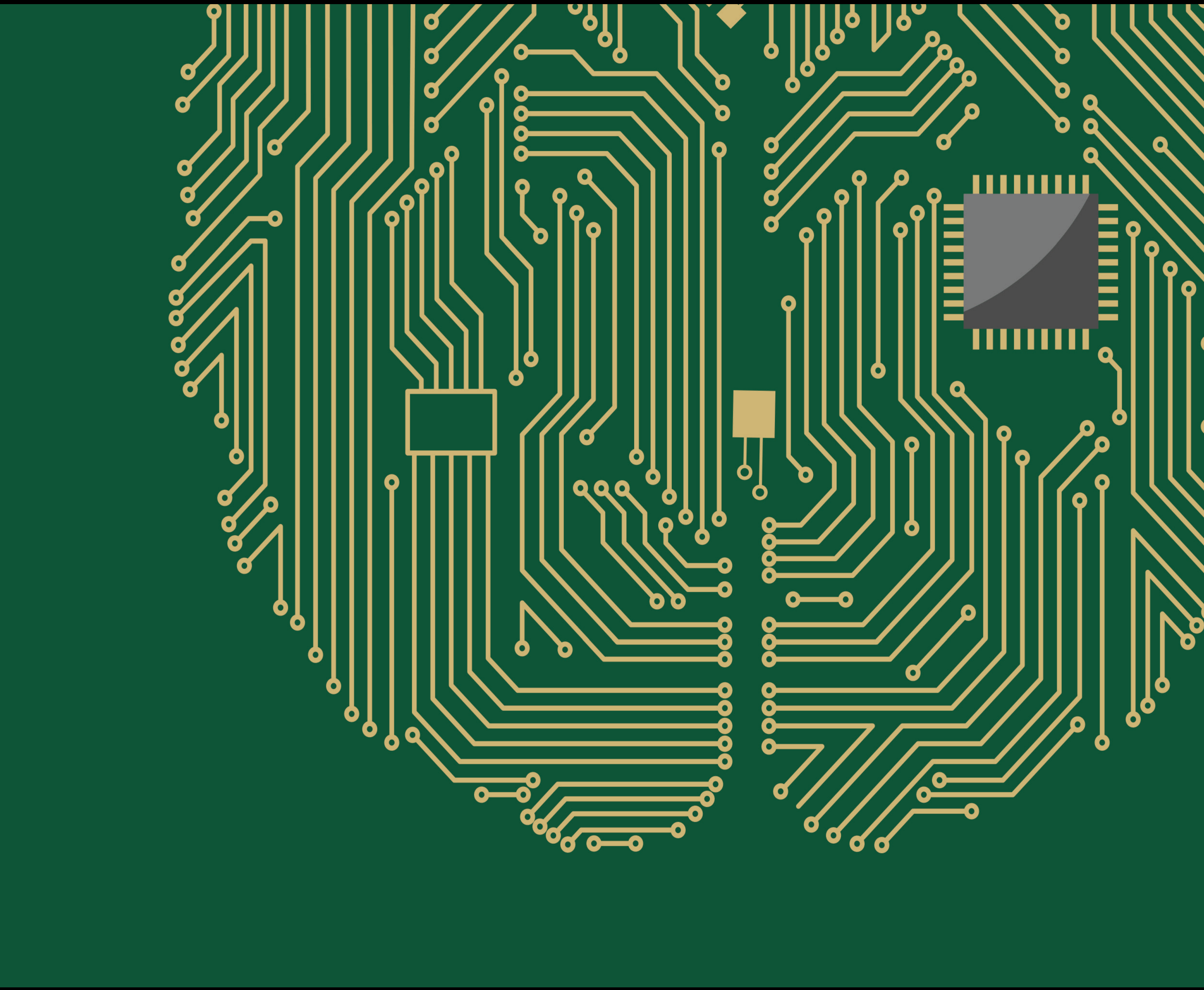
Computational Intelligence and Neuroscience

Advances in the Application of Human Activity Recognition

Lead Guest Editor: David Gil
Guest Editors: Magnus Johnsson, Javier Medina, Jesús Peral, and Julian Szymanski



Advances in the Application of Human Activity Recognition

Computational Intelligence and Neuroscience

Advances in the Application of Human Activity Recognition

Lead Guest Editor: David Gil
Guest Editors: Magnus Johnsson, Javier Medina,
Jesús Peral, and Julian Szymanski

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Computational Intelligence and Neuroscience
Volume 2023, Article ID 1394882, 22 pages
https://doi.org/10.1155/2023/1394882

Research Article

Towards a Better Performance in Facial Expression Recognition: A Data-Centric Approach

Christian Mejia-Escobar,[1] **Miguel Cazorla**,[2] **and Ester Martinez-Martin**[2]

[1] *Central University of Ecuador, P.O. Box 17-03-100, Quito, Ecuador*
[2] *Institute for Computer Research, University of Alicante, P.O. Box 99. 03080, Alicante, Spain*

Correspondence should be addressed to Christian Mejia-Escobar; cimejia@uce.edu.ec

Received 30 July 2022; Accepted 16 September 2022; Published 3 November 2023

Academic Editor: Jianli Liu

Facial expression is the best evidence of our emotions. Its automatic detection and recognition are key for robotics, medicine, healthcare, education, psychology, sociology, marketing, security, entertainment, and many other areas. Experiments in the lab environments achieve high performance. However, in real-world scenarios, it is challenging. Deep learning techniques based on convolutional neural networks (CNNs) have shown great potential. Most of the research is exclusively model-centric, searching for better algorithms to improve recognition. However, progress is insufficient. Despite being the main resource for automatic learning, few works focus on improving the quality of datasets. We propose a novel data-centric method to tackle misclassification, a problem commonly encountered in facial image datasets. The strategy is to progressively refine the dataset by successive training of a CNN model that is fixed. Each training uses the facial images corresponding to the correct predictions of the previous training, allowing the model to capture more distinctive features of each class of facial expression. After the last training, the model performs automatic reclassification of the whole dataset. Unlike other similar work, our method avoids modifying, deleting, or augmenting facial images. Experimental results on three representative datasets proved the effectiveness of the proposed method, improving the validation accuracy by 20.45%, 14.47%, and 39.66%, for FER2013, NHFI, and AffectNet, respectively. The recognition rates on the reclassified versions of these datasets are 86.71%, 70.44%, and 89.17% and become state-of-the-art performance.

1. Introduction

Our facial gestures speak more than a thousand words. Among the dynamic activities of the human body, the muscular movements of the face have meaning and potential interpretation. Facial expressions associated with the emotional state of a person are considered universal and the main signal to manifest and infer our feelings and sensations [1, 2]. An important study [3] quantified the degree of influence of the elements involved in the communication of emotions, determining the nonverbal part (facial and body gestures) as the most influential with 55%, whereas the tone of voice with 38%, and only 7% for verbal language. In a conversational context, the exclusively verbal manifestation of anger or happiness must be accompanied by a facial gesture to convey the credibility and conviction of the interlocutor. Even the gesture would be enough to describe the emotion we are experiencing, as we often pay more attention to the face than to the words. The recent pandemic has shown that when a facial mask is present, the human capacity to infer emotions is reduced [4]. Therefore, facial expressions that communicate emotions are essential in daily life at the individual, interpersonal and social levels [5]. Apart from interacting with other people, we are increasingly surrounded by machines trying to imitate human behavior, so there is a need to interact. In near future, this will be a common practice and it is intended to make such interaction as natural as possible. In the same way that people can infer the emotional state of others from facial expressions, computers and robots may also be able to recognize expressions and interpret human emotions. In recent years, automatic facial expression recognition (FER) has become an important area of research and development to improve human-machine interaction (HMI), leading communication to a more emotional, affective, and intelligent level [6, 7]. This can be applied to many activities and fields such as human behavior, healthcare, medicine, psychology, psychiatry, marketing, digital advertisement, customer feedback assessment, video games, video security, video surveillance, mobile phone unlocking, crime investigation (lie detection), online learning, and automobile safety [8–11].

1.1. Problem. Humans can easily recognize facial expressions, however, it is still a challenge for machines [12]. Automatic FER is one of the key tasks in the field of computer vision. This problem has motivated competitions such as the one organized on the Kaggle platform [13]. A popular approach is to classify the facial expression in a static image of a human face and associate it with one of the seven basic universal human emotions: happiness, surprise, anger, sadness, fear, disgust, and neutral [14, 15]. Some models measure emotions with continuous values (e.g., valence and arousal). However, there are very limited annotated facial databases [16]. In contrast, for a discrete (categorical) model, a wider range of available datasets can be found. Deep learning is preferred for this task avoiding the high cost of time and effort of manually defining multiple and complex features of facial expressions. In particular, convolutional neural networks (CNNs) have shown promising results from different facial image datasets. Images captured in a specific and controlled environment (in the lab) are taken of a few people, do not present variations in environmental conditions, and gestures have a high degree of expressivity, so a good level of accuracy can be achieved. Another way is collecting images in real-world situations from the Internet, which is referred to as in the wild [17]. The heterogeneity of human faces, people less expressive than others, subtle differences between expressions, variations in head pose, different body postures, lighting changes in the environment, and occlusions, are some of the factors that make FER outside the laboratory a difficult task even for humans [9, 18–20].

1.2. Motivation. A deep learning solution consists of a model and data. The vast majority of work follows a model-centric approach, whose purpose is finding new algorithms to achieve better performance on a certain facial image dataset. Several CNN architectures have been proposed, both customized (created from scratch) and pretrained using transfer learning and fine tuning techniques. Each one tests different hyperparameters and includes regularization mechanisms such as data augmentation, dropout, and batch normalization [9]. In practice, this process is very time-consuming and has not achieved the aim of ideal performance. On the other side, there is research data-centric guided by the principle that data is the most important resource and its quality directly influences the performance of learning models. Very few studies have focused on improving FER datasets even though the same creators admit the problems in the quality of the data [8]. The lack of remarkable results of the model-centric approach, the little work focused on the data, and the premise that the data would be more important than the model, motivate us to propose a novel data-centric method to improve existing FER datasets to achieve better performance of recognition models.

1.3. Hypothesis. The quality of the dataset is a prerequisite for improving the accuracy of FER models. If the inherent drawbacks of the dataset are not reduced, it is very difficult to improve the performance of a FER system. In other words, better performance and higher accuracy are expected if the dataset is improved.

1.4. Method. Improving the main resource of a FER model, i.e., the dataset, implies improving the accuracy of the recognition. To validate our proposal, we used some representative datasets of this domain, which suffer from well-known problems such as imbalance, irrelevant images, and misclassified images. Our interest is to deal with misclassification, since balancing or removing irrelevant images would modify the size of the dataset. In contrast, a reclassification would generate a new distribution of the available images in a better-quality dataset. The strategy is a progressive refinement of the dataset over several trainings of the same CNN-based model. After each training, the prediction of all facial images is performed, and only the correct ones are selected to form the dataset for the next training. This process is repeated until there are few incorrect predictions, usually single-digit numbers. As a result, the last trained model achieves very high accuracy, so it is in charge of relabeling all the images of the original dataset. Therefore, a new distribution of the dataset is generated without altering its size or modifying the images. In the final step, the same CNN model is trained on the reclassified version of the dataset, and the accuracy is higher compared to the original dataset. The experiments performed in the present work show an increase of 20.45%, 14.47%, and 39.66% for the FER2013, NHFI, and AffectNet datasets, respectively. State-of-the-art performance was also achieved for these datasets.

1.5. Contributions. Our research work provides: (1) a novel data-centric method to reclassify the images of a dataset that allows a higher precision of a FER model, (2) a methodology applicable to other datasets from different domains and supported by computer tools, especially Python and deep learning libraries, and (3) a reclassified version of each dataset, which may be useful for further research, publicly available for FER2013 and NHFI, whereas for AffectNet this is not possible due to licensing restrictions.

The content of this work is organized as follows: Section 2 reviews the data-centric works. Section 3 presents the FER datasets. Section 4 describes in detail the methodology. Section 5 explains the experimentation, and the results obtained in Section 6. Finally, Section 7 includes the conclusions and mentions future work.

2. Related Work

Our bibliographic search on improving the performance of FER in the wild using deep learning reports supremacy of model-centric research. This approach focuses on better architectures, hyperparameter tuning, and regularization techniques [21]. However, no significant progress can be expected when the data used are not reliable. On the other hand, data-centric efforts are scarce. There are few studies that deal with the dataset to improve the performance of a FER system. After analyzing the related literature, we can say the techniques frequently used under this approach include: image preprocessing, removing noise, deleting images with errors, data augmentation, and reclassification. For instance, Liu et al. [11] analyzed expression recognition considering the importance of data preprocessing by improving the image contrast. More discriminative facial features are obtained using a hybrid method for extraction, and a classification network combining EGG-16 and ResNet. Experiments on three benchmark datasets: CK+, FER2013, and AR achieved state-of-the-art recognition rates: 98.6%, 94.5%, and 97.2%, respectively. Kim et al. [22] designed an image and video preprocessing system called FIT (facial image threshing) machine capable of eliminating irrelevant facial images, cropping, resizing, and reorganizing the classification of facial images before training the Xception algorithm, improving the validation accuracy by 16.95% with the FER2013 dataset. Mazen et al. [8] applied the following operations on the dataset: (1) nonface images, text images, and profile images are deleted, (2) wrongly labeled images are relabeled using a CNN, and (3) data augmentation to overcome the class imbalance, generating new face images for the minority classes with a cycle generative adversarial network (CycleGAN). As a result, the average test accuracy was increased from 64% for the original FER2013 dataset to 91.76% for the modified balanced version. The cited works address the preprocessing of the dataset before the training of a model, however, the operations applied to change the total number of images either by removing or augmenting. In addition, the images are modified by cropping, resizing, or retouching the contrast. Our goal is to preserve the images and size of the dataset, so we focus on misclassification, one of the most influential problems in the lower performance of the FER models. For instance, Kim and Wallraven [23] presented a study of the quality of the labeling on AffectNet. Due to the large size of the dataset, a subset with a total of 800 difficult-to-recognize images of the different categorical expressions was selected to be relabeled by 13 human annotators. After the crowd reannotation, 83.25% of the total number of votes did not match the original dataset labels. In addition, the predictions of several ResNets trained on the original AffectNet are compared with the labels assigned by the human crowd, finding that there is no good coincidence for categorical expression. This pilot test suggests the low labeling quality of the original dataset for these difficult facial images, influencing the poor performance of a deep learning model. It is mentioned that more extensive reannotation work is in progress to check more accurate performance, however, manual annotation demands great effort and time. Our work does not require any kind of preparation or modification of the images, and avoids decreasing or increasing their number. It aims to automatically reclassify images to reduce intraclass variability and interclass overlapping of the original dataset. As a consequence, improve recognition performance.

3. Datasets

There are multiple image datasets created for automatic emotion recognition based on facial expressions. We have considered FER2013, AffectNet, and NHFI (natural human face image), mainly due to availability, size, image format, and categories of facial expressions.

3.1. Characteristics. The FER2013 dataset (created by Pierre-Luc Carrier and Aaron Courville) and AffectNet (Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor) are standards taken as benchmarks for competitions [24], whereas NHFI (Sudarshan Vaidya) is a novel dataset, created for the purpose of providing more data with better manual annotation, which we propose to analyze in the present study. Table 1 summarizes the most relevant characteristics of these datasets.

The quality of the datasets is more affected as the size of the dataset increases, so we selected a dataset at different scales: small (thousands of images), mid (tens of thousands), and large scale (hundreds of thousands). The facial images included are static, not video sequences, with a 2D or flat appearance, in contrast to the 3D images that generate a perception of depth [1]. Each image has a facial expression category assigned to it, this is a task performed entirely by humans, except for AffectNet, where one part was manually annotated and the rest automatically annotated using a neural network trained on all manually annotated training set samples [16]. The datasets are not balanced, i.e., they do not have the same number of images for each category, or at least a similar number. This drawback is discussed later. To examine the influence of image color and size on recognition performance, we have images in grayscale and RGB mode, as well as in small and medium sizes. JPG and PNG are standard image formats and are easy to convert to each other. The datasets encompass difficult naturalistic conditions (in the wild), with images far from a controlled environment, closer to reality, different lighting levels, ages, poses, intensity of expression, and occlusions, making recognition a challenging task [19].

3.2. Acquisition. The FER2013 (https://www.kaggle.com/datasets/deadskull7/fer2013) and NHFI (https://www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition) datasets are publicly available in Kaggle, whereas AffectNet requires permission for use via a request form to the authors (request form: mohammadmahoor.com/affectnet-request-form/).
FER2013 can be obtained in a comma-separable value (CSV) format whose columns represent the following attributes:

Table 1: Datasets considered and their main characteristics.

Characteristic	FER2013	NHFI	AFFECTNET
Number of images	35886	5558	287401
Expression model	Discrete	Discrete	Discrete/continuous
Categories	7	8	8
Type	2D facial image	2D facial image	2D facial image
Labelers	Humans	Humans	Automated and humans
Balanced	No	No	No
Resolution (pixels)	48 × 48	224 × 224	224 × 224
Color	Grayscale	Grayscale	RGB color
Format	JPG	PNG	JPG
Space	300 MB	50 MB	4 GB
Availability	Free	Free	Under request
Data source	Internet	Internet	Internet
Size	Mid	Small	Large
Environment	In-the-wild	In-the-wild	In-the-wild
Year	2013	2020	2017
Structure	CSV file	Folders and files	Image and NumPy files
Subsets (%)	Train/test (80/20)	None	Train/val (99/1)

a value between 0 and 6 for each of the 7 possible emotions (0: angry, 1: disgust, 2: fear, 3: happy, 4: neutral, 5: sad, and 6: surprise), a list of 2304 integer values, each equivalent to one pixel of the image of size 48 × 48, and finally the subset to which it belongs: training or test. Since the images are not directly visible, we used a Python script with the Pandas and NumPy libraries to read the file, store the integer values as pixel arrays, and convert them to image files. A total of 35886 images are obtained after transforming the pixel arrays to image files in JPG format, in grayscale and with a resolution of 48 × 48 pixels, divided into two subsets: training and test, 28708 and 7178 images, respectively. Each subset includes 7 folders, each one for a particular type of facial expression. NHFI downloading is a compressed file, which after decompression generates 8 folders, whose names are practically the same as the previous dataset, only the "contempt" category is excluded for a fair comparison. Inside each folder are images in PNG format. In the case of AffectNet, the link provided in response to the request allows for the download of two compressed archives for training and validation. After the extraction of each archive, an "images" folder containing the JPG files and another one called "annotations" containing the NPY files of the corresponding labels are created. We developed a Python script (github.com/cimejia/FER-datasets/blob/main/createAffecnet.py) to read the facial expression category from the NPY file and move the JPG file to the corresponding folder. It is worth mentioning that AffectNet has two versions of the dataset, we used the small one containing only the manually annotated images with 8 labels (but contempt is omitted) released in March 2021. The full AffectNet dataset is huge (122 GB) and a specific request is necessary [16].

3.3. Drawbacks. Automatic collection from the Internet and label crowdsourcing are the main reasons for the quantity and quality drawbacks of FER datasets. Regarding quantity, the major disadvantage is the imbalance, even with categories that largely exceed the number of facial images in other categories. On the other hand, the quality of the content is highly affected by the presence of irrelevant images and misclassification. These problems are widely mentioned in the literature and increase as the size of the dataset grows [8].

3.3.1. Imbalance. An imbalanced dataset could lead to a recognition model biased in favor of the majority classes. Having the same number of images per category is a difficult task. Facial images are usually sourced from the Internet and collected manually or automatically through browser plug-ins or programming scripts. These images are posted by people who tend to show smiling or happy faces, so this category predominates, in contrast to categories such as disgust, anger, or sadness, which users do not usually post. Table 2 indicates the number of images per facial expression category in each dataset.

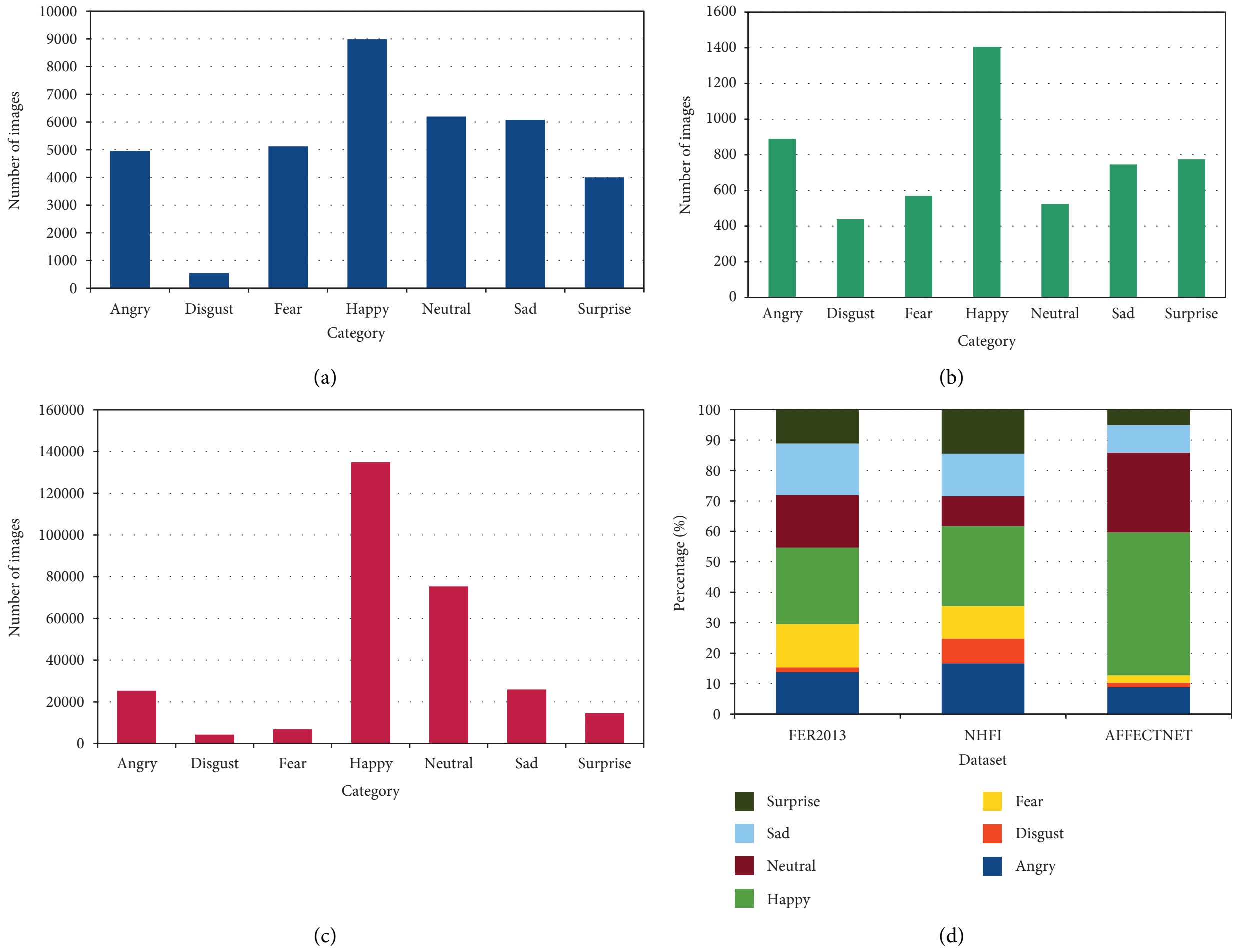
All three datasets show a significant imbalance (Figure 1). In FER2013 (Figure 1(a)), the "happy" category predominates, and the "disgust" category has few samples, and it is approximately regular for the rest of the categories. NHFI (Figure 1(b)) presents a similar behavior, but is less irregular. In AffectNet (Figure 1(b)), the difference in the number of images between all categories is much more pronounced.

Comparing the distributions on the same scale (Figure 1(d)), the imbalance is much more significant in AffectNet. A common pattern is the higher number of samples for the happy category and the lowest number for the disgust category. As mentioned before, this is because people tend to post images of happy faces and avoid showing other types of expression.

3.3.2. Misclassification and Irrelevant Images. Here, we join both problems related to the content of the datasets. Misclassification or mislabeling refers to placing facial images in the wrong directories. Among the factors that lead to this problem are: (a) emotions are subjective, it is common that two people to have different opinions on the

Table 2: Distribution of categories and number of images in FER datasets.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013	4953	547	5121	8988	6198	6077	4002	35886
NHFI	890	439	570	1406	524	746	775	5350
AffectNet	25382	4303	6878	134915	75374	25959	14590	287401



(a)

(b)

(c)

(d)

Figure 1: Imbalance in (a) FER2013; (b) NHFI; and (c) AffectNet; (d) overall.

same facial image, (b) there are slight differences between certain facial expressions, e.g., fear and surprise, disgust and anger, and contempt and sadness, (c) the degree of expressiveness varies from person to person, so gestures may appear exaggerated in one case and inhibited in others, and (d) human beings can feel multiple emotions in a given instant, something that is difficult to combine in a facial expression and can be confusing, e.g., smiling carrying tears is a combined emotion mistaken for sadness [9, 25, 26]. As irrelevant images are those with watermarks, occlusions, no faces, poorly visible or very dark, cartoons, text or symbols, half-side, sleeping faces or closed eyes, cropped, rotated, retouched, and duplicated images. It is important to check for these drawbacks in each dataset, however, an exhaustive manual and visual review of a large number of images are impractical. We designed the following procedure to easily locate such errors.

Search for facial images with errors follows the flowchart shown in Figure 2. We reused the CNN for facial expression recognition designed by Akshit Bhalla [27]. During the training on each dataset, we monitored the accuracy of the validation set at each iteration (epoch) to save the best model parameters. This model is used to perform the prediction on all the images of the validation set. The confusion matrix is obtained from these predictions, where the off-diagonal positions allow us to identify the failures and their corresponding images. As a result, we have a smaller set of images in each class that is stored in a separate folder. We then visually reviewed to select examples of mislabeling and irrelevant images with their respective file names (Figures 3–5).

In this section, we examined the problems of the FER datasets, which can be summarized as class imbalance, the existence of a significant number of images that are

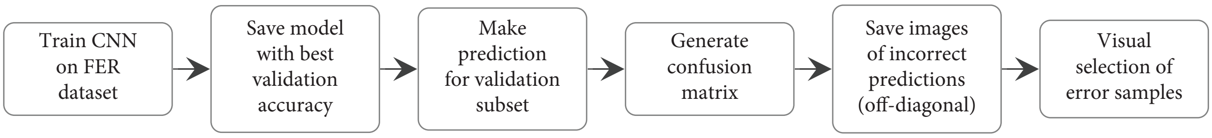
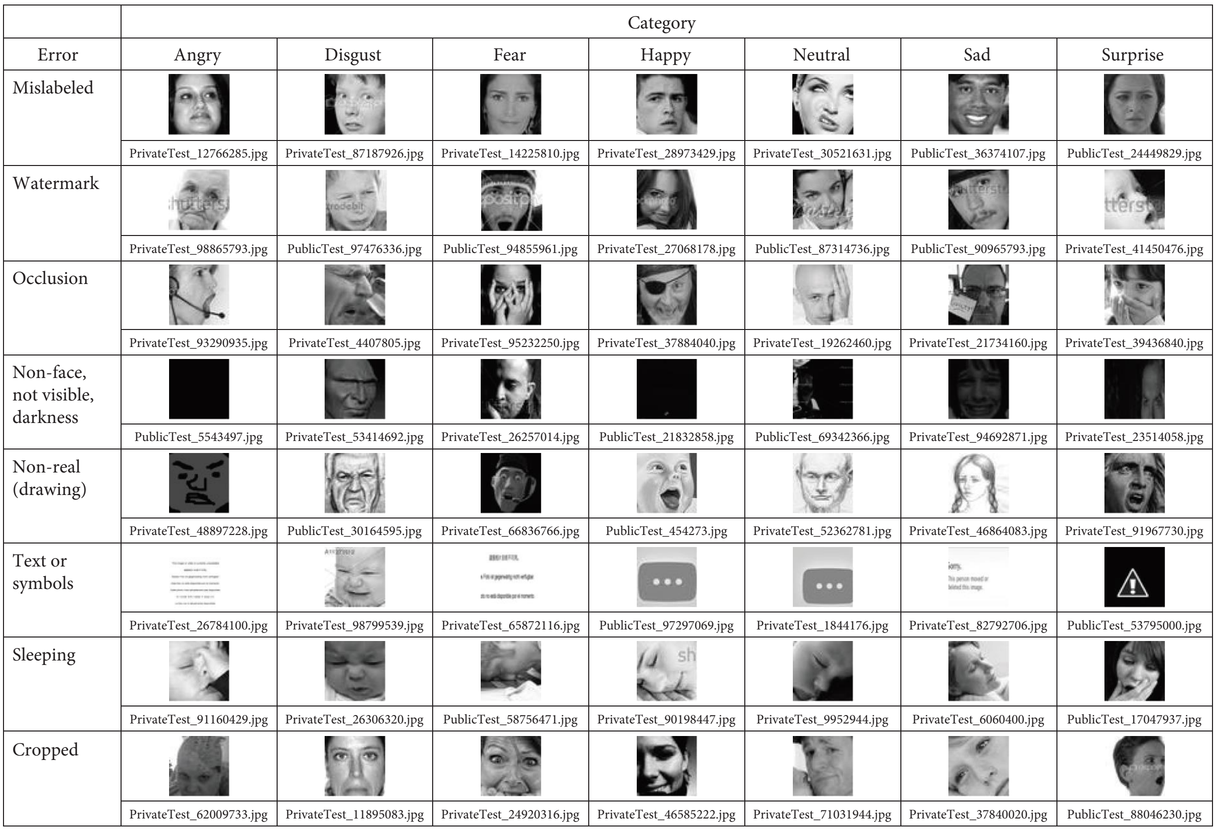

Train CNN on FER dataset
Save model with best validation accuracy
Make prediction for validation subset
Generate confusion matrix
Save images of incorrect predictions (off-diagonal)
Visual selection of error samples

Figure 2: Workflow for selecting and showing some error samples.



	Category						
Error	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Mislabeled	PrivateTest_12766285.jpg	PrivateTest_87187926.jpg	PrivateTest_14225810.jpg	PrivateTest_28973429.jpg	PrivateTest_30521631.jpg	PublicTest_36374107.jpg	PublicTest_24449829.jpg
Watermark	PrivateTest_98865793.jpg	PublicTest_97476336.jpg	PublicTest_94855961.jpg	PrivateTest_27068178.jpg	PublicTest_87314736.jpg	PublicTest_90965793.jpg	PrivateTest_41450476.jpg
Occlusion	PrivateTest_93290935.jpg	PrivateTest_4407805.jpg	PrivateTest_95232250.jpg	PrivateTest_37884040.jpg	PrivateTest_19262460.jpg	PrivateTest_21734160.jpg	PrivateTest_39436840.jpg
Non-face, not visible, darkness	PublicTest_5543497.jpg	PrivateTest_53414692.jpg	PrivateTest_26257014.jpg	PublicTest_21832858.jpg	PublicTest_69342366.jpg	PrivateTest_94692871.jpg	PrivateTest_23514058.jpg
Non-real (drawing)	PrivateTest_48897228.jpg	PublicTest_30164595.jpg	PrivateTest_66836766.jpg	PublicTest_454273.jpg	PrivateTest_52362781.jpg	PrivateTest_46864083.jpg	PrivateTest_91967730.jpg
Text or symbols	PrivateTest_26784100.jpg	PrivateTest_98799539.jpg	PrivateTest_65872116.jpg	PublicTest_97297069.jpg	PrivateTest_1844176.jpg	PrivateTest_82792706.jpg	PublicTest_53795000.jpg
Sleeping	PrivateTest_91160429.jpg	PrivateTest_26306320.jpg	PublicTest_58756471.jpg	PrivateTest_90198447.jpg	PrivateTest_9952944.jpg	PrivateTest_6060400.jpg	PublicTest_17047937.jpg
Cropped	PrivateTest_62009733.jpg	PrivateTest_11895083.jpg	PrivateTest_24920316.jpg	PrivateTest_46585222.jpg	PrivateTest_71031944.jpg	PrivateTest_37840020.jpg	PublicTest_88046230.jpg

Figure 3: Some errors in the FER2013 dataset.

irrelevant, or that do not correspond to the correct category. Combined or separately, these problems cause the performance of a FER model to degrade considerably, as well as learning to be biased in favor of the dominant classes [8, 9, 18, 22]. Therefore, the search for more convenient architectures and configurations for recognition models is a waste of time when the data used are of low quality. Firstly, it is necessary to address these problems to improve the datasets. Dealing with both the imbalance and the irrelevant images involves changing the size of the original dataset. Our work focuses on the problem of misclassification by keeping the number of available images of the dataset. To this end, we propose a novel data-centric method based on deep learning for the automatic relabeling of facial images.

