IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 June 2023, accepted 10 July 2023, date of publication 20 July 2023, date of current version 28 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297269

RESEARCH ARTICLE

Lightweight Deep Learning Framework for Speech Emotion Recognition

SAMSON AKINPELU, SERESTINA VIRIRI, (Senior Member, IEEE), AND ADEKANMI ADEGUN
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4041, South Africa
Corresponding author: Serestina Viriri (viriris@ukzn.ac.za)

ABSTRACT Speech Emotion Recognition (SER) system, which analyzes human utterances to determine a speaker's emotion, has a growing impact on how people and machines interact. Recent growth in human-computer interaction and computational intelligence has drawn the attention of many researchers in Artificial Intelligence (AI) to deep learning because of its wider applicability to several fields, including computer vision, natural language processing, and affective computing, among others. Deep learning models do not need any form of manually created features because they can automatically extract the prospective features from the input data. Deep learning models, however, call for a lot of resources, high processing power, and hyper-parameter tuning, making them unsuitable for lightweight devices. In this study, we focused on developing an efficient lightweight model for speech emotion recognition with optimized parameters without compromising performance. Our proposed model integrates Random Forest and Multi-layer Perceptron(MLP) classifiers into the VGGNet framework for efficient speech emotion recognition. The proposed model was evaluated against other deep learning based methods (InceptionV3, ResNet, MobileNetV2, DenseNet) and it yielded low computational complexity with optimum performance. The experiment was carried out on three datasets of TESS, EMODB, and RAVDESS, and Mel Frequency Cepstral Coefficient(MFCC) features were extracted with 6-8 variants of emotions namely, Sad, Angry, Happy, Surprise, Neutral, Disgust, Fear, and Calm. Our model demonstrated high performance of 100%, 96%, and 86.25% accuracy on TESS, EMODB, and RAVDESS datasets respectively. This revealed that the proposed lightweight model achieved higher accuracy of recognition compared to the recent state-of-the-art model found in the literature.

INDEX TERMS Deep learning, convolutional neural network, speech emotion, lightweight, human–computer interaction.

I. INTRODUCTION

The social structure, the demand for talent, and human-machine interaction have all altered because of the rapid growth of deep learning from artificial intelligence (AI), data science, and IoT technology [1], [2], [3]. Speech is an exceptionally straightforward and seamless mode of human interaction that has been proven to effectively and swiftly communicate information in this era. People now dedicate a lot of time to learning how to speak to a variety of smart devices and interact with them through speech signals. Currently, a wide range of voice assistants, including Amazon Alexa, Microsoft Cortana, and Samsung Bixby recognize human-generated information through interactive real-time intelligent conversation and realize automatic operation in accordance with the speech content [4].

The computer has given birth to many innovations through which traditional learning and communication have given way to animation, visual art, artificial reality, and other forms of computer-aided means of expression. As a result of the advancement in technology, the learning environment has evolved, and researchers' interest in the intelligent learning environment built around AI has increased significantly. Teachers conduct their educational activities online over the Internet in an adaptive learning environment, and students can readily learn new information through the network. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

all these teacher-learner communications are not without expression of one emotion or the other. Studies from the psychological domain have revealed that different emotions that emerge throughout the learning process can impact the learning outcome. Research has also demonstrated that while negative emotions [5], [6] like disgust, fear, and sadness can impede cognition, good emotions like happiness and joy that are produced throughout the learning process are beneficial to increasing learning interest [7], [8], [9].

Humans communicate naturally through speech. Communication through speech utterance has become a major necessity in human life by which messages and ideas are conveyed [10], [11]. Therefore, interacting with the computer rather than typing on a keyboard has become more pleasant for people. Speech Emotion Recognition (SER) refers to the process by which a computer is made to recognize the inherent emotion present in speech signals. If the SER system is successful, the computer will be able to engage with people on an entirely new level. For instance, in an e-learning system, the computer can identify the speaker's emotions and provide helpful answers by suggesting simpler learning steps than the ones already in place [12]. Additional uses of SER include health care systems, aiding fighter pilots in combat, mental disorder treatment, and caring for the elderly in society.

Deep learning differs from the conventional machine learning approach that is handcrafted in nature (Fig. 1). The machine learning approach usually consists of a separate segment for its feature extraction before introducing a machine learning algorithm [13] while deep learning does not require handcrafted feature extraction approach because the algorithm such as Convolutional Neural Network (CNN) searches what specific features is best for classifying images [14]. In other words, the machine learning approach tends to break the problem down into constituents' parts at first and later aggregate the result at the output stage whereas the deep learning model proffers a solution to a problem in an end-to-end manner.

Recent years have seen a rise in the use of deep learning models for the extraction of speech signals' emotions [15], [16]. For feature extraction, convolutional neural networks (CNNs) have been very effective [17], [15]. Researchers have also thought about combining CNNs with other machine learning techniques, such as support vector machines, to increase the effectiveness of emotion identification models [18], [19]. The development of compact models for usage on mobile and embedded systems has also received attention.

In general, deep learning generally requires a very huge volume of training data to avoid overfitting though not very much feasible in most SER systems (limited dataset) [20]. Researchers have attempted to perform a deeper convolution layer to increase accuracy and improve performance, however, they have discovered that resources for computation, like memory storage, have proven to be a significant roadblock. Computational complexity [21] has been the major drawback in this case because of several convolutional layers. Therefore, a lightweight model which is the focus of this study becomes paramount in increasing the efficiency and performance of the SER model on low low-memory devices, without comprising accuracy.

The following are the paper's main contributions: Speech emotion recognition is made possible by: (i). A lightweight model using deep learning techniques for feature extraction and the best-performing classifier for classification (ii) Reduction in the number of computing resources needed for deep learning model feature extraction through architectural depth optimization. (iii) Evaluation of our proposed lightweight model against existing deep learning models on Central Processing Unit(CPU) and low-memory devices, which performs significantly better in terms of accuracy of recognition and execution time than a complete deeply learn architecture.

II. RELATED WORKS

The creation of models for recognizing emotions from speech signals has garnered more attention in recent years. Both conventional techniques like hidden Makov Model(HMM) and Support Vector Machine (SVM) and more contemporary ones like CNNs, Recurrent Neural Networks (RNNs), and Long Short-Term Memory(LSTMs) have been investigated. In comprehending conversations more effectively, researchers have also considered leveraging transfer learning, domain adaptation, and Natural Language Processing (NLP) approaches, as well as utilizing the time-frequency information contained in voice signals for effective recognition of emotion.

Scholars have also introduced several traditional methods, such as HMMs and Random Forest, to extract acoustic information from speech signals [22]. Researchers have looked into recurrent neural networks (RNNs) and long short-term memory (LSTM) networks as viable methods for classifying emotions from voice inputs [23]. Also, some studies have examined how various forms of data augmentation [24] affect emotion classification models.

In addition to the approaches discussed above, researchers have also used transfer learning and domain adaptation strategies to improve the performance of emotion recognition models [25]. Other research has looked at exploiting the time-frequency information of speech signals [26], [27]. Furthermore, there have been studies exploring the use of biometric features of speech signals for emotion recognition [28], [29].

Several studies investigated the use of generative models for voice signal emotion identification [30], [31], [32], [33]. The use of NLP methods to extract features from conversations has also received attention [34]. Researchers have also looked into harnessing processes for self-attention to better recognize salient aspects in human speech [35] for accurate recognition of emotion. Moreover, experiments investigating the integration of deep learning and unsupervised learning techniques to enhance emotion identification models have been conducted [36].

