

Artificial Intelligence in Disease Diagnosis

Lead Guest Editor: Yan-Wu Xu

Guest Editors: Xiangjia Zhu, Weihua Yang, and Huiying Liu





Artificial Intelligence in Disease Diagnosis

Journal of Healthcare Engineering

Artificial Intelligence in Disease Diagnosis

Lead Guest Editor: Yan-Wu Xu

Guest Editors: Xiangjia Zhu, Weihua Yang, and
Huiying Liu



Copyright © 2023 Hindawi Limited. All rights reserved.

This is a special issue published in "Journal of Healthcare Engineering." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editors

Xiao-Jun Chen , China
Feng-Huei Lin , Taiwan
Maria Lindén, Sweden

Academic Editors




Cherif Adnen, Tunisia
Saverio Affatato , Italy
Óscar Belmonte Fernández, Spain
Sweta Bhattacharya , India
Prabadevi Boopathy , India
Weiwei Cai, USA
Gin-Shin Chen , Taiwan
Hongwei Chen, USA
Daniel H.K. Chow, Hong Kong
Gianluca Ciardelli , Italy
Olawande Daramola, South Africa
Elena De Momi, Italy
Costantino Del Gaudio , Italy
Ayush Dogra , India
Luobing Dong, China
Daniel Espino , United Kingdom
Sadiq Fareed , China
Mostafa Fatemi, USA
Jesus Favela , Mexico
Jesus Fontecha , Spain
Agostino Forestiero , Italy
Jean-Luc Gennisson, France
Badicu Georgian , Romania
Mehdi Gheisari , China
Luca Giancardo , USA
Antonio Gloria , Italy
Kheng Lim Goh , Singapore
Carlos Gómez , Spain
Philippe Gorce, France
Vincenzo Guarino , Italy
Muhammet Gul, Turkey
Valentina Hartwig , Italy
David Hewson , United Kingdom
Yan Chai Hum, Malaysia
Ernesto Iadanza , Italy
Cosimo Ieracitano, Italy

Giovanni Improta , Italy
Norio Iriguchi , Japan
Mihajlo Jakovljevic , Japan
Rutvij Jhaveri, India
Yizhang Jiang , China
Zhongwei Jiang , Japan
Rajesh Kaluri , India
Venkatachalam Kandasamy , Czech Republic
Pushpendu Kar , India
Rashed Karim , United Kingdom
Pasi A. Karjalainen , Finland
John S. Katsanis, Greece
Smith Khare , United Kingdom
Terry K.K. Koo , USA
Srinivas Koppu, India
Jui-Yang Lai , Taiwan
Kuruva Lakshmanna , India
Xiang Li, USA
Lun-De Liao, Singapore
Qiu-Hua Lin , China
Aiping Liu , China
Zufu Lu , Australia
Basem M. ElHalawany , Egypt
Praveen Kumar Reddy Maddikunta , India
Ilias Maglogiannis, Greece
Saverio Maietta , Italy
M.Sabarimalai Manikandan, India
Mehran Moazen , United Kingdom
Senthilkumar Mohan, India
Sanjay Mohapatra, India
Rafael Morales , Spain
Mehrbakhsh Nilashi , Malaysia
Sharnil Pandya, India
Jialin Peng , China
Vincenzo Positano , Italy
Saeed Mian Qaisar , Saudi Arabia
Alessandro Ramalli , Italy
Alessandro Reali , Italy
Vito Ricotta, Italy
Jose Joaquin Rieta , Spain
Emanuele Rizzuto , Italy

Dinesh Rokaya, Thailand
Sébastien Roth, France
Simo Saarakkala , Finland
Mangal Sain , Republic of Korea
Nadeem Sarwar, Pakistan
Emiliano Schena , Italy
Prof. Asadullah Shaikh, Saudi Arabia
Jiann-Shing Shieh , Taiwan
Tiago H. Silva , Portugal
Sharan Srinivas , USA
Kathiravan Srinivasan , India
Neelakandan Subramani, India
Le Sun, China
Fabrizio Taffoni , Italy
Jinshan Tang, USA
Ioannis G. Tollis, Greece
Ikram Ud Din, Pakistan
Sathishkumar V E , Republic of Korea
Cesare F. Valenti , Italy
Qiang Wang, China
Uche Wejinya, USA
Yuxiang Wu , China
Ying Yang , United Kingdom
Elisabetta Zanetti , Italy
Haihong Zhang, Singapore
Ping Zhou , USA







Contents

Multi-Layer Perceptron Classifier with the Proposed Combined Feature Vector of 3D CNN Features and Lung Radiomics Features for COPD Stage Classification

Yingjian Yang , Nanrong Zeng, Ziran Chen, Wei Li, Yingwei Guo, Shicong Wang, Wenxin Duan, Yang Liu, Rongchang Chen , and Yan Kang 



Research Article (15 pages), Article ID 3715603, Volume 2023 (2023)

Development and Application of a Standardized Testset for an Artificial Intelligence Medical Device Intended for the Computer-Aided Diagnosis of Diabetic Retinopathy

Hao Wang , Xiangfeng Meng , Qiaohong Tang , Ye Hao , Yan Luo , and Jiage Li 

Research Article (9 pages), Article ID 7139560, Volume 2023 (2023)

An End-to-End Data-Adaptive Pancreas Segmentation System with an Image Quality Control Toolbox

Yan Zhu , Peijun Hu, Xiang Li, Yu Tian, Xueli Bai, Tingbo Liang, and Jingsong Li 

Research Article (12 pages), Article ID 3617318, Volume 2023 (2023)

Pterygium Screening and Lesion Area Segmentation Based on Deep Learning

Shaojun Zhu , Xinwen Fang, Yong Qian, Kai He , Maonian Wu , Bo Zheng, and Junyang Song 




Research Article (9 pages), Article ID 3942110, Volume 2022 (2022)

Comparing Conventional and Deep Feature Models for Classifying Fundus Photography of Hemorrhages

Tamoor Aziz, Chalie Charoenlarnopparut , and Srijidtra Mahapakulchai

Research Article (9 pages), Article ID 7387174, Volume 2022 (2022)

Machine Learning-Based Prediction Model of Preterm Birth Using Electronic Health Record

Qi Sun , Xiaoxuan Zou, Yousheng Yan, Hongguang Zhang, Shuo Wang, Yongmei Gao, Haiyan Liu, Shuyu Liu, Jianbo Lu , Ying Yang , and Xu Ma 


Research Article (12 pages), Article ID 9635526, Volume 2022 (2022)

Multiview Volume and Temporal Difference Network for Angle-Closure Glaucoma Screening from AS-OCT Videos

Luoying Hao , Yan Hu , Risa Higashita , James J. Q. Yu , Ce Zheng , and Jiang Liu 

Research Article (9 pages), Article ID 2722608, Volume 2022 (2022)

Augmentation-Consistent Clustering Network for Diabetic Retinopathy Grading with Fewer Annotations

Guanghai Zhang, Keran Li, Zhixian Chen, Li Sun, Jianwei zhang, and Xueping Pan 

Research Article (10 pages), Article ID 4246239, Volume 2022 (2022)

Research Article

Multi-Layer Perceptron Classifier with the Proposed Combined Feature Vector of 3D CNN Features and Lung Radiomics Features for COPD Stage Classification

Yingjian Yang ^{1,2}, Nanrong Zeng,^{2,3} Ziran Chen,^{1,2} Wei Li,² Yingwei Guo,^{1,2} Shicong Wang,^{2,3} Wenxin Duan,^{2,3} Yang Liu,^{2,3} Rongchang Chen ^{4,5,6} and Yan Kang ^{1,2,3,7}

¹College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China

²College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China

³School of Applied Technology, Shenzhen University, Shenzhen 518060, China

⁴Shenzhen Institute of Respiratory Diseases, Shenzhen People's Hospital, Shenzhen 518001, China

⁵The Second Clinical Medical College, Jinan University 518001, Guangzhou, China

⁶The First Affiliated Hospital, Southern University of Science and Technology 518001, Shenzhen, China

⁷Engineering Research Centre of Medical Imaging and Intelligent Analysis, Ministry of Education, Shenyang 110169, China

Correspondence should be addressed to Rongchang Chen; chenrc@vip.163.com and Yan Kang; kangyan@sztu.edu.cn

Received 16 May 2022; Revised 2 August 2022; Accepted 25 April 2023; Published 3 November 2023

Academic Editor: Weihua Yang

Copyright © 2023 Yingjian Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Computed tomography (CT) has been regarded as the most effective modality for characterizing and quantifying chronic obstructive pulmonary disease (COPD). Therefore, chest CT images should provide more information for COPD diagnosis, such as COPD stage classification. This paper proposes a features combination strategy by concatenating three-dimension (3D) CNN features and lung radiomics features for COPD stage classification based on the multi-layer perceptron (MLP) classifier. First, 465 sets of chest HRCT images are automatically segmented by a trained ResU-Net, obtaining the lung images with the Hounsfield unit. Second, the 3D CNN features are extracted from the lung region images based on a truncated transfer learning strategy. Then, the lung radiomics features are extracted from the lung region images by PyRadiomics. Third, the MLP classifier with the best classification performance is determined by the 3D CNN features and the lung radiomics features. Finally, the proposed combined feature vector is used to improve the MLP classifier's performance. The results show that compared with CNN models and other ML classifiers, the MLP classifier with the best classification performance is determined. The MLP classifier with the proposed combined feature vector has achieved accuracy, mean precision, mean recall, mean *F1*-score, and AUC of 0.879, 0.879, 0.879, 0.875, and 0.971, respectively. Compared to the MLP classifier with the 3D CNN features selected by Lasso, our method based on the MLP classifier has improved the classification performance by 5.8% (accuracy), 5.3% (mean precision), 5.8% (mean recall), 5.4% (mean *F1*-score), and 2.5% (AUC). Compared to the MLP classifier with lung radiomics features selected by Lasso, our method based on the MLP classifier has improved the classification performance by 5.0% (accuracy), 5.1% (mean precision), 5.0% (mean recall), 5.1% (mean *F1*-score), and 2.1% (AUC). Therefore, it is concluded that our method is effective in improving the classification performance for COPD stage classification.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a common and non-infectious lung disease characterized by persistent airflow limitation [1–3]. Because of this characterization, the COPD stage is diagnosed from stage 0 to IV

according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria accepted by the American Thoracic Society and the European Respiratory Society [4]. GOLD is examined by the pulmonary function test (PFT) and diagnosed by the forced expiratory volume in 1 second/forced vital capacity (FEV1/FVC) and FEV1% predicted [1, 2]. PFT

can explain the impact on symptoms and life quality of COPD patients [5, 6], but it cannot reflect the change of the lung tissue in COPD patients with the COPD stage evolution. PFT changes from normal to abnormal occur when lung tissue is destroyed to a certain extent. Therefore, the PFT makes it challenging to identify the etiology of COPD.

Compared with the GOLD criteria and other imaging equipment, computed tomography (CT) has been regarded as the most effective modality for characterizing and quantifying COPD [7]. Compared with PFT, chest CT images can indicate that the patients have suffered from mild lobular central emphysema and decreased exercise tolerance in smokers without airflow limitation [8]. In addition, the chest CT images are also used to quantitatively analyze the bronchial, airway disease, emphysema, and vascular for COPD patients [7]. However, automatic multi-classification based on convolutional neural networks (CNNs) using chest CT images remains a challenging task for the COPD stage. One main reason is that the number of medical images is limited compared to natural images. In particular, few people seek medical treatment in the early stage of COPD and undergo CT scans simultaneously. Transfer learning [9] may solve the above problems. Since radiomics was proposed to mine more information from medical images using advanced feature analysis in 2007 [10], it has been widely used to analyze lung disease imaging [11–15]. However, radiomics features are extracted from medical images by specific calculation equations, preset types of images, and preset classes, limiting the forms of radiomics features. Some deep features from CNN (CNN features) are also needed to improve the classifier's performance in multi-classification. CNN features extracted from medical images will make up for the limitations of radiomics features.

Radiomics features in COPD develop slower than those in other lung diseases, such as lung cancer and pulmonary nodules. Until 2020, reference [16] points out that radiomics features in COPD have not been extensively investigated yet. Nevertheless, there are potential applications of radiomics features in COPD for the diagnosis, treatment, and follow-up of COPD and future directions [16]. A critical reason limiting the development of radiomics features in COPD is its diffuse distribution in the lung. At the same time, radiomics features need to be extracted from the region of interest (ROI) of the chest CT images. However, the diffuse distribution of COPD makes it difficult to determine ROI. COPD results from the joint action of the peripheral airway, pulmonary parenchyma, and pulmonary vessels [17–19]. Thus, the peripheral airway, pulmonary parenchyma, and pulmonary vessels as ROI to extracting lung radiomics features are reasonable for COPD stage classification.

Currently, radiomics features also have been used in COPD for survival prediction [20, 21], COPD presence prediction [22], COPD exacerbations [23], COPD early decision [4], and analysis of COPD and resting heart rate [3]. However, as mentioned above, lung radiomics features have not been applied in the COPD stage classification. On the other hand, radiomics based on machine learning (ML) and chest CT images based on CNN have been widely and respectively used in COPD and its

evaluation. However, the advantages of radiomics based on machine learning and medical images based on CNN need to be further integrated to improve the performance of COPD stage classification. Therefore, this paper proposes a feature combination strategy by concatenating three-dimension (3D) CNN features and lung radiomics features for COPD stage classification based on the multi-layer perceptron (MLP) classifier. Our contributions in this paper are briefly described as follows. (1) MLP classifier with the best classification performances is determined in the ML classifier for 3D CNN features or lung radiomics features. (2) Truncated transfer learning is proposed from the excellent segmentation model for generating nonlinear 3D CNN features. (3) The proposed feature combination strategy by concatenating 3D CNN features and lung radiomics features effectively improves the MLP classifier's performance.

2. Materials and Methods

2.1. Materials. The participants are enrolled by the national clinical research center of respiratory diseases, China, from May 25, 2009, to January 11, 2011. Finally, 465 Chinese subjects participated in the study after being strictly selected by the inclusion and exclusion criteria [24]. The 465 subjects underwent chest HRCT scans at the full inspiration state. In addition, the 465 subjects also underwent the PFT, and the COPD stage of each subject is diagnosed by PFT in Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria 2008 accepted by the American Thoracic Society and the European Respiratory Society.

Figure 1 shows the COPD stage distribution of the subjects in this study. There are 129, 108, 121, and 107 subjects in each COPD stage (GOLD 0, GOLD I, GOLD II, GOLD III, and GOLD IV). This study was approved by the ethics committee of the national clinical research center for respiratory diseases in China. In addition, all 465 subjects have been provided written informed consent to the first affiliated hospital of Guangzhou medical university before chest HRCT scans and PFT. Refer to our previous study [4] for a more detailed description of the materials.

2.2. Methods. Figure 2 shows the proposed method in this study. The main idea of the proposed method proposed in this paper is to combine 3D CNN features and lung radiomics features for COPD stage classification. When generating the 3D CNN features, we adopt a truncated transfer learning strategy that only intercepts the encoder backbone of the pretrained Med3d [25].

2.2.1. Lung Radiomics Features Extraction. First, 465 sets of chest HRCT images are automatically segmented by a trained ResU-Net [26], obtaining 465 sets of lung images with the Hounsfield unit (Hu) [27]. The lung images include the peripheral airway, pulmonary parenchyma, and pulmonary vessels. The architecture of the ResU-Net has been described in detail in our previous study [28]. Then, lung radiomics features of 465 subjects are extracted from the

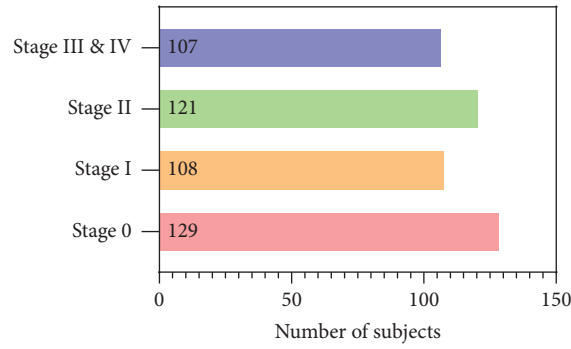


FIGURE 1: COPD stage distribution of the subjects in this study.

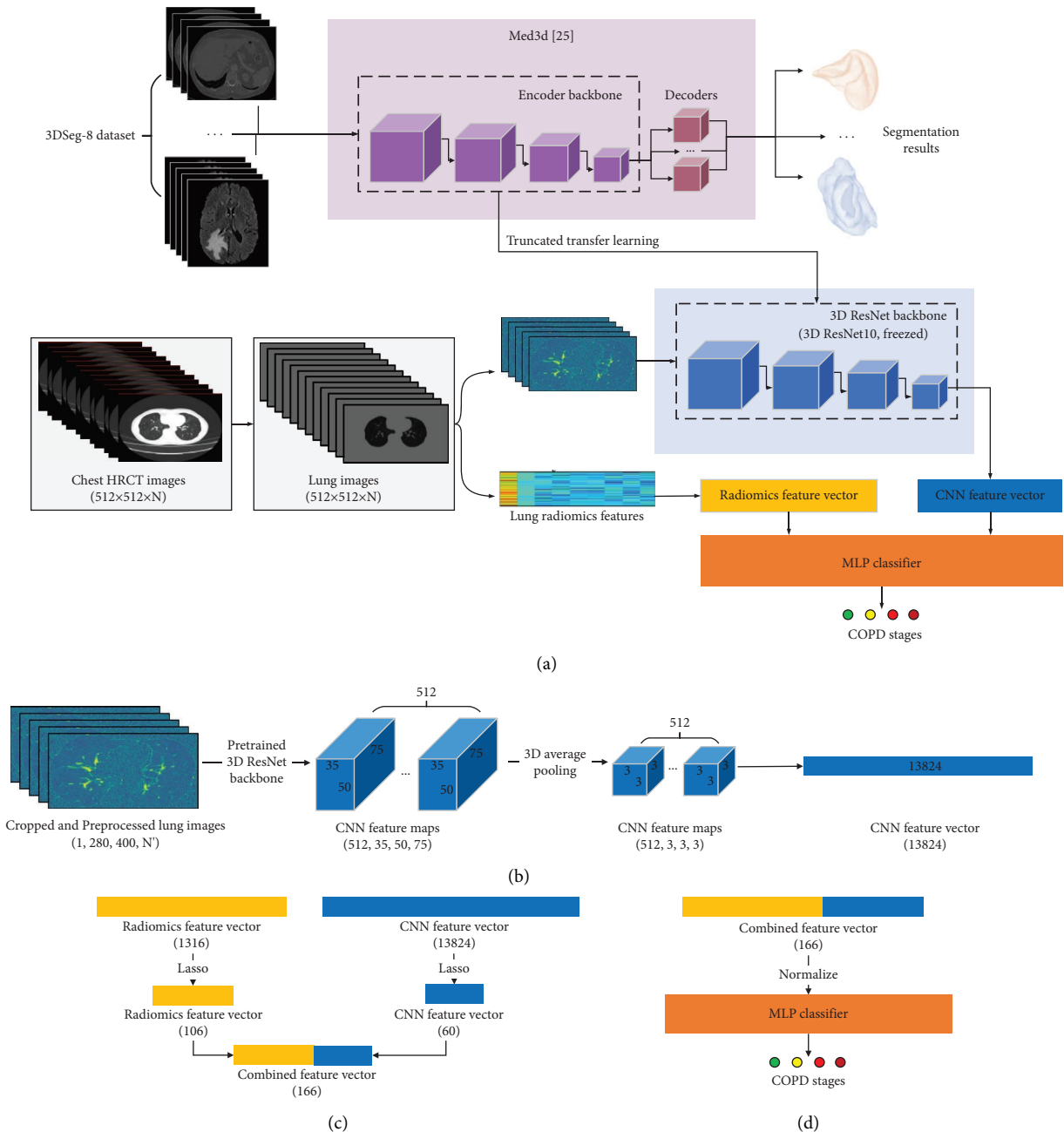


FIGURE 2: The proposed method in this study. (a) A constructed model of 3D CNN and MLP classifier for COPD stage classification. (b) CNN feature vector is generated by transfer learning from Med3D. (c) The combined feature vector is generated by concatenating the CNN feature vector and the radiomics feature vector. (d) The combined feature vector is used to classify the COPD stage based on MLP classifier.

lung images by PyRadiomics [29]. Refer to our previous study [4] for a more detailed description of the lung radiomics feature extraction.

2.2.2. 3D CNN Feature Extraction. A truncated transfer learning strategy is proposed to extract the 3D CNN features based on the pretrained Med3d [25]. Med3d, a heterogeneous 3D network, is used to extract general medical 3D features by building a 3Dseg-8 dataset with diverse modalities, target organs, and pathologies. Thus, we only transfer the encoder backbone of the pretrained Med3d (3D ResNet10) for generating the 3D CNN features, as shown in Figure 2(a).

Figure 2(b) shows that the 465 sets of lung images with Hu are input to the transferring encoder backbone, generating CNN feature vectors in detail. First, the lung images ($512 \times 512 \times N$) are cropped into the size $280 \times 400 \times N'$, retaining the lung region. The non-lung images are also deleted, so N changes into N' ($N' < N$). Second, the cropped lung images are preprocessed by the method in reference [25], normalizing the lung region and generating random values outside the lung region in accordance with Gaussian distribution. Equation (1) shows the mathematical form of normalization:

$$x' = \frac{x - \bar{x}}{\sigma}, \quad (1)$$

where x is the value of the lung region, \bar{x} is the mean value of the lung region, and σ is the mean square deviation of the lung region.

Third, the CNN feature maps ($512 \times 35 \times 50 \times 75$) are generated by the cropped and preprocessed lung images ($1 \times 280 \times 400 \times N'$) and the pretrained Med3d. Last, higher-order CNN feature maps ($512 \times 3 \times 3 \times 3$) need to be extracted from the CNN feature maps ($512 \times 35 \times 50 \times 75$) by 3D average pooling. Then, the higher-order CNN feature maps ($512 \times 3 \times 3 \times 3$) are flattened into the CNN feature vector. Finally, each CNN feature vector (per subject) includes 13824 3D CNN features ($512 \times 3 \times 3 \times 3 = 13824$).

2.2.3. Combined Feature Vector for COPD Classification.

Figure 2(c) shows that the combined feature vector is generated by concatenating the CNN feature vector and the radiomics feature vector. First, the CNN feature vector (13824) and the radiomics feature vector (1316) are selected by the least absolute shrinkage and selection operator (Lasso) [30], respectively. After Lasso, the number of the selected CNN feature vector and the selected radiomics feature vector is 60 and 106, respectively. A standard python package LassoCV, with tenfold cross-validation, is performed in this paper. Equation (2) shows the mathematical form of Lasso [4]:

$$A \leftarrow \arg \min \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=0}^p |\beta_j| \right\}, \quad (2)$$

where matrix A denotes the selected lung radiomics feature. x_{ij} denotes the lung radiomics features (the independent variable). y_i denotes the COPD stage (the independent variable). λ denotes the penalty parameter ($\lambda \geq 0$). β_j denotes the regression coefficient, $i \in [1, n]$, and $j \in [0, p]$.

Second, the combined feature vector is generated by concatenating the selected CNN feature vector and the selected radiomics feature vector. Finally, the combined feature vector is the size 1×166 per subject. Figure 2(d) shows that MLP [31, 32] with the combined feature vector is used to classify the COPD stage in this paper.

2.2.4. Experiments and Evaluation Metrics. Our proposed method uses the combined feature vector of 3D CNN features and lung radiomics features for COPD stage classification based on the MLP classifier. Our experiment includes five experiments in this section to verify the effectiveness of our proposed method.

Figure 3 shows the experimental design in this paper. End-to-end CNN models based on parenchyma images are used for COPD stage classification in experiments 1-2. Specifically, two classic CNN models, DenseNet and GoogleNet, based on parenchyma images, are adopted to compare the classification performance of the six different ML classifiers. The classification performance of DenseNet and GoogleNet has been evaluated by our previous study [33], which achieved the best classification performance for image classification. Furthermore, compared with experiment 1, multiple-instance learning (MIL) [34], a form of weakly supervised learning, is applied in experiment 2. Meanwhile, different ML classifiers based on different feature vectors are also used for COPD stage classification in experiments 3-5.

Specifically, the training parameters of 2D DenseNet and 3D DenseNet are set: 20/2 (batch size (2D/3D)), $512 \times 512 / 512 \times 512 \times 20^*$ (input size (2D/3D)), 50/50 (epoch (2D/3D)), and 0.5/0.2 (drop rate (2D/3D)) in experiment 1. The training parameters of 2D GoogleNet and 3D GoogleNet are set: 16/2 (batch size (2D/3D)), $512 \times 512 / 512 \times 512 \times 20^*$ (input size (2D/3D)), 50/50 (epoch (2D/3D)), and 0.2/0.2 (drop rate (2D/3D)) in experiment 1. *MIL: each case (a set of chest HRCT images) was equally divided into 20 segments, with one slice taken equidistantly to obtain 20 slices in each case. The training parameters of 2D DenseNet with MIL (2D DenseNet_MIL) and 3D DenseNet with MIL (3D DenseNet_MIL) are set: 16/2 (batch size (2D/3D)), $512 \times 512^{**} / 512 \times 512 \times 512 \times 16^{***}$ (input size (2D/3D)), 50/50 (epoch (2D/3D)), and 0.5/0.2 (drop rate (2D/3D)) in experiment 2. The training parameters of 2D GoogleNet with MIL (2D GoogleNet_MIL) and 3D GoogleNet with MIL (3D GoogleNet_MIL) are set: 16/2 (batch size (2D/3D)), $512 \times 512^{**} / 512 \times 512 \times 16^{***}$ (input size (2D/3D)), 50/50 (epoch (2D/3D)), and 0.2/0.2 (drop rate (2D/3D)) in experiment 2. **MIL: each case was equally divided into 10 bags, with one slice taken randomly to obtain 10 slices in each case. ***MIL: each case was equally divided into 16 bags, with one slice taken equidistantly to obtain 16 slices in each case.

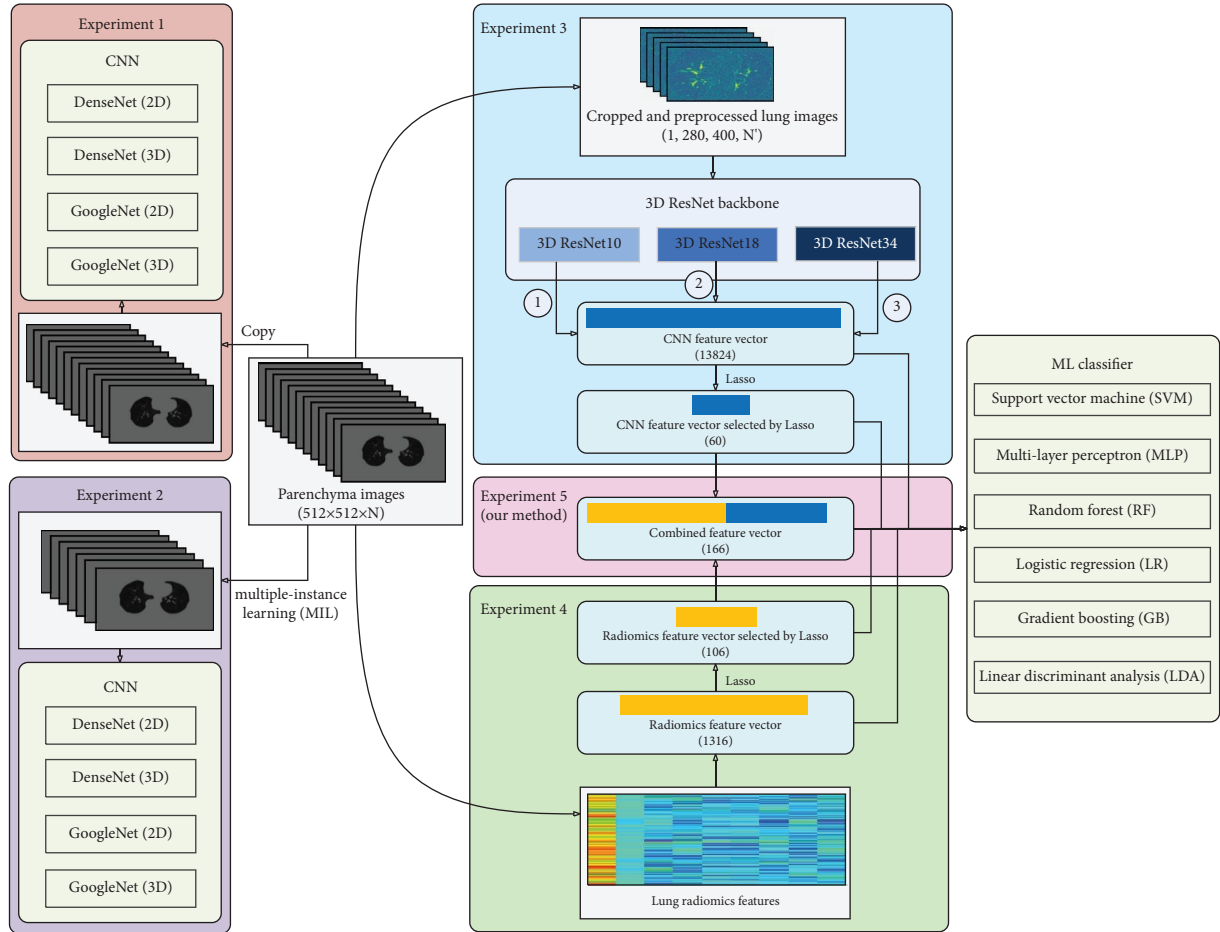


FIGURE 3: Experimental design in this paper.

Specifically, experiments 3–5 are designed to compare the classification performance of the six different classifiers based on the CNN feature vector (13824), radiomics feature vector (1316), their selected feature vector by Lasso, and the proposed combined feature vector (166), respectively. First, based on 3D ResNet10, we use six classic classifiers (SVM [35], MLP, RF [36], LR [37], GB [38], and LDA [39]) to determine the best COPD classification classifier by different feature vectors. Table 1 reports the six different classifiers with their definitions in this paper. The different feature vectors include the CNN feature vector (13824), CNN feature vector selected by Lasso (60), radiomics feature vector (1316), and radiomics feature vector selected by Lasso (106). The MLP classifier with the best classification performance is determined. Second, we further verify the proposed combined feature vector (166) to improve the MLP classifier's performance. Third, 3D ResNet18 and 3D ResNet34 are also transferred to generate the CNN feature vector, and the 3D ResNet10 is determined as the encoder backbone with the best performance on the MLP classifier. The 465 subjects are divided into the train set (70%) and the test set (30%). Figure 4 shows the detailed dataset division for training and test set in each COPD stage.

Standard evaluation metrics of the CNN and ML models include the accuracy, precision, recall, $F1$ -score, and area under the curve (AUC). The above standard evaluation metrics are defined as in equations (3)–(6). The evaluation metric AUC for multi-classification is calculated by the receiver operating characteristic curve (ROC) [40].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (5)$$

$$F1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

where the true positive (TP) and false positive (FP), respectively, represent the positive and negative samples classified to be positive by the CNN and ML models and the true negative (TN) and false negative (FN), respectively,

TABLE 1: The different classifiers with their definitions.

Classifier	Model definition in Python 3.6
SVM	<code>SVM sklearn.svm.SVC(kernel = "rbf",probability = true)</code>
MLP	<code>sklearn.neural_network.MLPClassifier (hidden_layer_sizes = (400, 100), alpha = 0.01, max_iter = 10000)</code>
RF	<code>sklearn.ensemble.RandomForestClassifier(n_estimators = 200)</code>
LR	<code>sklearn.linear_model.logisticRegressionCV(max_iter = 100000, solver = "liblinear")</code>
GB	<code>sklearn.ensemble.GradientBoostingClassifier()</code>
LDA	<code>sklearn.discriminant_analysis.()</code>

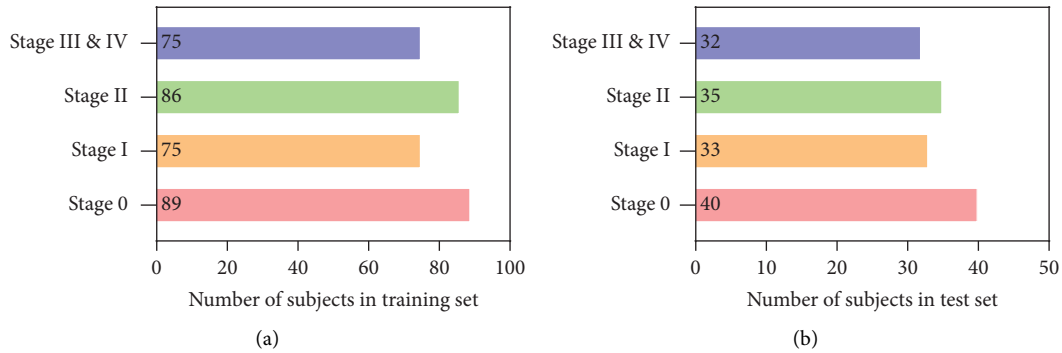


FIGURE 4: Dataset division in this paper. (a) Training set. (b) Test set.

represent the positive and negative samples classified to be negative by the CNN and ML models.

3. Results

This section reports the experimental results, including (1) the classification performance of the parenchyma images based on the DenseNet and GoogleNet; (2) the classification performance of the CNN feature vector and lung radiomics vector based on different classifiers; (3) the MLP classifier's performance with the combined feature vector; and (4) the MLP classifier's performance with combined feature vector based on different 3D ResNet.

3.1. The DenseNet and GoogleNet's Performance with Parenchyma Images. This section shows the classification performance of 2D/3D DenseNet, 2D/3D GoogleNet, 2D/3D DenseNet_MIL, and 2D/3D GoogleNet_MIL based on the parenchyma images, respectively.

Figure 5 intuitively shows the AUC of the CNN models by drawing the ROC curves. Tables 2 and 3 report the classification performance of CNN models. Specifically, Table 2 shows that 2D GoogleNet with parenchyma images performs the best in 2D CNN models, achieving 0.550 (accuracy), 0.562 (mean precision), 0.550 (mean recall), 0.553 (mean $F1$ -score), and 0.809 (AUC). In addition, Table 3 shows that 3D DenseNet with parenchyma images performs the best in 3D CNN models, achieving 0.579 (accuracy), 0.614 (mean precision), 0.579 (mean recall), 0.579 (mean $F1$ -score), and 0.787 (AUC).

3.2. The Classification Performance of CNN Feature Vector and Lung Radiomics Vector Based on Different Classifiers. This section shows the classification performance of the CNN feature vector (13824), the CNN feature vector selected by Lasso (60), the lung radiomics vector (1316), and the lung radiomics vector selected by Lasso (106) based on different classifiers, respectively.

Figure 6 intuitively shows the AUC of the different classifiers by drawing the ROC curves. Tables 4–7 show that the MLP classifier is the best classifier for COPD stage classification. Specifically, Table 4 reports the classification performance of the different classifiers with the CNN feature vector (13824), respectively. The best classifier is the MLP classifier with 0.793 (accuracy), 0.798 (mean precision), 0.793 (mean recall), 0.790 (mean $F1$ -score), and 0.790 (AUC), respectively. Table 5 reports that the classification performance of the MLP classifier with the CNN feature vector selected by Lasso has improved with 0.821 (accuracy), 0.826 (mean precision), 0.821 (mean recall), 0.821 (mean $F1$ -score), and 0.946 (AUC), respectively. Table 6 reports that the classification performance of the MLP classifier with the radiomics feature vector selected by Lasso has improved with 0.786 (accuracy), 0.784 (mean precision), 0.784 (mean recall), 0.784 (mean $F1$ -score), and 0.919 (AUC), respectively. Table 7 reports that the classification performance of the MLP classifier with the radiomics feature vector selected by Lasso has improved with 0.829 (accuracy), 0.828 (mean precision), 0.829 (mean recall), 0.824 (mean $F1$ -score), and 0.950 (AUC), respectively.

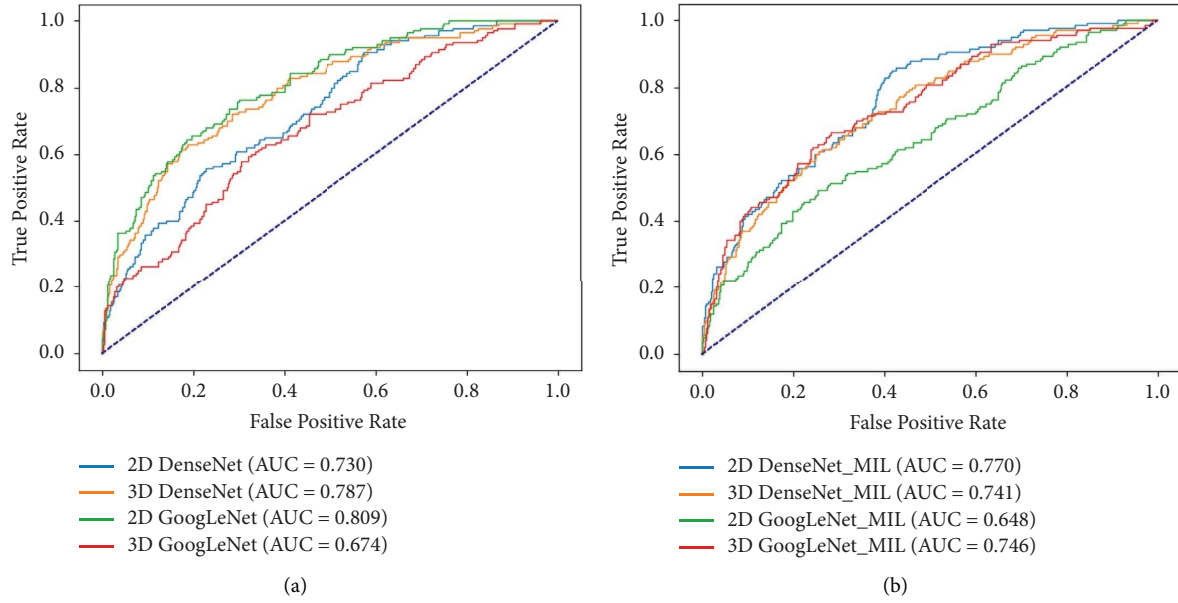


FIGURE 5: The ROC curves of the CNN model's performance with parenchyma images. (a) The ROC curves of the 2D/3D DenseNet and 2D/3D GoogLeNet. (b) The ROC curves of the 2D/3D DenseNet_MIL and 2D/3D GoogLeNet_MIL.

TABLE 2: The 2D DenseNet and 2D GoogLeNet's performance with parenchyma images in experiments 1 and 2.

CNN model	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
2D DenseNet	0.471	0.800/0.311/0.000/0.720/ (0.466)	0.500/0.848/0.000/0.562/ (0.471)	0.615/0.455/0.000/0.632/ (0.428)	0.730
2D GoogLeNet	0.550	0.788/0.419/0.385/0.622/ (0.562)	0.650/0.394/0.429/0.719/ (0.550)	0.712/0.406/0.405/0.667/ (0.553)	0.809
2D DenseNet_MIL	0.493	0.538/0.318/0.333/0.720/ (0.477)	0.875/0.424/0.057/0.562/ (0.493)	0.667/0.364/0.098/0.632/ (0.445)	0.770
2D GoogLeNet_MIL	0.414	0.418/0.444/0.368/1.000/ (0.545)	0.950/0.121/0.400/0.062/ (0.414)	0.580/0.190/0.384/0.118/ (0.333)	0.648

TABLE 3: The 3D DenseNet and 3D GoogLeNet's performance with parenchyma images in experiments 1 and 2.

CNN model	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
3D DenseNet	0.579	0.571/0.429/0.533/0.947/ (0.614)	0.800/0.455 0.457/0.562/ (0.579)	0.667/0.441/0.492/0.706/ (0.579)	0.787
3D GoogLeNet	0.393	0.463/0.333/0.279/0.833/ (0.471)	0.775/0.061/0.486/0.156/ (0.393)	0.579/0.103/0.354/0.263/ (0.338)	0.674
3D DenseNet_MIL	0.500	0.500/0.600/0.408/0.900/ (0.592)	0.950/0.091/0.571/0.281/ (0.500)	0.655/0.158/0.476/0.429/ (0.441)	0.741
3D GoogLeNet_MIL	0.486	0.471/0.413/0.200/0.789/ (0.463)	0.825/0.576/0.029/0.469/ (0.486)	0.600/0.481/0.050/0.588/ (0.432)	0.746

Table 5 also reports that Lasso only plays a role in improving the classification performance of the MLP classifier with the CNN feature vector. It does not improve the classification performance of other classifiers with the CNN feature vector. However, Table 7 reports that Lasso does play a role in improving the classification performance of all classifiers with the radiomics feature vector.

3.3. The MLP Classifier's Performance with Combined Feature Vectors. The best MLP classifier is determined with the CNN feature vector selected by Lasso (60) or the lung radiomics vector selected by Lasso (106) by Section 3.1. This section shows the classification performance of the MLP classifier with combined feature vectors.

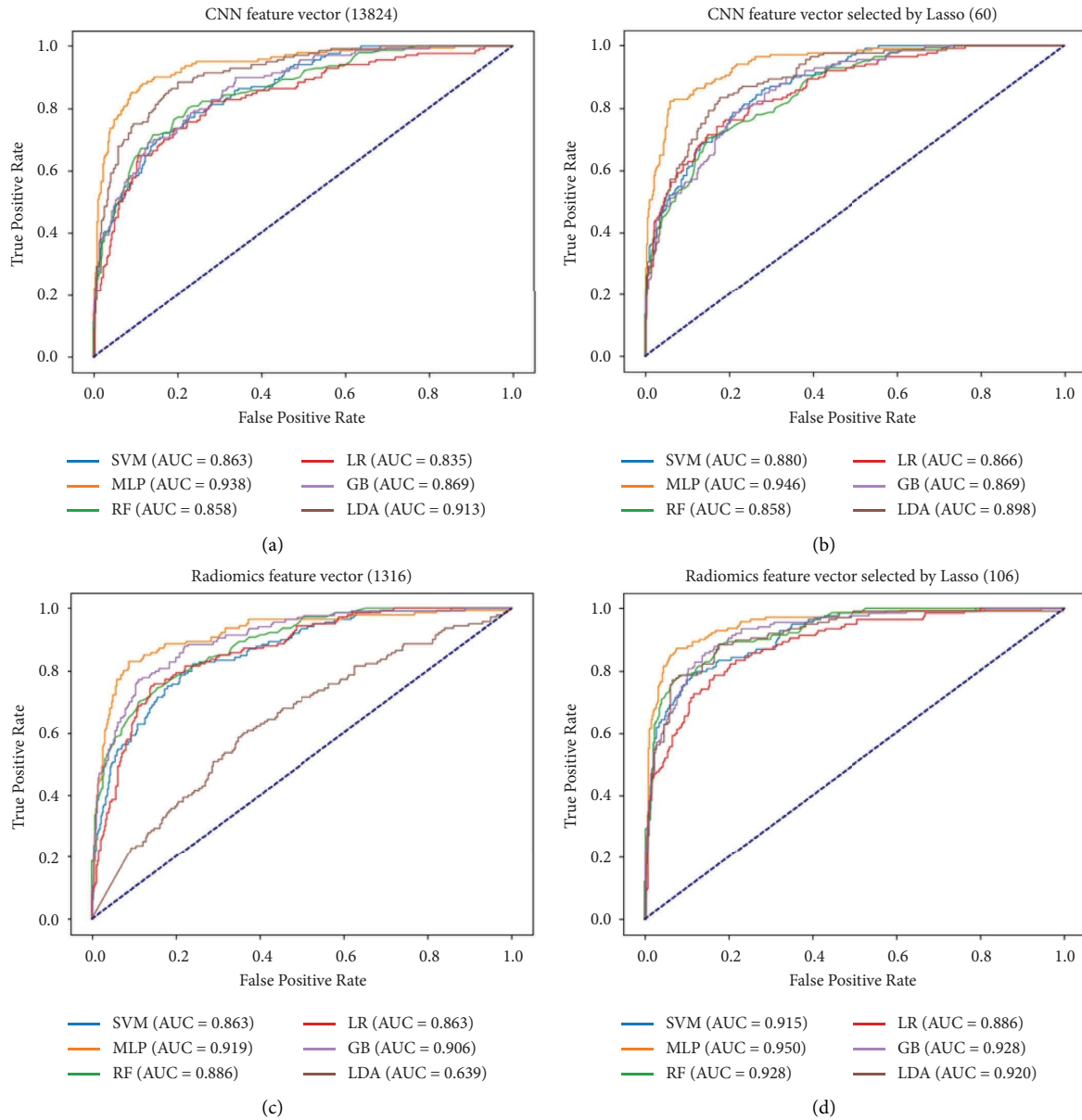


FIGURE 6: The ROC curves of the CNN feature vector and lung radiomics vector are based on different classifiers. (a) The ROC curves of the CNN feature vector (13824). (b) The ROC curves of the CNN feature vector selected by Lasso (60). (c) The ROC curves of the lung radiomics vector (1316). (d) The ROC curves of the lung radiomics vector selected by Lasso (106).

TABLE 4: The different classifiers’ performances based on CNN feature vector (13824) in experiment 3.

Classifier	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
SVM	0.629	0.763/0.556/0.514/0.690/(0.635)	0.725/0.606/0.543/0.625/(0.629)	0.744/0.580/0.528/0.656/(0.631)	0.863
MLP	0.793	0.829/0.815/0.806/0.732/(0.798)	0.850/0.667/0.714/0.938/(0.793)	0.840/0.733/0.758/0.822/(0.790)	0.938
RF	0.657	0.711/0.621/0.600/0.667/(0.652)	0.800/0.545/0.514/0.750/(0.657)	0.753/0.581/0.554/0.706/(0.652)	0.858
LR	0.650	0.689/0.621/0.630/0.641/(0.647)	0.775/0.545/0.486/0.781/(0.650)	0.729/0.581/0.548/0.704/(0.643)	0.835
GB	0.643	0.750/0.500/0.548/0.767/(0.644)	0.750/0.424/0.657/0.719/(0.643)	0.750/0.459/0.597/0.742/(0.641)	0.869
LDA	0.721	0.857/0.625/0.632/0.771/(0.726)	0.750/0.606/0.686/0.844/(0.721)	0.800/0.615/0.658/0.806/(0.722)	0.913

Figure 7 intuitively shows the confusion matrix and ROC curves of the MLP classifier with different feature vectors based on 3D ResNet10. The MLP classifier’s

performance with different feature vectors reported in Table 8 can be calculated from the confusion matrix. Table 8 reports that the proposed combined feature

TABLE 5: The different classifiers' performances based on CNN feature vector selected by Lasso (60) in experiment 3.

Classifier	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
SVM	0.629	0.811/0.450/0.552/0.706/(0.637)	0.750/0.545/0.457/0.750/(0.629)	0.779/0.493/0.500/0.727/(0.630)	0.880
MLP	0.821	0.919/0.722/0.833/0.811/(0.826)	0.850/0.788/0.714/0.938/(0.821)	0.883/0.754/0.769/0.870/(0.821)	0.946
RF	0.600	0.638/0.480/0.594/0.639/(0.590)	0.750/0.364/0.543/0.719/(0.600)	0.690/0.414/0.567/0.676/(0.591)	0.858
LR	0.650	0.714/0.500/0.538/0.771/(0.633)	0.875/0.455/0.400/0.844/(0.650)	0.787/0.476/0.459/0.806/(0.636)	0.866
GB	0.600	0.714/0.395/0.538/0.793/(0.613)	0.750/0.515/0.400/0.719/(0.600)	0.732/0.447/0.459/0.754/(0.602)	0.869
LDA	0.657	0.771/0.526/0.541/0.833/(0.670)	0.675/0.606/0.571/0.781/(0.657)	0.720/0.563/0.556/0.806/(0.662)	0.898

TABLE 6: The different classifiers' performances based on radiomics feature vector (1316) in experiment 4.

Classifier	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
SVM	0.643	0.784/0.514/0.514/0.793/(0.655)	0.725/0.576/0.543/0.719/(0.643)	0.753/0.543/0.528/0.754/(0.647)	0.863
MLP	0.786	0.857/0.731/0.692/0.848/(0.784)	0.900/0.576/0.771/0.875/(0.786)	0.878/0.644/0.730/0.862/(0.782)	0.919
RF	0.664	0.762/0.586/0.561/0.750/(0.668)	0.800/0.515/0.657/0.656/(0.664)	0.780/0.548/0.605/0.700/(0.664)	0.886
LR	0.679	0.850/0.567/0.564/0.710/(0.680)	0.850/0.515/0.629/0.688/(0.679)	0.850/0.540/0.595/0.698/(0.678)	0.863
GB	0.729	0.795/0.724/0.690/0.684/(0.727)	0.875/0.636/0.571/0.812/(0.729)	0.833/0.677/0.625/0.743/(0.724)	0.906
LDA	0.379	0.357/0.278/0.407/0.548/(0.395)	0.250/0.455/0.314/0.531/(0.379)	0.294/0.345/0.355/0.540/(0.377)	0.639

TABLE 7: The different classifiers' performances based on the radiomics feature vector selected by Lasso (106) in experiment 4.