4. Proposed Method

Our goal is to achieve increased accuracy in facial expression recognition through deep learning by previously improving the dataset used. We proposed a data-centric approach that specifically addresses the misclassification typically encountered in FER datasets. This drawback is likely the most influential in the lower performance of recognition models in the wild scenarios. Since a visual inspection of every facial image in a dataset would be an extremely time-consuming and tedious task, we designed a method to automatically reclassify images of a dataset and improve the performance of a FER model.

4.1. Workflow.

The proposed method consists of a series of steps represented by a workflow diagram in Figure 6.

(1) The dataset is organized in a folder-based structure, where each facial expression category is a folder containing the corresponding facial image files.

(2) Split the dataset into training and validation subsets, with the same folder and file structure. The training subset is larger and has the images to fit the model, whereas the validation subset is used to evaluate the model at training time. We omit a test subset because as many images as possible are needed for the next step. Thus, the input is ready for the deep learning model.

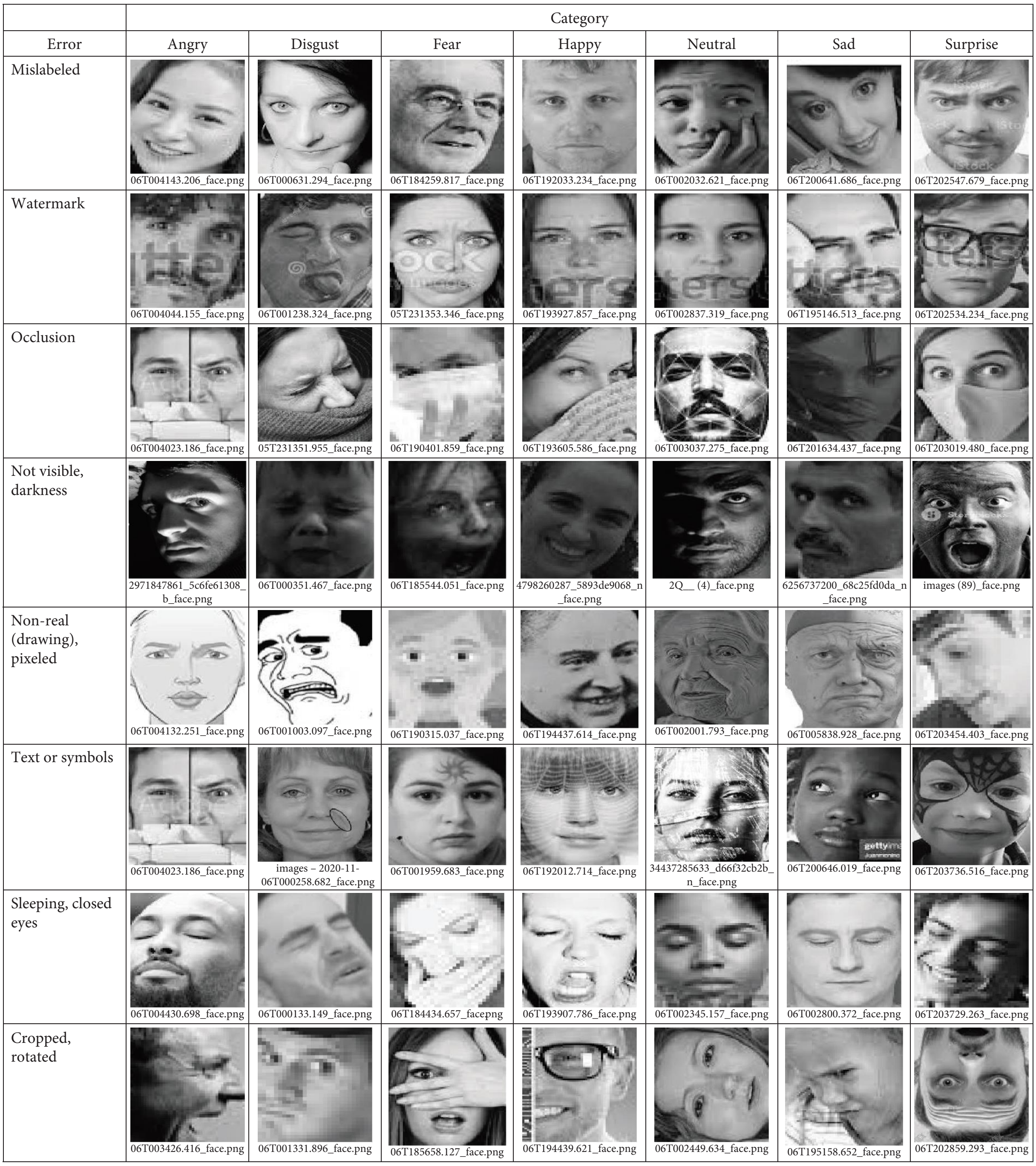


Figure 4: Some errors in the NHFI dataset.

(3) A CNN created from scratch or pretrained via the transfer learning technique is trained on the FER dataset. Both alternatives are shown in this work. Training is an iterative optimization process in which the model reduces an error as it learns to associate images and category labels.

(4) The training is monitored to save in a model file the parameters (weights and biases) corresponding to the iteration (epoch) of the best validation accuracy.

(5) The best model is used to perform the prediction of all facial images in the dataset. The results obtained allow us to generate the confusion matrix.

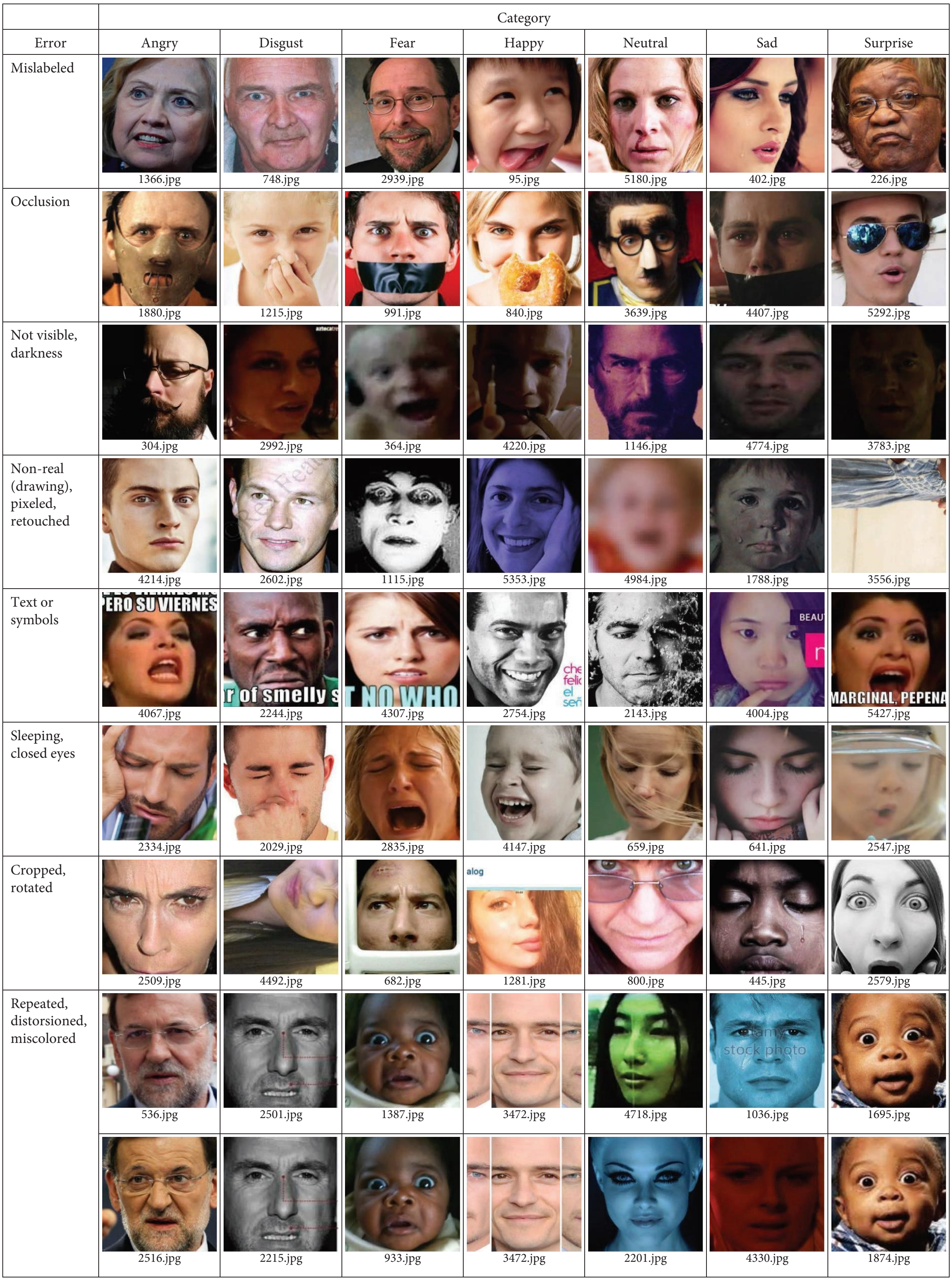


Figure 5: Some errors in the AffectNet dataset.

(6) The confusion matrix is evaluated considering a good dataset when the precision of each category exceeds 90% or the numbers outside the main diagonal are single digits. Several successive trainings will be necessary to meet these criteria.

(7) The correct predictions on the main diagonal of the confusion matrix allow us to select the corresponding facial images, which will form a smaller but much more reliable version of the dataset.

(8) The new version of the dataset is automatically divided into training and validation subsets, and training is performed with the same CNN. The process is repeated until the conditions established for a good dataset are reached.

(9) The last saved model performs the prediction of facial expression for all images in the original dataset. The result is the automatic reclassification generating a new distribution with all facial images.

In summary, we propose a process of iterative trainings to create successively more refined versions of the dataset. Each version is smaller, only the correct predictions of facial expression are included, but maintains a significant number of images. At the last training, a much more reliable dataset is obtained, as well as a model that produces a low number of incorrect predictions (single-digit values for each class). The convolutional network is fixed in terms of its architecture and hyperparameters along this process.

The key idea is that feature extraction is a crucial part of a FER system, and the expression classification accuracy will improve with an effective extraction of facial features [10, 11]. The progressive refinement of the dataset produces a smaller number of images in each training, but with less variability of the gestures of the faces. Therefore, the model can capture more distinctive features of each class gradually. As a consequence, it is possible to increase intraclass similarity and enlarge interclass differences within a dataset, thereby improving the accuracy of facial expression recognition in real-world scenarios.

4.2. Models. We leverage CNNs, current state-of-the-art tools in Computer Vision, for facial expression prediction in images. The design of CNNs imitates the human visual system, where a convolutional part would be the eyes of the network whereas a classifier part would be the brain, which decides the class of the object. CNNs can be created from scratch or pretrained using the transfer learning technique. In this work, we demonstrate the use of both alternatives, describing the architecture implemented for each of the datasets selected.

4.2.1. FER2013. We reutilized the CNN presented on the Kaggle site (https://www.kaggle.com/bhallaakshit/facial-expression-recognition), whose performance has shown good results in the task of facial expression recognition on this dataset (Figure 7).

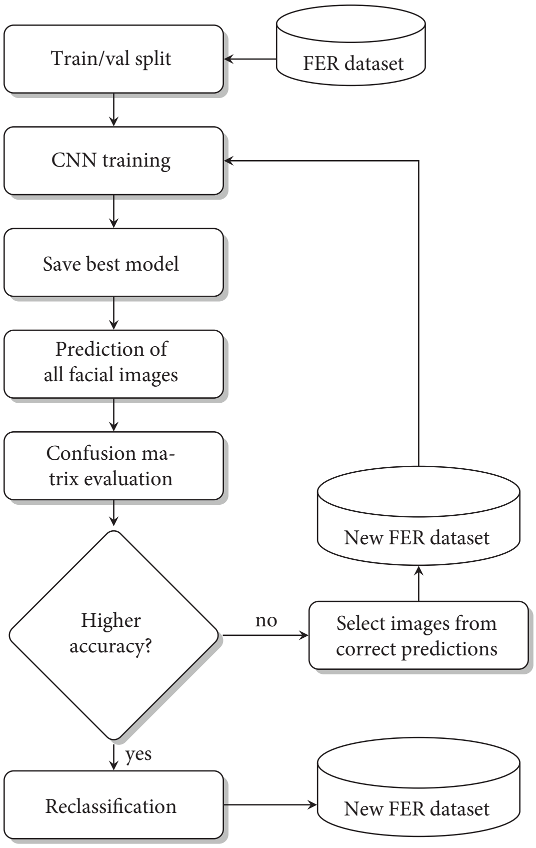


Figure 6: Workflow to automatically reclassify a FER dataset.

The 48 × 48 pixel grayscale input image is passed through 4 convolutional layers, each layer applies a number of filters (kernels) to generate feature maps that include hierarchically detected patterns, from the simplest to the most complex. Here, 64, 128, 512, and 512 filters of size 3 × 3, 5 × 5, 3 × 3, and 3 × 3 pixels, respectively, are applied. A ReLU activation function then turns the negative values to zero and maintains the positive values. Next, a max-pooling operation reduces the image dimensions by half, but preserves the found features. Batch normalization stabilizes the result of a convolution whereas dropout enables the active participation of all neurons in the learning process. Both are recommended regularization techniques to avoid possible overfitting. The flatten operation converts the feature maps into a vector of values suitable as input for the classifier, which is a traditional fully connected neural network with an input layer that receives the features in vector shape, two hidden layers of 256 and 512 neurons, and an output layer with a Softmax activation function for 7 probability values, one for each facial expression class.

4.2.2. NHFI. We tested the same CNN model with this dataset, however, the results after the first filtering indicated an insignificant increase in accuracy (approx. 1.5%) as shown in Table 3.

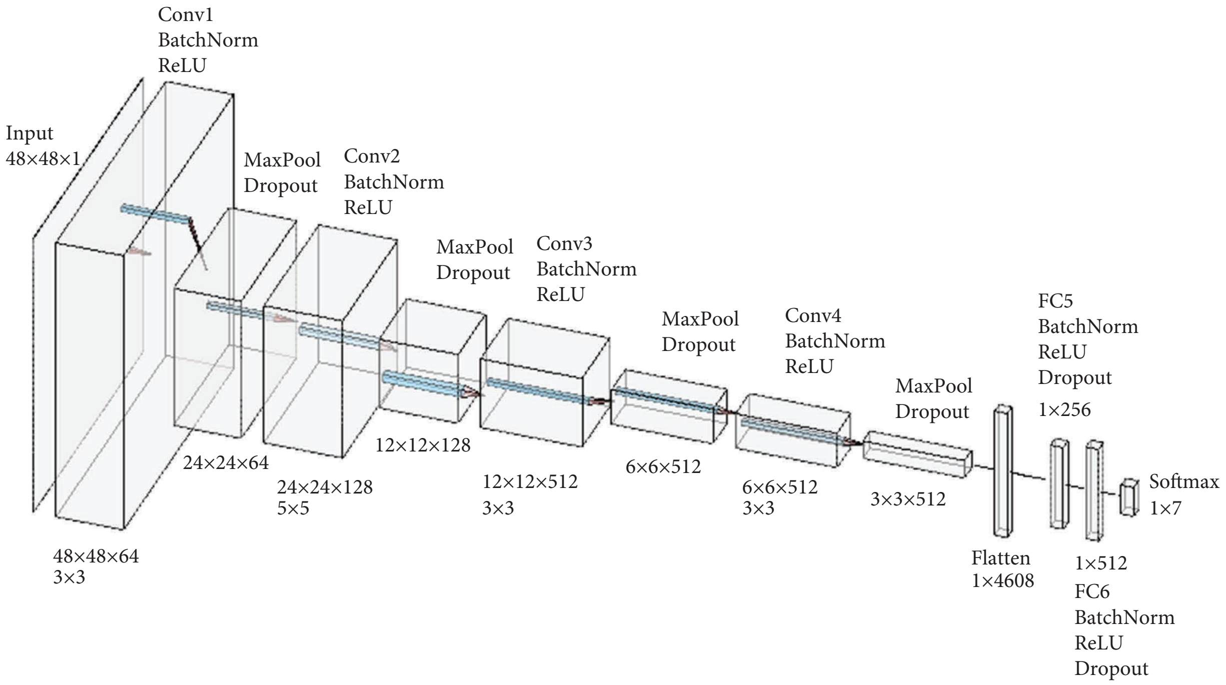


Figure 7: Architecture of the CNN for the FER2013 dataset.

Table 3: Refinement of the NHFI dataset using the CNN from scratch.

Training	Images (train)	Images (val)	Total	Accuracy
1	4278	1072	5350	0.5732
2	3211	616	3827	0.5885

Therefore, we searched for other architectures to achieve higher accuracy. A model using the transfer learning technique showed the best performance for this dataset. In the first filtering, the accuracy improved from 0.5597 to 0.8367 (27.7%) as opposed to 1.5% with the CNN from scratch. Thus, we were able to demonstrate that the proposed method works for both cases (pretrained and from scratch models). With transfer learning, the training phase will be much faster, since we only train the classifier parameters while keeping fixed the convolutional base that would have already learned features that are useful for most computer vision problems. The structure is presented in Figure 8.

The model is based on the $EfficientNet$, a very popular CNN pretrained on the ImageNet dataset [28]. We used version *B0*, whose convolutional base is kept for feature extraction. The advantage is that the image with the original size of 224 × 224 pixels is accepted as input. The classifier receives the features in the form of a flattened vector to decide the class to which the input image belongs by means of a fully connected neural network with two dense layers of 256 and 512 neurons, to which the ReLU activation function is applied, plus the batch normalization and dropout regularization techniques to reduce possible overfitting. The Softmax function in the last dense layer outputs a distribution of probabilities corresponding to each of the 7 categories of facial expression.

4.2.3. AffectNet. We performed several tries with different architectures to determine the most suitable CNN for this dataset. The best result was obtained with the CNN used for the FER2013 dataset (Figure 7). It is only necessary to change the size and color mode of the AffectNet images from 224 × 224 pixels in RGB to 48 × 48 pixels in grayscale. This conversion is performed automatically using the image generator of Python.

5. Experiments

The core of the experimentation is the run of trainings of each CNN-based model on the respective dataset. The main characteristics of the computational platform used are a processor Intel(R) Core(TM) i9-7920X, 2.90 GHz, RAM 64 GB, GPU NVIDIA GeForce RTX208 with RAM 12 GB, and the operating system Linux Ubuntu 18.04.5 LTS. The CNN architectures described in the previous section are implemented using Python version 2.7.17, supported by standard libraries such as OS, NumPy, and Matplotlib, to manage directories and files, numeric arrays, and visualization, respectively. For deep learning work, we used libraries such as TensorFlow, Keras, and scikit-learn, as well as the Image Data Generator utility for image preprocessing.

The learning process is aimed at model learning to associate facial images and labels of expression categories. A series of values known as hyperparameters must be explicitly defined by the programmer before training. There are no fixed rules for determining these values, they are the result of several tests to find the most convenient ones. Table 4 shows the hyperparameters for each model and dataset, which are maintained for all experiments.

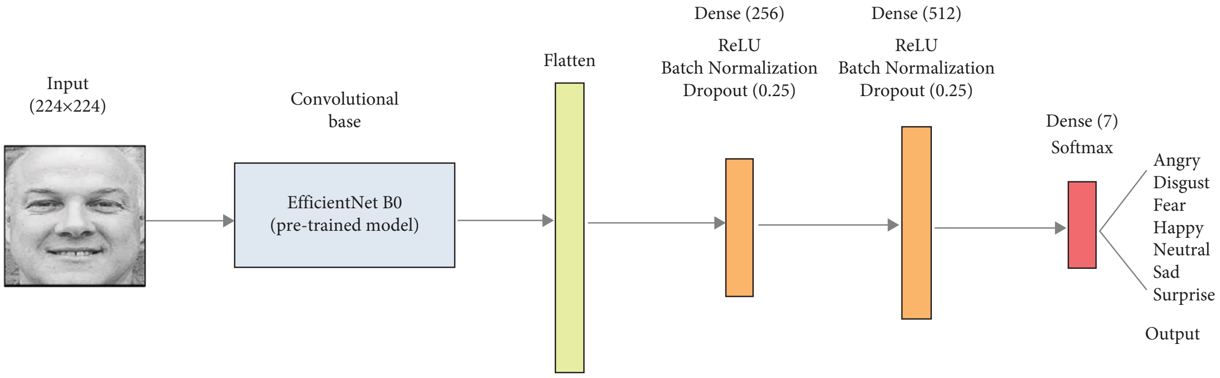


FIGURE 8: Architecture of the CNN with transfer learning for the NHFI dataset.

TABLE 4: Training hyperparameters set for our experiments.

Hyperparameter	FER2013	NHFI	AffectNet
Input shape	48, 48, 1	224, 224, 3	48, 48, 3
Train-val (%)	80–20	80–20	80–20
Batch size	64	64	64
Learning rate	0.01 to 0.00001	0.01 to 0.00001	0.001 to 0.00001
Optimizer	Adam	Adam	Adam
Loss function	categorical_cross entropy	categorical_cross entropy	categorical_cross entropy
Metrics	Loss and accuracy	Loss and accuracy	Loss and accuracy
Number of classes	7	7	7
Epochs	100	50	50
Data augmentation	Yes	No	No
Number of training	5	5	5

Our method of dataset refinement required five successive trainings for each dataset to meet the quality criteria. At each training, the model is fed with the facial images from the training subset of each dataset in batches of 64 images (batch size). We used the Image Data Generator utility from Keras to work with an image generator in batches, due to the large number and size of the images would cause a storage problem in memory. It also allows us to pass the images directly to the training model from directories, as well as automatically labeling the image with the respective category, and performing data augmentation. For each batch, predicted and actual labels are compared, obtaining a *loss* and an *accuracy* using the *categorical_cross entropy* function. Backpropagation and *Adam* (based on gradient descent) algorithms are applied to update the model weights according to the *learning rate* value. When all batches are completed, one *epoch* is accomplished, i.e., one iteration of all training images. The accuracy and loss values are measured after each epoch using the images from validation the subset. One hundred epochs have been run for FER2013, whereas for NHFI and AffectNet fifty epochs were sufficient to know the maximum level of accuracy since beyond this value, the behavior of the model remains practically stable and an improvement is not appreciable. The callback utility from Keras is leveraged to perform certain actions during training such as setting a checkpoint and reducing the learning rate. The model will only be saved to disk if the validation accuracy in the current epoch is greater than what it was in the last epoch. On the other hand, the learning rate tells us how much the weights will be updated each time, and is often between 0 and 1. It will decrease from an initial value to a minimum if the loss does not improve after a certain number of epochs, which usually results in better training.

6. Results

The results of the experimentation are presented graphically by means of learning curves and confusion matrices, whereas the numerical metric used for comparison is the validation accuracy. These tools allow us to evaluate the performance of the model and the improvement of the dataset. During the training and validation of each model, loss and accuracy values have been collected, respectively. This generates the so-called *learning curves*, where the horizontal axis represents the number of epochs and the vertical axis represents either the accuracy or the error. The confusion matrix, also known as the error matrix, is a table to visualize the model performance as presents information about actual and predicted classifications carried out by a classifier model. Rows represent the instances of actual classes, whereas columns represent the instances the classifier predicts [29]. From this matrix, several performance metrics can be obtained, however, we focus on *accuracy*, which compares the number of correct predictions (on diagonal) divided by the total number. The results obtained for each one of the analyzed datasets are presented next.

6.1. FER2013. The learning curves and confusion matrix for each of the five trainings required for the FER2013 dataset are shown in Figure 9. For each training (including validation), the following are presented: the accuracy curves (left), the loss curves (middle), and the corresponding confusion matrix (right). As more trainings are performed, the accuracy curves (training and validation) reach higher values, whereas the loss curves are decreasing in height and near to zero. In addition, the pairs of curves are very close to each other in all the graphs. Therefore, the accuracy of the model is higher, the error is lower, and there is no overfitting. This ideal behavior is the product of successive filtering of the dataset. The confusion matrices include the predictions of facial expressions for all the images in the dataset used for each training. The progressive trainings cause the desired effect in each matrix, that is, to reduce the values outside the main diagonal and to increase the values in this diagonal. The model is each time more accurate because wrong predictions are discarded in subsequent training. As a result, more distinctive features of each class are captured. In this way, the intraclass variability of the facial images is decreased and the interclass variability is increased.

The process of the dataset refinement is summarized in Table 5. Five trainings (four filtering operations) on the FER2013 dataset were necessary to achieve the expected performance metric (validation accuracy). Another training was not necessary because there is no significant improvement in the accuracy. The number of images gradually decreases, but it is still considerable for each training. The model with the highest accuracy (97.7%) has captured the most distinctive features of each facial expression category and is convenient for reclassifying all images in the dataset. The *predict()* method is used to assign the category of every facial image of the original dataset, generating a new distribution of the FER2013 dataset. The comparison is presented in Table 6.

Figure 10 shows that the categories “disgust” and “sad” have minimal variation, those of “angry,” “happy,” and “surprise” vary moderately, and the most affected categories are “fear” (decreasing) and “neutral” (increasing), indicating that the original FER2013 dataset suffers from misclassified facial images, especially between these two categories.

The decisive test of the effectiveness of our method is to train the same CNN on the reclassified FER2013 dataset. Figure 11 shows that better learning curves are obtained, as well as the confusion matrix indicates more correct and fewer incorrect predictions. Higher accuracy and lower loss are verified in Table 7.

The results confirm a more reliable dataset keeping the number of images. The reclassified FER2013 enabled a very significant increase in the validation accuracy of the model by 20.45% and the loss is much lower (0.34). The training accuracy is acceptable (88.76%), very close to the validation accuracy and the loss is lower. There is no overfitting and no significant difference between training loss and validation loss. All the categories show improved accuracy, in particular, there is a remarkable improvement for “angry” (an increase of 26%), “fear” (an increase of 38%), and “sad” (an increase of 25%), i.e., those that showed the most overlapping or confusion. According to the experiments, only 40 epochs in each training would be sufficient, since the behavior remains practically stable beyond this number.

6.2. NHFI. The accuracy curves for the NHFI dataset (left side in Figure 12) start quite separated from the other, evidencing the presence of overfitting, but as the trainings are performed, the curves become closer and reach high accuracy, similar to the loss curves, but in the opposite direction, becoming closer and nearer to the horizontal axis. The confusion matrices show higher values on the main diagonal and lower values of this diagonal, indicating the progressive improvement of the model accuracy, as well as the quality of the dataset used in each training. Despite successive discarding of incorrect predictions, the number of images is significant with respect to the original quantity. Table 8 shows the evolution of the trainings on the NHFI dataset.

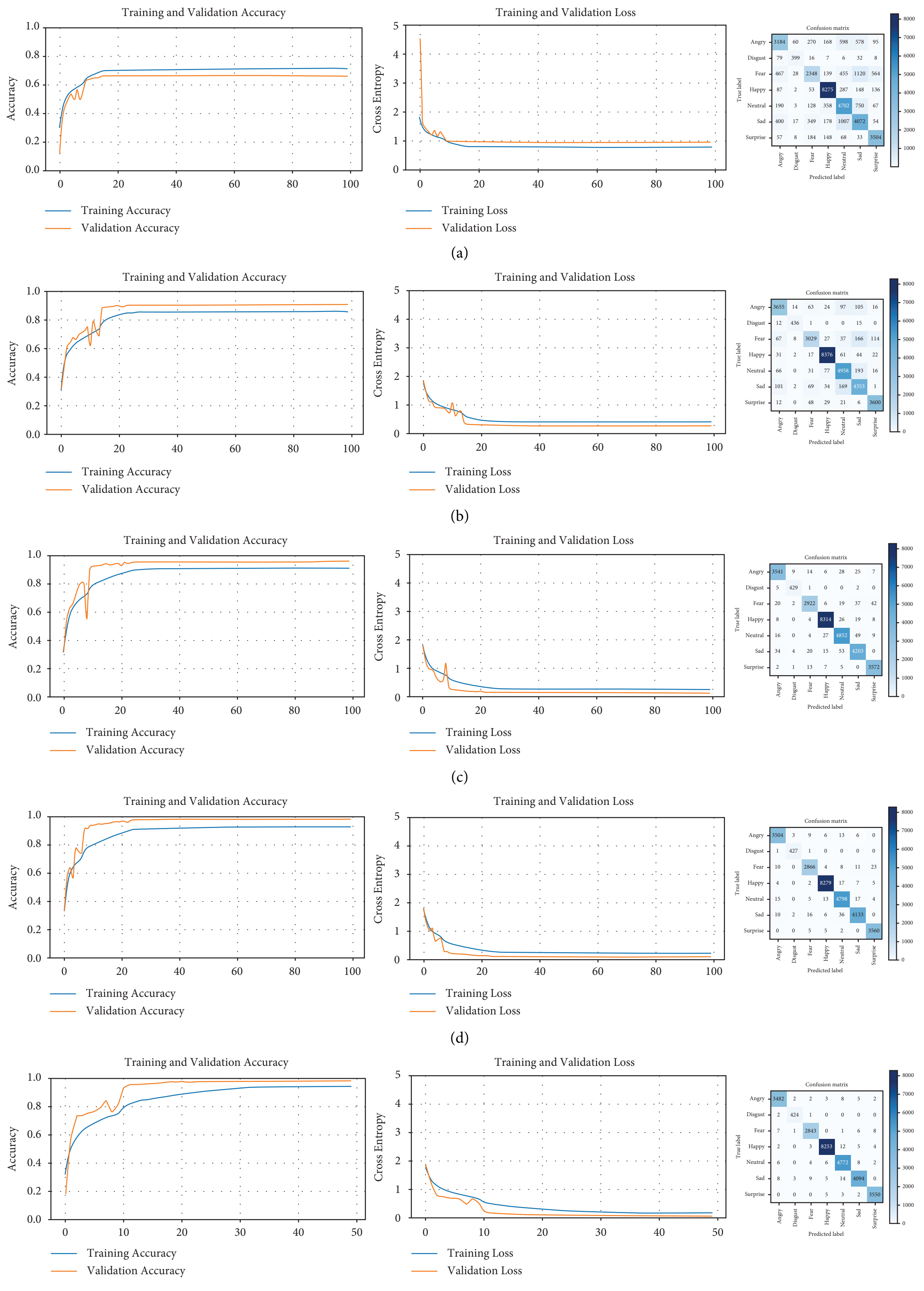
The reclassification of the original NHFI dataset is performed with the highest accuracy model (96.66%). A new distribution of the dataset is generated, which is shown in Table 9. In Figure 13, we can note that the “angry” and “neutral” categories had the greatest changes, indicating that these categories have the most intraclass variability in the original dataset.

To demonstrate improved recognition, the same CNN is trained on the reclassified NHFI dataset and the result is compared to the original dataset (Figure 14). The overfitting was not reduced, but the accuracy is higher, both in the training and validation subsets. The loss is decreased for the reclassified NHFI dataset, as well as the values off-diagonal from the confusion matrix.

The performance results for the original and reclassified distributions of the NHFI dataset are presented in Table 10. We have been able to significantly increase the accuracy in both training and validation subsets, by 18.74 and 14.47%, respectively. Except for the “angry” and “happy” categories, the validation accuracy is highly increased in the rest of the categories, particularly in the “sad” category from 49 to 85%. The methodology based on successive filtering with a transfer learning model has been successfully applied on a different dataset than FER2013.

6.3. AffectNet. The version of the AffectNet dataset we selected contains 287401 images with a large imbalance between the categories (Figure 1(c)). Training on this dataset can lead to biases and erroneous assessment of model accuracy. Therefore, we applied downsampling to balance all categories by considering the one with the lowest number of images. The “disgust” category limited the other categories with 4300 images, of which 3800 have been randomly selected for training by using the split-folders (https://pypi.org/project/split-folders/) library, whereas 500 images by default come with the dataset for validation. The balanced version is shown in Table 11.

