_{d,l}^{rs}, \forall l \ne k\right), \quad \forall r, s, d; k \in K_{rs}. \tag{12}$$

The above formula is calibrated according to the logit model [48], and the specific form of $P_{d,k}^{rs}$ can be obtained as follows:

$$P_{d,k}^{rs} = \frac{\exp\left(-\theta_d c_{d,k}^{rs}\right)}{\sum_{l \in K_{rs}} \exp\left(-\theta_d c_{d,l}^{rs}\right)}, \quad \forall k \in K_{rs}, \tag{13}$$

where a is the parameter to be calibrated in the logit model for class $P_{d,k}^{rs}$ travelers.

3.2.3. Road Network Traffic Distribution during Emergency. Considering the behavior of travelers' route selection in emergency situations, the traffic flow $f_{d,k}^{rs}$ of the kth route in one of OD pair rs and OD demand q_{rs} can meet the following conditions:

$$\begin{cases} f_{d,k}^{rs} = q_{rs} u_d^{rs} P_{d,k}^{rs}, \\ \sum_d u_d^{rs} = 1, \end{cases} \tag{14}$$

where u_d^{rs} is the proportion of type d^{th} travelers in OD pair rs.

From the above model, under the premise that q_{rs} and u_d^{rs} are known, the path flow $f_{d,k}^{rs}$ is related to the path selection probability $P_{d,k}^{rs}$, while $P_{d,k}^{rs}$ is determined by the parameter θ_d and the estimated value of path impedance $c_{d,k}^{rs}$, which is affected by the estimated value of travel time T_a^d, where T_a^d is the function of flow x_a; therefore, continuous selection can ensure that the network traffic finally reaches the user equilibrium state. The user equilibrium assignment model is given as follows [45]:

$$\min Z(f) = \sum_d \sum_r \sum_s \sum_{k \in K_{rs}} \frac{1}{\theta_d} f_{d,k}^{rs} \ln f_{d,k}^{rs} + \sum_a \int_0^{x_a} t_a(w) \mathrm{d}w, \tag{15}$$

$$\text{s.t. } x_a = \sum_d \sum_r \sum_s \sum_{k \in K_{rs}} f_{d,k}^{rs} \delta_{d,k}^{rs}, \tag{16}$$

$$\sum_d \sum_{k \in K_{rs}} f_{d,k}^{rs} = \sum_d q_{rs} u_d^{rs} P_{d,k}^{rs}, \tag{17}$$

$$\sum_d u_d^{rs} = 1, \tag{18}$$

$$f_{d,k}^{rs} \geq 0. \tag{19}$$

The above equations (16)–(19) are road flow constraint, path flow constraint, traveler proportion constraint, and nonnegative flow constraint, respectively, in which different travelers type d, proportion u_d^{rs}, and route selection probability $P_{d,k}^{rs}$ are considered. Obviously, when q_{rs}, u_d^{rs}, θ_d, and other parameters are known, the objective function of the model is strictly convex with respect to $f_{d,k}^{rs}$, and the constraints are linear, so it has a unique path flow solution.

Further analysis shows that parameter θ_d has the random characteristics of the whole model. When $\theta_d \longrightarrow \infty$ is satisfied, the objective function becomes a standard user's equilibrium (UE) problem. When $\theta_d \longrightarrow 0$ is satisfied, OD demand matrix $[q_{rs}]$ will be distributed evenly on the road network, and all path costs will be equal.

3.2.4. Design of Comprehensive Vulnerability Identification Algorithm. Under the impact of emergency situations, there is a sudden impact on road traffic at the beginning of the event, and subsequently to the traffic demand changes caused by road users' travel choice behavior after information diffusion, so that the traffic flow of each OD trip in the road network is redistributed. In this process, the traffic impedance and its induced travel efficiency will be highlighted, so that the extent of the road network vulnerability can be identified here. Based on the previous theoretical analysis and modeling, the vulnerability identification algorithm is described as follows:

Step 1: for the selected road network, the design and investigation plan shall be carried out: (i) SP investigation of travel selection behavior in the absence of emergency and (ii) in the case of emergency. Based on the survey data and the principle of 3.3, the parameters of the logit model are calibrated to obtain the trip selection probability of all kinds of type d travelers in the event of emergency.

Step 2: for the road network $G(V, E)$, according to the principle of 3.4, the demand q_{rs} of OD pair rs is allocated to the network using the random average assignment method, and the traffic volume x_i $(i \in L)$ of each road section is obtained;

Step 3: according to the principle of 3.2, calculate the impedance T_a^d and path impedance $c_{d,k}^{rs}$ of each road section, and then calculate the operation efficiency of the road network before the emergency according to equations (1)–(7).

Step 4: after an emergency occurs, the topological structure of the road network is broken down one by one according to its impact on nodes (intersections) and road sections, and then the path impedance of each road section is calculated step by step; finally, the maximum road network loss caused by the emergency is calculated according to equations (1)–(7).

Step 5: judge whether the vulnerability index calculation of all key nodes and key sections is complete. If so, go to Step 6; otherwise, return to Step 4.

Step 6: terminate the algorithm and output the ranking results of vulnerability indicators.

4. Application

4.1. Data Preparation. In this study, the road network of the Nanshan Science and Technology Park area in Shenzhen City, Guangdong Province, China, was used to verify the above research methods, models, and algorithms. The Science and Technology Park was funded and constructed in 2001 and covers an area of 706,000 square meters and a construction area of 3 million square meters. The park was in an advantageous geographical environment, rich cultural atmosphere, and numerous scientific research institutes. The road traffic network is an important infrastructure to support the social and economic development of the area.

According to statistics, at present, there are approximately 450,000 people employed in the Science and Technology Park area in Nanshan District of Shenzhen City, with a construction area of about 11 million square meters. It is estimated that by 2025, it will reach 800,000 people and 22.5 million square meters, and the existing road carrying capacity can only meet approximately 71% of the current travel demand. At present the average speed of the road network in the area is lower than the international congestion warning line (20 km per hour). According to the calculation, the traffic capacity of the main roads (approximately 28,000

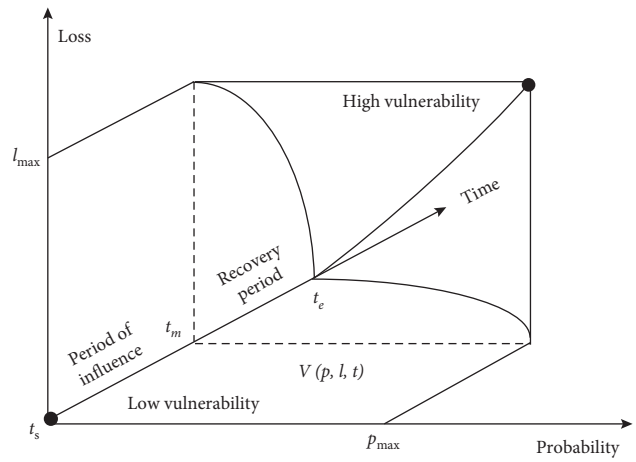


Figure 1: Three-dimensional model of road network vulnerability concept.

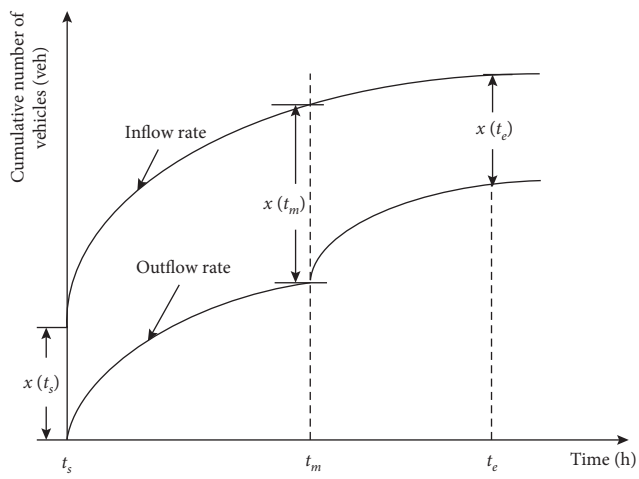


Figure 2: Schematic diagram of road traffic volume change after emergency.

vehicles/peak hour) cannot meet the demand of vehicles (approximately 40,000//peak hour). Figure 3 shows the electronic map of the local road network in this area as extracted by ArcGIS software. According to the research scheme, the backbone road network was selected on the electronic map, and the corresponding road network topology was constructed using the vision software. Figure 4 shows that there are 35 nodes and 59 arc segments, including two starting points (origin points) and two ending points (destination points). The arc segment of a single arrow indicated that the road section was one-way traffic, and the arrow at both ends indicates that the road section was two-way traffic. Based on the data provided by the department in charge of urban road traffic, the research group obtained the road traffic data of the area in combination with the field investigation (see Table 2).

4.2. Emergency Scenario and Impact Analysis. Emergency mainly refers to natural and accidental disasters, public health, and social security events that suddenly occur or may cause serious social harm where emergency measures are required to deal with them. These events were characterized by strong randomness, wide coverage, and significant negative diffusion effect. When an emergency occurs, on the one hand, it leads to interruption and changes the topological structure of the relevant road network; on the other hand, due to these changes, traffic and transportation activities are affected, and the travel demand also changes accordingly. With different degrees of information suspension and different audiences, the impact of emergency on road traffic activities would be different.

According to the stated preference (SP) survey, when the accident occurs, the road users in the regional road network could be divided into two types: (1) those who wait for the road section to be repaired, reopen, and drive according to their original road plan after receiving the information; (2) those who to return to the nearest exit through the reverse evacuation of traffic to find the shortest alternative path. Taking the topological road network shown in Figure 4 as example, it was assumed that there was a serious traffic accident on section 37 of the road network, which may produce the following two effects: (1) once the accident occurred, section 37 was interrupted. The vehicles originally

passing through section 37 would be diverted to other nearby roads, which would exert pressure on related traffic, and the traffic in this area would be redistributed; (2) after the accident, intersection 25 and 13 would be blocked. Due to the interruption of section 37, the traffic flow of two related intersections was blocked, and the intersection was affected by vehicle queuing, resulting in cascading failure. As a consequence, the topological structure of the road network would change to some extent.

4.3. Investigation of Road Users' Travel Choice Behavior. To obtain the characteristic parameters of road users' travel choice behavior under the emergency conditions and to obtain the probability of the same after the accident and its impact on the road traffic operation state, we investigated road users' travel choice behavior based on the set emergency scenario. The survey's purpose was to identify how those who originally planned to use the road network would adapt and select an alternative path once the original planned path had been interrupted. The possible situations include three travel plan choices: cancel travel, suspend travel, and continue with normal travel. In terms of travel mode selection, let us assume that there were three modes of transportation: private car, bus, and taxi. The survey content mainly included the influence of personal attributes (mainly occupation and age) and the travel purpose of road users, and the travel time and travel mode choice behavior under different travel purposes under the characteristics of road users' personal information and emergency.

Based on the survey, this project designed an SP questionnaire. Through the combination of online and offline surveys, 218 valid samples were finally recovered, which met the minimum requirements of statistical survey samples (157), and reflected the travel choices of different individuals in emergency situations. To quantitatively analyze the impact of personal attributes and options on travel results, this project used the stochastic utility logit model to establish the probability model of travel time and travel mode selection.

4.3.1. Determination of the Utility Function and Parameter Estimation. The previous discussion shows that if N was used to represent the set of possible travel options, and I was used to represent the set of traveler types, then the random utility U_n^i of type i travelers choosing the n travel option could be expressed as follows:

$$U_n^i = V_n^i + \varepsilon_n^i, \quad n \in N, i \in I, \tag{20}$$

where V_n^i was the determinable utility of type i travelers for the n travel scheme, ε_n^i was the random error term, and both obeyed the Gumbel distribution. According to the multinomial logit (ML) model, the number of samples is set as I, and δ_{in} as the probability variable. The probability of simultaneous implementation of each option was as follows:

$$p_{1n}^{\delta_{1n}}, p_{2n}^{\delta_{2n}} \cdots p_{in}^{\delta_{in}} \cdots p_{In}^{\delta_{In}} = \prod_{i \in I} p_{in}^{\delta_{in}}. \tag{21}$$

So, the simultaneous probability L^* of travelers was as follows:

$$L^* = \prod_{n=1}^{N} \prod_{i \in I} p_{in}^{\delta_{in}}. \tag{22}$$

Equation (22) represents the likelihood function of ml. According to equation (13), the log likelihood function could be obtained as follows:

$$L = \ln L^* = \sum_{n=1}^{N} \sum_{i \in I} \delta_{in} \ln p_{in} = \sum_{n=1}^{N} \sum_{i \in I} \delta_{in} \left(\theta X_{in} - \ln \sum_{j \in I} e^{\theta X_{jn}} \right). \tag{23}$$

In the above equation, X was the characteristic variable of road users' travel choice, θ was the parameter to be estimated corresponding to the characteristic variable, and L was a convex function about θ, and using θ to derive the two sides of the equation and setting the value to 0, the maximum likelihood estimation value $\widehat{\theta}$ of θ could be obtained. Here, the parameters of the model were estimated using SPSS software.

4.3.2. Parameter Calculation and Result Analysis. Taking cancellation, postponement, private car, bus, and taxi as travel options, through the correlation analysis of the survey data, the characteristics of the options and that of the travelers were finally determined (see Table 3). The values of the characteristic variables were obtained through actual investigation.

The logistic module of the SPSS software was used to estimate the model parameters. The model characteristic variables corresponding to the normal travel time selection and the parameter estimation values of the corresponding characteristic variables of other transportation modes with the bus as the travel reference in the travel mode selection were obtained. Tables 4 and 5 show the calculation results.