Classifier	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
SVM	0.736	0.816/0.606/0.694/0.818/(0.737)	0.775/0.606/0.714/0.844/(0.736)	0.795/0.606/0.704/0.831/(0.736)	0.915
MLP	0.829	0.864/0.840/0.750/0.857/(0.828)	0.950/0.636/0.771/0.938/(0.829)	0.905/0.724/0.761/0.896/(0.824)	0.950
RF	0.786	0.809/0.750/0.774/0.794/(0.783)	0.950/0.636/0.686/0.844/(0.786)	0.874/0.689/0.727/0.818/(0.781)	0.928
LR	0.693	0.800/0.667/0.630/0.636/(0.689)	0.900/0.485/0.486/0.875/(0.693)	0.847/0.561/0.548/0.737/(0.680)	0.886
GB	0.736	0.766/0.708/0.686/0.765/(0.732)	0.900/0.515/0.686/0.812/(0.736)	0.828/0.596/0.686/0.788/(0.729)	0.928
LDA	0.786	0.829/0.706/0.774/0.824/(0.785)	0.850/0.727/0.686/0.875/(0.786)	0.840/0.716/0.727/0.848/(0.784)	0.920

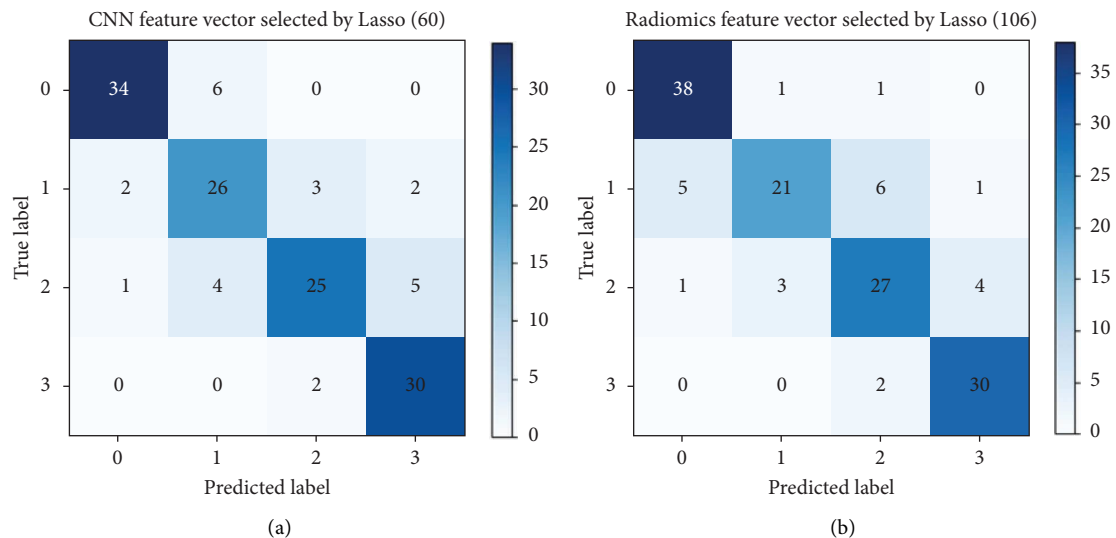


FIGURE 7: Continued.

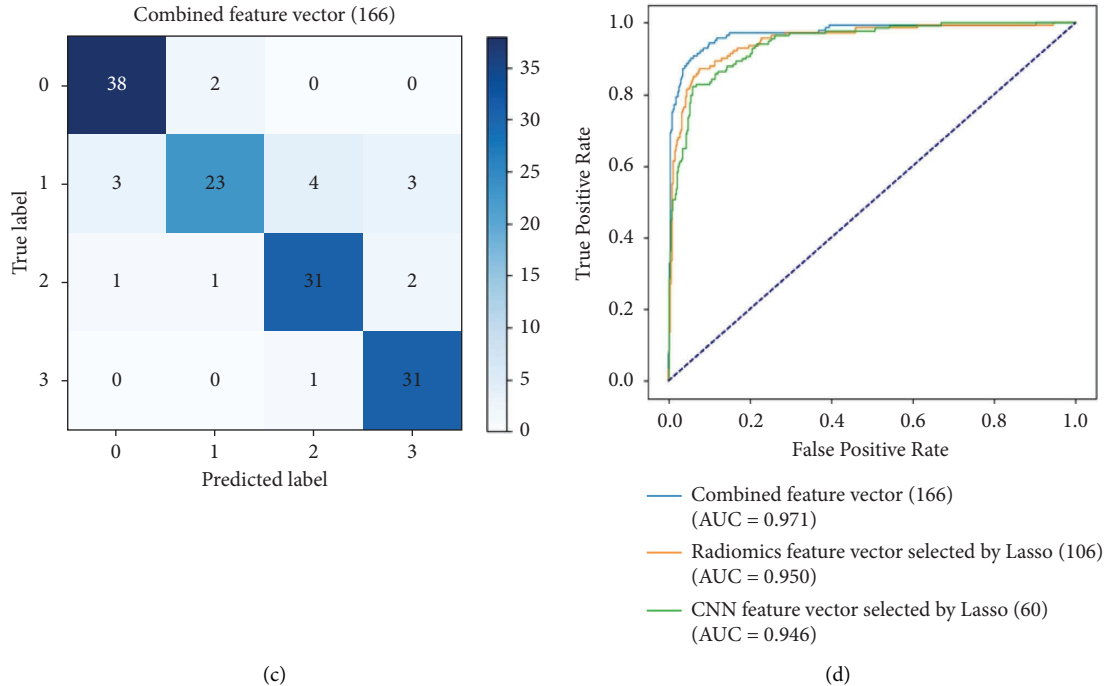


FIGURE 7: The confusion matrix and ROC curves of the MLP classifier with different feature vectors based on 3D ResNet10. (a) The confusion matrix of the MLP classifier with CNN feature vector selected by Lasso (60). (b) The confusion matrix of the MLP classifier with radiomics feature vector selected by Lasso (106). (c) The confusion matrix of the MLP classifier with combined feature vector (166). (d) The ROC curves of the MLP classifier with these feature vectors.

vectors improve the MLP classifier's performance, achieving 0.879 (accuracy), 0.879 (mean precision), 0.879 (mean recall), 0.875 (mean F1-score), and 0.971 (AUC), respectively.

3.4. The MLP Classifier's Performance with Combined Feature Vector Based on Different 3D ResNet. The best MLP classifier is determined with the CNN feature vector selected by Lasso (60) or the lung radiomics vector selected by Lasso (106) by Section 3.1. This section shows the classification performance of the MLP classifier with combined feature vectors.

Figure 8 intuitively shows the confusion matrix and ROC curves of the MLP classifier with combined feature vectors based on different 3D ResNet. The MLP classifier's performance with combined feature vectors based on different 3D ResNet reported in Table 7 can be calculated from the confusion matrix. Table 9 reports that the MLP classifier with combined feature vectors based on 3D ResNet10 achieves the best classification performance.

4. Discussion

This paper proposes a features combination strategy by concatenating 3D CNN features and lung radiomics features for COPD stage classification based on the MLP classifier. Three sections are discussed in this section, and we also point out the limitations in this study and the future direction.

First, 2D GoogleNet with parenchyma images performs the best in 2D CNN models. The main reason is that 2D

GoogleNet is designed for 2D natural image classification (RGB images). Therefore, it achieves the best classification performance in 2D parenchyma images. Meanwhile, because of the ability to extract interlayer information, 3D DenseNet with parenchyma images performs the best classification in 3D CNN models. However, CNN models with parenchyma images fail to classify the COPD stage. One main reason is that the chest HRCT image cannot fully reflect COPD's characteristics for the CNN models. Specifically, the gold standard of COPD classification is characterized by airflow restriction with a slight difference in the chest HRCT image. The slight difference in COPD is mainly caused by small airway disease with an airway diameter < 2 mm [17]. Because of the limitation of HRCT resolution, the above differential features of the small airway will be further blurred in the chest HRCT image. Another reason is that chest HRCT images can reflect the COPD anatomical characteristics, but COPD patients are with high heterogeneity and different phenotypes [1]. The heterogeneity and different phenotypes often result in different features of the chest HRCT images in the same stage. Therefore, it is hard for CNN models to learn specific COPD characteristics, resulting in bad classification performance. At the same time, a set of standard medical images is not as easy to obtain as natural images, and the number of chest HRCT images also restricts CNN models for COPD stage classification. Therefore, compared with CNN models, the ML classifier can realize the COPD stage classification with a small number of samples. This paper determines the MLP classifier with 3D CNN features or lung radiomics features, which performs the best for COPD stage

TABLE 8: The MLP classifier's performance with different feature vectors in experiment 5.

Feature vectors	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV mean))	AUC
CNN feature vector selected by Lasso (60)	0.821	0.919/0.722/0.833/0.811/(0.826)	0.850/0.788/0.714/0.938/(0.821)	0.883/0.754/0.769/0.870/(0.821)	0.946
Radiomics feature vector selected by Lasso (106)	0.829	0.864/0.840/0.750/0.857/(0.828)	0.950/0.636/0.771/0.938/(0.829)	0.905/0.724/0.761/0.896/(0.824)	0.950
Combined feature vector (166)	0.879	0.905/0.885/0.861/0.861/(0.879)	0.950/0.697/0.886/0.969/(0.879)	0.927/0.780/0.873/0.912/(0.875)	0.971

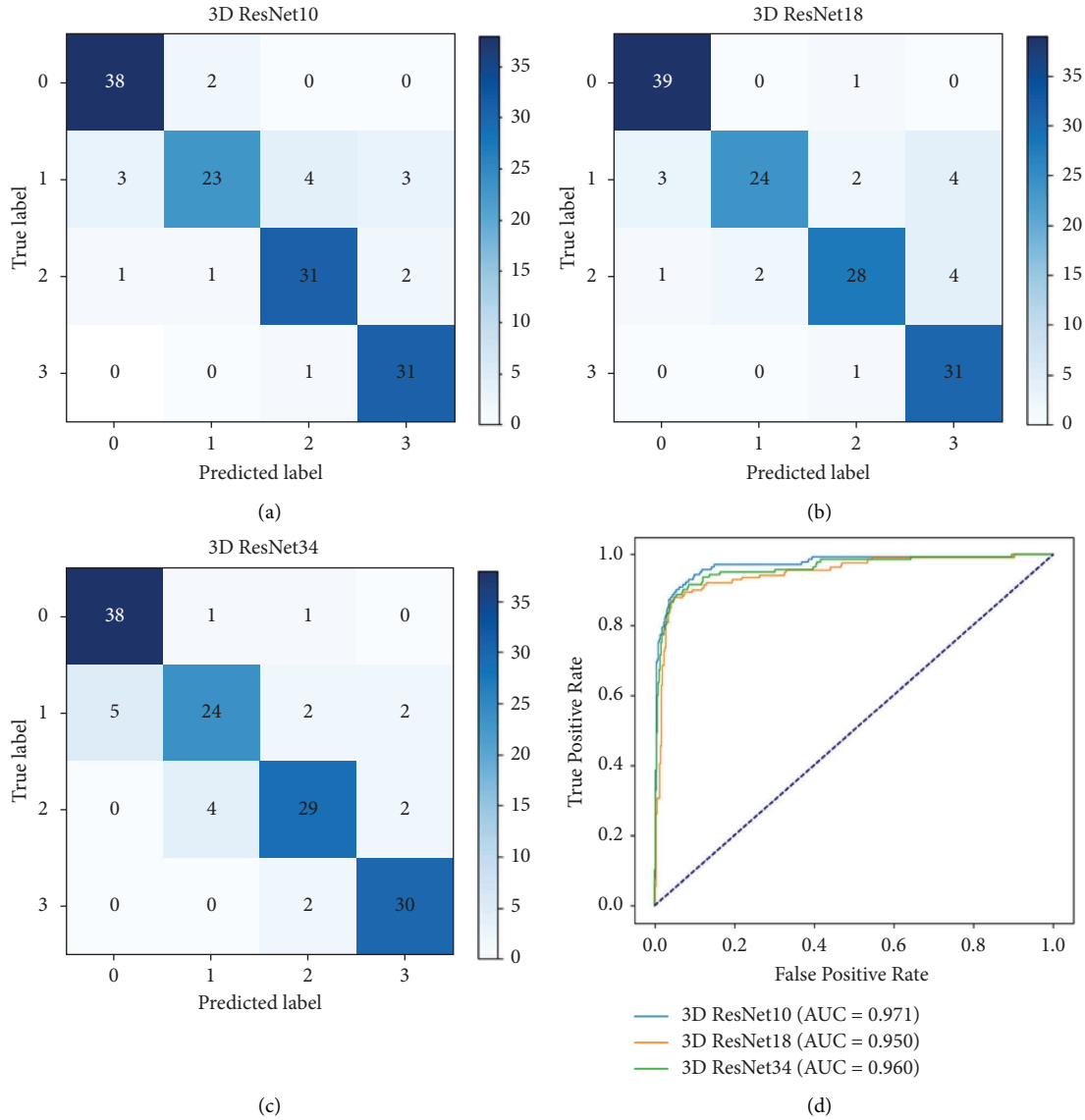


FIGURE 8: The confusion matrix and ROC curves of the MLP classifier with combined feature vectors based on different 3D ResNet. (a) The confusion matrix of the MLP classifier with combined feature vector based on 3D ResNet10. (b) The confusion matrix of the MLP classifier with combined feature vector based on 3D ResNet18. (c) The confusion matrix of the MLP classifier with combined feature vector based on 3D ResNet34. (d) The ROC curves of the MLP classifier with combined feature vectors based on 3D ResNet.

TABLE 9: 3D ResNet's performance based on MLP classifier with the combined feature vector (166).

3D ResNet	Accuracy	Precision (GOLD 0/I/II/III&IV (mean))	Recall (GOLD 0/I/II/III&IV (mean))	F1-score (GOLD 0/I/II/III&IV (mean))	AUC
3D ResNet10	0.879	0.905/0.885/0.861/0.861/(0.879)	0.950/0.697/0.886/0.969/(0.879)	0.927/0.780/0.873/0.912/(0.875)	0.971
3D ResNet18	0.871	0.907/0.923/0.875/0.795/(0.877)	0.975/0.727/0.800/0.969/(0.871)	0.940/0.814/0.836/0.873/(0.869)	0.950
3D ResNet34	0.864	0.884/0.828/0.853/0.882/(0.862)	0.950/0.727/0.829/0.938/(0.864)	0.916/0.774/0.841/0.909/(0.862)	0.960

classification. In addition, compared with the convolution layer in the CNN models, the MLP classifier is composed of three full connection layers, which is more efficient and more suitable for modeling long-range dependencies. The MLP classifier also can handle complex nonlinear features and discover dependencies between different input features compared with other classifiers [31, 32]. Meanwhile, the

objective evaluation of the COPD stage is only the degree of airflow limitation tested by GOLD criteria [1, 2, 4]. COPD is a heterogeneous disease [41], resulting in differences in features (3D CNN features or lung radiomics features extracted from chest HRCT images) with the same degree of airflow limitation. Therefore, a nonlinear relationship exists between 3D CNN features or lung radiomics features and

the COPD stage. Because of this, the MLP classifier is suitable for classifying the COPD stage and has achieved an excellent result in COPD stage classification.

Second, Lasso can improve the classification performance of the MLP classifier with the 3D CNN features and the lung radiomics features. Lasso is often used with survival analysis models to determine variables and eliminate the collinearity problem between variables [30, 42]. The results show that Lasso also can improve the MLP classifier's classification performance by establishing the relationship between the independent variables (3D CNN features or lung radiomics features extracted from chest HRCT images) and dependent variables (the COPD stages). Furthermore, Lasso selects 3D CNN features or lung radiomics features related to COPD stages to reduce the complexity of the MLP classifiers and avoid overfitting [43]. While reducing the complexity of the MLP classifiers, the MLP classifiers can focus on the selected lung radiomics features (the radiomics feature vector selected by Lasso) or the selected 3D CNN features (the CNN feature vector selected by Lasso) and improve the classification performance. From the results of the Lasso, the number of the CNN feature vector selected by Lasso is 60, and that of the radiomics feature vector selected by Lasso is 106. We are surprised that the number of collinearity features in the CNN feature vector is more than that in the radiomics feature vector. This further shows that feature selection of 3D CNN features or the radiomics features is necessary for the COPD stage classification, especially in clinical applications.

Third, the proposed feature combination strategy can further improve the classification performance of the MLP classifier. This paper does not improve the existing classic classifiers and starts with the classification features to enhance the classifier's performance. Many nonlinear classification features, the 3D CNN features, are obtained by a truncated transfer learning strategy. We concatenate the CNN feature vector and the radiomics feature vector for the COPD stage classification, which improves the MLP classifier's performance. The MLP classifier is good at handling complex nonlinear features by itself [31, 32]. Therefore, based on the radiomics feature vector, we add the nonlinear CNN feature vector to the radiomics feature vector, generating a combined feature vector. The combined feature vector with the nonlinear CNN feature vector enhances the MLP classifier's performance. Therefore, this fits the essence of the MLP classifier and is interpretable [44]. The selected encoder backbone of the pretrained Med3D is also directly related to the classification performance. Compared with the MLP classifier with 3D ResNet18 or 3D ResNet34, the MLP classifier with 3D ResNet10 performs the best, consistent with the results of multi-class segmentation task (left lung, right lung, and background) in reference [25].

Finally, this study has some limitations, and we point out the future direction. First, from the materials used in this study, there are not enough cases at the COPD stages III and IV. Second, the existing classic classifiers are not improved. Third, the classification performance of the ML classifier with the 3D CNN features is also limited by the encoder backbone of the pretrained Med3d. In our future work, the

recent networks, an auto-metric graph neural network [45], will be further attempted and modified for COPD stage classification based on the 3D CNN features and/or the lung radiomics features.

5. Conclusions

This paper proposes a feature combination strategy by concatenating 3D CNN features and lung radiomics features for COPD stage classification based on the MLP classifier. First, the 3D CNN features are extracted from the lung region images based on a truncated transfer learning strategy. Then, the lung radiomics features are extracted from the lung region images by PyRadiomics. Compared with CNN models and other ML classifiers, the MLP classifier with the best classification performance is determined by the 3D CNN features and the lung radiomics features. Lasso plays a role in improving the classification performance of the MLP classifier with the CNN feature vector and the radiomics feature vector. The proposed combined feature vector also improves the MLP classifier's performance. The MLP classifier with the proposed combined feature vector has accuracy, mean precision, mean recall, mean *F1*-score, and AUC of 0.879, 0.879, 0.879, 0.875, and 0.971, respectively. This shows that our method effectively improves the classification performance for COPD stage classification.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding authors on reasonable request.

Disclosure

Yingjian Yang and Nanrong Zeng are co-first authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yingjian Yang and Nanrong Zeng contributed equally to this work.

Acknowledgments

Thanks are due to the Department of Radiology, the First Affiliated Hospital of Guangzhou Medical University, for providing the dataset. This research was funded by the National Natural Science Foundation of China (grant no. 62071311), Stable Support Plan for Colleges and Universities in Shenzhen of China (grant no. SZWD2021010), Scientific Research Fund of Liaoning Province of China (grant no. JL201919), Natural Science Foundation of Guangdong Province of China (grant no. 2019A1515011382), and Special Program for Key Fields of Colleges and Universities in

Guangdong Province (Biomedicine and Health) of China (grant no. 2021ZDZX2008).

References

- [1] D. Singh, A. Agusti, A. Anzueto et al., "Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019," *European Respiratory Journal*, vol. 53, no. 5, Article ID 1900164, 2019.
- [2] M. C. Matheson, G. Bowatte, J. L. Perret et al., "Prediction models for the development of COPD: a systematic review," *International Journal of Chronic Obstructive Pulmonary Disease*, vol. 13, pp. 1927–1935, 2018.
- [3] Y. Yang, W. Li, Y. Kang et al., "A novel lung radiomics feature for characterizing resting heart rate and COPD stage evolution based on radiomics feature combination strategy," *Mathematical Biosciences and Engineering*, vol. 19, no. 4, pp. 4145–4165, 2022.
- [4] Y. Yang, W. Li, Y. Guo et al., "Early COPD risk decision for adults aged from 40 to 79 Years based on lung radiomics features," *Frontiers of Medicine*, vol. 9, Article ID 845286, 2022.
- [5] P. W. Jones, "Health status measurement in chronic obstructive pulmonary disease," *Thorax*, vol. 56, 2001.
- [6] C. D. Brown, J. O. Benditt, F. C. Sciruba et al., "Exercise testing in severe emphysema: association with quality of life and lung function," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 5, no. 2, pp. 117–124, 2008.
- [7] D. A. Lynch, "Progress in imaging COPD, 2004–2014," *Chronic Obstructive Pulmonary Diseases Journal of the Copd Foundation*, vol. 1, no. 1, pp. 73–82, 2014.
- [8] P. J. Castaldi, R. San José Estépar, C. S. Mendoza et al., "Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers," *American Journal of Respiratory and Critical Care Medicine*, vol. 188, no. 9, pp. 1083–1090, 2013.
- [9] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "Finding a moral compass without a compass: evolution and ethics," *Evolutionary Anthropology*, vol. 25, no. 1, pp. 1–5, 2016.
- [10] P. Lambin, E. Rios-Velazquez, and R. Leijenaar, "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 43, no. 4, pp. 441–446, 2007.
- [11] A. N. Frix, F. Cousin, T. Refaee et al., "Radiomics in lung diseases imaging: state-of-the-art for clinicians," *Journal of Personalized Medicine*, vol. 11, no. 7, p. 602, 2021.
- [12] S. M. Rezaei, R. Abedi-Firouzjah, and M. Ghorvei, "Screening of COVID-19 based on the extracted radiomics features from chest CT images," *Journal of X-Ray Science and Technology*, vol. 29, no. 4, pp. 1–15, 2021.
- [13] F. Xiao, R. Sun, W. Sun et al., "Radiomics analysis of chest CT to predict the overall survival for the severe patients of COVID-19 pneumonia," *Physics in Medicine and Biology*, vol. 66, no. 10, 2021.
- [14] F. Xiong, Y. Wang, T. You et al., "The clinical classification of patients with COVID-19 pneumonia was predicted by Radiomics using chest CT," *Medicine*, vol. 100, no. 12, Article ID e25307, 2021.
- [15] M. Tamal, M. Alshammari, M. Alabdullah, R. Hourani, H. A. Alola, and T. M. Hegazi, "An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from Chest X-ray," *Expert Systems with Applications*, vol. 180, Article ID 115152, 2021.
- [16] T. Refaee, G. Wu, and A. Ibrahim, "The emerging role of radiomics in COPD and lung cancer," *Respiration*, vol. 99, no. 2, pp. 1–9, 2020.
- [17] R. A. O'donnell, C. Peebles, and J. A. Ward, "Relationship between peripheral airway dysfunction, airway obstruction, and neutrophilic inflammation in COPD," *Thorax*, vol. 59, no. 10, pp. 837–842, 2004.
- [18] J. L. Wright and A. Churg, "Advances in the pathology of COPD: Advances in the pathology of COPD," *Histopathology*, vol. 49, no. 1, pp. 1–9, 2006.
- [19] V. I. Peinado, S. Pizarro, and J. A. Barbera, "Pulmonary vascular involvement in COPD," *Chest*, vol. 134, no. 4, pp. 808–814, 2008.
- [20] L. Huang, W. Lin, D. Xie et al., "Development and validation of a preoperative CT-based radiomic nomogram to predict pathology invasiveness in patients with a solitary pulmonary nodule: a machine learning approach, multicenter, diagnostic study," *European Radiology*, vol. 32, no. 3, pp. 1983–1996, 2022.
- [21] R. C. Au, W. C. Tan, J. Bourbeau, J. C. Hogg, and M. Kirby, "Impact of image pre-processing methods on computed tomography radiomics features in chronic obstructive pulmonary disease," *Physics in Medicine and Biology*, vol. 66, no. 24, Article ID 245015, 2021.
- [22] J. Yun, Y. H. Cho, S. M. Lee et al., "Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease," *Scientific Reports*, vol. 11, no. 1, pp. 15144–15149, 2021.
- [23] R. C. Au, W. C. Tan, J. Bourbeau, J. C. Hogg, and M. Kirby, "Radiomics analysis to predict presence of chronic obstructive pulmonary disease and symptoms using machine learning," *TP121. TP121 Copd: From Cells To The Clinic*, p. A4568, American Thoracic Society, New York, NY, USA, 2021.
- [24] Y. Zhou, P. Bruijnzeel, C. Mccrae et al., "Study on risk factors and phenotypes of acute exacerbations of chronic obstructive pulmonary disease in Guangzhou, China-design and baseline characteristics," *Journal of Thoracic Disease*, vol. 7, no. 4, pp. 720–733, 2015.
- [25] S. Chen, K. Ma, and Y. Zheng, "Med3d: transfer learning for 3d medical image analysis," 2019, <https://arxiv.org/abs/1904.00625>.
- [26] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, pp. 1–13, 2020.
- [27] Y. Yang, Y. Guo, J. Guo, Y. Gao, and Y. Kang, "A method of abstracting single pulmonary lobe from computed tomography pulmonary images for locating COPD," in *Proceedings of the Fourth International Conference on Biological Information and Biomedical Engineering*, pp. 1–6, Chengdu China, July 2020.
- [28] Y. Yang, Q. Li, Y. Guo et al., "Lung parenchyma parameters measure of rats from pulmonary window computed tomography images based on ResU-Net model for medical respiratory researches," *Mathematical Biosciences and Engineering*, vol. 18, no. 4, pp. 4193–4211, 2021.
- [29] J. J. M. Van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.
- [30] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

- [31] S. Wan, Y. Liang, Y. Zhang, and M. Guizani, "Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones," *IEEE Access*, vol. 6, pp. 36825–36833, 2018.
- [32] M. Taki, A. Rohani, F. Soheili-Fard, and A. Abdeshahi, "Assessment of energy consumption and modeling of output energy for wheat production by neural network (MLP and RBF) and Gaussian process regression (GPR) models," *Journal of Cleaner Production*, vol. 172, pp. 3028–3041, 2018.
- [33] Q. Li, Y. Yang, Y. Guo et al., "Performance evaluation of deep learning classification network for image features," *IEEE Access*, vol. 9, pp. 9318–9333, 2021.
- [34] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: a survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.
- [35] V. Jakkula, "Tutorial on support vector machine (svm)," *School of EECS, Washington State University*, vol. 37, 2006.
- [36] Y. Qi, "Random forest for bioinformatics," *Ensemble machine learning*, Springer, Boston, MA, USA, pp. 307–323, 2012.
- [37] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
- [38] V. K. Ayyadevara, "Gradient boosting machine," in *Pro Machine Learning Algorithms*, pp. 117–134, Springer, Berlin, Germany, 2018.
- [39] W. Hu, W. Hu, and S. Maybank, "Adaboost-based algorithm for network intrusion detection," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 2, pp. 577–583, 2008.
- [40] Y. Yang, W. Li, Y. Guo et al., "Lung radiomics features for characterizing and classifying COPD stage based on feature combination strategy and multi-layer perceptron classifier," *Mathematical Biosciences and Engineering*, vol. 19, no. 8, pp. 7826–7855, 2022.
- [41] C. Casanova, J. P. de Torres, A. Aguirre-Jaime et al., "The progression of chronic obstructive pulmonary disease is heterogeneous: the experience of the BODE cohort," *American Journal of Respiratory and Critical Care Medicine*, vol. 184, no. 9, pp. 1015–1021, 2011.
- [42] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.
- [43] D. M. McNeish, "Using Lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences," *Multivariate Behavioral Research*, vol. 50, no. 5, pp. 471–484, 2015.
- [44] J. B. Lont, *Analog CMOS Implementation of a Multi-Layer Perceptron with Nonlinear synapses*, ETH Zurich, Zürich, Switzerland, 1993.
- [45] X. Song, M. Mao, and X. Qian, "Auto-metric graph neural network based on a meta-learning strategy for the diagnosis of alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3141–3152, 2021.

Research Article

Development and Application of a Standardized Testset for an Artificial Intelligence Medical Device Intended for the Computer-Aided Diagnosis of Diabetic Retinopathy

Hao Wang ¹, Xiangfeng Meng ¹, Qiaohong Tang ¹, Ye Hao ¹, Yan Luo ²,
and Jiage Li ¹

¹Institute for Medical Device Control, National Institutes for Food and Drug Control, 31 Huatuo Rd, Beijing 102629, China

²State Key Laboratory of Ophthalmology, Image Reading Center, Zhongshan Ophthalmic Center, Sun Yat-Sen University, No. 54 Xianlie South Road, Yuexiu District, Guangzhou 510060, Guangdong, China

Correspondence should be addressed to Yan Luo; luoyan2@mail.sysu.edu.cn and Jiage Li; lijiage@nifdc.org.cn

Received 1 April 2022; Revised 21 May 2022; Accepted 24 November 2022; Published 8 February 2023

Academic Editor: Yanwu Xu

Copyright © 2023 Hao Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Objective. To explore a centralized approach to build test sets and assess the performance of an artificial intelligence medical device (AIMD) which is intended for computer-aided diagnosis of diabetic retinopathy (DR). **Method.** A framework was proposed to conduct data collection, data curation, and annotation. Deidentified colour fundus photographs were collected from 11 partner hospitals with raw labels. Photographs with sensitive information or authenticity issues were excluded during vetting. A team of annotators was recruited through qualification examinations and trained. The annotation process included three steps: initial annotation, review, and arbitration. The annotated data then composed a standardized test set, which was further imported to algorithms under test (AUT) from different developers. The algorithm outputs were compared with the final annotation results (reference standard). **Result.** The test set consists of 6327 digital colour fundus photographs. The final labels include 5 stages of DR and non-DR, as well as other ocular diseases and photographs with unacceptable quality. The Fleiss Kappa was 0.75 among the annotators. The Cohen's kappa between raw labels and final labels is 0.5. Using this test set, five AUTs were tested and compared quantitatively. The metrics include accuracy, sensitivity, and specificity. The AUTs showed inhomogeneous capabilities to classify different types of fundus photographs. **Conclusions.** This article demonstrated a workflow to build standardized test sets and conduct algorithm testing of the AIMD for computer-aided diagnosis of diabetic retinopathy. It may provide a reference to develop technical standards that promote product verification and quality control, improving the comparability of products.

1. Introduction

As an emerging branch of the medical device, the AIMD, along with increasing applications of deep learning [1, 2], has demonstrated significant potential in medical imaging, image reconstruction, and postprocessing [3–16]. While hundreds of AIMDs have been approved [17, 18], the verification and validation of such devices are mainly conducted by manufacturers spontaneously, leading to variation in evaluation metrics and data sets [19]. Stakeholders show rising concern on the quality of the AIMD, such as its comparability [20] and transparency [21], which poses considerable challenges to

regulation compared to a conventional medical device. In the past several years, special guidelines for the AIMD have been published [22, 23]. There are increasing efforts to establish standards for the AIMD [24–27]. The topics include terminology, performance testing, dataset quality management, and quality systems.

To support standard development, it would be helpful to explore the approach to build and apply standardized test sets. While the literature reports existing public datasets for medical AI [28, 29], they are more appropriate for model training or competition [5, 8] rather than testing. On the one hand, the design of public datasets usually occurs before the

research and development of the AIMD, and they may not match the application scenario of the AIMD. On the other hand, test sets have special requirements. They should be independent from manufacturers or developers in order to verify the generalizability of AI. The capacity and diversity of data samples should be similar to the intended patient population. Standard operation protocols should be followed during the lifecycle. A systematic annotation process is needed to provide the reference standard.

This article demonstrates a case study to build test sets for computer-assisted diagnosis of DR, which is a common application of the AIMD. It is reported that deep learning algorithms can differentiate referable DR patients from nonreferable DR patients by reading colour fundus photographs [5, 7, 9, 10, 12]. Indeed, annual DR screening using digital photographs of the retina has long been recommended by several major governmental or professional organizations, including the UK National Health Service [10, 30], the American Diabetes Association [31], and other international societies [32].

In this article, a standardized approach is proposed to compose test sets for DR. The major procedure is described, including data collection, curation, and annotation. The test set is applied in the testing of AUTs. The advantages and practical issues of this approach are discussed, which may provide a reference for the development of technical standards.

2. Materials and Methods

2.1. Framework for Dataset Construction. The framework to build the test set is illustrated in Figure 1. It depicts a workflow, including design input, requirement specification, data collection, data curation, data annotation, and quality inspection. Risk management and personnel management are also considered and integrated into the workflow.

2.2. Design Input and Requirement Specification. To initiate dataset construction, the design input is firstly clarified. The intended use of this test set is to verify algorithm performance on classification of diabetic retinopathy by comparing algorithm outputs with the reference standard. The test set represents colored fundus photographs of diabetic patients from hospitals. Common image formats such as JPEG and BMP are accepted.

Requirement specification of this test set further describes dataset composition, classification, and data inclusion/exclusion criteria. This study uses colored photographs taken by fundus cameras that are officially approved to enter the market with a field of view no less than 45°. Photographs taken under near-infrared illumination are not included. According to the common intended use of AIMD products and the clinical guidelines for DR [33, 34], the images in the test set should include 7 categories (shown in Table 1): no apparent DR, mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, proliferative DR (PDR), other fundus diseases, and ungradable images (low image quality). No apparent DR and mild NPDR are considered nonreferable. Moderate NPDR, severe NPDR, and

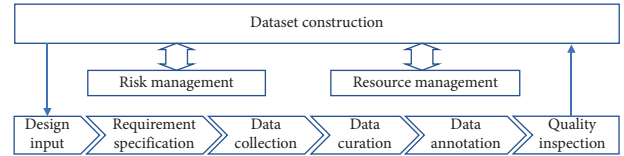


FIGURE 1: standardized framework for dataset construction.

TABLE 1: The categorization of the test set.

Class	Meaning
0	No apparent DR
1	Mild NPDR
2	Moderate NPDR
3	Severe NPDR
4	PDR
5	Other fundus diseases
6	Ungradable images

PDR are considered referable. The proportion of referable DR in the test set should be similar to the prevalence in the patient population.

Notably, the above categorization method is a result of justification since many AI products in China were designed according to the Guidelines for Diabetic Retinopathy Diagnosis and Treatment in China [33], which has referenced a previous version of the guidelines published in 1985 and ICO guidelines for diabetic eye care. The current guideline [33] divides DR based on severity into 6 stages as shown in Table 2. DR phases 0–III in Table 2 are equivalent to Classes 0–3 in Table 1. Since the treatment scheme of DR phases IV–VI is similar and the referral strategy is identical, the test set consolidates these stages into Class 4, which is compatible with ICO guidelines and practical in a clinical scenario.

Fundus diseases other than DR are classified as Class 5, which include but are not limited to hypertensive retinopathy [35], age-related macular degeneration [36], suspect glaucoma [37, 38], retinal vein occlusion [39], pathologic myopia [40], and optic nerve diseases [41]. Although these ocular diseases are not necessarily claimed by AIMD products, they may be imported into AIMDs in the real world. Therefore, they serve as negative controls in the test set.

Ungradable images are classified as Class 6. Image quality is given special attention in the development of the test set. DR screening is often performed in out-patients, sometimes on patients with undilated pupils. The colour retinal photographs are obtained using low levels of illumination. Also, human factors such as movement and positioning in addition to ocular factors such as cataracts and reflections from retinal tissues can produce defects. Especially, without pupillary dilatation, artifacts are observed in 3–30% of retinal images to the extent that they impede annotation [42]. Therefore, in this test set, ungradable images are also included, with conditions ranging from over darkness/saturation, out of focus, wrong positioning, lens contamination, to anterior segment images.

If an image only has minor quality problems that do not disturb annotation, it will be annotated and assigned to category 0–5. Images with photocoagulation marks and

TABLE 2: Definition of DR phases.

DR phases and findings observable on fundus photos [33]	Classes in ICO guidelines [32]
0: no abnormalities	No apparent DR
I: microaneurysms only	Mild NPDR
II: microaneurysms and other signs (e.g., dot and blot hemorrhages, hard exudates, and cotton wool spots), but less than severe nonproliferative DR	Moderate NPDR
III: moderate nonproliferative DR with any of the following: (1) Intraretinal hemorrhages (≥ 20 in each quadrant) (2) De nite venous beading (in 2 quadrants) (3) Intraretinal microvascular abnormalities (in 1 quadrant) (4) No signs of proliferative retinopathy	Severe NPDR
IV: neovascularization of the optic disc or elsewhere. When accompanied by vitreous/preretinal hemorrhage, it is defined as high risk PDR	Proliferative DR (PDR)
V: fibrous membrane could be accompanied by preretinal hemorrhage or vitreous hemorrhage	
VI: traction retinal detachment, combined with fibrous membrane, combined with/without vitreous hemorrhage, and neovascularization of the iris and the anterior chamber angle	

other treatment marks are annotated according to their posttreatment features. The comparison between pretreatment and posttreatment images is not within the scope of the test set.

2.3. Risk Management. Data security, patient privacy, and data bias are the major risks considered in this study. To ensure data security, all activities are conducted on the local area network with controlled user access. Data are stored in servers independent from algorithms under testing. Data annotation tools are not allowed to export images. To protect patient privacy, only deidentified images with ethical approval are accepted in this test set. To minimize data biases such as selection bias and coverage bias, the diversity of positive and negative samples is highlighted in the requirement specification.

2.4. Data Collection. During data acquisition, deidentified fundus photographs are collected retrospectively from partner hospitals with ethical approval from local institutional review boards. The raw images are submitted in JPEG formats. No modification or processing, such as filtering, smoothing, clipping, and contrast enhancing, is allowed. Additional information on image sources, including data collection sites, manufacturers of fundus cameras, and models of fundus cameras, is recommended and submitted.

2.5. Data Curation. Data curation is the process to ensure data safety and quality. First, the status of deidentification and ethical approval proof are manually confirmed. Second, data vetting is conducted to exclude problematic images, including unreadable files, incomplete images, and images

that compromise privacy information. After curation, the images are stored, indexed, and submitted to the image annotation process. Additional data preprocessing is not implemented in this study.

2.6. Resource Management. Dataset construction relies on resource management, especially personnel management and tool management.

Personnel management focuses on annotator recruitment, qualification, and management. The annotation task needs both junior annotators and senior annotators. All junior annotator candidates are publicly recruited. The basic qualification is a board-certified ophthalmologist with at least 5 years of clinical experience. All candidates receive annotation instructions in advance to clarify the classification rule according to the literature on DR [33, 34] and other fundus diseases [35–41]. After the training, the candidates attend an exam to classify 100 fundus photographs (18% nonreferrable DR, 45% referable DR, 32% other ocular diseases, and 6% ungradable images). Those who achieve greater than 80% accuracy pass the exam. They are given an additional training session.

Senior annotators should have professional certification as image readers and receive special training to promote consistency. In this article, senior annotators all have NHS (UK National Health Service) certification.

Tool management focuses on software tools that facilitate data processing and annotation. In this study, a custom-built annotation software is used. The main functions include image preview, contrast adjustment, image magnification, filter selection, task assignment, and progress monitoring. Annotators can add, edit, and submit annotation results. Reviewers and arbitrators can visit their

results and make corrections or justifications. The software only exports annotation results. No modifications are made to images.

2.7. Data Annotation. The reference standard is based on the combined decisions of junior annotators and arbitration experts. The image annotation is conducted in a laboratory environment. The annotation workflow is summarized in Figure 2. The annotation process includes two rounds:

2.7.1. First Round (Initial Annotation). Each batch of images is assigned to a team of 3 annotators. The annotators independently annotate images in a blinded way. If their classification result on an image is fully in agreement, such images are categorized as the prequalified pool. Images with discordant classifications are categorized as the arbitration pool. 10% of the prequalified pool is randomly sampled and submitted to the second round. The annotations of the rest of the prequalified pool are accepted conditionally. The arbitration candidate group are also submitted to the second round.

2.7.2. Second Round (Review and Arbitration). This step is carried out by a team of three senior annotators, one of whom acts as the team leader. The team leader has served as the director of an image reading center in a top ophthalmological hospital. They review all images submitted to this round so as to resolve the final annotation in the arbitration pool and review the samples from the prequalified pool. If sampled annotation results in the prequalified pool cannot pass the review, more samples will be submitted to the arbitration pool. Feedback may be given to annotators in the first round. Senior experts can justify the number of samples in the prequalified pool for inspection.

All images are stored, accessed, previewed, and manually classified using a custom-built annotation software.

2.8. Quality Inspection. After data annotation, quality inspection is conducted to examine the dataset's quality. The annotation records, including initial annotation, review, and arbitration, are reviewed and compared on each image to avoid inconsistencies and mistakes. Images that pass quality inspection are enrolled in the test set. The percentage of diabetic retinopathy subtypes is calculated. Usability and validity of each image are also examined manually.

2.9. Algorithm Testing. Five algorithm models intended to classify fundus photographs are enrolled as AUTs. They are trained by different manufacturers or developers. They all claim to use deep learning, but details such as the neural network structure, weights, and training sets are beyond the scope of this article. The test set is imported into each AUT. The output of AUTs is compared with the final annotation results. The overall accuracy, sensitivity, and specificity used to differentiate referable DR from nonreferable images are

reported. The performance of AUTs is further compared across the 7 subtypes separately.

3. Results

3.1. Diversity of the Test Set. The test set contains 6327 images from 11 hospitals in 10 provinces. Among them, 9 hospitals are tertiary hospitals and contribute 71.2% of the images, while the rest are secondary hospitals and contribute 28.8%. No primary hospitals or community clinics are involved. Since the images are deidentified, the location of the hospital is used to indicate geological distribution of patients. The provincial distribution of images is shown in Table 3, which demonstrates that representative provinces in Northeast China, North China, Central China, East China, Southeast China, and South China are involved.

The images are acquired by more than 13 types of fundus cameras made by 9 manufacturers, all in compliance with an ISO standard on fundus cameras [43]. The field of view is 45°. The optical resolution is between 80 and 120 pairs s/mm. All images are larger than 1000 pixel by 1000 pixel. The difference in image size, detector, light source, and embedded software may add more diversity to image quality and features.

In this test set, all fundus photographs are rectangular images with a pure background (either dark or white pixels) enveloping the round-shaped images of interest. The ratio between the pure background area and the whole area of each photograph is also considered an important source of image variation.

3.2. Performance of Annotators. During the recruitment of annotators, 47 ophthalmologists registered and attended the exam to classify 120 fundus images, including 63 DR images. 15 candidates finally passed and joined the annotation. Their average professional experience is above ten years. They are from 15 different hospitals in 7 provinces. Their accuracies in the exam range from 80% to 87%. The interannotator agreement is evaluated by calculating Fleiss' kappa. The result is 0.75, which is considered substantial given the fact that annotators come from different hospitals and regions. The intraannotator agreement is evaluated by calculating intraclass correlation, which is >85% for all qualified ophthalmologists. Additional training is given before the centralized annotation to reinforce the guidelines and minimize misunderstandings.

3.3. Annotation Results. In the first round, 15 annotators are evenly divided into 5 groups randomly. Individual workload is between 1000 and 1500 images. 3694 images yield concordant results, and 369 images are submitted to the second round as samples for inspection. 2356 images are graded with a majority opinion reached within each grading group and submitted to the second round for arbitration. 277 images yield totally diverse results within each group and are sent for arbitration too.

In the second round, the images are read by two NHS certified retinal experts and a senior expert with an NHS

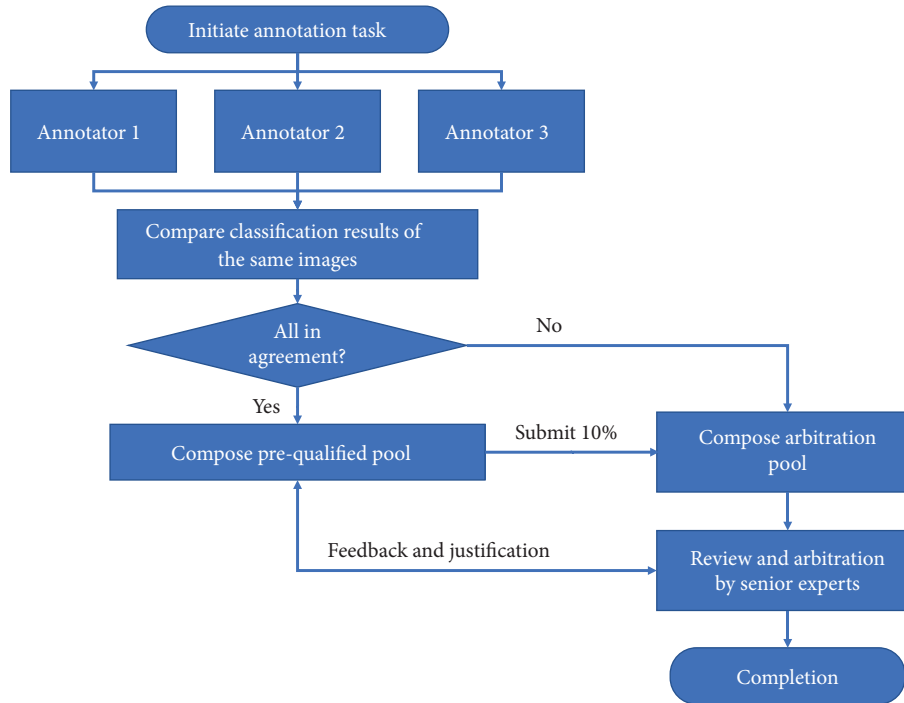


FIGURE 2: The annotation workflow.

TABLE 3: Geological distribution of image sources.

Region	Province	Percentage
North China	Beijing	16
Northeast China	Heilongjiang	13
	Liaoning	22
Central China	Henan	2
	Hubei	2
East China	Shanghai	9
	Zhejiang	4
	Anhui	3
Southeast China	Fujian	15
South China	Guangdong	14
Total		100

certificate independently in a blinded way. Then, they discuss all results and reach consensus on the final annotation results. According to the final results, 55.41% of images are directly determined by the consensus within each group in the first round. 16.02% of the images are graded according to the major opinion within each group in the first round. 26.81% of the images are graded with a reference to the minor opinion in each group in the first round. Only 1.76% of the images are graded only by the arbitrators.

Using the final annotation results as the reference standard, the accuracy of each annotator is calculated. The average accuracy is 83%. The minimum is 75%, while the maximum is 90%. 13 out of 15 annotators have accuracy higher than 80%. The performance of the 15 annotators comports with their qualification exam results and is considered satisfactory in comparison with the commonly accepted diagnostic accuracy by single-field fundus photography [42].

The composition of the annotated images is described in Table 4. The overall proportion of DR is 39.51%, comparable with the prevalence of DR in the Chinese DM population (24.7%–37.5%) [33]. The prevalence of other fundus diseases is 41.08%. This test set balances the proportion between DR and other fundus diseases that may be assessed by future AIMD products.

The classification of the current test set can be expressed in a simplified manner. Class 0 and Class 1 in Table 1 are consolidated into nonreferrable DR. Class 2 to Class 4 in Table 1 are consolidated into referable DR. Class 5 and Class 6 may remain independent or be consolidated into a certain type. In the following algorithm testing, they are considered nonreferrable.

3.4. Comparison with Raw Labels. During data collection, partner hospitals submitted raw labels, which were annotated by local annotators without centralized examination or training. The number of annotators deployed in each hospital varied from 1 to 3. The requirement for annotator qualification was different among partner hospitals. The minimum requirement was graduate student level, and the maximum requirement was associate professor level. Using the final annotation results as the reference standard, the overall accuracy of raw labels is 61.64%, and Cohen's Kappa is 0.5173, indicating the quality problems with raw labels.

3.5. Algorithm Testing Results. The overall accuracy, sensitivity, and specificity to differentiate referable DR from nonreferrable images are calculated and compared among the 5 AUTs. Table 5 shows the results of the 5 AUTs. The accuracy ranges from 0.77 to 0.88. The sensitivity ranges

TABLE 4: The distribution of annotated images.

Class	Number	Percentage
0: no apparent DR	873	13.798
1: mild NPDR	262	4.141
2: moderate NPDR	1118	17.670
3: severe NPDR	579	9.151
4: PDR	540	8.535
5: other fundus diseases	2600	41.094
6: ungradable	355	5.611
Total	6327	100

TABLE 5: Comparison of overall performance metrics.

Metrics	AUT1	AUT2	AUT3	AUT4	AUT5
Sensitivity	0.861422	0.814484	0.831024	0.802861	0.851587
Specificity	0.884597	0.820782	0.890465	0.799267	0.728851
Accuracy	0.876403	0.818555	0.869448	0.800537	0.772246

from 0.80 to 0.86. The specificity ranges from 0.73 to 0.89. AUT1 shows the highest accuracy and sensitivity among the 5 AUTs.

The capability of the algorithm to correctly classify images of a specific class as referable or nonreferable is also calculated. For class 2–class 4, it is represented as the number of true positives over the total number of samples in this category, which is equivalent to sensitivity. For other classes, the specificity of each category is calculated instead. Table 6 compares the performance of 5 AUTs on each specific class. It provides more details to demonstrate the variation in algorithm performance. For class 0, class 3, and class 4, the capability of all AUTs is above 95% on average. For class 1, the capability of AUT1 is significantly lower than the rest (on average above 90%). For class 2, the capability ranges from 0.64 to 0.75, indicating a common weakness among all 5 AUTs. For class 5, the capabilities of AUT1 and AUT3 significantly outweigh the rest of the AUTs. For class 6, AUT1 shows the top capability among the 5 AUTs. No AUTs in this experiment shows homogeneous capability to classify all 7 classes.