The refinement process is performed on this balanced version of the AffectNet dataset. The accuracy curves for the training and validation subsets (Figure 15) start with a small



(a)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3184	60	270	168	598	578	95
Disgust	79	399	16	7	6	32	8
Fear	467	28	2348	139	455	1120	564
Happy	87	2	53	8275	287	148	136
Neutral	190	3	128	358	4702	750	67
Sad	400	17	349	178	1007	4072	54
Surprise	57	8	184	148	68	33	3504

(b)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3655	14	63	24	97	105	16
Disgust	12	436	1	0	0	15	0
Fear	67	8	3029	27	37	166	114
Happy	31	2	17	8376	61	44	22
Neutral	66	0	31	77	4958	193	16
Sad	101	2	69	34	169	4353	1
Surprise	12	0	48	29	21	6	3600

(c)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3541	9	14	6	28	25	7
Disgust	5	429	1	0	0	2	0
Fear	20	2	2922	6	19	37	42
Happy	8	0	4	8314	26	19	8
Neutral	16	0	4	27	4852	49	9
Sad	34	4	20	15	53	4203	0
Surprise	2	1	13	7	5	0	3572

(d)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3504	3	9	6	13	6	0
Disgust	1	427	1	0	0	0	0
Fear	10	0	2866	4	8	11	23
Happy	4	0	2	8279	17	7	5
Neutral	15	0	5	13	4798	17	4
Sad	10	2	16	6	36	4133	0
Surprise	0	0	5	5	2	0	3560

(e)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3482	2	2	3	8	5	2
Disgust	2	424	1	0	0	0	0
Fear	7	1	2843	0	1	6	8
Happy	2	0	3	8253	12	5	4
Neutral	6	0	4	6	4772	8	2
Sad	8	3	9	5	14	4094	0
Surprise	0	0	0	5	3	2	3550

Figure 9: Learning curves and confusion matrices for five successive trainings of the FER2013 dataset. (a) Training #1, (b) training #2, (c) training #3, (d) training #4, and (e) training #5.

Table 5: Summary of experimental results for the FER2013 dataset.

Training	Images (train)	Images (val)	Total	Accuracy
1	28708	7178	35886	0.6702
2	25415	4810	30225	0.9089
3	24001	4379	28380	0.9582
4	23654	4179	27833	0.9761
5	23488	4079	27567	0.9770

Table 6: Distribution of the original and reclassified FER2013 dataset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
FER2013 (original)	4953	547	5121	8988	6198	6077	4002	35886
FER2013 (reclassified)	4817	532	3842	9202	7074	6090	4329	35886

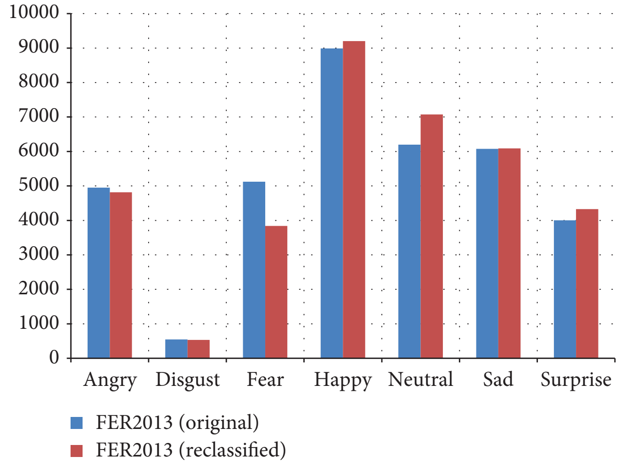


Figure 10: Graphical comparison of both distributions.

separation, which decreases as successive trainings are performed, even the validation curve finishes outperforming the training curve in accuracy. The same behavior, but in the opposite direction, is presented for the loss curves. The values on the main diagonal of the confusion matrix increase with each training and decrease off this diagonal, indicating a higher accuracy of the model due to a better dataset. The evolution of the successive training is summarized in Table 12.

The model in the last training reaches a higher validation accuracy (95.9%), which allows us to reclassify the balanced dataset. A new distribution of the AffectNet of 30100 images is generated, whose number of images per category is presented in Table 13.

After the reclassification of the balanced dataset, the new distribution is imbalanced (Figure 16). The categories of happy and fear have increased significantly, whereas the category of anger has increased slightly. In the remaining cases, there is a decrease, mainly in the categories of disgust and surprise.

Next, the CNN-based model is trained on the new version of the AffectNet dataset to verify that our method works. Figure 17 presents the learning curves of both versions of the dataset, where the new AffectNet (Figure 17(b)) allows to achieve better performance with higher accuracy.

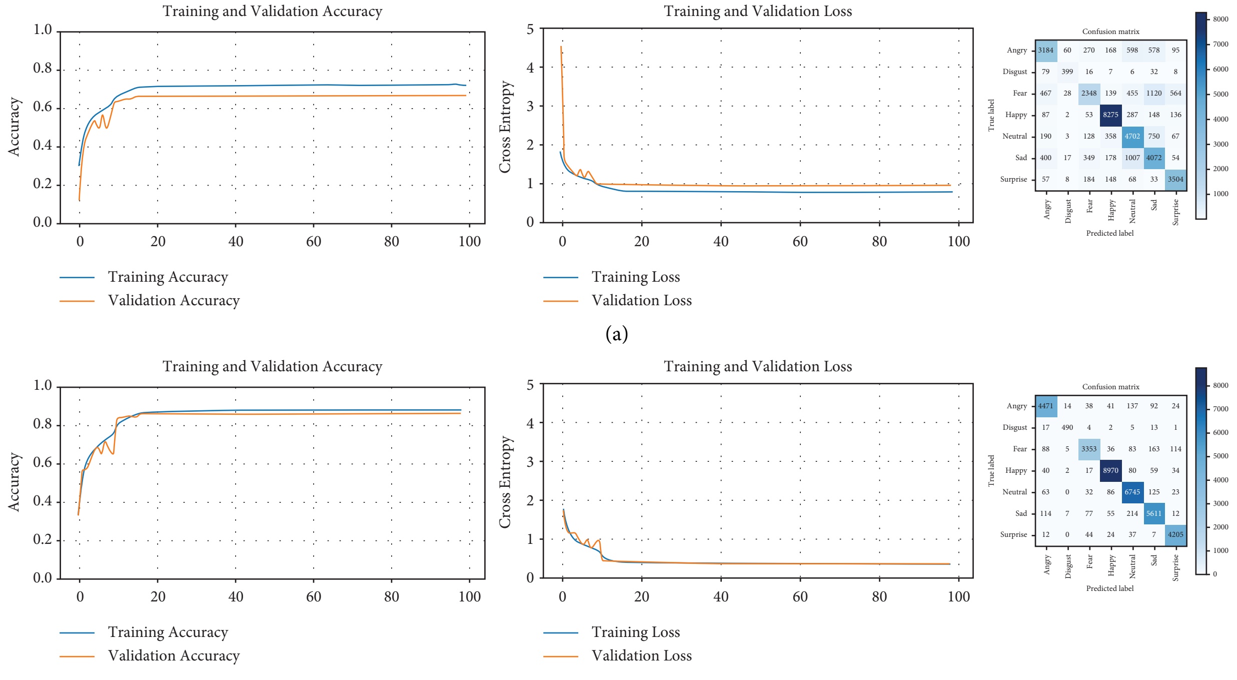
There is a notable improvement in accuracy compared to the first training of the balanced dataset (Table 14). Due to downsampling, the split ratio is 88 and 12%, for the training and validation subsets, respectively. For the new version of the dataset, the proportion is 80 and 20% and having more validation images, the accuracy percentage is almost duplicated (39.66%).

We successfully applied our method to a smaller and more balanced version of the AffectNet dataset. The purpose is to improve the original AffectNet dataset, which is larger and imbalanced. To this end, the last trained model is used to reclassify the facial images in the full version of AffectNet. The new distribution is presented in Table 15.

The bar plot in Figure 18 shows that the shape of the distribution of the new reclassified AffectNet is similar, however, there is a clear increase of images in the categories of fear, disgust, and surprise. This suggests that many facial images of these categories were misclassified as happy or neutral.

The following demonstrates the improved performance in facial expression recognition. The reclassified version of the AffectNet dataset is used to train the same CNN-based model, resulting in the learning curves and confusion matrix displayed in Figure 19. The accuracy curves of the training and validation subsets are increasing from the first epoch and reach a very high level, close to 90%. Also, both curves stay very near to each other. The error curves decrease together to levels near to zero, which is desirable. By using the validation results as suggested by the creators of the dataset, we calculated the accuracy with the *evaluate()* method and generated the normalized confusion matrix. The accuracy on the reclassified validation set is 89.17%, and for each facial expression, category fluctuates between 86% and 96%, which demonstrates a high rate of recognition and no bias for any of the categories as opposed to the original dataset. This behavior confirms better FER performance on the reclassified AffectNet dataset.

Finally, in Table 16, the results of the proposed method are compared with the state-of-the-art performance on the same datasets used in the present work. These are single network models that did not use extra images to the existing ones in the datasets. In all cases, our reclassified versions of the datasets allow us the highest accuracy values for both the



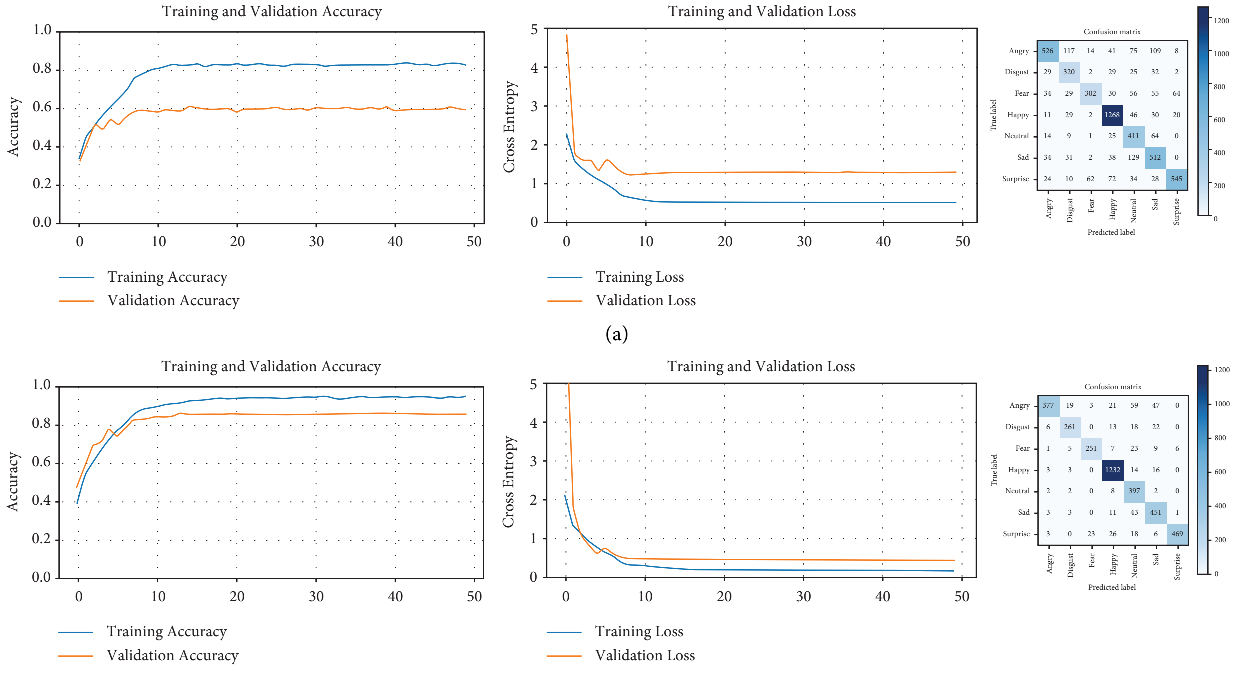
(a)

(b)

Figure 11: Comparison between the original and reclassified FER2013 dataset. (a) FER2013 dataset (original) and (b) FER2013 dataset (reclassified).

Table 7: Comparison of the training results for the original and reclassified FER2013 datasets.

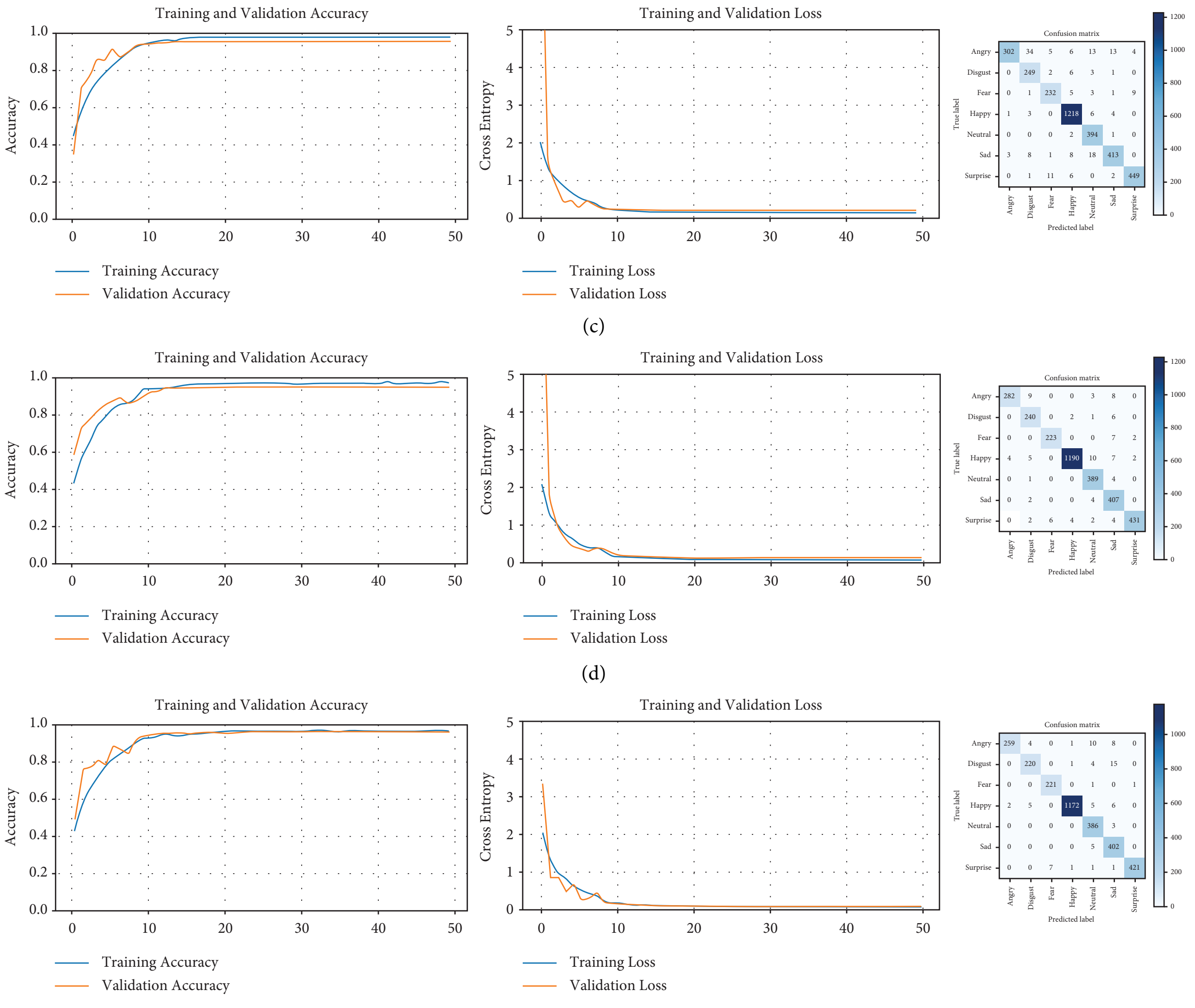
Dataset	Images (training)	Images (validation)	Total	Accuracy
FER2013 (original)	28708	7178	35886	0.6626
FER2013 (reclassified)	28708	7178	35886	0.8671



(a)

(b)

Figure 12: Continued.



(c)

(d)

(e)

Figure 12: Learning curves and confusion matrices for five successive trainings of the NHFI dataset. (a) Training #1, (b) training #2, (c) training #3, (d) training #4, and (e) training #5.

Table 8: Summary of experimental results for the NHFI dataset.

Training	Images (train)	Images (val)	Total	Accuracy
1	4278	1072	5350	0.5597
2	3284	600	3884	0.8367
3	2936	502	3438	0.9382
4	2786	471	3257	0.9533
5	2713	449	3162	0.9666

Table 9: Distribution of the original and reclassified NHFI dataset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
NHFI (original)	890	439	570	1406	524	746	775	5350
NHFI (reclassified)	336	514	383	1585	1042	962	528	5350

model from scratch and transfer learning. For the novel NHFI dataset, there is no formal report on classification accuracy, so we set as a baseline the accuracy achieved by the transfer learning model on the original dataset and contrast it with the reclassified version. These results demonstrate the effectiveness of our data-centric method, as it improves the performance of the FER models even achieving state-of-the-art accuracy values.

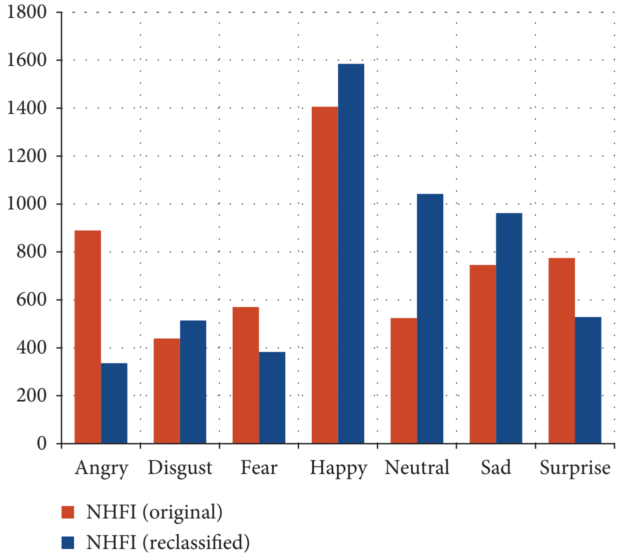
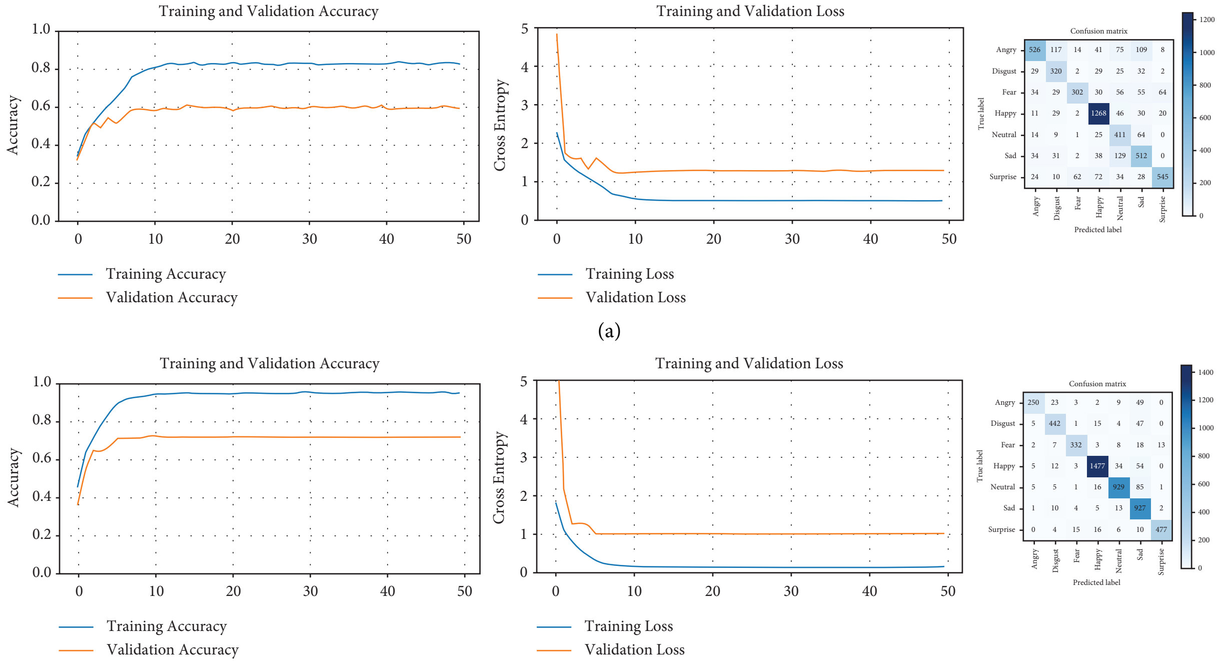


Figure 13: Graphical comparison of both distributions.



(a)

(b)

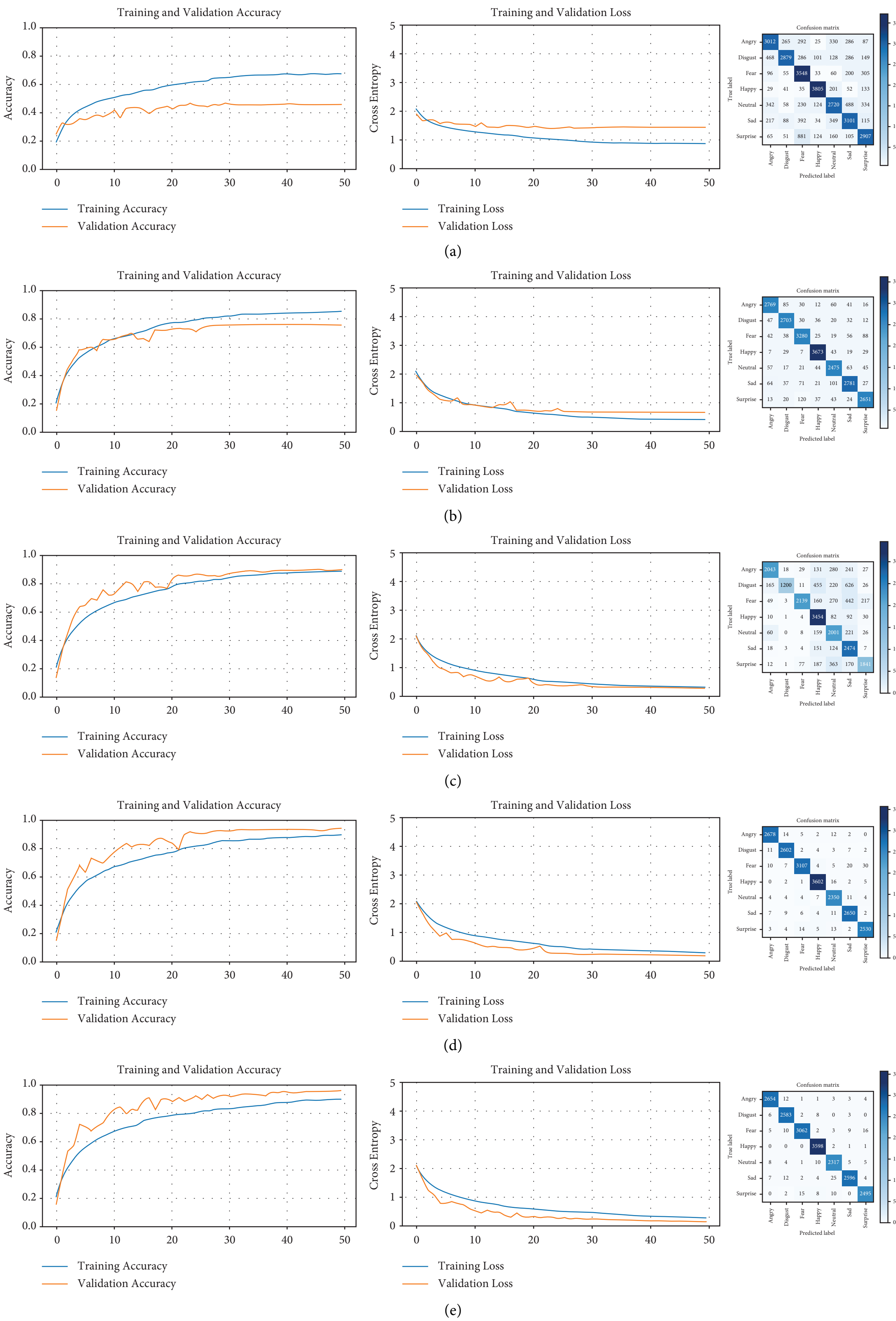
Figure 14: Comparison between the original and reclassified NHFI dataset. (a) NHFI dataset (original) and (b) NHFI dataset (reclassified).

Table 10: Comparison between the original and reclassified NHFI dataset.

Dataset	Images (train)	Images (val)	Total	Train_acc	Val_acc
NHFI (original)	4278	1072	5350	0.7676	0.5597
NHFI (reclassified)	4278	1072	5350	0.9550	0.7044

Table 11: Balanced distribution of the AffectNet dataset.

Subset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
Train set	3800	3800	3800	3800	3800	3800	3800	26600
Val set	500	500	500	500	500	500	500	3500
Total	4300	4300	4300	4300	4300	4300	4300	30100



(a)

(b)

(c)

(d)

(e)

Figure 15: Learning curves and confusion matrices for five successive trainings of the AffectNet dataset. (a) Training #1, (b) training #2, (c) training #3, (d) training #4, and (e) training #5.

Table 12: Summary of results for the balanced AffectNet dataset.

Training	Images (train)	Images (val)	Total	Accuracy
1	26600	3500	30100	0.4686
2	17587	4393	21980	0.7612
3	16268	4064	20332	0.9016
4	15843	3956	19799	0.9401
5	15617	3902	19519	0.9590

Table 13: Distribution of the balanced and reclassified AffectNet dataset.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
AffectNet (balanced)	4300	4300	4300	4300	4300	4300	4300	30100
AffectNet (reclassified)	4394	3893	5004	4594	4224	4071	3920	30100

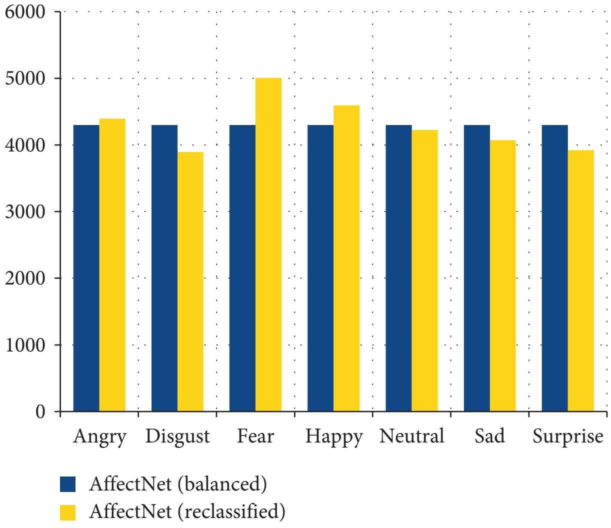
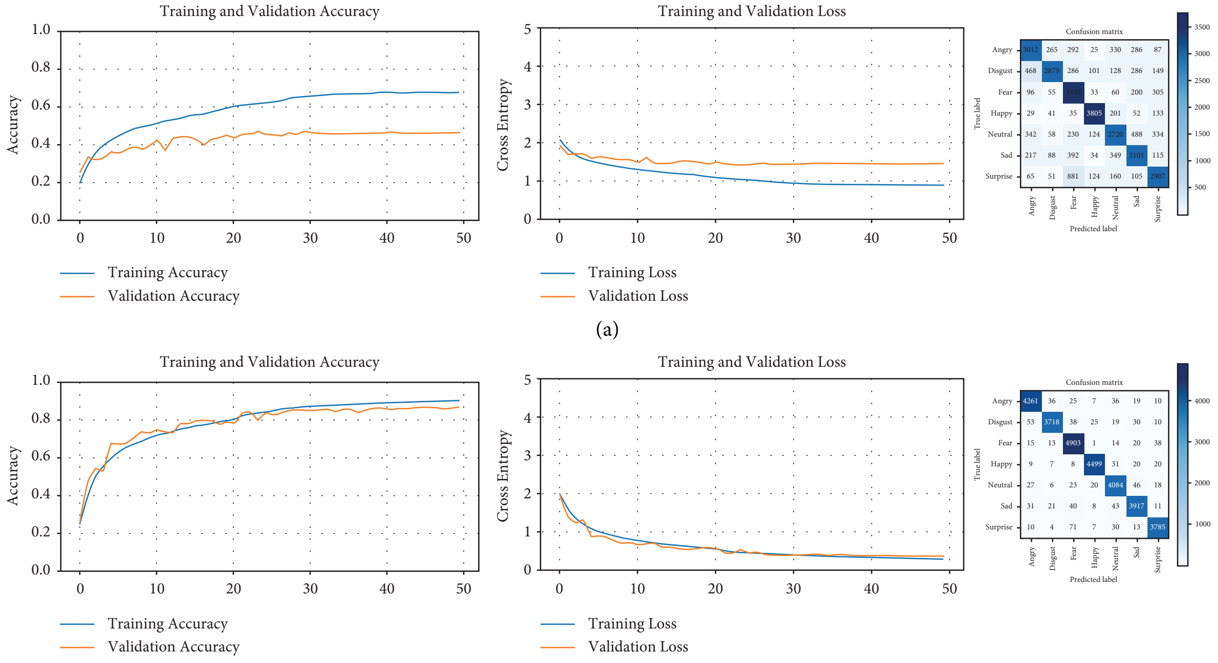


Figure 16: Graphical comparison of both distributions.



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	3012	265	292	25	330	286	87
Disgust	468	2879	286	101	128	286	149
Fear	96	55	3548	33	60	200	305
Happy	29	41	35	3805	201	52	133
Neutral	342	58	230	124	2720	488	334
Sad	217	88	392	34	349	3101	115
Surprise	65	51	881	124	160	105	2907

(a)

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	4261	36	25	7	36	19	10
Disgust	53	3718	38	25	19	30	10
Fear	15	13	4903	1	14	20	38
Happy	9	7	8	4499	31	20	20
Neutral	27	6	23	20	4084	46	18
Sad	31	21	40	8	43	3917	11
Surprise	10	4	71	7	30	13	3785

(b)

Figure 17: Comparison between the balanced and reclassified AffectNet dataset. (a) AffectNet dataset (balanced), (b) AffectNet dataset (reclassified).

Table 14: Comparison of the training results for the balanced and reclassified AffectNet datasets.

Dataset	Images (train)	Images (val)	Total	Train_acc	Val_acc
AffectNet (balanced)	26600	3500	30100	0.6763	0.4686
AffectNet (reclassified)	24084	6016	30100	0.9013	0.8652

Table 15: Distribution of the original and new AffectNet datasets.

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Total
AffectNet (original)	25382	4303	6878	134915	75374	25959	14590	287401
AffectNet (new)	30827	17475	19145	114275	52760	29160	23759	287401

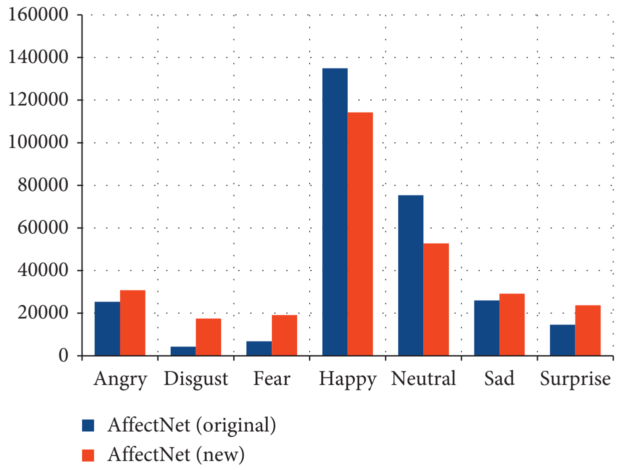


Figure 18: Graphical comparison of both distributions.