According to the principle of statistics, when the degree of freedom is 1 and the confidence is 0.05, the critical value of the Wald parameter test is 3.841. The larger the Wald test value, the more significant the correlation between independent and dependent variables. From the above Wald test value, we observed that the selected characteristic variables were the factors that had a significant impact on road users' travel choice behavior. Simultaneously, travel time, vehicle ownership, and age had a significant impact on road users' travel mode choice. Table 6 shows the test results for the overall likelihood ratio test of the model.

The confidence interval of the model parameter estimation was 95%, which is shown in Table 6 from the likelihood ratio test results, which demonstrated that the model had significant importance. According to the conditions of superior ratio, the fitting effect of the model was good. For the whole model, the value of MC Fadden's coefficient p^2 was between 0.2 and 0.4, which demonstrates good accuracy. Table 7 shows the calibration results of the traveler's nonintegrator parameters.

Table 1: Vulnerability scale of key nodes or road sections.

	Vulnerability level				
	1	2	3	4	5
Grade scale	[0, 0.51]	[0.51, 0.63]	[0.63, 0.80]	[0.80, 0.91]	[0.91, 1]
Vulnerability	Very low	Low	Medium	High	Very high

Figure 3: Electronic map of road network.

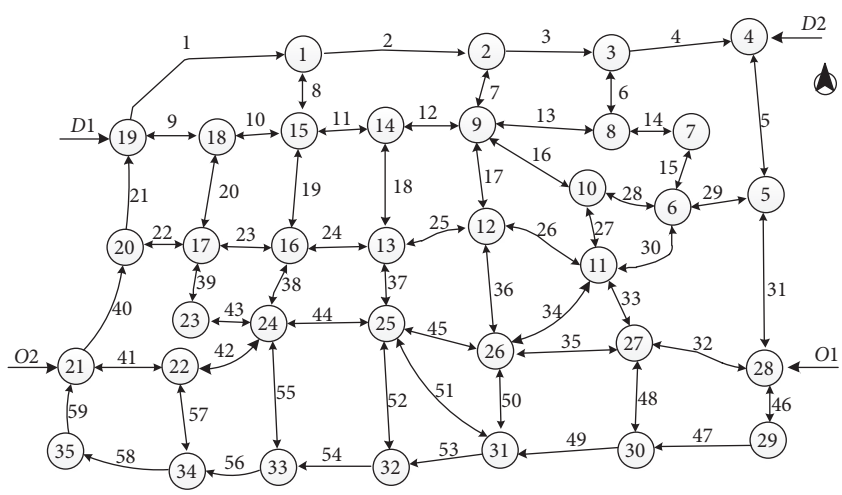


Figure 4: Backbone road network topology.

Table 2: Survey results of road traffic network data.

Nodes (r)	Nodes (s)	Arcs (a)	Lane	Length (m)	Designing velocity (km/h)	Free travelling time (h)	Designing capacity (pcu/h)
19	1	1	4	0.697	60	0.0116	3100
1	2	2	4	0.609	60	0.0102	3100
2	3	3	3	0.386	60	0.0064	2325
3	4	4	3	0.454	60	0.0076	2325
4	5	5	4	0.554	60	0.0092	3100
5	4	5	4	0.554	60	0.0092	3100
3	8	6	4	0.212	40	0.0053	2900
8	3	6	4	0.212	40	0.0053	2900
2	9	7	4	0.238	60	0.004	3100
9	2	7	4	0.238	60	0.004	3100
1	15	8	4	0.256	50	0.0051	3000
15	1	8	4	0.256	50	0.0051	3000
18	19	9	6	0.32	60	0.0053	4650
19	18	9	6	0.32	60	0.0053	4650
18	15	10	6	0.258	60	0.0043	4650
15	18	10	6	0.258	60	0.0043	4650
15	14	11	6	0.258	60	0.0043	4650
14	15	11	6	0.258	60	0.0043	4650
9	14	12	6	0.353	60	0.0059	4650
14	9	12	6	0.353	60	0.0059	4650
8	9	13	6	0.389	60	0.0065	4650
9	8	13	6	0.389	60	0.0065	4650
7	8	14	4	0.566	50	0.0113	3000
8	7	14	4	0.566	50	0.0113	3000
6	7	15	6	0.259	40	0.0065	4400
7	6	15	6	0.259	40	0.0065	4400
9	10	16	4	0.269	40	0.0067	2900
10	9	16	4	0.269	40	0.0067	2900
9	12	17	6	0.325	60	0.0054	4650
12	9	17	6	0.325	60	0.0054	4650
13	14	18	6	0.373	60	0.0062	4650
14	13	18	6	0.373	60	0.0062	4650
15	16	19	6	0.379	60	0.0063	4650
16	15	19	6	0.379	60	0.0063	4650
17	18	20	4	0.383	60	0.0064	3100
18	17	20	4	0.383	60	0.0064	3100
20	19	21	6	0.38	60	0.0063	4650
20	17	22	6	0.249	40	0.0062	4400
17	20	22	6	0.249	40	0.0062	4400
16	17	23	6	0.271	40	0.0068	4400
17	16	23	6	0.271	40	0.0068	4400
13	16	24	6	0.335	40	0.0084	4400
16	13	24	6	0.335	40	0.0084	4400
12	13	25	6	0.359	40	0.009	4400
13	12	25	6	0.359	40	0.009	4400
11	12	26	6	0.49	40	0.0123	4400
12	11	26	6	0.49	40	0.0123	4400
10	11	27	4	0.535	50	0.0107	3000
11	10	27	4	0.535	50	0.0107	3000
6	10	28	4	0.486	40	0.0122	2900
10	6	28	4	0.486	40	0.0122	2900
5	6	29	6	0.296	60	0.0049	4650
6	5	29	6	0.296	60	0.0049	4650
6	11	30	4	0.302	40	0.0076	2930
11	6	30	4	0.302	40	0.0076	2930
5	28	31	6	0.535	60	0.0089	4650
28	5	31	6	0.535	60	0.0089	4650
27	28	32	4	0.435	60	0.0073	3100
28	27	32	4	0.435	60	0.0073	3100
11	27	33	4	0.254	50	0.0051	3000
27	11	33	4	0.254	50	0.0051	3000
11	26	34	6	0.537	40	0.0134	4400

Table 2: Continued.

Nodes (r)	Nodes (s)	Arcs (a)	Lane	Length (m)	Designing velocity (km/h)	Free travelling time (h)	Designing capacity (pcu/h)
26	11	34	6	0.537	40	0.0134	4400
26	27	35	6	0.441	60	0.0074	4650
27	26	35	6	0.441	60	0.0074	4650
12	26	36	6	0.422	60	0.007	4650
26	12	36	6	0.422	60	0.007	4650
13	25	37	6	0.243	60	0.0041	4650
25	13	37	6	0.243	60	0.0041	4650
16	24	38	6	0.428	60	0.0071	4650
24	16	38	6	0.428	60	0.0071	4650
17	23	39	6	0.393	60	0.0066	4650
23	17	39	6	0.393	60	0.0066	4650
20	21	40	6	0.42	60	0.007	4650
21	20	40	6	0.42	60	0.007	4650
21	22	41	6	0.355	60	0.0059	4650
22	21	41	6	0.355	60	0.0059	4650
22	24	42	6	0.361	60	0.006	4650
24	22	42	6	0.361	60	0.006	4650
23	24	43	6	0.278	40	0.007	4400
24	23	43	6	0.278	40	0.007	4400
24	25	44	6	0.387	40	0.0097	4400
25	24	44	6	0.387	40	0.0097	4400
25	26	45	6	0.473	50	0.0095	4500
26	25	45	6	0.473	50	0.0095	4500
28	29	46	6	0.264	60	0.0044	4650
29	28	46	6	0.264	60	0.0044	4650
29	30	47	8	0.419	60	0.007	6200
27	30	48	6	0.344	60	0.0057	4650
30	27	48	6	0.344	60	0.0057	4650
30	31	49	8	0.444	60	0.0074	6200
26	31	50	6	0.351	60	0.0059	4650
31	26	50	6	0.351	60	0.0059	4650
25	31	51	6	0.588	50	0.0118	4500
31	25	51	6	0.588	50	0.0118	4500
25	32	52	6	0.351	60	0.0059	4650
32	25	52	6	0.351	60	0.0059	4650
31	32	53	8	0.36	60	0.006	6200
32	33	54	8	0.38	60	0.0063	6200
24	33	55	6	0.463	60	0.0077	4650
33	24	55	6	0.463	60	0.0077	4650
33	34	56	8	0.291	60	0.0049	6200
22	34	57	6	0.329	60	0.0055	4650
34	22	57	6	0.329	60	0.0055	4650
34	35	58	8	0.368	60	0.0061	6200
35	21	59	6	0.283	60	0.0047	4650

According to the calculation results of variable parameters, the probability calculation model of travel time selection under the sudden road fuel leakage and major traffic accidents can be determined, and the expression is as follows:

$$\ln\left(p_{ij}\right) = \sum_{k=1}^{6} \theta_{kj} X_{ikj}, \quad j = 1, 2, 3, 4, 5, \tag{24}$$

where $\sum_{j=1}^{5} P_{ij} = 1$, $P_{i1} \sim P_{i5}$ represents the probability of cancelling, delaying, and travelling by private car, bus, and taxi for type i travelers, and θ_{kj} represents the estimated parameters corresponding to different characteristic variables under travel choice.

4.4. Calculation of OD Traffic Generation in Residential Area. According to the travel time and travel mode selection probabilities obtained above, the traffic generation of a certain type of traffic mode in a certain period can be determined in the second district under an emergency, based on the personal classification method in the calculation of total traffic generation, as follows:

Table 3: Selection of model characteristic variables.

Utility	Selection of characteristic variables					Traveler characteristic variables				
	Intrinsic dumb element				Common variable					
	X_{in1}	X_{in2}	X_{in3}	X_{in4}	Travel time (X_{in5})	Gender (X_{in6})	Age (X_{in7})	Occupation (X_{in8})	Vehicle ownership (X_{in9})	Travel characteristics (X_{in10})
Cancel (V_{1n})	1	0	0	0	X_{1n5}	X_{1n6}	X_{1n7}	X_{1n8}	X_{1n9}	X_{1n10}
Delay (V_{2n})	0	1	0	0	X_{2n5}	X_{2n6}	X_{2n7}	X_{2n8}	X_{2n9}	X_{2n10}
Private car (V_{3n})	0	0	1	0	X_{3n5}	X_{3n6}	X_{3n7}	X_{3n8}	X_{3n9}	X_{3n10}
Bus (V_{4n})	0	0	0	1	X_{4n5}	0	0	0	0	X_{4n10}
Taxi (V_{5n})	0	0	0	0	X_{5n5}	0	0	0	0	X_{5n10}
Pending parameters	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}

Table 4: Estimation results of characteristic variable parameters of the travel time choice model.

	Cancel travel		Delay travel	
	Parameter	Wald	Parameter	Wald
Travel time	−0.17	4.83	−0.21	3.96
Gender	0.65	3.81	0.42	2.01
Age	0.53	2.01	0.25	1.86
Occupation	−0.08	3.82	−0.17	3.92
Vehicle ownership	0.34	5.45	2.81	5.33
Travel characteristics	3.11	2.45	2.34	4.09

Table 5: Estimation results of characteristic variable parameters of the travel mode selection model.

	Private car		Taxi	
	Parameter	Wald	Parameter	Wald
Travel time	0.21	4.53	0.19	0.08
Gender	−0.06	1.87	−0.72	3.85
Age	0.22	2.11	−0.64	4.76
Occupation	0.37	3.93	0.07	2.11
Vehicle ownership	0.42	5.95	3.21	0.32
Travel characteristics	0.32	4.28	3.45	0.42

Table 6: Model likelihood ratio test.

	X^2	d_f	Significant level (sig.)	ρ^2
Travel time selection	89.154	29	0.000	0.351
Choice of travel mode	81.326	28	0.000	0.296

$$q_{rn} = N_r \sum_i a_{ir} S_i p_{in}, \quad i \in I, \tag{25}$$

where q_{rn} was the traffic generating capacity of all types of road users in community r choosing travel mode n, N_r was the total population of all types of road users in community r, a_{ir} was the proportion of class i road users in community r, S_i was the average number of trips per person of class i road users in a certain period of time, and p_{in} was the probability of class i road users choosing travel mode.

4.5. Traffic Vulnerability Identification of Road Network. According to the design of the vulnerability comprehensive identification index, the traffic VURN under emergency situations could be identified by the comprehensive measurement index of the traffic VURN, and the comprehensive measurement index was obtained by a weighted synthesis of three types of indexes, which were road network connectivity and its change index under emergency conditions and road network vulnerability under the same, network efficiency and its change index, and network traffic

Table 7: Traveler type and nonaggregate parameter calibration.

Travel choice	Proportion (u_d^{rs})	Parameter (θ)	Constant (θ_0)	Selection probability (P_{in})
Cancel	0.245	1.15	0.14	0.72
Suspend	0.376	1.52	0.09	0.88
Private car	0.151	0.94	0.12	0.61
Bus	0.211	1.623	0.23	0.89
Taxi	0.017	0.264	0.04	0.45

performance and its change index in an emergency. Therefore, it was necessary to calculate three types of indicators and subsequently synthesize these to obtain the final vulnerability identification results.