4. Discussion

This article demonstrates a centralized pathway to build test sets and conduct third party testing of AIMD products. The test set is composed of 6327 images, which are annotated into 7 classes covering all stages of DR according to ICO guidelines, as well as “other fundus diseases” and “ungradable images.” The diversity of the test set considers data sources (11 hospitals from 10 provinces), fundus cameras (>13 models from 9 manufacturers), and image parameters (image sizes, detectors, and light sources).

The pathway for test set construction in this article is different from that in algorithm challenges, where test sets and training sets are usually constructed under the same protocol or as subsets of a larger dataset. This pathway relies on independent data collection, curation, annotation, and storage, which decreases the possible similarity between this test set and training sets owned by developers of AUTs and

TABLE 6: Comparison of decision capability among 5 AUTs.

Class	AUT1	AUT2	AUT3	AUT4	AUT5
0	0.983963	0.989691	0.988545	0.988545	0.934708
1	0.557252	0.958015	0.912214	0.885496	0.889313
2	0.752236	0.645796	0.677102	0.639534	0.746869
3	0.982729	0.991364	0.993092	0.984455	0.977547
4	0.957407	0.974074	0.975926	0.946296	0.933333
5	0.893846	0.801923	0.889231	0.761153	0.642308
6	0.814085	0.442254	0.642254	0.549295	0.738028

promotes the verification of AI algorithm generalizability. It may be suitable for third party testing laboratories to conduct conformity assessment.

According to the literature [5, 9, 10, 44], the pathway to form the reference standard in other studies is based on various combinations of annotators and reviewers. In this study, a combination of prequalified annotators and arbitrators conducted data annotation. Under this scheme, the annotators’ performance is estimated quantitatively (Fleiss Kappa = 0.75, individual accuracy >80%, and intra-class correlation >85%). During the annotation process, each image in the test set is reviewed by 3–6 experienced professionals, and 98.2% are determined by the major decision (3 votes out of 3 annotators or >4 votes out of 6). Only 1.76% are determined by the arbitration experts. The results show that the annotation scheme helps enhance consensus among annotators.

On the other hand, the raw labels from partner hospitals show significantly lower accuracy and consistency compared to the final annotation results. According to information provided by partner hospitals, the raw labels are annotated by an inconstant number of annotators, ranging from 1 to 3, including graduate students, residents, and junior and senior ophthalmologists. It suggests the importance to organize annotation task systematically and the necessity to establish consistent annotation rules among different hospitals. Otherwise, the discrepancy in data annotation may impact dataset quality and further inhibit the quality of the AIMD.

Using the annotated test set, the performance of 5 AUTs is tested quantitatively as technical demonstration. It is straightforward to compare the overall accuracy, sensitivity, and specificity in the scenario of DR classification. Algorithm performance can be further observed on subgroups of the test set. However, no AUT in this experiment shows homogeneous capability to classify different categories of images. While public stakeholders pay attention to algorithm fairness and generalizability, this study shows the necessity to reveal and understand how the AI algorithm performs differently on subtypes of diabetic retinopathy images. It also indicates that algorithm performance may change with the proportion of these categories. A strategy to tune the composition of test sets in a flexible manner is needed to guide future testing.

This work explores practical approach and issue in advancing the standardized testing of the AIMD. But due to time and resource constraints, it has limitations in the following aspects:

First, the test set is based on retrospective data collection. Although data are randomly sampled by partner hospitals, control measures should be taken to limit bias. Continuous sampling of data within a period may help.

Second, the proportion of mild NPDR is much smaller than that of other DR subtypes. One possible reason is that without compulsory DR screening, patients with mild NPDR are unlikely to take fundus photographs, which results in the relative scarcity of mild NPDR photographs. Increment of mild NPDR not only decreases the sampling errors of SE and SP but also improves the balance between different stages of DR. In fact, from the annotator's perspective, it is important to differentiate microaneurysm in mild NPDR from blot hemorrhages in moderate NPDR. Therefore, more cases of mild NPDR should be added to the current test set.

Third, as a colour fundus photograph dataset, it is difficult to use the test set alone to annotate important diseases among the 41.09% "other diseases" that may be assessed by AI in the near future. Colour fundus photographs are incapable of thickness measurement, which inhibits detection of certain diseases such as AMD and glaucoma. Images from additional imaging modalities such as OCT should be added to the test, but the cost will increase significantly.

Fourth, the diversity of this test set still needs improvement. Partner hospitals in this study are mostly tertiary hospitals, without community-level hospitals. As a result, most photographs are acquired by high-end fundus cameras. Handheld fundus cameras, which may be more popular in community-level clinics and rural areas, have minor contribution to data collection. More data should be added to compensate for this scenario and enrich data diversity.

To promote standardization of AIMD testing, reliability and comparability of test sets need to be addressed in the future research. Test sets built by different organizations may have different data sources, data inclusion/exclusion criteria, annotation resources, and procedures, which would cause inconsistent dataset quality. Transparent description of data sets should be normalized. Consensus standards on dataset construction and annotation are needed to guide the procedure. It would be necessary to conduct sample inspection and comparison among test sets, similar to proficiency testing [45] by interlaboratory comparison.

5. Conclusions

This article proposes a practical approach to build test sets for third-party testing of the AIMD. It takes quality control measure during data collection, curation, and annotation. It demonstrates the benefit of centralized data annotation in comparison with individual annotators and spontaneous annotation from single hospitals. The application of such a test set reveals algorithm performance and weakness in a comparative and straightforward manner, providing helpful information for regulation of such medical devices.

Data Availability

The data supporting the findings of the current study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors want to show their gratitude to all annotators participating in the study. The authors would like to thank Dr. Haiping Ren who provided helpful advice on the study design and manuscript preparation and Dr. Yifan Xiang who provided useful feedback to the study. This research was sponsored by the National Key R&D Program of China under grant nos. 2019YFC0118801 and 2019YFB1404805.

References

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 05/27/online 2015.
- [3] M. D. Abramoff, J. M. Reinhardt, S. R. Russell et al., "Automated early detection of diabetic retinopathy," *Ophthalmology*, vol. 117, no. 6, pp. 1147–1154, 2010.
- [4] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Y. K. Ng, and A. Laude, "Computer-aided diagnosis of diabetic retinopathy: a review," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2136–2155, 2013.
- [5] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [6] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [7] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
- [8] A. A. A. Setio, A. Traverso, T. de Bel et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017.
- [9] D. S. W. Ting, C. Y. L. Cheung, G. Lim et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
- [10] A. Tufail, C. Rudisill, C. Egan et al., "Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human annotators," *Ophthalmology*, vol. 124, no. 3, pp. 343–351, 2017.
- [11] D. S. Kermany, M. Goldbaum, W. Cai et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.



- [12] A. A. van der Heijden, M. D. Abramoff, F. Verbraak, M. V. van Hecke, A. Liem, and G. Nijpels, "Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System," *Acta Ophthalmologica*, vol. 96, no. 1, pp. 63–68, 2018.
- [13] S. Wang, C. Li, R. Wang et al., "Annotation-efficient deep learning for automatic medical image segmentation," *Nature Communications*, vol. 12, no. 1, Article ID 5915, 2021.
- [14] W. Huang, H. Yang, X. Liu et al., "A coarse-to-fine deformable transformation framework for unsupervised multi-contrast MR image registration with dual consistency constraint," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2589–2599, 2021.
- [15] S. Wang, H. Cheng, L. Ying et al., "Deepcomplexmri: exploiting deep residual network for fast parallel mr imaging with complex convolution," *Magnetic Resonance Imaging*, vol. 68, 2020.
- [16] S. Wang, S. Tan, Y. Gao et al., "Learning joint-sparse codes for calibration-free parallel MR imaging," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 251–261, 2018.
- [17] U.S. Food and Drug Administration, "Artificial intelligence and machine learning(AI/ML)-enabled medical devices," [EB/OL], 2022, <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-ai-ml-enabled-medical-devices>.
- [18] J. H. Woo, E. C. Kim, and S. M. Kim, "The current status of breakthrough devices designation in the United States and innovative medical devices designation in Korea for digital health software," *Expert Review of Medical Devices*, vol. 19, no. 3, pp. 213–228, 2022.
- [19] L. Wang, H. Wang, C. Xia et al., "Toward standardized premarket evaluation of computer aided diagnosis/detection products: insights from FDA-approved products," *Expert Review of Medical Devices*, vol. 17, no. 9, pp. 899–918, 2020.
- [20] M. Roberts, D. Driggs, M. Thorpe et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, pp. 199–217, 2021.
- [21] F. Cabitza, R. Rasoini, and G. F. Gensini, "Benefits and risks of machine learning decision support systems—reply," *JAMA*, vol. 318, no. 23, pp. 2356–2357, 2017.
- [22] U.S. Food and Drug Administration, "Good machine learning practice for medical device development: guiding principles," 2021, <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>.
- [23] Artificial Intelligence Medical Devices (AIMD) Working Group, *IMDRF/AIMD WG/N67 Machine Learning-enabled Medical Devices: Key Terms and Definitions*, IMDRF, 2022, <https://www.imdrf.org/sites/default/files/2022-05/IMDRF%20AIMD%20WG%20Final%20Document%20N67.pdf>.
- [24] IEC TC62 PT8, "PWI 62-3 artificial intelligence/machine learning-enabled medical device – performance evaluation process," [EB/OL], International Electrotechnical Commission, 2021, https://www.iec.ch/dyn/www/?p=103:38:411041400161435:::FSP_ORG_ID,FSP_APEX_PAGE,FSP_PROJECT_ID:1245,23,107066.
- [25] IEC TC62, IEC 63450 ED1, "Testing of artificial intelligence/machine learning-enabled medical devices," *International Electrotechnical Commission*, 2022, https://www.iec.ch/dyn/www/?p=103:38:401670179546963:::FSP_ORG_ID,FSP_APEX_PAGE,FSP_PROJECT_ID:1245,23,109273.
- [26] "IEEE 2801-2022 recommended practice for the quality management of datasets for medical artificial intelligence," *IEEE Engineering in Medicine and Biology Society/Standards Committee*, 2022, <https://standards.ieee.org/ieee/2801/7459/>.
- [27] "IEEE 2802-2022 Approved Draft Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology," *IEEE Engineering in Medicine and Biology Society/Standards Committee*, 2022, <https://standards.ieee.org/ieee/2802/7460/>.
- [28] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging*, vol. 5, no. 3, Article ID 036501, 2018 Jul.
- [29] E. Decencière, X. Zhang, G. Cazuguel et al., "Feedback on a publicly distributed database: the Messidor database," *Image Analysis and Stereology*, vol. 33, no. 3, pp. 231–234, aug 2014.
- [30] A. Tufail, V. V. Kapetanakis, S. Salas-Vega et al., "An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness," *Health Technology Assessment*, vol. 20, no. 92, pp. 1–72, Dec 2016.
- [31] American College of Physicians, American Diabetes Association, and American Academy of Ophthalmology, "Screening guidelines for diabetic retinopathy," *Annals of Internal Medicine*, vol. 116, pp. 683–685, 1992.
- [32] International Council of Ophthalmology, *ICO Guidelines for Diabetic Eye Care*, International Council of Ophthalmology, Brussels, Europe, 2017.
- [33] Vitreo-Retina Society, "Chinese society of ocular fundus diseases the guidelines for diabetic retinopathy diagnosis and treatment in China," *Chinese Journal of Ophthalmology*, vol. 50, no. 15, 2014.
- [34] Vitreo-Retina Society, "Chinese Society of Ocular Fundus Diseases Guidelines for image collection and reading of the diabetic retinopathy screening in China," *Chinese Journal of Ophthalmology*, vol. 53, 2017.
- [35] T. Y. Wong and P. Mitchell, "Hypertensive retinopathy," *New England Journal of Medicine*, vol. 351, no. 22, pp. 2310–2317, 2004.
- [36] Vitreo-Retina Society, "Chinese society of ocular fundus diseases clinical pathway of age-related macular degeneration in China," *Chinese Journal of Ocular Fundus Diseases*, vol. 29, no. 13, 2013.
- [37] International Council of Ophthalmology, *ICO Guidelines for Glaucoma Eye Care*, International Council of Ophthalmology, Brussels, Europe, 2016.
- [38] Glaucoma Society, "Chinese society of ocular fundus diseases diagnosis and treatment of primary glaucoma: expert consensus," *Chinese Journal of Ophthalmology*, vol. 50, no. 2, 2014.
- [39] A. Berger, A. Cruess, F. Altomare et al., "Optimal treatment of retinal vein occlusion: canadian expert consensus," *Ophthalmologica*, vol. 234, 2015.
- [40] Chinese Optometric Association, "Chinese ophthalmological society, "consensus: prevention and control of high myopia," *Chinese Journal of Optometry Ophthalmology and Visual Science*, vol. 19, no. 5, 2017.
- [41] Chinese Neural-Ophthalmology Association, "Chinese ophthalmological society, "diagnosis and treatment of optic neuritis: expert consensus," *Chinese Journal of Ophthalmology*, vol. 50, no. 6, 2014.
- [42] A. Govinda and R. de Verteuil, "Systematic review of the diagnostic accuracy of the single, two and three field digital retinal photography for screening diabetic retinopathy," *JB*

Database of Systematic Reviews and Implementation Reports, vol. 9, no. 16, pp. 491–537, 2011.

- [43] International Organization for Standardization, *ISO 10940: 2009 Ophthalmic Instruments -- Fundus Cameras, ISO Standard*, International Organization for Standardization, Geneva, Switzerland, 2009.
- [44] H. V. Nguyen, G. S. W. Tan, R. J. Tapp et al., “Cost-effectiveness of a national telemedicine diabetic retinopathy screening Program in Singapore,” *Ophthalmology*, vol. 123, no. 12, pp. 2571–2580, 2016.
- [45] International Organization for Standardization, *ISO 13528: 2015 Statistical methods for use in proficiency testing by interlaboratory comparison*, International Organization for Standardization, Geneva, Switzerland, 2015.

Research Article

An End-to-End Data-Adaptive Pancreas Segmentation System with an Image Quality Control Toolbox

Yan Zhu ¹, **Peijun Hu**,² **Xiang Li**,^{3,4} **Yu Tian**,¹ **Xueli Bai**,^{3,4} **Tingbo Liang**,^{3,4}
and **Jingsong Li** ^{1,2}

¹Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China

²Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 311100, China

³Department of Hepatobiliary and Pancreatic Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310006, China

⁴Zhejiang Provincial Key Laboratory of Pancreatic Disease, Hangzhou 310006, China

Correspondence should be addressed to Jingsong Li; ljs@zju.edu.cn

Received 6 June 2022; Revised 18 August 2022; Accepted 24 November 2022; Published 24 January 2023

Academic Editor: Weihua Yang

Copyright © 2023 Yan Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of radiology and computer technology, diagnosis by medical imaging is heading toward precision and automation. Due to complex anatomy around the pancreatic tissue and high demands for clinical experience, the assisted pancreas segmentation system will greatly promote clinical efficiency. However, the existing segmentation model suffers from poor generalization among images from multiple hospitals. In this paper, we propose an end-to-end data-adaptive pancreas segmentation system to tackle the problems of lack of annotations and model generalizability. The system employs adversarial learning to transfer features from labeled domains to unlabeled domains, seeking a dynamic balance between domain discrimination and unsupervised segmentation. The image quality control toolbox is embedded in the system, which standardizes image quality in terms of intensity, field of view, and so on, to decrease heterogeneity among image domains. In addition, the system implements a data-adaptive process end-to-end without complex operations by doctors. The experiments are conducted on an annotated public dataset and an unannotated in-hospital dataset. The results indicate that after data adaptation, the segmentation performance measured by the dice similarity coefficient on unlabeled images improves from 58.79% to 75.43%, with a gain of 16.64%. Furthermore, the system preserves quantitatively structured information such as the pancreas' size and volume, as well as objective and accurate visualized images, which assists clinicians in diagnosing and formulating treatment plans in a timely and accurate manner.

1. Introduction

Pancreatic cancer is a malignant tumor and is recognized with a low survival rate, and pancreatic diseases are characterized by rapid occurrence and continuous progression [1, 2]. Even among all types of cancer, it is one of the most dangerous and deadly. It is estimated that, in 2021, statistically, about 60,430 new cases of pancreatic cancer would be diagnosed in the US, and 48,220 people would die from this disease [3]. Because pancreatic cancer is difficult to diagnose in its early stage, the rate of diagnosis is almost

as high as the mortality rate. If not treated timely, a significant portion of patients with pancreas disease would be diagnosed with metastatic symptoms [4]. For pancreatic cancer, the common treatment options are surgical resection, neoadjuvant radiotherapy, radiotherapy, and chemotherapy [5, 6], of which surgery is acknowledged to achieve a better prognosis but with somewhat invasiveness. Regardless of the chosen treatment option, precise localization and segmentation of the pancreas are crucial for physicians to diagnose and assess the patient's condition early in the treatment phase [7].

Computed tomography (CT) and magnetic resonance imaging (MRI) are particularly important examination procedures and tools in diagnosing pancreatic diseases [8]. CT, especially contrast-enhanced CT, is the first choice for pancreatic examinations in hospitals and has the advantages of more rapid imaging and clearer strips than MRI. However, the irregularity and variability of the pancreas in morphology and low contrast in its surrounding tissues lead to a high demand for experience and prior knowledge of radiologists in the early diagnosis of pancreatic diseases [9–12]. Nowadays, precision medicine requires the clinical process upgrading from qualitative observation to quantitative analysis and diagnosis [13, 14]. Therefore, accurate automatic segmentation of pancreatic tissues in CT images can greatly accelerate the early diagnosis process for physicians and assist in the appropriate treatment process.

In recent years, with technological breakthroughs in deep learning, computer-aided pancreas segmentation has yielded a series of promising methodological studies [15–26]. Limited by the small object characteristics of the pancreas, several studies have employed the two-stage cascade approach for semantic segmentation [7, 26]. These methods first locate the pancreas in the entire CT sequence with bounding boxes and then perform pixel-level segmentation of the pancreas on the basis of box localization. nnU-Net adopts a two-stage strategy to perform segmentation on pancreatic healthy tissues and lesion tissues autonomously by configuring itself and achieves the best accuracy in the Medical Segmentation Decathlon challenge [27]. Zhu et al. [15], Fu et al. [17], and Oktay and Chen [18] introduced attention modules and spatial information to a three-dimensional (3D) segmentation structure to capture the consistency information of pancreatic tissue in CT images. More recently, a series of AI-based semantic segmentation or object detection methods has been applied to clinical practice for quantitative data measurement and morphometric analysis [28–30]. For example, pancreatic fistula prediction after pancreaticoduodenal surgery is based on quantitative pancreatic volume measurements in CT images, and gallbladder resection is guided by fat pairs from quantitative measurements of the pancreas in CT images.

However, the more challenging issue in the clinical application of computer-aided pancreas segmentation is that the heterogeneity of across domain images leads to poor generalization of models [31]. This is manifested by the fact that supervised models trained on a single dataset, even if trained with accurate expert manual annotations, are subject to significant model inference errors once they are deployed to other medical centers [32–38]. Normalized images are essential for good performance of deep learning models. The variation in medical images in terms of populations, scanning devices, scanning parameters, or imaging protocols will lead to varying quality [33, 39, 40]. This heterogeneity exists in both CT and MRI images [33, 34, 36]. Besides, pancreatic disease and pancreatic cancer dramatically change the morphology of the pancreatic tissue, specifically demonstrated by diffuse enlargement, inhomogeneous density, and

ambiguous boundaries, which make the data quality inconsistent among medical centers [41]. Liu et al. [33] and Wang et al. [36] illustrated the heterogeneity across medical sites in terms of the patient cohort and image quality. Several research studies on medical image segmentation have previously demonstrated significant performance degradation of single-site models when deployed to other sites. Lerousseau and Xiao [42] proposed a new weakly supervised multi-instance learning method as a tool for pancreas tumor segmentation which achieved promising performance by taking full advantage of less annotated data at the pixel level. However, this method yielded about 15–26% performance degradation when it was tested for other publicly available datasets. Obviously, interdomain generalization greatly limits the clinical application of automatic pancreas segmentation models. In the field of natural images, several studies have addressed this issue by using transfer learning or federation learning algorithms [43–47]. However, the effectiveness of approaches based on natural images is unsatisfactory or even worse for CT images due to the gap between the two types of images [48]. There is one research study on semisupervised segmentation pancreas tasks on the NIH-TCIA dataset and achieved a dice similarity coefficient score of 78.27% by using a portion of annotated images [49]. Moreover, to date, few research studies related to unsupervised segmentation or domain adaptation of the pancreas have been conducted.

Therefore, to tackle this serious real-clinical problem, we propose an end-to-end data-adaptive pancreas segmentation system with an image quality control toolbox in this paper. The system focuses on the pancreas segmentation model construction with heterogeneous cross-domain CT images in the presence of insufficient annotations in medical centers. The main contributions of this paper are summarized as follows: (1) The system utilizes adversarial learning to construct data-adaptive segmentation models with the assistance of domain discriminators. Besides, we employ a collaboration center to perform feature-level transfer learning without data sharing across domains. (2) A multifunctional image quality control toolbox is designed to standardize the quality of images from various medical centers in terms of the intensity range, field of view, region of interest, etc. (3) The system works in an end-to-end mode, which only requires physicians to select images, set up personalized parameters, and then wait for automatic model construction and inferences on unlabeled data. (4) The system offers a variety of pancreas-related features including textual information and imaging results, which can assist physicians in quantitative and precise clinical diagnosis of the pancreas.

We experimentally demonstrate the effectiveness of the system on a public dataset and an in-hospital dataset and validate the robustness of the system on a small dataset. The data-adaptive pancreas segmentation system we developed is able to diagnose a larger number of people quickly and effectively in the clinical practice and to obtain meaningful pancreas segmentation results.

2. Methods

Accurate segmentation of pancreatic tissue is an essential stage in clinical diagnosis of pancreatic diseases. The variability of CT images from different medical centers affects the generalizability of automatic pancreatic segmentation tools. Subtle differences in image features result in sudden decreases in segmentation accuracy for deep learning models. To address such a problem, improvements can be carried out in terms of both data alignment and segmentation model construction methods, respectively. On the one hand, some image quality control approaches for pancreas CT images are used to process different source images. On the other hand, the automatic pancreatic segmentation model should have the ability to adapt to variations in the data domain in terms of the methodology and system design. To address these issues, a novel end-to-end data-adaptive segmentation system for the pancreas with an image control toolbox is proposed for pancreatic data quality normalization and assisting in pancreas segmentation model generalization. In this chapter, the overall framework of the system and the construction and functions of each module are shown.

2.1. Framework. The overall framework of the proposed data-adaptive pancreas segmentation system embedded with a data quality control toolbox is shown in Figure 1. The system consists of two parts: the local clients of medical centers and a collaborative center on the cloud server. The primary tasks of the local client are integrating and processing data, constructing segmentation models, and visualizing segmentation results. The collaborative center is mainly responsible for transfer learning of image features among multiple medical centers.

The medical center client consists of four modules, namely, data organization module, image quality control module, segmentation module, and visualization module. The data organization module is mainly operated by imaging physicians to establish the patients' cohort to study and extract the pancreas CT images from the in-hospital database. The image quality control module performs standardized preprocessing on the previously selected pancreas CT images in order to normalize the data and reduce the variation of images from different sources. The segmentation module is mainly equipped with the predefined semantic segmentation model in the system to train annotated data and cooperates with the collaborative center to assist in the construction of the data-adaptive segmentation model for unannotated data. Besides, the segmentation module uses the well-trained model to predict the mask of pancreas tissues in input images. The visualization module presents multidimensional results of the segmented pancreas tissues in terms of visualized images and structured text.

The collaborative center mainly contains feature discriminators. The feature transmission and learning process between the medical center and the collaborative center is encapsulated as the transfer learning module. The transfer learning module mainly accomplishes adversarial learning

between image features from multiple centers. The feature discriminator optimizes the segmentation network of unlabeled images by balancing the Nash equilibrium of the domain classification loss and the segmentation loss. Thus, the segmentation network is adaptive to the new images without the need of annotation.

2.2. Image Quality Control Toolbox. The image quality control toolset provides a variety of image processing means to standardize various qualities of images from multiple sources, mainly including the intensity value cutoff, rotation augmentation, and superresolution reconstruction. The CT intensity of abdominal organs is in the range of $(-160, 240)$ HU, and the range for the pancreas is kept in $(-100, 240)$ HU. This scale preserves pancreatic tissue features and removes background information. Rotational augmentation refers to amplify CT images by rotating them at the axial plane with degrees in the range of $(\pm 5^\circ, \pm 10^\circ)$. This operation is not performed in the ordinary sense of augmenting data to improve model performance but rather to attenuate the angle bias introduced by the field of view or body position during scanning so as to eliminate heterogeneity. Super-resolution reconstruction could effectively improve the image quality, thus reducing the quality inconsistency caused by scanning devices, imaging protocols, slice thickness, and so on. In this study, it is concerned that the effective abdominal region in CT images fluctuates because of different scanning fields of view, so the first step in superpixel reconstruction is framing out the region of interest. Pancreatic CT images were binarized to measure image region properties. Then, the maximum connected region containing the region of interest was found by the region-growing algorithm. The rectangular area bounded by the diagonal vertices of the maximum connected region in the image is considered the valid abdominal area. Then, the truncated 3D volume is interpolated in 3D cubic interpolation (system default settings) and reconstructed to the same resolution to feed into the network. The default size is 512×512 resolution in an axial plane and 1 mm thickness in a sagittal direction. Furthermore, the system provides various image interpolation methods to system users. The reconstruction algorithm library includes nearest neighbor interpolation, bilinear interpolation, Lanczos interpolation, and bicubic interpolation to support multiple requirements for medical studies.

2.3. Segmentation Module. The segmentation module is embedded with a deep learning semantic segmentation network applicable to the pancreas segmentation task. This module is mainly responsible for the training of labeled data models, the construction of adaptive unlabeled data models in cooperation with discriminators, and the prediction of pancreas masks for input samples.

In this paper, a 3D ResUNet structure integrated with an attention mechanism is designed as the backbone model of the segmentation module. U-Net [50] is a widely used semantic segmentation network, which has the advantages of the small amount of data needed, high data utilization, and

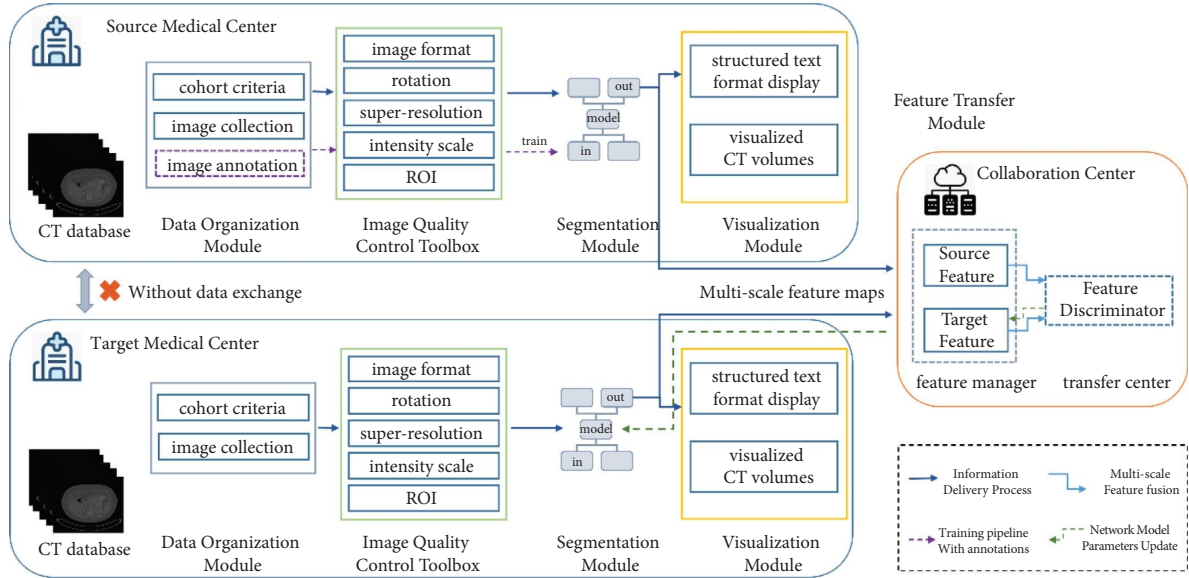


FIGURE 1: Overall framework of the data-adaptive pancreas segmentation system. The system consists of two parts: the local clients of medical centers and the collaborative center on the cloud server.

short training periods. On the basis of U-Net, the ResUnet model designed in this research introduces a deep residual structure which enables weighted interactions of image features at different scales, thus improving the segmentation performance of small targets like pancreatic tissue. The residual connection simplifies training and eliminates the degradation of the transmission of information among low and high scales, resulting in fewer parameters in networks. Besides, an attention mechanism is introduced into 3D spatial channels to enlarge the model capacity and enhance the network feature representation capability. The attention structure is inspired by the design paradigm of the squeeze and excitation (SE) network [50] and extended to 3D to handle 3D image representations. The specific structure is represented as follows: the input 3D feature map is squeezed based on pooling operations to obtain the feature vector in channels and feature activation is performed by the multilayer perception (MLP) operation of fully connected layers. After obtaining the weighting coefficients of the important channels in the feature map, the obtained coefficients are then linearly weighted into the input 3D feature map in a dot-product manner.

The feature maps at four scales (32×32 , 64×64 , 128×128 , and 256×256) of the decoder are transmitted to the collaborative center for adversarial learning. The feature maps at various scales contain not only the shallow boundary information but also the precise pancreas target information so as to ensure the effectiveness of domain adaptation. Moreover, to more strictly constrain the segmentation task on small object tasks, the model employs a linear combination of the dice loss and cross-entropy loss function as optimization criteria. The loss function is formulated as

$$\text{Loss}_{\text{all}} = \alpha \text{Loss}_{\text{CE}} + \lambda \text{Loss}_{\text{dice}}, \quad (1)$$

where α and λ are the linear coefficients that range from 0 to 1. These parameters can be customized by system users according to the needs of practical research purposes. In the experiments of this paper, $\alpha = 0.8$ and $\lambda = 0.5$.

2.4. Feature Transfer Module. The feature transfer module is mainly responsible for adversarial learning of image features among centers. Adversarial learning is performed by the discriminator in the collaborative center. The multiscale image features generated by the segmentation module of the medical center are transferred to the collaborative center as the four inputs of the discriminator. Each of the input feature maps sequentially undergoes a 3D convolutional layer and an activation function, where the step size of the 3D convolutional layer is set to 2. Therefore, the spatial scale of image features is decreased by half and thus concatenated with the next-scale feature map and fed to the next layer. Upon weighted feature fusion of multiple scales, the features are fed into the average pooling layer and the fully connected layer in turn to obtain the final domain classification results.

The workflow of feature transfer is shown in Figure 2. The labeled source-domain data are denoted as x , and the labeled dataset is defined as $S(x)$; the unlabeled target-domain data are denoted as z , and the unlabeled dataset is defined as $T(z)$. At first, the labeled dataset is trained with annotated images by the segmentation module to obtain an initial pancreas segmentation model. Then, x and z are fed into the segmentation model in pairs to generate multiscale feature maps $\mathcal{F}_s(x)$ and $\mathcal{F}_T(z)$, respectively, which are then transmitted to the collaborative center. In the collaborative center, the feature maps from the two branches are trained by the discriminator $\mathcal{H}(\cdot)$ for domain identification. Given feature maps from the source domain with label 1 and target-domain feature maps with label 0, then the optimization condition of the discriminator is to seek the weights

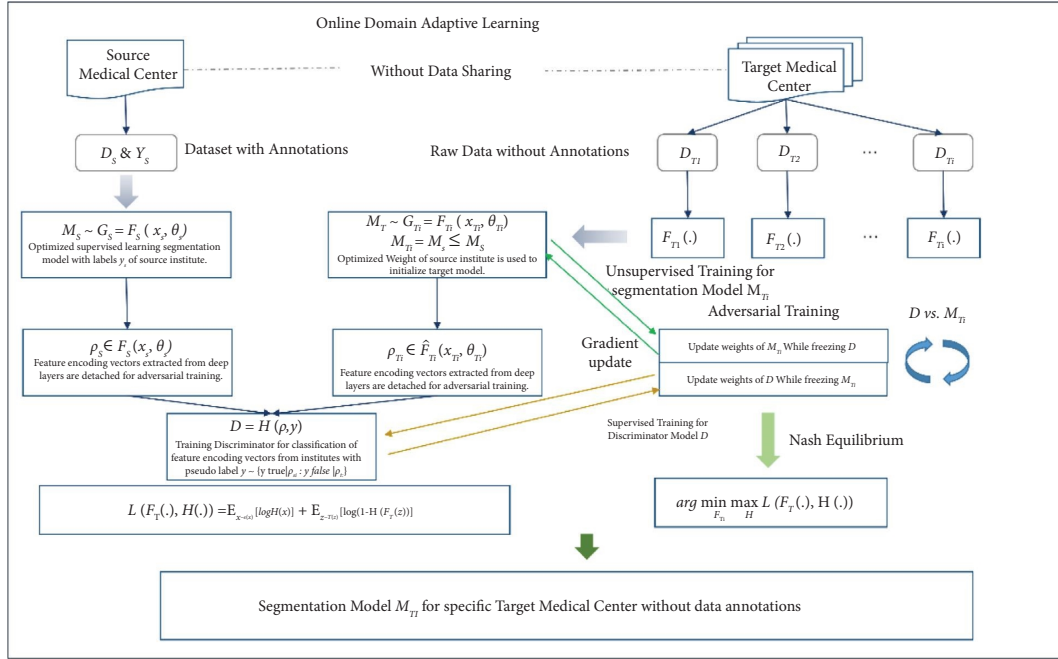


FIGURE 2: Workflow for the construction of a data-adaptive pancreas segmentation model for unlabeled data.

that maximize the difference of domain features, and the loss function is defined as

$$\mathcal{L}(\mathcal{F}_T(\cdot), \mathcal{H}(\cdot)) = \mathbb{E}_{x \sim S(x)} [\log (\mathcal{H}(\mathcal{F}_S(x)))] + \mathbb{E}_{z \sim T(z)} [\log (1 - \mathcal{H}(\mathcal{F}_T(z)))], \quad (2)$$

where $\mathbb{E}_{x \sim S(x)}$ and $\mathbb{E}_{z \sim T(z)}$ denote the mathematical expectations for the two parts.

Meanwhile, the feature maps from the target-domain images are labeled with a change to 1 and are fed into the discriminator in a single branch. The discriminator back-propagates the source-domain features with respect to the labels, thus amplifying such common image features and updating the pancreas segmentation model for the target domain. In the subsequent feature transfer process, the discriminator and the target domain segmentation model are continuously updated and frozen alternately as in the above step, thus searching for the Nash equilibrium of the two optimization functions in the adversarial process. The final optimization purpose of adversarial learning is to discover general image features between domains and utilize them to guide the segmentation of the pancreas.

2.5. Visualization Module. When the pancreas segmentation model for unannotated images is constructed by the feature transfer module, doctors could select CT images to study from the hospital local database and perform standardized preprocessing for images with custom parameters in the image quality control toolbox. Subsequently, the visualization module performs postprocessing on the pancreas mask output by the segmentation module and displays CT images and corresponding segmentation results. The presentation includes visualized images and structured text information.

Segmentation is essentially the prediction of whether each pixel in the image is a target foreground or background, so there are usually a number of isolated and noisy points in the segmentation mask. In this study, a conditional random field (CRF) model and a hole-filling algorithm are used as postprocessing operations in visualization modules to further optimize the segmentation mask to eliminate the anomalous structure and smooth the boundaries. In addition, the module offers a parallel visualization comparison between segmentation results and physician annotations in each slice-of-interest so that the physician can check annotations. The visualized image results include original CT images and segmented pancreas masks in the form of 2D slices and 3D volumes. Besides, the 3D reconstruction of surface distances between masks and annotations are also displayed. The module with visualized images also provides support for window dragging, rotating, zooming, and other operations to display images more comprehensively. Structured text information covers volume, size, and occupancy depth of pancreas tissues, surface distances between masks and annotations, etc.

2.6. Experimental Results

2.6.1. Datasets. The NIH-TCIA dataset [23] is employed as the labeled source-domain dataset in this study. The NIH-TCIA dataset is collected by the National Institutes of Health Clinical Center, which is currently the authoritative and commonly adopted public dataset for pancreas segmentation. We employed 70 cases of in-hospital CT images, collected from the First Affiliated Hospital of the Zhejiang University School of Medicine, as the unlabeled target domain dataset, noted as the Zheyi dataset. The annotations of Zheyi images were all manually outlined and cross-validated

by professional physicians. Notably, the annotations of the Zheyi dataset are not only used for the training process of the system but are also used for evaluation. The NIH-TCIA dataset contains 82 enhanced CT sequences, and the Zheyi dataset includes 70 instances. The axial resolution of CT images in the two datasets is 512×512 . Slice thickness of CT images in the NIH-TCIA dataset ranges from 0.5 mm to 1 mm, and the number of slices is in the range of 181 to 466, which are relatively high-resolution CT images. In contrast, the CT images of the Zheyi dataset range from 2.5 mm to 3 mm in layer thickness, and the slice number varies from 76 to 107.

2.6.2. Experimental Details. The system previously presented is expected to be able to cope with new sources of unlabeled CT images to construct an effective segmentation network. Taking NIH-TCIA as the labeled data center and Zheyi as the unlabeled data center, we validated the data adaptability and pancreas segmentation performance of the designed system for Zheyi images. We utilized PyTorch [51] in the Python environment to implement models and algorithms. The experiments were carried out with an NVIDIA TITAN V GPU with 12 GB memory and 2 Intel Xeon E5-2630 v4 CPUs. To guarantee the stability and reliability of the system, all trials are performed with a 5-fold cross-validation approach.

The execution time for necessary steps in the data-adaptive chain is listed in Table 1. These time data are statistically derived from the mean time of all sequences of the NIH-TCIA dataset. As can be seen in the table, the time for the data quality control module is mainly distributed over superresolution reconstruction. The duration of the complete data processing is about 16 seconds. The inference time of the segmentation model is only 2.37 seconds, and the postprocessing time consumes an average of 6.77 seconds. While 3D reconstruction takes up the majority of the time cost of the visualization module, text analysis is relatively less time consuming.

The experiments are conducted in the following steps: (1) The predefined model in the segmentation module performs supervised learning on the NIH-TCIA center to obtain the original baseline model. (2) The optimized baseline models

TABLE 1: Execution time for processing steps in the system.

Processing phase	Execution time (s)
CT format conversion	0.34
Intensity rescale	0.05
Rotation in angle ranges	2.18
ROI location	0.24
Superpixel reconstruction	13.44
Segmentation interference	2.37
Postprocessing	6.77
Visualization reconstruction	5.91
Textural metric analysis	0.025

are derived as in the first step from NIH-TCIA images with various image quality control measures. (3) The original and optimized baseline models are tested on Zheyi images, which undergo the same quality control measures as NIH-TCIA images corresponding to the model, respectively. In this way, the segmentation performance without the proposed data-adaptive system can be observed. (4) Data adaptation training is carried out on NIH-TCIA and Zheyi images to get the segmentation model applicable to unannotated Zheyi data, so as to investigate the segmentation effectiveness after data adaptation. The same data adaptation trainings are also performed on images on which various processing means in the quality control toolbox have been taken.

The results are mainly evaluated by the dice similarity coefficient (DSC) and mean intersection over union (mIoU), which indicate the similarity between the pancreas mask generated by the segmentation model and ground truth. The Hausdorff distance measures the deviation between the predicted mask and ground truth mask and is calculated as the distance of points in the two masks to each other's surfaces. The DSC and mIoU are defined as follows:

$$\text{DSC}(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|}, \quad (3)$$

$$\text{mIoU}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|},$$

where X represents the pancreas mask generated by the segmentation model and Y is the ground truth.

The Hausdorff distance is computed as

$$D_H(X, Y) = \{\max(\min_{x \in S(X)} d(x, S(Y))), \max(\min_{y \in S(Y)} d(y, S(X)))\}, \quad (4)$$

where $S(X)$ is the point set of the predicted pancreas mask and x represents the points in it. Similarly, $S(Y)$ is the point set of the ground truth mask and y represents the points in it. In the formula, $d(x, S(Y))$ indicates the distance from the point x to the surface formed by the point set $S(Y)$.

We display the average DSC, mIoU, and the Hausdorff distance on the test samples to demonstrate the average performance of the proposed pancreas segmentation system. In addition, considering the comprehensive presentation of segmentation masks, the visualization module will present the textual information and multidimensional images of the

pancreas segmentation results. The textual information contains the volume difference, Hausdorff distance, center-of-mass distance, and average symmetric surface distance (ASSD) to evaluate the segmentation results from multiple perspectives.

2.6.3. Baseline Performance. Figure 3 displays CT images from NIH and Zheyi datasets processed with multiple quality control methods. As can be observed, the raw data from the two datasets exhibit differences in various aspects

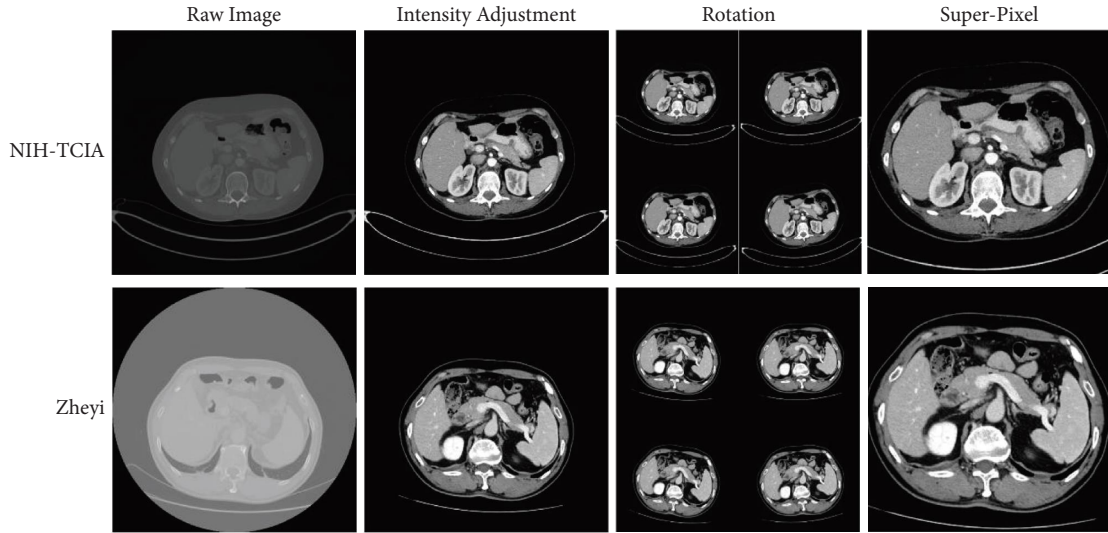


FIGURE 3: CT slice presentation with various image quality control means from two domains.

like the intensity distribution and scanning field of view (FOV) upon observation by a human. Moreover, it is also evident from the statistics shown in Table 2 that CT images of the two datasets differ significantly in terms of layer thickness. The range of intensity preserves most of information of abdominal organs and decreases the noise interference. The rotation augmentation with minor angles increases the diversity of FOV. The superresolution reconstruction algorithm frames out valid abdominal regions and eliminates redundant background information, so as to diminish heterogeneity caused by the clinical imaging process. It is observed from Figure 3 that, with the data quality control means, the visualized CT sequences narrowed the variation within and between datasets.

We first train the supervised baseline models on the NIH-TCIA dataset using raw and processed images and compare segmentation performance with existing pancreas segmentation methods. It is observed that the models trained with quality-controlled images outperform those trained on raw images, demonstrating that the image quality control toolbox designed in this paper is effective in decreasing intradomain heterogeneity. As summarized in Table 3, our model achieves a DSC of 85.45% after corresponding quality control means, which is superior to the performance of existing methods. The results show the mean value and standard deviation of the 5-fold cross-validation, which more reliably reflect the model’s performance on the whole dataset. The results indicate that the designed network in segmentation modules is of sufficient confidence to serve as the baseline models for subsequent adaptation experiments.

2.6.4. Data-Adaptive Performance. Table 4 lists the results of the unlabeled Zheyi dataset of directly tested and with data adaptation by the proposed system. When the NIH-TCIA baseline model is tested directly on the Zheyi dataset, only a DSC of 58.79% is obtained, indicating that almost half of the tissues are segmented incorrectly. Figure 4 presents the

TABLE 2: Datasets in the experiment.

Categories	Numbers	Axial resolution	Slice numbers	Slice thickness
NIH-TCIA	82	512×512	[181, 466]	[0.5, 1]
Zheyi	70	512×512	[76, 107]	[2.5, 3]

TABLE 3: Baselines segmentation performance on the NIH-TCIA dataset, where “sup-pixel” refers to superpixel reconstruction.

Model	Image options	Mean DSC (%)
Li et al. [19]	Raw images	83.06 ± 5.57
Cai et al. [20]	Raw images	82.40 ± 6.70
Yu et al. [16]	Raw images	84.50 ± 4.97
Ours	Raw images	83.13 ± 6.93
	Rotation	84.99 ± 4.86
	Sup-pixel	84.35 ± 6.10
	Rotation + sup-pixel	85.45 ± 4.51

The best performance value is presented in bold.

segmentation masks for three CT slices selected from the Zheyi dataset. As can be clearly observed, the model trained on the NIH-TCIA dataset exhibits significant degradation in pancreas segmentation on unseen images with heterogeneity. The masks in such cases do not effectively capture the pancreatic tissue information without data adaptation.

After data-adaptive training, the DSC score increases to 72.73% (a gain of 13.94%). Notably, the models trained with quality-controlled data demonstrate better performance when oriented to images from new sources. With rotation augmentation and superpixel reconstruction, the performance increases from 61.95% to 75.43% (a gain of 16.64%). To ensure the reliability of the experimental results, paired *t*-tests were performed on segmentation results of both the models with and without data adaptation. As listed in Table 4, the segmentation performance is significantly improved as observed, by the *p* value less than 0.01. Moreover,

TABLE 4: Adaptation performance of the system on the Zheyi dataset, where “sup-pixel” refers to superpixel reconstruction and “w/o” indicates without. The maximum value is presented in bold.

Labeled dataset	Unlabeled dataset	Image process options	Test w/o adaptation			Test with adaptation			p value of DSC
			Mean DSC (%)	mIoU (%)	D_H (mm)	Mean DSC (%)	mIoU (%)	D_H (mm)	
NIH-TCIA	Zheyi	Raw images	58.79	41.63	12.93	72.73	57.15	7.26	$3.13e-4$
		Rotation	62.73	45.70	9.37	74.91	60.00	5.57	$3.26e-7$
		Sup-pixel	61.58	44.49	11.44	73.64	58.28	6.89	$4.55e-3$
		Rotation + sup-pixel	61.95	44.88	10.16	75.43	60.55	6.34	$7.85e-6$

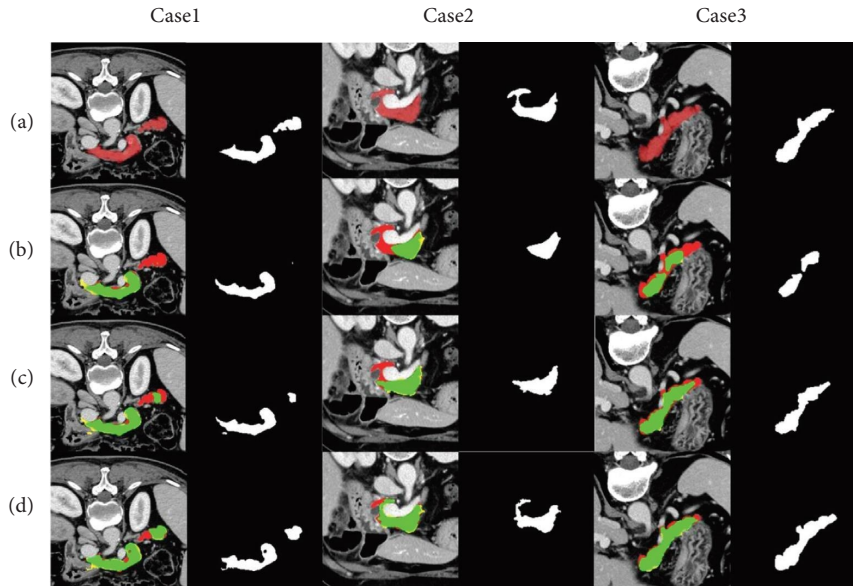


FIGURE 4: Qualitative segmentation results with different methods on several axial slices from the Zheyi dataset: manual label (red), prediction label (yellow), and overlap (green). (a) Ground truth. (b) Without adaptation. (c) With adaptation without the image quality control. (d) With adaptation with the image quality control.

after the image quality control, the difference in segmentation performance is more significant between before and after data adaptation. As can be observed in Figure 4, the pancreatic tissue can be correctly segmented out after the designed system. As expected, the model with data adaptation constructed by the system is capable of efficient segmentation for pancreas tissues and is of great significance in clinical decision-making.