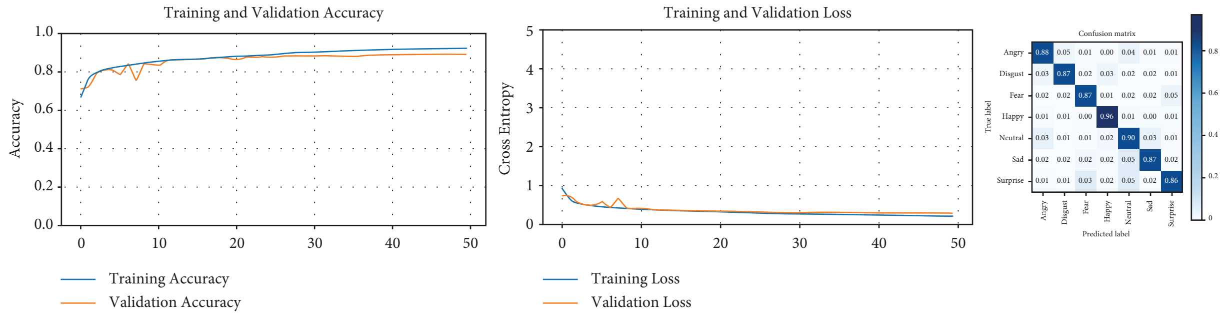


Figure 19: The learning curves and confusion matrix for the reclassified AffectNet dataset.

Table 16: Comparison of state-of-the-art performance on the FER datasets considered.

Dataset	Work	Model	Accuracy (%)
FER2013	[9]	VGG fine tuning	73.28
FER2013 (reclassified)	Ours	CNN from scratch	86.71
NHFI	Ours	EfficientNet-B0 transfer learning	55.97
NHFI (reclassified)	Ours	EfficientNet-B0 transfer learning	70.44
AffectNet	[30]	CNN-attention mechanism	65.69
AffectNet (reclassified)	Ours	CNN from scratch	89.17

7. Conclusions and Future Work

Facial expression recognition in the wild is a challenging problem for computer systems. Promising results have been achieved with deep learning methods, where the model and the data share responsibility. The vast majority of the research is oriented towards designing better models, which is not sufficient when the data suffers from drawbacks. One of the most influential problems in FER datasets is misclassification. In this work, we presented and implemented a method to reclassify all the facial images of a dataset by generating a new distribution that increases the accuracy of the FER models. The proposed method keeps the convolutional network fixed and iteratively improves the data over successive trainings. After each training, the dataset is evaluated with the confusion matrix, and the facial images corresponding to the correct predictions (on-diagonal) are selected to form the subsequent training data. This process gradually generates a more accurate model and more distinctive features for each category of facial expression. The model from the last training is used to reclassify all the images creating a new distribution of the dataset. We experimented with popular FER datasets and CNNs created from scratch and Transfer Learning. The increase in validation accuracy by 20.45%, 14.47%, and 39.66%, for FER2013, NHFI, and AffectNet, respectively, corroborates the efficacy of the proposed method. The results suggest that the quality and size of the dataset determine the most appropriate type of model. NHFI is a small and better-annotated dataset, so a pretrained model is convenient, unlike larger and lower quality datasets, which need a model from scratch, with longer training and more parameters. The reclassified versions of these datasets maintain the same number of images as the original dataset, but with less overlapping between categories, and less variability within the same category of facial expression. This allows us to achieve the state-of-the-art performance of single network FER models with 86.71%, 70.44%, and 89.17%, for FER2013, NHFI, and AffectNet, respectively. The recognition rates improved most significantly for the largest and lowest classified datasets, i.e., the proposed method works best for datasets with a high level of misclassified images. The refinement process of the dataset would enable several models to work well, not only diverse architectures of CNN, but others such as the transformer. Our proposal, beyond the application to the FER domain, is also useful for a variety of computer vision problems when the data are images. Furthermore, it can serve as a debugging tool in the automatic collection of image datasets. We maintained the size of the dataset, considering that quantity is important. However, there are irrelevant images that should be removed and the imbalance could be addressed with data augmentation or GANs. We believe that these contributions would improve the quality of the dataset and the accuracy of the models. Therefore, a methodology for automatic learning should consider the quality of the dataset as a prerequisite to the search for better network architectures and model configurations.

Data Availability

The code developed from this study is available in the GitHub repository (https://github.com/cimejia/FER-datasets/), whereas the datasets generated (except for the reclassified version of AffectNet) are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by grant CIPROM/2021/17 awarded by the Prometeo program from Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital of Generalitat Valenciana (Spain), and partially funded by the grant awarded by the Central University of Ecuador through budget certification no. 34 of March 25, 2022, for the development of the research project with code: DOCT-DI-2020-37.

References

[1] J. Yang, "2D+3D facial expression recognition via discriminative dynamic range enhancement and multi-scale learning," 2020, https://arxiv.org/abs/2011.08333.

[2] F. Ana Raquel, "A global perspective on an emotional learning model proposal," *Telematics and Informatics*, vol. 34, no. 6, pp. 824–837, 2017.

[3] M. Albert, "Communication without words," 1968.

[4] M. Gori, L. Schiatti, and M. Amadeo, "Masking emotions: face masks impair how we read emotions," in *Frontiers in Psychology*vol. 12, May 2021.

[5] H. Hwang and D. Matsumoto, "Functions of emotions," 2022, https://noba.to/w64szjxu.

[6] L.-F. Chen, "Emotion recognition and understanding for emotional human-robot interaction systems," *Emotion*, p. 354, Jan. 2021.

[7] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, p. 401, 2018.

[8] F. M. A. Mazen, A. A. Nashat, and R. A. A. A. A. Seoud, "Real time face expression recognition along with balanced FER2013 dataset using CycleGAN," *International Journal of Advanced Computer Science and Applications*, vol. 12, p. 6, 2021.

[9] Y. Khaireddin and Z. Chen, "Facial emotion recognition: state of the art performance on FER2013," 2021, https://arxiv.org/abs/2105.03588.

[10] Y. Wang, "Facial expression recognition based on random forest and convolutional neural network," *Information*, vol. 12, pp. 2078–2489, 2019, https://www.mdpi.com/2078-2489/10/12/375.

[11] J. Liu, H. Wang, and Y. Feng, "An end-to-end deep model with discriminative facial features for facial expression recognition," *IEEE Access*, vol. 9, pp. 12158–12166, 2021.

[12] A. Kandeel, "Facial expression recognition using a simplified convolutional neural network model," in *Proceedings of the ICCSPA 2020—4th International Conference on Communications, Signal Processing, and their Applications*, p. 57, Sharjah, UAE, March 2021.

[13] F. Z. Canal, T. R. Muller, J. C. Matias et al., "A survey on facial emotion recognition techniques: a state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022, https://www.sciencedirect.com/science/article/pii/S0020025521010136.

[14] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, pp. 124–129, 1971.

[15] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: a database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, 2019.

[17] Y. Wang, W. Song, W. Tao et al., "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 8, pp. 19–52, 20.

[18] L. Zho, “Discriminative attention-augmented feature learning for facial expression recognition in the wild,” in *Neural Computing and Applications*, pp. 1–12, 2021.

[19] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: state of the art,” 2016, https://arxiv.org/abs/1612.02903.

[20] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: a survey of registration, representation, and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[21] M. Harl, “A light in the dark: deep learning practices for industrial computer vision,” 2022, https://arxiv.org/abs/2201.02028.

[22] J. H. Kim, A. Poulose, and D. S. Han, “The extensive usage of the facial image threshing machine for facial emotion recognition performance,” *Sensors*, vol. 21, pp. 2026–8220, 6.

[23] D. Y. Kim and C. Wallraven, “Label quality in AffectNet: results of crowd-based re-annotation,” in *Asian Conference on Pattern Recognition*, pp. 518–531, Springer, 2022.

[24] I. J. Goodfellow, “Challenges in representation learning: a report on three machine learning contests,” in *International Conference on Neural Information Processing*, pp. 117–124, Springer, 2013.

[25] Z. He, “Bayesian based facial expression recognition transformer model in uncertainty,” in *2021 International Conference on Digital Society and Intelligent Systems*, pp. 157–161, 2021.

[26] S. Vaidya, “Detecting human emotions - facial expression recognition,” 2020, https://sudarshanvaidya.medium.com/detecting-human-emotions-facial-expression-recognition-ebf98fdf87a1.

[27] A. Bhalla, “Facial expression recognition,” 2020, https://www.kaggle.com/code/bhallaakshit/facial-expression-recognition/.

[28] M. Tan and V. Quoc, “EfficientNet: Rethinking model scaling for convolutional neural networks,” 2019, https://arxiv.org/abs/1905.11946.

[29] S. Haghighi, M. Jasemi, S. Hessabi, and A. Zolanvari, “PyCM: multiclass confusion matrix library in Python,” *Journal of Open Source Software*, vol. 3, no. 25, p. 729, 2018.

[30] Z. Wen, “Distract your attention: multi-head cross attention network for facial expression recognition,” 2021.

Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 1034983, 12 pages
https://doi.org/10.1155/2022/1034983

Research Article

Forecasting Unplanned Purchase Behavior under Buy-One Get-One-Free Promotions Using Functional Near-Infrared Spectroscopy

SuJin Bak,[1] **Minsun Yeu**,[2] **and Jichai Jeong**[1]

[1]*Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea*
[2]*College of Business Administration, University of Ulsan, Ulsan 44610, Republic of Korea*

Correspondence should be addressed to Minsun Yeu; minsunyeu@ulsan.ac.kr and Jichai Jeong; jcj@korea.ac.kr

Received 28 June 2022; Revised 4 October 2022; Accepted 8 October 2022; Published 7 November 2022

Academic Editor: Jesus Peral

It is very important for consumers to recognize their wrong shopping habits such as unplanned purchase behavior (UPB). The traditional methods used for measuring the UPB in qualitative and quantitative studies have some drawbacks because of human perception and memory. We proposed a UPB identification methodology applied with the brain-computer interface technique using a support vector machine (SVM) along with a functional near-infrared spectroscopy (fNIRS). Hemodynamic signals and behavioral data were collected from 33 subjects by performing Task 1 which included the Buy-One-Get-One-Free (BOGOF) and Task 2 which excluded the BOGOF condition. The acquired data were calculated with 6 time-domain features and then classified them using SVM with 10-cross validations. Thereafter, we evaluated whether the results were reliable using the area under the receiver operating characteristic curve (AUC). As a result, we achieved average accuracy greater than 94%, which is reliable because of the AUC values above 0.97. We found that the UPB brain activity was more relevant to Task 1 with the BOGOF condition than with Task 2 in the prefrontal cortex. UPBs were sufficiently derived from self-reported measurement, indicating that the subjects perceived increased impulsivity in the BOGOF condition. Therefore, this study improves the detection and understanding of UPB as a path for a computer-aided detection perspective for rating the severity of UPBs.

1. Introduction

Consumers have experienced financial problems such as excessive consumption, household debt, and monetary losses because of unplanned purchases [1]. Many studies have shown that reasonable consumption is difficult because unplanned purchase behavior (UPB) occurs emotionally or impulsively [2]. Many studies have reported that UPB can occur under situations that encourage people's impulsiveness such as price discounts and time pressures [3, 4]. UPBs are defined as a purchase of any item that consumers had not planned to purchase before entering the shops [5]. UPBs are increased by promotion strategies [6] such as price discounts, coupons, and money-back guarantee. Especially, "Buy-One-Get-One-Free" (BOGOF) is one of the most popular promotion strategies. A previous study found that over 53.3% of 192 respondents preferred BOGOF over other promotions [7]. This promotion strategy plays an important role in eliciting consumer's UPBs.

To assess whether consumers' UPB is, there are traditional research methods such as interviews, surveys, and questionnaires [8]. However, they rely on consumers' subjective perceptions and memories [9]. Furthermore, there is still a lack of tools and equipment for empirically measuring UPBs. To solve this issue, there are recent studies that have reported empirical evidence for unplanned purchases through brain signal measuring equipment [10]. There are noninvasive equipment for brain measurements such as electroencephalogram (EEG) [11, 12] and functional magnetic resonance imaging (fMRI) [13]. Figure 1 describes the noninvasive brain signal measurement equipment that is harmless to the human body. EEG records voltage

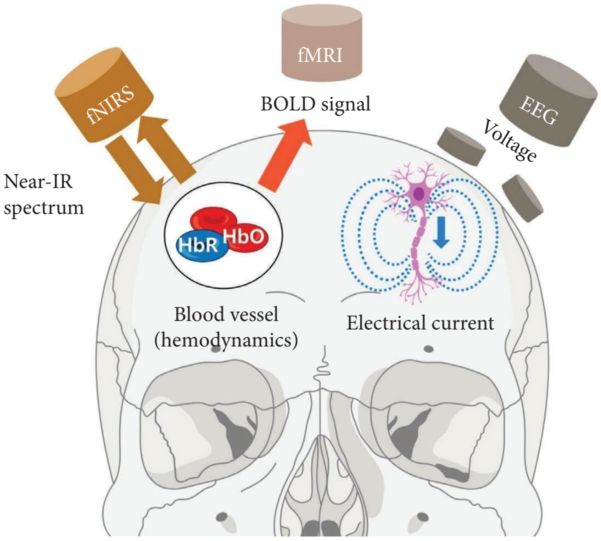


Figure 1: Noninvasive mapping of brain function using neuroimaging technologies. EEG records voltage fluctuations caused by electrical currents flowing through the brain because of active neurons using an array of electrodes placed on the scalp. fMRI measures hemodynamic responses associated with brain activity by relying primarily on the local blood-oxygen-level-dependent (BOLD) signal, which detects changes in blood oxygenation caused by neural activity. fNIRS measures the changes in oxygenation and deoxygenation hemoglobin concentrations in the brain using near-infrared light.

fluctuations caused by electrical currents flowing through the brain cortex because of neural activity [14]. fMRI uses the blood-oxygen-level-dependent (BOLD) contrast to detect changes in blood oxygenation that occur in response to neural activity, and it has become the most common method for imaging brain functions in vivo [15]. However, fMRI is unsuitable for certain research applications and various clinical applications because fMRI is physically prone to mixing motion artifacts, exposes to loud noises, and is expensive. To compensate for the shortcomings of fMRI, fNIRS has become a promising imaging modality for UPB evaluation as well as for reducing the physical space constraint and costs of fMRIs [16, 17]. fNIRS is one of the state-of-the-art brain signal measurement equipment, especially with scalability, convenience in use, and portability [18, 19]. With the benefits of the fNIRS, we can provide valuable insights into the consumers' UPB by using the brain-computer interface (BCI) technique [20]. We utilize the BCI technique to explain customer behaviors in detail [21], and it will benefit to neuromarketing industries using fNIRS utilities [22].

A general scheme of BCI can be explained using five main steps, as illustrated in Figure 2. In Step 1, the people should perform cognitive tasks with (or without) the BOGOF condition. The brain signal changes according to the cognitive tasks and these brain signals are acquired by an fNIRS device and then gets transmitted to the next step. In Step 2, the fNIRS signals are digitized, amplified, and filtered to delete undesired signals called artifacts such as physiological noise. Then, the clean signals go to Step 3, where the clean signals extract features to be used as a descriptor of the fNIRS signals for classifying the UPB patterns. The features are classified into UPB and non-UPB states in Step 4. Finally, the message which indicates the UPB classification result is presented on a computer screen in Step 5. Through these processes, this study shows that the UPB and non-UPB can be distinguished because of brain signals that reflect people's actual cognitive consequences.

Accordingly, we proposed a methodology to measure the UPB by the BCI technique. For this purpose, we acquired fNIRS signals during the cognitive tasks with and without BOGOF at online shopping shops and then converted them into preprocessed feature vectors by six time-domain feature extraction methods. To classify UPBs and non-UPBs, we adopted SVM, which is a widely used supervised learning approach with 10-fold cross-validation. We also used the "area under the receiver operating characteristic (AUC)," a measurement method for determining whether the results of SVM classifying UPB and non-UPB were reliable. As a result, we achieved an average accuracy of above 94% for classifying UPBs across all subjects, which also ensured the reliability of the results by obtaining an AUC value above 0.97. We observed that low brain activities were exhibited during Task 1 which included the BOGOF condition, and high brain activities were exhibited during Task 2 which excluded the BOGOF condition at the PFC. It is interpreted that there is a clear difference in the fNIRS signals depending on the BOGOF conditions. Furthermore, our experimental tasks are well designed because self-reported results indicate that these experimental tasks sufficiently induced the consumers' impulsiveness. Therefore, we believe that this study can be applied to a variety of applications by improving the accuracy of detecting UPB patterns under BOGOF conditions.

2. Materials and Methods

2.1. Subject. The study was approved by the Korea University Institutional Review Board (KUIRB-2022-0126–01) and then written informed consent was obtained from all subjects. Considering possible dropouts, we recruited 38 healthy adults but 5 people were excluded because of insufficient signal quality, and the remaining 33 subjects (mean ± standard deviation aged 24 ± 2.64 years) completed the entire study. There were 12 males (aged 24 ± 1.75 years) and 21 females (aged 24 ± 3.03 years) with normal or corrected to normal eyesight. All subjects were right-handed to minimize variability in brain signals. The subjects had no previous history of physical, mental, or psychological disabilities. All subjects were asked to minimize head movements and actively take part in the experiment as much as possible.

2.2. Experimental Procedures. To investigate brain activation patterns, we designed two experimental tasks depending on the presence or absence of BOGOF. Figure 3 illustrates the overall experiment protocol. Each experimental task comprised 5 trials where the subjects were asked to decide on

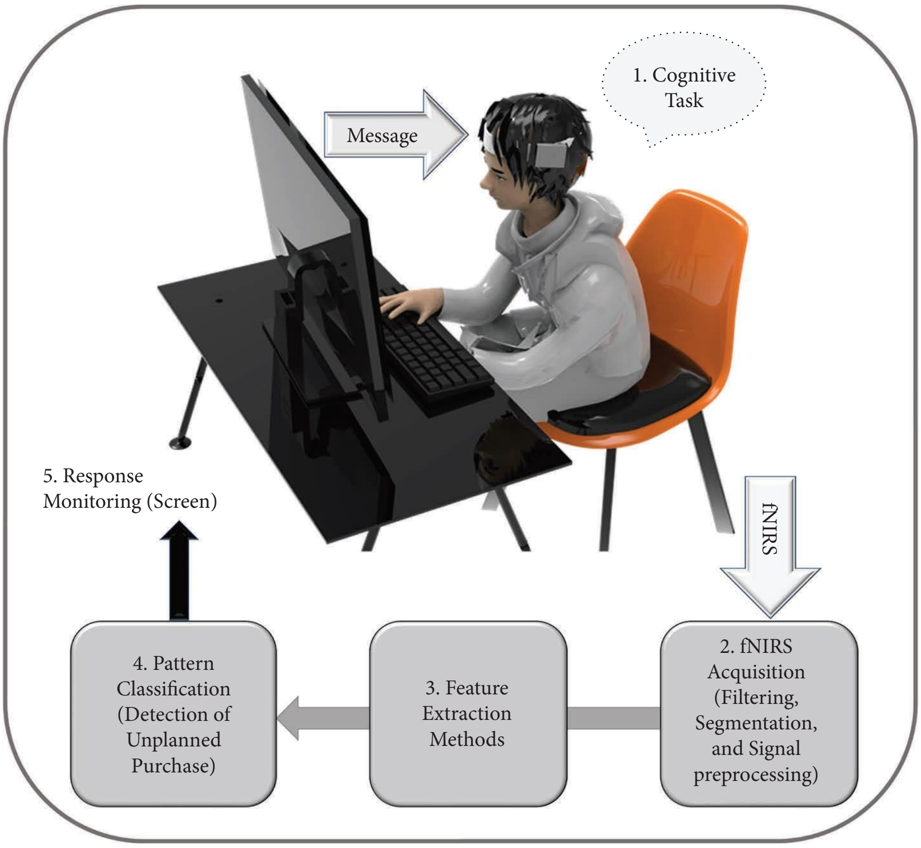

1. Cognitive Task
Message
fNIRS
5. Response Monitoring (Screen)
4. Pattern Classification (Detection of Unplanned Purchase)
3. Feature Extraction Methods
2. fNIRS Acquisition (Filtering, Segmentation, and Signal preprocessing)

Figure 2: Typical scheme of a BCI system consisting of five main stages. The process is as follows: (1) cognitive task; (2) fNIRS acquisition; (3) feature extraction methods; (4) pattern classification; (5) response monitoring. We went through this whole process to detect UPB.

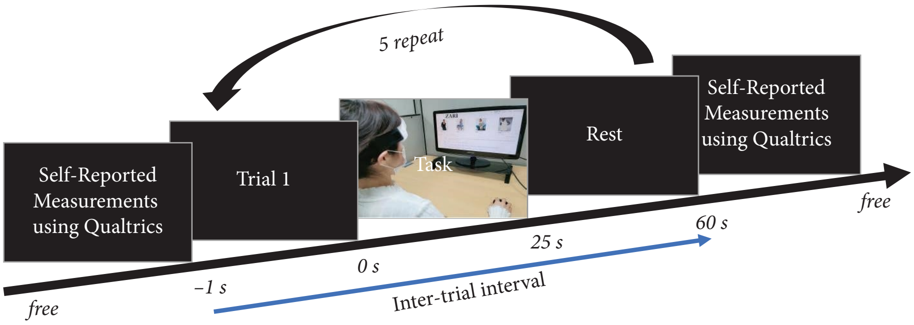

5 repeat
Self-Reported Measurements using Qualtrics
Trial 1
Task
Rest
Self-Reported Measurements using Qualtrics
free
−1 s
0 s
25 s
60 s
free
Inter-trial interval

Figure 3: Overall procedures of performing Tasks 1 and 2. Participants are free to choose what clothes they want to buy from online shopping malls, which are divided into Task 1 with BOGOF conditions and Task 2 without BOGOF conditions. The whole experiment consists of trial number mark, experimental task, and rest. Each task is conducted for 25 s followed by a rest for 30 s. A total of five trials were conducted. The tasks are randomized and counterbalanced in sequence. Before and after the experiment, a self-reported measurement related to UPB was conducted using the popular survey platform, Qualtrics.

purchasing displayed products under Task 1 (including BOGOF condition) and Task 2 (excluding BOGOF condition). In each task, the participants are free to choose the clothes they want [23]. The selected clothes brand is ZARA [24], known as the global specialty retailer of private label apparel fashion brand, which was selected by a Google survey in Koreapas [25], which is one of the major online communities for Korean university students. The clothing lines are divided into 5 major categories, which were displayed on the screen in the following order: knitwear, coat, vest, pants, and suit. There were 4 products in each group. For example, the coat group has four products (i.e., wool coat, classic long trench coat, wool mannish coat, and checked coat). The subjects can freely purchase up to 4 products they want in each group; in other words, they can purchase products from 0 to 20 per task. Each product was presented only once during the experiment. Each task lasted for 5 minutes and included the following stages (trial number display (1 s), task (25 s), rest (30 s)), and a brief buzzer (58.4 dBA, sound level meter, YATO, China). The order of the tasks was randomized and counterbalanced. All the subjects completed a self-reported measurement using Qualtrics, which is a popular survey platform, before and after the experiment.

2.3. fNIRS Equipment. To measure the brain's hemodynamic responses in the prefrontal cortex (PFC), we used an fNIRS device (NIRSIT Lite, OBELAB Inc., Korea). To provide detailed guidance on the specifications of fNIRS, Figure 4 illustrates the fNIRS channel configuration covering the forehead and an example of source-detector pairs in detail. On the left panel, the fNIRS device has a total of 15 fNIRS channels composed of 5 sources (the grey circles) and 7 detectors (the orange circles). The probe sets are symmetrically arranged at FPz between Chs. 7 and 10 according to the 10–20 international systems. They were divided into 4 regions: the Dorsolateral prefrontal cortex (DLPFC), Ventrolateral prefrontal cortex (VLPFC), Medial prefrontal cortex (mPFC), and Orbitofrontal cortex (OFC). In these four areas, brain activations are known to primarily inhibit impulsivity [26]. Among them, VLPFC has separate left and right functions, and the left VLPFC is related to the reward system [27]. These results can help interpret hemodynamic activation patterns in the PFC that occur when performing our experimental tasks. On the right panel, the detector measures the lights from a diffuse volume of tissue in accordance with the model of light propagation. These lights can reach 8 mm into the brain cortex while maintaining a distance of 3 cm between the source and detector. The fNIRS device reflects the absorption properties of living tissues to measure changes in the local concentrations of oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) within the crescent-shaped near-infrared region through the skull [28, 29]. The crescent-shaped paths represent the near-infrared light (NIR) photons' traveling area, while the blue dotted arrows represent light scattering. The red-colored arrows show the distance traveled by photons, which is corrected by the differential path length factor. Consequently, fNIRS can measure the hemodynamic changes quantitatively by absorbing near-infrared rays into the scalp and by measuring the emitted light emitted.

Based on the aforementioned principles of fNIRS, we recorded the optical density data at a frequency of 8.138 Hz and configured it to detect hemodynamic activity at wavelengths of 780 nm and 850 nm. The optical density data were bandpass filtered digitally in the range of 0.01–0.1 Hz to eliminate possible physiological signals such as respiration, heart rate, and unwanted noise. Filtered signals were converted to oxygenated and deoxygenated concentration changes using the modified Beer–Lambert law [30], and then the data were segmented into epochs ranging from −1 to 60 s relative to the task onset (0 s). The epoch was subjected to a baseline correction to subtract the mean value within a reference interval from −1 to 0 s. The temporal means of the fNIRS data in each channel were calculated by averaging the fNIRS data from the start to the end time (0–60 s) in each epoch. In this study, we handle only HbO signals because they have a higher signal-to-noise ratio [31, 32]. It means that the signal strength is stronger than the noise intensity. HbO is also regarded as a more reliable indicator for analyzing the PFC activation [33]. The acquired fNIRS dataset, as well as all related information, can be downloaded from https://github.com/SujinBak/BOGOF.

2.4. Extraction of Six Time-Domain Features: Mean, Variance, Kurtosis, Skewness, Slope, and Area. Feature extraction is an important step in extracting and maximizing the information that describes the unique property of the fNIRS signals. This step forms the features extracted from the brain signals into vectors. These feature vectors are recognized by the classifier, which makes it easier to classify two or more classes. The widely used feature extraction methods were divided into three main categories: time-domain analysis; frequency-domain analysis; and time-frequency domain analysis. We focused on time-domain analysis, which facilitates the understanding of the transient characteristics of physiological signals, including fNIRS signals [34]. It has been reported that time-domain features can improve the classification accuracy between different cognitive states [35]. Especially, the time-domain features represent the property difference between the measured signals, which is visually recognizable when an unexpected abnormality appears in the signals [36]. Accordingly, we adopted the framework for feature extraction used by Park and Dong. [37] and then calculated 6 time-domain features (mean, variance, kurtosis, skewness, slop, and area) to extract information across data [38]. We denote mean, variance, kurtosis, skewness, slope, and area as SM, SV, KR, SK, SS, and SA, respectively. Signal mean (SM) can be calculated by the following equation:

$$\mathrm{SM} = \frac{1}{N} \sum_{n=0}^{N-1} x[n], \tag{1}$$

where $x[n]$ is the input signal (ΔHbO) at the time index of n, and N is the total length of the signals. Signal variance (SV) is calculated as follows:

$$\mathrm{SV} = \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \mu)^2, \tag{2}$$

where $\mu (= \mathrm{SM})$ is the mean found from (1). For signal kurtosis (KR), it is calculated by the following equation:

$$\mathrm{KR} = E\left[\left(\frac{x[n] - \mu}{\sigma}\right)^4\right], \tag{3}$$

where E is the expected value and $\sigma (= \mathrm{SV})$ is the standard deviation. Similarly, signal skewness (SK) is the asymmetry of values relative to normal distribution around the mean, hence calculated in the following equation:

$$\mathrm{SK} = E\left[\left(\frac{x[n] - \mu}{\sigma}\right)^3\right]. \tag{4}$$

Signal slope (SS) is calculated by the following equation:

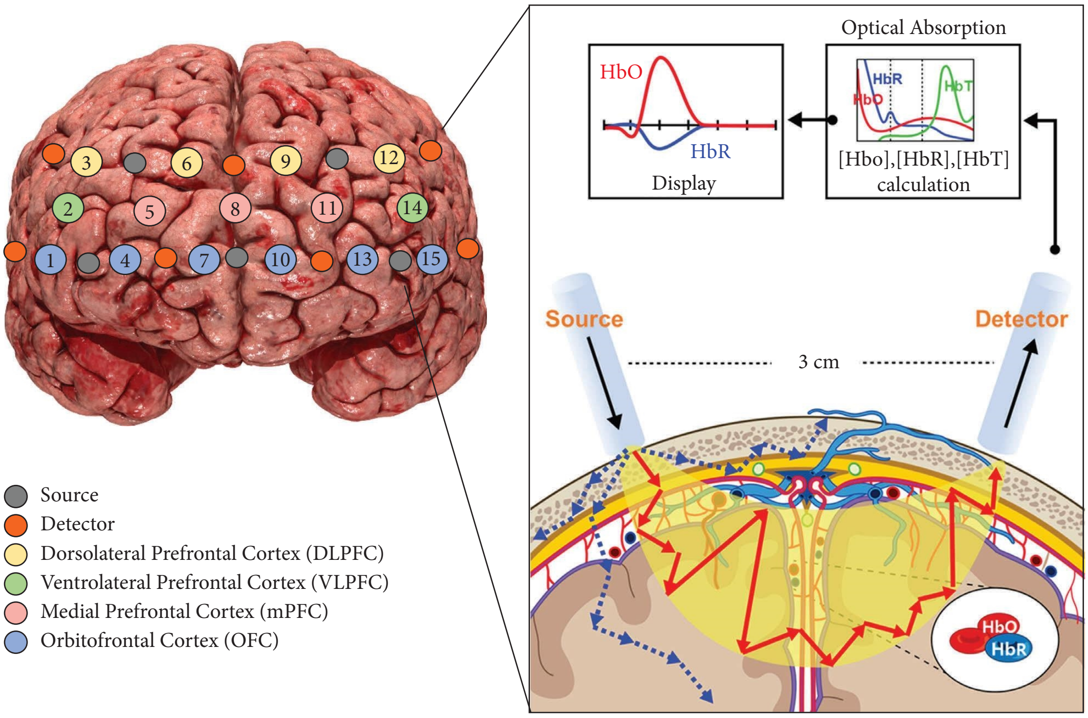

3
6
9
12
2
5
8
11
14
1
4
7
10
13
15
Source
Detector
Dorsolateral Prefrontal Cortex (DLPFC)
Ventrolateral Prefrontal Cortex (VLPFC)
Medial Prefrontal Cortex (mPFC)
Orbitofrontal Cortex (OFC)
Optical Absorption
HbO
HbR
Display
HbR
HbO
HbT
[Hbo],[HbR],[HbT]
calculation
Source
Detector
3 cm
HbO
HbR

Figure 4: fNIRS channel configuration (a) and source-detector pair (b). In the left panel, the grey circles indicate the sources, whereas the orange circles represent the detectors, resulting in a total of 15 fNIRS channels in the prefrontal cortex (PFC). According to the 10–20 international system, the probe sets are symmetrically placed at FPz between Chs. 7 and 10. On the right panel, the source-detector pair measures lights from a diffuse volume of tissue beneath the pair as shown in the model of light propagation. These lights can reach approximately 8 mm into the brain cortex at a source-detector spacing of 3 cm. Lights at two wavelengths (780 nm and 850 nm) are used to reconstruct changes in oxy- and deoxy-hemoglobin concentrations from the modified Beer–Lambert law. A detector captures the lights resulting from the interaction with HbO and HbR, following a crescent-shaped path back to the surface of the skin. The crescent-shaped paths depict the traveling area of the near-infrared light (NIR) photons, while the blue dotted arrows indicate the light scattering. The red-colored arrows show the extra distance traveled by photons, which is corrected by the differential path length factor.

$$\mathrm{SS} = \frac{x[n] - x[n-1]}{\Delta n}, \tag{5}$$

where $x[n]$ is the xvalue at the current time, and $x[n\text{-}1]$ is the x value at the previous time. Δn is the sampling time interval. Moreover, signal area (SA) is obtained by the following integral function expression:

$$\mathrm{SA} = \sum_{n=0}^{N-1} x[n]\Delta n. \tag{6}$$

All statistical features were rescaled between 0 and 1 to normalize the size of the extracted feature vector using the following equation:

$$Z' = \frac{Z - \min(Z)}{\max(Z) - \min\ (Z)}, \tag{7}$$

where Z' is the rescaled feature vector and Z refers to the original feature vector.