4.6. Road Network Connectivity Changes during Emergency. We constructed the node adjacency matrix of the road network in the Nanshan Science and Technology Park area of Shenzhen City, analyzed the topological characteristics before and after the emergency using the connectivity or network efficiency index, and calculated the changes of the connectivity or network efficiency index of the point or edge before and after the emergency according to equations (1) and (2). Prior to emergency, there were 35 nodes and 59 road sections in the road network, and the connectivity $G = 0.596$. After emergency, after each road node was interrupted one by one, the connectivity G' of the road network after the accident was calculated, and the statistical results are shown in Figure 5. Figure 5 shows the changes of network connectivity of road network nodes before and after the emergency. The nodes with $\Delta G > 0.03$ were nodes 9, 11, 24, 25, and 26. Relatively speaking, these nodes were vulnerable and required supervision. Table 8 shows the corresponding road sections.

4.7. Road Network Efficiency Changes during Emergency. The change of road network efficiency under emergency could also reflect the vulnerability of the road network from one side. According to the previous assumption, we considered that each road section in Figure 4 had a major accident, which led to the interruption of the road section and the traffic efficiency changed. Then, we calculated the traffic efficiency E after the accident according to equations (3) and (4). To calculate road network traffic efficiency, it was necessary to obtain the shortest path between OD pairs. Here, the Dijkstra algorithm was used for this purpose.

In Figure 6, the calculation shows that the network efficiency of the topology was 0.6742 prior to emergency. After the nodes were interrupted, the efficiency of the road network was E'. By comparing the efficiency before and after the emergency, the change ΔE was obtained, and Figure 6 shows the calculation results statistically.

Figure 6 shows a three-dimensional statistical chart composed of road sections (arcs), traffic efficiency vulnerability index (E') after the accident, and traffic efficiency vulnerability index (ΔE) before and after the accident. The list on the right side shows some statistical ranking results of link vulnerability. Figure 6 shows that section 42 had the greatest impact on traffic efficiency on the whole road network, played an important role, and was the most vulnerable section. Except for section 42, sections 37, 30, 10, and 16 with the change of road network efficiency more than 0.015 were arranged according to the influence of traffic efficiency from large to small. These sections played an important role in the stability of the road network and required high attention in daily maintenance.

4.8. Road Network Traffic Performance Changes during Emergency. The investigation showed that emergency could easily lead to urban road damage and traffic interruption. Among them, road damage would lead to changes in the structure of the road network, which would affect the connectivity and traffic efficiency. In addition, the accident would impact the traffic operation of the road unit, such that the road transportation function would be affected. In this case, for travelers, according to the amount of information they receive, they would choose to cancel, suspend, or travel through other modes of transportation, and the on-the-way traffic would find alternative routes, which would lead to the redistribution of urban road network traffic. For a certain period, the corresponding road sections would frequently develop long queues. Figure 2 shows the overall operation characteristics of network traffic and the distribution characteristics of section traffic flow. To analyze the vulnerability based on the traffic flow operation model, the following assumptions for the construction of the model were made:

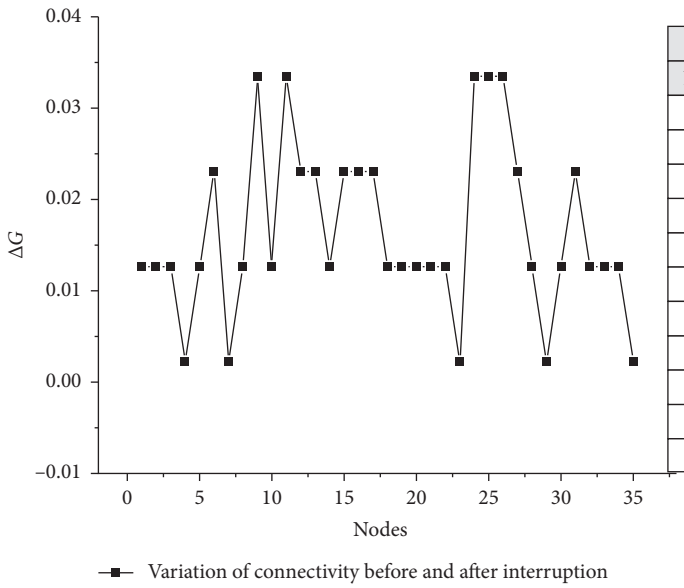
(1) Suppose the queue was at the vehicle point, and the length of each vehicle is 0

(2) The queuing capacity of each road section in the network was limited, which could not hold an infinite number of vehicles in line

(3) The section capacity would rapidly decrease to 0 under the condition of sudden congestion

(4) If the same lane ran in a single row, and there was no multivehicle parallel situation, the traffic capacity of the congested road section would be approximately equal to the maximum queuing capacity when congestion ceased

According to the road network distribution in Figure 4 and the survey data in Table 2, three typical OD pairs were selected to load. Using MATLAB, the Dijkstra algorithm was used to calculate the corresponding shortest path of each OD pair under normal conditions. Then, according to the set emergency, the above algorithm was used to interrupt the road operation after each incident, and the traffic



Partial data	
Nodes	ΔG
9	0.0334
11	0.0334
24	0.0334
25	0.0334
26	0.0334
6	0.0230
12	0.0230
13	0.0230
15	0.0230
16	0.0230
–	–

Figure 5: Changes of the connectivity of network before and after emergency.

distribution results after traffic allocation, after emergency, were obtained. This involved equations (5) through (7) and the algorithm designed earlier in this paper. In the calculation process, parameters α and β of the BPR function were 0.15 and 4, respectively. Finally, the change values of all traffic performance parameters were obtained, Table 8 shows the statistics.

Table 8 shows that after emergency, the interruption caused traffic demand change, which led to the redistribution of road traffic to a new equilibrium state, resulting in travel time loss. Among them, the total travel loss time of sections 23, 5, 25, 12, and 44 after interruption was more than 31.2 min. This was the key section of the whole road network, which in turn required significant attention from the relevant department. Figure 7 shows the convergence of the vulnerability comprehensive identification model and the algorithm. The iterative gap value decreased rapidly with the increase of the iteration time. After 10 iterations, the satisfactory equilibrium numerical solution was achieved, which demonstrated that the model solution algorithm proposed was highly efficient.

5. Analysis

In the previous part of network vulnerability identification, we analyze the topological connectivity, traffic efficiency, and performance of the network during emergency. The comprehensive identification of traffic VURN was analyzed from two perspectives. First, the traffic vulnerability of the road network was identified by considering both topological connectivity and network efficiency. Second, the traffic vulnerability was identified by considering topological connectivity, network efficiency, and traffic performance. On this basis, the results of road network vulnerability identification from two different perspectives were compared and analyzed.

5.1. Analysis of Vulnerability Identification Results considering Network Connectivity and Efficiency. The results calculated in the previous section were counted in descending order, including the change value of road network connectivity (ΔG), and that of road network efficiency (ΔE) and the corresponding nodes and sections. Table 9 only lists the first 10 rows of the results due to spacing issues. From these results, we observed that the top 10 nodes were nodes 9, 11, 24, 25, 26, 6, 12, 13, 15, and 16, which indicated that the vulnerability of these nodes had changed from strong to weak. According to the correlation between nodes and connected edges, when these nodes were impacted by emergency, their corresponding connected edges were also affected. Therefore, the corresponding connected edges of these nodes would also show the corresponding vulnerability. Based on this, the connected edges of these nodes were statistically analyzed.

The connection edges corresponding to the above nodes had a strong correlation with the connection edges after the statistics of the change value of network efficiency (ΔE). In other words, the nodes corresponding to the connection edges obtained by sorting the change value of network efficiency (ΔE) from large to small were highly consistent with the nodes arranged by the change value of the network connectivity index (ΔG) from large to small. From the top 10 nodes listed, there were seven nodes in total, including 9, 11, 24, 25, 6, 13, and 15, which coincided with the nodes corresponding to the network efficiency connecting edge. This indicated that prior to and after emergency, the change value of network connectivity (ΔG) and that of network efficiency

Table 8: Calculation results of traffic performance change value of road network under emergency.

Arcs	ΔL (h)
1	0.499829543
2	0.517159174
3	0.506221114
4	0.511633593
5	0.526580431
6	0.488086389
7	0.479349968
8	0.513842596
9	0.498912608
10	0.50785149
11	0.508517681
12	0.52126487
13	0.496948847
14	0.491859617
15	0.497418344
16	0.486644673
17	0.483751788
18	0.479275825
19	0.497191832
20	0.50924756
21	0.5004505
22	0.4893875
23	0.5279817
24	0.5117103
25	0.5251609
26	0.5162785
27	0.4788009
28	0.4833101
29	0.4913568
30	0.5049395
31	0.486535
32	0.471266
33	0.498022
34	0.505899
35	0.511477
36	0.469642
37	0.476885
38	0.487584
39	0.462275
40	0.488921
41	0.505747
42	0.512273
43	0.510029
44	0.521001
45	0.490435
46	0.506765
47	0.490421
48	0.489485
49	0.480163
50	0.495179
51	0.503999
52	0.496315
53	0.500324
54	0.467583
55	0.516451
56	0.499224
57	0.50053
58	0.473564
59	0.487039

(ΔE) caused by these nodes were large, indicating that these nodes and related road sections had high vulnerability.

5.2. Analysis of Vulnerability Identification Results considering Connectivity, Network Efficiency, and Traffic Performance. The vulnerability of the road network under the change of road traffic demand was identified, Table 8 shows the results. However, in the actual operation of the urban road network, when an emergency occurs, it would not only lead to a change of road traffic demand, but also to that of the road network structure. Therefore, it was necessary to comprehensively consider the changes of road network connectivity and efficiency as well as the traffic performance when analyzing the traffic VURN under emergency conditions. Based on this, the vulnerability of the road network was comprehensively identified.

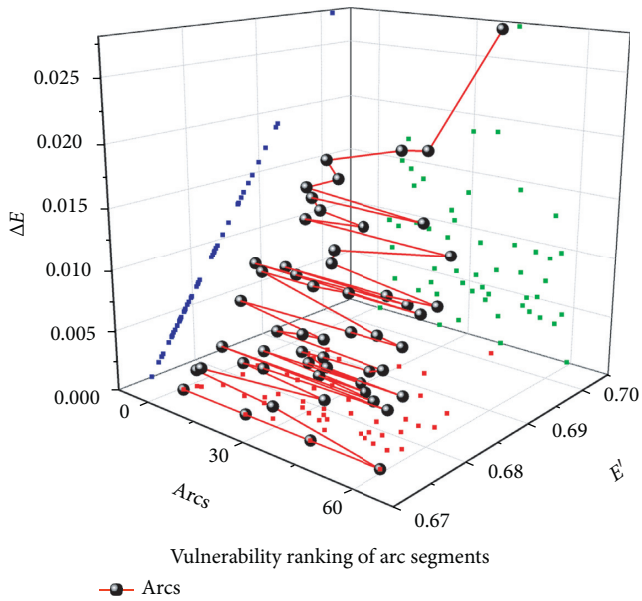
In Section 3.1, "Index Design," the vulnerability comprehensive identification method based on road network connectivity and efficiency as well as traffic performance were discussed, and the vulnerability comprehensive identification index including road network connectivity, efficiency, and traffic performance was designed (v) (see equation (7)). In this equation, three model parameters were involved, namely, ε, $\varnothing$, and γ, and the other three model variables are the road network connectivity change value ($\Delta G'$), network efficiency change value ($\Delta E'$), and traffic performance change value ($\Delta L'$) after data standardization. Among them, the model parameters required specific calibration. Here, the Delphi method was used for this purpose. Through the investigation of 15 transportation experts, parameter calibration results were obtained, and the parameters ε, $\varnothing$, and γ were 0.42, 0.23. and 0.35, respectively. By substituting them into equation (7), the following results were obtained:

$$v = 0.42\Delta G' + 0.23\Delta E' + 0.35\Delta L'. \tag{26}$$

Since the change value of network efficiency ($\Delta e'$) and the change value of traffic performance ($\Delta L'$) were based on the road section, to unify into equation (26), it was necessary to convert the change value of road network connectivity ($\Delta G'$) based on the node to the road section and subsequently calculate on this basis to obtain the results shown in Table 10.

According to the scale range of the vulnerability level of key nodes or sections of the road network set in Table 1, the corresponding vulnerability level was given in the last column as shown in Table 10. Among them, the vulnerability level of road section 12 was "high," sections 23, 37, 11, and 24 was "medium," and the remaining sections were below "low." Figure 8 shows the statistical chart of some data, and Figure 9 shows the distribution of the vulnerability level in the road network.

Compared with Tables 9 and 10, the former only considered the changes of network connectivity and network efficiency under emergency conditions. This distribution only reflected the vulnerability of a static road network,



Arcs	E'	ΔE
42	0.701	0.027
37	0.693	0.018
30	0.692	0.018
10	0.691	0.016
16	0.69	0.015
8	0.689	0.014
11	0.688	0.014
44	0.688	0.014
15	0.687	0.013
29	0.687	0.013
12	0.686	0.012
54	0.686	0.012
24	0.685	0.011
25	0.684	0.010
—	—	—

Figure 6: Road network vulnerability identification based on efficiency.