The achievement of the Deep Neural Network(DNN) in SER task cannot be overemphasized, though with a few peculiar limitations. Unlike any other image recognition task, speech signal differs in terms of environment, style, language, and content of the speaker. Also, DNN is prone to learning a high-level feature from common Low-Level Descriptors (LLD), a technique that is less sufficient in extracting all emotional features from speech signals. This is what paved the way for the use of the Mel Frequency Cepstral Coefficient (MFCC) by researchers in representing speech signals [37]. Two axes (vertical and horizontal) exist in any MFCC representation. While the horizontal axis carries information that is time-domain specific, the vertical axis usually carries information that has to do with the frequency of the signal. Thus, positioning MFCC has a unique representation of speech signals that possess essential speech emotional features. Therefore, CNN has achieved improved performance in the SER domain because it extracts emotional features from MFCC in an automatic manner.

In Muyawei et al. [38], a distributed CNN and bidirectional recurrent neural network were utilized in obtaining emotional features from human raw speech. They adopted the attention mechanism technique of focusing on the most useful section of emotion and achieved a weighted accuracy of 64.08% on the IEMOCAP dataset. The author in [39] proposed a combination of attention-based RNN and convolutional neural networks for SER. Their method achieved a significant performance on IEMOCAP and FAU datasets [40] as indicated in the result obtained. A parallelized CRNN (Convolutional Recurrent Neural Network) is proposed by Jiang et al., [41] to acquire more salient emotional features from human speech. In their methodology, LSTM was utilized to learn frame-level features and CNN was used to learn other features from the log Mel-spectrogram, thereafter, all the features were fused. Finally, a softmax classifier was adopted to classify emotion on four public speech emotion datasets. Their experimental result showed superior performance in previous work. Prau et al. [42] proposed presented a neural network classifier and RNN for speech emotion recognition. The author applied a denoising technique in removing noise from the speech dataset, through the median filter. However, RNN is prone to dependency problems, even though it has excellent performance on time series data. Also, no record of testing their model on any of the publicly available datasets.

Chimthankar [43] proposed an innovative deep learning technique that combines CNN and LSTM on MFCC features extracted from four popular speech datasets (TESS, RAVDESS, SAVEE, and CREMA-D). Their model achieved 67.58% validation accuracy and 71.28% testing accuracy on a German based (audio samples) separate dataset which was not introduced to the model during training. However, an improvement in the performance of the recorded experimental result can still be achieved and computational complexity peculiar to the CNN model. Atila et al. [44] proposed a 3D CNN-LSTM with an attention mechanism model for SER. Four features including MFCC, fractal dimension, etc. were used in their study. The method showed an improved result, but the number of parameters generated tends to increase complexity which may not be suitable for low-memory devices.

Aggarwal et al. [45] applied principal component analysis (PCA)-DNN and pre-trained VGG-16 model as a two-way feature extraction approach for speech emotion recognition. Their extensive experimental result over two datasets with in-depth analysis yielded an improved accuracy on the RAVDESS dataset alone. The author showed that their model achieved better performance compared to the one-way feature extraction approach with DNN. However, better performance over only one dataset poses a limitation on the generalizability of this work. A gender-dependent CNN model for SER was proposed in [46]. The author captured two different emotional intensities (normal and strong) on six emotions (Sad, Fear, angry, happy, disgust, and calm) from the RAVDESS speech dataset, with an improved result. MFCC features and its variant was used in their study with an in-depth performance comparison showing a relative improvement over the baseline system that utilized emotional features such as Chromogram, mel-spectrogram, MFCC, and Spectral contrast.

Building a lightweight model with limited data has been a challenging task in SER. The author in [47] proposed a lightweight architecture for speech emotion recognition using convolution that is separable, inverted residuals and attention mechanism. They achieved 71.72% and 90.1% on two well-established datasets (IEMOCAP and EMODB) respectively. Atsavasirilert et al., [48] also presented a lightweight DCNN model for SER using the EMOD dataset only and they achieved the highest accuracy of 87.16% accuracy. Speech emotion recognition based on lightweight CNN was proposed in Anvarjon et al. [49]. Deep features from the spectrogram were extracted from IEMOCAP and EMODB datasets, and they achieved 77.01% and 92.02% accuracy respectively on both datasets. Inspired by the need to build a more lightweight deep learning architecture for SER, coupled with the fact that much research has a lesser focus on SER for low-memory devices, this paper proposed an optimized and state-of-the-art lightweight methodology for speech emotion recognition with low computational complexity and higher recognition accuracy.

III. PROPOSED METHODOLOGY

The proposed lightweight architecture is designed to carry out speech emotion recognition tasks directly from speech signals irrespective of environmental background or language. One major pre-processing required is to reshape the MFCC speech feature image to 224×224 as input to our model for standardizing all the speech datasets used. The remaining part of this section gives detail of how feature extraction is performed using VGGNet and classifiers used in classifying emotion as shown in Figure 1. The input to our model is at first subject to a 2D convolutional layer with 2 strides for

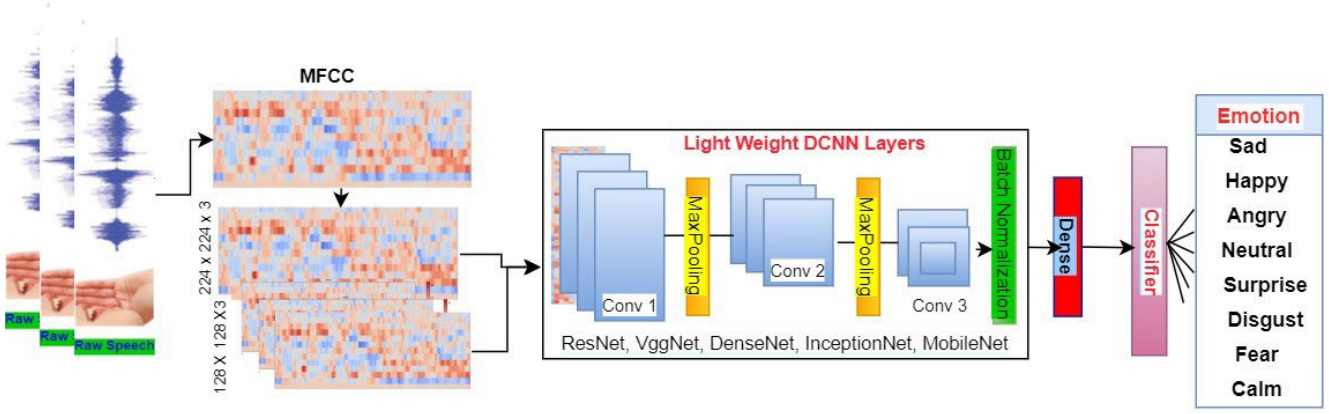

FIGURE 1. Lightweight SER architectural framework.

the extraction of distinct feature representations. A higher-level feature produced through the discriminative features is passed as input to the classifier for eventual speech emotion recognition. The result from other experimental studies is showcased as well to explain the rationale behind the choice of VGGNet and random forest classifier.

A. SPEECH FEATURE EXTRACTION

In this study, the emotional feature utilized is MFCC. Much research in SER has adopted MFCC as an efficient feature when it comes to emotion recognition. It can adequately describe the human vocal tract in speech utterance especially when speech sounds exhibit different kinds of emotion [50]. The mechanism behind the MFCC [51] extraction from the raw speech is to pass the audio spectrum into a Meyer filter bank to filter the audio spectrum in the frequency domain by the characteristics of human perception of sound frequency. Equation 1 illustrates the actual relationship between the filter's center frequency $Mel(f)$ and frequency f.