2.5. SVM for Detecting UPB Patterns. The most popular supervised learning model, SVM, has already demonstrated its excellent performance (i.e., classification accuracy) compared to other classifier models in many studies [39–41]. SVM can explicitly control errors by maximizing margins between two or more classes, known as support vectors [42, 43], as illustrated in Figure 5. The green squares and the pink circles are support vectors. Thus, SVM estimates the optimal hyperplane with the maximum margin (2l) between two classes. The optimal hyperplane is defined as follows:

$$d(x) = W^T x + b = 0, \tag{8}$$

where x represents the input values and becomes $x = (x[0], x[1], \ldots, x[N-1])^T$. W refers to the hyperplane's direction as a normal vector of the hyperplane and transposes it to W^T. b is the position. The optimal hyperplane is determined through W and b. To calculate W and b, the margin is defined as the distance between the nearest data points of either class measured perpendicular to the hyperplane. This means maximizing margins while minimizing generalized errors. To reduce errors, we calculated $2l$ by substituting W vectors obtained from (8) into (9). Ultimately, we can calculate an optimized hyperplane that maximizes margins between support vectors, which is important to determine the classification accuracy of SVM as follows:

$$\text{Maximum margin}\,(2l) = \max_{W,b} \frac{2}{\|W\|_2}, \tag{9}$$

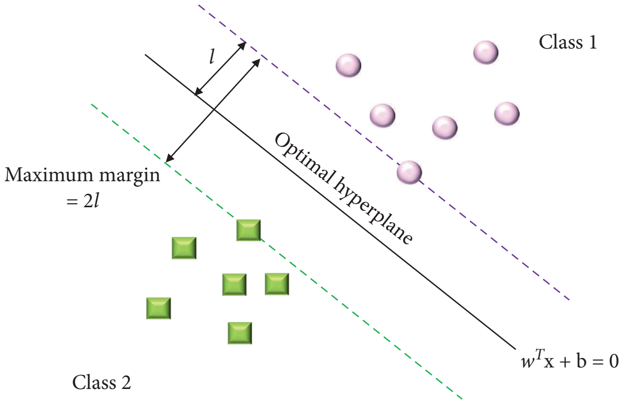


Figure 5: Concept of support vector machine (SVM). The green squares and the pink circles are support vectors. SVM should find the optimal hyperplane (solid black line) divided into Class 1 and Class 2 with a maximum margin (2*l*). Hence, SVM estimates the hyperplane in the two-dimensional space and classifies two classes.

where $\|W\|_2 = \sqrt{W[0]^2 + W[1]^2 + \ldots + W[N-1]^2}$. The SVM model was assessed by 10-fold cross-validation to avoid overfitting known as learning biases caused by the classifier's excessive dependency on training data. The training dataset is split into 10-folds containing an equal number of the training dataset. We divided them into the ratio of 8 train sets and 2 test sets for the cross-validation and then tested them 30 times to estimate the variability of the classification accuracies. We subsequently calculated the mean classification accuracy and a standard error of the mean (SEM).

However, this model can lead to an imbalance problem where one of the two classes has more data than the other classes [44]. Thus, we evaluated the reliability of the classification results in the following section to determine whether the classifier results were affected by the imbalance problem.

2.6. Reliability of Calculated Classification Results Using AUC. AUC is primarily used to validate the reliability of the results classified by the SVM [45–47]. Figure 6 depicts the typical receiver operating characteristic (ROC) curves and their AUCs which include a true-positive rate (TPR) and false-positive rate (FPR). These statistical indexes such as TPR and FPR are essential for interpreting the reliability of the calculated classification results. AUC estimates the whole two-dimensional area underneath the whole ROC curve (i.e., a kind of integral calculation) from (0,0) to (1,1). Hence, AUC is the range from 0 to 1, and the classification results are the most reliable with an AUC value of 1. The reliability of the results calculated by the classifiers is better as the FPR is lower and TPR is higher. In other words, the closer the AUC is to 1, the better the reliability of the results. If the AUC area is less than 0.5, the calculated classification results are not reliable. After all, it is important to find the largest AUC (close to 1). To quantify the AUC value, we first calculated the TPR and FPR using equations (10) and (11).

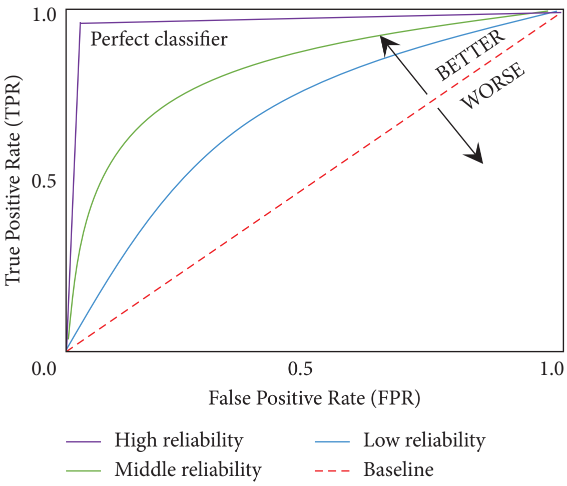


Figure 6: Typical ROC curves and their AUCs. The ROC curves show the relationship between true-positive rate (TPR) and false-positive rate (FPR) to validate the reliability of results calculated by the SVM classifier. TPR is the ratio of correctly judging the UPB elicited by BOGOF as the UPB. In contrast, FPR is the ratio of incorrectly judging the non-UPB as UPB. Through these TPRs and FPRs, the AUC can be calculated by quantifying the entire 2-D region under the ROC curve. The red dotted diagonal line (AUC = 0.5) depicts a baseline. The AUC of the purple line (AUC = 1.0) is considered to be the best reliability in the tested classifier results.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad (10)$$

where TP refers to the parameter in which the UPB is correctly classified as UPB, and FN indicates that the classifier incorrectly classified UPBs as non-UPBs. Thus, the TPR is the ratio of correctly judging the UPBs evoked by BOGOF as UPBs. In contrast, FPR is the ratio of incorrectly judging the non-UPB as the UPB.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \quad (11)$$

where TN represents that the classifier correctly classified non-UPBs as non-UPBs. FP means that non-UPB is misclassified as UPB.

2.7. Detection of UPB and Non-UPB. After completing the SVM classification process, MATLAB® App Designer, a fully integrated development environment, was used to represent UPB classification results by SVM on the computer screen. It is divided into two messages based on the classification accuracy between UPBs and non-UPBs. If the classification accuracy exceeds 80%, the message appears that it is ready to detect UPB patterns; otherwise, the message appears that it cannot detect UPB patterns.

2.8. Self-Reported Measurement. In this study, self-reported measurement was used to determine whether impulsivity increased in the BOGOF condition as perceived by the subjects. The subjects answered whether impulsivity was induced in each of the purchase situations of the two tasks. The subjects perceived that impulsivity was induced when BOGOF was present (compared to the absence), suggesting that the experiment was well designed.

2.9. Statistical Analyses. All statistical analyses were conducted using the Statistical Package for the Social Sciences, version 25.0. (SPSS Inc., Chicago, IL). Variables were calculated such as normality, means (μ), standard deviation (σ), and standard error of the mean (SEM) for each task. We used an independent sample t-test [48] to compare the difference in the number of clothes purchased by the subjects between Task 1 and Task 2, statistically.

2.10. Behavioral Analyses. People who overspend are more likely to make unplanned purchases [49]. Thus, we concentrated on the number of clothes that the subjects intended to buy to investigate the subject's UPB caused by BOGOF condition. The unplanned purchase ratio was calculated using the average and the sum of the clothing purchased by each subject. Thereafter, the independent sample t-test is used to investigate a statistical difference in the number of clothing purchased by the subjects.

3. Results

3.1. Self-Reported Results. A self-reported measurement method was used to determine whether there was a difference in the subjects' perceived impulsivity with and without BOGOF. Many researchers use an alternative variable, a compulsive desire to buy, to measure UPBs. Three items were used to measure the subjects' compulsive desire to buy [50], made on a 5-Likert scale.

There was a significant difference in the perceived impulse purchase intention according to the presence or absence of BOGOF. Subjects perceived impulsivity in the BOGOF condition ($\mu = 3.53$; $\sigma = 0.80$; t(64) = 5.375; $^{***}p < 0.001$) than in the non-BOGOF condition ($\mu = 2.40$; $\sigma = 0.90$; *N.S.*). It also suggests that our experiment tasks are well designed to compare subjects' impulsiveness and nonimpulsiveness.

3.2. Behavioral Results. The behavioral results were obtained from solely the number of clothing purchased by the subjects. The average number of the purchased clothes in Task 1 ($\mu \pm \sigma$: 6.67 ± 2.27) was higher than that in Task 2 (3.36 ± 2.41). An independent sample t-test shows statistically significant differences in the number of clothes purchased between the two tasks ($t = 5.649, ^{***}p < .001$), indicating that there is a difference in the UPB pattern between the two tasks. Specifically, in Task 1, the purchased clothes have the sum and standard deviation as follows: knitwear (43 ± 0.60), coat (43 ± 0.70), vest (45 ± 0.66), pants (51 ± 0.76), and suit (38 ± 0.52). Task 2 includes knitwear (31 ± 0.57), coat (19 ± 0.66), vest (25 ± 0.65), pants (24 ± 0.90), and suit (12 ± 0.49). Hence, the difference in the total number of clothes purchased between the tasks were knitwear ($t = 2.472, ^{*}p < 0.05$), coat ($t = 4.386, ^{***}p < 0.001$), vest ($t = 3.742, ^{***}p < 0.001$), pants ($t = 3.890, ^{***}p < 0.001$), and suit $t = 6.425, ^{***}p < 0.001$). This means that the total number in each clothing group can be revealed between the two tasks.

3.3. Analyses of Brain HbO Activity in PFC. We investigated the differences in the presence or absence of UPBs in connection to brain activity. Figure 7 depicts the topographical maps of averaged HbO activities across all subjects in the PFC areas, and Figures 7(a) and 7(b) correspond to Task 1 (including BOGOF) and Task 2 (excluding BOGOF), respectively.

Except for the left VLPFC, brain activation hardly occurred in Figure 7(a). In contrast, Figure 7(b) indicates the significant brain activations in several regions such as OFC, mPFC, and VLPFC regions, which showed particularly strong activations in the OFC area. Although the DLPFC showed little activation, significant brain activations occurred in the OFC, mPFC, and VLPFC areas, which are known to inhibit impulsivity. As a result, we revealed that Task 2 allows for reasonable consumption as opposed to Task 1.

3.4. Classification Results between UPB and Non-UPB Using SVM. We used SVM to calculate the accuracies for binary classification between Task 1, which elicits UPBs by BOGOF, and Task 2, which serves as a control task. Figure 8 exhibits the classification accuracies between UPB and non-UPB using SVM for each subject during cognitive tasks in accordance with the BOGOF. Especially, "A" on the x-axis represents the overall average classification accuracy of 94.23% for 33 subjects. The error bars represent SEMs, and the average error bar of "A" is 0.03. All subjects accounted for higher than 86% classification accuracy, which ranged from 86.42% ± 0.02 (accuracy (%) ± SEM) to (99.90% ± 0.01). These provide empirical evidence for differentiating UPBs from non-UPBs.

3.5. Reliability Verification of Classified Results Using AUC. AUC is used to determine the reliability of the classification results, which gives us an intuitive view of the entire spectrum of FPR (x-axis) and TPR (y-axis). Table 1 presents the AUC values of all subjects who participated in this experiment. The averaged AUC value is 0.97 across all subjects. Moreover, their AUCs lie between 0.85 and 1.00, indicating that the SVM model is trained perfectly, and their results are highly reliable. More specifically, Figure 9 illustrated the ROCs and their AUCs of two representative subjects with the lowest and highest AUCs among all subjects. Figure 9(a) refers to the subject's ROC and AUC (0.85) with the lowest accuracy value of 86.42%, and Figure 9(b) illustrated the subject's ROC and AUC (1.00) with the highest accuracy value of 99.90%. As a result, the curves are located above the baseline in both subjects, and their AUC values fully guarantee the reliability of the UPB detection results. In simple words, the larger the AUC value the higher is the reliability of the SVM results, in

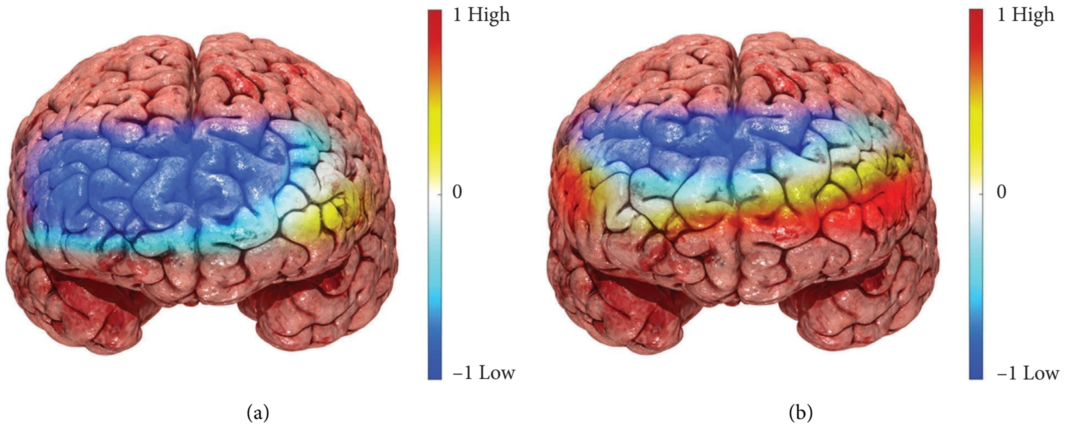


Figure 7: Topographical maps of averaged HbO activities under (a) Task 1 and (b) Task 2. Most areas show little activation except for the left VLPFC in (a). On the other hand, extensive brain activations appear in mPFC, VLPFC, and OFC in (b). These regions have the function of inhibiting UPB as an important predictor of impulsiveness. Thus, these findings provide empirical evidence that the BOGOF condition sufficiently encourages UPB to differentiate it from non-UPB.

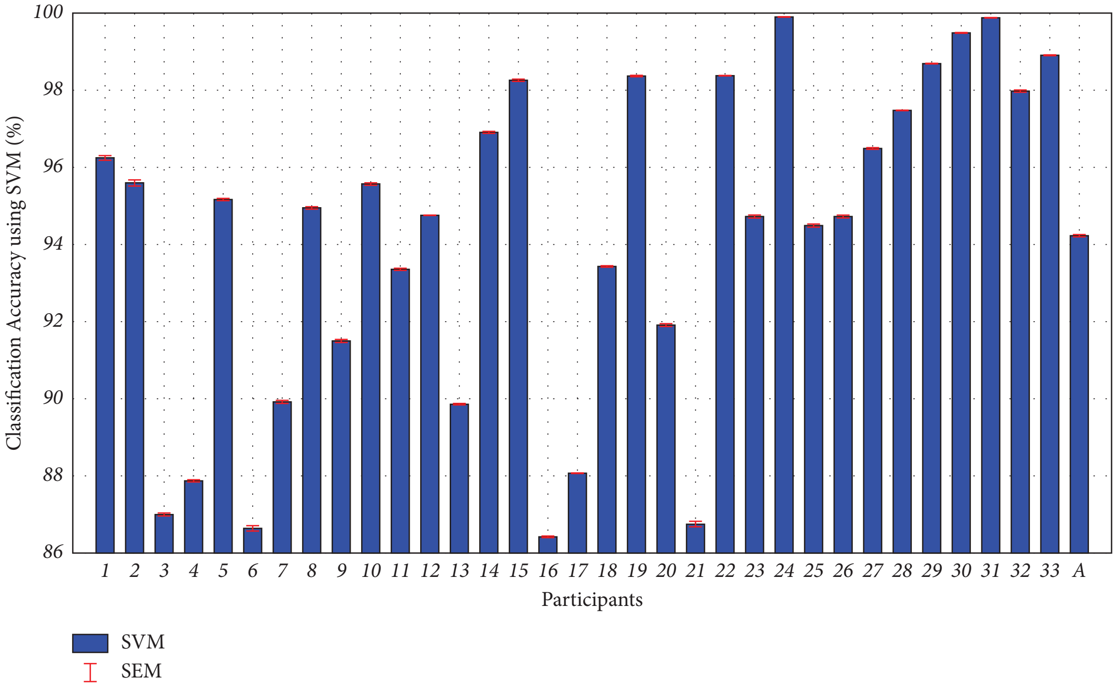


Figure 8: Classification results between UPB and non-UPB using SVM for each subject during cognitive tasks in accordance with the BOGOF condition. The red error bars refer to the standard error of mean (SEM). "A" in the x-axis indicates the average classification accuracy across 33 subjects with average accuracy and SEM (94.23% ± 0.73). The accuracies of all subjects are ranging from 86.42 to 99.90%, suggesting that the fNIRS data could be used as a biomarker to differentiate between UPB and non-UPB.

which both the UPBs and non-UPBs are trustworthily separable. Thus, our study denotes that the SVM model provides high accuracies which are reliable.

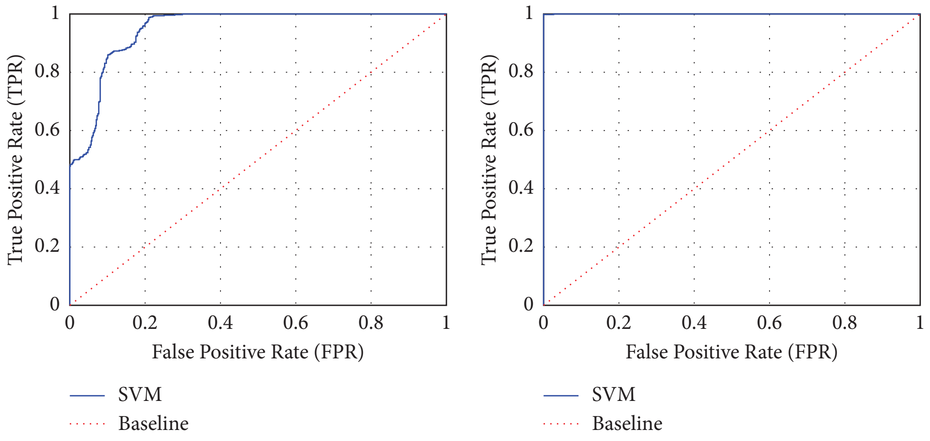
3.6. Detection Results of UPB and Non-UPB Patterns. Figure 10 illustrates the screenshots of the detection results for UPB patterns. All subjects received a message that this system can detect UPB patterns because each subject had reached a classification accuracy of more than 86% in this experiment.

4. Discussion

4.1. Proposed UPB Identification Methodology with BCI and Self-Reported Measurement. Research related to promotion strategies focuses on detecting and predicting people's UPB patterns [1]. However, it is difficult to measure the actual UPB in the lap setting. Therefore, UPB has so far relied on qualitative and quantitative research such as interviews and surveys. In line with this trend, this study also confirmed that a compulsive desire to buy increased during BOGOF using self-report measurement. It also demonstrates that

Table 1: AUC results across 33 subjects.

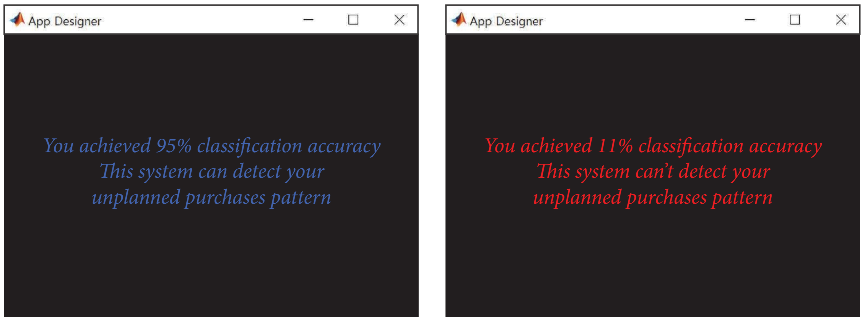
Subjects	AUC	Subjects	AUC	Subjects	AUC
Subject 1	0.99	Subject 13	0.95	Subject 25	0.98
Subject 2	0.96	Subject 14	0.99	Subject 26	0.97
Subject 3	0.92	Subject 15	0.99	Subject 27	0.98
Subject 4	0.95	Subject 16	0.95	Subject 28	0.97
Subject 5	1.00	Subject 17	0.85	Subject 29	1.00
Subject 6	0.93	Subject 18	0.99	Subject 30	1.00
Subject 7	0.96	Subject 19	1.00	Subject 31	1.00
Subject 8	0.94	Subject 20	0.97	Subject 32	0.99
Subject 9	0.96	Subject 21	0.97	Subject 33	1.00
Subject 10	0.99	Subject 22	1.00	Average	0.97
Subject 11	0.99	Subject 23	0.99	SD	0.03
Subject 12	0.97	Subject 24	1.00	SEM	0.01



(a)

(b)

Figure 9: ROC curves and their AUCs of two representative subjects with (a) the lowest accuracy and (b) the highest accuracy. The red dotted diagonals represent the baseline. The reliability of the classification results is not guaranteed if the SVM curves (blue lines) locate below the baseline, but the classification results are reliable if the SVM curves locate above the baseline. Therefore, all SVM results are trustworthy, and (b) is more reliable than (a).



You achieved 95% classification accuracy
This system can detect your
unplanned purchases pattern

You achieved 11% classification accuracy
This system can't detect your
unplanned purchases pattern

(a)

(b)

Figure 10: Screenshots of the detection results for UPB patterns. If the classification accuracy is above 80%, a message indicates that you are ready to detect unplanned purchase patterns (a); otherwise, indicates that you are not able to detect unplanned purchase patterns (b). In this experiment, all subjects received a message stating that this system can detect UPB patterns because each subject achieved a classification accuracy of higher than 86%.

our experiments are well designed. Several studies, however, emphasize that these tools still must be used with caution because they are influenced by the subjects' perceptions and memories [8]. To supplement the shortcomings of traditional marketing research methods, we present the UPB identification methodology to classify between a UPB and a non-UPB as illustrated in Figures 7 and 8. Eventually, we can identify the UPBs through a machine learning-based classification approach using fNIRS-SVM along with the self-reported results.

4.2. High Classification Accuracy in the Proposed Methodology Compared to the Existing Research Methodology. In our study, we proposed the optimal measurement methodology based on fNIRS-SVM, which would aid and improve the correct identification of UPBs caused by BOGOF. The proposed method along with self-reported measurements can serve as an optimal measurement tool to detect unplanned and impulsive purchase patterns. Likewise, in previous studies, measurement tools for impulsive detection have also existed, including clinical and neuropsychiatric tests. According to a previous study [51], psychometrical questionnaires such as Barrat's impulsiveness scale version 11 and the International Personality Disorder Evaluation Screening Questionnaire were used to detect people's impulsivity. These results are consistent with the SVM results by obtaining the impulsivity classification accuracy above 76%. Another study has reported the potential of the fNIRS-SVM classification approach between impulsive and nonimpulsive adolescents, which achieves a classification accuracy greater than 90%. This result was identical to the clinical assessment results by showing a significant difference in scores between the two adolescent groups [52]. Similarly, we achieved an average accuracy of 94% by detecting UPBs across 33 people, which was higher than the results of the previous study [52] as illustrated in Figure 8. Our study shows the highest achievement among previous achievements for detecting UPBs.

4.3. Detection of Low Brain Activity in the UPB. Many studies have reported that the more impulsive buying tendency people have, the lower is the brain activity in their PFC. For example, typical symptoms related to impulsiveness include obesity and binge-eating disorder [53]. They found that the obesity group had a lower fNIRS-based PFC response than the normal-weight group, indicating a connection between impulsiveness and a specific obesity phenotype. Another study found that the control group had higher prefrontal activation than ADHD children with high impulsiveness [54, 55]. Similarly, our study illustrates that Figure 7 depicts low brain activation at the PFC as a result of BOGOF. We have also confirmed that BOGOF elicits UPBs from the subjects' self-reported results. Thus, we revealed low brain activity because of UPBs at the PFC for the first time.

5. Conclusions

We proposed the optimal measurement methodology applied with fNIRS-SVM that can classify UBP patterns caused by BOGOF tasks and non-UBPs caused by control tasks and then validate their excellent classification results by AUC. As a result, we achieved an average accuracy above 94% by utilizing patterns of promotion strategy for UBPs. The classification result's reliability is validated by satisfying the AUC values above 0.97. We also found that the brain activity for UPBs was lower during the BOGOF tasks than during the control tasks at the PFC. This is consistent with the self-reported results that the subjects perceived an increase in impulsivity when they were exposed to BOGOF. Therefore, this study raises awareness of consumers' UPB and shows the possibility of applying optimized UPB measurement methodology to various applications such as mobile and PC in terms of computer-aided detection.

Data Availability

The raw data used in the study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by an Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government(MSIT) (Grant no. 2017-0-00451); Development of BCI-based Brain and Cognitive Computing Technology for recognizing User's Intentions using Deep learning. The authors would like to thank ESSAYREVIEW (https://www.essayreview.co.kr) for English language editing.

References

[1] A. Shoham and M. Makovec Brenčič, "Compulsive buying behavior," *Journal of Consumer Marketing*, vol. 20, no. 2, pp. 127–138, 2003.

[2] H. Dittmar and J. Drury, "Self-image–is it in the bag? A qualitative comparison between "ordinary" and "excessive" consumers," *Journal of Economic Psychology*, vol. 21, no. 2, pp. 109–142, 2000.

[3] I. Muratore, "Teens as impulsive buyers: what is the role of price?" *International Journal of Retail & Distribution Management*, vol. 44, no. 11, pp. 1166–1180, 2016.

[4] M. Mihić and I. Kursan, "Assessing the situational factors and impulsive buying behavior: market segmentation approach," *Management: Journal of Contemporary Management Issues*, vol. 15, no. 2, pp. 47–66, 2010.

[5] C. J. Cobb and W. D. Hoyer, "Planned versus impulse purchase behavior," *Journal of Retailing*, vol. 62, pp. 384–409, 1986.

[6] S. K. Hui, J. J. Inman, Y. Huang, and J. Suher, "The effect of in-store travel distance on unplanned spending: applications to mobile promotion strategies," *Journal of Marketing*, vol. 77, no. 2, pp. 1–16, 2013.

[7] C. Marquis and A. Park, "Inside the Buy-One Give-One Model," *Stanford Social Innovation Review, Winter*, pp.22–83, https://ecommons.cornell.edu/bitstream/handle/1813/36442/Inside_the_Buy_One_Give_One_Model.pdf?sequence=1, 2014.

[8] A. Nissen and C. Krampe, "Why he buys it and she doesn't–Exploring self-reported and neural gender differences in the perception of eCommerce websites," *Computers in Human Behavior*, vol. 121, Article ID 106809, 2021.

[9] L. Bell, J. Vogt, C. Willemse, T. Routledge, L. T. Butler, and M. Sakaki, "Beyond self-report: a review of physiological and neuroscientific methods to investigate consumer behavior," *Frontiers in Psychology*, vol. 9, p. 1655, 2018.

[10] A. M. Petrotta and A. Galli, "The Rhythm of the Buying: An Empirical Investigation of the Influence of Background Music Tempo on Online Impulse Buying," 2020, http://hdl.handle.net/10589/175234.

[11] X. Gu, C. Zhang, and T. Ni, "A hierarchical discriminative sparse representation classifier for EEG signal detection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1679–1687, 2021.

[12] W. Zhao, W. Zhao, W. Wang et al., "A Novel Deep Neural Network for Robust Detection of Seizures Using EEG Signals," vol. 2020, Article ID 9689821, 9 pages, 2020.

[13] R. Sitaram, A. Caria, R. Veit, T. Gaber, G. Rota, and A. Kuebler, "FMRI Brain-Computer Interface: A Tool for Neuroscientific Research and Treatment," vol. 2007, Article ID 025487, 2007.

[14] D. L. Schomer and F. L. Da Silva, *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related fields*, Lippincott Williams & Wilkins, Philadelphia, United States, 2012.

[15] V. Scarapicchia, C. Brown, C. Mayo, and J. R. Gawryluk, "Functional magnetic resonance imaging and functional near-infrared spectroscopy: insights from combined recording studies," *Frontiers in Human Neuroscience*, vol. 11, p. 419, 2017.

[16] S. Dong and J. Jeong, "Onset classification in hemodynamic signals measured during three working memory tasks using wireless functional near-infrared spectroscopy," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–11, 2019.

[17] A. Sassaroli, M. Pierro, P. R. Bergethon, and S. Fantini, "Low-frequency spontaneous oscillations of cerebral hemodynamics investigated with near-infrared spectroscopy: a review," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 18, no. 4, pp. 1478–1492, 2012.

[18] V. O. Korhonen, T. S. Myllyla, M. Y. Kirillin et al., "Light propagation in NIR spectroscopy of the human brain," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 2, pp. 289–298, 2014.

[19] A. M. Batula, J. A. Mark, Y. E. Kim, and H. Ayaz, "Comparison of Brain Activation during Motor Imagery and Motor Movement Using fNIRS," vol. 2017, Article ID 5491296, 12 pages, 2017.

[20] B. Blankertz, C. Sannelli, S. Halder et al., "Neurophysiological predictor of SMR-based BCI performance," *NeuroImage*, vol. 51, no. 4, pp. 1303–1309, 2010.

[21] K.-S. Hong, U. Ghafoor, and M. J. Khan, "Brain–machine interfaces using functional near-infrared spectroscopy: a review," *Artificial Life and Robotics*, vol. 25, no. 2, pp. 204–218, 2020.

[22] K.-S. Hong and M. A. Yaqub, "Application of functional near-infrared spectroscopy in the healthcare industry: a review," *Journal of Innovative Optical Health Sciences*, vol. 12, no. 06, Article ID 1930012, 2019.

[23] E. Joo Park, E. Young Kim, and J. Cardona Forney, "A structural model of fashion-oriented impulse buying behavior," *Journal of Fashion Marketing and Management: International Journal*, vol. 10, no. 4, pp. 433–446, 2006.

[24] ZARA. [cited 2021 15 November]; Available from:, 2021, https://www.zara.com/kr/ko/woman-trousers-l1335.html?v1=2025899.

[25] An online clothing brand survey. [cited 2021 30 November]; Available from:, 2021, https://www.koreapas.com/bbs/view.php?id=pashion&page=1&sn1=&divpage=17&sn=off&ss=on&sc=on&no=87003.

[26] A. Sebastian, P. Jung, A. Krause-Utz, K. Lieb, C. Schmahl, and O. Tuscher, "Frontal dysfunctions of impulse control - a systematic review in borderline personality disorder and attention-deficit/hyperactivity disorder," *Frontiers in Human Neuroscience*, vol. 8, p. 698, 2014.

[27] M. A. Bertocci, H. W. Chase, S. Graur et al., "The impact of targeted cathodal transcranial direct current stimulation on reward circuitry and affect in bipolar disorder," *Molecular Psychiatry*, vol. 26, no. 8, pp. 4137–4145, 2021.

[28] D. A. Boas, A. M. Dale, and M. A. Franceschini, "Diffuse optical imaging of brain activation: approaches to optimizing image sensitivity, resolution, and accuracy," *NeuroImage*, vol. 23, pp. S275–S288, 2004.

[29] H. Obrig and A. Villringer, "Beyond the visible—imaging the human brain with light," *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, vol. 23, no. 1, pp. 1–18, 2003.

[30] A. Villringer and B. Chance, "Non-invasive optical spectroscopy and imaging of human brain function," *Trends in Neurosciences*, vol. 20, no. 10, pp. 435–442, 1997.

[31] G. Strangman, J. P. Culver, J. H. Thompson, and D. A. Boas, "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," *NeuroImage*, vol. 17, no. 2, pp. 719–731, 2002.

[32] A. Watanabe, N. Kato, and T. Kato, "Effects of creatine on mental fatigue and cerebral hemoglobin oxygenation," *Neuroscience Research*, vol. 42, no. 4, pp. 279–285, 2002.