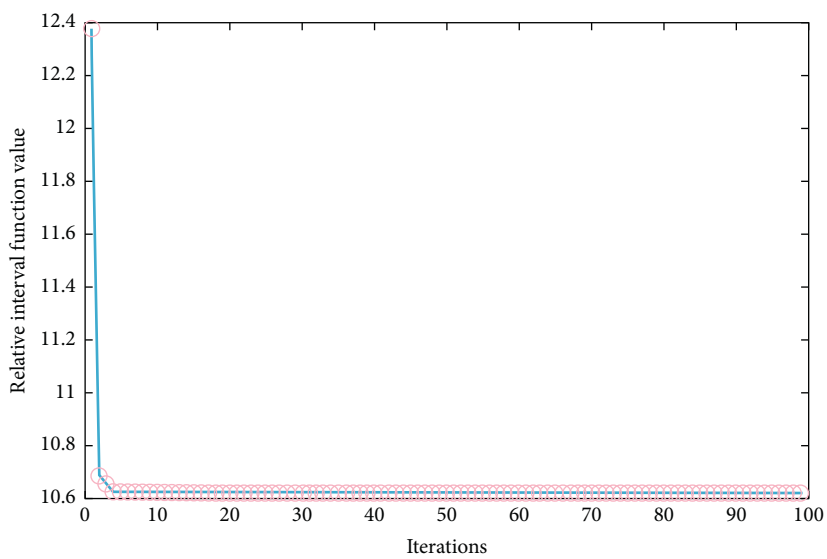


Figure 7: Convergence of the algorithm.

Table 9: Vulnerability identification results based on network connectivity and network efficiency (partial data).

Network connectivity		Network efficiency			
Nodes	Connectivity change (ΔG)	Nodes	Nodes	Road section	Network efficiency change (ΔE)
9	0.033459596	22	24	42	0.0277
11	0.033459596	13	25	37	0.0188
24	0.033459596	6	11	30	0.0185
25	0.033459596	15	18	10	0.0169
26	0.033459596	9	10	16	0.0158
6	0.023042929	1	15	8	0.0148
12	0.023042929	14	15	11	0.0142
13	0.023042929	24	25	44	0.0142
15	0.023042929	6	7	15	0.0135
16	0.023042929	5	6	29	0.0131

Table 10: Comprehensive identification results of road network vulnerability based on connectivity, network efficiency, and traffic performance.

Arcs	$\Delta G'$	$\Delta E'$	$\Delta L'$	v	Vulnerability level
12	1	0.458484	0.897775	0.839672622	4
23	0.666656	0.310469	1	0.701403462	3
37	1	0.6787	0.222352	0.653924174	3
11	0.666656	0.512635	0.703774	0.64422256	3
24	0.666656	0.404332	0.752363	0.63631881	3
38	1	0.238267	0.385181	0.609614902	2
15	0.666656	0.487365	0.534852	0.579287444	2
33	0.666656	0.31769	0.544039	0.543477668	2
5	0.333312	0.231047	0.978674	0.535667593	2
10	0.333312	0.610108	0.693635	0.523088253	2
27	1	0.036101	0.25151	0.516331787	2
36	1	0.151625	0.112119	0.494115447	1
17	0.666656	0.32491	0.326858	0.469125175	1
49	0.666656	0.32852	0.27224	0.450839102	1
2	0.333312	0.050542	0.83529	0.443967135	1
53	0.333312	0.31769	0.579073	0.415735252	1
56	0.333312	0.34296	0.562332	0.415688156	1
18	0.666656	0.140794	0.258738	0.402936491	1
13	0.333312	0.33574	0.527706	0.401908448	1
16	0.333312	0.570397	0.370886	0.40099232	1
3	0.333312	0.057762	0.668822	0.387364017	1
1	0.333312	0	0.571548	0.340032836	1
4	0	0.32491	0.751196	0.337647674	1
48	0.333312	0.129964	0.414113	0.314822252	1
31	0.333312	0.148014	0.369216	0.303260111	1
40	0.333312	0	0.405529	0.281926296	1
54	0.333312	0.454874	0.080783	0.272886104	1
46	0	0.144404	0.6771	0.270197893	1
14	0	0.101083	0.450253	0.180837486	1
39	0	0.086643	0	0.019927798	1

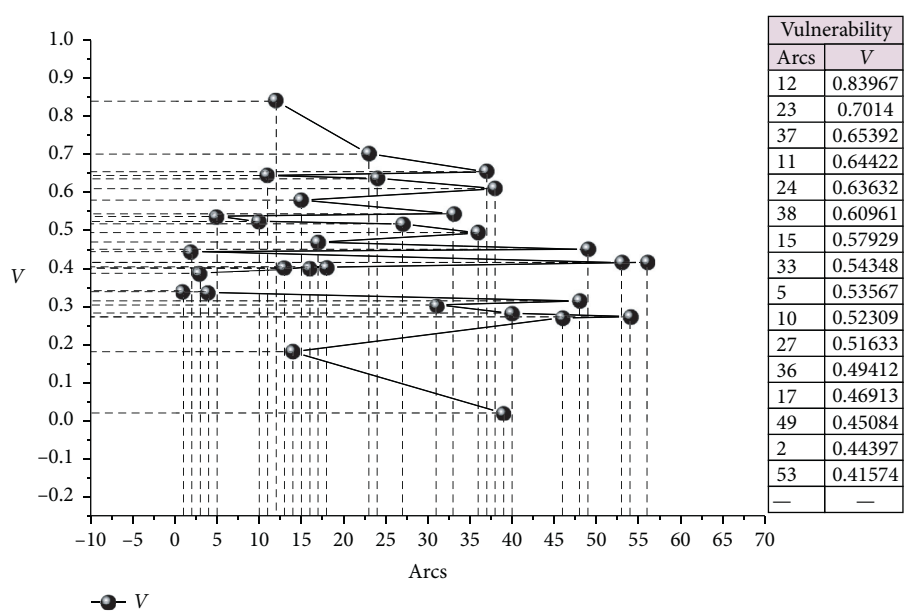


Figure 8: Vulnerability statistics after comprehensive identification.

including the connectivity of road sections in the network topology and the change of network efficiency characterized by the shortest path distribution. The latter did not consider the changes of static structure to the network under emergency conditions, nor did it consider the changes of the same in the network due to the accidents that change traffic

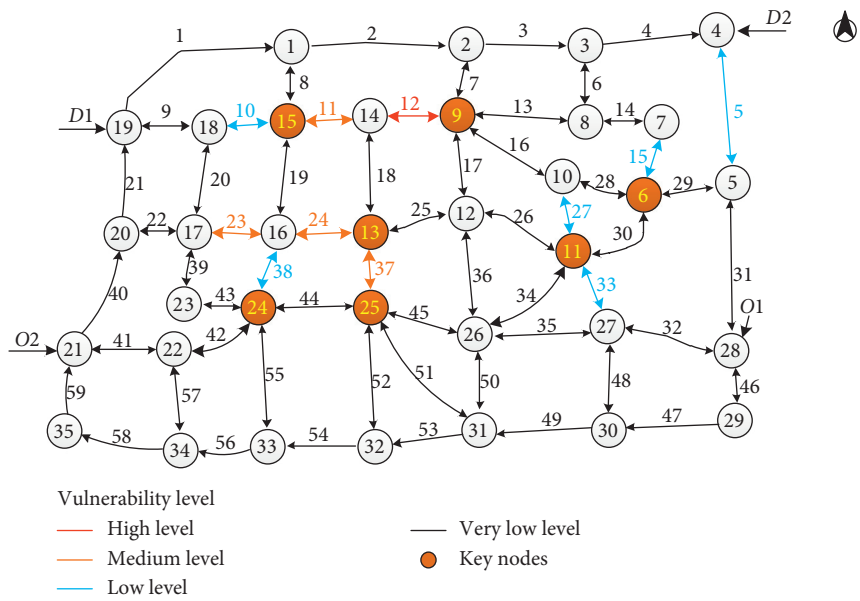


Figure 9: Vulnerability level display.

performance (change to traffic impedance of the road network) caused by the change in network traffic demand (change of road users' travel choice) was closer to the real road traffic situation.

Due different considerations, there was variation in the results of different discriminant indexes. From the identification results, only considering the changes of network connectivity and efficiency under emergency conditions, it was concluded that the vulnerability of nodes 9, 11, 24, 25, 6, and 13 and that of road sections 7, 12, 37, 24, 16, and 17 were relatively large. If the changes of road network connectivity and efficiency as well as the traffic performance were considered simultaneously, it was concluded that the vulnerability of sections 12, 23, 37, 11, 24, 38, and 15 was relatively high. Compared with the others, road sections 12 and 37 were considered two sections with greater vulnerability. Therefore, the two indicators had certain directivity. From another point of view, they also showed that sections 12 and 37 should be considered two sections with outstanding vulnerability. In the daily maintenance and after the occurrence of the emergency, these sections and nodes were the key links that departments needed to focus on. The comprehensive identification of road network traffic vulnerability based on a variety of discriminant indicators could represent the vulnerability from different angles, and various discriminant indicators could complement each other.

In the identification of urban road network traffic vulnerability, considering the network structure and traffic operation conditions under emergency conditions, the conclusion is closer to that of the actual road traffic situation. However, this comprehensive identification method based on multiple evaluation indexes was also limited by the model accuracy and operation efficiency. At present, room for improvement remains in the application of the proposed solution based on multiple discriminant indicators proposed in this study in terms of solving the large-scale vulnerability identification and evaluation.

6. Conclusions

The accurate, effective, and reliable identification of traffic VURNs can provide scientific guidance for road traffic planning, design, construction, operation, management, evaluation, and emergency rescue. Although the academic community has contributed much research into this, there is no unified and definite definition for the concept of road traffic vulnerability. Because of different interpretations regarding this, the research ideas proposed are dissimilar. Based on summarizing the existing research results, we proposed three core identification indexes from the perspectives of static road network structure and dynamic traffic performance, and a comprehensive identification index of urban road network traffic vulnerability based on this. Hence, the vulnerability identification model of the urban road traffic performance index and its solution algorithm were studied. Finally, to verify the designed vulnerability identification method, taking the road network of Nanshan Science and Technology Park area in Shenzhen City as an example, the single vulnerability identification index and comprehensive vulnerability identification index of an urban road network were verified, respectively. The results demonstrated that the proposed vulnerability comprehensive evaluation index had greater advantages than a single type of vulnerability evaluation index, because there was better complementarity between them.

The results of this study could be applied in the following contexts. First, it could be used to guide road management departments in recognizing the vulnerable links of the traffic system, assist in determining the priority of road maintenance and repair, and reduce the vulnerability of important traffic infrastructure. Second, it is conducive to the need to optimize the operation and management of daily traffic networks and reduce traffic congestion and time delay. Third, it is beneficial to evaluate the potential consequences

and overall effects of various control and management strategies and enhance the ability to prevent and respond to disasters and emergency situations. Fourth, it is beneficial to evaluate the connectivity vulnerability of various road network planning schemes and enhance the prevention and response capabilities through traffic network planning.

Because research on the vulnerability of urban road networks under emergency conditions involves too many uncertain factors, including engineering, technical, economic, social, and other aspects of urban development, it was difficult to provide a definitive solution. However, future work will concentrate on the vulnerability of key coupling networks in an urban multimode traffic network, with a focus on improving the practical application of the research.

Data Availability

The raw data supporting the research of this article will be made available to any qualified researcher by the authors.

Conflicts of Interest

The author declares no conflicts of interest.

Authors' Contributions

H. Xiang conceived the research project and proposed the algorithm and conducted the experiments.

Acknowledgments

This article was partially supported by the Science and Technology Projects of Shenzhen Science and Technology Innovation Committee (no. JSGG20170822093602485) and Shenzhen Fundamental Research Project (no. JCYJ20180305163701198). The author would like to thank Editage (https://www.editage.cn) for English language editing.

References

[1] A. D'Andrea, S. Cafiso, and A. Condorelli, "Methodological considerations for the evaluation of seismic risk on road network," *Pure and Applied Geophysics*, vol. 162, no. 4, pp. 767–782, 2005.

[2] B. Y. Chen, W. H. K. Lam, A. Sumalee, Q. Li, and Z.-C. Li, "Vulnerability analysis for large-scale and congested road networks with demand uncertainty," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 3, pp. 501–516, 2012.

[3] C.-H. Hsieh and C.-M. Feng, "The highway resilience and vulnerability in Taiwan," *Transport Policy*, vol. 87, pp. 1–9, 2020.

[4] I. Cambridge Systematics, *An Initial Assessment of Freight Bottlenecks on Highways*, Federal Highway Administration, U.S. transportation of Department, Washington, DC, USA, 2005, https://www.fhwa.dot.gov/policy/otps/bottlenecks/.

[5] A. Nagurney, Q. Qiang, and L. S. Nagurney, "Environmental impact assessment of transportation networks with degradable links in an era of climate change," *International Journal of Sustainable Transportation*, vol. 4, no. 3, pp. 154–171, 2010.

[6] K. Berdica, "An introduction to road vulnerability: what has been done, is done and should be done," *Transport Policy*, vol. 9, no. 2, pp. 117–127, 2002.

[7] F. Rupi, S. Bernardi, G. Rossi, and A. Danesi, "The evaluation of road network vulnerability in mountainous areas: a case study," *Networks and Spatial Economics*, vol. 15, no. 2, pp. 397–411, 2015.

[8] L. Gao, X. Liu, Y. Liu et al., "Measuring road network topology vulnerability by Ricci curvature," *Physica A: Statistical Mechanics and Its Applications*, vol. 527, Article ID 121071, 2019.

[9] E. Jenelius, T. Petersen, and L.-G. Mattsson, "Importance and exposure in road network vulnerability analysis," *Transportation Research Part A*, vol. 40, no. 7, 2005.

[10] Z. Zheng, Z. Huang, F. Zhang, and P. Wang, "Understanding coupling dynamics of public transportation networks," *EPJ Data Science*, vol. 7, no. 1, 2018.