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{1}$$

where f is the frequency in Hz

The MFCC approach is based on the idea of Mel frequency, which is one of the most popular and useful ways to characterize parameters and correct for convolutional channel distortion. The Fast Fourier Transform (FFT), Meyer-filter-bank, Normalization, Framing, Windowing, and Discrete Cosine Transform (DCT) are all parts of the MFCC extraction procedure as shown in Figure 2. For the extraction of the MFCC feature in this study, we first scanned through all the audio (WAV) files of our dataset, RAVDESS, TESS, and EMODB, and all the voice paths were saved. This was necessary to label each file using the path truncation approach, read the audio files to acquire information about them, and thereafter extract the MFCC from the audio files. We extracted 13 MFCC features as a standard practice for SER task and the dataset used. At the pre-emphasis stage, we employed a first-order high-passed filter of 100KHz to emphasize components of higher frequency. Each pre-processed audio signal is divided into equal lengths of frames. We utilize a hamming window of $25ms$ length, $16kHz$ sampling frequency in carrying out the filtering process of the short fast Fourier transform (STFT). The windowing is performed using equation 2 to prevent spectral leakage. The final MFCC coefficients were obtained by applying the DCT to the log filter-bank energies. The higher-order coefficients were left out and we maintained the first 13 coefficients as our features. The choice of MFCC features in this study is due to the fact that they are highly compact in terms of speech signal representation (emotion-rich feature), robust to noise variability, and computationally efficient which are beneficial to the lightweight deep learning model as compared to spectrogram.

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left[2\pi n/(n-1)\right] & 0 \le n = N-1 \\ 0 & other \end{cases} \tag{2}$$

where f is the frequency in Hz

Speech signal has a peculiarity of continuous form in the time domain, and to better capture this continuous nature, a first-order and second-order difference method on the MFCC is utilized as shown in equation 3:

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{3}$$

where c_t represents the speech signal data point.

B. CONVOLUTIONAL NEURAL NETWORK LAYERS

The deep learning model with CNN provides advantages in terms of feature extraction. With the help of its distinctive convolutional kernel, it is possible to extract both local and global emotional information efficiently. To increase recognition accuracy, the pooling operation can simultaneously adjust to varying speech speeds and shifts in speech positions, to handle temporal variations in speech duration and align the extracted features effectively. CNN, on the other hand, has its foundation in sharing of weight and local

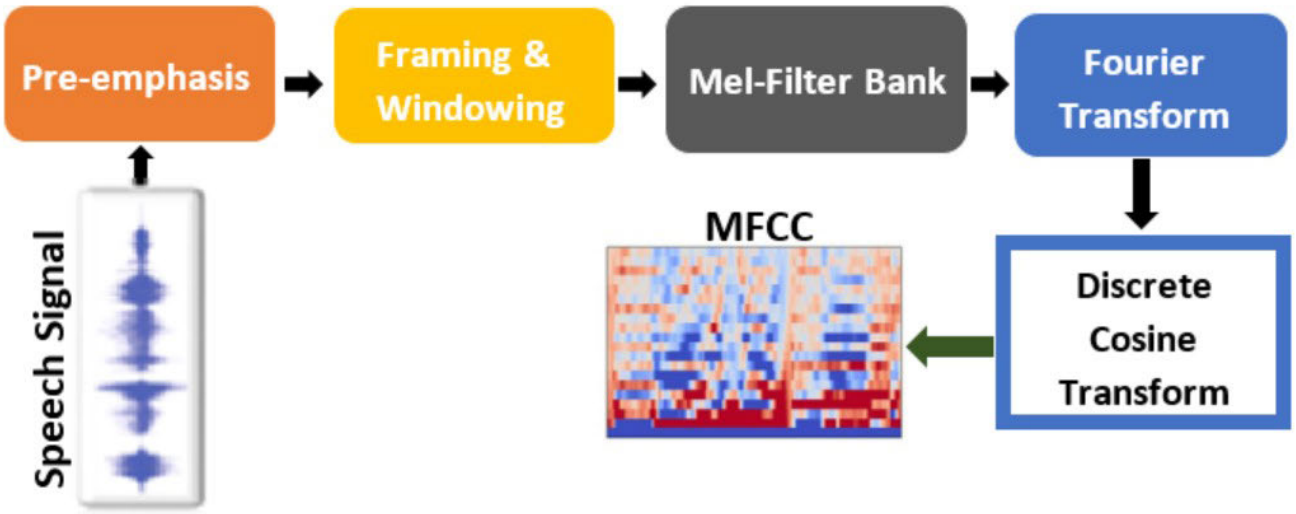

FIGURE 2. MFCC feature extraction from raw speech.

receptive fields [52]. The CNN model with fewer parameters, per the rules, requires comparatively less training data. Undoubtedly, a deeper convolutional network may extract high-dimensional aspects of emotion more effectively, but this typically comes at a hefty computational cost and is unsuitable for lightweight devices.

While some pre-trained deep learning models (VggNet, ResNet, MobileNet, etc.) and their variants enhance the flow of gradients and propagation of features through the sum of an identity function, a recent study [53] has revealed that a vast number of convolutional layers contribute very little to the outcomes. There are tons of trainable parameters generated by this. Contrarily, DenseNet places a strong emphasis on feature reuse through dense connectivity to address the issue and boost parameter effectiveness. Meanwhile, extensive connectivity enhances the flow of gradients and propagation of features even more.

Our proposed lightweight model is built on the VGGNet architecture for feature extraction, as shown in Figure 1. Figure 3 illustrates a comparison between the original VGGNet and our lightweight model. Three convolution blocks, one layer of batch normalization, one dense layer, and an input layer make up the proposed model. While the third convolution blocks include three convolution layers and one dropout layer, the first two convolution blocks only have two convolution layers and a max pooling layer. After passing through each convolution block, the input MFCC image's channel number steadily rises while its width and height are halved. The channel initially goes from 3 to 64, then to 128, 256, and finally to 512 channels, in that order. In generating the output image of $C_w/2$ and $C_h/2$, the max pooling operation divides the input channel's width, C_w, and height, C_h, into nearby pixels of 2×2 size. The fourth and fifth convolution block from the pre-trained VGGNet model has been eliminated from the proposed model to reduce the memory overhead. The top layer consisting of two fully connected and flatten layers were frozen from the conventional VGGNet. We adopted the weight from the original pre-trained VGGNet model on ImageNet which has been rained over four million images to prevent our model from training from scratch. The proposed model is designed to ensure adequate learning capacity with no overfitting [29], [54]

After the fourth convolution block, a layer of batch normalization is introduced as a way of normalizing the output and preventing model overfitting. Any layer of the neural network can undergo batch normalization, with the main goal being to achieve stable activation values that will lessen the inner covariate shift and prevent the over-fitting issue. The normalization of each dimension in the m-dimensional input, $x = (x^{(1)} \ldots x^{(d)})$, is computed in equation 4-6.

$$\hat{x}^{(m)} = \frac{x^{(m)} - E\left[x^{(m)}\right]}{\sqrt{Var\left[x^{(m)}\right]}} \tag{4}$$

where $x^{(m)}$ denotes each activation and

$$E\left[x^{(m)}\right] = \frac{1}{n}\sum_{j=1}^{n} x_j^m \tag{5}$$

denotes the mean while

$$Var\left[x^{(m)}\right] = \frac{1}{n}\sum_{j=1}^{n}\left(E\left[x^{(m)}\right]\right)^2 + \epsilon \tag{6}$$

denotes the variance and ϵ is the numerical constant added for stabilizing the output.