2.6.5. Visualization. The visualization module provides structured textual information about the segmentation results and visualized images to assist physicians. We selected a CT sequence from the Zheyi dataset and allowed the physician to make corrections to annotations. In the designed system, the visualization results of this sample are shown in Figure 5. For the segmentation mask, statistics such as pancreas volume and size are calculated and displayed, and pancreas images in multiple dimensions are reconstructed. For physician manual corrections, the system then indicates the deviation between manual and system corrections in metrics such as the volume difference, mass distance, average symmetric surface distance (ASSD), Hausdorff

distance, and dice similarity coefficient. In addition, we calculate the 3D surface distances between two masks and reconstruct them as images.

3. Discussion

3.1. System Functionalities. In this research, a multifunctional image quality control toolbox was developed to standardize CT images from various aspects. The general rotation augmentation is employed to enhance the abundance of samples. However, the pancreas as a segmentation target is relatively small in volume in comparison with the whole abdomen. Therefore, the ordinary rotation augmentation operation is of little significance for the task. In this paper, an augmentation approach with minor rotation angles is designed to solve this problem. First, the statistical analysis of images is performed first to obtain the angle bias range of the abdomen, and then, we set a reasonable range of rotation angles with regard to data distribution characteristics. This operation reduces the discrepancy in images caused by various scanning fields of view and patient body positions with precise angle settings and enriches the samples in quantity. Furthermore, superresolution

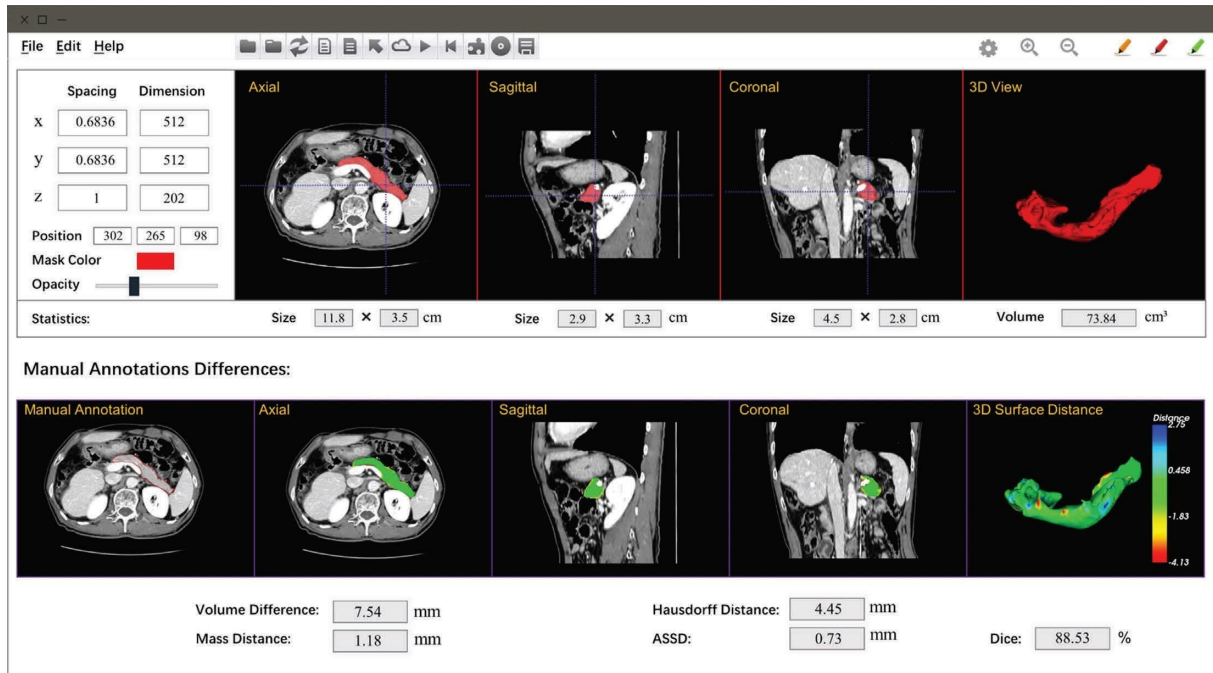


FIGURE 5: System interface display: visualization of multidimensional segmentation results in multiple dimensions and comparison with manual annotations.

reconstruction serves two functions. On the one hand, it standardizes the images in terms of resolution, layer thickness, field of view, etc., to minimize the heterogeneity among data domains. On the other hand, more fine-grained CT images are provided in the visualization module by improving the spatial resolution of the image, which is convenient for radiologists to review and diagnose.

As can be seen in Tables 3 and 4, the images with the quality control show less heterogeneity not only in intradomains but also in interdomains. The normalization of the images in terms of the angle, intensity scale, and region of interest lays the foundation for the subsequent transfer of the segmentation model. We performed paired t -tests for the ablation study in the image quality control module, and the statistical results are presented in Table 5. The results are statistically significantly improved by rotation and rotation plus superpixel reconstruction with a p value less than 0.01. The superpixel reconstruction only yields obviously improved results with a p value less than 0.05.

Moreover, in the visualization module, the reconstructed pancreas tissue displayed in Figure 5 of the visualization module demonstrates better spatial continuity, which facilitates the radiologist's diagnosis.

3.2. System Robustness. With the development of technology, the standardization and industrialization of deep learning are becoming more and more mature. As stated in the introduction, deep learning models have excellent performance in real-world applications but still face many challenges. This research improves the generalization of the segmentation model in the presence of new datasets, which addresses the potential problems of automatic pancreas

TABLE 5: Paired t -test results for the ablation study of raw images and images with the quality control, where "sup-pixel" refers to superpixel reconstruction.

Image options	Rotation	Sup-pixel	Rotation + sup-pixel
p value	$1.95e-3$	$3.41e-2$	$7.28e-4$

segmentation systems in practical deployment and applications. However, the robustness of the deep learning model is also an issue that needs to be focused.

Robustness typically denotes the property of a system to maintain its primary performance in the presence of fluctuations in some parameters [52, 53]. Normally, robustness is used to evaluate how stable a system is against uncertain utilization environments. It is widely known that deep learning systems are driven by big data, and thus, more data lead to richer feature extraction and higher quality model construction [54–56]. Therefore, when attention is paid to the small amount of data, we investigate whether the system is affected by performance degradation.

We adopted an external dataset for the robustness experiments. The dataset was collected by Vanderbilt University and contained 30 sequences, which was marked as the BTCV dataset [57, 58]. The resolution of CT images in this dataset is 512×512 pixels, the slice number ranges in [85, 198], and the layer thickness is in the interval of [2.5, 5] mm. The experiments were performed with the NIH dataset as a labeled center and the BTCV dataset as an unlabeled center. As can be seen in Table 6, after the proposed data-adaptive pancreas segmentation system with the image quality control, the performance on the BTCV dataset reached a DSC of 74.97%, which is 14.83% improvement

TABLE 6: Robust performance of the system on the BTCV dataset with 30 images, where “sup-pixel” refers to superpixel reconstruction and “w/o” indicates without.

Labeled dataset	Unlabeled dataset	Image process options	Performance by mean DSC (%)		<i>p</i> value
			Test w/o adaptation	Test with adaptation	
NIH-TCIA	BTCV	Raw images	60.14	71.19	$7.91e-3$
		Rotation	63.04	74.66	$1.82e-4$
		Sup-pixel	62.35	72.63	$9.54e-4$
		Rotation + sup-pixel	63.92	74.97	$6.34e-6$

The best performance values are presented in bold.

compared to a DSC of 60.14%, with directly testing original images with the model trained on the NIH dataset. The paired *t*-test indicates the statistically significant improvement with a *p* value less than 0.01. It is obvious that this result is consistent with that when the Zheyi dataset served as an unlabeled dataset, which means that the system still achieves a superior data adaptation performance. This result indicates that the designed system maintains a robust performance with respect to small datasets.

3.3. System Effectiveness. In clinical diagnosis, there is a demand for modern data analysis technology to mine CT image information and assist clinicians in improving diagnosis efficiency, thus refining the medical treatment process. The data-adaptive pancreas segmentation system proposed in this study utilizes the domain-invariant features from source-domain images with annotations for transfer learning to adapt the model to CT data features from different domains, thus implementing data-adaptive pancreatic segmentation. We developed a comprehensive image quality control toolbox to rectify data quality differences among different data domains. The image quality is controlled in several dimensions, including the image intensity range, scanning field of view, CT layer thickness, and valid region of interest. In addition, the image quality control toolbox provides physicians with more discriminative CT images. In terms of system architecture, we employ an adversarial learning scheme to implement feature distribution acquisition and feature adaptation among domains. In addition, this system fully takes into consideration the difference in the contribution of semantic information at different scales for domain adaptation and adopts weighted connections to realize the stitching of multiscale information to achieve a more stable and smooth adversarial learning structure. The effectiveness of the system has been validated with public datasets and real in-hospital data, and the robustness of the system has been demonstrated on a small dataset.

In summary, the system enables the establishment of the data-adaptive segmentation model by transfer learning, both interhospital and intrahospital across time lengths. The system eliminates the need for time-consuming and tedious annotation work by radiologists, which is of significant relevance to the automation of hospital treatment processes in real-time medical scenarios. In addition, it is a meaningful research area to combine image semantic segmentation techniques with other text analysis tasks to design an automatic pancreatic disease diagnosis system applicable to richer medical scenarios. In the subsequent research, we will

combine natural language processing and image interpretability to further improve the system, optimize the pancreatic disease diagnosis process, and promote the efficiency of physicians.

4. Conclusions

In this paper, we designed an end-to-end data-adaptive pancreas segmentation system with an image quality control toolbox. The system aims to address the problem of poor generalization capability exhibited by existing pancreas segmentation networks when oriented to data from different medical centers. For the visual task of label-free semantic segmentation, this research utilizes an adversarial learning method to obtain domain-invariant supervised information and construct the data-adaptive pancreas segmentation model. In addition, a functional image quality control toolbox was designed to provide multiple image preprocessing methods. The system works in an end-to-end manner and is easy to operate by physicians. The experimental results of public datasets and in-hospital datasets demonstrated that the end-to-end data-adaptive pancreas segmentation tool proposed in this paper can effectively assist in pancreas segmentation, and the generalization of segmentation networks was enhanced when facing images from different sources. This system is of considerable relevance in medical diagnosis and treatment and greatly promotes the development of precision and automated medical processes.

Data Availability

The CT sequences in the Zheyi dataset used to support the findings of this study are restricted by the First Affiliated Hospital of the Zhejiang University School of Medicine to protect the patient privacy. The data are not publicly available due to privacy restrictions.

Conflicts of Interest

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

Authors' Contributions

Yan Zhu and Peijun Hu equally contributed to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 12101571, 82172069, and

81702332), the Major Scientific Project of Zhejiang Lab (No. 2020ND8AD01), the Zhejiang Provincial Natural Science Foundation of China (No. LQ20H180001), and the Zhejiang Provincial Key Research and Development Program (No. 2020C03117).





References

- [1] X. Y. Jia, P. Ku, K. Wu et al., "Pancreatic cancer mortality in China: characteristics and prediction," *Pancreas*, vol. 47, no. 2, pp. 233–237, 2018.
- [2] L. C. Chu, S. Park, S. Kawamoto et al., "Utility of CT radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue," *American Journal of Roentgenology*, vol. 213, no. 2, pp. 349–357, 2019.
- [3] T. Tarver, "Cancer facts & figures 2012. American cancer society (ACS)," *Journal of Consumer Health on the Internet*, vol. 16, no. 3, pp. 366–367, 2012.
- [4] P. E. Oberstein and K. P. J. T. A. i. G. Olive, "Pancreatic cancer: why is it so hard to treat?" *Therapeutic Advances in Gastroenterology*, vol. 6, no. 4, pp. 321–337, 2013.
- [5] C. Yip, C. Dinkel, A. Mahajan, M. Siddique, G. J. R. Cook, and V. Goh, "Imaging body composition in cancer patients: visceral obesity, sarcopenia and sarcopenic obesity may impact on clinical outcome," *Insights into Imaging*, vol. 6, no. 4, pp. 489–497, 2015.
- [6] M. Ducreux, A. S. Cuhna, C. Caramella et al., "Cancer of the pancreas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, vol. 26, no. 5, pp. v56–v68, Sep 2015.
- [7] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille, "Deep supervision for pancreatic cyst segmentation in abdominal CT scans," in *Proceedings of the International Conference on Medical Image Computing & Computer-assisted Intervention*, Beijing China, October 2017.
- [8] M. Palmowski, N. Hacke, S. Satzler et al., "Metastasis to the pancreas: characterization by morphology and contrast enhancement features on CT and MRI," *Pancreatology*, vol. 8, no. 2, pp. 199–203, 2008.
- [9] A. Farag, L. Lu, H. R. Roth, J. Liu, E. Turkbey, and R. M. J. I. T. o. I. P. Summers, "A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 386–399, 2017.
- [10] M. S. Petrov, "Harnessing analytic morphomics for early detection of pancreatic cancer," *Pancreas*, vol. 47, no. 9, pp. 1051–1054, Oct 2018.
- [11] R. G. Singh, N. N. Nguyen, S. V. DeSouza, S. A. Pendharkar, and M. S. Petrov, "Comprehensive analysis of body composition and insulin traits associated with intra-pancreatic fat deposition in healthy individuals and people with new-onset prediabetes/diabetes after acute pancreatitis," *Diabetes, Obesity and Metabolism*, vol. 21, no. 2, pp. 417–423, 2019.
- [12] M. Oda, N. Shimizu, K. Karasawa, Y. Nimura, and K. Mori, "Regression forest-based atlas localization and direction specific atlas generation for pancreas segmentation," *Advanced Data Mining and Applications*, pp. 556–563, 2020.
- [13] Z. Li, J. Zhang, T. Tan et al., "Deep learning methods for lung cancer segmentation in whole-slide histopathology images -- the ACDC@LungHP challenge 2019," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 429–440, 2021.
- [14] S. Geert, *A Survey on Deep Learning in Medical Image Analysis*, Medical Image Analysis, China, 2017.
- [15] Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille, "A 3D coarse-to-fine framework for volumetric medical image segmentation," in *Proceedings of the 2018 International Conference on 3D Vision (3DV)*, Beijing China, April 2018.
- [16] Q. Yu, L. Xie, W. Yan, Y. Zhou, and A. L. Yuille, "Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach California, June 2018.
- [17] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach California, June 2020.
- [18] O. Oktay and W. Chen, *Attention U-Net: Learning where to Look for the Pancreas*, Medical Image Analysis Special Issue, China, 2018.
- [19] M. Li, F. Lian, and S. Guo, "Pancreas segmentation based on an adversarial model under two-tier constraints," *Physics in Medicine and Biology*, vol. 65, no. 22, Article ID 225021, 2020.
- [20] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," 2017, <https://arxiv.org/abs/1707.04912>.
- [21] C. Fang, G. Li, C. Pan, Y. Li, and Y. Yu, "Globally guided progressive fusion network for 3D pancreas segmentation," *Advanced Data Mining and Applications*, pp. 210–218, 2019.
- [22] H. R. Roth, L. Lu, N. Lay et al., "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical Image Analysis*, vol. 45, pp. 94–107, 2018.
- [23] H. R. Roth, L. Lu, A. Farag et al., "DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation," *Lecture Notes in Computer Science*, vol. 9349, pp. 556–564, 2015.
- [24] B. Giddwani, S. Pandey, H. Tekchandani, and S. Verma, "CSTA-2PID UNet: consecutive spatio-temporal attention for multi-scale 3D pancreas segmentation," in *Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, July 2020.
- [25] W. Wang, Q. Song, R. Feng, T. Chen, and J. Wu, "A fully 3D cascaded framework for pancreas segmentation," in *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, April 2020.
- [26] P. Hu, X. Li, Y. Tian et al., "Automatic pancreas segmentation in CT images with distance-based saliency-aware DenseASPP network," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1601–1611, 2021.
- [27] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [28] Z. C. Xu, Q. Zhang, F. Xu, D. Li, and R. Zhang, "Anthropometric measurements in 126 microtia reconstructions," *Facial Plastic Surgery*, vol. 29, no. 04, pp. 321–326, Aug 2013.
- [29] U. Koc, O. J. J. o. M. I. Taydas, and R. Sciences, "Investigation of the relationship between fatty pancreas and cholecystectomy using noncontrast computed tomography," *Journal of Medical Imaging and Radiation Sciences*, vol. 50, no. 2, pp. 220–226, 2019.
- [30] Y. H. Roh, B. K. Kang, S. Y. Song, C. M. Lee, Y. K. Jung, and M. Kim, "Preoperative CT anthropometric measurements and pancreatic pathology increase risk for postoperative

- pancreatic fistula in patients following pancreaticoduodenectomy,” vol. 15, no. 12, Article ID e0243515, 2020.
- [31] S. European, “Imaging what the radiologist should know about artificial intelligence - an ESR white paper,” *Insights into Imaging*, vol. 1, 2019.
- [32] G. Campanella, M. G. Hanna, L. Geneslaw et al., “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
- [33] Q. Liu, Q. Dou, L. Yu, and P. A. J. I. T. o. M. I. Heng, “MS-net: multi-site network for improving prostate segmentation with heterogeneous MRI data,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.
- [34] E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, and D. C. Barratt, “Inter-site variability in prostate segmentation accuracy using deep learning,” in *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018*, Granada, Spain, September 2018.
- [35] K. V. Sarma, S. Harmon, T. Sanford et al., “Federated learning improves site performance in multicenter deep learning without data sharing,” *Journal of the American Medical Informatics Association*, vol. 28, no. 6, pp. 1259–1264, 2021.
- [36] Z. Wang, Q. Liu, Q. J. I. J. o. B. Dou, and H. Informatics, “Contrastive cross-site learning with redesigned net for COVID-19 CT classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2806–2813, 2020.
- [37] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, “Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, 2017.
- [38] J. A. Onofrey, D. I. Casetti-Dinescu, A. D. Lauritzen, S. Sarkar, and X. Papademetris, “Generalizable multi-site training and testing of deep neural networks using image normalization,” in *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, NY China, June 2019.
- [39] F. J. Brooks and P. W. J. B. M. I. Grigsby, “Quantification of heterogeneity observed in medical images,” *BMC Medical Imaging*, vol. 13, no. 1, pp. 7–12, 2013.
- [40] L. Gao, T. Jiao, Q. Feng, and W. J. O. I. Wang, “Application of artificial intelligence in diagnosis of osteoporosis using medical images: a systematic review and meta-analysis,” *Osteoporosis International*, vol. 32, no. 7, pp. 1279–1286, 2021.
- [41] S. Kaur, M. J. Baine, M. Jain, A. R. Sasson, and S. K. J. B. i. M. Batra, “Early diagnosis of pancreatic cancer: challenges and new developments,” *Biomarkers in Medicine*, vol. 6, no. 5, pp. 597–612, 2012.
- [42] M. Lrousseau and L. Xiao, *Weakly Supervised Pan-Cancer Segmentation Tool*, Springer, Heidelberg, Germany, 2021.
- [43] E. Ferrante, O. Oktay, B. Glocker, and D. H. Milone, “On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains,” *Advanced Data Mining and Applications*, pp. 294–302, Beijing, 2018.
- [44] R. J. Xu, Z. L. Chen, W. M. Zuo, J. J. Yan, and L. Lin, “Deep cocktail network: multi-source unsupervised domain adaptation with category shift,” in *Proceedings of the presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Cvpr)*, New York, October 2018.
- [45] M. Dunnhofer, N. Martinel, and C. Micheloni, “Weakly-supervised domain adaptation of deep regression trackers via reinforced knowledge distillation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5016–5023, 2021.
- [46] L. Liu, L. Yang, and B. J. K.-B. S. Zhu, “Sparse feature space representation: a unified framework for semi-supervised and domain adaptation learning,” *Knowledge-Based Systems*, vol. 156, pp. 43–61, 2018.
- [47] M. Wang and W. Deng, “Deep visual domain adaptation: a survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [48] Y. F. Zhang, Y. Wei, Q. Wu et al., “Collaborative unsupervised domain adaptation for medical image diagnosis,” *IEEE Transactions on Image Processing*, vol. 29, pp. 7834–7844, 2020.
- [49] X. Luo, J. Chen, T. Song, G. Wang, and S. Zhang, “Semi-supervised medical image segmentation through dual-task consistency,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8801–8809, 2021.
- [50] O. Ronneberger, P. Fischer, and T. J. S. I. P. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing, Heidelberg, Germany, 2015.
- [51] A. Paszke, S. Gross, F. Massa, A. Lerer, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, China, 2019.
- [52] W. Zhou, J. S. Berrio, S. Worrall, and E. J. I. T. o. I. T. S. Nebot, “Automated evaluation of semantic segmentation robustness for autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1951–1963, 2020.
- [53] H. D. Tran, N. Pal, P. Musau, D. M. Lopez, and T. T. Johnson, “Robustness verification of semantic segmentation neural networks using relaxed reachability,” *Advanced Data Mining and Applications*, pp. 263–286, NY China, 2021.
- [54] Q. Zhang, L. T. Yang, Z. Chen, and P. J. I. F. Li, “A survey on deep learning for big data,” *Information Fusion*, vol. 42, pp. 146–157, 2018.
- [55] X. W. Chen and X. J. I. A. Lin, “Big data deep learning: challenges and perspectives,” *IEEE Access*, vol. 2, no. 2, pp. 514–525, 2014.
- [56] A. M. J. I. J. o. E. R. Zain and P. Health, “Detection of COVID-19 in chest X-ray images: a big data enabled deep learning approach,” *International Journal of Environmental Research and Public Health*, vol. 18, 2021.
- [57] X. Z. Landman Ba, J. E. Igelsias, M. Styner, T. R. Langerak, and A. Klein, “MICCAI Multi-Atlas Labeling beyond the Cranial Vault - Workshop and challenge,” 2015, <https://arxiv.org/abs/1711.06853>.
- [58] Z. Xu, C. P. Lee, M. P. Heinrich et al., “Evaluation of six registration methods for the human abdomen on clinically acquired CT,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1563–1572, 2016.

Research Article

Pterygium Screening and Lesion Area Segmentation Based on Deep Learning

Shaojun Zhu ^{1,2}, Xinwen Fang,¹ Yong Qian,³ Kai He ¹, Maonian Wu ^{1,2}, Bo Zheng,^{1,2}
and Junyang Song ⁴

¹School of Information Engineering, Huzhou University, Huzhou 313000, China

²Zhejiang Province Key Laboratory of Smart Management & Application of Modern Agricultural Resources, Huzhou University, Huzhou 313000, China

³Jiangsu Testing and Inspection Institute for Medical Devices, Nanjing 210000, China

⁴Department of Ophthalmology, The First People's Hospital of Huzhou, Huzhou 313000, China

Correspondence should be addressed to Junyang Song; ze8068e@163.com

Received 9 March 2022; Accepted 18 April 2022; Published 21 November 2022

Academic Editor: Yanwu Xu

Copyright © 2022 Shaojun Zhu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A two-category model and a segmentation model of pterygium were proposed to assist ophthalmologists in establishing the diagnosis of ophthalmic diseases. A total of 367 normal anterior segment images and 367 pterygium anterior segment images were collected at the Affiliated Eye Hospital of Nanjing Medical University. AlexNet, VGG16, ResNet18, and ResNet50 models were used to train the two-category pterygium models. A total of 150 normal and 150 pterygium anterior segment images were used to test the models, and the results were compared. The main evaluation indicators, including sensitivity, specificity, area under the curve, kappa value, and receiver operator characteristic curves of the four models, were compared. Simultaneously, 367 pterygium anterior segment images were used to train two improved pterygium segmentation models based on PSPNet. A total of 150 pterygium images were used to test the models, and the results were compared with those of the other four segmentation models. The main evaluation indicators included mean intersection over union (MIOU), IOU, mean average precision (MPA), and PA. Among the two-category models of pterygium, the best diagnostic result was obtained using the VGG16 model. The diagnostic accuracy, kappa value, diagnostic sensitivity of pterygium, diagnostic specificity of pterygium, and F1-score were 99%, 98%, 98.67%, 99.33%, and 99%, respectively. Among the pterygium segmentation models, the double phase-fusion PSPNet model had the best results, with MIOU, IOU, MPA, and PA of 86.57%, 78.1%, 92.3%, and 86.96%, respectively. This study designed a pterygium two-category model and a pterygium segmentation model for the images of the normal anterior and pterygium anterior segments, which could help patients self-screen easily and assist ophthalmologists in establishing the diagnosis of ophthalmic diseases and marking the actual scope of surgery.

1. Introduction

Pterygium is a common and frequently occurring disease in ophthalmology that affects the fibrovascular tissue on the ocular surface, resulting in eye irritation and inflammation [1, 2]. It can cause visual impairment or even blindness when the lesion covers most of the cornea [3, 4]. Corresponding treatment methods can be used to control pterygium development in the early stage. However, in the later stage, only surgery can be used to respect the lesion area for

treatment [5–7]. The diagnosis and surgery of pterygium require the localization of the lesion area. Currently, the most commonly used method is manual positioning by ophthalmologists based on anterior segment images. Manual positioning is slow and not precise, and different doctors may position different lesion ranges. Simultaneously, the early detection, diagnosis, and treatment of pterygium can better control or treat the disease. Therefore, a pterygium two-category model and a pterygium lesion area segmentation model were designed, which could initially screen the

pterygium and segment the lesion area accurately. These models can assist ophthalmologists in establishing the diagnosis of ophthalmic diseases and marking the scope of surgical resection.

With the close integration of artificial intelligence and ophthalmology, many studies have used deep learning models to assist in the diagnosis of ophthalmic diseases [8–13]. In terms of lesion segmentation, most studies have diagnosed glaucoma by segmenting the optic disc [14–16], and there have also been some studies on segmenting the blood vessels of fundus images to screen for related diseases [17–19]. Regarding the studies conducted on pterygium, some researchers used traditional machine learning [20] and deep learning methods to classify [21, 22] pterygium as normal and pterygium disease. A three-category pterygium model on normal, pterygium observation, and pterygium surgery periods was studied by some researchers [23]. Related studies have also been conducted on the localization and segmentation of pterygium lesions [24]. The above studies on pterygium classification and segmentation were conducted separately. In this study, the two studies were combined. The two-category model of pterygium was used on the anterior segment image, and the lesion area was segmented according to the pterygium image.

In this study, four deep-learning models were used to realize the two categories of pterygium for preliminary screening. Simultaneously, the team's improved models were used to segment the pterygium lesion area accurately, which could not only help patients understand the progression of pterygium but could also assist ophthalmologists in establishing the diagnosis of ophthalmic diseases and marking accurate lesion localization before surgery.

2. Materials and Methods

2.1. Data Source. The Affiliated Eye Hospital of Nanjing Medical University provided 1034 anterior segment images for this study. The data were obtained using two different brands of slit-lamp digital microscopes, and the quality of the images was high. Relevant personal information of the patient was removed from the image data provided. Therefore, it did not violate the patient's privacy. This study had no restrictions on the sex and age of patients, and the data provided did not contain related information of patients. Hence, this study had no relevant statistics.

The anterior segment images provided by the hospital in this study were of a single type of pterygium, which can only be diagnosed as normal or pterygium. The corresponding label (normal or pterygium) of each anterior segment image and lesion area annotation map of the pterygium anterior segment image along with the image were provided by the hospital. The marking standard for pterygium was as follows [25]: the normal anterior segment was characterized by the absence of evident hyperemia or proliferative bulge in the conjunctiva, with a transparent cornea. Figure 1 shows the images of the normal anterior segment Figure 1(a), the anterior segment of the pterygium Figure 1(b), and the labeling map of the lesion area Figure 1(c). Two professional ophthalmologists independently diagnosed the same

anterior segment. If the diagnosis results were consistent, it was the final diagnosis result. If the diagnosis results were inconsistent, the final diagnosis result was decided by an expert ophthalmologist. Labeling of the pterygium lesion area was performed by a trained professional ophthalmologist and confirmed by an expert ophthalmologist. If the lesion area was marked incorrectly, it was revised and reconfirmed until it was correct.

The pterygium two-category models were trained using 734 anterior segment images and were tested using 300 anterior segment images. The normal anterior segment and pterygium images in the training and test image data were equally divided. The pterygium lesion area segmentation models were trained using 367 pterygium images and tested using 150 pterygium images.

2.2. Classification Model Training. Deep learning classical classification models mainly include AlexNet [26], VGG16 [27], ResNet18 [28], and ResNet50 [28]. This study used the above four classical models to design two-category models on normal and anterior pterygium segment images. The network structures of these classical models are similar. The backbone networks of AlexNet and VGG16 include convolutional, pooling, and fully connected layers. ResNet adds a residual network structure. The model network structure is shown in Figure 2.

The aforementioned classical models require an input image size of 224×224 pixels. In this study, the adaptive average pooling method was added before the fully connected layer of the classical models. Therefore, the input size could be adjusted to the required size. The input image size was set to 336×224 pixels to adapt to the size of the original anterior segment image.

Normal and pterygium anterior segment images were divided into the training and validation sets in a 9:1 ratio. When training the pterygium two-category model, the original image was resized to 336×224 . The preprocessing method adopted a random rotation of $-3^\circ - 3^\circ$. The parameters trained by several models in the ImageNet [29] dataset were used as the initial parameters for the corresponding models. The loss function was the cross-entropy loss function. The learning rate of AlexNet and VGG16 was 0.001, the epoch was 30, the learning rate of ResNet18 and ResNet50 was 0.01, and the epoch was 100. The training parameters of the four models were iteratively updated to obtain the best model for the validation set as the final pterygium two-category model for each model.

2.3. Segmentation Model Training. Classical semantic segmentation models include U-Net [30], DeepLabv3+ [31], and PSPNet [32] models. The PSPNet and its improved models were used to segment the pterygium lesion areas in the anterior segment images of the pterygium. The results were compared with those of other segmentation models.

MobileNet [33] was used as the backbone network of PSPNet to extract features and obtain the feature map of the input image. Average pooling was used on the feature map at four different scales: 1×1 , 2×2 , 3×3 , and 6×6 . Subsequently,

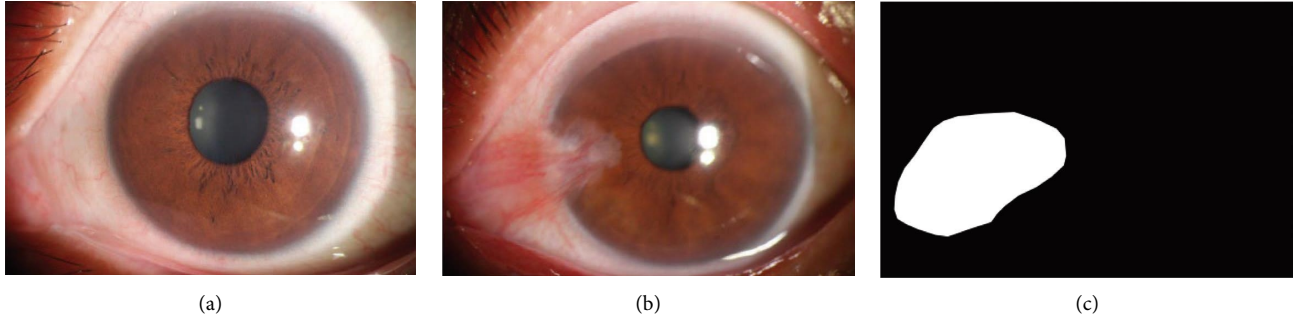


FIGURE 1: Images of normal anterior segment, pterygium anterior segment, and the labeling map of the lesion area.

the maps after average pooling with the same size as the feature map were obtained through bilinear interpolation. The feature map and maps after average pooling were spliced; finally, the segmented prediction map was obtained. As shown in Figure 3, PSPNet consists of Figures 3(a)–3(d) and 3(f), excluding Figure 3(g) and the stage upsampling module in PPM+.

The backbone network MobileNet was replaced by ResNet50 in the PSPNet, which can obtain better mean intersection over union (MIOU) and IOU results. Two improvements were made to the PSPNet model using ResNet50 as the backbone network. The first improvement was to increase the stage upsampling module, which first upsampled the feature map (1) to $\times 2$ through bilinear interpolation and then added the sampled feature map and feature map (2). The added feature map was upsampled and then added to the feature map (3) element by element. The added feature map was upsampled and then added to the feature map (4) element by element. The final added feature map was upsampled to 30×30 pixels. The feature map obtained after the stage upsampling module continued to be stacked to Figure 3(e) to obtain a new feature map. Therefore, a new pyramid pooling module (PPM+) was obtained, and the final prediction map through convolution was obtained. The first improvement model, called phase-fusion PSPNet, and the structure of this model are shown in Figure 3.

The second improvement was mainly aimed at the feature extraction of the ResNet50 network. The shallow feature maps of the ResNet50 third-layer input were input into the PPM+ module, and the results obtained after convolution were the same as those obtained after PPM+ and convolution in the phase-fusion PSPNet. Feature maps were added, and the final prediction map was obtained after upsampling. As shown in Figure 4, box A in the figure represents the newly added feature extraction and fusion module in the phase-fusion PSPNet.

A total of 367 pterygium anterior segment images were selected to train the segmentation models, of which 330 and 37 were used as the training and validation sets, respectively. Both sides of the short side of the input image were lengthened so that the length of the short side was the same as the length of the long side. Then, the image became a square, and the increased part was filled with gray (R, G, B are all 128), and the square image size was resized to

473×473 as the input image for training. The number of training epochs was 80, and the model with the best validation result was selected as the final segmentation model.

2.4. Statistical Analyses. The Statistical Package for the Social Sciences version 22.0 software was used for statistical analyses of the two-category models. The count data are expressed as the number and percentage of images. The sensitivity, specificity, F1-score, area under the curve (AUC), kappa value, and other indicators were used to evaluate the diagnosis results of the expert diagnosis and model groups. A receiver operating characteristic (ROC) curve was drawn to compare the results of the models. Segmentation of pterygium lesions was evaluated using four indicators: IOU, MIOU, PA, and MPA.

2.5. Calculation Methods. The calculation methods of IOU, MIOU, PA, and MPA are as follows:

$$\begin{aligned} \text{IOU} &= \frac{p_i \cap g_i}{p_i \cup g_i}, \\ \text{MIOU} &= \frac{1}{k+1} \sum_{i=0}^k \frac{p_i \cap g_i}{p_i \cup g_i}, \\ \text{PA} &= \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}}, \\ \text{MPA} &= \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}}, \end{aligned} \quad (1)$$

where p_i is the segmented area, g_i is the real area, k is the number of classes (excluding background classes), p_{ii} is the number of correctly predicted pixels, and p_{ij} and p_{ji} are the numbers of incorrectly predicted pixels.

3. Results

3.1. Results of Classification. In this study, four models were tested with 150 images of normal and pterygium anterior segments, and the VGG16 model had the best results, with an accuracy of 99% and a kappa value of 98%. The sensitivities of diagnosing normal and pterygium were 99.33%

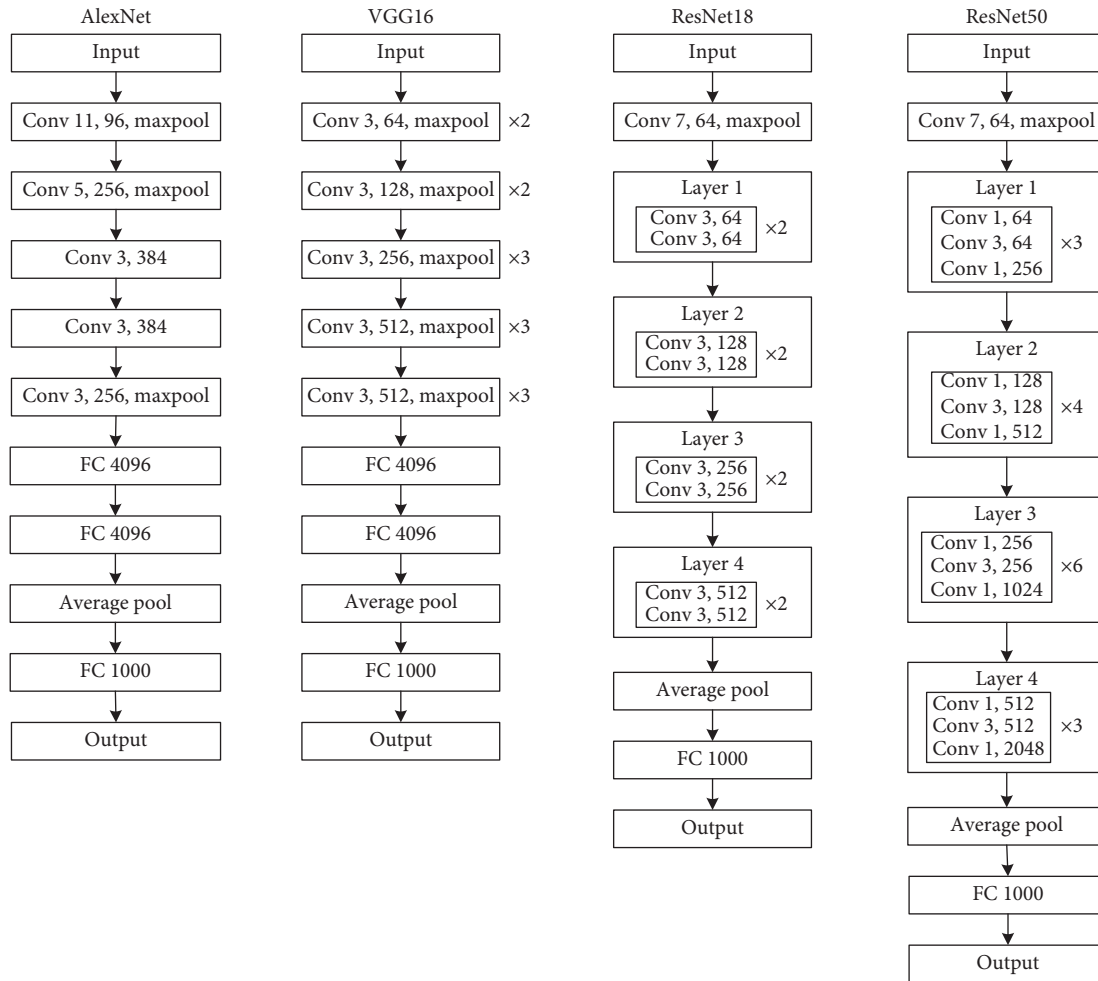


FIGURE 2: The model network structures of AlexNet, VGG16, and ResNet.

and 98.67%, respectively, the specificities were 98.67% and 99.33%, and the AUCs were 98.67% and 99.33%, respectively. The diagnostic results and evaluation indicators of the four models are shown in Tables 1 and 2, respectively, and the ROC curve is shown in Figure 5.

3.2. Results of Segmentation Models. A total of 150 pterygium anterior segment images were used to test U-Net, DeepLabv3+, PSPNet (based on MobileNet and ResNet50), and the two improved models based on PSPNet. The pterygium segmentation results for the six models are presented in Table 3.

As shown in Table 3, the PSPNet model based on ResNet50 performed better than the U-Net, DeepLabv3+, and MobileNet-based PSPNet models for the MIOU, IOU, and MPA indicators. The double phase-fusion PSPNet was obtained after two improvements on the ResNet50-based PSPNet; its MIOU, IOU, MPA, and PA were 86.57%, 78.1%, 92.3%, and 86.96%, respectively. The result of the PA was slightly worse than that of the PSPNet model based on MobileNet, but other indicators yielded the best results. The segmentation results of the phase-fusion and double phase-fusion PSPNets are shown in Figure 6.

4. Discussion

Most patients with pterygium are outdoor workers, such as fishermen and farmers [34]. In the early stage of the disease, there will be no significant effect on the patient, and the symptoms are similar to ordinary inflammation, which will not attract the attention of the patient. Thus, the disease gradually develops to the stage where surgical treatment is necessary. The pterygium two-category and lesion segmentation model can help patients screen for the disease by themselves and pay attention to the progress of the lesion area. Therefore, the patient has an intuitive understanding of the disease's progress and then immediately visits a hospital for diagnosis and treatment, finally obtaining a good therapeutic effect.

Four classical classification models were selected to diagnose whether the anterior segment images were normal or pterygium images. The normal anterior segment was clearly distinguished from the anterior segment of the pterygium. Subsequently, the features can be extracted better without a complex network structure. Therefore, the VGG16 model yielded the best results. ResNet18 and ResNet50 have more complex network structures, whereas the AlexNet network structure is slightly simpler; therefore, the diagnosis results of these models were both worse than those of VGG16.

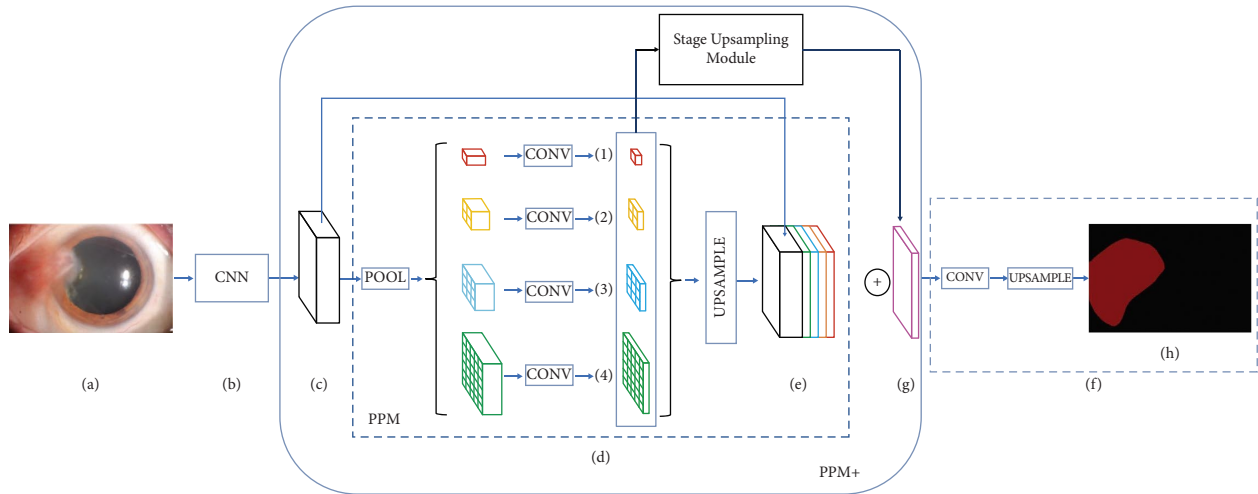


FIGURE 3: The structures of PSPNet and phase-fusion PSPNet. (a) represents the input image; (b) represents the feature extraction network, the feature extraction part of MobileNet or ResNet50; (c) represents the feature map extracted by the feature extraction network; (d) represents the pyramid pooling module; (e) represents the feature map output by the pyramid pooling module; (f) represents the output image; (g) represents the feature map formed by stage upsampling module; (h) represents the output image.

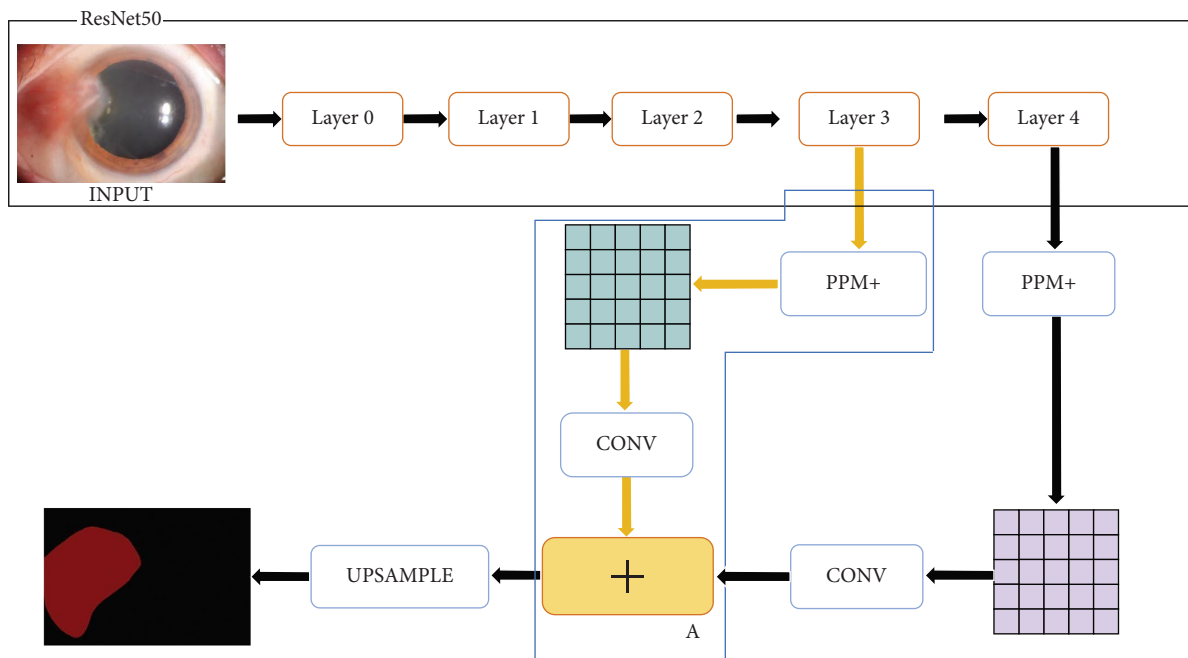


FIGURE 4: The structure of the double phase-fusion PSPNet.

TABLE 1: Diagnostic results of the four models.

Clinical	AlexNet diagnosis		VGG diagnosis		ResNet18 diagnosis		ResNet50 diagnosis		Total
	Normal	Pterygium	Normal	Pterygium	Normal	Pterygium	Normal	Pterygium	
Normal	147	3	149	1	143	7	140	10	150
Pterygium	6	144	2	148	12	138	11	139	150
Total	153	147	151	149	155	145	151	149	300

In 2018, Wan Zaki et al. [20] used support vector machine (SVM) and artificial neural network methods to study the two categories of pterygium. The data used in the study were obtained from four datasets, including 2692 and 325

images of the normal anterior and pterygium anterior segments, respectively. The result obtained using the SVM method was better, with sensitivity, specificity, and AUC values of 88.7%, 88.3%, and 0.956, respectively. In 2019,

TABLE 2: Evaluation index results of the four models.

Model	AlexNet		VGG16		ResNet18		ResNet50	
	Normal	Pterygium	Normal	Pterygium	Normal	Pterygium	Normal	Pterygium
Sensitivity	98.00%	96.00%	99.33%	98.67%	95.33%	92.00%	93.33%	92.67%
Specificity	96.00%	98.00%	98.67%	99.33%	92.00%	95.33%	92.67%	93.33%
F1-score	97.03%	96.97%	99.00%	99.00%	93.77%	93.56%	93.02%	92.98%
AUC	0.97		0.99		0.94		0.93	
95%CI	0.95–0.99		0.98–1		0.91–0.97		0.90–0.96	
Kappa	94.00%		98.00%		87.33%		86.00%	
Accuracy	97.00%		99.00%		93.67%		93.00%	

AUC: area under the curve; CI: confidence interval.

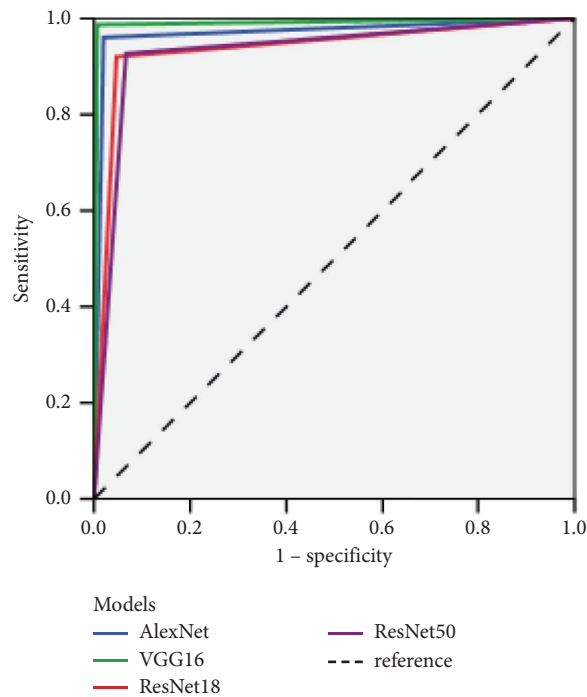


FIGURE 5: Receiver operating characteristic curve of the four models.

TABLE 3: Evaluation index results of the six models.

Model	MIOU (%)	IOU (%)	MPA (%)	PA (%)
U-Net	83.33	72.77	89.5	81.5
DeepLabv3+	83.91	73.98	91.45	86.39
PSPNet (MobileNet)	74.25	60.38	89.52	88.89
PSPNet (ResNet50)	85.4	76.27	91.92	86.7
Phase-fusion PSPNet	86.31	77.64	91.91	86.1
Double phase-fusion PSPNet	86.57	78.1	92.3	86.96

MIOU: mean intersection over union; IOU: intersection over union; MPA: mean average precision; PA: average precision.

Zulkifley et al. [21] used the convolutional neural network method to diagnose pterygium based on 60 normal and anterior pterygium segment images, with diagnostic sensitivity and specificity of 95% and 98.3%, respectively. In this study, the sensitivity, specificity, and AUC of the VGG16 model for the diagnosis of pterygium were 98.67%, 99.33%, and 0.99, respectively, which are higher than those reported by other researchers. The VGG16 model can better extract image features. The training data were balanced, and the

number of training images was greater than that in the literature [21]; thus, better results were obtained.