[11] M. Weber, E. Saucan, and J. Jost, "Characterizing complex networks with forman-ricci curvature and associated geometric flows," *Journal of Complex Networks*, vol. 5, no. 4, pp. 527–550, 2017.

[12] M. Li, H. Wei, Y. Li, and S. Liu, "Identifying influential nodes in complex networks based on local and global methods," *Journal of Physics: Conference Series*, vol. 1738, no. 1, 2021.

[13] Y. Duan and F. Lu, "Robustness of city road networks at different granularities," *Physica A: Statistical Mechanics and Its Applications*, vol. 411, pp. 21–34, 2014.

[14] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: the role of clustering," *PLoS One*, vol. 8, no. 10, Article ID e77455, 2013.

[15] H.-Y. Yin and X. U. Li-Qun, "A model for identifying vulnerable links of road networks based on bayesian networks," *Journal of Systems& Management*, vol. 19, no. 6, pp. 656–661, 2010, in Chinese.

[16] E. L. D. Oliveira, L. D. S. Portugal, and W. P. Junior, "Determining critical links in a road network: vulnerability and congestion indicators," *Procedia-Social and Behavioral Sciences*, vol. 162, pp. 158–167, 2014.

[17] S. Jung, S. Lee, O. Kwon, and B. Kim, "Grid-based traffic vulnerability analysis by using betweenness centrality," *Journal of the Korean Physical Society*, vol. 77, no. 7, pp. 538–544, 2020.

[18] R. Milo, S. Itzkovitz, N. Kashtan et al., "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, 2004.

[19] M. Snelder, H. J. Van Zuylen, and L. H. Immers, "A framework for robustness analysis of road networks for short term variations in supply," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 5, pp. 828–842, 2012.

[20] D. M. Scott, D. C. Novak, L. Aultman-Hall, and F. Guo, "Network robustness index: a new method for identifying critical links and evaluating the performance of transportation networks," *Journal of Transport Geography*, vol. 14, no. 3, pp. 215–227, 2006.

[21] Z.-P. Du and A. Nicholson, "Degradable transportation systems: sensitivity and reliability analysis," *Transportation Research Part B: Methodological*, vol. 31, no. 3, pp. 225–237, 1997.

[22] A. Chen, H. Yang, H. K. Lo, and W. H. Tang, "Capacity reliability of a road network: an assessment methodology and numerical results," *Transportation Research Part B: Methodological*, vol. 36, no. 3, pp. 225–252, 2002.

[23] E. Jenelius, T. Petersen, and L.-G. Mattsson, "Importance and exposure in road network vulnerability analysis,"

Transportation Research Part A: Policy and Practice, vol. 40, no. 7, pp. 537–560, 2006.

[24] J. Sohn, "Evaluating the significance of highway network links under the flood damage: an accessibility approach," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 6, pp. 491–506, 2006.

[25] M. A. P. Taylor, S. V. C. Sekhar, and G. M. D'Este, "Application of accessibility based methods for vulnerability analysis of strategic road networks," *Networks and Spatial Economics*, vol. 6, no. 3-4, pp. 267–291, 2006.

[26] K. Berdica and L. G. Mattsson, "Vulnerability: a modelasedcase study of the road network in stockholm," *Critical Infrastructure: Reliability and Vulnerability*, A. T. Burray and T. H. Grubesic, Eds., Springer, Berlin, Germany, 2007.

[27] A. Chen, C. Yang, S. Kongsomsaksakul, and M. Lee, "Network-based accessibility measures for vulnerability analysis of degradable transportation networks," *Networks and Spatial Economics*, vol. 7, no. 3, pp. 241–256, 2007.

[28] L.-G. Mattsson and E. Jenelius, "Vulnerability and resilience of transport systems - a discussion of recent research," *Transportation Research Part A: Policy and Practice*, vol. 81, pp. 16–34, 2015.

[29] A. J. Holmgren, "A framework for vulnerability assessment of electric power systems," *Reliability and Vulnerability in Critical Infrastructure: A Quantitative Geographic Perspective*, A. Murray and T. Grubesic, Eds., Springer, Berlin, Germany, 2007.

[30] J. Husdal, "The vulnerability of road networks in a cost-benefit perspective," 2005.

[31] M. A. P. Taylor and G. M. D'Este, "Transport network vulnerability: a method for diagnosis of critical locations in transport infrastructure systems," *Critical Infrastructure: Reliability and Vulnerability*, A. T. Murray and T. H. Grubesic, Eds., Springer, Berlin, Germany, 2007.

[32] A. Erath, J. Birdsall, K. W. Axhausen et al., "Vulnerability assessment of the swiss road network," in *Proceedings of the 88th Transportation Research Board Annual Meeting*, pp. 1–17, Washington DC, USA, March 2009.

[33] E. Jenelius and L.-G. Mattsson, "Road network vulnerability analysis: conceptualization, implementation and application," *Computers, Environment and Urban Systems*, vol. 49, pp. 136–147, 2015.

[34] Y. Lu-Ping and D.-L. Qian, "Vulnerability analysis of road networks," *Journal of Transportation Systems Engineering and Information Technology*, vol. 12, no. 2, pp. 105–110, 2011, in Chinese.

[35] M. Lujak and S. Giordani, "Centrality measures for evacuation: finding agile evacuation routes," *Future Generation Computer Systems*, vol. 83, pp. 401–412, 2018.

[36] United Nations International Strategy for Disaster Reduction, *Living with Risk: A Global Review of Disaster Reduction Initiatives*, United NationsInternational Strategy for Disaster Reduction, Geneva, Switzerland, 2004.

[37] Y. Liu, "Emergency response facility location in transportation networks: a literature review," *Journal of Traffic and Transportation Engineering*, vol. 3, 2021, English Edition.

[38] R. Z. Farahani, M. M. Lotfi, A. Baghaian et al., "Mass casualty management in disaster scene: a systematic review of OR&MS research in humanitarian operations," *European Journal of Operational Research*, vol. 287, no. 3, Article ID 787e819, 2020.

[39] A. P. Mera and C. Balijepalli, "Towards improving resilience of cities: an optimisation approach to minimising vulnerability to disruption due to natural disasters under budgetary constraints," *Transportation*, vol. 47, no. 4, pp. 1809–1842, 2020.

[40] F. Makoto, "A study of vulnerability of emergency transport road network to various hazards," *Građevinar*, vol. 70, no. 12, pp. 1065–1074, 2018.

[41] H. Zhang and Y. Yao, "An integrative vulnerability evaluation model to urban road complex network," *Wireless Personal Communications*, vol. 107, no. 1, pp. 193–204, 2019.

[42] A. B. Morelli and A. L. Cunha, "Measuring urban road network vulnerability to extreme events: an application for urban floods," *Transportation Research Part D: Transport and Environment*, vol. 93, Article ID 102770, 2021.

[43] J. Liu, Z. Shi, and X. Tan, "Measuring the dynamic evolution of road network vulnerability to floods: a case study of Wuhan, China," *Travel Behaviour and Society*, vol. 23, pp. 13–24, 2021.

[44] M. Ansari Esfeh, A. E. Salari, L. Kattan, W. H. K. Lam, R. Ansari Esfe, and M. Salari, "Compound generalized extreme value distribution for modeling the effects of monthly and seasonal variation on the extreme travel delays for vulnerability analysis of road network," *Transportation Research Part C: Emerging Technologies*, vol. 120, Article ID 102808, 2020.

[45] M. M. De Brito, M. Evers, and B. Höllermann, "Prioritization of flood vulnerability, coping capacity and exposure indicators through the Delphi technique: a case study in Taquari-Antas basin, Brazil," *International Journal of Disaster Risk Reduction*, vol. 24, pp. 119–128, 2017.

[46] L. Jun-Qiang, A. Zhai Jing, L. Qian-Wen, and L. Zhao, "Construction of road network vulnerability evaluation index based on general travel cost," *Physica A*, vol. 493, pp. 421–429, 2018.

[47] N. He and S. Zhao, "Discussion on influencing factors of free-flow travel time in road traffic impedance function," *Procedia - Social and Behavioral Sciences*, vol. 96, pp. 90–97, 2013.

[48] T. P. Van Oijen, W. Daamen, and S. P. Hoogendoorn, "Estimation of a recursive link-based logit model and link flows in a sensor equipped network," *Transportation Research Part B: Methodological*, vol. 140, pp. 262–281, 2020.

Hindawi
Complexity
Volume 2021, Article ID 5592272, 11 pages
https://doi.org/10.1155/2021/5592272

Research Article

An Innovation Design Approach for Product Service Systems Based on TRIZ and Function Incentive

Jie Jiang,[1] Yan Li,[2] Lidan Li,[3] Changchun Zhou,[1] Yuxiang Huo,[1] and Qian Li[1]

[1]*College of Environment and Civil Engineering, Chengdu University of Technology, Chengdu 610059, China*
[2]*College of Mechanical Engineering, Sichuan University, Chengdu 610065, China*
[3]*Sichuan Jiuzhou Electric Appliance Group Co., Ltd., Mianyang 621000, China*

Correspondence should be addressed to Jie Jiang; 110991785@qq.com

Received 25 January 2021; Revised 22 February 2021; Accepted 2 March 2021; Published 20 March 2021

Academic Editor: Long Wang

Good balance between product and service is the key in the innovative design of product service systems (PSS). In this study, the evolution route of the PSS based on Teoriya Resheniya Izobretatelskikh Zadatch ideal final result was provided. The function model of the PSS was constructed according to the service blueprint and function system diagrams. On this basis, an innovation design method of the PSS based on function incentive was established. The function incentive strategies included function synergy, function supplement, and function substitution. Finally, the PSS design process of agricultural machinery based on computer-aided innovation platform was analyzed to verify this method.

1. Introduction

Product service systems (PSS) originated in the 1990s. The primary goal of PSS is to help manufacturing enterprises focus on the integration of physical products and behavioral services to meet customers' needs instead of only producing and selling physical products [1]. It can satisfy customers' needs more effectively with less environmental damage than traditional production and consumption models. With dwindling natural resources and increasingly serious environmental pollution, many countries and enterprises have realized the importance of reducing the usage of resources and improving environmental protection. Hence, their ultimate goal is to design and manufacture products that can maximize energy and reduce emissions and, at the same time, meet customers' needs. For this reason, PSS have gradually received attention and have been under development in various countries since the 21st century.

The introduction of service into a product created new development opportunities for manufacturing enterprises and improved the diversity and competitiveness of design solutions. With the continued integration of services and products, determining how to provide customers with diverse and practical solutions has become a challenge. Owing to the unique characteristics of services—intangibility, timeliness, production, and consumption at the same time [2]—the design process of PSS is more complex and innovative than a single product or a single-service design. By integrating the design of a product and a service, a growing number of PSS design solutions have been generated to meet customers' diverse needs. PSS design should begin with the design of the function [3]. Clarifying the functional interaction between product and service is the essential stage in PSS design.

With the growing amount of research on PSS, mainstream studies were mainly divided into three aspects. (1) The product-oriented design methodology involves the integration and coordination of a service and a product in the whole life cycle of an existing product to obtain the PSS solution [4, 5]. (2) The service-oriented design methodology explores the needs of various stakeholders in the PSS, a PSS design framework was constructed, and various contacts in the PSS framework were designed [6–8]. (3) Parallel product-and-service design methodology is the equal interaction between a product and a service in PSS design to meet customers' needs [9, 10]. However, from the practical

perspective, further studies are needed to address two issues. (1) Avoiding excessive pursuit of service maximization reduces resource consumption but ignores current production and service capacity of manufacturers and service providers, respectively. (2) The design process should be shortened, and the evolution capability should be improved. Therefore, PSS needs to evolve not only in response to changes in market demands but also before potential market demands are discovered. Through the change and reconstruction of PSS, a new design method can be formed. Therefore, the research direction is to use existing product innovation design methods and theories (Teoriya Resheniya Izobretatelskikh Zadatch (TRIZ), Quality Function Deployment (QFD), Design Structure Matrix, etc.) to carry out functional design to the internal system of the PSS and to find the best balance state between product and service.

Recent studies investigated the promising methodology called TRIZ, which is a Russian acronym for "Theory of Inventive Problem Solving." TRIZ is as a set of analytical tools that detect contradictions in the problem-solving process by eliminating or attenuating conflicts [11] and generating innovation [12]. As a structured innovation design method, TRIZ has been applied in PSS design. Lee et al. proposed a generic and systematic service design method with a customer-centric and adductive reasoning logic based on TRIZ; the design ontology was constructed on the basis of 40 principles of TRIZ [13]. Wang and Zhang proposed the product emotional intention to elaborate the invention principle of the product surface structure and functional technology based on TRIZ, and the explanation of the improvement was put forward on the basis of further contradiction analysis [14]. Wang et al. used TRIZ to eliminate the contradiction among these service design requirements and the TRIZ innovation principles as the service resolution in the problem resolution phase [15]. Song and Sakao proposed a design framework with a design process by using different methods in which TRIZ was employed to solve design conflicts [16]. Yang and Xing proposed an eco-innovative method by using TRIZ to design an eco-leasing PSS [17]. Chen and Liu proposed a substance-field model grounded on results from system layer analysis maps, aiming to obtain innovative and improved low-carbon PSS ideas [18]. The analysis of the literature suggested that TRIZ could be effectively applied to PSS design.

