IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 16, 2022, accepted April 27, 2022, date of publication May 5, 2022, date of current version May 12, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172954

Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm

ABDELAZIZ A. ABDELHAMID[1,2], EL-SAYED M. EL-KENAWY[3,4], (Senior Member, IEEE), BANDAR ALOTAIBI[5,6], (Member, IEEE), GHADA M. AMER[7], MAHMOUD Y. ABDELKADER[1], ABDELHAMEED IBRAHIM[8], (Member, IEEE), AND MARWA METWALLY EID[4]

[1]Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[2]Department of Computer Science, College of Computing and Information Technology, Shaqra University, Riyadh 11961, Saudi Arabia
[3]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology (DHIET), Mansoura 35111, Egypt
[4]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 35712, Egypt
[5]Department of Information Technology, Faculty of Computers and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia
[6]Sensors Networks and Cellular Systems Research Center, University of Tabuk, Tabuk 71491, Saudi Arabia
[7]Department of Electrical Engineering, Faculty of Engineering, Benha University, Benha 13511, Egypt
[8]Department of Computer Engineering and Control Systems, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

Corresponding author: Abdelaziz A. Abdelhamid (abdelaziz@cis.asu.edu.eg)

This work was supported in part by the Sensor Networks and Cellular Systems (SNCS) Research Center under Grant 1443-001.

ABSTRACT One of the main challenges facing the current approaches of speech emotion recognition is the lack of a dataset large enough to train the currently available deep learning models properly. Therefore, this paper proposes a new data augmentation algorithm to enrich the speech emotions dataset with more sam Department, College of Computing and ples through a careful addition of noise fractions. In addition, the hyperparameters of the currently available deep learning models are either handcrafted or adjusted during the training process. However, this approach does not guarantee finding the best settings for these parameters. Therefore, we propose an optimized deep learning model in which the hyperparameters are optimized to find their best settings and thus achieve more recognition results. This deep learning model consists of a convolutional neural network (CNN) composed of four local feature-learning blocks and a long short-term memory (LSTM) layer for learning local and long-term correlations in the log Mel-spectrogram of the input speech samples. To improve the performance of this deep network, the learning rate and label smoothing regularization factor are optimized using the recently emerged stochastic fractal search (SFS)-guided whale optimization algorithm (WOA). The strength of this algorithm is the ability to balance between the exploration and exploitation of the search agents' positions to guarantee to reach the optimal global solution. To prove the effectiveness of the proposed approach, four speech emotion datasets, namely, IEMOCAP, Emo-DB, RAVDESS, and SAVEE, are incorporated in the conducted experiments. Experimental results confirmed the superiority of the proposed approach when compared with state-of-the-art approaches. Based on the four datasets, the achieved recognition accuracies are 98.13%, 99.76%, 99.47%, and 99.50%, respectively. Moreover, a statistical analysis of the achieved results is provided to emphasize the stability of the proposed approach.

INDEX TERMS Speech emotions, deep learning, stochastic fractal search optimization, guided whale optimization algorithm.

I. INTRODUCTION

Speech emotion recognition (SER) has received much attention in recent years [1], [2]. Although human emotions are hard to characterize and categorize, research on machine understanding of human emotions is rapidly advancing. The recognition of speech emotions usually includes extracting paralinguistic features from speech. These features should be independent of the speaker and lexical content of the speech signal. Generally, the information embedded in speech signals can be categorized into paralinguistic information and linguistic information. Paralinguistic information refers

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

to implicit features, such as the emotions harnessed in the speech signal, which is the domain of SER [3]. On the other hand, linguistic information refers to the context and meaning of the speech signal, which is the domain of interest in speech recognition.

To recognize the embedded emotions in speech, many distinguishing features can be extracted. These features include spectral features, qualitative features, and continuous features [4]. Many researchers have investigated the application of these features in SER. On the other hand, other researchers investigated the advantages and disadvantages of these features; however, the best features that can be used for this task cannot be identified easily. These features are usually referred to as handcrafted features. The accuracy of these features is relatively high; however, professional knowledge is required for extracting these features. Consequently, deep learning is introduced to model the extraction of high-level features from lower-level features to save the efforts needed for extracting the handcrafted features [5].