Proposed			Original VGGNet		
Layer(type)	**Output Shape**	**Parameters**	**Layer(type)**	**Output Shape**	**Parameters**
input_1(Input Layer)	[(None,224,224,3)]	0	input_1(Input Layer)	[(None,224,224,3)]	0
block1_conv1(Conv2D	[(None,224,224,64)]	1792	block1_conv1(Conv2D)	[(None,224,224,64)]	1792
block1_conv2(Conv2D)	[(None,224,224,64)]	36928	block1_conv2(Conv2D)	[(None,224,224,64)]	36928
block1_pool(MaxPooling2D)	[(None,112,112,64)]	0	block1_pool(MaxPooling2D)	[(None,112,112,64)]	0
block2_conv1(Conv2D)	[(None,112,112,128)]	73856	block2_conv1(Conv2D)	[(None,112,112,128)]	73856
block2_conv2(Conv2D)	[(None,112,112,128)]	147584	block2_conv2(Conv2D)	[(None,112,112,128)]	147584
block2_pool(MaxPooling2D)	[(None,56,56,128)]	0	block2_pool(MaxPooling2D)	[(None,56,56,128)]	0
block3_conv1(Conv2D)	[(None,56,56,256)]	295168	block3_conv1(Conv2D)	[(None,56,56,256)]	295168
block3_conv2(Conv2D)	[(None,56,56,256)]	590080	block3_conv2(Conv2D)	[(None,56,56,256)]	590080
block3_conv3(Conv2D)	[(None,56,56,256)]	590080	block3_conv3(Conv2D)	[(None,56,56,256)]	590080
Dropout_6(Dropout)	[(None,56,56,256)]	0	block3_conv3(Conv2D)	[(None,56,56,256)]	0
batch_normalization_3	[(None,56,56,256)]	1024	block4_conv4(Conv2D	[(None,28,28,512)]	1180160
dense_1(Dense)	[(None,7,7,512)]	1799	block4_conv4(Conv2D)	[(None,28,28,512)]	2359808
			block4_conv4(Conv2D)	[(None,28,28,512)]	2359808
			block4_pool(MaxPooling2D)	[(None,14,14,512)]	0
			block5_conv1(Conv2D	[(None,14,14,512)]	2359808
			block5_conv1(Conv2D)	[(None,14,14,512)]	2359808
			block5_conv1(Conv2D)	[(None,14,14,512)]	2359808
			block4_pool(MaxPooling2D)	[(None,7,7,512)]	0

FIGURE 3. The proposed model and original VGGNet architecture.

At the final layer, we added a dense layer that aggregates the input from the preceding layer and output feature vector according to the number of emotions (sad, angry, happy, disgust, calm, etc.) to be recognized. Output from this layer is fed into the classifier for the final recognition of emotion.

C. EMOTION RECOGNITION CLASSIFIERS

In recognition of emotion from this study, two classifiers were employed in this study which is Random Forest and Multilayer perceptron. The output from the dense layer after several convolutions is passed to the classifier for accurate classification of emotion into various categories. The description of the classifier is as follows:

1) RANDOM FOREST(RF) CLASSIFIER

Random forest is an ensemble learning technique for classification that is based on the class that most decision trees have chosen as their target as shown in Figure 4. In this situation, the features produced by convolution layers can be handled by downscaling variables since random forest excels at managing enormous input variables. Any size can be used because there is no need for a cross-validation set to guarantee an impartial estimate.

2) MULTI-LAYER PERCEPTRON (MLP) CLASSIFIER

In this study, a multi-layer perceptron (MLP) classifier is used for recognizing emotion as it receives input from the last layer of our proposed model. This feedforward artificial neural network model functions as a feedforward network-based classifier, mapping a set of appropriate outputs from a set of input datasets. Each layer that makes up an MLP is completely coupled to the layer below it. The nodes of the layers represent neurons with nonlinear activation functions,

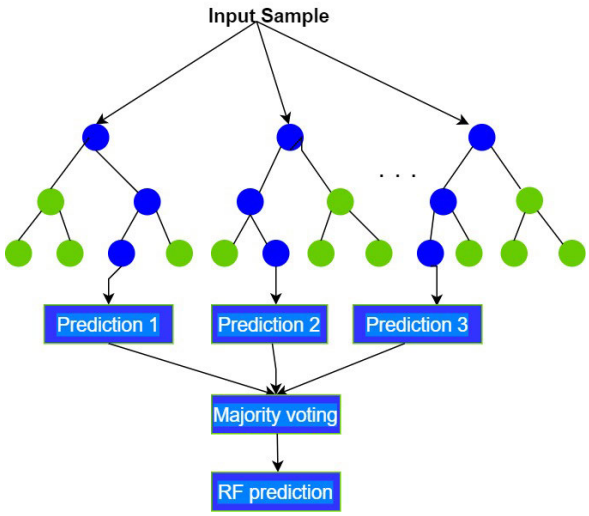

FIGURE 4. Random forest classifier structural view.

except the input layer nodes. We employed three hidden layers in our MLP classifier, with the first layer, second and third layers containing 128, 64 and 32 hidden units respectively.

IV. EXPERIMENTS AND RESULTS

We present significant details about the experimental setting, our experiments, and the analyses' findings in this section of the paper.

A. TESS DATASET

Many SER tasks have employed the Toronto English Speech Set [55], or TESS for short, one of the largest publicly accessible datasets. TESS speech samples were captured in 2010 at Northwestern University's Auditory Laboratory.

TABLE 1. Description of TESS speech dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	400	14.28
Sad	400	14.28
Disgust	400	14.28
Fear	400	14.28
Happy	400	14.28
Neutral	400	14.28
Surprise	400	14.28

TABLE 2. Description of RAVDESS speech dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	192	13.33
Sad	96	6.66
Fear	192	13.33
Boredom	192	13.33
Disgust	192	13.33
Happy	192	13.33
Neutral	192	13.33
Calm	192	13.33

Two actresses were instructed to speak a few of the 200 words during the spontaneous occurrence, and their voices were captured, creating a comprehensive collection of 2800 speech utterances. There were seven various emotions observed in the scenario, including joyful, furious, fear, disgust, pleasant, surprise, sad, and neutral. The description of TESS is illustrated in Table 1.

B. RAVDESS DATASET

A new English-language scripted emotional corpus, known as RAVDESS or Ryerson's audio-visual dataset of emotional song and speech [56], was completed in 2018. It is the most widely used dataset for identifying emotionally charged songs and speech. The recommended corpus, which consists of eight distinct emotions, was recorded by 24 professional individuals-12 men and 12 women-speaking scripts with altered emotions. The RAVDESS speech corpus is now mostly utilized for comparative analysis, which demonstrates the model's generalization as a result of the frequent use of emotions. It comprises 1440 sounds in all, collected at a sample rate of 48000 Hz. Table 2 provides a full description of the categories, audio utterances, and participation rate in percentage.

C. EMODB DATASET

This data set, also known as the Berlin emotion dataset [60] or the EMO-DB, is one of the most widely used. There are 535 speech utterances with seven different emotions in this well-known and popular speech emotion dataset. Ten expert people read prescript sentences and record different emotions for the suggested dataset, five males and five females. In the EMO-DB corpus, time is recorded with a sampling rate of $16kHz$ and an average of 2 to 3 seconds. A large number of emotion recognition techniques are based on the EMO-DB corpus, which is widely utilized in the SER domain.

TABLE 3. Description of EMODB speech emotion dataset.

Emotion	Audio files Used	Percentage Ratio (%)
Angry	127	23.73
Sad	62	11.58
Disgust	46	8.59
Neutral	79	14.76
Happy	71	13.27
Fear	69	12.89
Boredom	81	15.14

TABLE 4. Hyperparameter setting.

Hyperparameter	Value
Input size	224x224 x3
Optimizer	Adam
Learning rate	5e-5
Loss function	Sparse categorical crossentropy
Classifier random state	50
Number of estimator	100

An overview of selected emotions, total utterances, and participation ratio is shown in Table 3.

1) EXPERIMENTAL SETUP

The experiment for this study out was carried on 8GB RAM, 64bit OS, Intel core i7 device with Python 3.9 programming software. Table 4 shows the summary of the hyperparameter setting for the experiment. In selecting the best lightweight model architecture, we experimented with several deep learning lightweight architectures. Among them are DenseNet, VGG16, InceptionNetV3, ResNet, and MobileNet. Out of these, the VGGNet outperformed the rest in terms of accuracy and speed of emotion recognition. The feature map produced by the proposed lightweight model (Figure 5) for speech emotion recognition, it indicates whether the model is creating discriminative and meaningful representations. We can qualitatively evaluate if the model is capturing relevant speech features that are instructive for emotion recognition by looking at the feature maps. We gain insights into the learned representations and see how the model alters the input features to capture pertinent patterns and features for the speech emotion recognition task by visualizing the feature maps at various layers. Table 5 shows the performance of our model in comparison with other architecture tested, where the EMODB dataset (represented as DT1), RAVDESS dataset (represented as DT2) and TESS dataset (represented as DT3) respectively with highest accuracy recorded on DT3 when estimators value is 100 and random state set to 50. The dataset is split into an 80:20 ratio for the training and testing set. Both the test and training sets of data had their pixel values normalized to range from 0 to 1. Besides, the model size is minimal with optimum accuracy compared to the existing ones.