[33] M. M. Plichta, M. Herrmann, C. Baehne et al., "Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable?" *NeuroImage*, vol. 31, no. 1, pp. 116–124, 2006.

[34] C. Altın and O. Er, "Comparison of different time and frequency domain feature extraction methods on elbow gesture's EMG," *European journal of interdisciplinary studies*, vol. 5, no. 1, pp. 35–44, 2016.

[35] M. Diykh, Y. Li, and P. Wen, "EEG sleep stages classification based on time domain features and structural graph similarity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1159–1168, 2016.

[36] S.-i. Kim, Y. Noh, Y. J. Kang, S. Park, and B. Ahn, "Fault classification model based on time domain feature extraction of vibration data," *Journal of the Computational Structural Engineering Institute of Korea*, vol. 34, no. 1, pp. 25–33, 2021.

[37] S. Park and S.-Y. Dong, "Effects of daily stress in mental state classification," *IEEE Access*, vol. 8, Article ID 201360, 2020.

[38] N. Naseer, N. K. Qureshi, F. M. Noori, and K. S. Hong, "Analysis of Different Classification Techniques for Two-Class Functional Near-Infrared Spectroscopy-Based Brain-Computer Interface," vol. 2016, Article ID 5480760, 2016.

[39] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1882–1889, 2003.

[40] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335–1343, 2004.

[41] Q. She, Y. Ma, M. Meng, and Z. Luo, "Multiclass Posterior Probability Twin SVM for Motor Imagery EEG Classification," vol. 2015, Article ID 251945, 95 pages, 2015.

[42] R. A. Khan, N. Naseer, and M. J. Khan, "Drowsiness detection during a driving task using fNIRS," in *Neuroergonomics*, pp. 79–85, Elsevier, Amsterdam, Netherlands, 2019.

[43] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[44] Q. Wang, “A hybrid sampling SVM approach to imbalanced data classification,” in *Abstract and Applied Analysis,* Hindawi, London, 2014.

[45] D. Avci, “An automatic diagnosis system for hepatitis diseases based on genetic wavelet kernel extreme learning machine,” *Journal of Electrical Engineering and Technology*, vol. 11, no. 4, pp. 993–1002, 2016.

[46] S. Wang, X. Wang, J. Chen et al., “Optical visualization of cerebral cortex by label-free multiphoton microscopy,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–8, 2019.

[47] Z. Chen, Y. Yan, E. Wang, H. Jiang, Y. Tang, and X. Yu, “Detecting Abnormal Brain Regions in Schizophrenia Using Structural MRI via Machine Learning,” vol. 2020, Article ID 6405930, 13 pages, 2020.

[48] S. Wang, X. Sun, Z. Chen et al., “Label-free detection of the architectural feature of blood vessels in glioblastoma based on multiphoton microscopy,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 27, no. 4, pp. 1–7, 2021.

[49] K. D. Vohs and R. J. Faber, “Spent resources: self-regulatory resource availability affects impulse buying,” *Journal of Consumer Research*, vol. 33, no. 4, pp. 537–547, 2007.

[50] S.-W. Lin and L. Y.-S. Lo, “Evoking online consumer impulse buying through virtual layout schemes,” *Behaviour & Information Technology*, vol. 35, no. 1, pp. 38–56, 2016.

[51] D. Delgado-Gomez, H. Blasco-Fontecilla, A. A. Alegria, T. Legido-Gil, A. Artes-Rodriguez, and E. Baca-Garcia, “Improving the accuracy of suicide attempter classification,” *Artificial Intelligence in Medicine*, vol. 52, no. 3, pp. 165–168, 2011.

[52] S. B. Erdoğan, G. Yukselen, M. M. Yegul et al., “Identification of impulsive adolescents with a functional near infrared spectroscopy (fNIRS) based decision support system,” *Journal of Neural Engineering*, vol. 18, no. 5, Article ID 056043, 2021.

[53] S. A. Rösch, R. Schmidt, M. Luhrs, A. C. Ehlis, S. Hesse, and A. Hilbert, “Evidence of fnirs-based prefrontal cortex hypoactivity in obesity and binge-eating disorder,” *Brain Sciences*, vol. 11, no. 1, p. 19, 2020.

[54] A. Güven, M. Altinkaynak, N. Dolu et al., “Combining functional near-infrared spectroscopy and EEG measurements for the diagnosis of attention-deficit hyperactivity disorder,” *Neural Computing & Applications*, vol. 32, no. 12, pp. 8367–8380, 2020.

[55] M. Altinkaynak, A. Guven, N. Dolu, M. Izzetoglu, E. Demirci, and S. Ozmen, “Investigating prefrontal hemodynamic responses in ADHD subtypes: a fNIRS study,” in *Proceedings of the 2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, IEEE, Bursa, Turkey, November 2017.

Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 1808990, 21 pages
https://doi.org/10.1155/2022/1808990

Research Article

HIT HAR: Human Image Threshing Machine for Human Activity Recognition Using Deep Learning Models

Alwin Poulose,[1] **Jung Hwan Kim,**[2] **and Dong Seog Han**[2]

[1] *Center for ICT and Automotive Convergence, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea*
[2] *Graduate School of Electronic and Electrical Engineering, Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Republic of Korea*

Correspondence should be addressed to Dong Seog Han; dshan@knu.ac.kr

Received 22 June 2022; Revised 19 August 2022; Accepted 8 September 2022; Published 6 October 2022

Academic Editor: Javier Medina

In recent days, research in human activity recognition (HAR) has played a significant role in healthcare systems. The accurate activity classification results from the HAR enhance the performance of the healthcare system with broad applications. HAR results are useful in monitoring a person's health, and the system predicts abnormal activities based on user movements. The HAR system's abnormal activity predictions provide better healthcare monitoring and reduce users' health issues. The conventional HAR systems use wearable sensors, such as inertial measurement unit (IMU) and stretch sensors for activity recognition. These approaches show remarkable performances to the user's basic activities such as sitting, standing, and walking. However, when the user performs complex activities, such as running, jumping, and lying, the sensor-based HAR systems have a higher degree of misclassification results due to the reading errors from sensors. These sensor errors reduce the overall performance of the HAR system with the worst classification results. Similarly, radiofrequency or vision-based HAR systems are not free from classification errors when used in real time. In this paper, we address some of the existing challenges of HAR systems by proposing a human image threshing (HIT) machine-based HAR system that uses an image dataset from a smartphone camera for activity recognition. The HIT machine effectively uses a mask region-based convolutional neural network (R-CNN) for human body detection, a facial image threshing machine (FIT) for image cropping and resizing, and a deep learning model for activity classification. We demonstrated the effectiveness of our proposed HIT machine-based HAR system through extensive experiments and results. The proposed HIT machine achieved 98.53% accuracy when the ResNet architecture was used as its deep learning model.

1. Introduction

The human healthcare systems have a vital role in our daily life. Due to the busy lifestyle, these days, the lack of exercise causes serious health issues. Emerging technologies such as human activity recognition (HAR) systems [1] can monitor the users' activities in the healthcare system. Recent research trends in HAR show its wide variety of applications that include health and fitness monitoring [2], assisted living [3], context-enabled games and entertainment [4], social networking [5], and sports tracking [6]. In HAR, the system tracks the user's movements and classifies the user's activities based on the sensor reading. The existing HAR system includes vision-based [7], radiofrequency-based [8], or wearable sensor-based approaches [9]. The most common and low installation cost-based HAR technique is the wearable sensor-based approach. The sensor-based technique is location independent, and the user can easily hold the sensor during their activities. The sensor-based HAR approaches achieved a remarkable classification accuracy, and smartphone or smartwatch-based HAR is the most common system used for activity recognition. However, the sensor errors, sensor type, sensor position in the human body, and user's complex activities make the system more challenging for activity recognition. The HAR system has worst classification results when the user is in complex activity motion. On the other side, when the HAR system uses radio frequency (RF) signals for activity recognition, the system takes advantage of the wireless communication features to classify the user's activities. Compared with the sensor-based HAR approach, RF-based HAR is device-free, and the system does not need any physical sensing module. The device-free characteristics of radio frequency-based HAR provide reduction in energy consumption and privacy protection compared with the sensor or vision-based HAR systems. However, indoor channel conditions, non-line of sight conditions, and signal interference affect the performance of HAR, and the system faces difficulties in maintaining high accuracy levels. Besides these HAR approaches, the vision-based HAR system uses a camera that records the user's activities in a video sequence. The vision-based approach uses computer vision algorithms for activity recognition. Based on the camera type used in the HAR system, the video sequence from the vision approach is in the form of RGB videos [10], depth videos [11], or RGB-D videos [12]. Compared with sensor-based or radio frequency-based HAR approaches, the vision-based approach shows higher classification results for users' complex activities. However, user privacy, energy consumption, and deployment cost are the main challenges for the vision-based HAR approaches. In this paper, our research focuses on the vision-based HAR approach, and we propose a human image threshing (HIT) machine-based HAR system that addresses some of the existing vision-based HAR challenges. Our HIT machine-based HAR system uses a smartphone camera as an input device to record the users' activities. A mask region-based convolutional neural network (R-CNN) further processes the recorded activity videos for human body detection, a facial image threshing machine (FIT) for image cropping and resizing [13], and a deep learning model for activity recognition. Our HIT machine can generate HAR images from activity videos, human body detection from images, data cleaning and removal of irrelevant data, and activity classification using a deep learning model. We tested our HIT machine with different HAR experiments based on deep learning models, including visual geometry group (VGG) [14], Inception [15], ResNet [16], and EfficientNet [17] models. The results from the HIT machine show that the system always maintains the classification accuracy for activity recognition. We analyzed our HIT machine results with conventional HAR approaches that include inertial measurement unit (IMU) and stretch sensor-based approaches. The results show that the HIT machine outperforms the traditional sensor-based approaches with a higher level of accuracy for activity recognition. We also tested our pre-trained deep learning models with unseen HAR datasets and analyzed the classification performance. The key contributions from our HIT machine are stated as follows:

(i) We created a HAR dataset using a smartphone camera, IMU sensor, and stretch sensor. Our dataset consists of nine activities: sitting, standing, lying, walking, push up, dancing, sit-up, running, and jumping. It has 36, 558 image samples from smartphone cameras, 97,454 data samples from IMU sensors, and 7,850 data samples from stretch sensors. We used these datasets to validate our HIT machine, and the deep learning models can use our HAR datasets for training and testing without any computational complexity. We also collected HAR datasets for unseen datasets and tested them with pre-trained deep learning models.

(ii) We proposed a HIT machine for activity recognition, and our HIT machine shows accurate classification results for basic (sitting, standing, and walking) and complex (running, jumping, and lying) activities. We tested our HIT machine with different deep learning models and analyzed the classification performance in terms of a confusion matrix, accuracy, loss, precision, recall, and F1 score. We also tested the pre-trained models with unseen HAR datasets and compared the performance of each model. We validated our HIT machine results with sensor-based HAR results and proved the impact of the HIT machine for activity recognition.

The rest of the paper is organized as follows: Section 2 discusses the existing HAR systems, recently proposed HAR systems with their advantages, and current HAR challenges for practical implementation. Section 3 presents our proposed HIT machine-based HAR system, including mask R-CNN, FIT machine, and deep learning models. Section 4 discusses our HAR experiments with the validation of our HIT machine in terms of the impact of various deep learning models, analysis of unseen datasets for pre-trained models, and the result comparison with conventional HAR approaches. Finally, Section 5 concludes our HIT machine-based HAR approach with future research directions.

2. Related Work

HAR has been studied for applications in healthcare monitoring, smart homes, security, medical imaging, robot/human interaction, personal assistants, and surveillance [18–20]. Many researchers have discussed various HAR approaches based on the technologies or algorithms used for activity recognition [21–25]. In this paper, our literature focuses on related work for HAR approaches that include sensors [26, 27], Wi-Fi [28], Wi-Fi, and sensors [29], vision [30, 31], and RFID [32]-based activity recognition. The HAR approaches from [26–32] provide significant performance improvements for HAR applications. However, the diversity of age, gender, and number of subjects, postural transitions, number of sensors and type of sensors, different body locations of wearable sensors or smartphones, missing values or labeling error, similar postures and datasets having complex activities, lack of ground truths, selection of appropriate datasets, and selection of sensors [33, 34] create challenges to the HAR implementation. This paper proposes a HIT machine-based HAR system to address some of these challenges with higher classification results.

The sensors-based HAR approaches are the most common and popular HAR systems. In sensor-based HAR, the system uses wearable sensors, smartphones, or smartwatches to collect data and identify the users' activity based on the sensor readings. Some of the recent HAR systems which take advantage of wearable sensors are discussed in [35–39]. These systems achieved a remarkable recognition accuracy in real time. However, mounting a wearable sensor in the human body is challenging, and the wearable sensor's position determines the system's performance. The wearable sensor-based HAR systems still need to optimize the location of sensors in the human body for complex activity. An alternative method for activity recognition is the smartphone-based HAR systems [40–43]. In smartphone-based HAR, the user holds the smartphone and performs the activities. Compared with wearable sensor-based approaches, the smartphone-based method is simple and easy to implement in any place without any external sensors. However, the position in which the smartphone is held and the modes such as texting and calling affect the system's performance. The smartphone or wearable sensors-based HAR approach still needs to improve the classification performance at a certain level, and current systems use deep learning models for activity recognition [44–47]. The deep learning HAR-based systems include convolutional neural network (CNN) [48], long short-term memory (LSTM) [49], LSTM-CNN [50], deep recurrent neural networks (DRNN) [51], generative adversarial networks (GAN) [52], extreme learning machine (ELM) [53], graph neural network (GNN) [54], and semi-supervised deep learning models [55, 56]. These systems use raw sensor reading or extract the signal features in the time/frequency domain for activity recognition. When the system uses the signal in the time domain, it extracts the variance, mean, maximum, minimum, and range values and uses these features as model inputs. On the other hand, If the signal is in the frequency domain, the system extracts the amplitude, skewness, kurtosis, and energy information as to its features and uses this input to the model. Compared with the raw input signal-based deep learning HAR approach, the feature-based approaches show better classification results [2]. However, the deep learning-based HAR approaches are not free from challenges. A large number of data samples for training, training time, the complexity of feature extraction, and human resources required for data collection are some of the main challenges of deep learning-based HAR approaches. These challenges reduce systems performance and require further classification improvements.

The RF-based HAR approaches use physical sensors, such as pressure, proximity, FM radio, microwave, or RFID for activity recognition [57–61]. In a radio frequency-based approach, the system takes advantage of the body attenuation and the channel fading characteristics for activity recognition. The basic principle of RF-based HAR systems is that the propagation of RF signals is affected by the human body movement, resulting in attenuation, refraction, diffraction, reflection, and multipath effects. These pattern differences in the received RF signals are the key ideas for activity recognition. Different activities lead to various patterns inside RF signals, and the system can use these features for classification. The RF-based systems consist of signal selection, model, signal processing, segmentation, feature extraction, and activity classification. In signal selection, the system uses Wi-Fi, ZigBee [63], RFID [64], frequency-modulated continuous-wave radar (FMCW) or acoustic devices. The system uses phase, frequency, amplitude, or raw signal for activity recognition depending on the signal selection. These factors determine the model of the HAR system. When the model is defined, the system uses signal processing techniques, including noise reduction, calibration, and redundant removal. After this, the system uses signal segmentation in the time or frequency domain. When segmentation performed, the time domain, frequency domain, time-frequency domain, or spatial domain features are extracted for classification. The deep learning models use extracted features for activity recognition. Compared with the wearable sensor-based HAR approach, the RF-based approach exploits the wireless communication features for activity recognition. These systems do not use any physical sensing module, thus reducing energy consumption and user privacy concern. Some of the RF-based HAR approaches are discussed in [65–68]. The RF-based systems discussed here have enhanced the HAR classification performance and opened many applications for detection, recognition, estimation, and tracking. However, the wireless channel conditions, signal interference, non-line-of-sight (NLOS) conditions, multi-user activity sensing, and limited sensing range make the systems more challenging. They require new theoretical models and open datasets for accurate classification.

The system uses a video sequence for activity monitoring when considering a vision-based HAR approach for activity recognition [69, 70]. The vision-based approach is best for multi-user activity recognition when privacy is not a significant concern. These systems use different computer vision algorithms on activity videos to predict the user's activities from videos or images. Some of the vision-based HAR approaches are proposed in [71–77]. These vision-based systems effectively use the video or image sequences and classify the users' activity by taking advantage of the recent deep learning models. Several review papers on the vision-based HAR systems are discussed in [78–80]. From vision-based HAR review discussions, the authors from [81] focus on the high level of visual processing, including human body modeling, understanding of human actions, and approaches to human action recognition. In [82], the authors presented the current state-of-the-art development of automated visual surveillance systems. They discussed the necessity of intelligent visual surveillance in commercial, law enforcement, and military applications. In [83], the paper reviews the advances in human motion capture and analysis from 2000 to 2006 and discusses the problems for future research to achieve automatic visual analysis of human movement. The review paper [84] analyzes the approaches taken to date within the computer vision, robotics, and artificial intelligence communities to represent, recognize, synthesize, and understand action. In [84], the authors pay more attention to identifying actions at different levels of complexity. Machine recognition of human activities is reviewed in [85], and the authors present a comprehensive survey of efforts to address the vision-based HAR systems. The paper [80] focuses on pedestrian detection, and [86] introduces a HAR system that recognizes the human behaviors from transit scenes. The most recent HAR systems are presented in [87–89]. These systems tried to improve the feature extraction techniques by introducing object detection, skeleton tracking, and human body poses. The vision-based HAR systems discussed here still have some challenges, such as processing high-quality videos or images, the complexity of the vision algorithms, the requirement for a higher graphics processing unit (GPU) processing power, the installation cost of the camera, and challenges from vision systems such as camera viewpoint, lighting, human body appearance, occlusion, and background clutter. These challenges make it more difficult for the vision-based approaches for real-time health monitoring.

So far, we have discussed different types of HAR approaches based on their technologies and algorithms used for activity recognition. In this paper, our research mainly focuses on the vision-based HAR approach, and we used our smartphones for data collection. We also collected data using IMU and stretch sensors, and the results from these sensors are compared with our proposed HIT machine. The experiment results show that the HIT machine is a practical HAR approach for healthcare applications and needs only a basic smartphone model for activity recognition.

3. Proposed HIT Machine-Based HAR System

The HIT machine consists of HAR dataset creation, data preprocessing, human body detection using mask R-CNN, image cropping and resizing, data cleaning and removal of irrelevant data, deep feature extraction, model building, and activity classification. Figure 1 shows the framework of our proposed HIT machine-based HAR system.

We first started our data collection in the HIT machine by using android and iOS smartphones that record activity videos. Next, the HIT machine performs the data aggregation on the activity video sequences. The data aggregation gathers all activity data and presents it in a summarized format. Followed by the data aggregation process, our system uses a mask R-CNN algorithm for human body detection. After this, the HIT machine operates the FIT machine for image cropping and resizing when the human body is identified from images. The cropped and resized activity images are ready for the model to use for training and testing. Our HIT machine also used a data cleaning process that removes the unnecessary images from the HAR dataset. After the data cleaning process, the images are ready to be used for model training and testing. We extracted the features from the activity images and created a deep learning model that classifies user activities into nine groups. The output of the HIT machine is the classification results of user activities which include sitting, standing, lying, walking, push up, dancing, sit-up, running, and jumping. The flowchart of the proposed HIT machine is presented in Figure 2.

In the flowchart, the system starts with HAR datasets. The datasets include HAR images from smartphones, accelerometer and gyroscope readings from IMU sensors, and stretch sensor readings. The HAR image dataset is then divided into training, testing, and unseen datasets. We used our HIT machine in the image HAR dataset for human body detection and activity recognition. The HIT machine includes human body detection, data preprocessing using a FIT machine, and deep learning models for classification. A mask R-CNN-based object detection algorithm is used for human body detection. A FIT machine is used for data preprocessing, including image cropping, resizing, data cleaning, and data segregation. A deep learning model is used for the training, and the model classifies the user activities into different categories. The system uses deep learning models of VGG, Inception, ResNet, and EfficientNet. On the other hand, conventional HAR approaches use IMU and stretch sensor data for activity recognition with a CNN model. The CNN model also uses the HAR image dataset for activity recognition, and we compared the effect of our HIT machine (with and without HIT machine) for activity recognition. Further discussions of mask R-CNN, FIT machine operation, and the deep learning models are added in the following subsections.

3.1. *Mask R-CNN.*

In computer vision, mask R-CNN is widely used for object detection tasks [90]. The mask R-CNN separates different objects from a video or an image. The algorithm provides the object bounding boxes, classes, and mask information, and our HIT machine can effectively utilize this information for human body detection. The mask R-CNN from our HIT machine operates in two stages. First, the algorithm generates proposals about the regions where an object is located in the input image. Second, the algorithm predicts the object class and refines the bounding box. The algorithm also adds a mask in the pixel level of the object based on the first stage proposal. Compared with Fast/Faster R-CNN-based object detection approaches, the mask R-CNN-based approach has additional features such as a binary mask for each region of interest (RoI). Our system utilizes this binary mask feature for human body detection. Figure 3 shows the structure of mask R-CNN.

The mask R-CNN consists of a backbone, a region proposal network (RPN), a region of interest alignment layer (RoTAlign), an object detection head, and a mask generation head. The backbone of mask R-CNN is the primary feature extractor which uses residual networks (ResNets) with or without feature pyramid networks [91]. When our HAR images are fed into a ResNet backbone, the images go through multiple residual bottleneck blocks and turn into a feature map. The feature map contains the abstract information of input images, including different object instances, classes, and spatial properties. The feature map data are then fed into the RPN layer. In this layer, the network scans the feature map and RoI where the human body is located. The next step is to find each RoI from the feature map. This process is referred to as RoIAlign in Figure 3. The RoIAlign extracts the feature vectors from the feature map based on the RoI suggested by the RPN layer. The feature vectors are then converted into a fix-sized tensor for further processes. The outputs from RoIAlign are then processed by two parallel branches: object detection branch and mask generation branch. The object detection branch is a fully-connected layer that maps the feature vectors to the final classes and bounding box coordinates. The mask generation

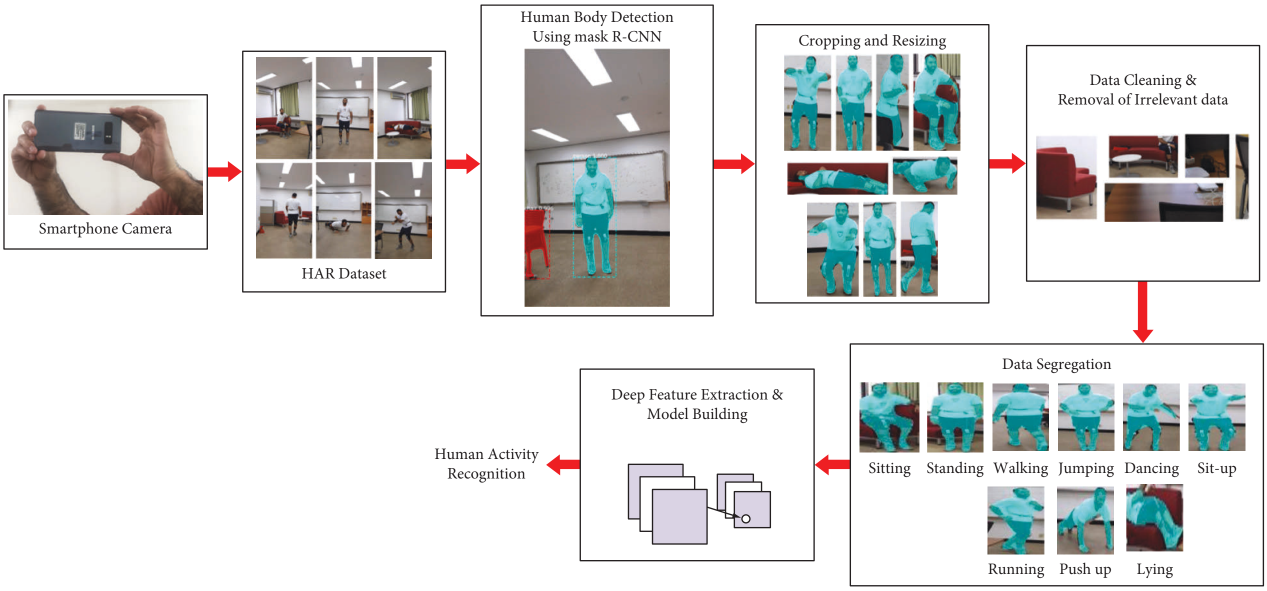

Smartphone Camera
HAR Dataset
Human Body Detection Using mask R-CNN
Cropping and Resizing
Data Cleaning & Removal of Irrelevant data
Data Segregation
Sitting
Standing
Walking
Jumping
Dancing
Sit-up
Running
Push up
Lying
Deep Feature Extraction & Model Building
Human Activity Recognition

FIGURE 1: Proposed HIT machine-based HAR system.

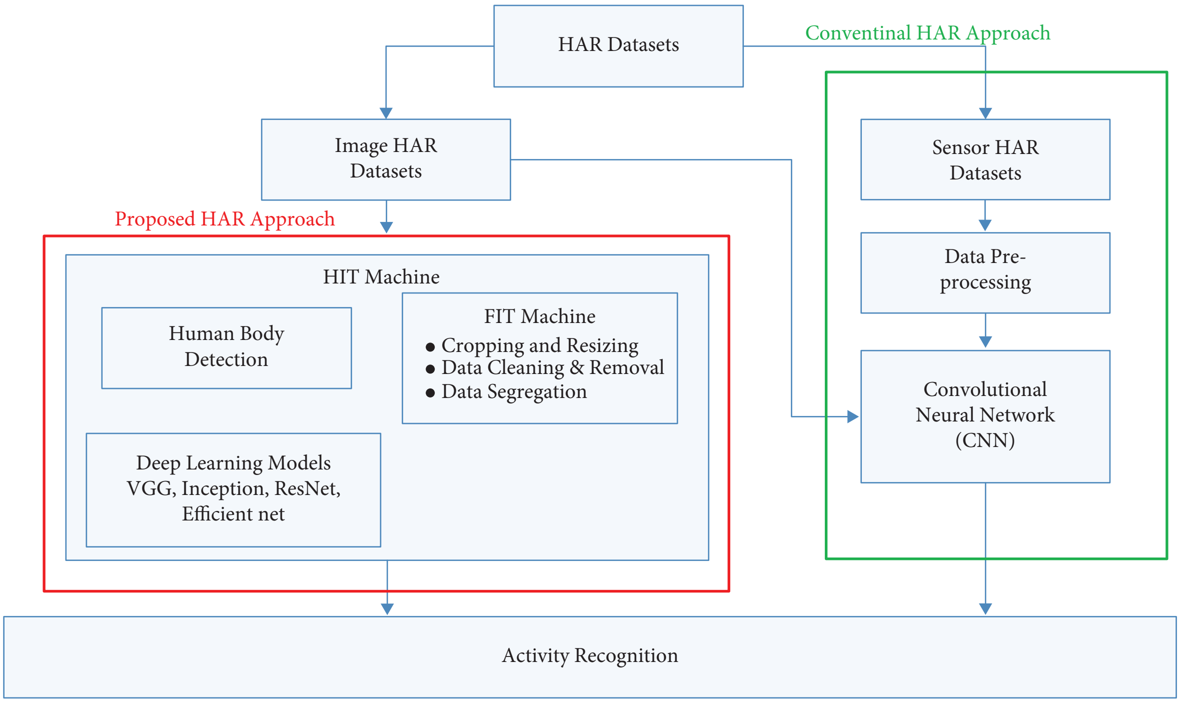
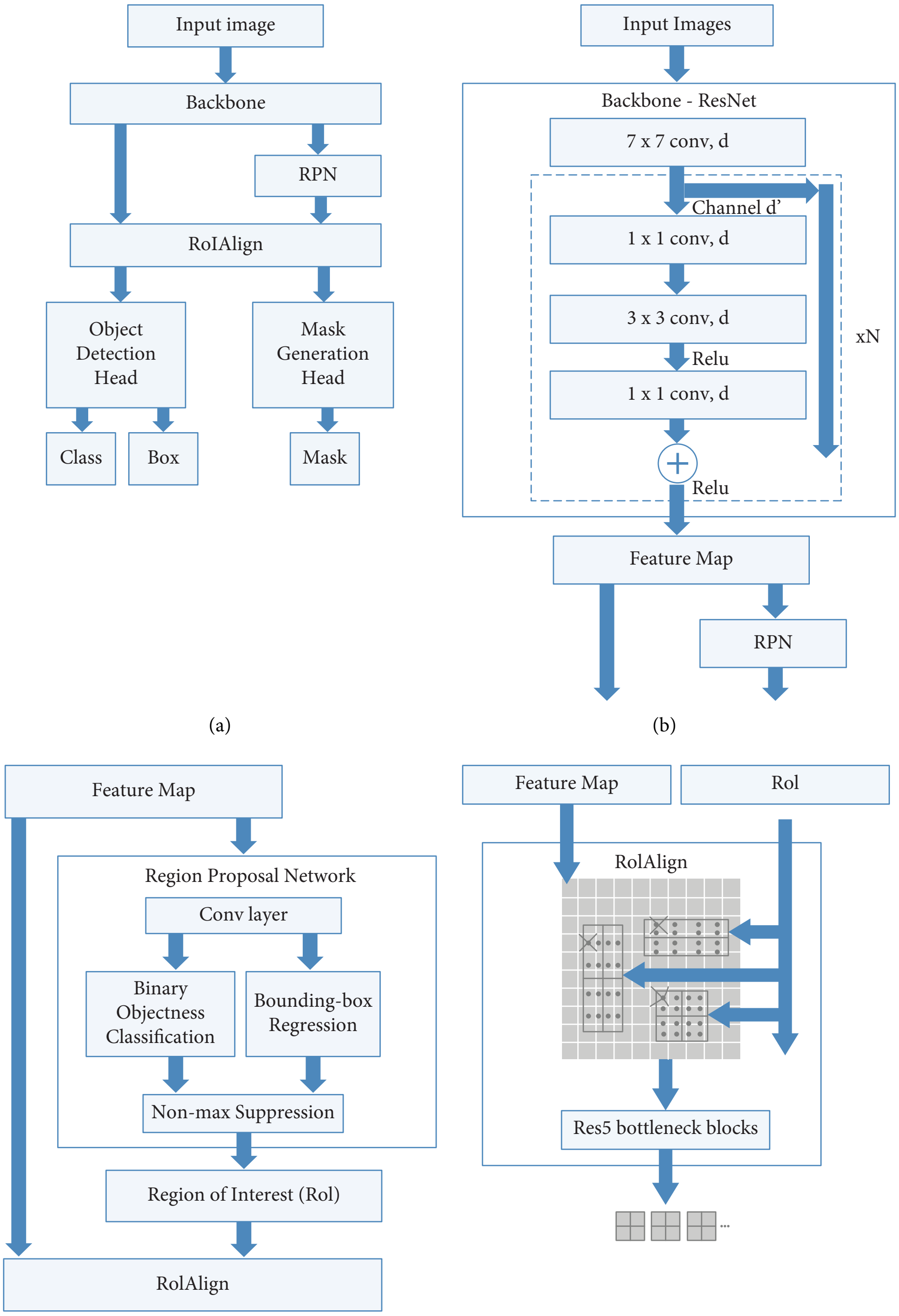


FIGURE 2: Flowchart of the proposed HIT machine.

branch feeds the feature map into a transposed convolutional layer and convolutional layer. The output of mask generation branch is one binary segmentation mask that is generated for one class. Then the system picks the output mask based on the class prediction from the object detection branch. Figure 4 shows the human body detection using our HIT machine for nine activities.

As shown in Figure 4, the mask R-CNN accurately detects the human body for nine activities without any detection error. The mask R-CNN used here is straightforward and has a small computational overhead that enables a fast system and rapid experimentation. For more details on mask R-CNN and its implementation, refer to [92–94].

3.2. FIT Machine. The HIT machine effectively uses our previously proposed FIT machine for image cropping and resizing [13]. The FIT machine is used to correct missing HAR datasets, remove irrelevant data, merge datasets on a massive scale, and crop and resize images. Our FIT machine



(a)

(b)

(c)

(d)

Figure 3: Continued.