Currently, deep learning approaches are employed to solve many critical problems. The strength of deep learning comes from its ability to learn high-level features. Therefore, many researchers have introduced these approaches to recognize speech emotions based on many deep learning architectures. These architectures could achieve reasonable accuracy for the task of SER. However, more efforts are still required to improve the recently achieved performance [6], [7].

Deep learning greatly improved the performance of speech signal processing frameworks. Excellent results are achieved by researchers in this field based on the application of convolutional neural networks (CNNs), deep belief networks (DBNs), and long short-term memory (LSTM) [8], [9]. Special processing is required for speech signals to model their time-varying nature. Therefore, LSTM is more suitable to extract the long-term contextual dependencies in the input speech. One of the most effective features that can be used in SER is time-frequency decomposition, which is represented by a spectrogram. These features are proven to give significant recognition accuracy compared to using the raw speech signal when used to train deep learning frameworks [10].

The hyperparameters of deep learning models affect their performance to a certain extent. The selection of proper values of these parameters usually forms a challenge in utilizing deep learning models for different tasks. Recently, many optimization techniques have emerged to optimize the parameters of various models. These optimization techniques include particle swarm optimization (PSO) [11], whale optimization algorithm (WOA) [12], gray wolf optimization (GWO) [13], dipper throated optimization (DIP) [14], etc. In this research, we adopted the WOA, as an example, for optimizing the hyperparameters of the proposed deep learning model. Other types of optimizers will be considered in the future perspectives of this research.

There are many beneficial usages of SER in various applications that are based on the interaction between humans and computers. These applications include customer service, speech synthesis, medical analysis, forensics, and smart education. These applications highlight the significance of the automatic recognition of speech emotions and the necessity for achieving high recognition accuracy to realize these applications properly.

This paper presents an accurate approach for recognizing speech emotions using an optimized deep learning model based on cascaded layers of CNN+LSTM and stochastic fractal search guided whale optimization algorithm (SFS-Guided WOA). The effectiveness of the proposed approach is validated in terms of four standard speech emotion datasets, namely, IEMOCAP [15], Emo-DB [16], RAVDESS [17], and SAVEE [18]. In addition, the results of the proposed approach are compared with the results achieved by the other competing approaches in the literature to prove its superiority. Moreover, statistical analysis is performed to confirm the stability of the performance of the proposed approach.

The structure of this paper is organized as follows. A literature review is presented in section II. The proposed approach, along with the system architecture, is then explained in section III. Section IV presents and discusses the results of the conducted experiments. Finally, the conclusions and future perspectives are given in section V.

II. LITERATURE REVIEW

Speech emotion recognition (SER) is addressed by many researchers in the literature. In this section, we discuss some of these research efforts focusing on their achievements.

Aharon *et al.* [19] employed a deep neural network to recognize speech emotions from paralingual information. This deep network consists of convolutional and recurrent layers to learn the inherent representations of speech emotions. This approach utilizes the speech signal spectrogram to achieve this goal. The processing of speech signals is performed based on small segments with non-overlapping parts. This approach was tested on the IEMOCAP dataset and achieved recognition accuracy of 68% when the deep network was combined with a high-complexity convolutional LSTM.

Jonathan *et al.* [20] proposed an improved approach based on two machine learning approaches. They employed both multitask machine learning and deep convolutional generative adversarial networks to generate a set of unlabeled data. Using these approaches, they could leverage the size of the speech emotions training corpus to 100 hours. This large corpus could improve the performance of speech emotion classifiers, and the achieved performance was better than that of the baseline systems. The percentage of the achieved improvement reached 43.88%, which competes with the methods.

Chen *et al.* [21] hypothesized that measuring deltas and delta-deltas for customized characteristics not only retains successful emotional information but also reduces the impact of emotionally irrelevant variables, resulting in less misclassification. Furthermore, SER is often plagued by silent frames and emotionally meaningless frames. In the meantime, the attention mechanism has shown exceptional abilities in studying relevant feature representations for complex tasks. They considered using the Mel spectrogram with deltas and delta-deltas as data to train 3-D attention-dependent convolutional recurrent neural networks (ACRNNs) to learn discriminative features for SER. Experiments on the Emo-DB and IEMOCAP corpora reveal that the suggested method works well and achieves best-in-class unweighted average recall.