2) EXPERIMENTAL RESULTS

The performance comparison of our proposed model with other studies is highlighted in Table 7 following the speech

FIGURE 5. Feature map from the first convolutional layer.

TABLE 5. Lightweight model selection.

Model	Size in MB (Lightweight)	No. of Parameters	Dataset	Accuracy(%)	Average Accuracy
DenseNet	7.04	7,043,654	DT1 DT2 DT3	86.13 79.24 91.60	85.65
VGGNet	**7.64**	**7,640,903**	**DT1** DT2 DT3	**96.03** 86.25 100.00	**94.09**
InceptionNet	9.64	9,604,544	DT1 DT2 DT3	62.30 75.60 88.20	75.36
MobileNet	4.80	4,806,855	DT1 DT2 DT3	48.50 81.32 31.96	53.92
ResNet	9.12	9,116,032	DT1 DT2 DT3	87.12 83.90 98.00	89.67

TABLE 6. Low-memory device configuration.

Device	Specification
Hardware	Raspberry Pi 4 Model B
Processor	Quad-core Cortex-A72 (ARMv8) 64-bit SoC @ 1.5GHz
RAM	2GB LPDDR4
Operating System	Raspbian OS
Software Frameworks	TensorFlow Lite, Python 3.7
Compilation	Model optimized using TensorFlow Lite for ARM architecture

TABLE 7. Performance comparison.

Reference	Year	Dataset	Accuracy Reported (%)
[57]	2018	**EMODB**	82.82
[41]	2019		84.49
[58]	2019		88.99
[59]	2020		85.57
[60]	2020		90.01
[49]	2020		92.02
[61]	2021		93.00
[62]	2022		90.01
Proposed	**2023**		**96.03**
		RAVDESS	
[63]	2019		75.79
[64]	2019		67.14
[59]	2020		77.01
[65]	2022		81.82
Proposed	**2023**		**86.25**
		TESS	
[66]	2017		96.00
[67]	2018		89.96
[68]	2019		89.96
[69]	2021		93.30
[70]	2022		96.10
Proposed	**2023**		**100.00**

emotion database used. Our model outperforms other proposed methods in all three datasets from an accuracy and computational complexity point of view. The highest accuracy recorded by other researchers on EMODB was 93.00%, 81.82% on RAVDESS, and 96.10% on TESS, but our model supersedes them all. Crucial to the recognition of speech emotion in real time-application is time, and our proposed model has an average clock of 0.07 seconds (CPU time) in the classification of a single emotional utterance because the number of parameters has been reduced drastically. Besides, our model outperforms what was obtained in [57] when comparing the model size(323.46 to 7.94) and accuracy(82.82% to 96.03%) of emotion recognition on the EMODB dataset. To evaluate the performance and efficiency of our lightweight SER model on low-resource devices, we conducted an experiment using the hardware and configuration as shown in Table 6. We obtain an average processing time of 0.15 seconds per utterance and 0.02 seconds standard deviation. The processing time measurement includes all the steps from pre-processing to emotion recognition. Specifically, it incorporates feature extraction, model loading, and forward pass through the model.

3) DISCUSSION

The significance of feature extraction in the speech emotion recognition model is very important. However, some features

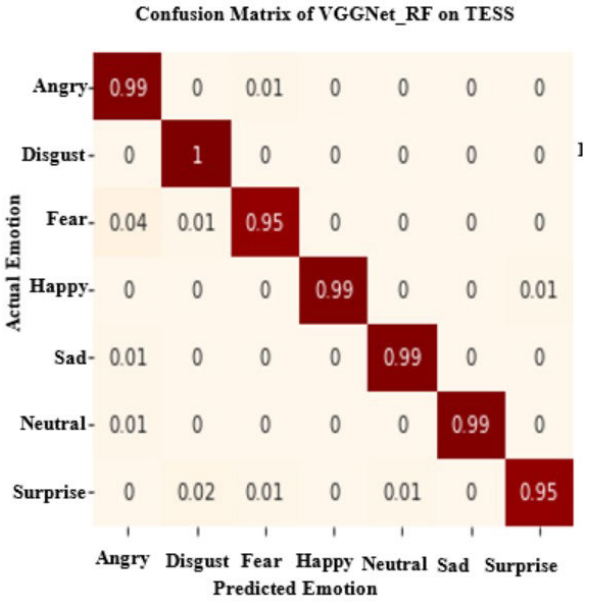

FIGURE 6. Confusion matrix of the proposed model on TESS datasets with an average recognition accuracy of 98%.

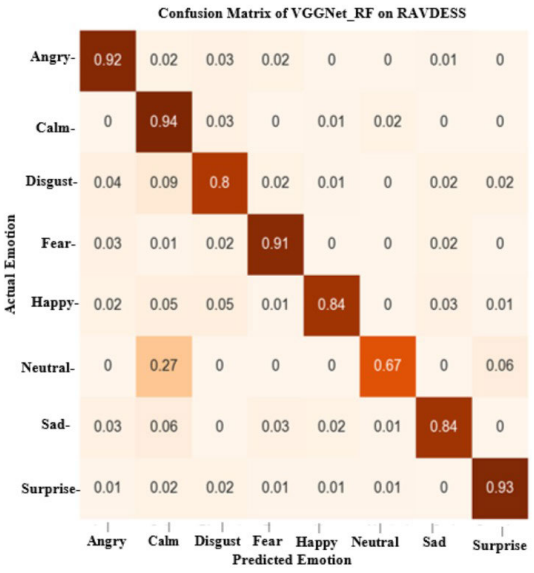

FIGURE 8. Confusion matrix of the proposed model on RAVDESS datasets with an average recognition accuracy of 85.6%.

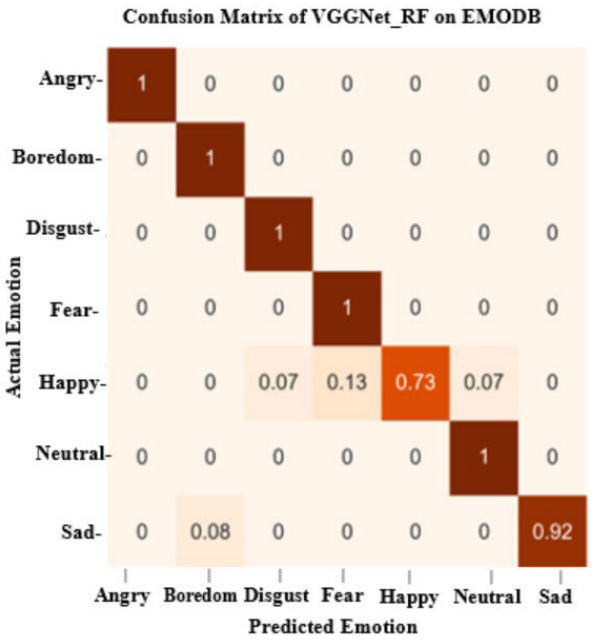

FIGURE 7. Confusion matrix of the proposed model on EMODB datasets with an average recognition accuracy of 95.5%.

can be difficult to extract owing to environmental factors and the expected number of features to be extracted. From the input MFCC image, the features extracted by each layer of the convolutions differ. The deeper the convolution, the higher the number of features that will be extracted. Many a time these features range from generic ones to more specific ones (high-level features). However, computational complexity in terms of resources, space, and training time has been a major drawback. Our lightweight VGGNet utilized the transfer learning techniques, where weights from the original pre-trained VGGNet model that has been trained on millions of datasets was transferred. Two major criteria were used in optimizing our proposed model, which are training-time and accuracy. Our extensive experiment indicated that at the third convolutional block with additional drop out, batch, and dense layer, an optimal accuracy was recorded. Hence, the remaining convolutional layer and block from the main VGGNet have been expunged.