Classical (U-Net, DeepLabv3, PSPNet) and improved models based on PSPNet (phase-fusion PSPNet and double phase-fusion PSPNet) were used to segment pterygium. According to Table 3, the improved model had better segmentation results. The improved model extracted more features from the pterygium image, which can fully combine local features, global features, and features at different levels

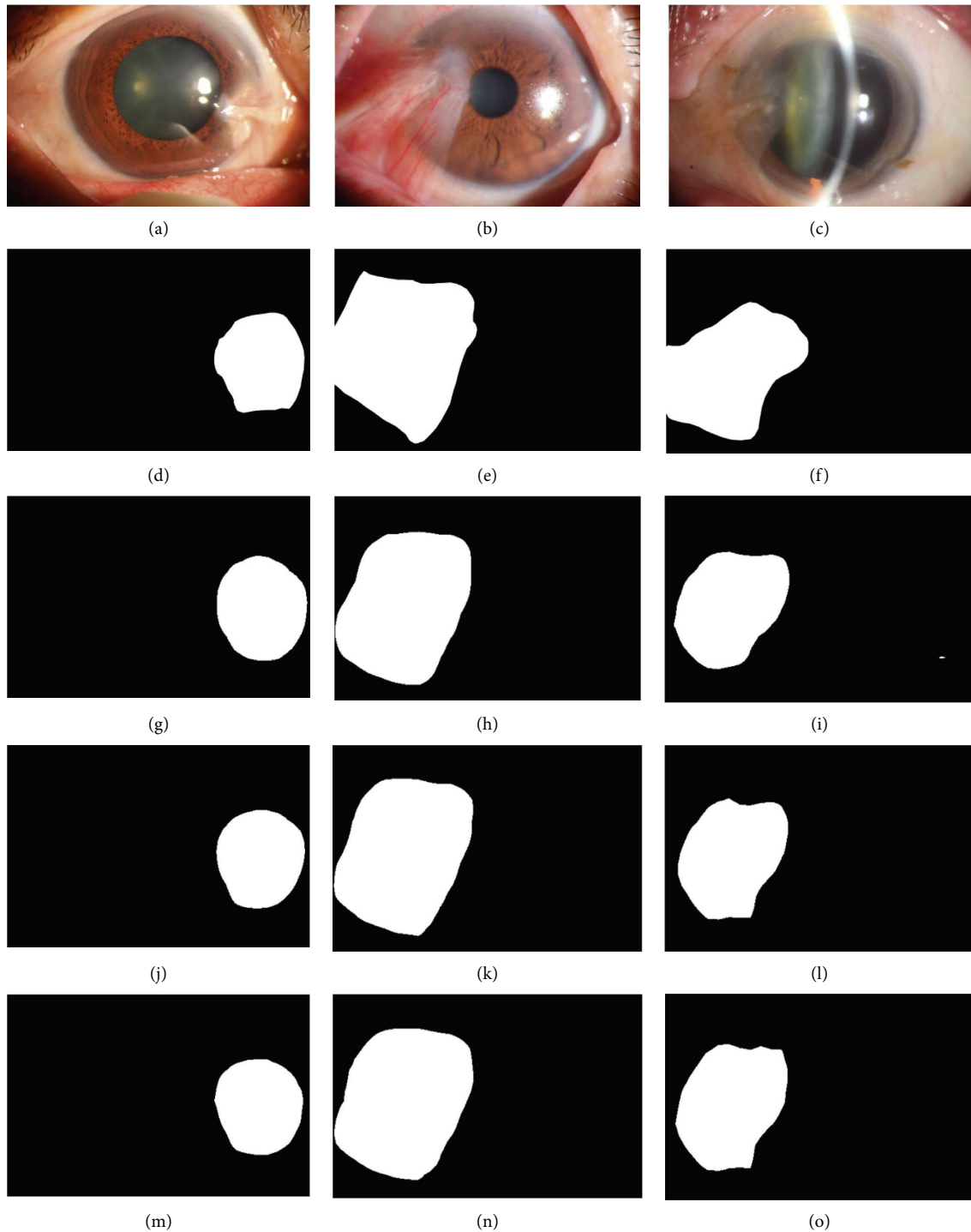


FIGURE 6: The segmentation results of the phase-fusion PSPNet and double phase-fusion PSPNet. Figures (a)–(c) show the original pterygium images; figures (d)–(f) show the real label of the pterygium lesion area of figures (a)–(c); figures (g)–(i) show the segmentation results of the PSPNet (ResNet50) model; figures (j)–(l) show the segmentation results of the phase-fusion PSPNet model; figures (m)–(o) show the segmentation results of the double phase-fusion PSPNet model.

in the feature extraction network. Their structures can lose less feature information and obtain better segmentation results.

Abdani et al. [24] used Dense Deeplabv2 to segment pterygium in 2020. Compared with the Deeplabv1, Dense Deeplabv1, and Deeplabv2 models, the best MIOU result

was 83.81%. The same team designed Group-PPM-Net to segment pterygium in 2021, and the best MIOU result was 86.32% [35]. Cai et al. [36] used DRUNet and SegNet to segment pterygium, and the best IOU was 60.8%. The MIOU and IOU results obtained using the double phase-fusion PSPNet in this study were 86.57% and 78.1%, respectively.

The study in [24, 35] had 328 pterygium images, which are less than this study in terms of the number of training images. Simultaneously, the improved model can better extract image features and obtain better results.

Figure 6 shows that there is a certain gap between the segmentation and real results. The models can only assist physicians in determining the position before the surgery. Physicians also need to calibrate and confirm its boundary and range. More labeled data are required to further train the models, or a more sensitive and efficient model is expected. Therefore, the predicted segmentation results are closer to the real segmentation results.

5. Conclusions

A pterygium two-category model and a pterygium segmentation model for the images of the normal anterior and pterygium anterior segments were designed in this study, which could help patients self-screen easily and assist ophthalmologists in establishing the diagnosis of ophthalmic diseases and marking the actual scope of surgery. The VGG16 model can obtain the best diagnostic result among the four two-category models, and the double phase-fusion PSPNet model had the best results among the pterygium segmentation models. The two models could help patients self-screen easily and assist ophthalmologists in marking the actual scope of surgery.

Data Availability

The data used in this study can obtain from the corresponding author with a reasonable request.

Conflicts of Interest

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

Acknowledgments

The study was supported by the National Natural Science Foundation of China (No. 61906066), Natural Science Foundation of Zhejiang Province (No. LQ18F020002), and Science and Technology Planning Project of Huzhou Municipality (No. 2016YZ02).

References

- [1] J. Chen and W. Li, "Progress in pathogenesis and treatment of pterygium," *Modern Journal of Integrated Traditional Chinese and Western Medicine*, vol. 29, no. 12, pp. 1364–1368, 2020.
- [2] L. Jin, Z. P. Yan, N. Chen et al., "Preliminary study of pterygium intelligent diagnosis system based on deep learning," *Artificial Intelligence*, vol. 3, no. 3, pp. 48–55, 2021.
- [3] X. P. Pan, L. L. Huang, W. H. Yang et al., "Study on application of quality nursing in dry eye after the surgery of primary pterygium," *China Modern Doctor*, vol. 58, no. 19, pp. 172–175, 2020.
- [4] X. L. Zhang, X. Yang, and M. Zhang, "Evaluation of therapeutic contact lenses used in pterygium surgery combined with limbal conjunctival autograft transplantation," *International Eye Science*, vol. 19, no. 19, pp. 867–869, 2019.
- [5] L. H. Tao and W. H. Yang, "A study on changes in corneal refractive and corneal surface regularity before and after pterygium excision surgery," *China Medical Herald*, vol. 15, no. 15, pp. 115–118, 2018.
- [6] Q. Shen and W. H. Yang, "Effect of bandage contact lens on pain and corneal epithelium healing condition after pterygium excision combined with autologous conjunctival flap graft transplantation," *China Medical Herald*, vol. 14, no. 15, pp. 131–134, 2017.
- [7] Y.-C. Wang, F. K. Zhao, Q. Liu, Z.-Y. Yu, J. Wang, and J.-S. Zhang, "Bibliometric analysis and mapping knowledge domain of pterygium: 2000-2019," *International Journal of Ophthalmology*, vol. 14, no. 6, pp. 903–914, 2021.
- [8] B. Zheng, Q. Jiang, B. Lu et al., "Five-category intelligent auxiliary diagnosis model of common fundus diseases based on fundus images," *Translational Vision Science & Technology*, vol. 10, no. 7, p. 20, 2021.
- [9] L. P. Cen, J. Ji, J. W. Lin et al., "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Communications*, vol. 12, no. 1, p. 4828, 2021.
- [10] W.-H. Yang, B. Zheng, M.-N. Wu et al., "An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research," *Diabetes Therapy*, vol. 10, no. 5, pp. 1811–1822, 2019.
- [11] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [12] M. Mateen, J. Wen, D. Nasrullah, S. Song, and Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, p. 1, 2019.
- [13] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, p. 2054, 2017.
- [14] L. Dong and F. Li, "Joint optic disk and cup segmentation for glaucoma screening using a region-based deep learning network," *Scientific Journal of Control Engineering*, vol. 18, pp. 1–10, 2021.
- [15] H. P. Liu, Y. H. Zhao, H. X. Hou et al., "Optic disc and cup segmentation by combining context and attention," *Journal of Image and Graphics*, vol. 26, no. 5, pp. 1041–1057, 2021.
- [16] N. Thakur and M. Juneja, "Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma," *Biomedical Signal Processing and Control*, vol. 42, pp. 162–189, 2018.
- [17] Z. Tang, X. Zhang, G. Yang et al., "Automated segmentation of retinal nonperfusion area in fluorescein angiography in retinal vein occlusion using convolutional neural networks," *Medical Physics*, vol. 48, no. 2, pp. 648–658, 2021.
- [18] G. Z. Xu, W. J. Lin, S. Chen, W. Kuang, B. Lei, and J. Zhou, "Fundus vessel segmentation method combined with U-Net and adaptive threshold pulse coupled neural network," *Journal of Computer Applications*, vol. 40, pp. 825–832, 2021.
- [19] J. Morano, Á. S. Hervella, J. Novo, and J. Rouco, "Simultaneous segmentation and classification of the retinal arteries and veins from color fundus images," *Artificial Intelligence in Medicine*, vol. 118, pp. 102–116, 2021.
- [20] W. M. D. Wan Zaki, M. Mat Daud, S. R. Abdani, A. Hussain, and H. A. Mutalib, "Automated pterygium detection method of anterior segment photographed images," *Computer*

- Methods and Programs in Biomedicine*, vol. 2018, no. 154, pp. 71–78, 2018.
- [21] M. A. Zulkifley, S. R. Abdani, and N. H. Zulkifley, “Pterygium-Net: a deep learning approach to pterygium detection and localization,” *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 34563–34584, 2019.
- [22] S. R. Abdani, M. A. Zulkifley, and A. Hussain, “Compact convolutional neural networks for pterygium classification using transfer learning,” in *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 140–143, IEEE, Kuala Lumpur, Malaysia, 17 September 2019.
- [23] B. Zheng, Y. Liu, K. He et al., “Research on an intelligent lightweight-assisted pterygium diagnosis model based on anterior segment images,” *Disease Markers*, vol. 2021, p. 7651462, 2021.
- [24] S. R. Abdani, M. A. Zulkifley, and A. M. Moubark, “Pterygium tissues segmentation using densely connected deeplab,” in *Proceedings of the 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 229–232, Malaysia, 18 April 2020.
- [25] G. Koray, E. Kuddusi, T. Duygu, and J. Colin, “Effect of pterygia on refractive indices, corneal topography, and ocular aberrations,” *Cornea*, vol. 30, no. 1, pp. 24–29, 2011.
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, Curran Associates Inc., Lake Tahoe Nevada, 3 December 2012.
- [27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, <https://arxiv.org/pdf/1409.1556.pdf>.
- [28] K. He, X. Zhang, S. Ren, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, Las Vegas, NV, USA, 27 June 2016.
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: a large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pp. 248–255, IEEE, Miami, FL, USA, 20–25 June 2009.
- [30] 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, IEEE, Cambridge, UK, 19 September 2015.
- [31] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, Springer, Munich, Germany, 8 September 2018.
- [32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, IEEE, Honolulu, HI, USA, 21 July 2017.
- [33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: inverted residuals and linear bottlenecks,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, IEEE, Salt Lake City, UT, USA, 18 June 2018.
- [34] W.-P. Zhou, Y.-F. Zhu, B. Zhang, W.-Y. Qiu, and Y.-F. Yao, “The role of ultraviolet radiation in the pathogenesis of pterygia (Review),” *Molecular Medicine Reports*, vol. 14, no. 1, pp. 3–15, 2016.
- [35] S. R. Abdani, M. A. Zulkifley, and N. H. Zulkifley, “Group and shuffle convolutional neural networks with pyramid pooling module for automated pterygium segmentation,” *Diagnostics*, vol. 11, no. 6, pp. 1–16, 2021.
- [36] W. Cai, J. Xu, K. Wang et al., “EyeHealer: a large-scale anterior eye segment dataset with eye structure and lesion annotations,” *Precision Clinical Medicine*, vol. 4, no. 2, pp. 85–92, 2021.

Research Article

Comparing Conventional and Deep Feature Models for Classifying Fundus Photography of Hemorrhages

Tamoor Aziz,¹ Chalie Charoenlarnnoppa¹ ,¹ and Srijidtra Mahapakulchai²

¹School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum-Thani, Thailand

²Department of Electrical Engineering, Kasetsart University, Bangkok, Thailand

Correspondence should be addressed to Chalie Charoenlarnnoppa; chalie@siit.tu.ac.th

Received 11 February 2022; Revised 27 March 2022; Accepted 8 April 2022; Published 19 November 2022

Academic Editor: Huiying Liu

Copyright © 2022 Tamoor Aziz et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Diabetic retinopathy is an eye-related pathology creating abnormalities and causing visual impairment, proper treatment of which requires identifying irregularities. This research uses a hemorrhage detection method and compares the classification of conventional and deep features. Especially, the method identifies hemorrhage connected with blood vessels or residing at the retinal border and was reported challenging. Initially, adaptive brightness adjustment and contrast enhancement rectify degraded images. Prospective locations of hemorrhages are estimated by a Gaussian matched filter, entropy thresholding, and morphological operation. Hemorrhages are segmented by a novel technique based on the regional variance of intensities. Features are then extracted by conventional methods and deep models for training support vector machines and the results are evaluated. Evaluation metrics for each model are promising, but findings suggest that comparatively, deep models are more effective than conventional features.

1. Introduction

Diabetic retinopathy (DR) is a prevalent cause of vision loss among working-age adults. The statistics of DR patients have been projected to be 191 million by the year 2030 [1]. Initially, its diagnosis is almost impossible due to the absence of distinct symptoms. DR identification is crucial in the early phase because its timely treatment and medication may reduce the progression rate by 57% [2], approximately. Therefore, an annual examination is recommended for diabetes patients. Several surveys were conducted which highlighted that diabetes patients refused to have regular checkups because of lack of symptoms, time-consuming diagnostic process, and limited access to ophthalmologists [3]. DR falls into two main categories: nonproliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). NPDR weakens capillary walls and yields leakage of blood from vessels that compile microaneurysms (MAs). Later, ruptures turn MAs into hemorrhages (HEs). MAs and HEs are often term as red lesions. When the disease

progresses, the NPDR turns into PDR and angiogenic factors originate from new blood vessels, called neovascularization.

Eye experts use fluorescein angiography (FA), optical coherence tomography (OCT), and fundus photography for the screening of DR [4]. FA is used to identify locations where blood vessels are closed or ruptured. OCT screening method provides a cross-sectional overview to determine the amount of fluid in retinal tissue and is used to evaluate the effectiveness of the adopted treatment. Next, fundus photography is an easy and immediate screening technique for documentation of DR progression and its improvement over time. Laser treatment, eye injections, or eye surgery can be recommended by an ophthalmologist in a case when DR is intimidating to the eyesight [5]. Laser treatment helps to cure the neovascularization of blood vessels at the back of the eye. It stabilizes the changes that occur because of diabetes. Eye injection is used in the case of PDR to stop the emergence of new blood vessels. The benefit of this method is the improvement in eyesight. However, steroid injection produces excessive pressure inside the eye that may cause

blood clots. Eye surgery is performed on an eye when a massive amount of blood accumulates in the vitreous humour. The eye specialist removes some jelly-like substance that fills the space back of the eye.

Retinal fundus imaging is preferred for the initial screening phase because of its easy assessment and it is less expensive. Ophthalmologists capture retinal images using a fundus camera with an appropriate field of view (FOV). Early signs of DR are observed to determine its stage for medical prescription. Contrary to benefits, HEs detection is challenging due to certain impediments. Factors like blurriness and poor illumination may reduce diagnostic accuracy. Uneven lighting conditions may produce dark shades in retinal images, which misleads detection. Blood vessels share intensity characteristics with HEs because of their similar appearance. Sometimes, HEs can be adjoined with blood vessels because they originate from them. Detection of those HEs is imperative for early screening of DR. HEs that reside at the retinal periphery are blended with the black background and are problematic to identify for computer-aided automatic detection. Appropriate selection of a deep network for classifying HEs is crucial to obtain promising results. Hence, these constraints cause HEs detection to be a challenging task. Figure 1 shows the characteristics of fundus images.

The risk of human interpretation necessitates an efficient algorithm that can segment and classify hemorrhages effectively. The computer-based second interpreter expedites the diagnostic process and assists ophthalmologists in assessment. The proposed methodology addresses the problems of fundus images. A novel gradient-based adaptive gamma correction adjusts the brightness of fundus images adaptively. An automatic detection scheme is proposed by image calibration. The proposed smart-window-based adaptive thresholding (SWAT) segments the objects while isolating hemorrhages from blood vessels and the retinal periphery. Objects are classified based on the intuitive selection of conventional features by manipulating the visual appearance of hemorrhage in retinal fundus images. The statistical comparison of features for HEs classification using conventional and deep models is provided. This research study uses various architectures of deep models to analyze which is suitable for HEs classification. Identification and detection of hemorrhages that resided at the retinal periphery and connected with blood vessels are the hallmarks of the proposed algorithm.

2. Related Work

N. Figueiredo et al. [6] proposed an algorithm to detect retinal abnormalities at the early stage of DR. This technique uses three classifiers for detection, including HEs. Novel features based on the inherent properties of lesions are used for classification. These features are extracted from wavelet bands, Hessian multiscale analysis, variational segmentation, and texture decomposition. The sensitivity and specificity of HEs detection are 86% and 90%, respectively. Tang et al. [7] propounded a splat feature classification method for HEs detection. The retinal image is partitioned into

nonintersected segments called splats. The formation of each splat is based on similar color and spatial information. Shape, texture, the intersection of neighboring splats, and filter bank information are used. Later, optimum features are selected using the filter approach. This method achieves a 0.96 receiver operating characteristic curve (ROC). Detection of early signs of DR was proposed by Junior and Welfer [8]. The technique is based on mathematical morphology to remove fovea and blood vessels because they share the intensity characteristics with HEs. This approach achieves 87.69% sensitivity and 92.44% specificity. The gradual elimination of blood vessel-based HEs detection technique is presented by Zhou et al. [9]. This technique deals with the HEs that are attached to the blood vessels by segmenting the dark regions, retinal vasculature, and HEs candidates. A binary image is manipulated further for providing good vascular connectivity and then removed gradually. A support vector machine (SVM) is trained using 49 features to classify candidates into non-HEs and HEs. The technique benchmarks promising results for two datasets. Karkuzhali and Manimegalai detect retinal abnormalities to classify fundus images into various DR stages [10]. Median filter, shade correction, Gaussian, and modified Kirsch filter are used to suppress noise and quality enhancement in the preprocessing stage. The image is divided into nonoverlapping patches of similar gray information. The Super-pixel method is applied to obtain the uneven grids. The gradient magnitude with toboggan segmentation is used for HEs segmentation. Feature vector and classifier mark images into various stages of DR.

The automatic segmentation of retinal lesions is presented by Tan et al. [11] using a novel single convolutional neural network (CNN). The proposed CNN model consists of 10 layers that classify retinal lesions simultaneously. The technique normalizes input images before network training. The proposed CNN model marks 0.6257 sensitivity on a large dataset. Another automatic detection of retinal lesions is proposed by Lam et al. [12]. The technique uses 1,324 image patches for the training of the deep network. The sliding window method considers all the patches from the testing image to generate the probability map. This CNN model provides promising results for each type of lesion. A deep learning approach was propounded by Islam et al. [13] for the detection and grading of DR. The technique focuses on early DR detection using a novel CNN network. The method is tested on a publicly available Kaggle dataset and reports a 0.851 quadratic weighted kappa score and 0.844 area under the curve. The technique for the detection of red lesions using the You Look Only Once (YOLO-V3) algorithm is proposed by Pal et al. [14]. The contrast of the green channel is enhanced and then the bounding boxes of red lesions are obtained using the YOLO algorithm. Detection is performed using Darknet53, and logistic regression provides the confidence level of an object. The model is trained using Adam optimization and tested for red lesion detection. Objectness threshold is employed to reduce the false predictions. This technique scores 83.33% of the average precision. A synergy deep learning model is presented by Shankar et al. [15] to classify fundus images into DR stages.

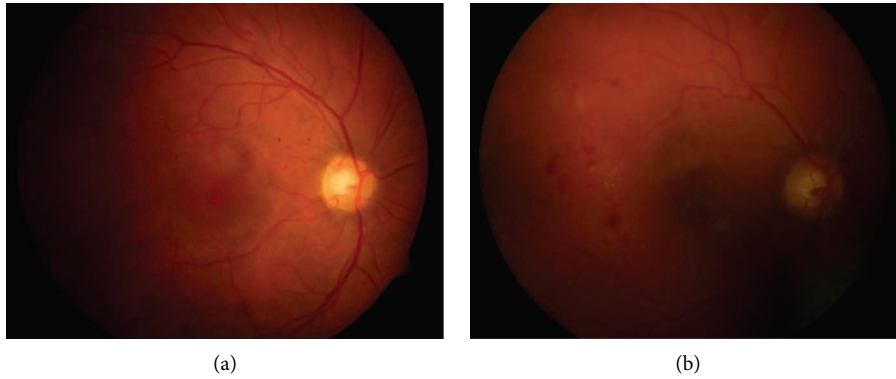


FIGURE 1: Degraded retinal fundus images: (a) uneven illumination; (b) the dark image because of low lighting condition.

This technique removes noise from the edges in the preprocessing stage. Then, histogram-based segmentation obtains regions for the classification. Synergy deep learning model classifies images into severity levels. The algorithm is benchmarked on the Messidor dataset which shows promising results.

3. Method

This section provides a detailed explanation of the detection scheme. Figure 2 shows the steps of the propounded HES detection scheme. First, the image is preprocessed to enhance the quality. Then the prospective hemorrhage candidates are estimated. The objects are segmented using smart window-based adaptive thresholding. Finally, the objects are classified into hemorrhage and nonhemorrhage classes using features.

3.1. Dataset Description. The algorithm is trained and tested on the DIARETDB1 dataset [16]. The dataset contains 89 fundus images, of which five images are normal and the rest have various DR pathological symptoms. These images are captured by the 50-degree field of view using a fundus camera under different illumination conditions.

3.2. Preprocessing. Few images of the DIARETDB1 dataset have good brightness levels and contrast, while the majority of them are dark with low contrast. The quality of fundus images is enhanced using contrast limited adaptive histogram equalization (CLAHE) [17], gradient-based adaptive brightness adjustment (GAGC) [18], and nonlinear unsharp masking [19]. CLAHE enhances contrast and reduces the effects of over-saturation by clipping intensity peaks. Our GAGC utilizes Sobel gradient information. Gamma correction [20] is applied using the adjusted threshold value of the Sobel operator. HES can be attached to blood vessels and can only be separated when their regions are clearly defined. Therefore, fuzzy logic-based image sharpening using a nonlinear filter is employed to sharpen the image. This method determines a fuzzy relationship between central and adjacent pixels in a 3×3 window. Sharpening filters work efficiently, but they

introduce the noise in the image. The nonlinear property sharpens images and produces less noise than linear filters. The result of the preprocessing stage is provided in Figure 3(b).

3.3. Seed Points Extraction. The detection process can be time-consuming if the entire image is considered for the search operation during detection. A good approach is to obtain prospective locations of objects to be detected and eliminate redundant information. This approach expedites detection with high accuracy. A similar technique manipulates the intensity profile of HES in our work. HES are dark objects surrounded by bright regions and share intensity characteristics with blood vessels and dark shades. This property suggests an inverted Gaussian matched filter [21], whose intensity values are low at the center and grow gradually beyond the center. This filter enhances HES and blood vessels due to high correlation and yields low response wherever applied to the rest of the image, and is depicted in Figure 4(a).

The redundant information is further reduced using the thresholding method. It depicted from the matched-filtered image that low and high responses are close to each other. Therefore, entropy thresholding is employed [22] that eliminates unrequired information efficiently. This thresholding method finds cross entropies between quadrants of gray level co-occurrence matrix (GLCM). The optimum threshold value from the gray range is selected successively, which minimizes the objective function. Figure 4(b) is a sample image of cross-entropy thresholding.

Elimination of blood vessels may also remove some of the HES attached to them. Therefore, consideration of objects that correspond to blood vessels is imperative. The morphological opening is applied to break the vasculature structure. This maneuver provides seed points for all types of HES, including those that are attached to the blood vessels. Conversely, it increases the number of seed points for subsequent segmentation and classification stages and can be depicted in Figure 4(c).

3.4. Image Calibration. The HES can be present at a jelly-like surface called the vitreous humour, and the black

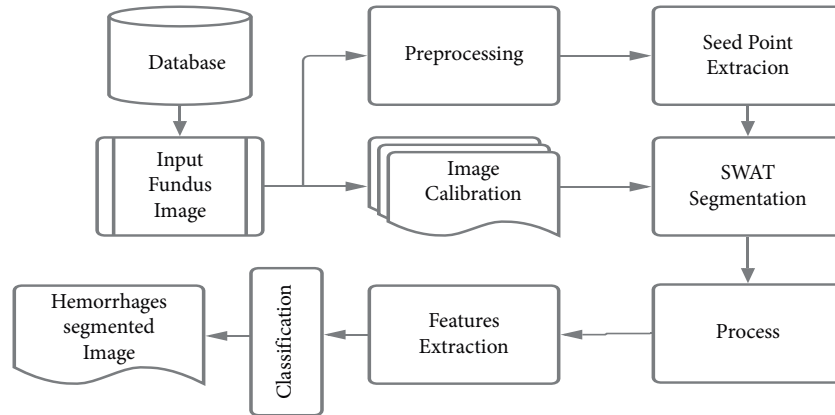


FIGURE 2: Illustration of the proposed detection technique.

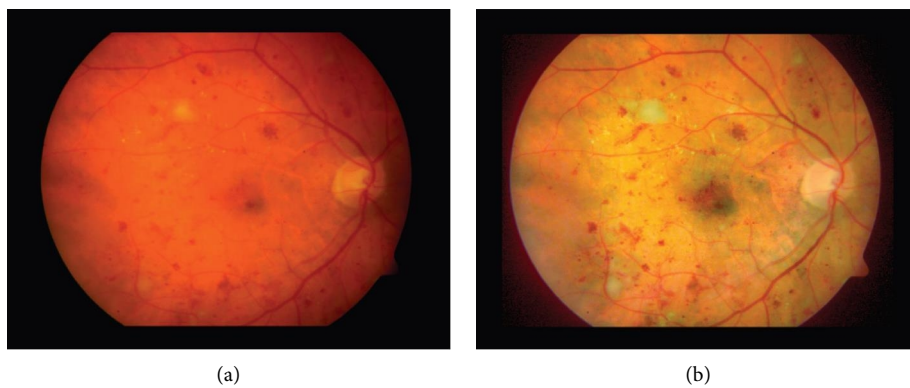


FIGURE 3: Preprocessing of the retinal fundus image: (a) input fundus image; (b) enhanced fundus image.

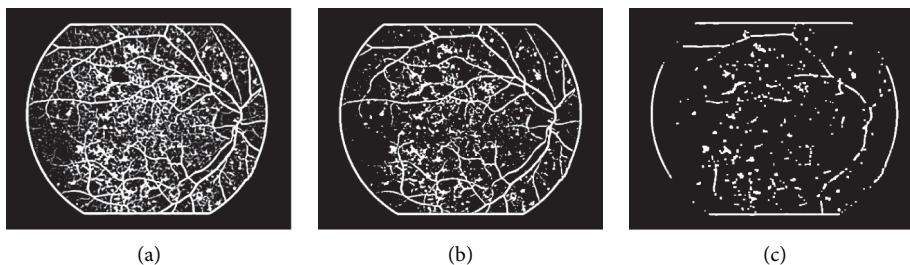


FIGURE 4: Seed points extraction: (a) response of matched filter; (b) cross-entropy thresholding; (c) morphological opened image.

background does not contribute to the detection phase. A black background is darker than HEs and misleads the detection process. Therefore, it impedes the automatic detection of those HEs that reside at the retinal border. The black background is illuminated for effective and automatic detection. First, a median filter is applied on a green channel to suppress intensity variation in the background and then binarized. The resultant image is called the retinal mask that highlights the retinal area. Later, an eroded mask is subtracted from the retinal mask to get the retinal boundary. Calibrated image is obtained by adding an enhanced green channel, complemented retinal mask, and retinal border. A sample of the calibrated image is depicted in Figure 5 and is used for segmenting HEs.

3.5. Smart Window-Based Adaptive Thresholding Segmentation. A segmentation method is sensitive to the dissimilarity of objects and their surroundings, and dissimilarity can be in terms of intensities or textures. There are two challenges for segmenting HEs. First is a segmentation of HEs blended with the black background and located at the retinal rim. This background has been illuminated using image calibration. The second is a segmentation of HEs that are attached to blood vessels. Blood vessels and HEs share intensity characteristics and they are known as dark smooth regions. Therefore, a novel smart window-based adaptive thresholding (SWAT) is proposed that isolates HEs from blood vessels. This method is adaptive and segments HEs encompassed by various bright regions. A

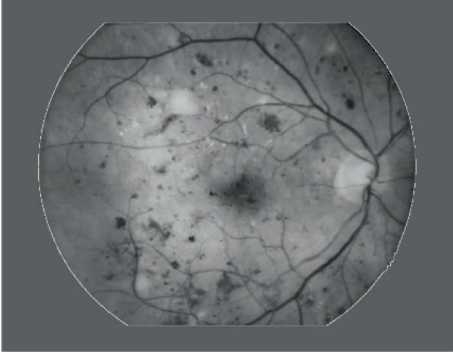


FIGURE 5: Image calibration for feature extraction.

search space is defined for automatic detection to constrain segmentation within image range. Complemented binary mask, obtained in the previous section, is expanded 80 pixels wide to provide sufficient space for HEs residing at the retinal border.

Segmentation using a threshold value is obtained by maximizing inter-region variance from image histogram [23]. This method determines the weighted variance $\sigma_B^2(j)$ between regions for a given threshold value j as

$$\sigma_B^2(j) = \sum_{z=1}^R \omega_z (\mu_z - \mu_T)^2, \quad (1)$$

where μ_T is mean value of an original image, ω_z is the total probability of individual region z , and μ_z is the mean value of individual regions in $R = \{2, 3, \dots, 19, 20\}$ after thresholding. An optimum threshold value is taken successively by maximizing the inter-region variance as

$$\sigma_B^2(\tau^*) = \max_{1 \leq j \leq L} \sigma_B^2(j), \quad (2)$$

$$j = \{0, 1, 2, \dots, L - 1\}.$$

The effectiveness η of an optimum threshold τ^* depends upon a selection of an appropriate number of regions from R . An appropriate number of regions provides maximum effectiveness. η is a ratio of weighted variance $\sigma_B^2(\tau^*)$ to the total variance σ_T^2 of an image that can be calculated as

$$\eta = \frac{\sigma_B^2(\tau^*)}{\sigma_T^2}. \quad (3)$$

SWAT initiates from seed points to segment retinal structures from the calibrated image, Figure 5. The search process starts from the bounding box of a seed point. The calibrated image is cropped using the vertices $V = \{v_1, v_2, v_3, \text{ and } v_4\}$ of a seed point. The cropped window $W_1(x, y)$ is thresholded iteratively until the appropriate number of regions from R is selected as

$$\vartheta = \begin{cases} R \longrightarrow R + 1, & \text{if } \eta < 0.8, \text{ and } R \leq 20, \\ \text{stop}, & \text{otherwise,} \end{cases} \quad (4)$$

where ϑ is a vector that contains $R - 1$ threshold values. (4) provides robustness in accordance with the regional diversity of HEs and foregrounds. In the case of bright

foreground, fewer iterations are required to approach the stopping criteria that yield few numbers of regions. For the dark foreground, more iterations are required to reach $\eta \geq 0.8$, which requires more numbers of regions to perform effective segmentation. HEs are dark objects surrounded by various bright regions. The window is thresholded as

$$W_2(x, y) = \begin{cases} 0, & \text{if } W_1(x, y) > \min(\vartheta), \\ 1, & \text{else,} \end{cases} \quad (5)$$

where $\min(\vartheta)$ is the minimum threshold value of the vector ϑ . There is a possibility that a window may have many HEs or dark objects after thresholding, so priority is given to the biggest ones because they are more dangerous for eyesight than the smaller HEs. Therefore, two large objects are kept and the rest are removed based on their area. This maneuver is applied such dark shades, often bigger sizes than HEs, cannot mislead the segmentation and actual HEs can be retained within the window. Furthermore, an object closer to the center of the window is more likely a HE than the other one. This probability criterion is proposed because seed points are extracted using the matched filter that models the intensity characteristic of HEs. Therefore, the object is eliminated using distance transform except one with minimum distance from the center of the window. The distance d_i of the i_{th} object from the center $W_2(x_c, y_c)$ of the window is calculated using

$$d_i = \min \sqrt{\{W_2(x_c) - I_i(x)\}^2 + \{W_2(y_c) - I_i(y)\}^2}, \quad (6)$$

where $I_i(x)$ and $I_i(y)$ denote the x and y spatial locations of i_{th} object, respectively, and $i = \{1, 2\}$. The sizes of the HEs are bigger than the size of the window because they initiated from a seed point. The window must be expanded to capture the complete HEs using

$$V = \begin{cases} v_1 \longrightarrow v_1 - 5, & \text{if } q_1 = 1 \text{ AND } v_1 \cap S, \\ v_2 \longrightarrow v_2 - 5, & \text{if } q_2 = 1 \text{ AND } v_2 \cap S, \\ v_3 \longrightarrow v_3 + 10, & \text{if } q_3 = 1 \text{ AND } v_3 \cap S, \\ v_4 \longrightarrow v_4 + 10, & \text{if } q_4 = 1 \text{ AND } v_4 \cap S. \end{cases} \quad (7)$$

$Q = [q_1, q_2, q_3, q_4]$ contains information of border pixels. Binary variables $q_1, q_2, q_3,$ and q_4 correspond to left, top, right, and bottom border pixels, respectively. If all these variables are 0, then no further iteration is required because it shows the complete segmentation of the object. If any variable in Q has a value of 1, it guarantees that the size of an object is bigger than the size of the window towards a particular direction. The window is expanded using equation (7) and the calibrated image is cropped by the updated vector V .

The search space assists in performing segmentation automatically. Some of the seed points are redundant and belong to blood vessels and dark shades. A window may go beyond image range when segmenting blood vessels or dark shades. The condition on vector S in (7) determines whether vertices of vector V lie within search space. Windows containing HEs and non-HEs objects are classified using features in the next section.

3.6. Features Extraction and Classification Stage. Support vector machine (SVM) is a statistical learning model used for classification by placing a hyperplane between positive and negative examples. Three sets of deep features were obtained from the hidden layers of VGG16 [24], ResNet50 [25], and AlexNet [26]. Four SVMs trained using conventional features and deep features to classify objects into HEs and non-HEs categories. Conventional features manipulate the visual appearance of HEs. For instance, HEs have sharp edges than macula, known as central vision. So, Laplacian-gradient features differentiate HEs from the macula. Blood vessels are line-shaped objects and HEs are comparatively circular objects. Therefore, connected component descriptors are useful to classify them. Color features help to distinguish dark shades from HEs. Opened or closed object's contour, number of corner points, and the spatial distance from the corners to the object's center are hand-crafted features. Hence, connected component [27], texture [28], color [29], and hand-crafted features are extracted to train SVM. While the VGG16, ResNet50, and AlexNet CNN models provide deep features for SVMs training.

4. Results, Comparison, and Discussion

The findings of the propounded detection scheme are reported in this section. Illustrations of performance metrics and the statistical comparison of various deep models are presented. The results can be pictorially be depicted in Figure 6.

4.1. Data Preparation and Evaluation Metrics. The DIA-RETDB1 dataset is employed to detect HEs and compare various feature extraction models. This dataset is divided into training and testing subsets. The training subset is further separated into training and validation subsets. Windows obtained by the SWAT segmentation were annotated using ground truths. Twenty images are used to benchmark the performances of classifiers. The classification results are compared using sensitivity (SE) and specificity (SP) [30].

$$\begin{aligned} SE &= \frac{TP}{TP + FN}, \\ SP &= \frac{TN}{TN + FP}, \end{aligned} \quad (8)$$

where true-positive (TP) and true-negative (TN) are the truly predicted measurements by the classifier. TP is the rate of truly classified hemorrhage, while TN is the correct prediction rate of the negative class. Conversely, false-positive (FP) and false-negative (FN) are the measurements of the false predictions of the classifier. FP wrongly indicates that an object belongs to a hemorrhage, but actually, it does not. FN shows that hemorrhage is not present while the window contains a hemorrhage.

4.2. Results. Results represent that the false-negative (FN) rate of conventional features is higher than deep features. It states that conventional methods cannot identify some of the

HEs. Conversely, deep models are more capable of HEs identification. The classification results of deep and conventional models are provided in Table 1, while visually they are depicted in Figure 6.

5. Discussion

The SE and SP observe the performances of the classification models. All the features' extraction models show promising results and are applicable for the detection of DR. However, the FN rate of the SVM trained by the conventional features is the highest, resulting in a minimum SE than the other methods. The reason is obvious, the small arteries of the blood vessels. Blood vessels are classified using connected components, but the small arteries also share these properties. Therefore, their similar appearance concerning the intensity and connected component characteristics misleads the classifier because they are labeled as a negative class. The conventional features of HEs and the arteries overlap, therefore, the highest misdetection rate. It marks 88.98% SE and 97.67% SP. VGG16 CNN model has better SE that can detect more HEs than the conventional method. It marks the highest SE but low SP, which states that its false-positive (FP) rate is the highest compared to the other methods. The worst performance of VGG16 might be its high convergence rate toward the solution. The increased convergence rate has the drawback of oscillatory behavior around the optimum solution. Therefore, the network cannot converge to the optimum point for useful features. The SE and SP of this deep model are 95.88% and 94.87%, respectively. The performances of the ResNet50 and AlexNet are mediocre. They provide better SE than the convention method and better SP than the other two models. Evaluation metrics output 92.24% SE and 97.81% SP by the ResNet50. While the AlexNet also has a similar behavior for SE, the SP is considerably higher than ResNet50. Effectively, it yields the highest SP among all the methods. The statistics of AlexNet for SE and SP are 92.21% and 98.24%, respectively. Furthermore, the assessment of ResNet50 and AlexNet architectures reveals that the ResNet50 is unnecessarily deep. ResNet50 contains fifty layers, while AlexNet is eight layers deep. Therefore, AlexNet can be a good choice for HEs classification because it marks competitive results.

The assessment of the deep feature extraction models reveals that the arrangement of layers in deep models is crucial for a particular application. The increasing number of deep layers may not yield good results, instead, it increases the training time. For instance, AlexNet is shallower than VGG16 and ResNet50 but provides the highest SP of 98.24 and competitive SE of 92.21. While VGG16 is deeper than AlexNet and shallower than ResNet50, it yields the highest SE of 95.88 and lowest SP of 94.87. ResNet50 is the deepest and marks mediocre results.

Furthermore, It is observed from the architectures of the predefined networks that the filter sizes of the first convolution layers of VGG16, ResNet50, and AlexNet are 3×3 , 7×7 , and 11×11 , respectively. The small filter's size is appropriate for HEs classification because of its homogeneous property. HEs are regarded as dark smooth regions. VGG16 has the smallest filter size and identifies more HEs

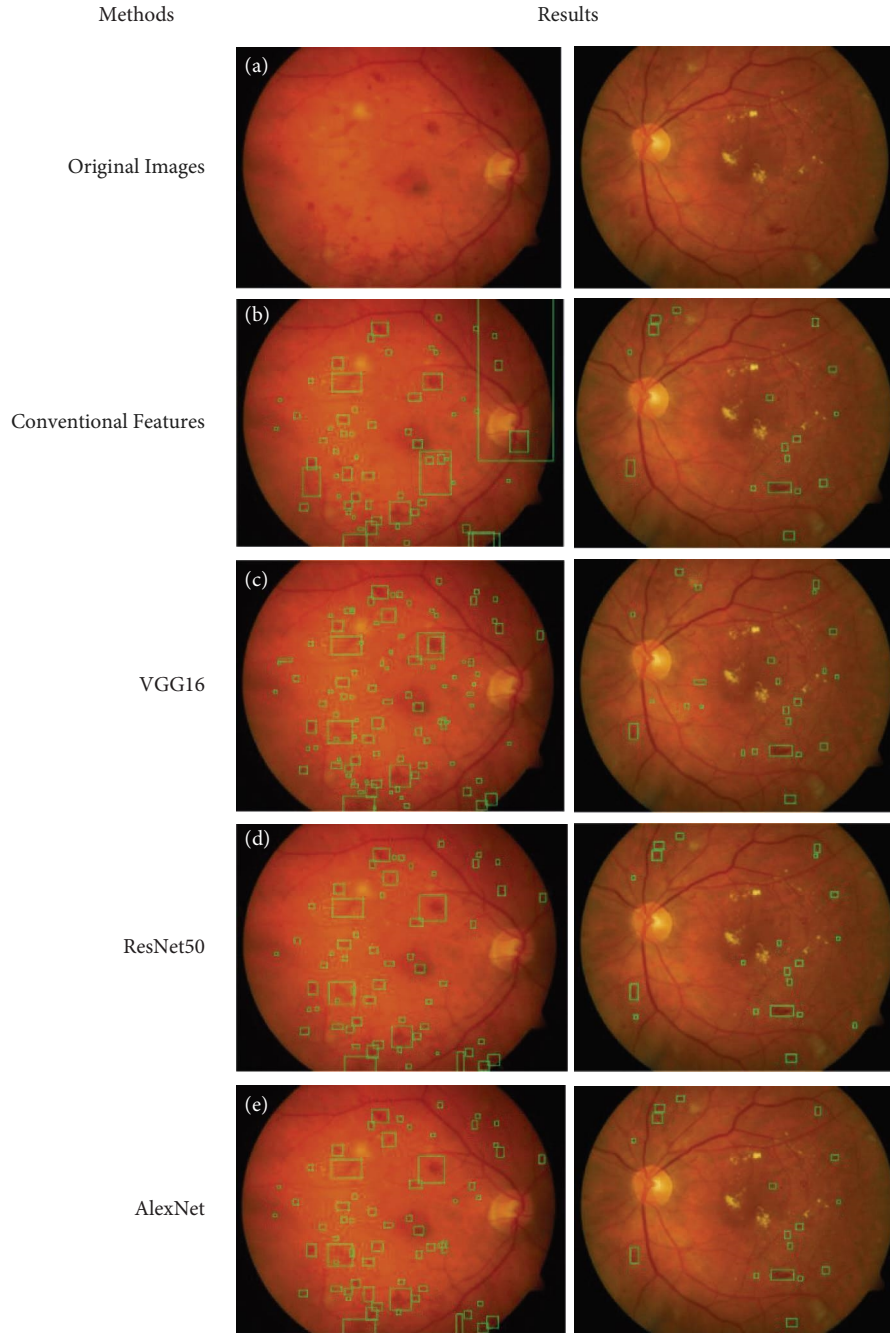


FIGURE 6: Classification results: (a) original images, (b) classification using conventional features, (c) classification using VGG16, (d) classification using ResNet50, and (e) classification using AlexNet.

TABLE 1: Comparison of the classification stage using various models on the DIARETDB1 dataset.

Methods	SE (%)	SP (%)
SVM	88.98	97.67
VGG16	95.88	94.87
ResNet50	92.24	97.81
AlexNet	92.21	98.24

because its FN rate is the lowest among other deep models. Conversely, dark shades and small blood vessels mislead VGG16, resulting in the highest FP rate.

The analysis of the classification results recommends the deep model for the HEs classification. The reason could be that some conventional features may not be effective and mislead the classifier. The deep networks provide relevant features because they learn incrementally from the data. Therefore, no feature can mislead the classification stage. On the contrary, the CNN models take more time to learn from the windows and recognize them. They often need large numbers of training examples, depending upon the complexity of the data, for better performance. The conventional method needs comparatively less time and training examples to obtain the statistical features.

6. Conclusion

This research presents an automatic detection technique to compare various deep-learning-based models with the conventional features extraction approach. The method first enhances the quality of the fundus images for a better appearance of pathological symptoms in the preprocessing stage. Then, the locations of the hemorrhages are estimated using seed points extraction that expedites the detection process. Deep and conventional features classify the objects into hemorrhages and nonhemorrhages. The research concept emerged from the problem highlighted by the research community that two types of hemorrhages are challenging to detect. First, the hemorrhages that are associated with the blood vessels. Second, the hemorrhages that are located at the retinal border. Our detection scheme is suitable for all types, including those hemorrhages that reside in the vitreous humour. This study also prescribes that the deep features can better classify hemorrhages than the conventional methods; hence they are more efficient and suitable for the hemorrhages classification.

The assessment of performance metrics of deep modalities reveals that a shallow network produces competitive results compared to deep models. An intense deep network may not yield significant improvements but increases training time. In this study, AlexNet shows promising results despite the shallowest network. Therefore, a suitable network with its appropriate parameters is critical.

The research work's intuition is to present a fully-automated scheme for reducing the misdetection rate of hemorrhages by ophthalmologists interpreting fundus photographs. The method identifies hemorrhages in an interactive way that is easy to interpret for diabetic retinopathy diagnosis. Furthermore, the locations of hemorrhages are highlighted, which might help the clinicians conclude the severity levels of the disease.

Data Availability

Publicly available datasets were analyzed in this study. This data can be found as follows: <https://www.it.lut.fi/project/imageret/diaretdb1/index.html>.

Conflicts of Interest

All authors in this study declare no conflicts of interest.

Authors' Contributions

TA performed the investigation, methodology, software implementation, writing the original draft, conceptualization, and formal analysis. CC was involved in funding acquisition, project administration, supervision, review, and editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported by Thammasat University Research Fund and SIIT Research Fund.