Log-Mel spectrograms and high-level features are learned from raw audio clips by Zhao *et al.* [22]. In this research, the authors created a combined convolutional neural network (CNN) with two branches, namely, 1D CNN and 2D CNN. There are two stages in constructing the combined deep CNN. The two designed architectures' hyperparameters are chosen using Bayesian optimization in training. After designing and evaluating one 1D CNN and one 2D CNN architecture, the two CNN architectures were combined after removing the second dense layer. Transfer learning was added to the training to help speed up the training of the combined CNN. The first two CNNs to be trained were the 1D and 2D CNNs. The 1D and 2D CNN's learned features were then repurposed and converted to the combined CNN. The final step was to fine-tune the merged deep CNN that had been initialized with migrated functionality. Experiments show that combining deep CNNs will significantly boost emotion classification results when tested on two benchmark datasets.

Yenigalla *et al.* [23] suggested a phoneme-based and spectrogram-based approach for speech emotion detection. The phoneme sequence and spectrogram both preserve the emotional content of expression, lost as it is translated to text. They used various deep neural networks with phonemes and spectrograms as inputs to conduct multiple experiments. Three of these network architectures are discussed there, and compared to state-of-the-art approaches on a comparison dataset, they helped to achieve better precision. The phoneme and spectrogram hybrid CNN model was the most reliable model for understanding feelings on IEMOCAP data. Compared to current state-of-the-art approaches, the average class accuracy and the overall accuracy are improved.

Sarma *et al.* [24] used the IEMOCAP database to analyze many DNN architectures for emotion recognition. First, they contrast different function extraction front ends: they contrast time-domain and frequency-domain with high-dimensional Mel-frequency cepstral coefficient (MFCC) input (equivalent to filter banks) approaches to learning filters as part of the network. The time-domain filter-learning technique gives them the best outcomes. The researchers then looked at various methods for aggregating data throughout a speech. They experimented with approaches that use time aggregation within the network and single label per utterance and approaches that use a label that is replicated with each frame. The best design they tried interleaves time-restricted self-attention with time-delay neural network (TDNN) + LSTM and achieves a weighted precision of 70.6% percent, compared to 61.8% achieved by the most promising method presented previously that was based on Fourier log-energy input with 257 dimensions.

Latif *et al.* [25] used a novel transition learning methodology in cross-language and cross-corpus situations to enhance the accuracy of SER systems. Compared to support vector machines (SVMs) and sparse autoencoders, deep belief networks (DBNs) offer greater accuracy on cross-corpus emotion detection than previous approaches on five different corpora in three different languages. The results also show that using many languages for training and only a small portion of the target data in training will greatly improve accuracy compared to the baseline, including for corpora with few examples for training.

Zhao *et al.* [26] proposed two CNN+LSTM networks, one 1D CNN+LSTM network, and one 2D CNN+LSTM network, to learn local and global emotion-related features from speech and log-Mel spectrograms, respectively. The architecture of the two networks is identical, with four local function learning blocks (LFLBs) and one LSTM layer in each. LFLB is designed to learn local correlations and derive hierarchical correlations, and it consists primarily of one convolutional layer and one max-pooling layer. The LSTM layer is used to learn long-term dependencies from the local learned functions.

Sun *et al.* [27] presented a new algorithm that incorporates both a sparse autoencoder and a method for focusing attention. The goal is to use an autoencoder to learn from both labeled and unlabeled data and to use the attention function to focus on speech frames with strong emotional content. Such nonemotional speech frames can also be overlooked. Three online databases with a cross-language system are used to test the proposed algorithm. Compared to current speech emotion detection algorithms, experimental findings reveal that the proposed algorithm provides substantially more reliable predictions.

Jiang *et al.* [28] suggested a feature representation extraction method based on deep learning from heterogeneous acoustic feature groups that could include redundant and irrelevant content, resulting in poor emotion recognition output in their research. A fusion network is learned to jointly learn the discriminative acoustic feature representation and SVM as the final classifier after the informative features are obtained. The proposed architecture increased recognition efficiency by 64% compared to current state-of-the-art methods, according to experimental findings on the IEMOCAP dataset.