The confusion matrix for this study is shown in Figure 6-9. With our Lightweight model, we achieved the highest accuracy of recognition (100%) on TESS with MLP classifier with 7 emotions. With the confusion matrix, there is a deep insight that showcases the misperception between the predicted emotional classes and the actual emotional classes, along with other emotions at corresponding rows. The two axes that existed in the confusion matrix represent expected predictions (x-axis) and actual predictions (y-axis).

Our proposed SER lightweight model achieved overall recognition accuracy of 100%, 96.03%, and 86.25% for TESS, EMODB, and RAVDESS speech datasets respectively. To further the investigation of the model performance, Figure 10 illustrate the emotional-level prediction for each of the dataset. Angry, boredom, and fear emotional class shows the highest recognition rate of 100% on the EMODB dataset with both classifiers, while only Angry indicates the highest recognition accuracy with both classifiers for the TESS dataset. For RAVDESS, angry and calm emotions show the highest accuracy of 94% from both classifiers. The lowest accuracy recorded was on Neutral emotion with a random forest classifier on the RAVDESS dataset. The experimental results obtained have in no doubt, shown the efficiency and

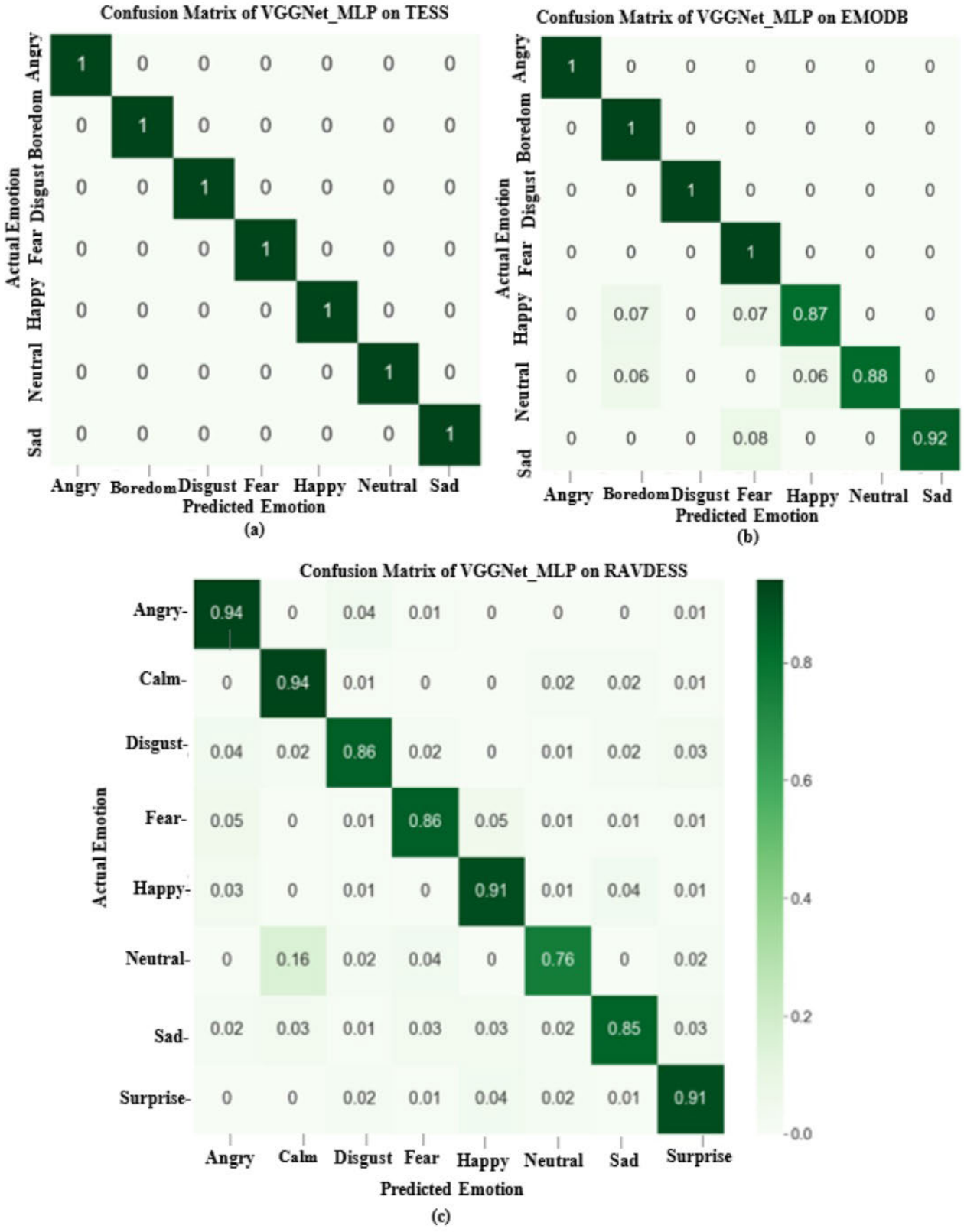

FIGURE 9. Confusion matrix of the proposed model on three datasets with MLP Classifier with 100%, 90.83% and 87.87% average recognition accuracy respectively.