References

- [1] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review," *Clinical and Experimental Ophthalmology*, vol. 44, no. 4, pp. 260–277, 2016.
- [2] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.
- [3] C.-F. Chou, C. E. Sherrod, X. Zhang et al., "Barriers to eye care among people aged 40 Years and older with diagnosed diabetes, 2006-2010," *Diabetes Care*, vol. 37, no. 1, pp. 180–188, 2014.
- [4] B. Corcóstegui, S. Durán, M. O. González-Albarrán et al., "Update on diagnosis and treatment of diabetic retinopathy: a consensus guideline of the working group of ocular health (Spanish society of diabetes and Spanish vitreous and retina society)," *Journal of ophthalmology*, vol. 2017, 2017.
- [5] M. W. Stewart, "Treatment of diabetic retinopathy: recent advances and unresolved challenges," *World Journal of Diabetes*, vol. 7, no. 16, p. 333, 2016.
- [6] I. N. Figueiredo, S. Kumar, C. M. Oliveira, J. D. Ramos, and B. Engquist, "Automated lesion detectors in retinal fundus images," *Computers in Biology and Medicine*, vol. 66, pp. 47–65, 2015.
- [7] L. Tang, M. Niemeijer, J. M. Reinhardt, M. K. Garvin, and M. D. Abramoff, "Splat feature classification with application to retinal hemorrhage detection in fundus images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 364–375, 2012.
- [8] S. B. Júnior and D. Welfer, "Automatic detection of microaneurysms and hemorrhages in color eye fundus images," *International Journal of Computer Science and Information Technology*, vol. 5, no. 5, pp. 21–37, 2013.
- [9] L. Zhou, P. Li, Q. Yu, Y. Qiao, and J. Yang, "Automatic hemorrhage detection in color fundus images based on gradual removal of vascular branches," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 399–403, IEEE, Phoenix, AZ, USA, September 2016.
- [10] S. Karkuzhali and D. Manimegalai, "Distinguishing proof of diabetic retinopathy detection by hybrid approaches in two dimensional retinal fundus images," *Journal of Medical Systems*, vol. 43, no. 6, pp. 1–12, 2019.
- [11] J. H. Tan, H. Fujita, S. Sivaprasad et al., "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Information Sciences*, vol. 420, pp. 66–76, 2017.
- [12] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Investigative Ophthalmology & Visual Science*, vol. 59, no. 1, pp. 590–596, 2018.
- [13] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," 2018, <https://arxiv.org/abs/1812.10595>.
- [14] P. Pal, S. Kundu, and A. K. Dhara, "Detection of red lesions in retinal fundus images using YOLO V3," *Curr. Indian Eye Res. J. Ophthalmic Res. Group*, vol. 7, p. 49, 2020.
- [15] K. Shankar, A. R. W. Sait, D. Gupta, S. K. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using

- synergic deep learning model,” *Pattern Recognition Letters*, vol. 133, pp. 210–216, 2020.
- [16] J. Du, B. Zou, P. Ouyang, and R. Zhao, “Retinal microaneurysm detection based on transformation splicing and multi-context ensemble learning,” *Biomedical Signal Processing and Control*, vol. 74, Article ID 103536, 2022.
- [17] M. J. Alwazzan, M. A. Ismael, and A. N. Ahmed, “A hybrid algorithm to enhance colour retinal fundus images using a wiener filter and CLAHE,” *Journal of Digital Imaging*, vol. 34, pp. 1–10, 2021.
- [18] T. Aziz, A. E. Ilesanmi, and C. Charoenlarnnopparut, “Efficient and accurate hemorrhages detection in retinal fundus images using smart window features,” *Applied Sciences*, vol. 11, no. 14, p. 6391, 2021.
- [19] F. Russo, “Design of fuzzy relation-based image sharpeners,” in *New Advances in Intelligent Signal Processing*, pp. 115–131, Springer, Berlin, Germany, 2011.
- [20] M. Veluchamy and B. Subramani, “Image contrast and color enhancement using adaptive gamma correction and histogram equalization,” *Optik*, vol. 183, pp. 329–337, 2019.
- [21] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, “Detection of blood vessels in retinal images using two-dimensional matched filters,” *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 263–269, 1989.
- [22] F. Nie, C. Gao, Y. Guo, and M. Gan, “Two-dimensional minimum local cross-entropy thresholding based on co-occurrence matrix,” *Computers & Electrical Engineering*, vol. 37, no. 5, pp. 757–767, 2011.
- [23] C. Sha, J. Hou, and H. Cui, “A robust 2D Otsu’s thresholding method in image segmentation,” *Journal of Visual Communication and Image Representation*, vol. 41, pp. 339–351, 2016.
- [24] D. M. S. Arsa and A. A. N. H. Susila, “VGG16 in batik classification based on random forest,” vol. 1, pp. 295–299, in *Proceedings of the 2019 International Conference on Information Management and Technology (ICIMTech)*, vol. 1, IEEE, Jakarta/Bali, Indonesia, August 2019.
- [25] D. Theckedath and R. R. Sedamkar, “Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks,” *SN Computer Science*, vol. 1, no. 2, pp. 1–7, 2020.
- [26] S. Lu, Z. Lu, and Y.-D. Zhang, “Pathological brain detection based on AlexNet and transfer learning,” *Journal of computational science*, vol. 30, pp. 41–47, 2019.
- [27] H. Chatbri and K. Kameyama, “Document image dataset indexing and compression using connected components clustering,” in *Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 267–270, Tokyo, Japan, May 2015.
- [28] Y. Hu and Y. Zheng, “A GLCM embedded CNN strategy for computer-aided diagnosis in intracerebral hemorrhage,” 2019, <https://arxiv.org/abs/1906.02040>.
- [29] S. Murali and V. K. Govindan, “Shadow detection and removal from a single image using LAB color space,” *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 95–103, 2013.
- [30] S. Wan, Y. Liang, and Y. Zhang, “Deep convolutional neural networks for diabetic retinopathy detection by image classification,” *Computers & Electrical Engineering*, vol. 72, pp. 274–282, 2018.

Research Article

Machine Learning-Based Prediction Model of Preterm Birth Using Electronic Health Record

Qi Sun ^{1,2}, Xiaoxuan Zou,³ Yousheng Yan,⁴ Hongguang Zhang,¹ Shuo Wang,³ Yongmei Gao,³ Haiyan Liu,³ Shuyu Liu,³ Jianbo Lu ^{1,2}, Ying Yang ¹, and Xu Ma ^{1,2}

¹National Human Genetics Resource Center, National Research Institute for Family Planning, Beijing 100081, China

²Graduate School of Peking Union Medical College, Beijing 100730, China

³Haidian Maternal and Child Health Hospital, Beijing 100080, China

⁴Beijing Obstetrics and Gynecology Hospital, Beijing 100010, China

Correspondence should be addressed to Jianbo Lu; jblu@lsec.cc.ac.cn, Ying Yang; angela-yy65@hotmail.com, and Xu Ma; nfpcc_ma@163.com

Received 11 December 2021; Revised 26 February 2022; Accepted 14 March 2022; Published 13 April 2022

Academic Editor: Weihua Yang

Copyright © 2022 Qi Sun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Objective. Preterm birth (PTB) was one of the leading causes of neonatal death. Predicting PTB in the first trimester and second trimester will help improve pregnancy outcomes. The aim of this study is to propose a prediction model based on machine learning algorithms for PTB. **Method.** Data for this study were reviewed from 2008 to 2018, and all the participants included were selected from a hospital in China. Six algorithms, including Naive Bayesian (NBM), support vector machine (SVM), random forest tree (RF), artificial neural networks (ANN), K-means, and logistic regression, were used to predict PTB. The receiver operating characteristic curve (ROC), accuracy, sensitivity, and specificity were used to assess the performance of the model. **Results.** A total of 9550 pregnant women were included in the study, of which 4775 women had PTB. A total of 4775 people were randomly selected as controls. Based on 27 weeks of gestation, the area under the curve (AUC) and the accuracy of the RF model were the highest compared with other algorithms (accuracy: 0.816; AUC = 0.885, 95% confidence interval (CI): 0.873–0.897). Meanwhile, there was positive association between the accuracy and AUC of the RF model and gestational age. Age, magnesium, fundal height, serum inorganic phosphorus, mean platelet volume, waist size, total cholesterol, triglycerides, globulins, and total bilirubin were the main influence factors of PTB. **Conclusion.** The results indicated that the prediction model based on the RF algorithm had a potential value to predict preterm birth in the early stage of pregnancy. The important analysis of the RF model suggested that intervention for main factors of PTB in the early stages of pregnancy would reduce the risk of PTB.

1. Introduction

Preterm birth (PTB) is defined as births before 37 completed weeks of gestation [1]. The PTB studied in this study was for 28–37 weeks of gestational age. Based on gestational age at delivery, PTB can be subdivided into very early preterm (<28 weeks), early preterm (28–31 weeks), moderate preterm (31–33 weeks), and late preterm (33–37 weeks) [2]. The global estimated prevalence of PTB was 11.1% (95% confidence interval [CI]: 9.1%–13.4%) [3]. The majority of PTB occurred in low- and middle-income countries [2], and the incidence of PTB in China was 6.9% in 2014 [4]. Although

the incidence of premature birth was relatively low in China, PTB had a considerable impact on the health of pregnant women and children. Evidence shows that PTB was the most common cause of neonatal death and the second most frequent cause of death in children aged <5 years [5]. Further studies found that gestational age at delivery was inversely associated with the risk of neonatal morbidity and mortality [6], and about 35.00% of deaths among newborns were caused by complications of PTB [7]. Preterm neonates who survived were vulnerable to diseases, including pulmonary hypertension [8], retinopathy [9], visual and hearing impairments [10], and mental health problem [11]. Moreover,

PTB not only caused death and diseases in the newborn, but also caused anxiety and depression in postpartum women [12]. Previous study showed that early screening of preterm birth pregnant women could reduce the incidence of preterm birth [13]. Therefore, a prediction model was needed to predict PTB.

Currently, numerous studies have attempted to predict preterm birth in pregnant women. Several studies supported that sonographic measurement of cervical length (CL) could be used for the prediction of PTB in the first trimester of pregnancy [14, 15], but other studies did not demonstrate the capability of CL in the screening of PTB [16, 17]. Fetal fibronectin had extensively used to predict PTB, but the sensitivity and positive predictive value of fetal fibronectin were low [18, 19]. In recent years, machine learning algorithms have been widely used in medicine with a better performance [20]. Compared with the logistic regression algorithm, the advantages of the machine learning were the ability to process higher-dimensional data and self-learn capacity [21]. Studies have shown that the use of machine learning algorithms improved the predictive accuracy of the prediction model for PTB [22, 23]. There are also some prediction models based on machine learning algorithms that have poor prediction accuracy. Weber et al. established a machine learning prediction model for preterm birth using demographic, maternal, and residency characteristics, but the predictive performance of the model was poor [24], which may be caused by inaccurate geographic information.

Inconsistent predictive power of machine learning in preterm birth. In this study, we try to use a new method to preprocess predictors. At the same time, we compared the predictive power of 6 machine learning algorithms in PTB.

2. Materials and Methods

2.1. Participants. Data for this study were reviewed from 2008 to 2018. All the participants included in this study were collected from Haidian Maternal & Child Health Hospital. The inclusion criteria of the PTB group were as follows: (1) signed informed consent; (2) gestational age between 28 and 37 weeks; and (3) maternal age older than 18 years. The exclusion criteria of the PTB group are as follows: (1) missing maternal age; (2) missing gestational age; and (3) chronic diseases such as diabetes, hypertension, and heart disease. Controls were selected from hospitals in the same period in a 1:1 ratio. The inclusion criteria of controls were as follows: (1) signed informed consent; (2) gestational age ≥ 37 weeks; and (3) maternal age ≥ 18 years. Exclusion criteria are as follows: (1) missing maternal age; (2) missing gestational age; (3) and chronic diseases such as diabetes, hypertension, and heart disease. The flowchart of the study is shown in Figure 1.

2.2. Feature Processing. Demographic factors (i.e., age), physical examination, blood test (red blood cells (RBC), white blood cell count (WBC), and plateletcrit (PCT)), urine test strip (urine pH, urine WBC, and glycosuria), and gynecological examination (bacterial vaginosis (BV), cleaning degree of vagina (CDV), and vaginal yeast infection (VYI))

were collected in our study. All participants had at least five antenatal check-ups before 27 weeks of gestation. For avoiding the overfitting of the model, variables that were measured multiple times were represented using the mean and mode, depending on the type of variable. With the increase in the gestational age, variables were more influence on the outcome. Therefore, we gave more weight to the later data. The equation is defined as

$$\text{var}_{\text{mean}}^{20} = \text{average}(\text{var}_{\text{week}_1}, \text{var}_{\text{week}_2}, \dots, \text{var}_{\text{week}_{20}}), \quad (1)$$

$$\text{var}_{\text{mean}}^i = \text{average}(\text{var}_{\text{mean}}^{i-2}, \text{var}_{\text{week}_{i-1}}, \text{var}_{\text{week}_i}), \quad (2)$$

$i = 22, 24, 26, 27$ weeks of gestation.

As shown in Figure 2, the variable processing process at each time point is determined by the values of the previous time point and the current time point. The dataset was divided into five datasets (20 weeks, 22 weeks, 24 weeks, 26 weeks, and 27 weeks of gestation dataset), according to the time of prenatal examination.

2.3. Machine Learning Algorithms. In this study, six algorithms, including Naive Bayesian (NBM), support vector machine (SVM), random forest tree (RF), artificial neural networks (ANN), K-means, and logistic regression, were used to predict PTB (Figure 3).

2.4. Outcome Measure. In this study, 4 metrics were used to measure the predictive performance of the model: accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. The accuracy is the proportion of correct predictions among the total number of cases examined (1). Sensitivity refers to the test's ability to correctly detect true positive (2). Specificity relates to the test's ability to correctly detect true negative (4). AUC is a comprehensive measure of the sensitivity and specificity of the model:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (3)$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4)$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (5)$$

TP = true positive; FP = false positive; TN = true negative; FN = false negative.

2.5. Statistical Analysis. The Kolmogorov-Smirnov test was used to test the normality of continuous variable. If the variable satisfies normal distribution, the mean \pm standard deviation was used to describe the continuous variable. Categorical variables were shown as numbers and percentages. Because our data were collected from electronic medical records, there were missing values in the dataset. Therefore, we excluded cases and variables that were missing

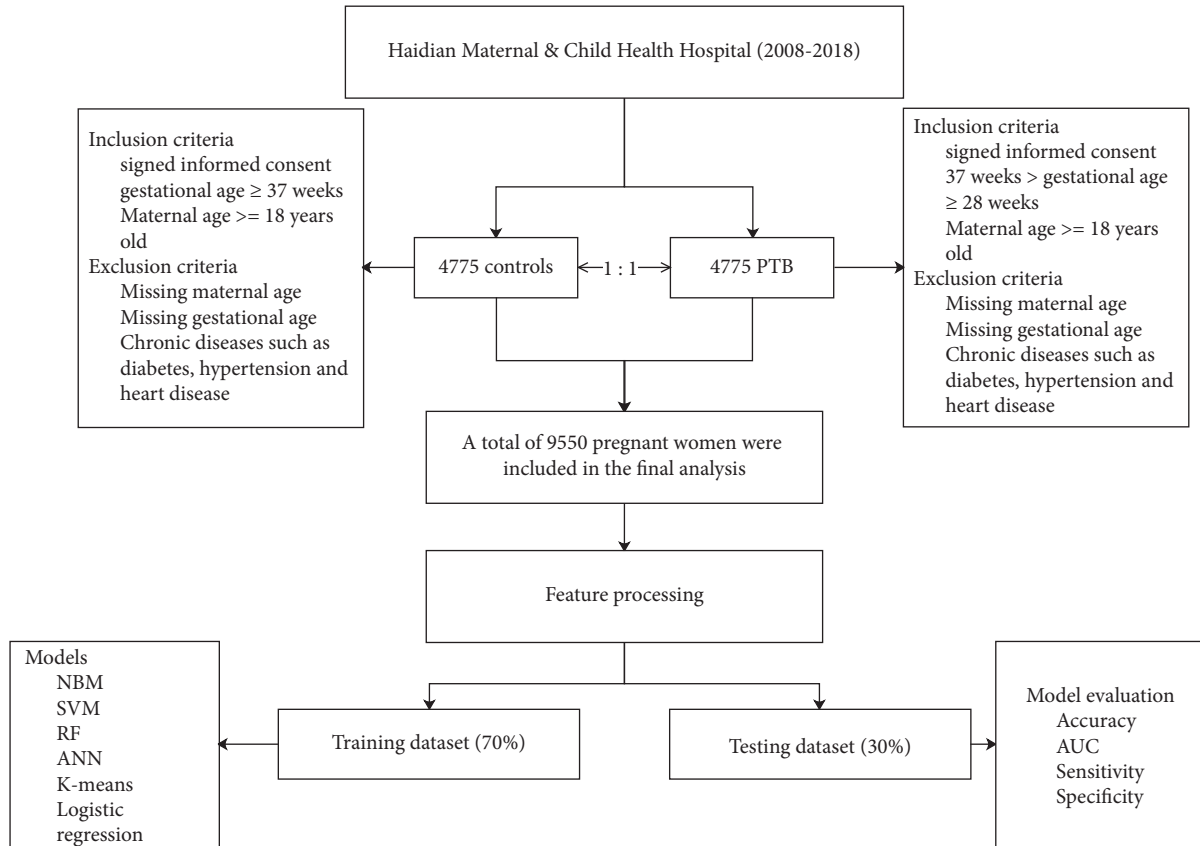


FIGURE 1: Workflow of this study.

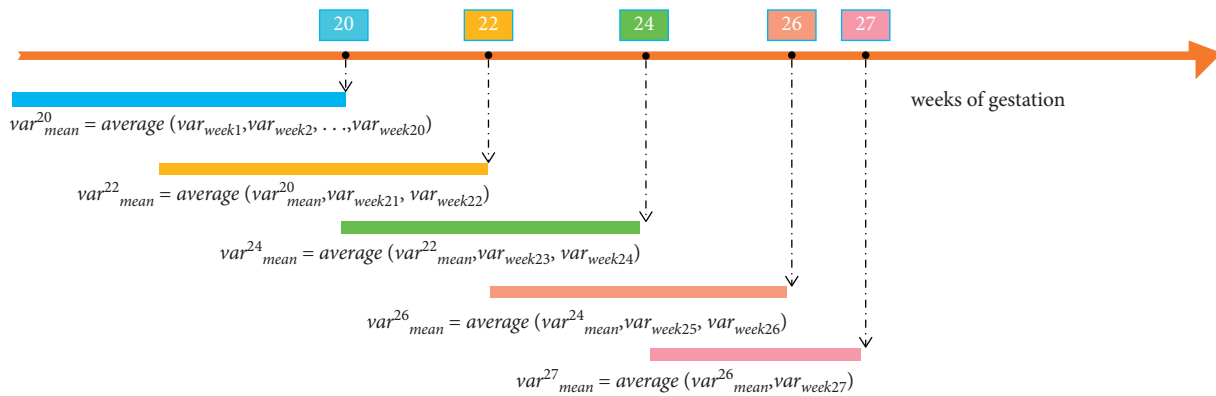


FIGURE 2: Preprocessing of variables. var_{week_i} is the measurement result of the variable in week i ; var_{mean}^{20} is the composite indicator representing the variable 20 weeks ago.

more than 10%. For categorical variables, mode was used to fill, and for continuous variables, mean was used to fill. Comparison between the outcome groups was made by the chi-square test or Fisher’s exact test for categorical variables and by the t -test or Wilcoxon test for continuous variables.

The dataset was randomly divided into a training set (70%) and a test set (30%). The training set was used to train the model, and the test set was used to evaluate the model. Four indicators, the area under the curve (AUC), accuracy, sensitivity, and specificity, were used to measure the

performance of the model. The importance of a variable was assessed by the decreased accuracy of the model after removing the variable. The higher the decreased accuracy of the model, the more important the variable. All statistical analyses were performed in R software (version 3.5.1) using the “e1071” (Naive Bayesian algorithm and support vector machine), “randomForest” (random forest tree), and “kkn” (K-means) packages. For all analyses, if the two-tailed P value < 0.05 , the result was considered statistically significant.

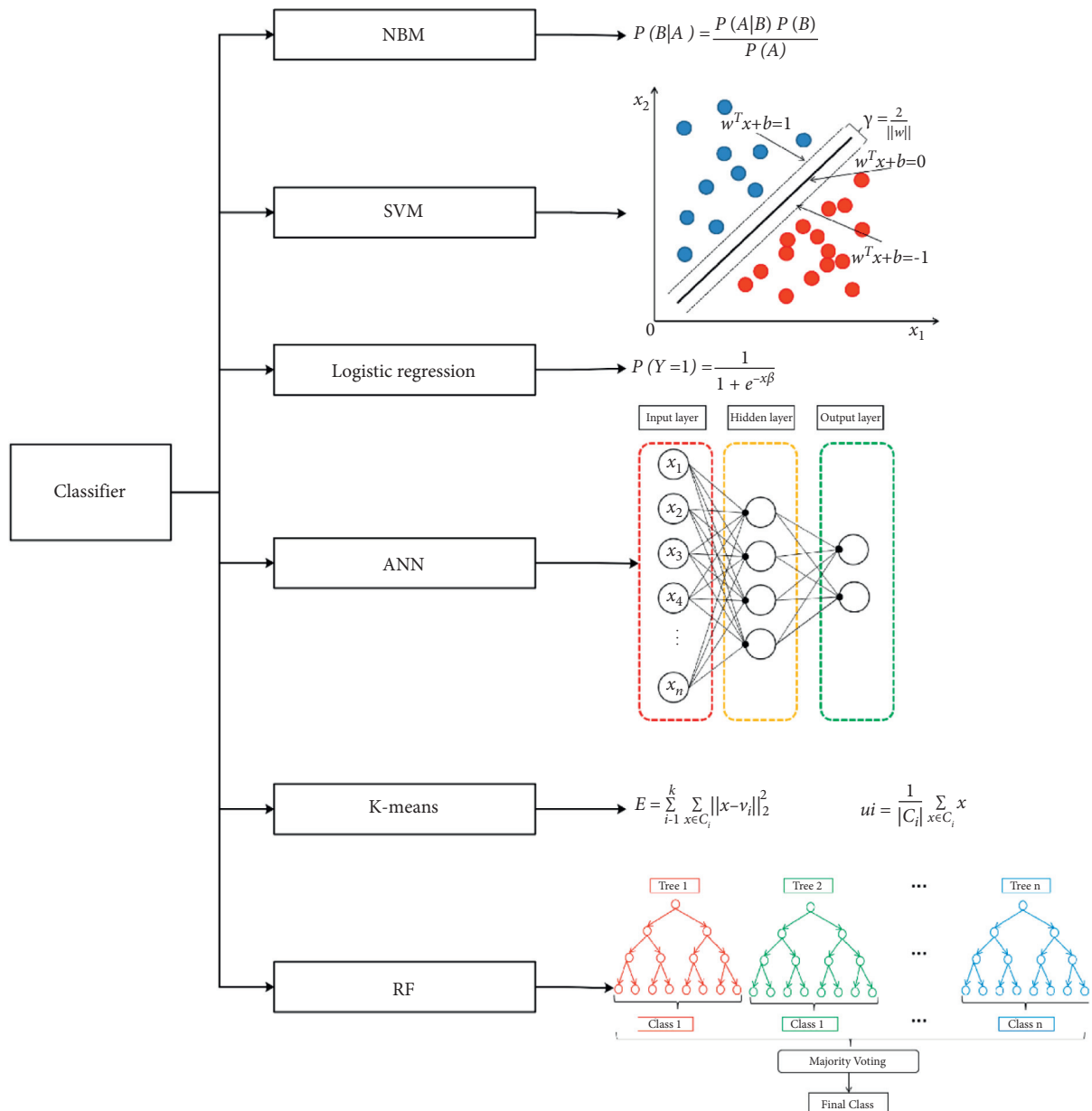


FIGURE 3: Classifiers used in this study. (1) Naive Bayesian (NBM): Naive Bayes calculates the posterior probability $P(B|A)$ from $P(A)$, $P(B)$ and $P(B|A)$; $P(B|A)$ is the posterior probability of class B and $P(A)$ is the prior probability of predictor A and $P(B)$ is the prior probability of class, and $P(B|A)$ is the probability of the predictor for the particular class. (2) Support Vector Machine (SVM); SVM outputs a hyperplane ($w^T x + b = 0$) that best separates the classes and has the largest separation of geometrical separations. (3) Logistic regression: The principle of logistic regression is to use a logistic function to map the results of linear regression between 0 and 1; X is the input features, and β is the weight of the features. $P(Y = 1)$ is the predicted probability of class 1. (4) Artificial Neural Networks (ANN): An artificial neural network consists of an input layer, a hidden layer, and an output layer, and its core component is an artificial neuron. Each neuron is summed by several other neurons multiplied by weights; x_i is the input features. (5) K-means: The K-Means algorithm minimizes the squared error for cluster C_i ; x is the unclassified sample, and C_i is the clusters, and u_i is the mean vector of clusters C_i . (6) Random Forest Tree (RF): Random forest is an algorithm that integrates multiple decision trees through the Bagging idea of ensemble learning. The principle of random forest bagging is to vote the classification results of several weak classifiers to form a strong classifier.

TABLE 1: Characteristics of mother and newborn between PTB and control group.

Variables		Control (4775)	Case (4775)	t/chi	P
Age, years		30.72 ± 4.00	29.94 ± 5.39	8.00	<0.001
Gestation, days		274.66 ± 7.15	251.19 ± 11.51	119.70	<0.001
Gravidity	1	3437 (0.72)	3644 (0.76)	25.08	<0.001
	2–3	1240 (0.26)	1063 (0.22)	25.08	<0.001
	>3	98 (0.02)	68 (0.01)	25.08	<0.001
Parity	1	4006 (0.84)	4176 (0.87)	24.37	<0.001
	>2	769 (0.16)	599 (0.13)	24.37	<0.001
Multiple birth	No	4763 (1.00)	4284 (0.90)	479.50	<0.001
	Yes	12 (0.00)	491 (0.10)	479.50	<0.001
Birth gender	Male	2464 (0.52)	2659 (0.56)	15.85	<0.001
	Female	2311 (0.48)	2116 (0.44)	15.85	<0.001
Birth weight, g		3410.68 ± 402.05	2691.13 ± 544.90	73.43	<0.001
Birth height, cm		50.38 ± 1.25	47.85 ± 2.82	56.81	<0.001
Apgar scores (1 min)	9.95 ± 0.71	9.70 ± 1.37	11.19	<0.001	
Apgar scores (5 min)	10.00 ± 0.66	9.82 ± 1.20	8.97	<0.001	
Apgar scores (10 min)	9.95 ± 0.54	9.77 ± 1.32	8.68	<0.001	

PTB: preterm birth.

2.6. *Statement of Ethics.* Ethics approval of this research was approved by the Institutional Research Review Board at National Research Institute for Family Planning and performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

3. Results

3.1. *Characteristics of Pregnant Women and Newborns.* A total 9550 of pregnant women (PTB: 4775, control: 4775) were included in our study. The mean ages of the PTB group were lower than those of the control group (PTB: 29.94 ± 5.39), control: 30.72 ± 4.00, $P < 0.001$). The gestation of pregnant women was 251.19 ± 11.51 days in the case group and 274.66 ± 7.15 days in the control group ($P < 0.001$). The gravidity and parity of pregnant women in the PTB group were lower than those in the control group (all $P < 0.001$). The weight and height of newborns in the control group were higher than those in the PTB group (all $P < 0.001$). The Apgar scores (1, 5, and 10 minutes) of newborns in the control group were higher than those in the case group (all $P < 0.001$). The characteristics of pregnant women and newborns were summarized in Table 1.

3.2. *Prenatal Testing of Pregnant Women before 27 Weeks of Gestation.* In the biochemical analysis, albumin, aspartate transaminase (AST), total serum iron (TSI), magnesium (Mg), and triglycerides (TG) levels were higher in the PTB group than those in the control group (all $P < 0.05$). Meanwhile, the plasma glucose (fasting) is lower in the PTB group than that in the control group (all $P < 0.05$). Total biliary acid (TBA) and urea levels were higher in the PTB group than those in the control group (all $P < 0.05$). Platelet, intermediate cell, lymphocyte (LY), monocytes (MO), neutrophil granulocytes (NE), red blood cell distribution width-SD (RDW-SD), and WBC levels were higher in the PTB group than those in the control group.

Mean cell hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and platelet distribution width (PDW) were lower in the PTB group than those in the control group. Waist size, fundal height, SBP, and DBP were higher in the PTB group than those in the control group. Fetal heart rate (FHR) in the PTB group was slower than that in the control group. Urine PH was higher in the PTB group than those in the control group. Pregnant women with blood type B were found to be more common in the case group than in the control group (Table 2). The results of prenatal testing at several other time points (20, 22, 24, and 26 weeks of gestation) were described in Supplementary Tables S1–S4.

3.3. *Performance of Prediction Models.* Six algorithms (NBM, SVM, RF, ANN, K-means, and logistic regression) were used to build the model based on five datasets (20, 22, 24, 26, and 27 weeks of gestation).

Table 3 depicts the performance of the six types of models. The results showed that the AUC and the accuracy of the RF model based on 27 weeks of gestation were the highest compared with other algorithms (accuracy: 0.816; AUC = 0.885, 95% (confidence interval) CI: 0.873–0.897). The sensitivity and specificity of the RF model based on 27 weeks of gestation were 0.751 and 0.882. Meanwhile, there was positive association between the accuracy and AUC of the RF model and gestational age (Figure 4). The sensitivity of the NBM model based on 24 weeks of gestation was 0.837, but the specificity was only 0.515. The specificity of the NBM model based on 26 weeks of gestation was 0.946, but the sensitivity was only 0.328. The receiver operating characteristic (ROC) curve of the models is shown in Figure 5.

The importance analysis of the RF model found that the top 10 most important variables were age, magnesium, fundal height, serum inorganic phosphorus, mean platelet volume, waist size, total cholesterol (TC), TG, globulins, and

TABLE 2: Prenatal testing of pregnant women before 27 weeks of gestation between PTB and control group.

Variables		Control (4775)	Case (4775)	t/chi	P	
Physical examination	Waist size, cm	82.68 ± 14.19	83.30 ± 13.74	-2.17	0.030	
	Fundal height, cm	20.57 ± 3.62	20.90 ± 3.84	-4.33	<0.001	
	SBP, mmHg	112.09 ± 10.26	113.34 ± 10.44	-5.86	<0.001	
	DBP, mmHg	69.47 ± 7.72	70.71 ± 16.09	-4.82	<0.001	
	FHR, times/min	145.50 ± 3.05	146.46 ± 17.27	-3.78	<0.001	
	Weight, kg	63.16 ± 9.15	63.04 ± 9.39	0.60	0.549	
	Edema	No	4759 (1.00)	4747 (0.99)	2.76	0.096
		Yes	16 (0.00)	28 (0.01)		
Blood test	BG	A	1238 (0.26)	1063 (0.22)	128.27	<0.001
		B	1571 (0.33)	2106 (0.44)		
		AB	484 (0.10)	417 (0.09)		
		O	1482 (0.31)	1189 (0.25)		
		Ne	24 (0.01)	15 (0.00)	1.65	0.199
	Blood RH	Po	4751 (0.99)	4760 (1.00)		
	ALB, g/L		41.24 ± 3.46	41.45 ± 2.75	-3.20	0.001
	ALT, U/L		20.65 ± 14.05	21.02 ± 13.69	-1.28	0.199
	AST, U/L		20.91 ± 7.81	22.14 ± 7.72	-7.74	<0.001
	Glu, mmol/L		4.57 [4.25, 4.93]	4.56 [4.23, 4.72]		<0.001
	Ca, mmol/L		2.30 ± 0.14	2.31 ± 0.12	-1.36	0.174
	Cr, umol/L		50.86 ± 7.61	51.17 ± 8.25	-1.92	0.055
	DB, umol/L		1.72 [1.10, 2.30]	1.74 [1.43, 1.90]		0.594
	TSI, umol/L		17.44 ± 3.33	17.60 ± 2.63	-2.61	0.009
	GLOB, g/L		27.28 ± 3.32	27.24 ± 2.43	0.73	0.466
	Mg, mmol/L		0.87 ± 0.13	0.88 ± 0.09	-4.43	<0.001
	IP, mmol/L		1.25 ± 0.15	1.25 ± 0.12	-1.61	0.108
	TBA, umol/L		3.83 [2.90, 5.10]	4.90 [3.32, 5.01]		<0.001
	TB, umol/L		11.28 ± 3.53	11.18 ± 2.64	1.67	0.095
	CHOL, mmol/L		4.78 ± 0.73	4.80 ± 0.38	-1.66	0.096
	TP, g/L		68.76 ± 4.84	68.78 ± 3.55	-0.22	0.829
	TG, mmol/L		1.52 ± 0.54	1.58 ± 0.41	-6.57	<0.001
	Urea, mmol/L		2.80 [2.38, 3.28]	2.84 [2.40, 3.10]		0.002
	UA, umol/L		199.75 ± 40.26	198.22 ± 39.37	1.88	0.060
	BA, 10e9/L		0.01 ± 0.03	0.01 ± 0.05	-1.01	0.314
	Plt, 10e9/L		220.31 ± 48.25	224.52 ± 48.60	-4.25	<0.001
	EOS, 10e9/L		0.09 ± 0.09	0.09 ± 0.07	0.43	0.665
	Hb, g/L		117.98 ± 8.56	117.69 ± 8.59	1.62	0.105
	MID, 10e9/L		0.55 ± 0.10	0.56 ± 0.12	-4.51	<0.001
	LY, 10e9/L		1.72 ± 0.40	1.75 ± 0.41	-2.92	0.004
	MCH, pg		31.49 ± 1.91	31.33 ± 1.85	4.15	<0.001
	MCHC, g/L		344.87 ± 10.25	343.39 ± 10.52	6.95	<0.001
	MCV, fL		91.31 ± 4.72	91.24 ± 4.50	0.76	0.445
MO, 10e9/L		0.53 ± 0.14	0.54 ± 0.14	-4.33	<0.001	
MPV, fL		8.58 ± 1.10	8.60 ± 1.09	-1.07	0.283	
NE, 10e9/L		7.23 ± 1.69	7.36 ± 1.72	-3.60	<0.001	
P-LCR, %		0.23 ± 0.05	0.23 ± 0.05	5.83	<0.001	
HCT, %		0.34 ± 0.02	0.35 ± 0.25	-1.18	0.238	
PCT, %		0.19 ± 0.04	0.19 ± 0.03	-0.23	0.819	
PDW, %		15.16 ± 2.25	14.83 ± 2.51	6.75	<0.001	
RDW-CV, %		0.16 ± 0.51	0.16 ± 0.35	0.81	0.416	
RDW-SD, fL		42.84 ± 2.46	43.45 ± 2.12	-13.07	<0.001	
RBC, 10e12/L		3.76 ± 0.31	3.77 ± 0.32	-1.51	0.131	
WBC, 10e9/L		9.58 ± 1.93	9.74 ± 1.97	-3.93	<0.001	

TABLE 2: Continued.

Variables			Control (4775)	Case (4775)	t/chi	P
Urine test strip	Urine pH		6.67 ± 0.46	6.73 ± 0.46	-6.77	<0.001
	USG		1.02 ± 0.01	1.02 ± 0.01	4.96	<0.001
	BIL	Ne	4737 (0.99)	4749 (0.99)	1.90	0.168
		Po	38 (0.01)	26 (0.01)		
	Glycosuria	Ne	3780 (0.79)	3820 (0.80)	0.98	0.322
		Po	995 (0.21)	955 (0.20)		
	KET	Ne	4593 (0.96)	4589 (0.96)	0.03	0.873
		Po	182 (0.04)	186 (0.04)		
	Nitrituria	Ne	4728 (0.99)	4740 (0.99)	1.49	0.222
		Po	47 (0.01)	35 (0.01)		
	Blood	Ne	4322 (0.91)	4397 (0.92)	7.22	0.007
		Po	453 (0.09)	378 (0.08)		
	Proteinuria	Ne	4729 (0.99)	4698 (0.98)	7.41	0.006
		Po	46 (0.01)	77 (0.02)		
	Bilirubinuria	Ne	4758 (1.00)	4755 (1.00)	0.11	0.742
Po		17 (0.00)	20 (0.00)			
Urine WBC	Ne	3490 (0.73)	3475 (0.73)	0.10	0.747	
	Po	1285 (0.27)	1300 (0.27)			
Gynecological examination	BV	Ne	4678 (0.98)	4719 (0.99)	10.63	0.001
		Po	97 (0.02)	56 (0.01)		
	CDV	1	854 (0.18)	975 (0.20)	60.20	<0.001
		2	2904 (0.61)	3066 (0.64)		
		3	845 (0.18)	590 (0.12)		
		4	172 (0.04)	144 (0.03)		
	VYI	Ne	4499 (0.94)	4549 (0.95)	5.05	0.025
Po		276 (0.06)	226 (0.05)			

ALB: serum albumin; ALT: alanine transaminase; AST: aspartate transaminase; BA: basophil granulocytes; BG: blood group; BIL: urine bilirubin; Blood RH: blood RH; BV: bacterial vaginosis; Ca: total calcium; CDV: cleaning degree of vagina, The higher the value, the worse the cleanliness; CHOL: total cholesterol; Cr: creatinine; DB: direct bilirubin; DBP: diastolic blood pressure; EOS: eosinophil granulocytes; FHR: fetal heart rate; GLOB: globulins; Glu: plasma glucose (fasting); Hb: hemoglobin; HCT: hematocrit; IP: serum inorganic phosphorus; KET: urine ketone bodies; LY: lymphocytes; MCH: mean cell hemoglobin; MCHC: mean corpuscular hemoglobin concentration; MCV: mean cell volume; Mg: magnesium; MID: intermediate cell; MO: monocytes; MPV: mean platelet volume; NE: neutrophil granulocytes; PCT: plateletcrit; PDW: platelet distribution width; P-LCR: mean platelet volume; Plt: platelet count; RBC: red blood cells; RDW-CV: red blood cell distribution width-CV; RDW-SD: red blood cell distribution width-SD; SBP: systolic blood pressure; TB: total bilirubin; TBA: total biliary acid; TG: triglycerides; TP: total protein; TSI: total serum iron; UA: uric acid; Urea: urea; Urine WBC: urine white blood cell; USG: urine specific gravity; VYI: vaginal yeast infection; WBC: white blood cell count; PTB: preterm birth. Variables that are not normally distributed were expressed as p50 [p25, p75].

TABLE 3: The performance of models in the test set.

	Models	Accuracy	AUC (95% CI)	Sensitivity	Specificity
Dataset 1	SVM	0.720	0.791 (0.771–0.811)	0.710	0.731
	RF	0.777	0.861 (0.841–0.871)	0.720	0.840
	NBM	0.677	0.741 (0.721–0.761)	0.705	0.646
	ANN	0.634	0.691 (0.671–0.711)	0.687	0.576
	K-means	0.611	0.681 (0.661–0.701)	0.794	0.412
	Log	0.610	0.701 (0.681–0.721)	0.378	0.861
Dataset 2	SVM	0.721	0.791 (0.781–0.811)	0.722	0.721
	RF	0.794	0.871 (0.851–0.881)	0.756	0.832
	NBM	0.682	0.771 (0.751–0.791)	0.785	0.581
	ANN	0.666	0.731 (0.711–0.751)	0.595	0.738
	K-means	0.602	0.681 (0.671–0.701)	0.811	0.393
	Log	0.606	0.701 (0.681–0.721)	0.364	0.847
Dataset 3	SVM	0.719	0.801 (0.781–0.811)	0.695	0.743
	RF	0.806	0.881 (0.871–0.901)	0.765	0.846
	NBM	0.674	0.791 (0.771–0.811)	0.837	0.515
	ANN	0.733	0.801 (0.791–0.821)	0.741	0.726
	K-means	0.612	0.711 (0.691–0.731)	0.824	0.405
	Log	0.633	0.701 (0.681–0.721)	0.421	0.839

TABLE 3: Continued.

	Models	Accuracy	AUC (95% CI)	Sensitivity	Specificity
Dataset 4	SVM	0.719	0.791 (0.781–0.811)	0.678	0.763
	RF	0.807	0.881 (0.871–0.891)	0.743	0.875
	NBM	0.626	0.741 (0.721–0.761)	0.328	0.946
	ANN	0.732	0.811 (0.801–0.831)	0.730	0.734
	K-means	0.626	0.721 (0.701–0.741)	0.801	0.436
	Log	0.611	0.701 (0.691–0.721)	0.361	0.880
Dataset 5	SVM	0.729	0.801 (0.781–0.811)	0.685	0.773
	RF	0.816	0.891 (0.871–0.901)	0.751	0.882
	NBM	0.622	0.741 (0.721–0.761)	0.315	0.937
	ANN	0.747	0.811 (0.801–0.831)	0.730	0.763
	K-means	0.609	0.701 (0.681–0.721)	0.780	0.434
	Log	0.623	0.691 (0.671–0.711)	0.391	0.861

NBM: Naive Bayesian; SVM: Support Vector Machine; RF: Random Forest Tree; ANN: Artificial Neural Networks; Log: Logistic regression; Dataset 1: 20 weeks gestation; Dataset 2: 22 weeks gestation; Dataset 3: 24 weeks gestation; Dataset 4: 26 weeks gestation; Dataset 5: 27 weeks gestation. AUC: the area under the curve; CI: confidence interval.

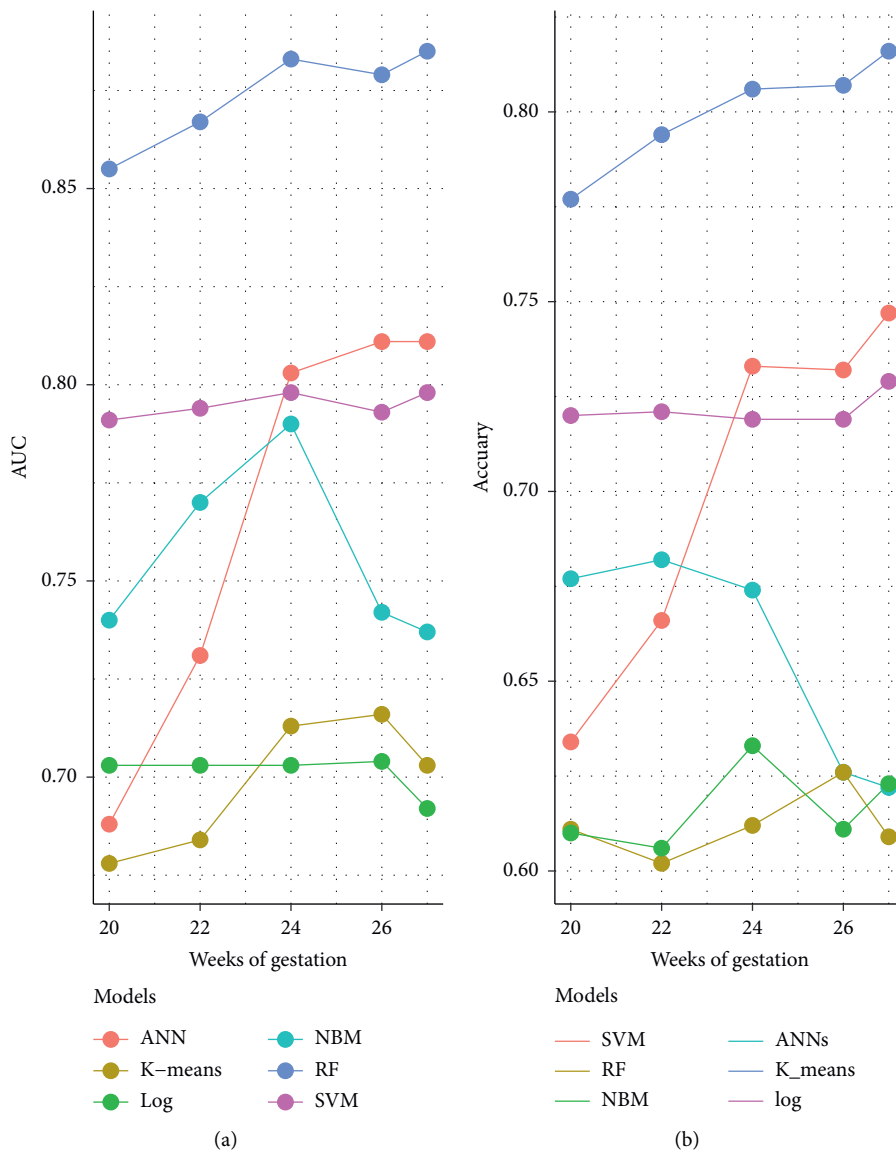


FIGURE 4: AUC (a) and accuracy (b) of models in different gestation times. (NBM: Naive Bayesian; SVM: Support Vector Machine; RF: Random Forest Tree; ANN: Artificial Neural Networks; Log: logistic regression; AUC: the area under the curve).

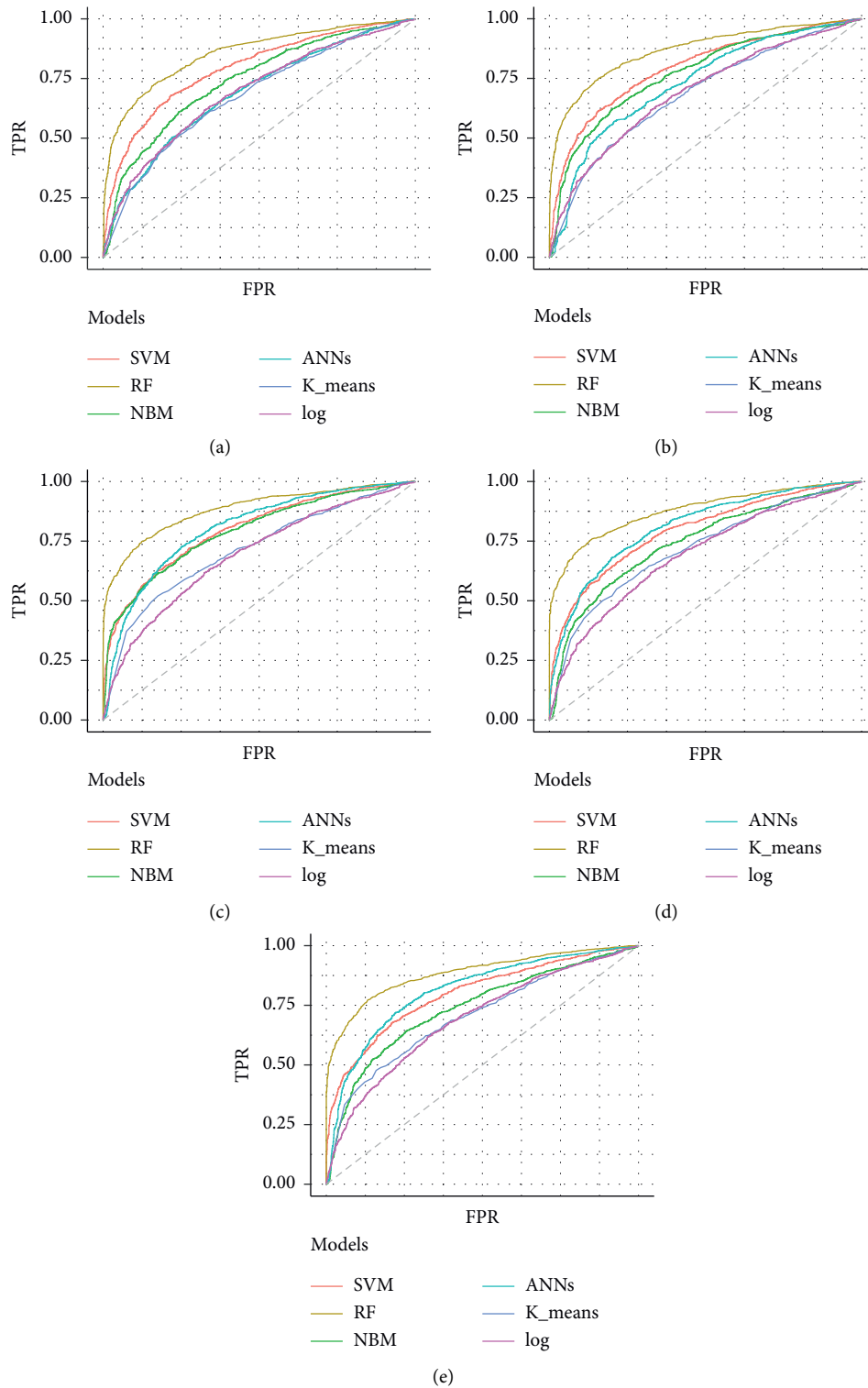


FIGURE 5: The ROC curve of the models. (a) Based on 20 weeks of gestation. (b) Based on 22 weeks of gestation. (c) Based on 24 weeks of gestation. (d) Based on 26 weeks of gestation. (e) Based on 27 weeks of gestation.

total bilirubin (TB) (Table 4). According to the importance of variables, we gradually increase the number of predictors, and the results show that the AUC of the model also increases gradually. The AUC of the model is stable when the number of predictors increases to 15 (Figure 6).

4. Discussion

In this study, six algorithms were used to establish the prediction model of premature birth in the early stage of gestation. The overall prediction effect of the RF model was

TABLE 4: The top 20 importance variables of RF model.

Variables	Decreased accuracy
Age (physical examination)	0.0251
Magnesium (blood test)	0.0098
Fundal height (physical examination)	0.0077
Serum inorganic phosphorus (blood test)	0.0038
Mean platelet volume (blood test)	0.0038
Waist size (physical examination)	0.0038
Total cholesterol (blood test)	0.0035
Triglycerides (blood test)	0.0031
Globulins (blood test)	0.0024
Total bilirubin (blood test)	0.0024
Neutrophil granulocytes (blood test)	0.0024
Red blood cell distribution width-SD (blood test)	0.0024
Bacterial vaginosis (gynecological examination)	0.0021
Urine bilirubin (urine test strip)	0.0021
Urine white blood cell (urine test strip)	0.0021
Diastolic blood pressure (physical examination)	0.0014
Blood group (blood test)	0.0014
Parity (physical examination)	0.0014
Eosinophil granulocytes (blood test)	0.0010
White blood cell count (blood test)	0.0010

RF: Random Forest tree.

better than that of other models. We also found that the predictive power of the RF model increased with the increase of gestational age. Age, magnesium, fundal height, serum inorganic phosphorus, mean platelet volume, waist size, TC, TG, globulins, and TB were found to be the main influencing factors of preterm birth.

In our study, we used the data from the production inspection to build the model based on the machine learning algorithm. The prediction performance of the model was relatively good, and the cost of the model was low. Ramkumar et al. using multivariate adaptive regression splines established a prediction model based on biomarkers (including IL-1RA, TNF- α , angiopoietin 2, TNFRI, IL-5, MIP1 α , IL-1 β , and TGF- α), resulting in a high AUC (train set: 0.82–0.98, test set: 0.66–0.86) [25]. Teresa et al. used cervical length at admission, gestational age, amniotic fluid glucose, and interleukin-6 to establish a prediction model, resulting in a high AUC (0.86, 95% CI: 0.77–0.95) [26]. Thuy et al. found that nine cell-free RNA could be used to predict gestational age and preterm delivery, and the AUCs of preterm delivery were 0.86 in the discovery cohort and 0.81 in the validation cohort [27]. In these studies, the prediction performance of the preterm birth model was better, but another clinical test was needed and expensive. Kamala et al. used a combination of neighborhood socioeconomic status and individual status to predict preterm birth, but the AUC (0.75) of the model was relatively low [28]. Liu et al. found that cervical elastography could be used as a predictive indicator, and the AUC of the model was 0.73 [29]. The above studies used a traditional biological algorithm, such as logistic regression, to build the model, but the predictive power of the model is relatively low.