Pandey *et al.* [29] provided an overview of deep learning strategies for extracting and classifying emotional states from speech utterances. They investigate the most commonly used simple deep learning architectures in the literature. On the two common datasets, Emo-DB and IEMOCAP, architectures such as CNN and LSTM were used to measure the emotion capture capability of various standard speech representations such as Mel-spectrograms, magnitude spectrograms, and MFCCs. The experiments' results and the reasoning behind them have been discussed to determine which architecture and function combination is best for speech emotion detection.

Meng *et al.* in [30] employed the bidirectional LSTM along with CNN to recognize speech emotions. In addition, they adopted the Mel-spectrogram features in the 3D space as the main features used to train the CNN network. That model was evaluated based on IEMOCAP and Emo-DB datasets. Although the results achieved by this model are promising, it lacks generalization, as the model performs well on the training data; however, the performance is worse on the test set.

Zhen *et al.* in [31] proposed a model composed of CNN, BLSTM, and SVM for recognizing the speech emotions based on log-Mel spectrogram features. The model is evaluated on the IEMOCAP dataset and shows better performance when compared with another approach in the literature. Despite the promising performance of the model, it still needs to be evaluated using other datasets to show its generalization capability. On the other hand, the study presented in [32] showed the performance of various models used in SER using six speech datasets. This study concluded that the CNN+LSTM model performs better than the other models for five out of the six datasets.

Lili Guo *et al.* [33] employed kernel extreme learning machine (KELM) for classifying classes of speech emotions. In this approach, a fusion of spectral features is used to train the presented model. The evaluation of this model is performed in terms of two datasets, Emo-DB and IEMOCAP. However, the presented results show promising performance on only one dataset, which means that the presented approach lacks proper generalization. In addition, the authors concluded that the fusion of the spectral features allows the models to achieve higher classification accuracy.

Misbah *et al.* in [34] investigated the application of a deep convolutional neural network (DCNN) to extract features from the log-Mel spectrogram of the raw speech. The study employed four datasets, IEMOCAP, Emo-DB, SAVEE, and RAVDESS. The classification of speech emotions is performed using four classifiers: SVM, random forest, k nearest neighbors, and neural networks. The performance of these classifiers is promising; however, no single classifier could perform well on the four datasets. This indicates that these classifiers lack generalization capability.

Sonawane *et al.* [35] demonstrated a deep learning approach for speech emotion understanding. For the classification of emotions such as positive, negative, indifferent, disgust, and surprise, a multilayer convolutional neural network is used with a basic K-nearest neighbor (KNN) classifier. The combination of MFCC-CNN and the KNN classifier performs better than the current MFCC algorithm, according to experimental findings on a real-time database obtained from the open-access social media site YouTube.

Sajjad *et al.* [36] presented a new SER system focused on Radial basis function network (RBFN) similarity calculation in clusters and the main sequence segment selection method. The STFT algorithm is used to transform the chosen sequence into a spectrogram, which is then fed into the CNN model, which extracts the discriminative and salient features from the speech spectrogram. Additionally, to ensure precise recognition performance, CNN features were normalized and fed to the deep bidirectional long short-term memory (BiLSTM) for emotion recognition based on the learned temporal information.

Kwon *et al.* [37] made significant contributions to (1) improving SER precision in comparison to other methods and (2) improving the complexity of the proposed SER model. They suggest an artificial intelligence-assisted deep stride convolutional neural network (DSCNN) architecture based on the simple net approach to learn salient and discriminative features from spectrograms of speech signals. The hidden local features are learned in convolutional layers rather than pooling layers, with unique strides to downsample the feature maps, and fully connected layers are used to learn the global features. This approach was based on a softmax classifier for classifying speech emotions. On the RAVDESS and IEMOCAP datasets, the proposed strategy improves the overall accuracy by 4.5% and 7.85%, respectively.

Vryzas *et al.* [38] developed and tested SER based on CNN. On consecutive time frames of continuous expression, emotion recognition is performed. The acted emotional speech dynamic database (AESDD) is the dataset used for training and analyzing the model and the techniques of data augmentation. The AESDD is subjected to arbitrary evaluations to act as a benchmark for human-level identification performance. In terms of precision, the CNN model outperforms the other models using SVM by 8.4%.