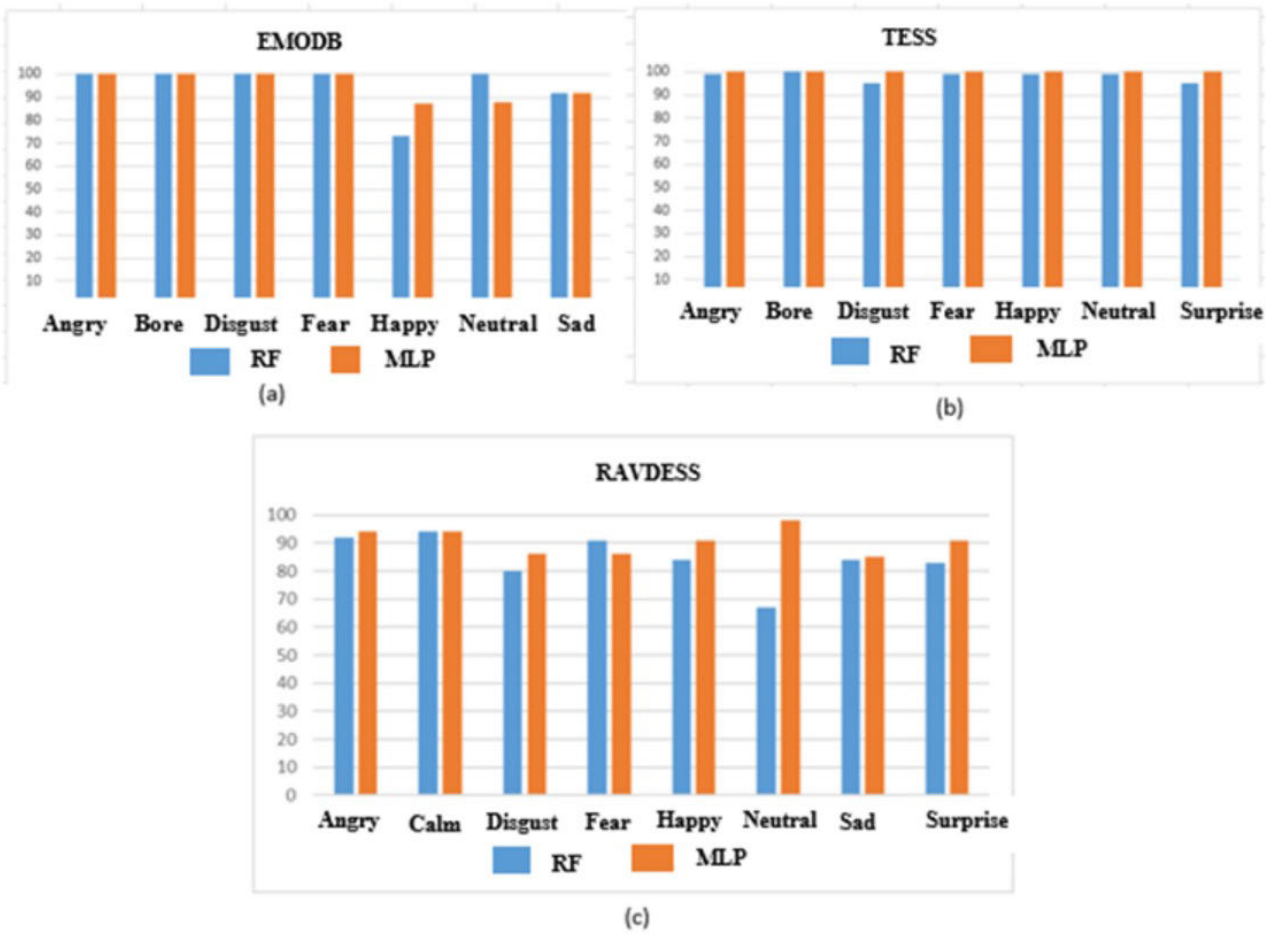

FIGURE 10. Emotion class prediction comparison on two classifiers with three datasets.

robustness of our proposed lightweight model for SER, with an improved performance over state-of-the-art techniques.

V. CONCLUSION

In this work, we have proposed a lightweight speech emotion recognition model that uses an end-to-end approach for feature extraction and classification through the best-performing classifier. Our VGGNet SER model has been optimized such that model size, computational complexity, and recognition time has been simplified, while the accuracy of recognition has been improved. The robustness and efficiency of our model on three popular datasets, TESS, EMODB, and RAVDESS achieved recognition accuracy of 100%, 96.03%, and 86.02% respectively. Our extensive experiments with the proposed model have proven that it is suitable for real-life application because it requires lesser time for recognition of emotion and has the capability for generalization. The proposed lightweight model has improved on emotion recognition systems especially for low-memory devices because of its moderate size (7.94Mb). However, in the future, we intend to investigate the possibility of integrating our model with other deep learning architecture (self-attention and transformer), and audio pre-trained models to extract more salient features from speech to improve emotion recognition. Besides, we will also take a closer at experimenting larger size emotional speech database to establish the performance of the model in SER-related tasks. The result of our experiment on a low-memory device highlights the practicality and efficiency of our proposed approach for real-time applications on such devices, however, we have a limitation on the number of these devices available at our disposal as at the time of this study, but we intend to explore more low-memory devices in our future research. The link to our implementation code to foster collaboration can be found here: https://github.com/samsoftcom1/Speech-Emotion2023.git and https://www.kaggle.com/code/samsonadebisi/speech-emotion-recognition-using-attention-network/.

REFERENCES

[1] T. S. Ustun, S. M. S. Hussain, L. Yavuz, and A. Onen, "Artificial intelligence based intrusion detection system for IEC 61850 sampled values under symmetric and asymmetric faults," *IEEE Access*, vol. 9, pp. 56486–56495, 2021.

[2] F.-K. Wang, T. Mamo, and X.-B. Cheng, "Bi-directional long short-term memory recurrent neural network with attention for stack voltage degradation from proton exchange membrane fuel cells," *J. Power Sources*, vol. 461, Jun. 2020, Art. no. 228170, doi: 10.1016/j.jpowsour.2020.228170.

[3] Y. Liu and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Future Gener. Comput. Syst.*, vol. 119, pp. 1–6, Jan. 2021, doi: 10.1016/j.future.2021.01.010.

[4] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic modeling based on deep learning for low-resource speech recognition: An overview," *IEEE Access*, vol. 8, pp. 163829–163843, 2020, doi: 10.1109/ACCESS.2020.3020421.

[5] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative emotion recognition using deep learning for Thai language," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng.*, Mar. 2020, pp. 71–74, doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[6] X. Lu, "Deep learning based emotion recognition and visualization of figural representation," *Frontiers Psychol.*, vol. 12, p. 818833, 2022, doi: 10.3389/fpsyg.2021.818833.

[7] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[8] S. Lugovic, I. Dunder, and M. Horvat, "Techniques and applications of emotion recognition in speech," in *Proc. 39th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2016, pp. 1278–1283.

[9] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov model," *Speech Commun.*, vol. 10, no. 4, pp. 603–623, 2003;.

[10] L. Trinh Van, T. D. T. Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," *Sensors*, vol. 22, no. 4, p. 1414, Feb. 2022, doi: 10.3390/s22041414.

[11] E. L. R. Ewe, C. P. Lee, L. C. Kwek, and K. M. Lim, "Hand gesture recognition via lightweight VGG16 and ensemble classifier," *Appl. Sci.*, vol. 12, no. 15, p. 7643, Jul. 2022, doi: 10.3390/app12157643.

[12] N. Kim and S. Kim, "A study on user experience of online education programs with elementary schools and art museums in non-face-to-face era," *J. Digit. Converg.*, vol. 19, no. 8, pp. 311–317, 2021.

[13] K. Feng and T. Chaspari, "A Siamese neural network with modified distance loss for transfer learning in speech emotion recognition," 2020, *arXiv:2006.03001*.

[14] O. Fagbuagun, O. Folorunsho, L. Adewole, and H. Akin-Olayemi, "Breast cancer diagnosis in women using neural networks and deep learning," *J. ICT Resour. Appl.*, vol. 16, no. 2, pp. 152–166, 2022, doi: 10.5614/itbj.ict.res.appl.2022.16.2.4.

[15] J. Amherst and J. Jhun, "Speech emotion recognition using biometric features," *IEEE Access*, vol. 8, pp. 12452–12463, 2020.

[16] W. Zheng, Z. Wenming, and Z. Yuan, "Multi-scale discrepancy adversarial network for cross-corpus speech emotion recognition," *Virtual Reality Intell. Hardw.*, vol. 3, no. 1, pp. 65–75, 2021, doi: 10.1016/j.vrih.2020.11.006.

[17] Y. Gou, S. Wang, and C. Feng, "Speech emotion recognition based on parallel convolutional neural networks," in *Proc. 2017 Int. Joint Conf. Neural Netw.*, 2017, pp. 3789–3794.

[18] Z. Wang, H. Seng-Beng, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools Appl.*, vol. 79, pp. 35553–35582, Jan. 2020, doi: 10.1016/j.vrih.2020.11.006.

[19] A. M. Badshah, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019.

[20] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, O. F. Afe, and K. I. Okesola, "Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches," *Int. J. Adv. Comput. Res.*, vol. 11, no. 53, pp. 51–60, Mar. 2021, doi: 10.19101/IJACR.2021.1152001.

[21] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review," *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.

[22] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[23] Y. Yu and Y.-J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," *Electronics*, vol. 9, p. 713, 2020, doi: 10.3390/electronics9050713.

[24] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/s22165941.

[25] Q. Lin, H. Feng, and H. Yin, "Emotion recognition from speech using convolutional neural network and transfer learning," *IEEE Access*, vol. 7, pp. 94059–94068, 2019.

[26] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition," *J. Neural Eng.*, vol. 19, no. 1, Feb. 2022, Art. no. 016012, doi: 10.1088/1741-2552/ac49a7.

[27] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[28] C. A. Kumar and K. A. Sheela, "Emotion recognition from speech biometric system using machine learning algorithms," in *Advances in Communications, Signal Processing, and VLSI*, T. Laxminidhi, J. Singhai, S. R. Patri, and V. V. Mani, Eds., vol. 722. Singapore: Springer, 2021, doi: 10.1007/978-981-33-4058-9_6.