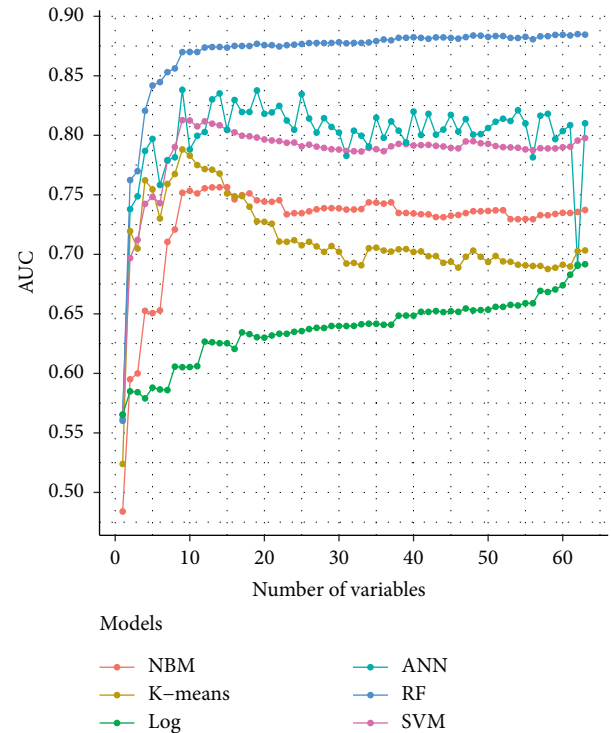


FIGURE 6: The AUC of the model increases with the number of predictors. (NBM: Naive Bayesian; SVM: Support Vector Machine; RF: Random Forest Tree; ANN: Artificial Neural Networks; Log: logistic regression; AUC: the area under the curve).

In this study, the results of the numerical experiments show that the AUC of SVM, RF, and ANN models were higher than logistic, NBM, and k-means. The possible reason for the low AUC of the NBM model is that the NBM model assumes that features are independent of each other, which is often not true in practice. For logistic regression and k-means algorithms, they were susceptible to outliers and noise that reduce prediction accuracy. For the other 3 machine algorithms, the AUC value of the RF model was the highest. The RF model is an ensemble learning method, which constructs a multitude of decision trees at training time and then sets up the trees to give the classification [30]. This ensemble strategy makes several weak classifiers form a strong classifier to improve the predictive ability of the model. In a recent study, the RF algorithm had also achieved a good predictive effect in fatty liver disease [31], suggesting that the RF algorithm had advantages in the processing of clinical electronic medical records. Moreover, we found that the prediction performance of RF was the best at 27 weeks of gestation. This may be due to alternation of biochemical indexes in pregnant women as delivery approached. The AUC of the model based on random forest in 20 weeks of gestation was 0.855 (95% CI: 0.841–0.869), suggesting that interventions could be performed before these biochemical indicators change.

In the importance analysis of the RF model, we found that age was the greatest effect on preterm birth. A case-control study showed that premature delivery was associated with greater maternal age [32]. We also found that serum magnesium had a great influence on the results of the model. A

double-blind study suggested that magnesium supplementation during pregnancy is associated with a reduction in preterm delivery [33]. Maternal fundal height was found to be a valuable predictor for PTB in our study. Previous study used maternal fundal height to predict fetal weight [34], suggesting that fundal height was a good predictor for PTB. The measurement of fundal height is susceptible to measurement personnel, which may limit its clinical use. Della Rosa et al. used 9 most informative predictors to build a preterm birth prediction model, and the AUC of the model reached 0.812 [35]. Our results show that using only 15 predictions can achieve better model predictions. Considering the cost effect, this result has important implications for guiding clinical practice.

There were some limitations in our study. First, our dataset, collection from electronic medical records, and lack of some data such as smoking, drinking, family income, method of conception, medication, and fetal fibronectin. The absence of these factors may underperform our model. Second, previous studies found that the conception method has an important effect on preterm birth [36, 37], but it was not included in our model, which may affect the prediction accuracy of our model. Third, controls of the study were matched 1:1 from contemporaneous hospitals, which may overestimate the performance of the model and may limit the use of the model to a normal proportion of the population.

5. Conclusions

Our results indicated that the prediction model based on the RF algorithm had a potential value to predict preterm birth early stage of pregnancy. The RF model also found the main influence factors of PTB, suggesting that intervention in the early stages of pregnancy could decrease the risk of preterm birth.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

X Ma, JB Lu, and Y Yang helped with protocol development. YS Yan and HG Zhang collected the data. S Wang, YM Gao, HY Liu, and SY Liu analyzed the data. Q Sun and XX Zou wrote the manuscript. Qi Sun and Xiaoxuan Zou contributed to the work equally.

Acknowledgments

The authors thank all pregnant women who participated in the study. This work was supported by the National Key Research and Development Program of China (2016YFC1000307) and the subproject of National Key Research and Development Program of China (2016YFC1000307-10).

Supplementary Materials

Table S1: prenatal testing of pregnant women before 20 weeks of gestation between the PTB group and the control group. Table S2: prenatal testing of pregnant women before 22 weeks of gestation between the PTB group and the control group. Table S3: prenatal testing of pregnant women before 24 weeks of gestation between the PTB group and the control group. Table S4: prenatal testing of pregnant women before 26 weeks of gestation between the PTB group and the control group. (*Supplementary Materials*)

References

- [1] R. L. Goldenberg, F. C. Jennifer, D. I. Jay, and R. Roberto, "Epidemiology and causes of preterm birth," *Lancet*, vol. 371, no. 9606, pp. 75–84, 2008.
- [2] H. Blencowe, S. Cousens, M. Z. Oestergaard, and D. Chou, "National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications," *Lancet*, vol. 379, no. 9832, pp. 2162–2172, 2012.
- [3] S. Beck, D. Wojdyla, L. Say et al., "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity," *Bulletin of the World Health Organization*, vol. 88, no. 1, pp. 31–38, 2010.
- [4] S. Chawanpaiboon, K. Watananirun, P. Lumbiganon et al., "Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis," *Lancet Global Health*, vol. 7, no. 1, pp. e37–e46, 2019.
- [5] H. A. Frey and M. A. Klebanoff, "The epidemiology, etiology, and costs of preterm birth," *Seminars in Fetal & Neonatal Medicine*, vol. 21, no. 2, pp. 68–73, 2016.
- [6] B. J. Stoll, N. I. Hansen, E. F. Bell et al., "Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network," *Pediatrics*, vol. 126, no. 3, pp. 443–456, 2010.
- [7] D. You, "Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation," *Lancet*, vol. 386, no. 10010, pp. 2275–2286, 2015.
- [8] E. Naumburg and L. Soderstrom, "Increased risk of pulmonary hypertension following premature birth," *BMC Pediatrics*, vol. 19, no. 1, p. 288, 2019.
- [9] A. M. Lynch, Wagner, Hodges, T. S. Thevarajah, McCourt, and Cerda, "The relationship of the subtypes of preterm birth with retinopathy of prematurity," *American Journal of Obstetrics and Gynecology*, vol. 217, no. 3, pp. 354 e1–354 e8, 2017.
- [10] M. Hirvonen, "Visual and hearing impairments after preterm birth," *Pediatrics*, vol. 142, no. 2, 2018.
- [11] T. M. Luu, M. O. Rehman Mian, and A. M. Nuyt, "Long-term impact of preterm birth: neurodevelopmental and physical health outcomes," *Clinics in Perinatology*, vol. 44, no. 2, pp. 305–314, 2017.
- [12] L. T. Singer, A. Salvator, and S. Guo, "Maternal psychological distress and parenting stress after the birth of a very low-birth-weight infant," *JAMA*, vol. 281, no. 9, pp. 799–805, 1999.
- [13] Y. Ville and P. Rozenberg, "Predictors of preterm birth," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 52, pp. 23–32, 2018.
- [14] E. Greco, R. Gupta, A. Syngelaki, L. C. N. Poon, and K. H. Nicolaides, "First-trimester screening for spontaneous

- preterm delivery with maternal characteristics and cervical length,” *Fetal Diagnosis and Therapy*, vol. 31, no. 3, pp. 154–161, 2012.
- [15] A. P. Souka, I. Papastefanou, V. Michalitsi, K. Salambasis, C. Chrelias, and G. Salamalekis, “Cervical length changes from the first to second trimester of pregnancy, and prediction of preterm birth by first-trimester sonographic cervical measurement,” *Journal of Ultrasound in Medicine*, vol. 30, no. 7, pp. 997–1002, 2011.
- [16] P. Tsikouras, G. Galazios, A. Zalvanos, A. Bouzaki, and A. Athanasiadis, “Transvaginal sonographic assessment of the cervix and preterm labor,” *Clinical & Experimental Obstetrics & Gynecology*, vol. 34, no. 3, pp. 159–162, 2007.
- [17] G. Conoscenti, Y. J. Meir, G. D’Ottavio, M. A. Rustico, R. Pinzano, L. F. Tamaro, and T. Stampalija, “Does cervical length at 13–15 weeks’ gestation predict preterm delivery in an unselected population?” *Ultrasound in Obstetrics and Gynecology*, vol. 21, no. 2, pp. 128–134, 2003.
- [18] D. Alexander, “The national Institute of Child health and human development and phenylketonuria,” *Pediatrics*, vol. 112, no. 6 Pt 2, pp. 1514–1515, 2003.
- [19] F. Lucaroni, “Biomarkers for predicting spontaneous preterm birth: an umbrella systematic review,” *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 31, no. 6, pp. 726–734, 2018.
- [20] O. I. Abiodun, A. J. Abiodun, and Dada, “State-of-the-art in artificial neural network applications: a survey,” *Heliyon*, vol. 4, no. 11, Article ID e00938, 2018.
- [21] K. Y. Ngiam and L. W. Khor, “Big data and machine learning algorithms for health-care delivery,” *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.
- [22] I. Vovsha, “Predicting preterm birth is not elusive: machine learning paves the way to individual wellness,” in *Proceedings of the aaai spring symposium*, Palo Alto, California, March 2014.
- [23] R. Raja, I. Mukherjee, and B. K. Sarkar, “A machine learning-based prediction model for preterm birth in rural India,” *J Healthc Eng*, vol. 2021, Article ID 6665573, 11 pages, 2021.
- [24] A. Weber, G. L. Darmstadt, S. Gruber, M. E. Foeller, S. L. Carmichael, and D. K. Stevenson, “Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women,” *Annals of Epidemiology*, vol. 28, no. 11, pp. 783–789.e1, 2018.
- [25] R. Menon, “Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth,” *Acta Obstetrica et Gynecologica Scandinavica*, vol. 93, no. 4, pp. 382–391, 2014.
- [26] T. Cobo, Aldecoa, and Herranz, “Development and validation of a multivariable prediction model of spontaneous preterm delivery and microbial invasion of the amniotic cavity in women with preterm labor,” *American Journal of Obstetrics and Gynecology*, vol. 223, no. 3, pp. 421.e1–421.e14, 2020.
- [27] T. T. M. Ngo, M. N. Moufarrej, M. L. H. Rasmussen, J. C. Soler, and W. Pan, “Noninvasive blood tests for fetal development predict gestational age and preterm delivery,” *Science*, vol. 360, no. 6393, pp. 1133–1136, 2018.
- [28] K. Adhikari, Patten, Williamson et al., “Does neighborhood socioeconomic status predict the risk of preterm birth? A community-based Canadian cohort study,” *BMJ Open*, vol. 9, no. 2, Article ID e025341, 2019.
- [29] L. Du, Zhang, Zheng, Xie, Gu, and Lin, “Evaluation of cervical elastography for prediction of spontaneous preterm birth in low-risk women: a prospective study,” *Journal of Ultrasound in Medicine*, vol. 39, no. 4, pp. 705–713, 2020.
- [30] T. K. Ho, “Random decision forests,” in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, IEEE, Montreal, Canada, August 1995.
- [31] N. Atabaki-Pasdar, Ohlsson, Viñuela, Frau, and Millan, “Predicting and elucidating the etiology of fatty liver disease: a machine learning modeling and validation study in the IMI DIRECT cohorts,” *PLoS Medicine*, vol. 17, no. 6, Article ID e1003149, 2020.
- [32] P. Stylianou-Riga, P. Kouis, P. Kinni, A. Rigas, T. Papadouri, and P. K. Yiallouroux, “Maternal socioeconomic factors and the risk of premature birth and low birth weight in Cyprus: a case-control study,” *Reproductive Health*, vol. 15, no. 1, p. 157, 2018.
- [33] L. Spatling and G. Spatling, “Magnesium supplementation in pregnancy. A double-blind study,” *British Journal of Obstetrics and Gynaecology*, vol. 95, no. 2, pp. 120–125, 1988.
- [34] D. Anggraini, M. Abdollahian, and K. Marion, “Accuracy assessment on prediction models for fetal weight based on maternal fundal height,” in *Information Technology: New Generations*, pp. 859–868, Springer, 2016.
- [35] P. A. Della Rosa, Miglioli, Caglioni, Tiberio, Mosser, and Vignotto, “A hierarchical procedure to select intrauterine and extrauterine factors for methodological validation of preterm birth risk estimation,” *BMC Pregnancy and Childbirth*, vol. 21, no. 1, p. 306, 2021.
- [36] P. Cavoretto, M. Candiani, V. Giorgione et al., “Risk of spontaneous preterm birth in singleton pregnancies conceived after IVF/ICSI treatment: meta-analysis of cohort studies,” *Ultrasound in Obstetrics and Gynecology*, vol. 51, no. 1, pp. 43–53, 2018.
- [37] P. I. Cavoretto, V. Giorgione, A. Sotiriadis et al., “IVF/ICSI treatment and the risk of iatrogenic preterm birth in singleton pregnancies: systematic review and meta-analysis of cohort studies,” *Journal of Maternal-Fetal and Neonatal Medicine*, pp. 1–10, 2020.

Research Article

Multiview Volume and Temporal Difference Network for Angle-Closure Glaucoma Screening from AS-OCT Videos

Luoying Hao ¹, Yan Hu ¹, Risa Higashita ², James J. Q. Yu ¹, Ce Zheng ³,
and Jiang Liu ^{1,4,5}

¹Department of Computer Science and Engineering, Southern University of Science and Technology, 518055 Shenzhen, China

²Tomey Corporation, 451-0051 Nagoya, Japan

³Department of Ophthalmology, Xinhua Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

⁴School of Ophthalmology & Optometry, School of Biomedical Engineering, Wenzhou Medical University, Zhejiang, China

⁵Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, 518055 Shenzhen, China

Correspondence should be addressed to Yan Hu; huy3@sustech.edu.cn

Received 16 December 2021; Accepted 15 March 2022; Published 7 April 2022

Academic Editor: Weihua Yang

Copyright © 2022 Luoying Hao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Precise and comprehensive characterizations from anterior segment optical coherence tomography (AS-OCT) are of great importance in facilitating the diagnosis of angle-closure glaucoma. Existing automated analysis methods focus on analyzing structural properties identified from the single AS-OCT image, which is limited to comprehensively representing the status of the anterior chamber angle (ACA). Dynamic iris changes are evidenced as a risk factor in primary angle-closure glaucoma. **Method.** In this work, we focus on detecting the ACA status from AS-OCT videos, which are captured in a dark-bright-dark changing environment. We first propose a multiview volume and temporal difference network (MT-net). Our method integrates the spatial structural information from multiple views of AS-OCT videos and utilizes temporal dynamics of iris regions simultaneously based on image difference. Moreover, to reduce the video jitter caused by eye movement, we employ preprocessing to align the corneal part between video frames. The regions of interest (ROIs) in appearance and dynamics are also automatically detected to intensify the related informative features. **Results.** In this work, we employ two AS-OCT video datasets captured by two different devices to evaluate the performance, which includes a total of 342 AS-OCT videos. For the Casia dataset, the classification accuracy for our MT-net is 0.866 with a sensitivity of 0.857 and a specificity of 0.875, which achieves superior performance compared with the results of the algorithms based on AS-OCT images with an obvious gap. For the Zeiss AS-OCT video dataset, our method also gets better performance against the methods based on AS-OCT images with a classification accuracy of 0.833, a sensitivity of 0.860, and a specificity of 0.800. **Conclusions.** The AS-OCT videos captured under changing environments can be a comprehended means for angle-closure classification. The effectiveness of our proposed MT-net is proved by two datasets from different manufacturers

1. Introduction

Glaucoma is an eye disease with extremely complex etiology, ranking second among the four major blinding eye diseases. By 2040, it is estimated that 112 million people in the world will be affected by this disease [1, 2]. Globally, glaucoma (40–80 years old) is estimated to increase to 66–80 million people worldwide by 2020, and 11 million of these patients will eventually become blind [1]. With the ageing of the population, the number of glaucoma patients is increasing year by year. In China, primary

angle-closure glaucoma (PACG) is more prevalent. But fortunately, it is preventable after early treatment of anterior chamber angle (ACA), such as laser peripheral iridotomy (LPI). Therefore, early screening and treatment are critical. Recently, anterior segment optical coherence tomography (AS-OCT) is widely accepted by ophthalmologists in glaucoma examination because of its efficient and noncontact imaging anterior chamber with depth information [3].

The shallow anterior chamber is an important risk factor for PACG [4–6], so ophthalmologists often judge the open

or closure status of ACA from AS-OCT. Some computer-aided angle-closure classification algorithms based on ACA are proposed to reduce the doctors' burdens based on machine learning [7–10] or convolutional neural network (CNN) [4, 11, 12]. Most of the present algorithms give out the classification results based on several statically captured AS-OCT images. However, static anatomical factors alone cannot fully explain the relatively high prevalence of PACG and dynamic changes of the anterior chamber structure are more convincing for the diagnosis [13]. For example, as shown in Figure 1, we randomly selected two video samples with PACG and normal ACA. Figure 1(a) is a PACG video sample with the angle status in dark (3rd frame) and bright conditions (55th frame), while Figure 1(b) shows a normal sample with the angle status in dark (4th frame) and bright conditions (34th frame). The video frames under light conditions in Figure 1 are compared when the pupil contracts to the maximum. For the two samples, it is noted that the ACA status is almost closed in dark environments, but after light illumination, it is changed to open. It will lead to inconsistent results for the same sample if only based on a single image.

Thus, it is difficult to distinguish the patients' types only by statically captured AS-OCT images, and most of the present angle-closure classification methods, only based on the angle status of a certain state, have certain limitations [14–16]. But it is correctly classified by the iris motion state (such as the iris motion information as shown in Figure 1, which also can better reflect the complete angle state of the eyes at different times). There is some research explaining this phenomenon. The iris is spongy and compressible in the eyes of healthy and PACG subjects, but it is incompressible in the eyes of PACG and suspected angle-closure [17]. Moreover, the movement features of angle-closure eyes and angle-opening eyes are researched, and the angle-closure group has a slower iris contraction speed in the reflection of light, which is faster after receiving effective treatment [18]. Iris elastic acceleration and pupil block acceleration are correlated with PACG [19]. Therefore, in this article, the angle-closure detection is based on the AS-OCT videos, which are captured in the dark-bright-dark changing environments. As far as we know, there is no research on angle-closure detection concerning the movement of iris based on AS-OCT videos.

In this article, a deep learning-based framework is proposed for angle-closure detection that makes use of AS-OCT videos. The contributions are summarized as follows: (1) we first propose to detect the chamber angle status based on AS-OCT videos in changing environments, which are proven to be more complete representation of the patients' anterior chamber. (2) We propose a multiview volume and

temporal difference network (MT-net) for ACA status detection, which integrates the spatial structural information from multiple views of AS-OCT videos and simultaneously utilizes temporal dynamics based on image difference. (3) We propose an automated AS-OCT video alignment algorithm based on the corneal part in video frames, to reduce the impacts of video jitter. Regions of interest (ROIs) in 3D appearance and dynamics are also detected based on the position of the scleral spur (SS) and image difference to enlarge the informative features. (4) We carry out comparison and ablation study experiments to demonstrate the effectiveness of our proposed algorithm by seven evaluation metrics based on two AS-OCT video datasets.

2. The Proposed Method

Figure 2 illustrates the framework of our proposed MT-net (short of multiview volume and temporal difference network). First, the AS-OCT video jitter is removed by the automated image registration method, and the ACA is located by extracting the position of the SS, while motion information is obtained by image difference. Then, the proposed MT-net is introduced that multiple views of ACA volumes are fed to extract spatial features, while the motion feature is input to study temporal information of iris dynamic. Finally, the prediction scores based on spatial and temporal information are integrated to further enhance the performance of angle-closure detection.

2.1. AS-OCT Video Alignment and ROIs Extraction

2.1.1. AS-OCT Video Alignment. Due to the impacts of involuntary eye movement and improper placement of the optical axis of the eye, misalignment exists between adjacent video frames. As shown in Figure 3, the corneal in the 1st and 38th frame cannot overlap, which may lead to the resulting video frame sequence being unreliable [20].

Assume a video contains N frames, and the frames are denoted by $f_i (i \in [1, N])$. To ensure the consistency of the placement of the anterior chamber structure in the video frames, we transform the frames $f_i (i \in [2, N])$ into the coordinate system of frame f_1 and crop the transformed frames to be the same dimension as f_1 . First, the multiscale face point features p^f and corner-like features p^c are extracted from the frames. Rotation, translation, and scale are considered the main changes between video frames; thus, the affine transformation parameters θ are estimated based on the similarity metrics, and an iterative optimization process is further used to refine the transformation, defined as follows:

$$\mathcal{E}(\theta; \zeta_f, \zeta_c) = \operatorname{argmin} \left(\sum_{(p_i^f, q_i^f) \in \zeta_f} \omega_f \rho(d_f(p_i^f, q_i^f, \theta)) + \sum_{(p_i^c, q_i^c) \in \zeta_c} \omega_c \rho(d_c(p_i^c, q_i^c, \theta)) \right), \quad (1)$$

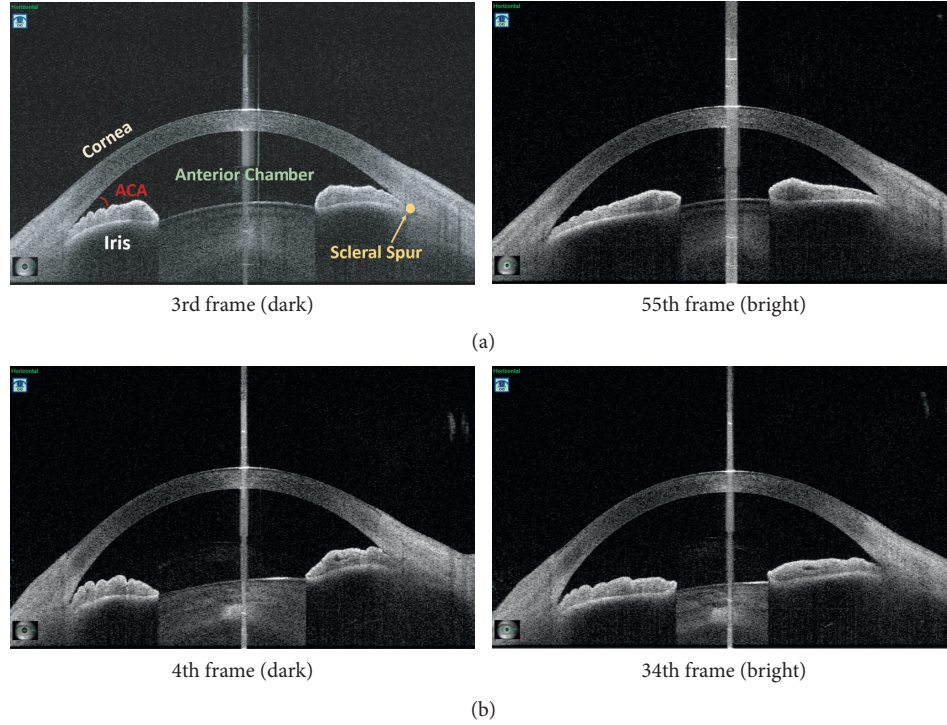


FIGURE 1: The example of (a) angle-closure and (b) open angle (normal) video.

where $d_f(\cdot)$ is the normal distances of pair face points, $d_c(\cdot)$ is the Euclidean distances of pair corner points, and $\rho(\cdot)$ is the Beaton–Tukey [21]. The face point feature matching sets and corner feature matching sets are denoted as $(p_i^f, q_i^f) \in \zeta_f$ and $(p_i^c, q_i^c) \in \zeta_c$, respectively. The ω_f and ω_c are the distance-based robust weight factors.

Besides, to speed up the alignment procedure, median filter and frame resize are adopted before alignment. As shown in Figure 3(b), the corneal is overlapped between frames after alignment.

2.1.2. ROIs Extraction. The ACA and iris region are the ROIs during ophthalmologists diagnosing PACG [22]. In this study, ROI extraction includes ACA extraction and image difference, which can reinforce ACA spatial and iris temporal representation.

(1) ACA Extraction. Locating local regions can retain more useful information at the last feature map of the backbone network [11, 23–25]. The SS is the key point of the ACA; thus, we obtain the ACA for angle status detection by SS localization in the article. We propose to use a UNet-like architecture based on nested and dense skip connections (UNet++) [26] to get accurate SS localization. Then, the ACAs are cropped directly from aligned videos and resized to one fixed resolution. In this way, the network can focus on visual contents by cropped bounding boxes. Moreover, the scenes of frame inputs are enlarged to capture more useful visual content.

(2) Image Difference. To better extract long-term temporal information, a motion representation is carried on to obtain iris motion first. For motion modelling, the optical flow has been used extensively as a motion representation [27, 28]. However, the extraction of optical flow is expensive in both time and space, which is often calculated in advance and then stored in hard drives. Motivated by this, efforts have been made to find good alternatives. Researchers [29, 30] found that the difference between adjacent frames, namely, image difference, can be useful instead of optical flow.

In this study, image difference, also known as the Eulerian motion, is used to represent the motion of images [31]. Instead of calculating the motion between consecutive frames in a video, this article puts the focus on the iris change compared to the first frame. As shown in Figure 2, the image difference of two images is defined as $V = I_t - I_1$, where I_t is a frame within the scope of $[2, N]$, while I_1 is the first frame in a video. Image differences can capture the short-term motion information to effectively facilitate to model longer-range temporal relations in videos.

2.2. MT-net Framework. The proposed MT-net framework is composed of two subnetworks, multiview volume subnetwork for spatial information (as shown in Figure 2(a)) and temporal difference subnetwork (as shown in Figure 2(b)) for temporal information.

2.2.1. Multiview Volume Subnetwork. The ACAs contain spatial information in the video frame sequence. In this work, the ACAs are composed as a volume with size $H \times W \times T$ as Figure 2(a)(a1), which provides context

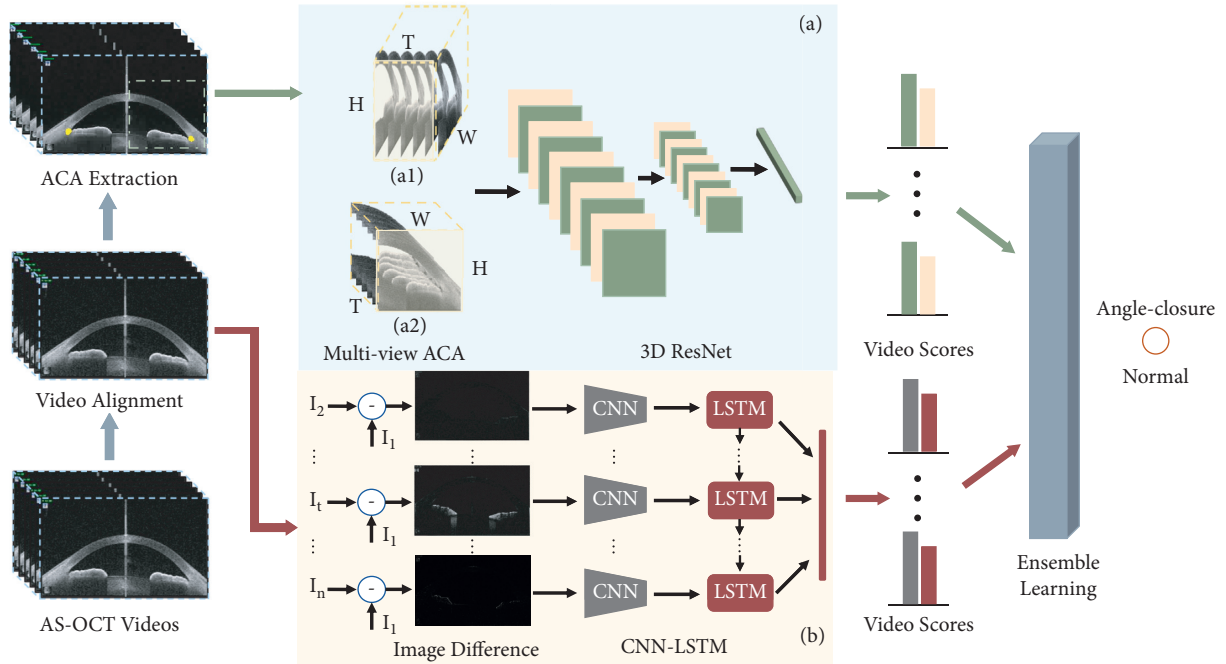


FIGURE 2: The pipeline of our MT-net architecture. The input AS-OCT videos are aligned to reduce the video jitter. Then, the ACA extraction and image difference are carried on for the two subnetworks: (a) multiview volume network and (b) temporal difference network. Finally, a soft voting-based ensemble model is adopted to incorporate the two subnetworks to output the final classification results.

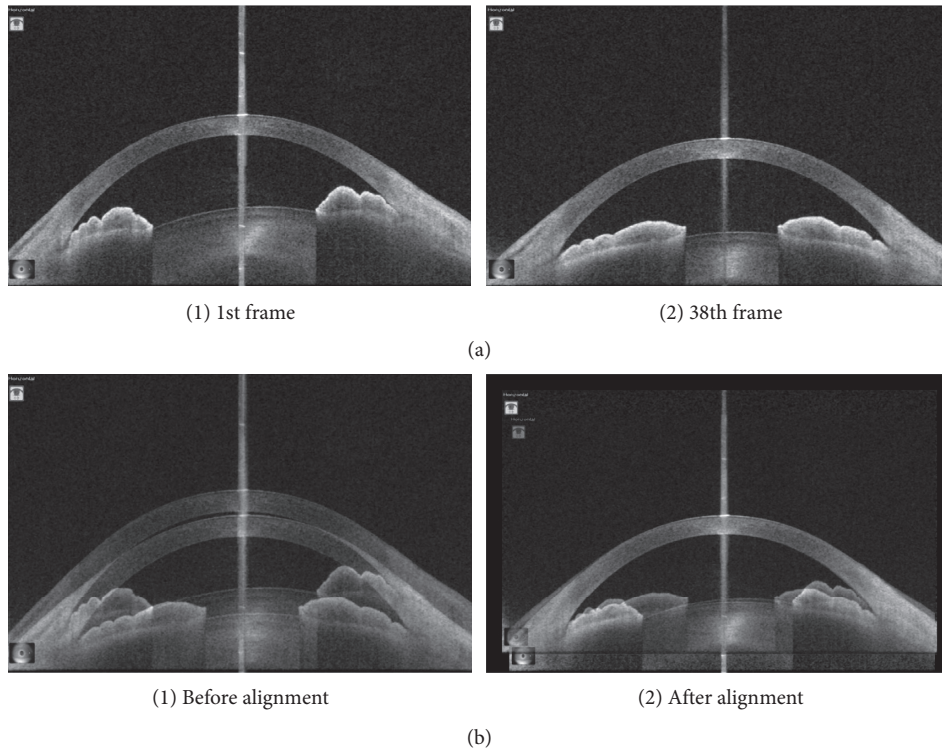


FIGURE 3: The example of an angle-closure video alignment. (a) Origin angle-closure video. (b) Alignment effect.

information of ACAs in the time dimension T . During the volume analysis, we find that when the volume is rotated with size $H \times T \times W$ as Figure 2(a)(a2), it reveals the fluctuation characteristics of ACAs. Thus, to adopt more useful

information for angle status classification, we propose a multiview volume subnetwork by integrating the above different-view volumes. The 3D ResNet is adopted as the backbone since it makes full use of the 3D context

information and is easier to optimize with high accuracy from considerably increased depth [32]. The sizes of convolutional kernels in 3D ResNets are $3 \times 3 \times 3$, and both the temporal and spatial stride are 2. The 16-frame ACA clips are input into the subnetwork with the size of $3 \times 16 \times 224 \times 224$. Since the small scale of the medical image dataset is the main reason for low classification accuracy, fine-tuning pretrained model on large-scale datasets becomes an effective way [33]. We also fine-tune the pretrained 3D ResNets model on Kinetics [34]. Also, identity connections and zero paddings for the shortcuts of the ResNet block are utilized to avoid the increasing number of parameters [35].

2.2.2. Temporal Difference Subnetwork. As the feature of iris dynamic movement under the dark-light-dark environment is helpful for the angle-closure state classification, temporal information of the AS-OCT videos is adopted in the article. To reduce the computation complexity of the subnetwork, we propose to apply a ResNet model to extract features of image difference. Then, the extracted features are input into the long short-term memory (LSTM) layer with batch normalization [36], which encodes the states and models the long-term dependencies between the feature map along the time axis. Finally, a fully connected layer on the top of LSTM output is adopted for multiway classification [34].

2.2.3. Angle-Closure Detection. Temporal information plays an important role in understanding the iris motion, while ACA volume provides anatomical features of the anterior segment at different times. We take into account two kinds of context information in our model: scene volume context and temporal changing information over the entire span of videos. Finally, we adopt the model ensemble, specifically the soft voting ensemble method [37], to integrate multifaceted contents and obtain a more comprehensive and accurate classification result. The soft voting ensemble method is a soft variant of a voting scheme that takes into account the class probabilities of each algorithm and combines these decisions through the averaging process, instead of hard voting through on-off decisions [37]. In this article, we independently train each subnetwork, get the probability distribution of the test set (As shown in Figure 2), and finally synthesize the performance of different classifiers of each subject to get the final classification results.

3. Experimental Results

3.1. Clinical AS-OCT Video Dataset. Our AS-OCT video datasets are collected by two devices: Swept-source OCT [38] (Casia Swept-source-1000 OCT, Tomey, Nagoya, Japan) and Visante OCT [39] (Visante OCT, Model 1000, software version 2.1; Carl Zeiss Meditec). We collect the AS-OCT videos of normal people and patients with PACG under a dark-light-dark environment. Subjects are recruited from the outpatient and inpatient departments of the Singapore National Eye Centre (SNEC) and joint Shantou International Eye Centre of Shantou University and the Chinese

University of Hong Kong, which include patients and volunteers aged over 40 years. In particular, the recording of the AS-OCT videos is started one minute after dark adaption using a standard protocol, and the light intensity is approximately 20 lux. The iris and anterior chamber changes between the dark and light environments are recorded. A single ophthalmologist performs all AS-OCT testing for data consistence. For each video, the ground-truth label of normal or angle-closure is determined from the majority diagnosis of senior ophthalmologists.

For the Casia dataset, it includes 148 videos, including 68 videos of normal eyes and 80 videos of eyes with PACG. The resolution of video frames is 1644×1000 . The Zeiss dataset consists of 194 videos, including 116 videos of normal eyes and 78 videos of eyes with PACG. The resolution of video frames is 600×300 . For the two datasets, Table 1 lists the maximum, minimum, and median of video frames. We equally and randomly divide 30 videos as the testing set, while the remaining videos are divided into the validation set and training set. The size of all input video frames for the deep learning network is fixed at 224×224 .

3.2. Implementation Details. The proposed architecture is implemented using the publicly available PyTorch Library. In the training phase, for the multiview volume subnetwork, we utilize stochastic gradient descent to optimize the model (200 epochs), with a gradually decreasing learning rate starting from 0.1, a momentum of 0.9, and a batch size of 128. For the temporal difference subnetwork, we employ an Adam optimizer to optimize the model (180 epochs), with a learning rate of 0.0001, a momentum of 0.01, and a batch size of 128. For all the processes of training and testing, we conduct them on one NVIDIA TITAN V GPU.

3.3. Experimental Criterion and Baseline. To measure the performance of our network, we employ seven evaluation criteria: balanced accuracy (B-Acc), precision (Pre), recall, F1 score, sensitivity (Sen), specificity (Spe), and Kappa analysis. Kappa analysis and F1 score are used to reflect the trade-offs between Sen and Spe.

As shown in Table 2, we use the basic subnetwork backbones of 3D CNN and CNN-LSTM to conduct training and testing on our private Casia dataset. For a small-scale medical image dataset, different proportions of validation set and training set affect the anterior chamber status classification. We conduct experiments for the two subnetworks with the proportion of validation set and training set to 5%, 10%, and 20%, and the results are shown in Table 2.

For 3D CNN, it can be seen from Table 2 that 3D ResNet18 has the highest B-Acc and F1 score of the three dataset splits. In the training process, the relatively shallow network is easier to converge than the deeper network. For the experiment of CNN-LSTM, the ResNets are fine-tuned from initialization with the pretrained deep model. As shown in Table 2, based on the same testing set, the B-Acc and F1 scores of this network are basically higher than that of 3D CNN. The possible reason is that CNN-LSTM models the global movement of the iris better, which also further proves

TABLE 1: The maximum, minimum, and median of video frames for the two datasets.

	Maximum	Minimum	Median
Casia dataset	121	21	53
Zeiss dataset	135	20	48

TABLE 2: Performance of different subnetworks on the private Casia video dataset.

3D CNN (B-Acc/F1 score)					
Splits	18-Layer	34-Layer	50-Layer	101-Layer	152-Layer
5%	0.638/	0.464/	0.625/	0.562/	0.558/
	0.632	0.282	0.627	0.430	0.463
10%	0.692/	0.518/	0.612/	0.594/	0.562/
	0.695	0.463	0.589	0.487	0.430
20%	0.589/	0.482/	0.509/	0.589/	0.562/
	0.589	0.437	0.492	0.514	0.430
CNN-LSTM (B-Acc/F1 score)					
Splits	18-Layer	34-Layer	50-Layer	101-Layer	152-Layer
5%	0.531/	0.643/	0.777/	0.607/	0.719/
	0.367	0.614	0.763	0.562	0.678
10%	0.500/	0.679/	0.781/	0.714/	0.656/
	0.371	0.662	0.789	0.707	0.589
20%	0.500/	0.754/	0.710/	0.714/	0.571/
	0.297	0.757	0.695	0.707	0.505

The bold values indicate the optimal results.

that iris motion features are important to predict the binary classification (angle status) result. The testing accuracy of CNN-LSTM shows the best performance at the 50th layer with the increase in depth. The performance of both 3D CNN and CNN-LSTM on the data splits of 5 % and 10 % is much better than those of 20 % . Therefore, in the follow-up experiments, we conduct training on the two dataset splits and take the average testing values as final results.

3.4. Ablation Study. To evaluate the effectiveness of four modules in our framework, including alignment, ACA extraction, image difference, 3D CNN, and CNN-LSTM, we provide an ablation study. Based on the baseline experiments, we employ 3D ResNet18 and ResNet50-LSTM as baselines in the following experiments, and the results are reported in Table 3.

The scleral spur localization is very important for the classification, Thus, in the article, we adopt UNet++ to get accurate SS localization. The model is trained based on the public AGE dataset [6], which is similar to our dataset. For very few video frames that cannot locate SS, we get it from the SS position of the frame preceding the current frame of the aligned video.

- (i) For the volume spatial information, video alignment and ACA region extraction improve the classification results of 3D CNN to a certain extent compared with the baselines. When the two preprocesses are combined, all the evaluation metrics increase. It is noted that the results combined with the multiviews are better than those from only one general view.
- (ii) For temporal information, it illustrates the importance of global change in the iris regions for

improving classification performance. For CNN-LSTM, although its testing performance is not promoted much after extracting the iris motion information (image difference), it significantly improves after the video is aligned. When image difference is combined with video alignment, the evaluation metrics further increase, which indicates the negative effect of video jitter on the extraction of iris dynamic features. The temporal information is helpful for the classification.

- (iii) For volume spatial and temporal information, the alignment, ACA extraction, and image difference improve the results, as shown in Table 3. The results in the last line achieve optimal performance by integrating the multiview spatial, temporal, and preprocessing, which is our proposed framework, MT-net.

3.5. Performance on the Two Private AS-OCT Video Datasets.

To prove the superiority of classification based on the AS-OCT videos, we compare our framework with the present algorithm based on single AS-OCT images. We select frames from the beginning and end of our videos taken under a dark environment, which is the same as the datasets of most of the present classification algorithms [4, 11, 12]. For the Casia dataset, the selected images are combined into a training set with a total of 2160 AS-OCT images (1230 angle-closure and 930 normal images) and a testing set with 520 AS-OCT images (250 angle-closure and 270 normal images) with the same distribution as the video dataset. For the Zeiss dataset, the extracted image dataset contains a training set with 3380 AS-OCT images (1360 angle-closure and 2020 normal images) and a testing set with 500 AS-OCT images (200 angle-closure and 300 normal images) with the same distribution as the video dataset.

We use 2D ResNet50, which has the best performance in the baseline experiments, as the comparison algorithm based on the AS-OCT image datasets. The ACA extraction is also combined with 2D ResNet50, and the results are shown in Table 4. To ensure the fairness of comparison, for AS-OCT image datasets, we get final classification results based on each video in the test stage; that is, if the number of correctly classified images accounts for more than 50 % of the total frames of the video, we will give the correct judgment.

As shown in Table 4, for the two datasets, the ACA extraction is helpful for the ACA status classification for all two datasets. But our proposed MT-net based on AS-OCT videos gives the best evaluation metrics. For the Casia dataset, the classification accuracy for our MT-net is 0.866 with a sensitivity of 0.857 and a specificity of 0.875, which achieves superior performance compared with the results of the algorithms based on AS-OCT images with an obvious gap. For the Zeiss dataset, our method based on AS-OCT videos also gets better performance against those based on AS-OCT images with a classification accuracy of 0.833, a sensitivity of 0.860 and a specificity of 0.800. Although the values of sensitivity and specificity are not the highest in Table 4 for the Zeiss dataset, we achieve the highest Kappa

TABLE 3: Classification performance of the angle-closure glaucoma by different module combinations on private Casia video dataset.

AL ¹	ACA	Diff ²	C3D ³	ConvL ⁴	B-Acc	Pre	Recall	F1 score	Sen	Spe	Kappa
			✓		0.692	0.704	0.697	0.695	0.718	0.673	0.370
✓			✓		0.712	0.735	0.703	0.701	0.848	0.576	0.493
	✓		✓		0.719	0.720	0.720	0.720	0.706	0.733	0.518
✓	✓		✓		0.755	0.756	0.753	0.754	0.777	0.732	0.587
✓	✓		✓ ⁵		0.763	0.767	0.767	0.766	0.714	0.813	0.529
				✓	0.781	0.823	0.777	0.789	0.821	0.625	0.545
✓				✓	0.813	0.819	0.816	0.817	0.750	0.860	0.629
		✓		✓	0.607	0.615	0.600	0.596	0.714	0.500	0.210
✓		✓		✓	0.830	0.834	0.833	0.833	0.786	0.875	0.664
			✓	✓	0.777	0.804	0.767	0.763	0.728	0.625	0.542
✓	✓	✓	✓	✓	0.820	0.838	0.817	0.814	0.857	0.780	0.636
✓	✓	✓	✓ ⁵	✓	0.866	0.867	0.867	0.867	0.857	0.875	0.732

¹AL: Alignment; ²Diff: Difference; ³C3D: 3D CNN; ⁴ConvL: CNN-LSTM; ⁵3D CNN (multiview). The bold values indicate the optimal results.

TABLE 4: Comparison of the classification performance on private two AS-OCT video datasets and image datasets.

Casia	B-Acc	Pre	Recall	F1 Score	Sen	Spe	Kappa
ResNet50 (images)	0.767	0.768	0.767	0.766	0.800	0.733	0.533
ACA + ResNet50 (images)	0.774	0.830	0.759	0.748	0.810	0.547	0.530
Our MT-net	0.866	0.867	0.867	0.867	0.857	0.875	0.732
Zeiss	B-Acc	Pre	Recall	F1 score	Sen	Spe	Kappa
ResNet50 (images)	0.750	0.775	0.750	0.744	0.900	0.600	0.500
ACA + ResNet50 (images)	0.795	0.804	0.800	0.798	0.714	0.875	0.594
Our MT-net	0.833	0.840	0.840	0.840	0.860	0.800	0.600

The bold values indicate the optimal results.

value and F1 score, which are used to reflect the trade-offs between sensitivity and specificity.

4. Discussion

In this study, after extracting multiview spatial information and modelling motion, we develop the MT-net to learn to discriminate 3D spatial and temporal features from AS-OCT videos. Our proposed method is shown to be a promising technology for serving clinicians in faithfully identifying angle-closure in AS-OCT videos with a high classification accuracy. The proposed framework opens the door to further enhance the screening ability of angle-closure-related disease from a brand new perspective. More research is needed to explore the employment of deep learning algorithms deployed in diverse population settings, with the use of multiple devices and larger AS-OCT datasets.

The effectiveness of our proposed MT-net is proved in the above experimental parts. The AS-OCT videos can be a more comprehensive means for angle-closure diagnosis. But the study still has two limitations. One limitation of this study is that it assesses two specific Asian populations (Chinese and Singaporeans) due to the high prevalence of primary glaucoma in Asia, so the results may not be applicable to other ethnic groups. But this effect can be mitigated by increasing the diversity of ethnic data. Another potential limitation is that the AS-OCT videos are captured from Casia and Zeiss, the two famous manufacturers in the world. Because of the difference

between the capturing machines, this may adversely affect the quality and performance when our network is applied to videos from other AS-OCT acquisition devices, which did not happen in our present two datasets. If more data can be acquired from other devices in the future, the performance of our model may become more stable and more powerful.

5. Conclusions

We first proposed to detect the ACA status based on light-changing AS-OCT videos in this article. A multiview volume and temporal difference framework (MT-net) is proposed to learn to discriminate spatial and temporal features on the ROIs of AS-OCT videos, which include ACA and iris dynamic changes in the dark-light-dark environment. The ablation experiments prove the effectiveness of our MT-net. The evaluation metrics based on videos are better than those based on 2D AS-OCT images, manifesting that the chamber angle status analysis in a changing environment could improve the ability of angle-closure related disease screening.

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to restrictions in the ethical permit but are partly available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (8210072776), the Science and Technology Innovation Committee of Shenzhen City (JCYJ20200109140820699 and 20200925174052004), Guangdong Basic and Applied Basic Research Foundation (2021A1515012195), Guangdong Provincial Department of Education (2020ZDZX3043), and Guangdong Provincial Key Laboratory (2020B121201001). The authors thank the doctors from Xinhua Hospital for the data collection and analysis. The authors also thank the help of our imed Group for the support.