QFD method and its adaptation to PSS development (QFD for PSS) is a customer-demand-oriented product design planning method, which was widely considered the optimization tool for product design planning. Haber and Fargnoli analyzed customers' demands by using the Kano model and transformed it into receiver state parameters [19]. Fuzzy analytic hierarchy process was used to evaluate the parameters and inherent uncertainties. Haber et al. proposed a method based on the synergy of QFD for PSS, Axiomatic Design (AD), and Service Blueprint [20]. By using the method, the complexity of meeting requirements in PSS design was reduced, the potential of PSS to meet market demands was maintained, and the risk of overdesign and design conflicts was reduced. Haber and Fargnoli proposed a QFD for PSS method and network analysis process method based on market demand and customer demand analysis to evaluate the interaction between product and service elements to elect customer needs and expectations [21].

Design structure matrix (DSM) is a common method to model systems in matrix form in product innovation design. Sakao et al. aimed at custom PSS, in which the interaction between service elements was considered and reflected in the DSM, to obtain service modules [22]. Lee and Abuali proposed an operable innovative design method based on creative tools such as design structure matrix, spatial mapping, and QFD [23]. Son et al. proposed a technology evolution method based on DSM to facilitate collaborative design among multiple stakeholders [24].

In this study, a PSS evolution curve was established by adopting TRIZ, and an interactive function model of the product and service was proposed using the PSS blueprint. The function incentive strategies were provided to achieve a PSS conceptual design. Through the selection and evaluation of conceptual design solutions, the PSS innovation design was finally generated. In addition, the PSS design module was realized in the prototype system, which was developed by our research group to assist in PSS design.

2. PSS Evolution Routes Based on the Ideal Final Result of TRIZ

In this study, PSS is defined as a balanced production system with appropriate proportion between products and services and generally optimized in the whole life cycle service of a product. In the PSS, the relationship between a product and a service is inseparable and interactive. Product is a platform that provides service, whereas service can create value without using materials. Determining how to find the balance between products and services to reduce material flow, to maximize resources, and to protect the environment is the research goal of this study. As the most basic element of conceptual design, function runs through all stages of conceptual design. Starting with function, this study seeks the interactive balance between product function and service function in the conceptual design stage. Hence, PSS innovative design can be linked with customer demands from the very beginning and eliminates the limitations of looking only at product, structure, and technical carrier.

The Ideal Final Result (IFR) of TRIZ [10] refers to the ideal solution that can eliminate the defects of the existing system while maintaining its advantages. The direction of the ideal solution is clearly defined to ensure that the system can always move forward in the process of problem solving. The disadvantage of not having a target in traditional innovative design methods can be avoided, and the efficiency of the innovative design can be improved. The important issue in this technical system is how to realize the function more scientifically rather than the system itself. Idealized improvement of this system is maximizing useful functions and minimizing harmful functions.

The IFR of TRIZ sets a series of ideal models at the beginning of problem solving. The ideal model contains all the elements involved in the problem, such as the ideal system, ideal process, ideal resources, ideal method, ideal

machine, and ideal materials. The final idealized state of the PSS is an ideal system. The system has no entity, no material, and no energy consumption, but it can achieve all required functions. The parameters that measure the idealization degree of the system are named as the idealization level. The TRIZ idealized level can be calculated as follows:

$$I = \frac{\sum \mathrm{UF}}{\sum \mathrm{HF}}. \tag{1}$$

Here, I represents the idealized level, $\sum$UF represents the sum of useful functions, and $\sum$HF represents the sum of harmful functions. According to the IFR of TRIZ, the idealization level is as follows:

$$I = \frac{\sum \mathrm{UF}}{\left(\sum \mathrm{cost} + \sum \mathrm{consumption}\right)}. \tag{2}$$

Here, I represents the idealized level, $\sum$UF represents the useful function of the PSS, $\sum$cost represents the total cost consumed by the PSS, and $\sum$consumption represents the total consumption to the environment. From formula (2), the idealized level of the PSS is directly proportional to useful functions. However, the idealized level of the PSS is inversely proportional to the total cost of consumption and the total consumption of the environment. The PSS design aims to increase the idealized level. According to formula (2), the evolution route of the PSS has the following states:

(1) Increase useful functions: service provides product life cycle benefits and increases useful functions of the PSS. In this evolutionary route, manufacturers change the single physical product form into a diversified product and service form. The product not only addresses customers' demands but also adds a range of services around the physical product to sustain and extend customers' requirements.

(2) Reduce costs: maximizing the efficiency of the physical product is the purpose of this route. In this evolutionary route, manufacturers try to reduce the amount of the physical product to reduce costs. Through rational use and distribution of the physical products, previously products owned by customers can be transformed into customer leasing and sharing.

(3) Reduce environmental consumption: service provides the result directly to customers. Manufacturers focus more on what the physical product can offer, but less on how to achieve it. As much as possible, tangible products are no longer needed, and customers' needs can be achieved in the form of service.

Figure 1 shows the evolution route of the PSS based on the IFR of TRIZ.

3. PSS Function Representation and Incentive Strategies

According to Umeda et al. [25], the function of the PSS was defined as a series of product behavior or service behavior descriptions from the perspective of customers. According to this definition, PSS functions are expressed as a series of actions in the form of "what to do."

3.1. PSS Function Modeling. Considering the heterogeneity and interactivity of service and product functions, an effective way is needed to represent the function interaction model of the product and the service in PSS design. A function representation method including the interactive relationship between the product and service was established according to the service blueprint method [26] and the system function diagram [27], which was called the PSS function blueprint method.

In this study, three lines were proposed to divide the whole PSS into four different function areas. The function model boundary was expanded from only the physical product to include the interaction of the customer, product, and service. The three dividing lines of the PSS are product service line, product use line, and service visualization line. First, the whole PSS function blueprint was divided into product function domain and service function domain by the product service line. The division describes the interaction between the product and the service. The product function domain contained the function set of the physical product, whereas the service function domain contained the function set of the intangible service. Second, the product use line divided the product function domain into product use domain and product management domain. The product use domain contained the primary function set that the physical product must have in the use process. The product management domain contained the support function set that can support the physical product to complete its essential functions through software during usage. The relationship between product function domain and product management domain can be represented by the input/output energy flow based on the system function diagram. The product use line describes the interaction between the hardware function and the software function during usage of the physical product. Finally, the service visualization line divided the service function domain into visual service domain and invisible service domain. The visual service domain contained the service function set that customers perceived and engaged with directly. By contrast, the invisible service domain contained the background service function set that customers cannot realize and can only obtain indirectly. The relationship between these two functions was represented by the input and output service flow based on the function system diagram, and the service visualization line described the interaction between the service that customers can directly perceive. Figure 2 depicts four domains of the PSS, as well as the standardized representation of different function types and input/output flow. The thick solid line represents the product service line; the thick dashed line represents the product use line; the thick dotted line represents the visual service line; and the thin solid/dashed line with an arrow represents the input/output flow between functions.

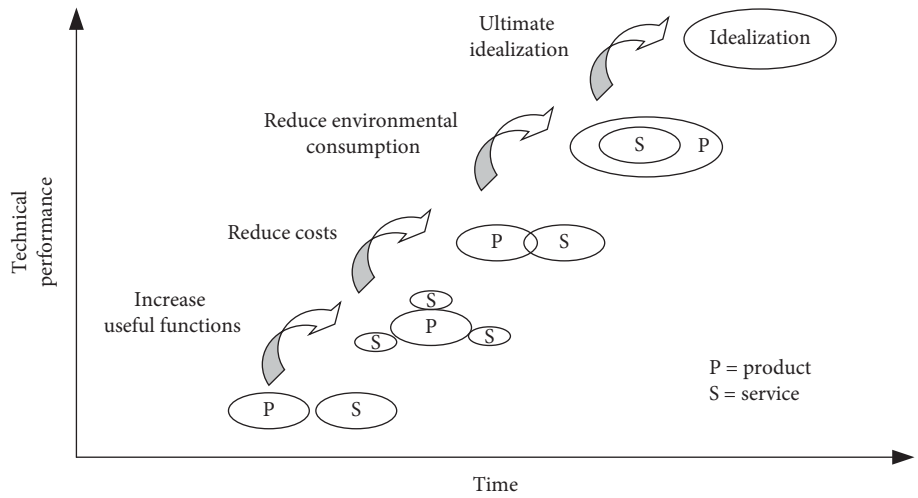


Figure 1: PSS evolution route.

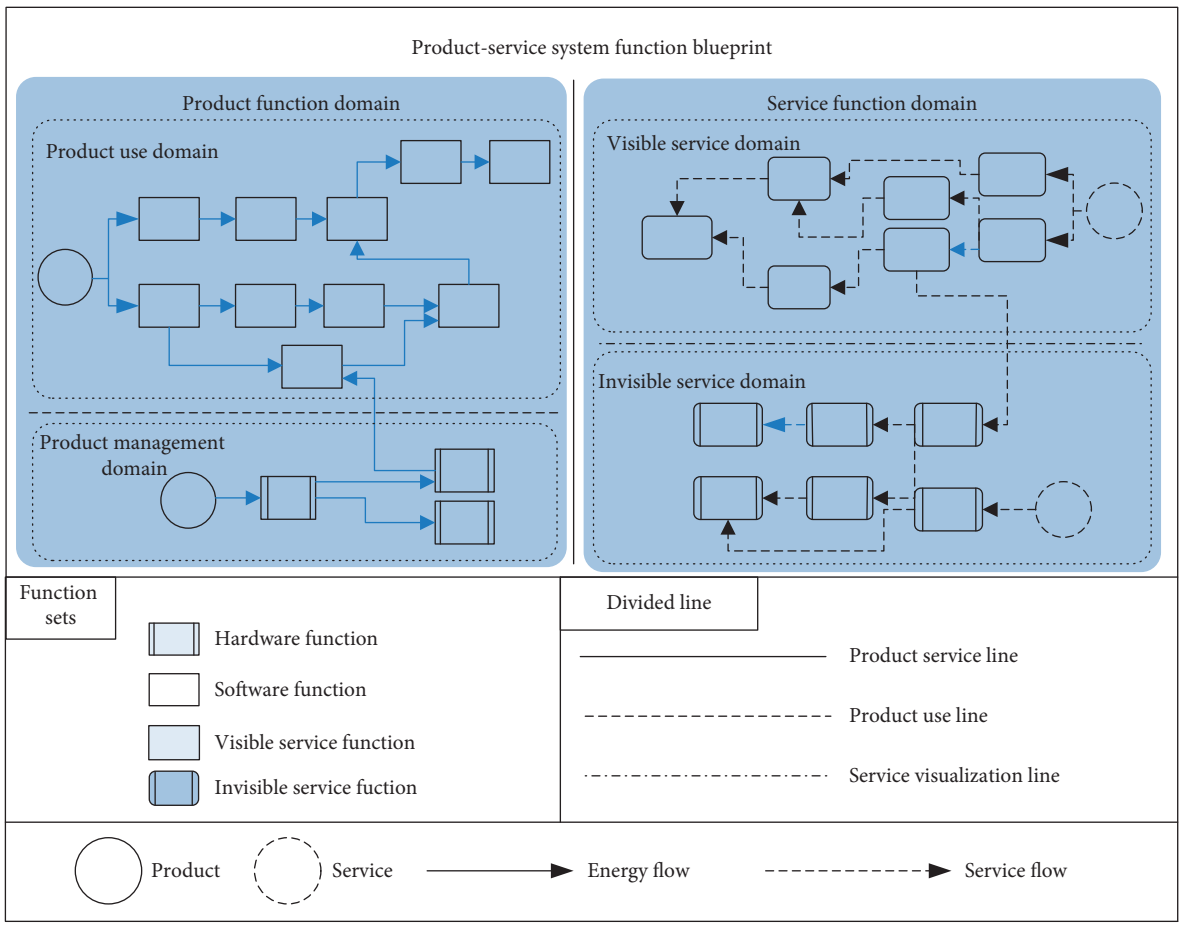


Figure 2: PSS function modeling.

3.2. Function Incentive Strategies of PSS. In this study, function incentive strategies under each evolution route of the PSS based on the IFR of TRIZ were proposed, and the innovative design under different routes was realized.

3.2.1. Synergetic Innovation. According to the initial state of the PSS evolution route, the initial evolution direction of the PSS was to increase its useful function by providing additional services for the physical product in the product life

cycle. Under this evolutionary route, the function incentive strategy was synergetic innovation. Synergetic innovation retained the original excellent characteristics of the existing product. Moreover, it added better functional characteristics after introducing service functions. Take the remote technical support provided by Lenovo, for example. The "maintenance function" in the visual service domain and the "Internet function" in the product use domain worked in coordination with each other. Customers made a corresponding remote service reservation via e-mail or app. In turn, technicians can directly troubleshoot software problems, solve technical problems remotely, and restore hard disk data through the network. Therefore, synergistic innovation tapped the potential functions of the PSS. In addition, it can open new market demand and improve the benefit of an existing product. Ultimately, it can increase useful functions of the PSS.

3.2.2. Supplement Innovation. According to the intermediate state of the PSS evolution route, the evolution direction of the PSS was to change a physical product into a service platform, and its purpose was to optimize usage efficiency. In this evolutionary route, the function incentive strategy was the supplement innovation. Traditional product functions cannot fulfill customers' requirements, and they cannot be further improved due to their technical limitations. Service functions were introduced to supplement product function. For example, traditional medical diagnosis only relied on doctors' previous diagnoses. Hence, different doctors give different conclusions. To address this limitation, GE Medical established an automatic medical support center. The "data interaction" in the invisible service domain and the "cause diagnosis" in the product use domain supplemented each other. Patient diagnosis results were fed back to the GE Medical online center. Through the big medical data system, similar electronic medical records can be retrieved, diagnosis and treatment modes of different doctors can be mined, complementary and interactive diagnosis for patients can be made, and traditional medical devices can be transformed into this platform to provide optimized medical services. Therefore, the service supplement product changed the potential form of the PSS, made the configuration of resources reasonable, and improved the usage efficiency of the physical product to achieve the goal of reducing the PSS cost.