[29] S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *Int. J. Intell. Syst.*, vol. 36, no. 9, pp. 5116–5135, Sep. 2021, doi: 10.1002/int.22505.

[30] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 1098–1110, Apr. 2022, doi: 10.1109/TAFFC.2020.2998118.

[31] A. Shahid, S. Latif, and J. Qadir, "Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation," 2023, *arXiv:2301.03751*.

[32] B. Pan and W. Zheng, "Emotion recognition based on EEG using generative adversarial nets and convolutional neural network," *Comput. Math. Methods Med.*, vol. 2011, Oct. 2021, Art. no. 2520394, doi: 10.1155/2021/2520394.

[33] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics*, vol. 11, no. 23, p. 3935, Nov. 2022, doi: 10.3390/electronics11233935.

[34] Y. Gao, D. Zhang, and H. Li, "Emotion recognition from conversation using natural language processing," *Brain Informat.*, vol. 8, no. 1, p. 5162, 2021.

[35] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y.-I. Cho, "Modeling speech emotion recognition via attention-oriented parallel CNN encoders," *Electronics*, vol. 11, no. 23, p. 4047, Dec. 2022, doi: 10.3390/electronics11234047.

[36] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[37] A. Winursito, "Improvement of MFCC feature extraction accuracy using PCA," in *Proc. Indonesian Speech Recognit. Int. Conf. Inf. Commun. Technol.*, 2018, pp. 379–383.

[38] L. Muyawei, G. Hernndez, C. Antonio, A. Carlos, W. Xuetian, and G. Hongmin, "Speech emotion recognition using convolutional-recurrent neural networks with attention model," in *Proc. 2nd Int. Conf. Comput. Eng., Inf. Sci. Internet Technol.*, 2017, pp. 341–350, doi: 10.12783/dtcse/cii2017/17273.

[39] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019, doi: 10.1109/ACCESS.2019.2928625.

[40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[41] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019, doi: 10.1109/ACCESS.2019.2927384.

[42] M. Prau, A. Tiwari, R. K. Singh, and R. Yadav, "Speech emotion recognition and features extraction based on NN classifier," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 2852–2857, 2020.

[43] P. P. Chimthankar. (2021). *Speech Emotion Recognition using Deep Learning*. School of Computing National College of Ireland. [Online]. Available: http://norma.ncirl.ie/5142/1/priyankaprashantchimthankar.pdf
[44] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," *Appl. Acoust.*, vol. 182, Nov. 2021, Art. no. 108260, doi: 10.1016/j.apacoust.2021.108260.
[45] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 2022, pp. 1–11, Jan. 2022, doi: 10.3390/s22062378.
[46] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," *Proc. Comput. Sci.*, vol. 218, pp. 2533–2540, Jan. 2023, doi: 10.1016/j.procs.2023.01.227.
[47] Y. Zhong, Y. Hu, H. Huang, and W. Silamu, "A lightweight model based on separable convolution for speech emotion recognition key laboratory of multilingual information technology," in *Proc. Interspeech*, 2020, pp. 3331–3335.
[48] K. Atsavasirilert, T. Theeramunkong, S. Usanavasin, A. Rugchatjaroen, S. Boonkla, J. Karnjana, S. Keerativittayanun, and M. Okumura, "A lightweight deep convolutional neural network for speech emotion recognition using Mel-spectrograms," in *Proc. 14th Int. Joint Symp. Artif. Intell. Natural Language Process. (iSAI-NLP)*, Oct. 2019, pp. 1–4.
[49] T. Anvarjon and S. Kwon, "Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.
[50] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, Nov. 2021, doi: 10.3390/s21227530.
[51] M. M. Lynn, C. Su, and K. K. Maw, "Efficient feature extraction for emotion recognition system," in *Proc. 4th Int. Conf. Converg. Technol. (I2CT)*, Oct. 2018, pp. 1–6, doi: 10.1109/I2CT42659.2018.9058313.
[52] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "RAVDNESS, CREMA-D, TESS based algorithm for emotion recognition using speech," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, Jan. 2022, pp. 1625–1631, doi: 10.1109/ICSSIT53264.2022.9716313.
[53] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," 2016, *arXiv:1603.09382*.
[54] F. Zhu-Zhou, R. Gil-Pita, J. García-Gómez, and M. Rosa-Zurera, "Robust multi-scenario speech-based emotion recognition system," *Sensors*, vol. 22, no. 6, p. 2343, Mar. 2022, doi: 10.3390/s22062343.
[55] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set (TESS)," *Can. Acoust. Acoust. Can.*, vol. 39, pp. 182–183, 2011.
[56] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.
[57] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[58] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[59] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[60] S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114177.
[61] S. Kwon, "ATT-Net: Enhanced emotion recognition system using lightweight self-attention module," *Appl. Soft Comput.*, vol. 102, Apr. 2021, Art. no. 107101, doi: 10.1016/j.asoc.2021.107101.
[62] E. Guizzo, T. Weyde, S. Scardapane, and D. Comminiello, "Learning speech emotion representations in the quaternion domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1200–1212, 2023.
[63] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.
[64] A. A. A. Zamil, S. Hasan, S. M. D. J. Baki, J. M. D. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST)*, Dhaka, Bangladesh, Jan. 2019, pp. 281–285, doi: 10.1109/ICREST.2019.8644168.
[65] C. Luna-Jimnez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernndez-Martnez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Appl. Sci.*, vol. 12, p. 327, Jan. 2022.
[66] D. Verma and D. Mukhopadhyay, "Age driven automatic speech emotion recognition system," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, Apr. 2016, pp. 1005–1010, doi: 10.1109/CCAA.2016.7813862.
[67] V. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1577–1587, 2018, doi: 10.3844/jcssp.2018.1577.1587.
[68] Y. Gao. (2019). *Speech-Based Emotion Recognition*. [Online]. Available: https://libraetd.lib.virginia.edu/downloads/2f75r8498?filename=1GaoYe2019MS.pdf
[69] P. Krishnan, A. J. Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," *Complex Intell. Syst.*, vol. 7, no. 4, pp. 1919–1934, 2021, doi: 10.1007/s40747-021-00295-z.
[70] S. Akinpelu and S. Viriri, "Robust feature selection-based speech emotion classification using deep transfer learning," *Appl. Sci.*, vol. 12, no. 16, p. 8265, Aug. 2022, doi: 10.3390/app12168265.

SAMSON AKINPELU received the B.Sc. and M.Tech. degrees in computer science from the University of Kwazulu-Natal, South Africa, where he is currently pursuing the Ph.D. degree in computer science. He has over three years of teaching experience with Federal University Oye Ekiti, Nigeria. His main research interests include artificial intelligence, computer vision, deep leaning, speech emotion recognition, pattern recognition, and natural language processing.

SERESTINA VIRIRI (Senior Member, IEEE) received the B.Sc. degree in mathematics and computer science and the M.Sc. and Ph.D. degrees in computer science. He is a Full Professor of computer science with the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, South Africa. He has been in academia, since 1998. He has published extensively in several artificial intelligence and computer vision-related accredited journals and international and national conference proceedings. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and other image processing related fields, such as biometrics, medical imaging, and nuclear medicine. He serves as a reviewer for several machine learning and computer vision-related journals. He has also served on program committees for numerous international and national conferences. He is a Rated Researcher by the National Research Foundation (NRF) of South Africa.

ADEKANMI ADEGUN received the B.Tech., M.Sc., and Ph.D. degrees in computer science. He has a lecturing experience in universities, for nearly ten years. He has also co-supervised M.Sc. and Ph.D. candidates in machine learning fields. He has published extensively in several artificial intelligence and computer vision-related accredited journals and international and national conference proceedings. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and natural language processing. He serves as a reviewer for some machine learning and computer vision-related journals.