Ngoc-Huynh *et al.* [39] presented a multimodal approach for recognizing speech emotions. The presented approach is based on a multi-Level multi-head fusion (MLMHF) attention mechanism, and recurrent neural network [44]. MFCC features are utilized in the presented approach. Three datasets are employed to evaluate the presented approach: IEMOCAP, MELD, and CMU-MOSEI. Despite the promising performance achieved by this approach, the performance varies greatly depending on the tested dataset. Therefore, it can be noted that this approach does not generalize well, based on the presented results.

Orhan *et al.* [40] presented a model based on 3D CNN+LSTM that an attention model guides. This model follows the approach of deep end-to-end learning. The features extracted from the speech signals to train the model are Mel-frequency coefficients. The presented model is evaluated using three datasets: RAVDESS, SAVEE, and RML. The achieved results by this model are 96.18%, 87.50%, and 93.32%, respectively.

Turker *et al.* [41] developed a nonlinear multi-level feature generation model is based on cryptographic structure. The performance of that model is validated using four speech emotion datasets, namely, RAVDESS, Emo-DB, SAVEE, and EMOVO. The presented model achieved 87.43%, 90.09%, 84.79%, and 79.08% classification accuracy based on

TABLE 1. Summary of the studies conducted on recognizing speech emotions.

Reference	Year	Methodology	Feature	Dataset	Accuracy (%)
[21]	2018	ACRNNs	Mel-spectrogram	IEMOCAP Emo-DB	68.43 75.47
[22]	2018	Merged deep CNN	Log-Mel spectrograms	IEMOCAP Emo-DB	89.77 92.71
[23]	2018	Multi-channel CNN	Spectrogram	IEMOCAP	73.90
[24]	2018	TDNN+LSTM	MFCC	IEMOCAP	70.10
[25]	2018	DBN	eGeMAPS	FAU-AIBO IEMOCAP Emo-DB SAVEE EMOVO	77.15 61.32 78.51 68.12 80.11
[26]	2019	CNN+LSTM	Log-Mel spectrogram	IEMOCAP Emo-DB	89.16 92.90
[27]	2019	Sparse autoencoder	Log-Mel spectrogram	CASIA Emo-DB IEMOCAP	83.30 89.70 69.50
[28]	2019	Hybrid DNN	Heterogeneous	IEMOCAP	64.00
[29]	2019	CNN +BLSTM	MFCC + Spectrogram	IEMOCAP Emo-DB	50.05 82.35
[30]	2019	ADRNN	Log-Mel spectrogram	IEMOCAP Emo-DB	74.96 90.78
[31]	2019	CNN+BLSTM+SVM	Log-Mel spectrogram	IEMOCAP	62.31
[32]	2019	CNN + LSTM	Mel Filter bank	RAVDESS SAVEE Emo-DB	65.67 72.66 69.72
[33]	2019	KELM	Spectrogram fusion	IEMOCAP Emo-DB	57.99 92.45
[34]	2020	DCNN + CFS + ML	Log-Mel spectrogram	RAVDESS IEMOCAP SAVEE Emo-DB	81.30 83.80 83.80 82.10
[35]	2020	ICNN	MFCC	CASIA	96.32
[36]	2020	RBFN+BiLSTM	Spectrogram	IEMOCAP Emo-DB RAVDESS	72.25 85.57 77.02
[37]	2020	DSCNN	Spectrogram	IEMOCAP RAVDESS	84.00 80.00
[38]	2020	CNN+SVM	Raw signal	AESDD	69.20
[39]	2020	MLMHF-attention	MFCC	IEMOCAP MELD CMU-MOSEI	76.98 63.26 99.19
[40]	2021	CNN+LSTM	Composite features	RAVDESS SAVEE RML	96.18 87.50 93.20
[41]	2021	SVM	Twine-shuf-pat	RAVDESS Emo-DB SAVEE EMOVO	87.43 90.09 84.79 79.08
[42]	2021	CNN+LSTM	MFCC	IEMOCAP	79.52
[43]	2022	StarGAN+DCNN	Log-Mel spectrogram	RAVDESS Emo-DB SAVEE	97.36 91.06 92.97

these datasets, respectively, using a 10-fold cross-validation strategy.

A summary of the relevant milestones of SER in the literature is presented in Table 3. This summary is presented in terms of the year of the publication, the proposed methodology, the type of features utilized in the research, the dataset employed, and the achieved accuracy corresponding to each dataset.