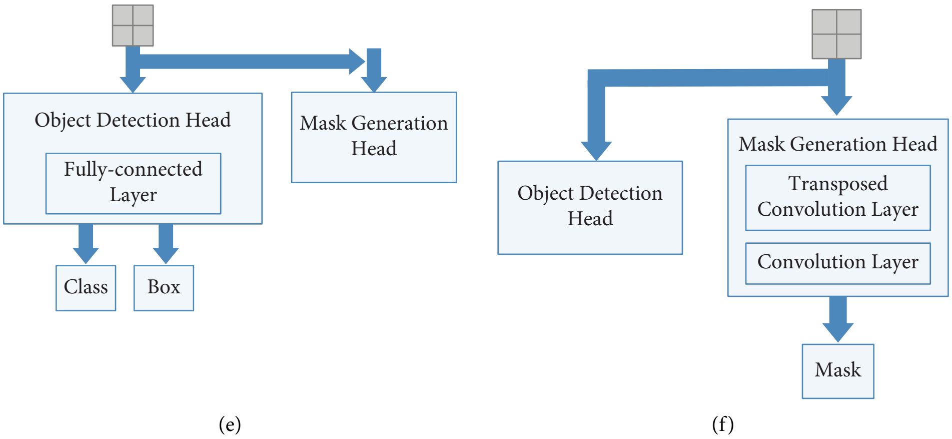


Figure 3: Mask R-CNN. (a) Structure. (b) Backbone. (c) RPN. (d) RoIAlign. (e) Object detection head. (f) Mask generation head.

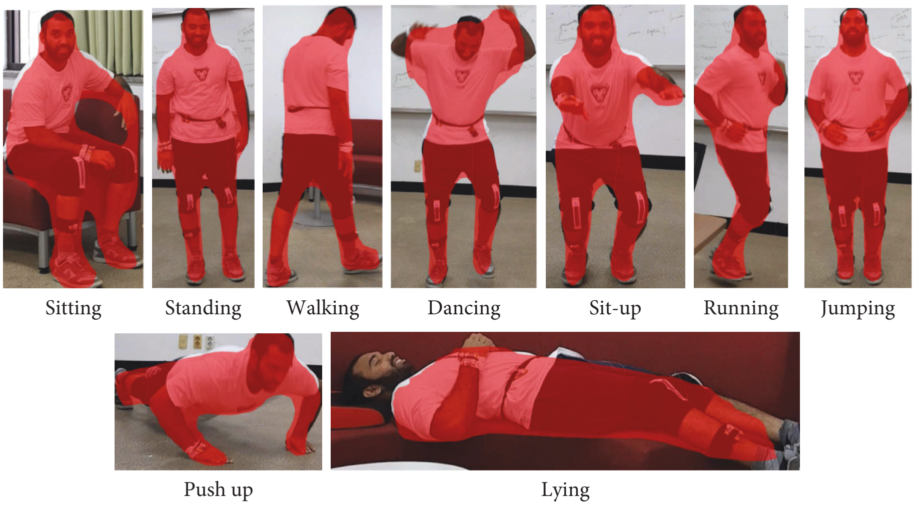
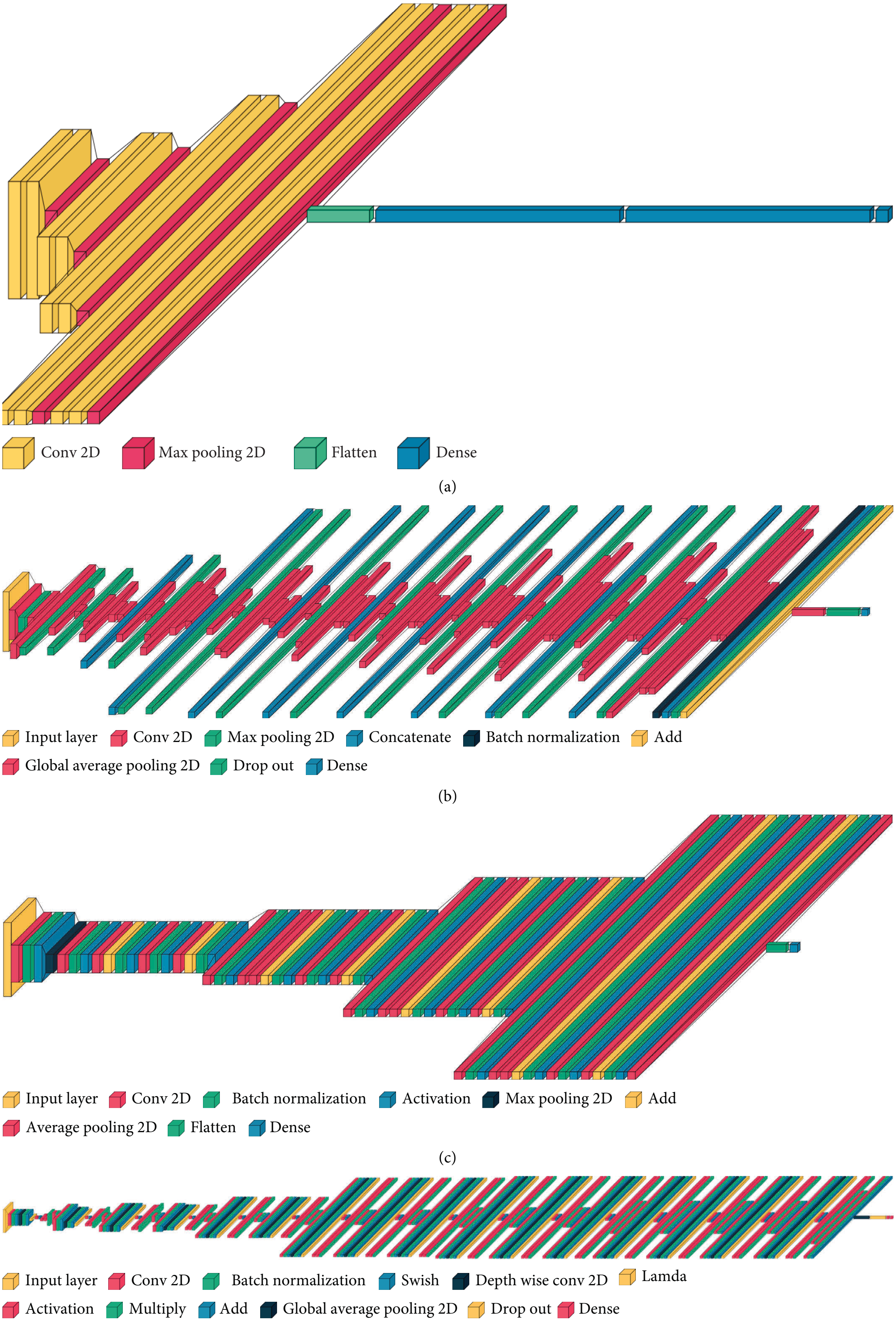


Figure 4: Human body detection using our HIT machine.

converts input activity video sequences into the image output samples that consist of cropped, resized, and categorized activity images. The FIT machine contains a data receiver, a multi-task cascaded convolutional network (MTCNN), an image resizer [95], and a data segregator as the pre-trained Xception algorithm model [96]. The data receiver converts activity video sequences into images, and the MTCNN identifies the human faces from the activity images. The MTCNN used here consists of P-Net, R-Net, and O-Net layers. When the architecture detects the human faces, the input images enter the P-Net layer, which chooses the possible face frames from the input images. The R-Net layer in the MTCNN uses P-Net outputs as its inputs. The R-Net layer inspects the given initial frames from P-Net, then removes the face frames that do not reach a threshold score. Followed by the R-Net, the O-Net uses the output from the R-Net at the end. In the O-Net layer, it selects the best face frames from the given output from R-Net. Next, the images are passed through an image resizer, reducing the image size to 224×224 pixels. The last part of the FIT machine is a data segregator, which segregates the activity images into adequately labeled directories. The data segregator contains a pre-trained Xception model made by a depth-wise separable convolution layer. The depth-wise separable convolution layer used in the model splits each channel of the input and filter separately. The layer convolves them by each channel and later separates one element of 3 channels to be convoluted until all aspects have been convoluted. The architecture also has some shortcut structure that skips over the block of the depth-wise separable convolution layers. The model uses a categorical cross-entropy loss function as the metric loss measurement. For more details on the FIT machine, refer to [13].

3.3. *Deep Learning Models.*

The last stage of the HIT machine is the deep learning models. Our HAR dataset is trained with deep learning models and classifies user activities into sitting, standing, lying, walking, push up, dancing, sitting, running, and jumping. The HAR dataset consists of image samples, and our system considers four image classification models VGG, ResNet, Inception, and EfficientNet, as the deep learning models. Figure 5 shows the deep learning models used by our HIT machine.



(a)

(b)

(c)

(d)

Figure 5: Deep learning models used in the HIT machine. (a) VGG. (b) Inception. (c) ResNet. (d) EfficientNet.

The most common image classification model is the VGG model introduced by the visual graphics at University of Oxford [14]. The VGG model consists of 13 convolution layers, five pooling layers, and three dense layers. The VGG model is sequential in nature and uses many filters one after another. The architecture uses a stack of convolutional layers with different depths in different architectures followed by three fully-connected (FC) layers. The first two FC layers have 4,096 channels each, and the third FC performs the 1,000-way classification. The last layer is the soft-max layer that is used to normalize the classification vector. All the hidden layers in the VGG architecture use rectified linear unit (ReLU) as the activation function. The ReLU activation function is computationally efficient, and its results are in faster learning. The ReLU function also reduces the likelihood of vanishing gradient problems and improves the classification performance. Figure 5(a) shows the architecture of the VGG network.

Next, our HIT machine used a deep learning model, which was developed by Google [16]. The GoogLeNet or Inception is a smaller network than the VGG model and uses an Inception module. The Inception module performs convolutions with different filter sizes on the input images, performs Max Pooling, and concatenates the result for the next Inception module. The architecture uses a 1×1 convolution operation which reduces the parameters drastically. This architecture is designed to solve the problem of computational expense, overfitting, and other deep learning model issues. The Inception model takes advantage of the multiple kernel filter sizes within the CNN, and rather than stacking them sequentially, it orders them to operate on the same level. Figure 5(b) shows the Inception architecture used by our HIT machine. The architecture has nine inception modules stacked linearly and has 22 layers deep (27, including the pooling layers). It uses global average pooling at the end of the last inception module. Compared with VGG networks, Inception networks are more computationally efficient in terms of the number of parameters generated by the network and the computational cost incurred. For more details on the Inception model, refer to [16].

Our HIT machine also analyzed the impact of the ResNet architecture for activity recognition. The main idea of ResNet architecture is to avoid poor accuracy when the model uses deeper layers. This model is mainly designed for the gradient vanishing problem. Figure 5(c) shows the ResNet architecture used by our HIT machine. The ResNet architecture is a 34-layer plain network inspired by VGG-19 networks, which adds shortcut connections. These shortcut connections then convert the ResNet architecture into the residual network. The first two layers of the model are the same as those of the Inception model. The model uses a 7×7 convolution layer with 64 output channels and a stride of 2 followed by the 3×3 maximum pooling layer. The major difference with ResNet is the batch normalization layer which is added after each convolutional layer. The inception model discussed previously uses four modules which are made up of Inception blocks. However, the ResNet architecture uses four modules which are made up of residual blocks. Each residual block uses several residual blocks with the same number of output channels. The first module from the architecture uses the number of channels that are the same as the input channel numbers. From the first residual block of each subsequent module, the number of channels is doubled compared with the previous module, and the height and width are halved. Compared with Inception architecture, the ResNet model is more straightforward, easy to modify, easy to optimize, and achieves higher accuracy when the depth of the network increases. For more details on ResNet architecture and its implementation, refer to [15].

At last, our HIT machine used a model called EfficientNet from Google for activity recognition [17]. In EfficientNet, a new scaling method called compound scaling is introduced. The model ResNet discussed before follows a conventional approach of scaling the dimensions arbitrarily and adding more layers. However, if the model scales the dimensions by a fixed amount simultaneously and does so uniformly, the model achieves better performance. The user can decide the scaling coefficients. EfficientNet architecture is a convolutional neural network architecture with different scaling methods. In EfficientNet, the architecture uniformly scales all depth/width/resolution dimensions using a compound coefficient. Compared with conventional ways that arbitrarily scale these factors, the scaling method in the EfficientNet architecture uniformly scales network width, depth, and resolution with a set of fixed-scaling coefficients. Figure 5(d) shows the EffientNet architecture used by our HIT machine. The main building block of this architecture consists of mobile inverted bottleneck Convolution (MBConv), to which squeeze-and-excitation optimization is added. The MBConv layer is similar to the inverted residual blocks used in MobileNet v2 [97]. The MBConv creates a shortcut connection between the beginning and end of a convolutional block. The input activation maps are first expanded using 1×1 convolutions, increasing the depth of the feature maps. 3×3 depth-wise convolutions and point-wise convolutions follow this, and this structure reduces the number of channels in the output feature map. The shortcut connections connect the narrow layers, while the wider layers are present between the skip connections. This form of structure decreases the overall number of operations required as well as the model size. For more details on the EfficientNet architecture and its implementation, refer to [17].

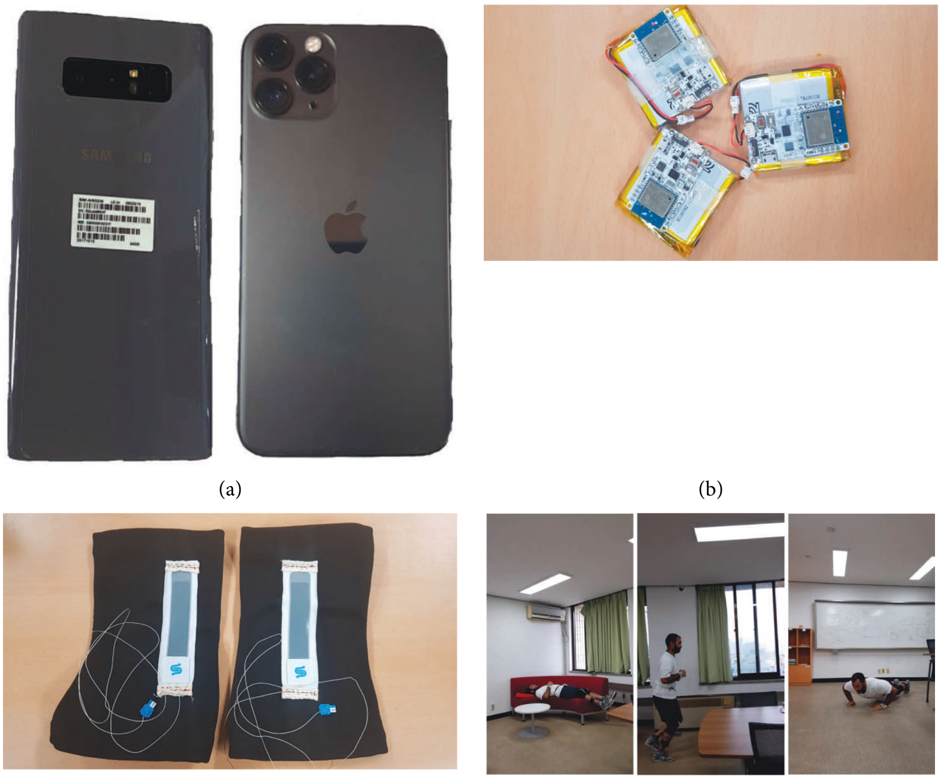
4. Experiment Results and Analysis

We collected HAR datasets from different users to validate our proposed HIT machine-based HAR approach. There were 10 volunteers for data collection, consisting of five members for the training dataset and five for the unseen dataset. The demographic information of participants is given in Table 1.

We used Samsung galaxy note eight and iPhone 11 pro smartphone models for video recording. The smartphones were kept stationary during the initial stage of the experiment and moved their positions based on the user's motions. The users made their activities within the 15 m experiment area. We also used the IMU and stretch sensors and recorded the sensor reading from the users' activities during the

Table 1: Demographics of participants.

Subjects	1	2	3	4	5	6	7	8	9	10
Age	30	21	35	22	28	26	30	32	35	23
Height (cm)	175	180	172	160	174	162	176	165	168	159
Weight (kg)	80	84	87	60	70	58	78	62	85	57
Gender	M	M	M	F	M	F	M	F	M	F
Training dataset	√	√		√			√			√
Unseen dataset			√		√	√		√	√	



(a) (b) (c) (d)

Figure 6: Experiment setup. (a) Smartphones. (b) IMU sensor. (c) Stretch sensor. (d) Experiment area.

experiment time. Our conventional HAR approaches use the sensor reading for activity recognition, and we compared these HAR results with our HIT machine approach. Figure 6 shows the smartphones, IMU and stretch sensors, and experiment area involved in the HAR data collection. Table 2 summarizes our system configurations and hyperparameters used for model training and testing.

We started the analysis of the HIT machine by implementing deep learning models, such as VGG, Inception, ResNet, and EfficientNet. We tested these models with our HAR dataset, and Figure 7 shows the classification results from each model. We used confusion matrices to analyze each model, summarizing the classification performance. The color bars indicate the number of samples populated in a specific area. When the data samples are higher, the color becomes lighter and vice versa. The results observed in confusion matrices show that the ResNet architecture has the highest classification performance compared with other models and achieved a 98.53% model accuracy, 0.20 model loss, 98.56% precision, 98.53% recall, and 98.54% F1 scores. The VGG model reached 96.38% model accuracy with 0.09 model loss, 96.58% precision, 96.38% recall, and 96.36% F1 score as shown in Figure 7(a). The VGG model has a higher classification accuracy for sitting, sit-up, standing, and walking activities. The model has the highest misclassification error for running. Some of the running activity is misclassified as walking. Figure 7(b) shows the classification results from the Inception model. This model achieved a 93.18% classification accuracy with 0.13 model loss, 93.18% precision and recall, and 93.11% F1 scores, which are worse performances than the results obtained by the VGG model. Furthermore, Figure 7(c) shows the best classification results from our HIT machine based on ResNet architecture. The ResNet architecture showed the best model accuracy with the least classification errors. However, the model loss is higher than other models and needs higher computation time than VGG and Inception models. This model maintains the classification accuracy for basic and complex activities,

Table 2: System configurations and hyperparameters used for model training and testing.

System configuration	Description
Processor	Intel® core^{TM} i7-11700k
RAM	32 GB
Graphics card	GeForce RTX^{TM} 3070 Ti
Python version	3.8
Tensorflow version	tf-nightly = = 2.6.0
Keras version	2.6.0
cuDNN library	cuDNN v8.1.0
CUDA version	CUDA toolkit 11.2.0
Model parameter	Value
Ratio of training data to overall data	0.70
Input image size	224 × 224
Number of channels	1
Optimizer	Adam
Learning rate	0.02
Batch size	128
Loss	Categorical cross-entropy
Number of classes	9
Epochs	25

and the model is the best choice for HIT machine-based activity recognition. Figure 7(d) shows our last deep learning model results from EfficientNet. The EfficientNet reached 89.94% for classification accuracy with 0.21 model loss, 90.19% precision, and 89.94% recall and F1 score, which has worse HAR performance than VGG, Inception, and ResNet models. The higher level of classification error from EfficientNet shows that this model is unsuitable for our HIT machine-based activity recognition. Figures 8 and 9 show the deep learning models accuracy and loss plots, and Table 3 summarizes their performance.

In Table 3, we used the accuracy, loss, precision, recall, and F1 score parameters for performance evaluation. The following equations from [98] define these parameters.

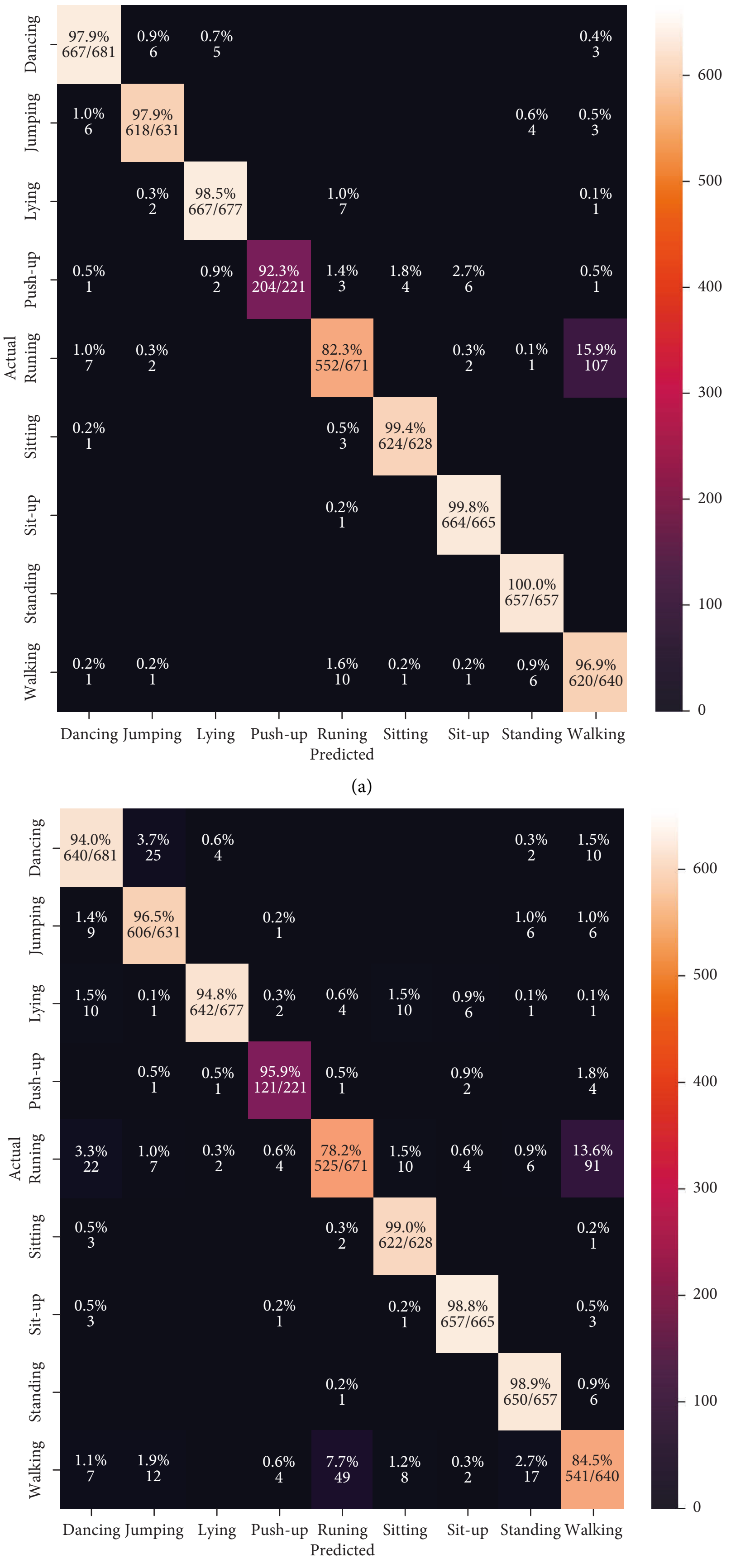
$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN},\\
\text{Loss} &= -\sum_{i-1}^{\text{outputsize}} y_i.\log\widehat{y}_i,\\
\text{Precision} &= \frac{TP}{TP + FP},\\
\text{Recall} &= \frac{TP}{TP + FN},\\
F1 &= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}},
\end{aligned} \tag{1}$$

where the variables TP, TN, FP, and FN are defined as true positive, true negative, false-positive and false-negative in a given experiment. In the loss function, y_i is the ith scalar value in the model output, $\widehat{y}_i$ is the corresponding target value, and the output size is the number of scalar values in the model output. From the results in Table 3, the ResNet architecture outperforms the other deep learning models with an average value of 98.53%. These results indicate that the system trained with the ResNet model is the best choice for activity recognition.

When we consider the training time results from Table 3, it shows that the ResNet-based HAR approach has a higher training time (600 s) than other models. This is due to the deep architecture of ResNet, and the system takes more time to train the model. However, the activity recognition results from ResNet compensate for the training time when considering the overall system performance (6.44% of classification improvements than EfficientNet model-based HAR system). In the case of VGG model-based HAR, the system achieved the most down training time (240 s) compared with other models and reached good classification results for activity recognition. The Inception model-based HAR system has a 60 s time difference for model training compared with the EfficientNet model. The EfficientNet has a lower training time (300 s) than the Inception model-based (360 s) HAR system. However, the EfficientNet-based HAR approach shows worst classification results than other models.

To further validate our HIT machine performance, we tested the pre-trained deep learning models with unseen HAR datasets. We collected another set of HAR datasets and tested them with our pre-trained models. Table 4 summarizes the results for unseen datasets from pre-trained models. The results in Table 4 show that the ResNet architecture achieved 72.13% for classification accuracy with 72.25% precision, 72.92% recall, and 72.95% F1 scores. These results outperformance the other pre-trained models. However, the computational complexity of this architecture makes it more practically challenging for real-time HAR applications. The classification accuracy from the Inception model shows that the model reached 65.17% for classification accuracy with 65.21% precision, 65.08% recall, and 65.59% F1 scores. The results from the Inception-based pre-trained model give better results than VGG and EfficientNet pre-trained models. In the case of VGG based pre-trained model, the system shows 61.85% for classification accuracy with 61.73% precision, 61.48% recall, and 61.47% F1 scores. The EfficientNet pre-trained model-based HAR system shows 57.42% for classification accuracy with 57.48% precision, 57.43% recall, and 57.72% F1 scores. These results show the worst classification results compared with other pre-trained models, and the approach is unsuitable for image-based HAR systems.

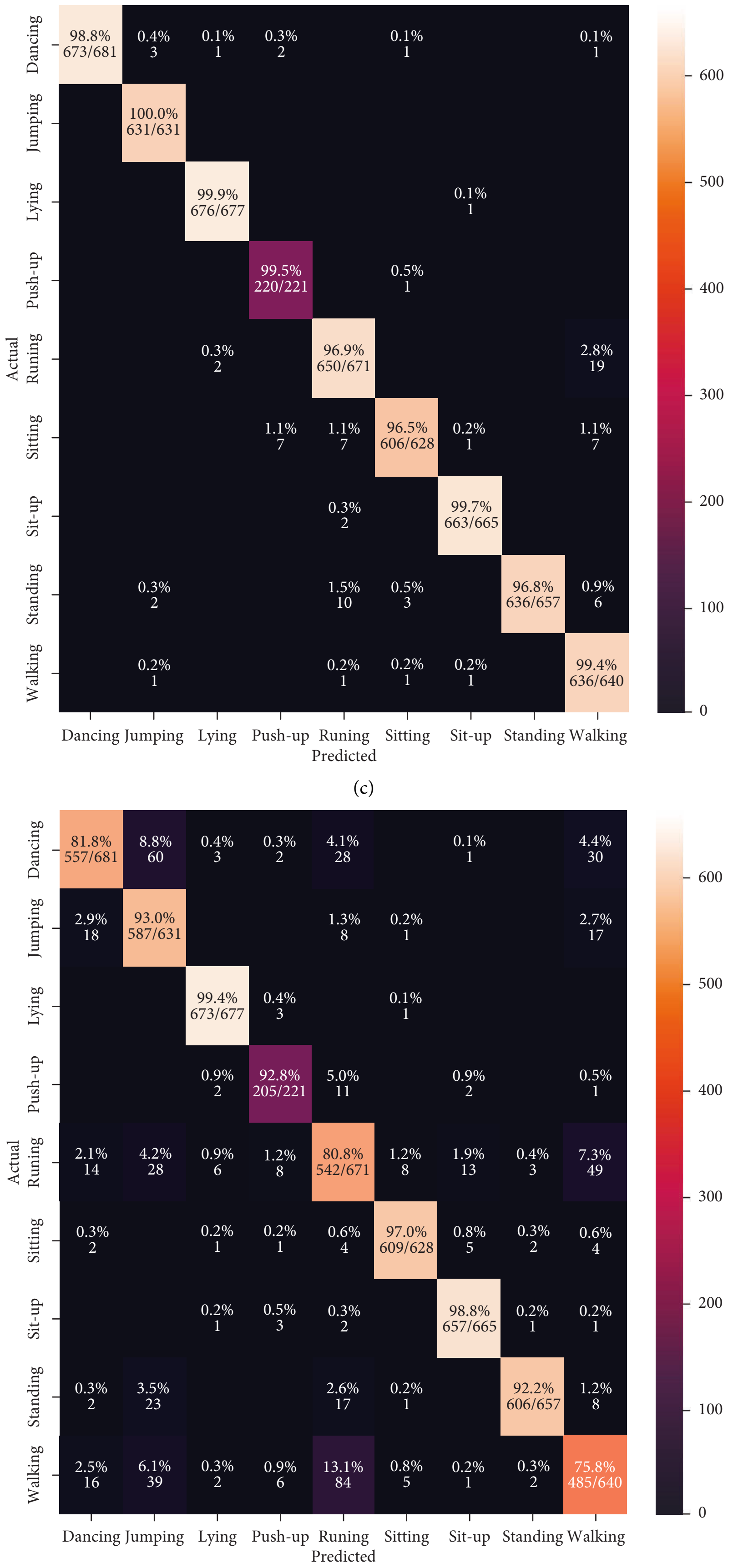
Next, we validated our HIT machine results with sensor-based HAR approaches and image-based HAR without HIT machine. Figure 10 shows the classification results from our HIT machine, HAR without HIT machine, and sensor-based techniques. This analysis uses a 2D CNN model for activity recognition. The CNN model is computationally lighter than other deep learning models and easily fits IMU and stretch sensor datasets. Figure 10(a) shows the classification results from IMU sensor-based HAR approach. The results show that the IMU sensor approach reached 90.71% of classification accuracy with 0.27 model loss, 90.47% precision, 90.71% recall, and 90.00% F1 scores. The activities that include running, sitting, sit-up, standing, and walking have higher classification errors due to the similarities of IMU



(a)

(b)

Figure 7: Continued.



(c)

(d)

Figure 7: The confusion matrix results. (a) VGG. (b) Inception. (c) ResNet. (d) EfficientNet.