References

- [1] X. Li, E. Chan, J. Liao, T. Wong, T. Aung, and C. -Y. Cheng, "Number of people with glaucoma in Asia in 2020 and 2040: a hierarchical bayesian meta-analysis," *Investigative Ophthalmology & Visual Science*, vol. 54, p. 2656, 2013.
- [2] Y.-C. Tham, L. Xiang, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.
- [3] D. H. W. Su, D. S. Friedman, J. L. S. See et al., "Degree of angle closure and extent of peripheral anterior synechiae: an anterior segment OCT study," *British Journal of Ophthalmology*, vol. 92, no. 1, pp. 103–107, 2008.
- [4] P. J. Foster, F. T. S. Oen, D. Machin et al., "The prevalence of glaucoma in Chinese residents of SingaporeA cross-sectional population survey of the tanjong pagar district," *Archives of Ophthalmology*, vol. 118, no. 8, pp. 1105–1111, 2000.
- [5] R. Sihota, D. Ghate, S. Mohan, V. Gupta, R. M. Pandey, and T. Dada, "Study of biometric parameters in family members of primary angle closure glaucoma patients," *Eye*, vol. 22, no. 4, pp. 521–527, 2008.
- [6] H. Fu, F. Li, X. Sun et al., "Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography," *Medical Image Analysis*, vol. 66, Article ID 101798, 2020.
- [7] M. E. Nongpiur, L. M. Sakata, D. S. Friedman et al., "Novel association of smaller anterior chamber width with angle closure in Singaporeans," *Ophthalmology*, vol. 117, no. 10, pp. 1967–1973, 2010.
- [8] Y. Xu, L. Jiang, J. Cheng et al., "Automated anterior chamber angle localization and glaucoma type classification in OCT images," in *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7380–7383, Osaka, Japan, July 2013.
- [9] Y. Xu, L. Jiang, W. K. W. Damon et al., "Similarity-weighted linear reconstruction of anterior chamber angles for glaucoma classification," in *Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 693–697, Prague, Czech Republic, April 2016.
- [10] H. Fu, Y. Xu, S. Lin et al., "Segmentation and quantification for angle-closure glaucoma assessment in anterior segment OCT," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1930–1938, 2017.
- [11] H. Fu, Y. Xu, S. Lin et al., "Angle-closure detection in anterior segment OCT based on multilevel deep network," *IEEE Transactions on Cybernetics*, vol. 2019, Article ID 2897162, 2019.
- [12] H. Fu, M. Baskaran, Y. Xu et al., "A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images," *American Journal of Ophthalmology*, vol. 203, no. 37–45, 2019.
- [13] H. A. Quigley, "Angle-closure glaucoma-simpler answers to complex mechanisms: LXVI Edward Jackson memorial lecture," *American Journal of Ophthalmology*, vol. 148, no. 5, pp. 657–669, 2009.
- [14] H. A. Quigley, D. M. Silver, D. S. Friedman et al., "Iris cross-sectional area decreases with pupil dilation and its dynamic behavior is a risk factor in angle closure," *Journal of Glaucoma*, vol. 18, no. 3, pp. 173–179, 2009.
- [15] A. Narayanaswamy, C. Zheng, S. A. Perera et al., "Variations in iris volume with physiologic mydriasis in subtypes of primary angle closure glaucoma," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 1, pp. 708–713, 2013.
- [16] F. Aptel and P. Denis, "Optical coherence tomography quantitative analysis of iris volume changes after pharmacologic mydriasis," *Ophthalmology*, vol. 117, no. 1, pp. 3–10, 2010.
- [17] H. A. Quigley, "The iris is a sponge: a cause of angle closure," *Ophthalmology*, vol. 117, no. 1, pp. 1–2, 2010.
- [18] C. Zheng, C. Y. Cheung, A. Narayanaswamy et al., "Pupil dynamics in Chinese subjects with angle closure," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 250, no. 9, pp. 1353–1359, 2012.
- [19] C. Zheng, C. Y. Cheung, T. Aung et al., "In vivo analysis of vectors involved in pupil constriction in Chinese subjects with angle closure," *Investigative Ophthalmology & Visual Science*, vol. 53, no. 11, pp. 6756–6762, 2012.
- [20] D. Williams, Y. Zheng, P. G. Davey et al., "Reconstruction of 3D surface maps from anterior segment optical coherence tomography images using graph theory and genetic algorithms," *Biomedical Signal Processing and Control*, vol. 25, pp. 91–98, 2016.
- [21] G. Yang, C. V. Stewart, M. Sofka, and C.-L. Tsai, "Registration of challenging image pairs: initialization, estimation, and decision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1973–1989, 2007.
- [22] J. Hao, F. Li, H. Hao et al., "Hybrid variation-aware network for angle-closure assessment in As-Oct," *IEEE Transactions on Medical Imaging*, vol. 41, 2021.
- [23] H. Hao, H. Fu, Y. Xu et al., "Open-appositional-synechial anterior chamber angle classification in as-oct sequences," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Manhattan, NY, USA, October 2020.
- [24] H. Fu, Y. Xu, S. Lin et al., "Multi-context deep network for angle-closure glaucoma screening in anterior segment OCT," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Manhattan, NY, USA, 2018.
- [25] H. Hao, Y. Zhao, Q. Yan et al., "Angle-closure assessment in anterior segment oct images via deep learning," *Medical Image Analysis*, vol. 69, Article ID 101956, 2021.
- [26] Z. Zhou, Md M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

- [27] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1933–1941, Las Vegas, NV, USA, September 2016.
- [28] Y.-H. Joe, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694–4702, Boston, MA, USA, March 2015.
- [29] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4597–4605, Santiago, Chile, December 2015.
- [30] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," in *Computer Vision - ECCV 2016*, Springer, Manhattan, NY, USA, 2016.
- [31] J. Yue-Hei Ng and L. S. Davis, "Temporal difference networks for video action recognition," in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1587–1596, IEEE, Lake Tahoe, NV, USA, March 2018.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [33] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet?" in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, Salt Lake City, UT, USA, June 2018.
- [34] W. Kay, J. Carreira, K. Simonyan et al., "The kinetics human action video dataset," 2017, <https://arxiv.org/abs/1705.06950>.
- [35] H. Kataoka, T. Wakamiya, K. Hara, and Y. Satoh, "Would mega-scale datasets further enhance spatiotemporal 3D CNNs?," 2020, <https://arxiv.org/abs/2004.04968>.
- [36] J. Donahue, L. A. Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, Boston, MA, USA, June 2015.
- [37] S. Karlos, G. Kostopoulos, and S. Kotsiantis, "A soft-voting ensemble based co-training scheme using static selection for binary classification problems," *Algorithms*, vol. 13, no. 1, 2020.
- [38] S. Liu, M. Yu, C. Ye, D. S. C. Lam, and C. Leung, "Anterior chamber angle imaging with swept-source optical coherence tomography: an investigation on variability of angle measurement," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 12, pp. 8598–8603, 2011.
- [39] Y. Zhang, S. Z. Li, L. Li, M. G. He, R. Thomas, and N. L. Wang, "Dynamic iris changes as a risk factor in primary angle closure disease," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 1, pp. 218–226, 2016.

Research Article

Augmentation-Consistent Clustering Network for Diabetic Retinopathy Grading with Fewer Annotations

Guanghua Zhang,¹ Keran Li,² Zhixian Chen,¹ Li Sun,^{3,4} Jianwei zhang,⁵ and Xueping Pan ⁶

¹Department of Intelligence and Automation, Taiyuan University, Taiyuan 030000, China

²The Affiliated Eye Hospital of Nanjing Medical University, Nanjing, China

³Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, Nanjing 210028, China

⁴Jiangsu Provincial Academy of Traditional Chinese Medicine, Nanjing 210028, China

⁵Technical Aspects of Multimodal Systems (TAMS), University of Hamburg, Hamburg 22527, Germany

⁶The First People's Hospital of Huzhou, Huzhou 313000, China

Correspondence should be addressed to Xueping Pan; panxueping1006@139.com

Received 26 December 2021; Revised 17 February 2022; Accepted 22 February 2022; Published 28 March 2022

Academic Editor: Yanwu Xu

Copyright © 2022 Guanghua Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Diabetic retinopathy (DR) is currently one of the severe complications leading to blindness, and computer-aided, diagnosis technology-assisted DR grading has become a popular research trend especially for the development of deep learning methods. However, most deep learning-based DR grading models require a large number of annotations to provide data guidance, and it is laborious for experts to find subtle lesion areas from fundus images, making accurate annotation more expensive than other vision tasks. In contrast, large-scale unlabeled data are easily accessible, becoming a potential solution to reduce the annotating workload in DR grading. Thus, this paper explores the internal correlations from unknown fundus images assisted by limited labeled fundus images to solve the semisupervised DR grading problem and proposes an augmentation-consistent clustering network (ACCN) to address the above-mentioned challenges. Specifically, the augmentation provides an efficient cue for the similarity information of unlabeled fundus images, assisting the supervision from the labeled data. By mining the consistent correlations from augmentation and raw images, the ACCN can discover subtle lesion features by clustering with fewer annotations. Experiments on Messidor and APTOS 2019 datasets show that the ACCN surpasses many state-of-the-art methods in a semisupervised manner.

1. Introduction

Diabetic retinopathy (DR) is one of the most prevalent complications caused by diabetes, which may cause intermittent or even permanent blindness [1–3]. Ophthalmologists often judge the severity of DR based on the features of the disease and the number of lesions, such as observing the characteristics of microaneurysms, hemorrhages, soft exudates, and hard exudates [4, 5]. Recognized by international authorities [6, 7], the severity of DR can be categorized into the following five levels: normal, mild, moderate, severe nonproliferative, and proliferative; these can be summarized into two main categories: normal and abnormal or

nonreferable and referable symptoms [7–9]. If the retina is in the pathological state of DR for a long time, the blood vessels in the eye will eventually become blocked, eventually leading to decreased vision and even blindness. Therefore, it is essential to detect DR early and grade the DR severity in patients because early correct and timely treatment can largely avoid the deterioration of the disease.

In clinical diagnosis, DR detection mainly relies on the careful comparison of colorful fundus images by ophthalmologists. Recently, as the number of diabetic patients has increased yearly, the number of subjects to be tested has become vast, bringing a significant burden on ophthalmologists and DR experts who waste much time observing

fundus images. Therefore, it is necessary to develop computer-aided diagnosing models to efficiently reduce the workload and inspection time for ophthalmologists and experts, achieving real-time DR diagnosis for patients.

To solve the automatic DR grading, early attempts [10–13] are inclined toward exploiting traditional machine learning methods on manual features, limited by specific feature extraction skills and experience. Aiming at this weakness, deep learning has become a popular solution for DR grading with many successful applications [14, 15] because it can automatically learn critical features from fundus images, supervised by accurate annotations. However, these models often depend on a large number of labeled fundus images, whose discriminant information only occurs in subtle blood vessels. The DR grading annotators must master the professional medical knowledge to support them, manually finding key features to decide on actual DR severity, which is a highly time-consuming workload. Thus, high-quality labeled data are scarce, making the supervised DR grading model hard to accomplish.

To save the expensive annotating work in real applications, this paper attempts to solve automatic DR grading in a semisupervised manner to integrate unlabeled data into the training stage because clinical inspection can produce many unlabeled fundus images containing important potential information. Thus, the most crucial task of this paper is to train a robust DR-grading model from massive unlabeled data assisted by fewer annotations, as shown in Figure 1. Extracting more identical information from unlabeled fundus images becomes a top priority, and the data consistency of unlabeled data is vital for feature learning in our work [16–19]. Inspired by previous works, we make more efforts to mine consistent correlations between raw fundus images and their augmentations, which preserve the consistent discriminative information but suffer from image transformations, such as geometric transformation, color space augmentation, random erasing, generative adversarial networks, and neural style transfer.

In this paper, we propose an augmentation-consistent clustering network (ACCN) to alleviate the laborious annotating workload in clinical application, which straightforwardly mines the consistent inner correlations among fundus image augmentations and dynamically conducts weight clustering to utilize the sufficient unlabeled data, absorbing fewer annotated fundus images. As the discriminant cues indicating DR grades are subtle in fundus images, the augmentations from raw images can help the ACCN spread the information from annotated data to unlabeled images. Besides, an online memory unit is introduced to dynamically update the clustering centroids, guaranteeing the global consistency between labeled and unlabeled fundus images in exploring critical information.

The main contributions of this article are summarized as follows:

- (1) We propose a brand-new, highly robust semi-supervised framework (ACCN) to solve the DR grading problem, inspired by the consistent discriminative correlations between labeled and unlabeled fundus images with different augmentations.

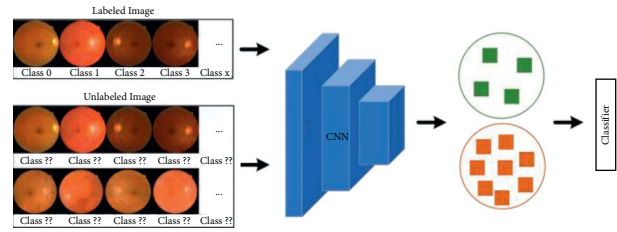


FIGURE 1: Analysis diagram of our semisupervised DR-grading solution.

- (2) We design a reasonable weight-clustering algorithm that benefits from an online memory unit to dynamically update the clustering centroids with global consistency, generating high-quality pseudolabels for unlabeled images and integrating annotated fundus images to explore discriminative information for DR grading.
- (3) We conducted experiments on the public data sets Messidor and APTOS 2019, and the results show that the ACCN is superior to many state-of-the-art DR grading methods.

2. Related Work

This section summarizes recent works on the diabetic retinopathy grading problem and introduces the successful computer-aided diagnosing applications of semisupervised learning.

2.1. Diabetic Retinopathy Grading. With the continuous development of deep learning, its application to retinal images has also achieved great success. Recently, some new research has been proposed [20–23]. For example, Sambyal et al. [20] proposed an aggregated residual transformation-based model for automatic multistage classification of diabetic retinopathy. Bhardwaj et al. [21] developed a hierarchical severity-level grading system to detect and classify DR ailments. Bodapati et al. [22] presented a hybrid deep neural network architecture with a gated attention mechanism for automated diagnosis of diabetic retinopathy. Math et al. [23] designed a segment-based learning approach for diabetic retinopathy detection, which mutually learns classifiers and features from the data and achieves significant development in diabetic retinopathy recognition.

However, the methods mentioned above require a large amount of labeling information. Medical labeling is well known to be expensive and time-consuming, which many institutions cannot afford. This significantly constrains the transferability of these developed DR grading systems.

2.2. Semisupervised Learning in Medical Image Classification. In recent years, medical imaging technology has been fully developed for clinical applications [24–26]. In medical image analysis, annotation is often difficult to obtain because it is expensive and labor-intensive. Semisupervised learning

to relieve the pressure of labeling has provided great help to a certain extent. In recent years, some studies have successfully applied the semisupervised framework to medical image analysis [27–31]. Wang et al. [27] incorporated virtual adversarial training on both labeled and unlabeled data into the course of training, self-training, and consistency regularization to effectively exploit useful information from unlabeled data. Calderon et al. [28] explored the impact of using unlabeled data through the implementation of a recent approach known as MixMatch for mammogram images. Pang et al. [29] developed a radionics model based on a semisupervised GAN method to perform data augmentation in breast ultrasound images. Liu et al. [30] proposed a self-supervised mean teacher for chest X-ray classification that combines self-supervised mean-teacher pretraining with semisupervised fine-tuning. Bakalo et al. [31] designed a deep learning architecture for multiclass classification and localization of abnormalities in medical imaging illustrated through experiments on mammograms.

In this paper, we propose a novel augmentation-consistent clustering network (ACCN) for semisupervised diabetic retinopathy grading on fundus images, exploring the discriminative information learned from plentiful unlabeled data and fewer annotated fundus images.

3. Method

Aiming to explore the discriminant information from massive unlabeled fundus images, we design a novel semisupervised DR grading approach, the augmentation-consistent clustering network (ACCN), to assist the supervised model trained by fewer annotated data. The ACCN utilizes consistent learning and weight clustering on easily accessible unlabeled data with the help of fewer annotations to achieve the semisupervised diabetic retinopathy grading task. In detail, the ACCN first considers the category correlations among unlabeled fundus images, maintaining consistency with different augmentations. Then the trained model from annotated fundus images is utilized as the baseline network, and the ACCN deploys a clustering algorithm to weight their CNN features to calculate the pseudolabels for unlabeled images. Finally, we utilize the real annotations and pseudoannotations to train the network parameters. The whole workflow for the ACCN is illustrated in Figure 2, and the symbols are summarized in Table 1.

3.1. Augmentation-Consistent Learning. In semisupervised DR grading work, the most crucial task is the exploration of unlabeled retinal images. At the same time, the augmentation in deep learning is a popular and easily conducted process to produce various transformations for unlabeled raw fundus images, containing consistent identity information but close to realistic scenarios [19, 32]. Thus, the ACCN first conducts reasonable augmentations for raw retinal images to generate diverse data with the same category and then employs a convolutional neural network to learn appearance feature representations for the augmented images.

In the ACCN, we adopt augmentation anchoring technology [19, 32] that utilizes the pseudolabels that come from weakly augmented samples as the “anchor” and align the strongly augmented samples to the “anchor.” Notably, the weak augmentation A_{weak} in our method contains a random cropping followed by a random horizontal flip, and the strong augmentation sequence $A_{\text{strong}} = \{A_{\text{strong}}^1, A_{\text{strong}}^2, \dots, A_{\text{strong}}^k\}$ is achieved by RandAugment and a fixed augmentation strategy that contains a sequence of image transformations.

Because the labeled images contain sufficient grading information to find samples in the same category, with no need to generate much more augmented images, we only process the annotated retinal image x_i^l by weak augmentation to produce an “anchor” \tilde{x}_i^l ,

$$\tilde{x}_i^l = A_{\text{weak}}(x_i^l), \quad (1)$$

while the unlabeled fundus image x_u^l should be transformed into an image sequence by strong augmentations to produce more strongly augmented samples to form sufficient training data in the same category. Thus, we utilize the strong augmentation series to generate their augmentations:

$$\tilde{X}_j^u = \{A_{\text{strong}}^k(x_j^u)\}_{k=1}^K, \quad (2)$$

where \tilde{x}^u denotes K strongly augmented unlabeled fundus images from A_{strong} .

Through the above-mentioned augmentations, we can obtain the weak augmented annotated image \tilde{x}_i^l and strong augmented unlabeled fundus images \tilde{X}_j^u , which are intended to supervise the model training to analyze the images from multiple angles and extract more critical features.

As for feature learning, the ACCN employs the ResNet-50 architecture [33] as the feature extractor for fundus images and their augmentations due to its excellent performance in medical imaging. Particularly, the feature extractor is defined by G for annotated and unlabeled retinal images, and the feature vector $G(\cdot)$ is transformed into a probability vector by a classifier F . Taking a retinal image x as an example, its prediction can be mathematically represented by

$$P(x) = F(G(x)). \quad (3)$$

Essentially, the weak augmented images enlarge the scale of labeled data to compose a labeled set $X^l = \{x_1^l, x_2^l, \dots, x_{N_l}^l\} \cup \{\tilde{x}_1^l, \tilde{x}_2^l, \dots, \tilde{x}_{N_l}^l\}$, training the feature extractor and classifier by labeled cross-entropy (lce) loss:

$$L_{lce} = - \sum_{x_i \in X^l} y_i^l \log F(G(x_i; W_G); W_F), \quad (4)$$

where W_G and W_F represent the network parameters of the feature extractor and the classifier, respectively.

Similarly, the strong augmentations for unlabeled images produce the transformed samples with the same category as raw images. Thus, we also introduce an augmentation-consistent (ac) loss to enforce that the

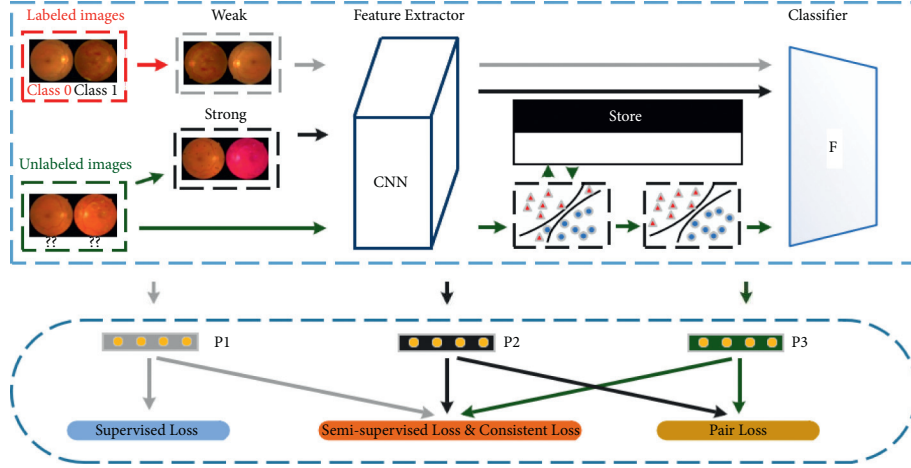


FIGURE 2: Scheme of the augmentation-consistent clustering network. First, different augmentations for annotated and unlabeled fundus images are generated in a weak and a strong manner, respectively, and consistent feature learning is conducted to train a robust feature extractor. Then, the unlabeled feature representations are fed into a weight-clustering unit to assign pseudolabels with dynamically updating memory in model training. Finally, the pseudolabels and corresponding unlabeled retinal images are utilized to optimize the whole network for solving the DR grading task with fewer annotations.

TABLE 1: The symbol summary.

Symbol	Meaning
x_i^l	The i -th annotated retinal image
x_j^u	The j -th unlabeled retinal image
A_{weak}	The weak augmentation
A_{strong}	The collection of strong augmentations
\tilde{x}_i^l	The weak augmented image for x_i^l
\tilde{X}_j^u	The collection of strong augmentations for x_j^u
G	The feature extractor
F	The classifier
X^l	The labeled raw images and their augmentations
X^u	The set of unlabeled raw images
X_u	The unlabeled raw images and their augmentations
c_k	The local centroid for k -th class
y_j^u	The generated pseudolabel
M^k	The global centroid

classifier predicts the consistent probability vectors for the correlated augmentation and raw fundus images:

$$L_{ac} = \sum_{x_j \in X^u, \tilde{x}_j \in \tilde{X}_j^u} \|P(x_j) - P(\tilde{x}_j)\|, \quad (5)$$

where $X^u = \{x_1^u, x_2^u, \dots, x_{N_u}^u\}$ denotes the set of unlabeled retinal images.

Benefiting from the labeled cross-entropy loss L_{lce} and augmentation-consistent loss L_{ac} , the feature extractor G and classifier F can learn a lot from the discriminative consistency between augmentations and raw images, especially from the unlabeled retinal images. Hence, the backbone network in the ACCN possesses quite an inferential capability for unknown retinal images.

3.2. Weight Clustering Unit. Even though the consistency information has been extracted from unlabeled images, accurate diabetic retinopathy grading cues are implied in the

annotations. In recent years, pseudolabels have become an essential research topic in unlabeled image analysis [34–36]. However, simply introducing a pretrained fully connected classifier F by the limited labeled data does not contain robust identification ability; thus, it cannot effectively extract the internal association between the unlabeled feature representations because the augmentation consistent loss is short of the annotations. To address this weakness, the ACCN designs a weight clustering unit to mine the mutual relationships between unknown samples and their pseudolabels.

Specifically, we calculate the estimated centroid c_k for each class according to the primary outputs from the trained classifier F :

$$c_k = \frac{\sum_{x_i^u \in X^u} \delta_k(F(G(x_i^u)))G(x_i^u)}{\sum_{x_i^u \in X^u} \delta_k(F(G(x_i^u)))}, \quad (6)$$

where δ_k corresponds to the k -th element output by softmax. Then, we calculate the distance between each unlabeled feature and each centroid to generate pseudolabels according to the nearest neighbor principle:

$$y_j^u = \arg \min_k d(G(x_j^u), c_k), \quad (7)$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance measure. In this way, we induce the prediction model focus on some samples around the decision boundary and explore more discriminative information by the weight clustering unit.

It should be noted that weight clustering is supported by iterative epochs to update the centroids. This means that multiple clustering is required in each batch, producing different local centroids. This may cause much more centroid deviation with wrong pseudolabeled annotations. To avoid this problem in our ACCN model, we design a dynamic centroid memory $\{M_k\}_{k=1}^{N_c}$ to store the temporary global centroids in each batch, where M_k is the k -th class center and N_c represents the number of image categories.

Besides, the updated strategy for the global centroid is as follows:

$$M_k = (1 - \eta_{t_k})M_k + \eta_{t_k}c_k, \quad (8)$$

where $\eta_{t_k} = e^{-t_k}$ represents the updating rate of grade k and t_k denotes the number of categories k that appeared in the previous batch.

Finally, we minimize the distance between the local and global centroids in each batch by a global consistent (gc) loss:

$$L_{gc} = \frac{1}{N_c} \sum_{k=1}^{N_c} \|M_k - c_k\|_2. \quad (9)$$

By advancing the above-mentioned relationship, we can alleviate the problem that wrong pseudolabeled samples cannot be correctly distinguished, which also improves the effect of diabetic retinopathy grading.

By the weight clustering unit, we can obtain reasonable pseudoannotation for the unlabeled retinal images. This supports us to conduct the annotation level supervised training from unlabeled fundus data and their strong augmentations $X^u = \{x_1^u, x_2^u, \dots, x_{N_u}^u\} \cup \{\bar{X}_1^u, \bar{X}_2^u, \dots, \bar{X}_{N_u}^u\}$ corresponding to their pseudolabels $\{y_1^u, y_2^u, \dots, y_{N_u}^u\}$, according to a pseudo-cross-entropy (pce) loss:

$$L_{pce} = - \sum_{x_j \in X^u} y_j^u \log F(G(x_j; W_G); W_F). \quad (10)$$

3.3. Final Loss for ACCN Model. As described above, our semisupervised diabetic retinopathy grading approach ACCN is composed of two crucial modules, namely, an augmentation-consistent learning and a weight clustering unit, attached with labeled cross-entropy loss L_{lce} , augmentation-consistent loss L_{ac} , global-consistent loss L_{gc} , and pseudo-cross-entropy loss L_{pce} .

To update all trainable parameters in the ACCN, we integrate the final loss into the network with balance parameters:

$$\min_{W_G, W_F} L = L_{lce} + \gamma_1 L_{ac} + \gamma_2 L_{gc} + \gamma_3 L_{pce}, \quad (11)$$

where γ_1 , γ_2 , and γ_3 are parameters to balance different loss functions.

4. Experiments

4.1. Database Description. In this section, we evaluate the proposed augmentation-consistent clustering network by training on the publicly available dataset Messidor [37]. In detail, Messidor [37] contains approximately 1200 digital fundus images obtained by using a Topcon TRC NW6 nonmydriatic camera. The sizes of fundus images are 440×960 , 2240×1488 , or 2304×1536 in, and ophthalmologists labeled each image. According to the DR severity, Messidor classifies the fundus images into one of the four grades, namely, normal and no lesion ($R0$), mild ($R1$), severe nonproliferative ($R2$), and proliferative ($R3$) retinal images.

The data distribution of Messidor in each grade is described in Table 2, and the popular DR grading task of normal/abnormal classification is summarized in Table 3. The distribution shows that the common challenging problem is the data imbalance, which may influence the model training.

4.2. Experimental Settings. This paper conducts normal/abnormal DR grading experiments, dividing the dataset into 600 training images and 600 testing samples. In detail, labeled retinal images in the training data contain 400 labeled fundus images, including 200 positive cases and 200 negative images. As for the unlabeled training data, they contain 46 positive cases and 154 negative images. In addition, we chose the left 600 retinal images as testing data, which contain 300 positive and 300 negative cases. The entire experimental process is completed using the PyTorch framework under GeForce 2080TI GPU. Precisely, each retinal image is adjusted to $512 * 512$ pixels before inputting it to the network, and the batch size is set to 8. Besides, we use ResNet-50 as the backbone, and the classifier is composed of linear layers. For parameter settings, the learning rate is set to 0.001, and balance parameters [λ_1 , λ_2 , and λ_3] are [0.6, 0.3, and 0.8, respectively] to perform the best DR grading results. In addition, the training process spends around 2.5 minutes per epoch, and the evaluation for testing images takes 5 milliseconds per fundus image.

To measure the experimental performance, we adopt the popular indicators to compare and evaluate our models: specificity (SPE), sensitivity (SEN), accuracy (ACC), and the area under the ROC curve (AUC).

4.3. Comparison with Other Methods

4.3.1. Performance on Messidor. In order to demonstrate the performance of the ACCN on DR grading, we compare with different baseline methods for the normal/abnormal DR grading task. As to the compared methods, we choose the manual grading results from two experts [38] and introduce two experimental methods used in [39], which emphasize the role of multiple filter sizes in learning fine-grained discriminant features and proposes two deep convolutional neural networks, combining kernels with a multiple loss network and a V_{gg} network. The normal/abnormal fundus image classification results on Messidor are reported in Table 4, and our ACCN framework achieves the highest accuracy of 89.8%, sensitivity of 93.0%, specificity of 86.7%, and AUC of 93.6%, outperforming the supervised DR grading model and experts. What needs to be emphasized is that our ACCN model only utilizes 400 annotated retinal images and other training data is unlabeled while the compared models require fully annotated retinal images and experts require long-term professional training. Therefore, the excellent performance of our ACCN in a semisupervised manner proves that it can save us from depending on expensive annotating networks in significant applications for DR grading.

Besides, we choose two existing semisupervised medical image classification methods [30, 41] to compare with our

TABLE 2: The class distribution of datasets.

Label	Messidor
DR 0	546
DR 1	153
DR 2	247
DR 3	254

TABLE 3: The popular classification task on DR grades.

Label	Description
DR grading	DR 0/DR 1/DR 2/DR 3
Normal/abnormal DR	DR 0/DR 1, DR 2, DR 3

TABLE 4: Compared performance on Messidor.

Methods	Accuracy	Sensitivity	Specificity	AUC
Expert A [38]	87.8	—	—	92.2
Expert B [38]	76.4	—	—	86.5
Holly et al. [39]	87.1	88.2	85.7	87.0
Holly et al. [39]	85.8	91.6	80.3	86.2
Odena et al. [40]	94.7	95.4	95.1	96.7
S ² MTS ² [30]	86.7	88.7	84.8	86.3
SRC-MT [41]	85.8	86.4	85.2	84.8
ACCN	89.8	93.0	86.7	96.0

ACCN model. S²MTS² [30] combines self-supervised mean-teacher pretraining with a semisupervised fine-tuning method to solve the multilabel chest X-ray classification; SRC-MT [41] proposes a sample relation data consistency paradigm to effectively extract unlabeled data by modeling the relationship information among different medical image samples. To compare the ACCN with them, we implement their public available code on the Messidor dataset with the same settings. The results are summarized in Table 4, proving that our ACCN approach is superior to those semisupervised medical image classification methods, with considerable improvements in each metric. Although our method outperforms some supervised methods, there is still a gap with advanced supervised methods, and the ACCN still has the potential to be explored to reach the supervised performance.

4.4. Visual Analysis for ACCN. This article outlines two popular visualizations for the ACCN to make it generally available for the diabetic retinopathy grading task. First, the ROC curve is shown in Figure 3, and our approach achieves an AUC of 0.96 on the Messidor dataset. Besides, we utilize 600 testing fundus images and illustrate the classification results in the confusion matrix (Figure 4). The confusion matrix can quickly visualize the proportion of various misclassified categories into other classes. From the results, the ACCN model correctly classifies the 279 abnormal and 261 normal fundus images, with 89.9% accuracy. Summarizing the above-mentioned visualization results, we can see that our ACCN model effectively utilizes a large amount of unlabeled data with fewer annotations to solve the semisupervised DR grading task well.

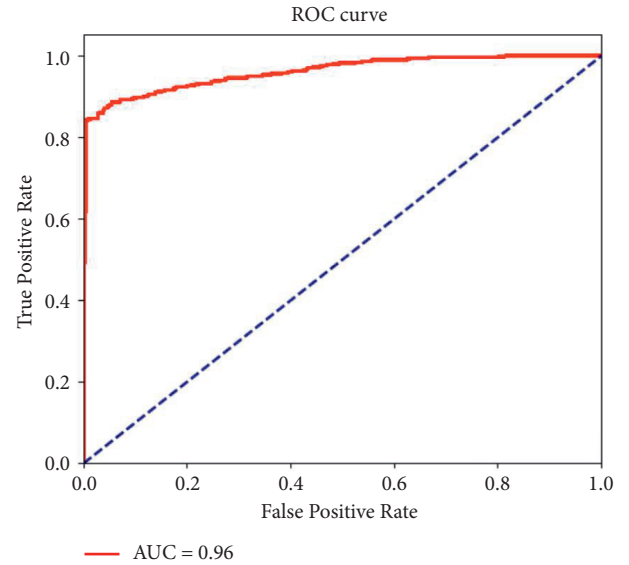


FIGURE 3: ROC curve of the proposed ACCN model for normal/abnormal DR grading on the Messidor dataset.

		Prediction	
		Normal	Abnormal
Ground Truth	Normal	261	39
	Abnormal	21	279

FIGURE 4: Normal/abnormal DR classification on the Messidor dataset.

At the same time, we calculate the loss reduction during model training, illustrated in Figure 5. The overall loss reveals a downward trend, and the regeneration of pseudo-labels causes the ups and downs in the first half by clustering within the batch. After adding the global-consistent loss, the clustering centroids are dynamically updated more reasonably, with stable loss convergence. This demonstrates that our ACCN can rapidly train a semisupervised DR grading model and the global-consistent loss significantly improves the convergence.

4.5. Performance on Other DR Grading Datasets. This article also chooses another publicly available DR grading dataset,

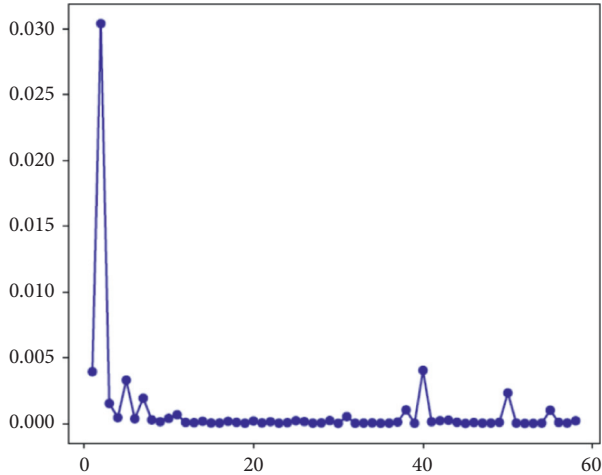


FIGURE 5: Loss curve of the ACCN for model training on the Messidor dataset.

APTOS 2019, in the normal/abnormal DR experiments to provide the transferability of the proposed ACCN approach. APTOS 2019 [42] was proposed in the APTOS 2019 diabetic retinopathy classification contest, which was organized by the Asia Pacific Tele-Ophthalmology Society. It comprises 3662 retinal images from fundus photography with available annotations captured from multiclinics with different imaging conditions at Aravind Eye Hospital in India. Concretely, this dataset contains five classes for training the ACCN, and the data are highly imbalanced, as summarized in Table 5. Compared to Messidor, APTOS 2019 is more challenging because it contains five grades on DR and it can prove the effectiveness of our ACCN model more sufficiently on normal and abnormal DR classification, and the detailed division of different DR grades can be found in Table 4.

From Table 6, it can be found that the ACCN has reached a high accuracy of 93.4%, sensitivity of 91.0%, specificity of 95.7, and AUC of 0.984. These results mean that the ACCN can effectively extract the internal connections among unlabeled retinal images in different datasets and it can successfully solve the DR grading problem with fewer annotations when transferred to other application scenarios.

5. Further Analysis

This section further discusses the impacts of major components and parameters on the ACCN approach to the semisupervised DR grading task, including the labeled data, augmentation-consistent learning, and the weight clustering unit.

5.1. The Impact of Labeled Fundus Images. This paper attempts to solve the DR-grading task with fewer annotations. Thus there are very few high-quality samples with accurate labels for DR diagnosis. To measure the impacts of labeled data, we use accuracy to test how the number of labeled retinal images influences the ACCN performance on the

TABLE 5: The class distribution of APTOS 2019.

Label	APTOS	Division
DR 0	1805	Normal
DR 1	370	Abnormal
DR 2	999	Abnormal
DR 3	193	Abnormal
DR 4	295	Abnormal

TABLE 6: Experimental results on APTOS 2019.

Methods	Accuracy	Sensitivity	Specificity	AUC
ACCN	93.4	91.0	95.7	98.4

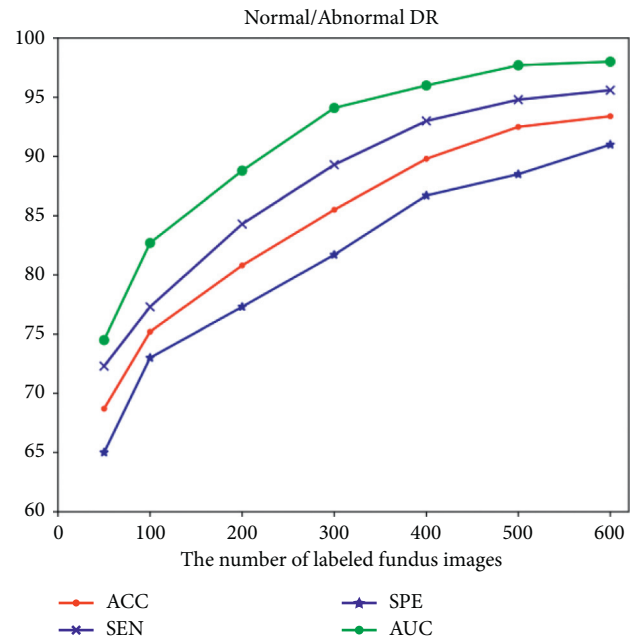


FIGURE 6: DR classification performance with different numbers of labeled data.

Messidor dataset. From the results in Figure 6, it can be observed that the DR grading accuracy rapidly increases from 68.7% to 75.2% as the number of labeled fundus images increases from 50 to 100 and it mildly increases from 75.2% to 89.8% when the number of labeled data is between 100 and 400. Finally, the ACCN model achieves an accuracy of 93.4% when it is fully supervised.

The above-mentioned experimental results show that the proposed semisupervised model can work well using a relatively small number of labeled samples, with fewer annotating costs than existing supervised DR grading models. However, using the proposed ACCN approach still requires a certain amount of labeled samples to obtain a higher classification accuracy. A similar trend and conclusion can also be observed from sensitivity, specificity, and AUC.

TABLE 7: The contributions of the major steps in ACCN (%).

Target	Accuracy	Sensitivity	Specificity	AUC
ACL	+13.5	+14.7	+12.4	+14.6
WLU	+8.1	+9.3	+7	+9.8

5.2. The Impact of Augmentation-Consistent Learning. The first dominating method in the ACCN is the augmentation-consistent learning module, which generates weak and strong augmentations for annotated and unlabeled training images, respectively, and conducts consistent feature learning for the raw images and their augmentations. To weigh the impact of this module, we only employ raw images to conduct the weight clustering network and assign pseudolabels. The results are reported in Table 7 (ACL). Concretely, the ACL module improves the DR grading performance with an accuracy of +13.5%, sensitivity of +14.7%, specificity of +12.4%, and AUC of +14.6%. This further certifies that the novelties of our proposed augmentation-consistent learning mechanism are beneficial to the semisupervised DR grading task.

5.3. The Impact of Weight Clustering. We then analyze the influence of the weight clustering module. We remove the entire clustering module and directly use the prediction vector of the high-confidence sample after the softmax output as the pseudolabel for training. The effect of normal/abnormal DR classification on the Messidor dataset is that the accuracy has dropped by 8.1%, which demonstrates that the ACCN employing a weight clustering unit to explore the internal relationship between unknown samples is effective in semisupervised DR grading task. Compared to the supervised models in the study by Holly et al. [39], our model achieves a competitive AUC of 86.2% when removing the WLU. It benefits from the proposed augmentation-consistent learning module and further proves the effectiveness of our semisupervised learning approach.

5.4. The Impact of Positive Cases in Unlabeled Data. The positive proportion of unlabeled data is an important factor affecting the final performance for the semisupervised diabetic retinopathy grading problem. We finally discuss the influence of the positive proportion of unlabeled training data by changing the proportion of positive cases in unlabeled data. The results on the Messidor dataset are summarized in Figure 7, revealing that the accuracy of performance decreases with increasing positive proportion in unlabeled training. This demonstrates that the positive cases in labeled training data provide more discriminative information than the ones in unlabeled data. Thus, the balanced distribution of negative and positive cases both in labeled and unlabeled data is important for the semisupervised diabetic retinopathy grading task. In addition, under the premise that the number of labeled samples remains unchanged, we record experimental results employing different proportions of positive samples (unlabeled). The result is shown in Figure 8.

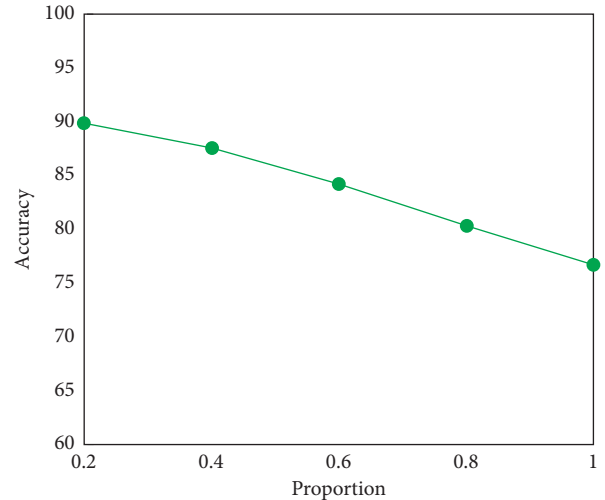


FIGURE 7: The accuracy results of different positive proportions in unlabeled training data.

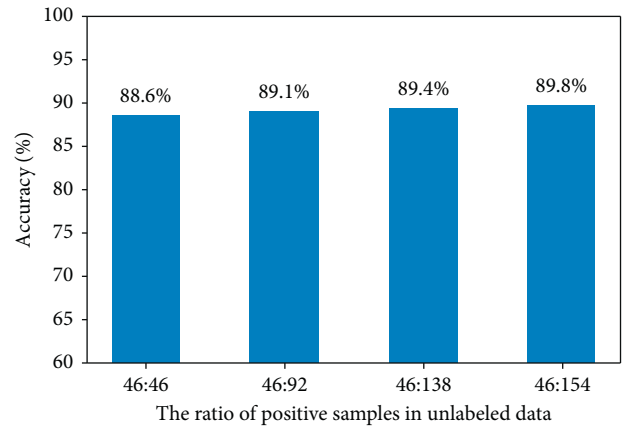


FIGURE 8: The accuracy results with different ratios of positive samples in unlabeled data.

6. Discussion and Conclusion

For the real application of diabetic retinopathy grading, the lack of labeled data is the main challenge that limits the application of deep learning. This is probably due to the following reasons. First, the lesion indicating DR is always subtle in digital fundus images, so labeling retinal images require expertise in long-term training, and hiring experts to annotate is very expensive and time-consuming. Second, medical data, especially images for human diseases, become difficult to collect due to rigorous privacy issues. Finally, the diseases that require the aid of computer vision are often complex, and the model training must use sufficient data, making the fundus image annotation more complicated.

To address the above-mentioned challenges, we propose an augmentation-consistent clustering network (ACCN) approach for semisupervised diabetic retinopathy grading, which can mine internal correlations among unknown samples assisted by fewer annotations. The proposed model can compensate for the lack of labeled data in the following

ways. (1) The augmentation-consistent learning generates weak and strong augmentations for annotated and unlabeled fundus images and provides inherent consistent information by labeled cross-entropy and augmentation-consistent losses. (2) A weight clustering unit is designed to calculate the pseudolabels for unknown retinal images with a dynamically clustering algorithm, which utilizes weight centroids to cluster in a global-consistent manner. (3) The DR classification model is further trained by combining annotated and pseudolabeled retinal images to achieve the semisupervised diabetic retinopathy grading task. Adequate experiments on the Messidor dataset prove that the ACCN can perform effective DR classification with limited labeled data, and the extensive experiments on APTOS 2019 demonstrate the scalability of our ACCN network to different domains.

In future, we will work on the unsupervised learning approach to conduct fundus image classification without any annotations. Besides, we will focus on diabetic retinopathy grading in multiple stages to provide a more accurate diagnosis for ophthalmologists.

Data Availability

The datasets used and/or analyzed during the present study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence our work; there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, the entitled manuscript.

Authors' Contributions

Guanghua Zhang, Keran Li, and Zhixian Chen contributed equally to this work.

Acknowledgments

This work was supported by the Research Funds of the Shanxi Transformation and Comprehensive Reform Demonstration Zone (Grant no. 2018KJXC04), the Fund for Shanxi "1331 Project," and the Key Research and Development Program of Shanxi Province (No. 201903D311009). The work was also partially sponsored by the Research Foundation of Education Bureau of Shanxi Province (Grant No. HLW-20132), the Scientific Innovation Plan of Universities in Shanxi Province (Grant no. 2021L575), and, the Shanxi Scholarship Council of China (Grant No. 2020-149). The work was also sponsored by the Zhejiang Medical and Health Research Project (2020PY027) and the Huzhou Science and Technology Planning Program (2019GY13).

References

- [1] N. H. Cho, J. E. Shaw, S. Karuranga et al., "Idf diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 2018.
- [2] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
- [3] H. Pratt, F. Coenen, D. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.
- [4] X. Li, X. Hu, and L. Yu, "Canet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1483–1493, 2019.
- [5] Y. Zhou, X. He, and L. Huang, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2079–2088, Long Beach, CA, USA, June 2019.
- [6] C. P. Wilkinson, F. L. Ferris III, R. E. Klein et al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.
- [7] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [8] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: deep mining lesions for diabetic retinopathy detection, medical image computing and computer assisted intervention – MICCAI 2017," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 267–275, Springer, Cambridge, UK, September 2017.
- [9] Y. Yang, F. Shang, and B. Wu, "Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image," *IEEE Transactions on Cybernetics*, vol. 2021, Article ID 3062638, 2021.
- [10] A. Sopharak, M. N. Dailey, B. Uyyanonvara et al., "Machine learning approach to automatic exudate detection in retinal images from diabetic patients," *Journal of Modern Optics*, vol. 57, no. 2, pp. 124–135, 2010.
- [11] R. Priya and P. Aruna, "Diagnosis of diabetic retinopathy using machine learning techniques," *ICTACT Journal on soft computing*, vol. 3, no. 4, pp. 563–575, 2013.
- [12] J. Krause, V. Gulshan, E. Rahimy et al., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.
- [13] M. Zhang, W. Meng, T. Davies, Y. Zhang, and S. Q. Xie, "A robot-driven computational model for estimating passive ankle torque with subject-specific adaptation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 814–821, 2015.
- [14] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.
- [15] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

- [16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International conference on machine learning*, pp. 1597–1607, PMLR, Las Vegas, Nevada, June 2020.
- [17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, Seattle, WA, USA, June 2020.
- [18] L. Zhang and G.-J. Qi, "Wcp: worst-case perturbations for semi-supervised deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3912–3921, Seattle, WA, USA, June 2020.
- [19] K. Sohn, D. Berthelot, C.-L. Li et al., "Fixmatch: simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.
- [20] N. Sambyal, P. Saini, R. Syal, and V. Gupta, "Aggregated residual transformation network for multistage classification in diabetic retinopathy," *International Journal of Imaging Systems and Technology*, vol. 31, no. 2, pp. 741–752, 2021.
- [21] C. Bhardwaj, S. Jain, and M. Sood, "Hierarchical severity grade classification of non-proliferative diabetic retinopathy," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2649–2670, 2021.
- [22] J. D. Bodapati, N. Shaik, and V. Naralasetti, "Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 1–15, 2021.
- [23] L. Math and R. Fatima, "Adaptive machine learning classification for diabetic retinopathy," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5173–5186, 2021.
- [24] Z. Jiang, Z. Li, M. Grimm et al., "Autonomous robotic screening of tubular structures based only on real-time ultrasound imaging feedback," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 7, 2021.
- [25] M. Thies and M. L. Oelze, "Combined therapy planning, real-time monitoring, and low intensity focused ultrasound treatment using a diagnostic imaging array," *IEEE Transactions on Medical Imaging*, vol. 2022, Article ID 3140176, 2022.
- [26] Z. Jiang, M. Grimm, M. Zhou, Y. Hu, J. Esteban, and N. Navab, "Automatic force-based probe positioning for precise robotic ultrasound acquisition," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 11, 2020.
- [27] X. Wang, H. Chen, and H. Xiang, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical Image Analysis*, vol. 70, Article ID 102010, 2021.
- [28] S. Calderón-Ramírez, D. Murillo-Hernández, K. Rojas-Salazar et al., "Improving uncertainty estimations for mammogram classification using semi-supervised learning," in *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Shenzhen, China, July 2021.
- [29] T. Pang, W. L. Ng, and C. Chan, "Semi-supervised ganbased radiomics model for data augmentation in breast ultrasound mass classification," *Computer Methods and Programs in Biomedicine*, vol. 203, Article ID 106018, 2021.
- [30] F. Liu, Y. Tian, F. R. Cordeiro, V. Belagiannis, I. Reid, and G. Carneiro, "Self-supervised mean teacher for semi-supervised chest x-ray classification," in *International Workshop on Machine Learning in Medical Imaging*, pp. 426–436, Springer, Cham, Switzerland, 2021.
- [31] B. Ran, J. Goldberger, and R. Ben-Ari, "Weakly and semi supervised detection in medical imaging via deep dual branch net," *Neurocomputing*, vol. 421, pp. 15–25, 2021.
- [32] D. Berthelot, N. Carlini, and D. Cubuk, "Remixmatch: semi-supervised learning with distribution alignment and augmentation anchoring," in *Proceedings of the International Conference on Learning Representation*, Addis Ababa, Ethiopia, April 2020.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, "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.
- [34] H. Feng, M. Chen, J. Hu, D. Shen, H. Liu, and D. Cai, "Complementary pseudo labels for unsupervised domain adaptation on person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2898–2907, 2021.
- [35] Z. Hu, Z. Yang, X. Hu, and N. Ram, "Simple: similar pseudo label exploitation for semi-supervised classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15099–15108, Nashville, TN, USA, June 2021.
- [36] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: revisiting pseudo-labeling for semi-supervised learning," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event*, pp. 6912–6920, AAAI Press, February 2021, <https://ojs.aaai.org/index.php/AAAI/article/view/16852>.
- [37] E. Decencière, X. Zhang, G. Cazuguel et al., "Feedback on a publicly distributed image database: the Messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, 2014.
- [38] CI. Sánchez, M. Niemeijer, AV. Dumitrescu, MS. Suttorp-Schulten, MD. Abramoff, and B. van Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Investigative Ophthalmology & Visual Science*, vol. 52, no. 7, pp. 4866–4871, 2011.
- [39] H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)*, pp. 209–215, IEEE, San Jose, CA, USA, December 2016.
- [40] A. Odena, "Semi-supervised learning with generative adversarial networks," 2016, <https://arxiv.org/abs/1606.01583>.
- [41] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3429–3440, 2020.
- [42] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, and A. Elbaz, "Deep learning based method for computer aided diagnosis of diabetic retinopathy," in *Proceedings of the 2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–4, IEEE, Abu Dhabi, UAE, December 2019.