3.2.3. Substitution Innovation. According to the final state of the evolution route, the evolution direction of the PSS was to provide the final result by providing direct service. In this evolutionary route, the function incentive strategy was the substitution innovation. Service functions not only met customers' needs but were also provided with the support of the physical product. For example, iCloud adopted a new way to store and access local files on the basis of network cloud technology. By providing each customer with a cloud terminal, the synchronization and push of information can be realized. iCloud avoided the hard disk drive purchase cost, which shortened the operation cycle and reduced consumption of resources. Therefore, according to substitution innovation, customers can obtain the result of the product directly without owning or purchasing the material form of the product, thus reducing environmental consumption.

4. PSS Innovation Design

In this study, the innovation design method of the PSS was systematically realized by different evolution routes. The function model of PSS was proposed, and the incentive strategies were established.

4.1. PSS Innovation Design Process. Figure 3 shows the innovative design process model of the PSS. The design process model includes the following main modules:

4.1.1. Function Modeling Module. In this module, stakeholders of an existing product were analyzed, and the functions of the product were deconstructed. Product functions and service functions were extracted and entered. The PSS function blueprint was constructed according to the function components and stakeholders. The interaction between product and service functions as well as the relationship between different stakeholders and the product/service was clarified in the PSS.

4.1.2. Evolution Route Selection Module. Designers obtained different customers' requirements through market survey, and the evolution route of the PSS was selected according to these requirements. The selected evolutionary route assisted the designer in clarifying the direction of the PSS innovation design.

4.1.3. Function Incentive Strategy Module. From the selected evolutionary route, the function components were analyzed according to the PSS function blueprint. The multiple possible function results were obtained through the function incentive strategy to obtain the corresponding original design solutions of the PSS.

4.1.4. Design Solution Generation and Evaluation Module. The original PSS solutions were evaluated on the basis of customers' needs. If the market demand and customer demand were not satisfied, then further evolution route and other function incentive strategies were chosen. If the market demand and customer demand were met, then the original PSS solutions were translated into the detailed PSS innovation design solutions with the help of existing knowledge resources. The detailed PSS innovation design solutions were evaluated and optimized according to the manufacturing cost, market condition, and the actual situation of the enterprise. A detailed PSS design solution with better innovation, practicality, and reliability was proposed.

4.2. PSS Innovation Design Based on CAIP 3.0. Computer-Aided Innovation Platform (CAIP 3.0) is an innovative design system. It converges cognitive science,

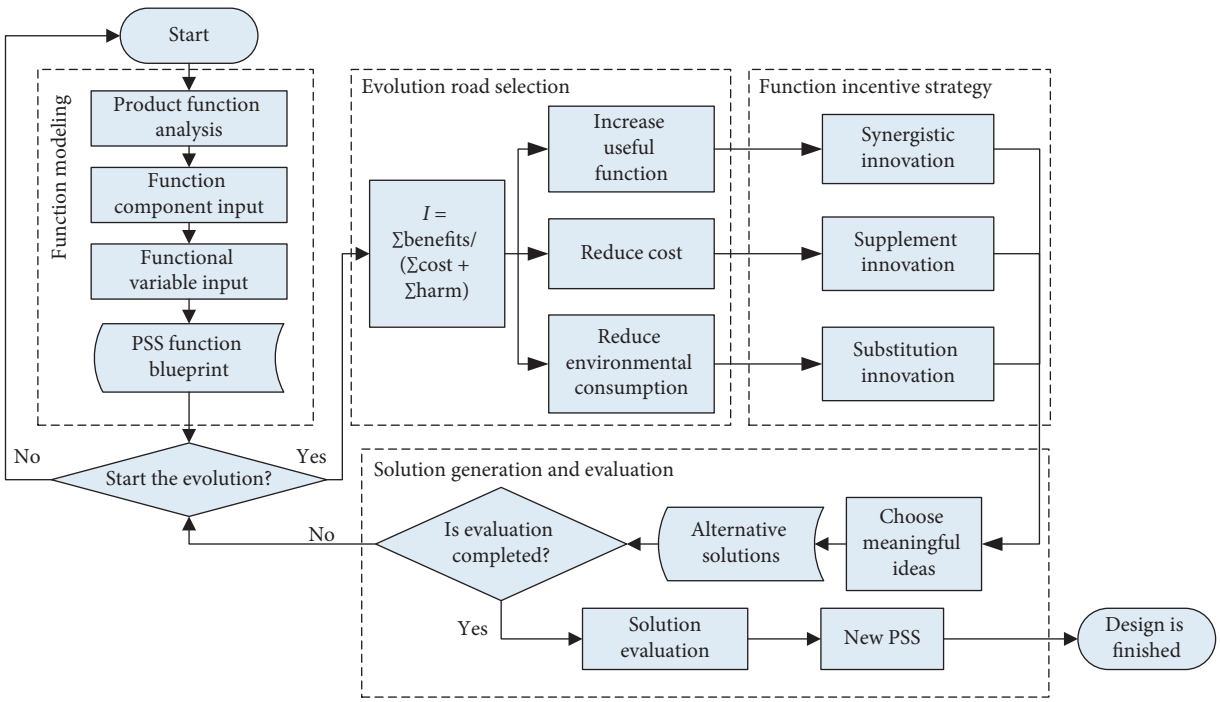


Figure 3: PSS innovation design process.

information technology, innovation design theory, and methodology to assist designers in creative design through the interaction of design process and knowledge base. The main development tools and running environment of CAIP 3.0 include Java, Eclipse, Microsoft Access, AWT, Swing, and Eclipse GEF. The development language is JAVA, and the integrated development environment is Eclipse. The background database management system is Microsoft Access. The central technology of the system interface is AWT and Swing. The graphical editing framework is Eclipse GEF. The CAIP 3.0 system is mainly composed of the following modules: project management module, problem analysis module, innovative technique module, knowledge base module, and solution management module.

The innovative design process of the PSS based on TRIZ and function incentives can be realized through the PSS design module of CAIP 3.0. PSS design module mainly includes demand analysis, function modeling, function incentives, and solution generation. The PSS function blueprint construction was implemented by the function modeling submodule in CAIP 3.0. The interaction of product functions and service functions was implemented by the function incentives submodule in CAIP 3.0. The design solutions generation and evaluation were realized through solution generation submodule according to the existing knowledge base of CAIP 3.0. In CAIP 3.0, designers first input product/service functions and function attributes. After clicking the "finish button," a PSS function blueprint is generated automatically. The PSS function modeling is finally realized. According to different evolution routes, different function incentive strategies were carried out on the PSS function blueprint. Finally, the PSS innovation design solutions were generated using the existing knowledge base of CAIP 3.0.

5. Agricultural Machinery PSS Innovation Design

China is a large agricultural country with 900 million farmers. It is currently transitioning from traditional agriculture into modern agriculture. The demand for agricultural machinery products is large, and agricultural machinery production accounts for a large proportion of China's manufacturing industry. Traditional agricultural machinery manufacturers had no other way to increase their profits except for profiting from selling machinery and providing simple maintenance service. However, farmers often only use single functions offered by traditional agricultural machines, and they do not receive professional training in farming. In the face of different crops and fertility levels, farmers can rely on their own experiences to cultivate. Moreover, agricultural machinery manufacturers urgently need agricultural machinery with innovative PSS design. With the transformation to the PSS, agricultural machinery manufacturers can abandon the traditional manufacturing model. Traditional agricultural machinery enterprises transformed into agricultural service enterprises by providing crop growth management and training. The proposed PSS design method in this study was verified through the agricultural machinery.

5.1. Agricultural Machinery Demand Analysis. First, the PSS design needed to acquire and analyze customers' demand and to obtain original customers' demand data (most of the data in this study came from the Internet) which were obtained by means of customer interviews and market feedback. These data were input into the demand analysis

submodule. At present, customers want agricultural machinery to offer multifunctional efficiency, high reliability, good operability, beautiful appearance, energy-saving features, and environmental protection. The preprocessing algorithm of customers' requirements in the submodule was used to classify and combine the original customers' requirements. The sorted customers' needs were classified according to basic needs, existing needs, expected needs, and irrelevant needs. Then, the analytic hierarchy process was used to rank the priority weight among the classified customers' needs. Figure 4 shows the analysis submodule of customers' demands on agricultural machinery.

5.2. Agricultural Machinery Function Model Construction. On the basis of existing agricultural machinery, the corresponding function model was constructed. Product functions were divided into product use functions and product management functions, whereas service functions were divided into visible service functions and the invisible service functions. Functions obtained by customers directly through agricultural machinery products were defined as product use functions. Examples of such functions are cultivating land, sowing seeds, irrigating crops, spraying pesticides, removing weeds, applying chemical fertilizer, and harvesting crops. Functions obtained by customers through software technology and manufacturer interactive access were defined as product management functions. These functions include remote monitoring, automatic positioning, sensing information, uploading data, and managing data. Functions obtained by customers through direct experience were defined as visible service functions, which include product trial, periodic maintenance, training, fault diagnosis, maintenance products, and recycling of waste products. Functions that cannot be directly perceived by customers through the third-party service provider were defined as invisible service functions, such as risk aversion, soil analysis, and fertilizer production and transportation. Figure 5 shows the function model construction of the agricultural machinery PSS based on CAIP 3.0.

5.3. Evolutionary Road Selection and Function Incentives. After function input, the PSS function blueprint was generated automatically using CAIP 3.0. The relationship between service function and product function was represented, and the function correlation matrix of the agricultural machinery was established. According to the function incentive strategies corresponding to different evolutionary routes, the functions of other regions in the PSS blueprint were selected and integrated. Figure 6 shows the function incentive selection of the agricultural machinery.

According to the following evolution routes, different function incentive strategies were adopted to the agricultural machinery PSS innovation design.

5.3.1. Evolutionary Road 1: Increase the Useful Function of the Existing Product. The corresponding function incentive strategy was synergetic innovation. The "fault diagnosis" function of the visual service functions and the "remote monitoring" function of the product use functions complemented each other. The operating condition information collected by traditional agricultural machinery was used for troubleshooting only when a problem occurred. Nonetheless, the preventive inspection and maintenance of the agricultural machinery could be done on the basis of these data. According to big data analysis technology, remote technical maintenance guidance and fault warning could be provided to farmers. The failure rate of the agricultural machinery and possible losses caused by defects were significantly reduced.

5.3.2. Evolution Road 2: Lower the Cost. The corresponding function incentive strategy was supplement innovation. The "soil professional analysis" function of the invisible service functions and the "data management" function of the product management functions collaborated with each other. The agricultural production information was collected to understand the impact of soil condition on crop yield, and the collected data were stored in the service database. The scanned land morphology was divided into small pieces, and the soil information of each small piece was collected and analyzed through the sensor installed on the agricultural machinery. The data on the impact of soil conditions on crop yield were stored in the database and analyzed. Optimal seed ratio, optimal fertilization scheme, optimal use of the pesticides with minimal environmental pollution, and most reasonable irrigation methods were determined, to provide a reference for farmers.

5.3.3. Evolution Road 3: Reduce Environmental Consumption. The corresponding function incentive strategy was substitution innovation. The "fertilize" function of the product use functions was replaced by the "fertilizer transport" function of the invisible service functions. Fertilizer was applied to the roots of the crops through the buried fertilization system. Chemical fertilizer producers established alliances with agricultural machinery manufacturers to obtain first-hand information about customers and their soils. According to the data, fertilizer producers provided a customer with a complete analysis of soils and crops. They can also provide customers with complete crop fertilization solutions including optimal ratio of required fertilizer, schedule of crop fertilization, and customized fertilizers for different crops. The fertilizer producers charged according to the area (square kilometers) that received fertilization services.

5.4. Function Solving and Conceptual Solution Evaluation. The preventive maintenance system of the agricultural machinery product was obtained by "supplement innovation." With the help of specialized knowledge and professional service, useful functions increased and the

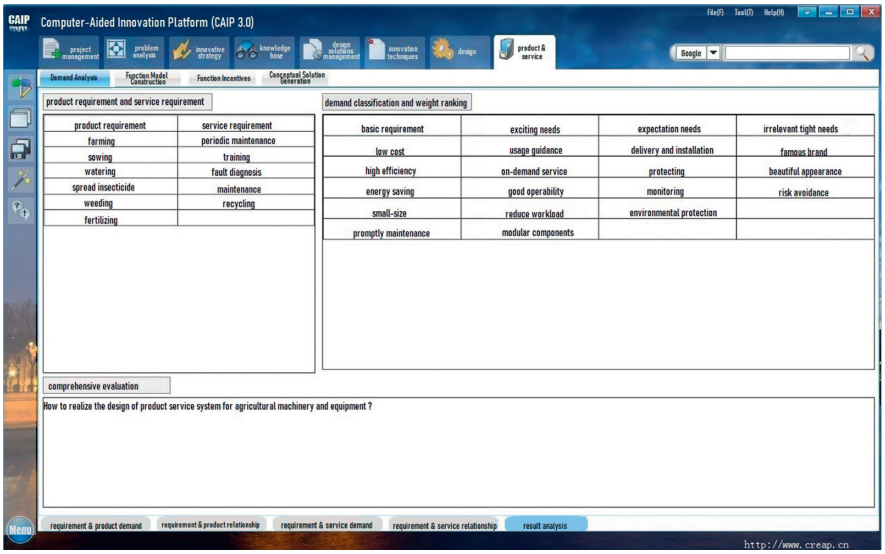


Figure 4: Analysis submodule of customers' demand of agricultural machinery based on CAIP 3.0.