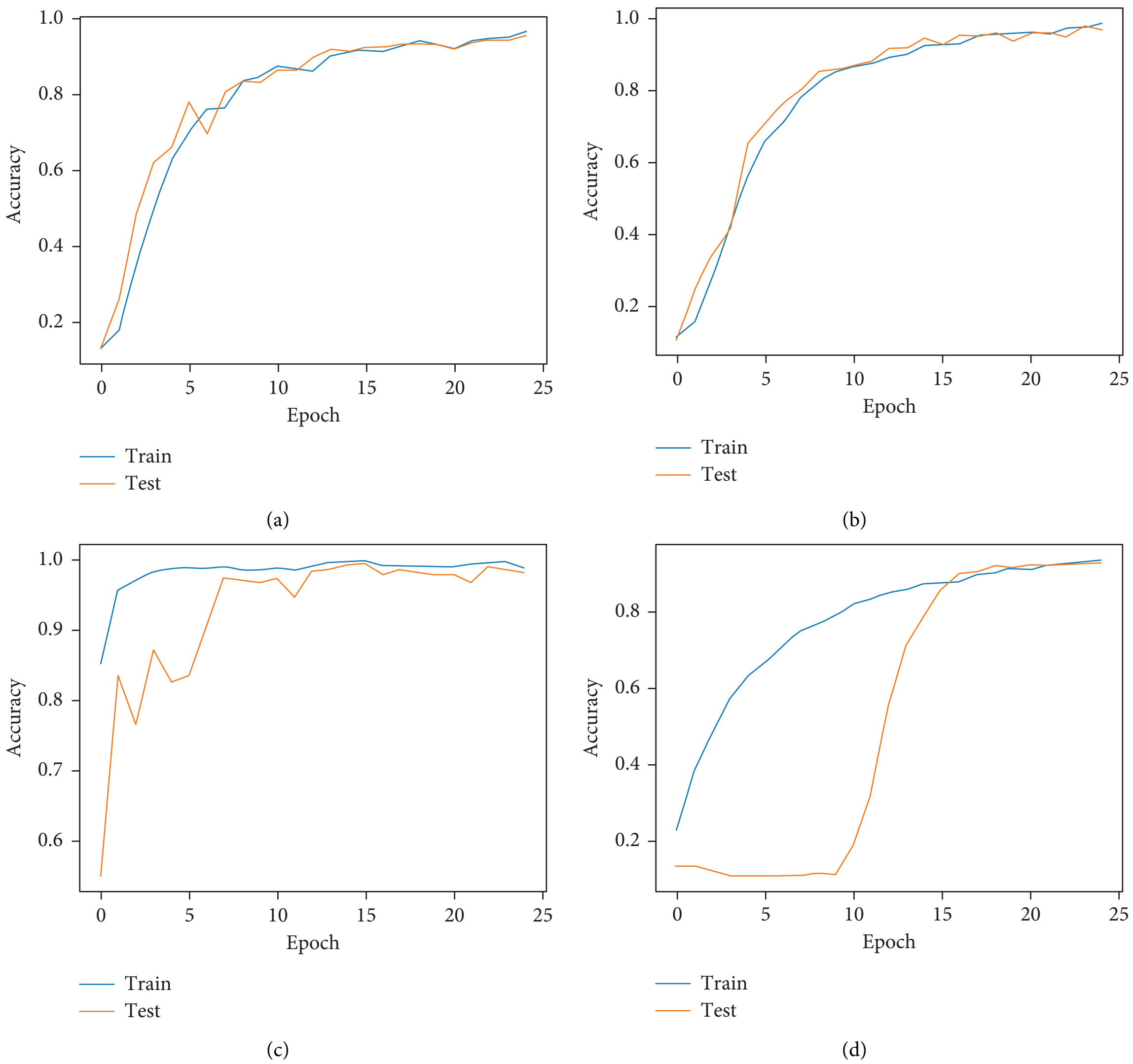


Figure 8: Deep learning models accuracy plots. (a) VGG. (b) Inception. (c) ResNet. (d) EfficientNet.

sensor data. The model fails to classify these activities, increasing the classification errors in the HAR system. When the system uses a stretch sensor instead of an IMU sensor, the classification performance has a 3% improvement. The stretch sensor-based HAR system achieved 93.80% of classification accuracy with 0.27 model loss, 94.16% precision, 93.80% recall, and 93.20% F1 scores. Figure 10(b) shows the classification results from stretch sensor-based HAR approach. The stretch sensor data are more stable than the IMU sensor and have accurate HAR results. The activities that include sitting and walking have higher classification errors than the IMU sensor-based approach. The stretch sensor-based HAR approach is reasonable if the system cost is not a primary concern. The prohibitive cost of the stretch sensor makes the system more challenging for practical health care applications. Next, we analyzed a HAR approach that uses image data without a HIT machine. Figure 10(c) shows the results from a HAR without HIT machine. The HAR system without HIT machine reached 90.98% of classification accuracy with 0.20 model loss, 91.24% precision, 90.98% recall, and 90.90% F1 scores. These results indicate the significance of the HIT machine. Compared with the results from Figure 10(d), the system without a HIT machine has a higher classification error and shows the worst performance for both basic and complex activities. The results from Figure 10(d) show the classification performance of the HIT machine, which has the best performance compared with other HAR approaches. The system achieved a 6.01% accuracy improvement compared with the IMU sensor-based approach and 2.4% accuracy improvement compared with the stretch sensor-based approach. The system also has a 5.3% accuracy improvement compared with the HAR approach without HIT machine. Our proposed HIT machine-based HAR system show 96.28% of classification accuracy with 0.09 model loss, 96.26% precision, 96.28% recall, and 96.27% F1 scores. Table 5 summarizes the performance of each approach in terms of accuracy, loss, precision, recall, and F1 score. From Table 5 results, the HIT machine shows the highest classification results than the sensor-based and without HIT

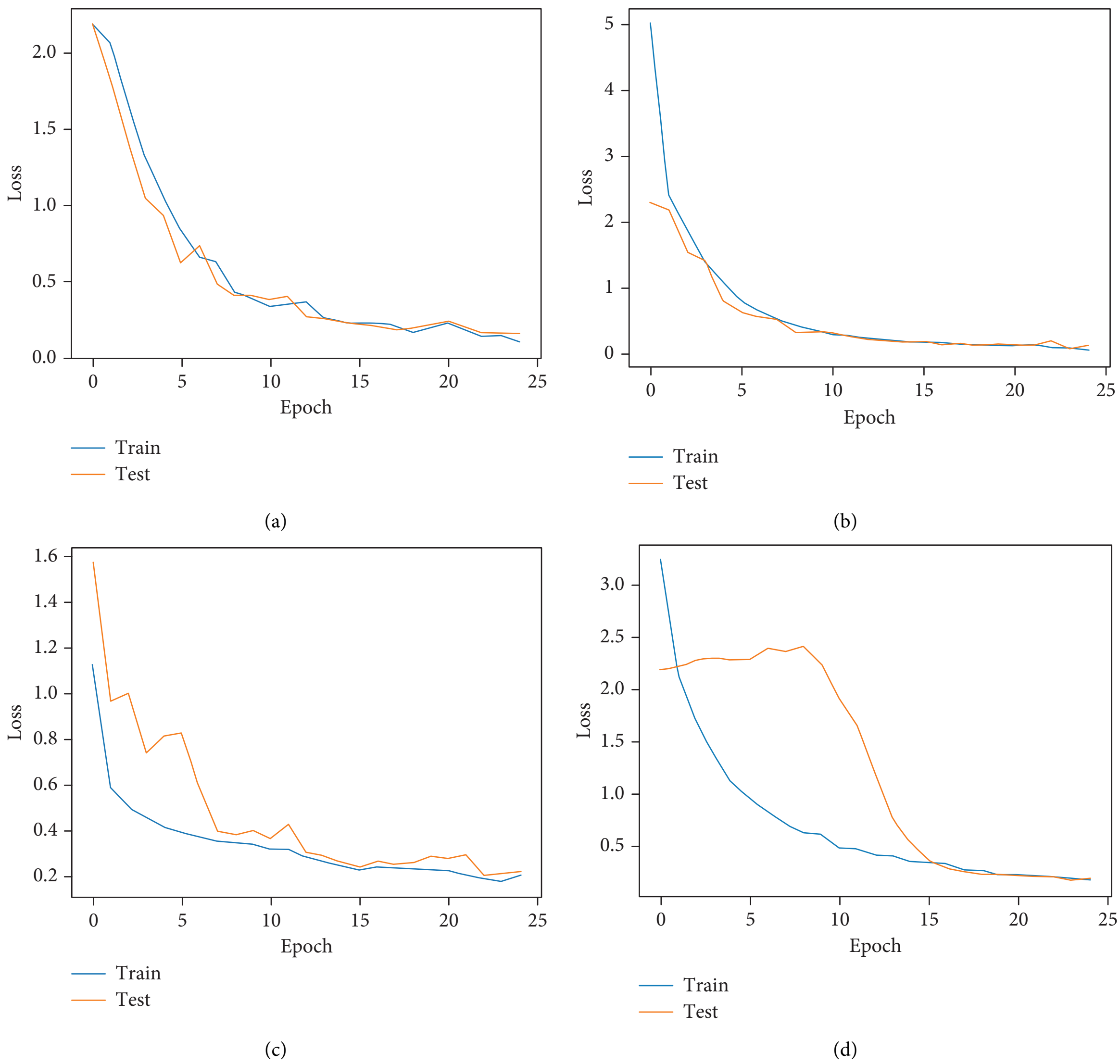


Figure 9: Deep learning models loss plots. (a) VGG. (b) Inception. (c) ResNet. (d) EfficientNet.

Table 3: Performance comparison of deep learning models used in the HIT machine.

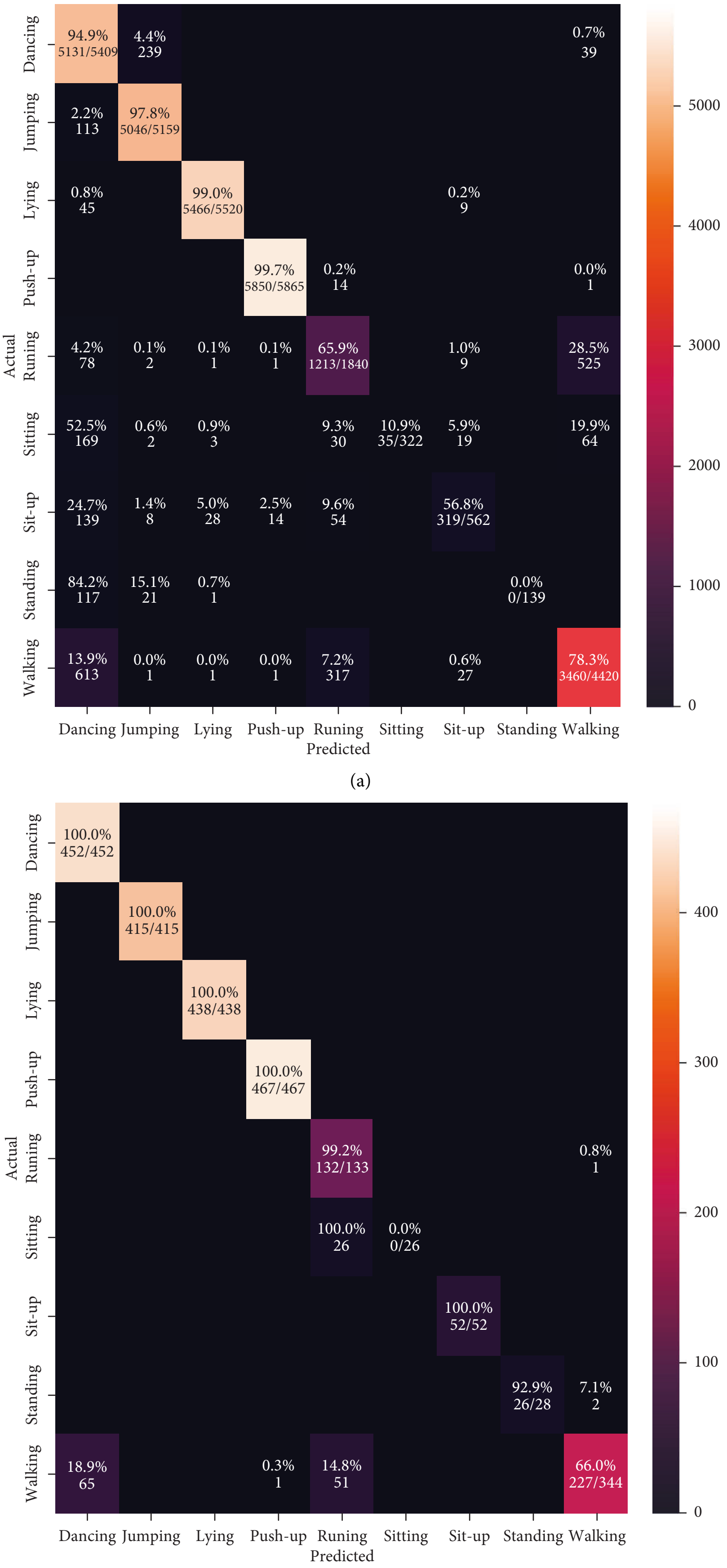
Deep learning model	Accuracy	Loss	Precision	Recall	F1 score	Training time (secs)
VGG	96.38	0.09	96.58	96.38	96.36	240
Inception	93.18	0.13	93.18	93.18	93.11	360
ResNet	98.53	0.20	98.56	98.53	98.54	600
EfficientNet	89.94	0.21	90.19	89.94	89.94	300

machine-based HAR approaches. The results indicate the impact of the HIT machine-based activity recognition for complex activities.

Table 4: Performance comparison of pre-trained deep learning models for unseen HAR datasets.

Pre-trained model	Accuracy	Precision	Recall	F1 score
VGG	61.85	61.73	61.48	61.47
Inception	65.17	65.21	65.08	65.59
ResNet	72.13	72.25	72.92	72.95
EfficientNet	57.42	57.48	57.43	57.72

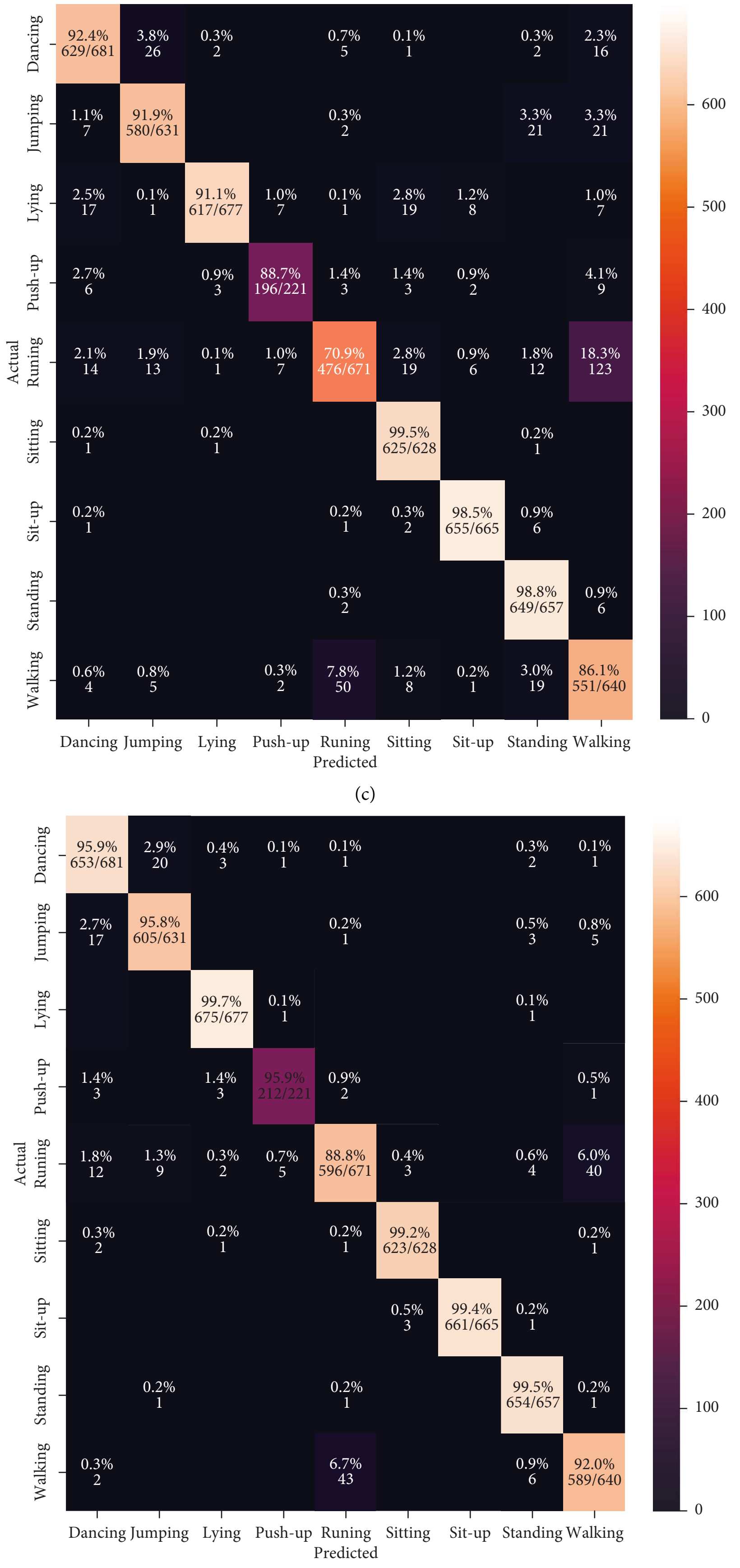
The training time results from Table 5 indicate that the stretch sensor-based HAR system shows the best training time (120 s) than the other HAR systems. This is due to the small number of data samples from the stretch sensor dataset. In the case of the IMU sensor-based HAR approach, the system has a 300 s training time, which is 180 s higher than the stretch sensor-based HAR approach. Also, the classification accuracy from the IMU sensor-based HAR approach is 3.09% lower than the stretch sensor-based approach. The proposed HIT machine-based HAR approach shows 340 s training time, which is lower than HAR without HIT machine-based approach (480 s). The training time



(a)

(b)

Figure 10: Continued.



(c)

(d)

Figure 10: The confusion matrix results. (a) IMU sensor-based approach. (b) Stretch sensor-based approach. (c) HAR without HIT machine. (d) Proposed HIT machine-based approach.

Table 5: Performance comparison of HAR approaches.

HAR approach	Accuracy	Loss	Precision	Recall	F1 score	Training time (secs)
IMU sensor	90.71	0.27	90.47	90.71	90.00	300
Stretch sensor	93.80	0.27	94.16	93.80	93.20	120
HAR without HIT machine	90.98	0.20	91.24	90.98	90.90	480
Proposed HIT machine	96.28	0.09	96.26	96.28	96.27	340

results from our proposed HIT machine indicate that the approach reduced 140 s of training time compared with HAR without a HIT machine-based approach.

From the experiment and result analysis, it can be seen that the HIT machine-based HAR approach has a significant role in activity recognition. The proposed HAR system addresses the primary vision-based HAR system's challenge, such as processing high-quality images. We used image cropping, resizing, and data cleaning to make the system can perform the high-quality images without compromising the classification results. Our system takes advantage of the mask R-CNN algorithm, which is computationally lighter than other vision algorithms. The proposed method also solves the camera viewpoint and background clutter issues by considering the smartphone camera's wide-angle feature. The classification results from the HIT machine show that the proposed HAR approach is a valid method for healthcare applications, including abnormal activity detection, elderly care in homes, and disabled assistance. The extended versions of HIT machines are helpful in other applications, including intelligent environments, indoor navigation [99], security and surveillance, and people monitoring [100].

5. Conclusion

This paper proposed a HIT machine-based HAR system for healthcare applications. The proposed HIT machine approach effectively utilizes the advantages of the mask R-CNN for human body estimation and enhances the performance of the HAR. The classification results from our experiments indicate that the proposed HIT machine has better classification results than conventional sensor-based HAR approaches. The traditional sensor-based HAR systems are not free from sensor errors, showing very poor classification results for complex activities. The proposed HIT-based HAR system is suitable for basic and complex user movements and maintains its classification accuracy in all user motions. Our HAR classification results and analysis show the influence of the HIT machine for activity recognition. The proposed HIT machine-based HAR system is a suitable healthcare option if HAR systems use a camera as their input device. We validated our proposed HIT machine-based HAR system for human activity recognition through extensive experiments and analysis. To improve the classification performance, we intend to use a sensor fusion technique that combines the image and sensor data for activity recognition in our future work. Furthermore, we will consider the most popular public datasets (UCI- human activity recognition using smartphones dataset) for future research and compare our HAR datasets' performance with public datasets.

Data Availability

The data used to support the findings of this study have not been made available because of the privacy of the research participant.

Conflicts of Interest

The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03043144).

References

[1] M. Ronald, A. Poulose, and D. S. Han, "iSPLInception: an inception-ResNet deep learning architecture for human activity recognition," *IEEE Access*, vol. 9, pp. 68985–69001, 2021.

[2] O. Steven Eyobu and D. S. Han, "Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network," *Sensors*, vol. 18, no. 9, p. 2892, 2018.

[3] T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse, "An activity monitoring system for elderly care using generative and discriminative models," *Personal and Ubiquitous Computing*, vol. 14, no. 6, pp. 489–498, 2010.

[4] N. T. H. Thu, D. S. Han, and H. HAR, "A hierarchical hybrid deep learning architecture for wearable sensor-based human activity recognition," *IEEE Access*, vol. 9, pp. 145271–145281, 2021.

[5] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942, Miami, FL, USA, June 2009.

[6] Y. Ohgi, M. Yasumura, H. Ichikawa, and C. Miyaji, *Analysis of Stroke Technique Using Acceleration Sensor IC in Freestyle Swimming*, pp. 503–511, Blackwell Science, Oxford, UK, 2000.

[7] H. A. Ullah, S. Letchmunan, M. S. Zia, U. M. Butt, and F. H. Hassan, "Analysis of Deep Neural Networks for Human Activity Recognition in Videos–A Systematic Literature Review," *IEEE Access*, vol. 12, 2021.

[8] J. Liu, G. Teng, and F. Hong, "Human activity sensing with wireless signals: a survey," *Sensors*, vol. 20, no. 4, p. 1210, 2020.

[9] S. Ashry, T. Ogawa, and W. Gomaa, "Charm-deep: continuous human activity recognition model based on deep

neural network using IMU sensors of smartwatch," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8757–8770, 2020.

[10] X. Ren and C. Gu, "Figure-ground Segmentation Improves Handled Object Recognition in Egocentric Video," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3137–3144, San Francisco, USA, June 2010.

[11] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient Model-Based 3D Tracking of Hand Articulations Using Kinect," *BmVC*, vol. 3, 2011.

[12] J. Lei, X. Ren, and D. Fox, "Fine-grained kitchen activity recognition using rgb-d," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 208–211, Pittsburgh, PA, USA, September 2012.

[13] J. H. Kim, A. Poulose, and D. S. Han, "The extensive usage of the facial image threshing machine for facial emotion recognition performance," *Sensors*, vol. 21, no. 6, p. 2026, 2021.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, https://arxiv.org/abs/1409.1556.

[15] 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. 2818–2826, Las Vegas, Nevada, United States, July 1 2016.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, Nevada, United States, June 26.

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

[18] J. M. Rodriguez-Borbon, X. Ma, A. K. Roy-Chowdhury, and W. A. Najjar, "Heterogeneous acceleration of HAR applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 888–902, 2020.

[19] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, 2016.

[20] G. Yuan, Z. Wang, F. Meng, Q. Yan, and S. Xia, "An Overview of Human Activity Recognition Based on Smartphone," *Sensor Review*, vol. 1, 2019.

[21] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: a survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[22] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.

[23] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[24] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: a comprehensive survey," *IEEE Access*, vol. 8, pp. 83791–83820, 2020.

[25] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13029–13040, 2021.

[26] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition–a systematic review of literature," *IEEE Access*, vol. 6, pp. 59192–59210, 2018.

[27] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[28] R. Alazrai, M. Hababeh, B. A. Alsaify, M. Z. Ali, and M. I. Daoud, "An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals," *IEEE Access*, vol. 8, pp. 197695–197710, 2020.

[29] M. Muaaz, A. Chelli, A. A. Abdelgawwad, A. C. Mallofré, and M. Pätzold, "WiWeHAR: multimodal human activity recognition using wi-fi and wearable sensing modalities," *IEEE Access*, vol. 8, pp. 164453–164470, 2020.

[30] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[31] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41-42, pp. 30509–30555, 2020.

[32] F. Wang, J. Liu, and W. Gong, "Multi-adversarial in-car activity recognition using RFIDs," *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2224–2237, 2021.

[33] M. A. R. Ahad, A. D. Antar, and M. Ahmed, "IoT Sensor-Based Activity Recognition," *IoT Sensor-Based Activity Recognition*, vol. 2, 2020.

[34] A. D. Antar, M. Ahmed, and M. A. R. Ahad, "Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review," in *Proceedings of the 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 134–139, Spokane, WA, USA, June 2019.

[35] P. Saengthong and S. Laitrakun, "Fusion approaches of heterogeneous multichannel CNN and LSTM models for human activity recognition using wearable sensors," in *Proceedings of the 2021 13th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 183–188, Chiang Mai, Thailand, October 2021.

[36] J. Lu, X. Zheng, M. Sheng, J. Jin, and S. Yu, "Efficient human activity recognition using a single wearable sensor," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11137–11146, 2020.

[37] C.-T. Yen, J.-X. Liao, and Y.-K. Huang, "Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms," *IEEE Access*, vol. 8, pp. 174105–174114, 2020.

[38] I. A. Lawal and S. Bano, "Deep human activity recognition with localisation of wearable sensors," *IEEE Access*, vol. 8, pp. 155060–155070, 2020.

[39] Y. Tang, Q. Teng, L. Zhang, F. Min, and J. He, "Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 1, pp. 581–592, 2021.

[40] W. Qi, H. Su, and A. Aliverti, "A smartphone-based adaptive recognition and real-time monitoring system for human activities," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 414–423, 2020.

[41] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[42] A. Wang, S. Zhao, C. Zheng, H. Chen, L. Liu, and G. Chen, "HierHAR: sensor-based data-driven hierarchical human

activity recognition," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3353–3365, 2021.

[43] Z. Chen, C. Jiang, S. Xiang, J. Ding, M. Wu, and X. Li, "Smartphone sensor-based human activity recognition using feature fusion and maximum full a posteriori," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 3992–4001, 2020.

[44] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, and M. Linden, "A comparative analysis of hybrid deep learning models for human activity recognition," *Sensors*, vol. 20, no. 19, p. 5707, 2020.

[45] X. Li, Y. Wang, B. Zhang, and J. Ma, "PSDRNN: an efficient and effective HAR scheme based on feature extraction and deep learning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6703–6713, 2020.

[46] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: a deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.

[47] I. K. Ihianle, A. O. Nwajana, S. H. Ebenuwa, R. I. Otuka, K. Owa, and M. O. Orisatoki, "A deep learning approach for human activities recognition from multimodal sensing devices," *IEEE Access*, vol. 8, pp. 179028–179038, 2020.

[48] E. Kim, "Interpretable and accurate convolutional neural networks for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 7190–7198, 2020.

[49] O. Barut, L. Zhou, and Y. Luo, "Multitask LSTM model for human activity recognition and intensity estimation using wearable sensor data," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8760–8768, 2020.

[50] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[51] M. Munoz-Organero, "Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs," *IEEE Access*, vol. 7, pp. 74422–74436, 2019.

[52] X. a. Li, J. Luo, and R. Younes, "ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition," in *Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pp. 249–254, Virtual Conference, September 2020.

[53] Z. Chen, C. Jiang, and L. Xie, "A novel ensemble ELM for human activity recognition using smartphone sensors," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2691–2699, 2019.

[54] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja, and R. Sarkar, "A new framework for smartphone sensor-based human activity recognition using graph neural network," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11461–11468, 2021.

[55] Q. Zhu, Z. Chen, and Y. C. Soh, "A novel semisupervised deep learning method for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3821–3830, 2019.

[56] Y.-L. Hsu, S.-C. Yang, H.-C. Chang, and H.-C. Lai, "Human daily and sport activity recognition using a wearable inertial sensor network," *IEEE Access*, vol. 6, pp. 31715–31728, 2018.

[57] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing daily activities with RFID-based sensors," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, pp. 51–60, Orlando Florida, September 2009.

[58] P. Hevesi, S. Wille, G. Pirkl, N. Wehn, and P. Lukowicz, "Monitoring household activities and user location with a cheap, unobtrusive thermal sensor array," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pp. 141–145, Seattle, US, September, 2014.

[59] P. Rashidi and D. J. Cook, "Mining Sensor Streams for Discovering Human Activity Patterns over Time," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 431–440, Sydney, NSW, Australia, December 2010.

[60] S. Shi, S. Sigg, and Y. Ji, "Joint localization and activity recognition from ambient FM broadcast signals," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, pp. 521–530, Zurich, Switzerland, September 2013.

[61] M. Sekine and K. Maeno, "Activity recognition using radio Doppler effect for human monitoring service," *Journal of Information Processing*, vol. 20, no. 2, pp. 396–405, 2012.

[62] D. Jiang, M. Li, and C. Xu, "WiGAN: a WiFi based gesture recognition system with GANs," *Sensors*, vol. 20, no. 17, p. 4757, 2020.

[63] B. Kellogg, V. Talla, and S. Gollakota, "Bringing gesture recognition to all devices," in *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, pp. 303–316, Seattle, WA, USA, April 2014.

[64] X. Qi, G. Zhou, Y. Li, and G. Peng, "Radiosense: exploiting wireless communication patterns for body sensor network activity recognition," in *Proceedings of the 2012 IEEE 33rd Real-Time Systems Symposium*, pp. 95–104, San Juan, PR, USA, December 2012.

[65] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, 2017.

[66] F. Wang, W. Gong, J. Liu, and K. Wu, "Channel selective activity recognition with WiFi: a deep learning approach exploring wideband information," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 181–192, 2020.

[67] F. Wang, W. Gong, and J. Liu, "On spatial diversity in WiFi-based human activity recognition: a deep learning-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2035–2047, 2019.

[68] S. Wang and G. Zhou, "A review on radio based activity recognition," *Digital Communications and Networks*, vol. 1, pp. 20–29, 2015.

[69] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, pp. 1–31, Article ID 3090343, 2017.

[70] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, pp. 88–131, 2013.

[71] A. Jalal, M. Z. Uddin, and T.-S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, 2012.

[72] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1383–1394, 2013.

[73] H. Zhang and L. E. Parker, “Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 541–555, 2016.

[74] S. Gaglio, G. L. Re, and M. Morana, “Human activity recognition process using 3-D posture data,” *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586–597, 2015.

[75] X. Yang and Y. Tian, “Super normal vector for human activity recognition with depth cameras,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 1028–1039, 2017.

[76] M. Ajmal, F. Ahmad, M. Naseer, and M. Jamjoom, “Recognizing human activities from video using weakly supervised contextual features,” *IEEE Access*, vol. 7, pp. 98420–98435, 2019.

[77] B. Ayhan, C. Kwan, B. Budavari, J. Larkin, D. Gribben, and B. Li, “Video activity recognition with varying rhythms,” *IEEE Access*, vol. 8, pp. 191997–192008, 2020.

[78] R. Raj and A. Kos, “Different techniques for human activity recognition,” in *Proceedings of the 2022 29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES)*, pp. 171–176, Wrocław, Poland, June 2022.

[79] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: a review,” *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[80] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detection: survey and experiments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[81] J. K. Aggarwal and S. Park, “Human motion: modeling and recognition of actions and interactions,” in *Proceedings of the 2nd International Symposium on 3D Data Processing*, pp. 640–647, IEEE, NY China, June 2004.

[82] M. Valera and S. A. Velastin, “Intelligent distributed surveillance systems: a review,” *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, 2005.

[83] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[84] V. Krüger, D. Kragic, A. Ude, and C. Geib, “The meaning of action: a review on action recognition and mapping,” *Advanced Robotics*, vol. 21, no. 13, pp. 1473–1501, 2007.

[85] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, “Machine recognition of human activities: a survey,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[86] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: a survey on human behavior-recognition algorithms,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 2010.

[87] Y. Liu, J. Yuan, and Z. Tu, “Motion-driven Visual Tempo Learning for Video-Based Action Recognition,” *IEEE Transactions on Image Processing*, vol. 21, 2022.

[88] M. M. E. Yurtsever and S. Eken, “BabyPose: real-time decoding of baby’s non-verbal communication using 2D video-based pose estimation,” *IEEE Sensors Journal*, vol. 22, no. 14, pp. 13776–13784, 2022.

[89] C. Han, L. Zhang, Y. Tang et al., “Understanding and improving channel attention for human activity recognition by temporal-aware and modality-aware embedding,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[90] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, October 2017.

[91] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936–944, Honolulu, HI, USA, July 2017.

[92] Z. Yin, X. Wang, and L. Li, “Optimization of human body attitude detection based on mask RCNN,” in *Proceedings of the 2020 8th International Conference on Orange Technology (ICOT)*, pp. 1–4, Daegu, South Korea, December 2020.

[93] Y. Wang, J. Wu, and H. Li, “Human detection based on improved mask R-CNN,” *Journal of Physics: Conference Series*, vol. 1575, no. 1, Article ID 012067, 2020.

[94] X. Li and S. Cheng, “Pedestrian gender detection based on mask R-CNN,” in *Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, Chengdu, China, December 2019.

[95] A. Rosebrock, “Deep learning for computer vision with Python: starter bundle,” *PyImageSearch*, 2017.

[96] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1251–1258, Honolulu, HI, USA, July 2017.

[97] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[98] Y. Tian and J. Zhang, “Optimizing sensor deployment for multi-sensor-based HAR system with improved glowworm swarm optimization algorithm,” *Sensors*, vol. 20, no. 24, p. 7161, 2020.

[99] A. Poulose and D. S. Han, “Hybrid indoor localization using IMU sensors and smartphone camera,” *Sensors*, vol. 19, no. 23, p. 5084, 2019.

[100] J. Van Hauwermeiren, K. Van Nimmen, P. Van den Broeck, and M. Vergauwen, “Vision-based methodology for characterizing the flow of a high-density crowd on footbridges: strategy and application,” *Infrastructure*, vol. 5, no. 6, p. 51, 2020.