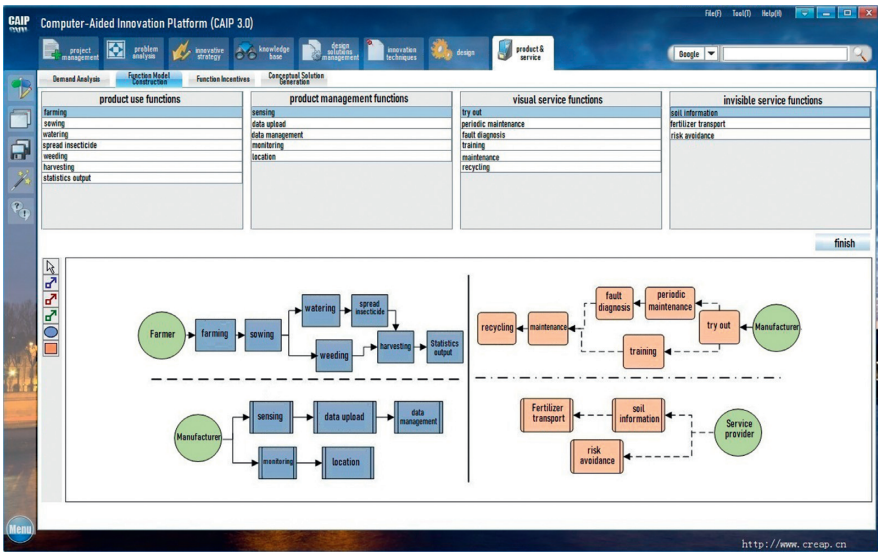


Figure 5: Function blueprint submodule of the agricultural machinery PSS based on CAIP 3.0.

value of the product was optimized. The crop growth management service system was obtained through "synergistic innovation." The output of crops on the land was maximized, and the cost was reduced as much as possible. The buried fertilization system was obtained through "substitution innovation." A complete crop fertilization scheme was provided to farmers directly, and the fertilizer plant charged per service. To gain more innovative solutions, the knowledge base was established to seek appropriate engineering examples. The original solutions were solved by function to principle mapping. Finally, the detailed design was carried out. The designer evaluated the new PSS solutions according to the potential market feasibility cost and the actual situation of the enterprise. The optimal PSS solution was determined by using the existing search algorithms, as shown in Figure 7.

6. Discussion of Results

6.1. Novelty and Advantages. The distinctive feature of TRIZ is systematic problem solving process. The whole problem-solving process was guided by TRIZ tools, which indicate the direction of the problem solving. Substantial research has been conducted on PSS design based on TRIZ [12, 28, 29], which adopted different tools in problem definition, problem solving, and solution evaluation stages. Existing studies mainly focused on the application of conflict matrix to define a problem and to solve the problem through innovation principle, separation principle, and standard solution. In this study, on the basis of the ideal solution method of TRIZ, the evolutionary route of the PSS innovation design was explored according to the proportion of products and services and the change of product ownership. In the conceptual design stage, the blindness caused by design can be

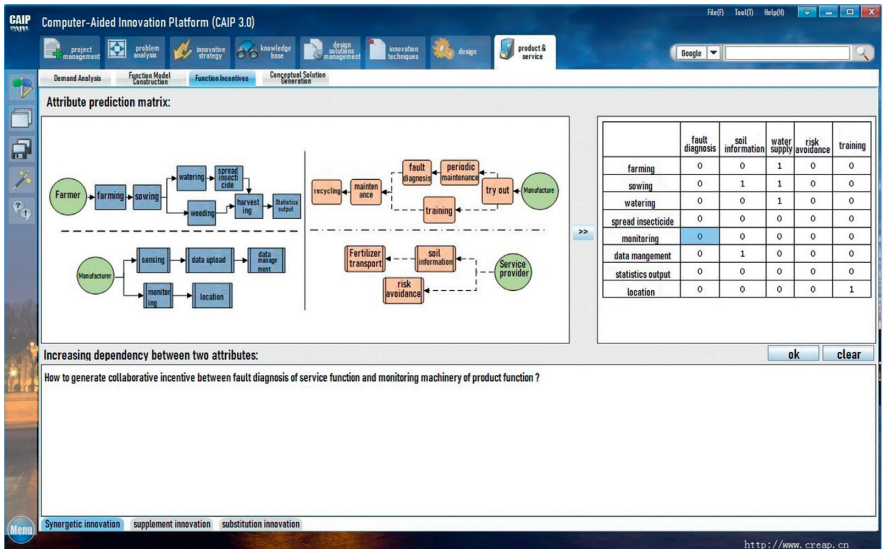


Figure 6: Function incentive selection submodule of the agricultural machinery PSS based on CAIP 3.0.

Figure 7: Design solutions submodule of the agricultural machinery PSS based on CAIP 3.0.

effectively improved by looking for the optimal integration method of product function and service function at different stages.

This study also constructed computer-aided PSS innovation design modules using CAIP 3.0. By building a function model, evolution route selection, incentive function and solving, and concept scheme evaluation, it provided a useful tool for designers in the conceptual design and scheme design stages. Moreover, it promoted designers' creativity in the creative design phase. CAIP 3.0 system was applied in an agricultural machinery manufacturing enterprise to guide designers in an enterprise to design differentiated products by combining services and to improve the product competitiveness of the enterprise.

6.2. Managerial Implications. This study introduced services in the concept design stage to reduce production costs and to increase the value of physical products. Reducing service costs was easier than reducing production costs. By reducing service costs and improving service quality, the market competitiveness of enterprises can be improved.

Under fierce market competition, products that cannot meet the market demand will be eliminated, and a differentiated design that comes only from a physical product cannot adapt in the market. Therefore, this study built differentiated products with services through product and service function interaction to improve the quality and adaptability of external services, to expand the scope of services, and ultimately to gain benefits.

With the growing environmental problems, the environmental protection requirements of enterprises are becoming increasingly high. Servicization of manufacturing enterprises is an alternative under environmental pressure. It can reduce the negative impact of tangible products on the environment by selling services instead of products.

6.3. Limitations. Despite these innovations, the present study has limitations. First, the cognitive mechanism of the conceptual solutions generated by the designer in the PSS concept design process and the cognitive basis of the service participants' behavior were not considered. Nevertheless, product designers' cognitive behavior can be used to produce comparable and commensurable results when analyzed by PSS designers. In addition, the PSS concept design lacked quantitative methods to select and evaluate the design solutions. The demands of users, manufacturers, service providers, and other stakeholders needed to be integrated, and a quantitative evaluation mechanism needed to be set to select and evaluate the PSS conceptual design from multiple perspectives. These two points above will be the authors' future research directions.

7. Conclusions

In this study, a PSS innovation design method based on IFR of TRIZ and function incentive strategy was proposed. In this innovation mode, three possible evolution routes of the PSS were proposed. The function model of the PSS was established by the PSS function blueprint and system function diagram. On the basis of the evolution routes, different function incentive strategies were carried out to obtain the conceptual design solutions. The best PSS solution was generated by using the existing knowledge base and search algorithm. The relationship between product and service was analyzed and clarified in the conceptual design stage. Material flow (the quantity of the product) was reduced, and environmental friendliness was increased through this method. Finally, the PSS design process of agricultural machinery based on CAIP 3.0 was analyzed as an example to verify this method.

Data Availability

The data can be found in http://www.creap.cn/

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] E. Manzini, C. Vezzoli, and G. Clark, "Product-service systems: using an existing concept as a new approach to sustainability," *Journal of Design Research*, vol. 1, no. 2, pp. 27–40, 2001.

[2] H. Meier, O. Völker, and B. Funke, "Industrial product-service systems (IPS2)," *The International Journal of Advanced Manufacturing Technology*, vol. 52, no. 9–12, pp. 1175–1191, 2011.

[3] T. Hara, T. Arai, and Y. Shimomura, "A CAD system for service innovation: integrated representation of function, service activity, and product behaviour," *Journal of Engineering Design*, vol. 20, no. 4, pp. 367–388, 2009.

[4] S. Jiang, D. Feng, C. Lu et al., "Research on the construction of the spiral evolutionary design methodology for a product service system based on existing products," in *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*vol. 234, no. 4, pp. 825–839, 2020.

[5] W. Li and F.-T. Chan, "Multi-objective configuration optimization for product-extension service," *Journal of Manufacturing Systems*, vol. 37, pp. 113–125, 2015.

[6] D. Yin, X. Ming, Z. Liu, and X. Zhang, "A fuzzy ANP-QFD methodology for determining stakeholders in product-service systems development from ecosystem perspective," *Sustainability*, vol. 12, no. 8, 2020.

[7] M. Sholihah, Y. Mitake, T. Nakada et al., "Innovative design method for a valuable product-service system: concretizing multi-stakeholder requirements," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 13, no. 5, 2019.

[8] T. Sakao and Y. Shimomura, "Service engineering: a novel engineering discipline for producers to increase value combining service and product," *Journal of Cleaner Production*, vol. 15, no. 6, pp. 590–604, 2007.

[9] G. Pezzotta, F. Pirola, R. Pinto, and F. Akasaka, "A Service Engineering framework to design and assess an integrated product-service," *Mechatronics*, vol. 31, pp. 169–179, 2015.

[10] N. Shimomura and M. Fargnoli, "Designing product-service systems: a review towards a unified approach," in *Proceedings of the International Conference On Industrial Engineering And Operations Management*, pp. 817–837, Rabat, Morocco, April 2017.

[11] G.-S. Altshuller, *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*, Technical Innovation Center, Inc., Shanghai, China, 1999.

[12] S. Kim and B. Yoon, "Developing a process of concept generation for new product-service systems: a QFD and TRIZ-based approach," *Service Business*, vol. 6, no. 3, pp. 323–348, 2012.

[13] C.-H. Lee, C.-H. Chen, and Y.-C. Lee, "Customer requirement-driven design method and computer-aided design system for supporting service innovation conceptualization handling," *Advanced Engineering Informatics*, vol. 45, p. 101117, 2020.

[14] J.-W. Wang and J.-M. Zhang, "Research on innovative design and evaluation of agricultural machinery products," *Mathematical Problems in Engineering*, vol. 2019, Article ID 8179851, 18 pages, 2019.

[15] Y.-H. Wang, C.-H. Lee, and A. J. C. Trappey, "Service design blueprint approach incorporating TRIZ and service QFD for a meal ordering system: a case study," *Computers & Industrial Engineering*, vol. 107, pp. 388–400, 2017.

[16] W. Song and T. Sakao, "A customization-oriented framework for design of sustainable product/service system," *Journal of Cleaner Production*, vol. 140, pp. 1672–1685, 2017.

[17] L. Yang and K. Xing, "Innovative conceptual design approach for product service system based on TRIZ," in *Proceedings of the 2013 10th International Conference on Service Systems and Service Management*, pp. 247–252, IEEE, Hong Kong, China, July 2013.

[18] J. L. Chen and Y. Liu, "Innovative design and assessment of low-carbon emission concept product service systems," *The Philosopher's Stone For Sustainability*, pp. 369–374, Springer, Berlin, Germany, 2013.

[19] N. Haber, M. Fargnoli, and T. Sakao, "Integrating QFD for product-service systems with the Kano Model and fuzzy AHP," *Total Quality Management and Business Excellence*, vol. 31, pp. 1–26, 2018.

[20] N. Haber, M. Fargnoli, and T. Sakao, "PSS modularization: a customer-driven integrated approach," *International Journal of Production Research*, vol. 57, pp. 1–17, 2018.

[21] N. Haber and M. Fargnoli, "A practical ANP-QFD methodology for dealing with requirements' inner dependency in PSS development," *Computers & Industrial Engineering*, vol. 127, pp. 536–548, 2019.

[22] T. Sakao, W.-Y. Song, and J. Matschewsky, "Creating service modules for customising product/service systems by extending DSM," *CIRP Annals-Manufacturing Technology*, vol. 66, pp. 21–24, 2017.

[23] J. Lee and M. Abuali, "Innovative Product Advanced Service Systems (I-PASS): methodology, tools, and applications for dominant service design," *The International Journal of Advanced Manufacturing Technology*, vol. 52, no. 9–12, pp. 1161–1173, 2011.

[24] H. Son, Y. Kwon, S.-C. Park et al., "Using a design structure matrix to support technology road mapping for product-service systems," *Technology Analysis & Strategic Management*, vol. 20, no. 3, pp. 223–250, 2018.

[25] Y. Umeda, M. Ishii, M. Yoshioka, and T. Tomiyama, "Supporting conceptual design based on the function-behavior-state modeler," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 10, no. 4, pp. 275–288, 1996.

[26] G.-L. Shimomura, "How to design a service," *European Journal Of Marketing*, vol. 16, 1982.

[27] K.-N. Otto, *Product Design: Techniques In Reverse Engineering And New Product Development*, Tsinghua University Press, Beijing, China, 2003.

[28] J. P. Dlego-Augusto, T. C. Carla-Schwengber, J. Carlos-Fernando et al., "State of the art on the role of the theory of inventive problem solving in sustainable product-service systems: past, present, and future," *Journal of Cleaner Production*, vol. 212, pp. 489–504, 2019.

[29] L.-J. Yang and K. Xing, "Innovative conceptual design approach for product service system based on TRIZ," in *Proceedings of the International Conference on Service Systems & Service Management*, IEEE, Hong Kong, China, July 2013.