III. PROPOSED METHODOLOGY

This section explains the proposed speech emotion recognition (SER) methodology. The proposed approach consists of a proposed data augmentation algorithm, a proposed CNN+LSTM deep neural network, and a proposed optimization approach using a stochastic fractal search-guided whale optimization algorithm (SFS-Guided WOA) for optimizing the parameters of the deep network. Figure 1 depicts the overall architecture of the proposed SER methodology.

A. DATA AUGMENTATION

A large amount of training data is usually required for deep learning to achieve better results. One way to increase the number of training samples is through data augmentation. In this paper, we propose a new data augmentation algorithm as presented in Algorithm (1). This algorithm creates additional training samples by carefully adding fractions of noise to the clean samples. The choice of this fraction is critical, as it may corrupt the signal content if the amount of noise is large or may be irrelevant if the amount of noise is too small. In this paper, we adopted the noise ratio as the 0.005 × *max* value in the speech signal. In this research, after performing data augmentation, each clean sample in the dataset will have three new samples generated by the augmentation algorithm. Therefore, the ratio of the clean to the newly generated samples in the augmented dataset is 1:3.

Algorithm 1 Data Augmentation

```
procedure Augment data
    Ratio $\leftarrow$ 0.005
    Max $\leftarrow$ np.amax(data)
    rUniform $\leftarrow$ np.random.uniform()
    NoiseFactor $\leftarrow$ Ratio $\times$ Max $\times$ rUniform
    Noise $\leftarrow$ np.random.randn(len(data))
    AugmentedData $\leftarrow$ data + NoiseFactor $\times$ Noise
    return AugmentedData
end procedure
```

[1] **np**: refers to the Python NumPy module.

The addition of this fraction of noise to the clean signal is significant to improve the generalization of the proposed deep learning model. On the other hand, the existence of the clean samples in the dataset makes the model capable of recognizing the speech emotions of a clean signal as well as the noisy signal.

B. FEATURE EXTRACTION

The features extracted from the speech dataset are represented in the 2D space as log-Mel spectra. These features are employed as a static input to the deep network to achieve a better distribution of emotional features. In addition, this representation of features can extract the features corresponding to the emotions of interest accurately when compared with the raw spectrum and with a reduction in the dimensionality of the feature space [30]. Moreover, the log-Mel spectrum helps to reduce the effect of interference that may occur in the frequency bands and improve the linearization of the frequency perception of human ears [43]. Consequently, the speed of training the classification model along with the recognition process can be significantly improved.

To map the signal frequency to a log-Mel spectrum, equation (1) is employed.

$$Mel(k) = 2,595 \times log\left(1 + \frac{f}{700}\right) \tag{1}$$

where k represents the frequency of the *Mel* scale and f denotes the frequency that moves on the scale of $0 \leq f \leq 22,050$.

The process of extracting the log-Mel spectrum is represented in the following steps.

- Framing and windowing: A window size of 25 ms or equivalently 256 samples is used as an analysis window. To smoothly cover the spectrum variation, a skip rate of 50% is also applied. The analysis window is applied in terms of the Hamming window to effectively reduce the signal distortions. The Hamming window is expressed as presented in equation (2) for the window length is denoted by N, and δ is usually set as 0.46.

$$w(n) = \begin{cases} (1-\delta)+\delta cos\left(\dfrac{2\pi n}{N-1}\right) & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases} \tag{2}$$

- Mel filter: To measure the energy of the speech signal, the modulus of the frequency spectrum is squared. Then, a set of triangles is applied to the Mel scale of the energy spectrum. The application of this set of filter banks helps reduce the harmonics and improve the smoothness of the frequency spectrum. In addition, these filter banks can reduce the time needed to calculate the resulting output while reducing the dimension of the feature space. In most speech processing approaches, the number of filter banks is usually 13. The output from each triangular filter is defined as shown in equation (3), as shown at the bottom of the next page, for $k \in [0, 255]$, $m \in [0, 13]$, and the image of the Mel-frequency filter bank is characterized by the function denoted by $f(.)$. Due to the relation between the methodology of Mel-spectrogram and its inspiration from the human auditory system, it is usually used in several operation of speech processing, such as speech recognition, speech synthesis, speech emotion, etc.

C. THE PROPOSED CNN+LSTM

To understand speech emotions, researchers have one key challenge, which is the extraction of most distinctive

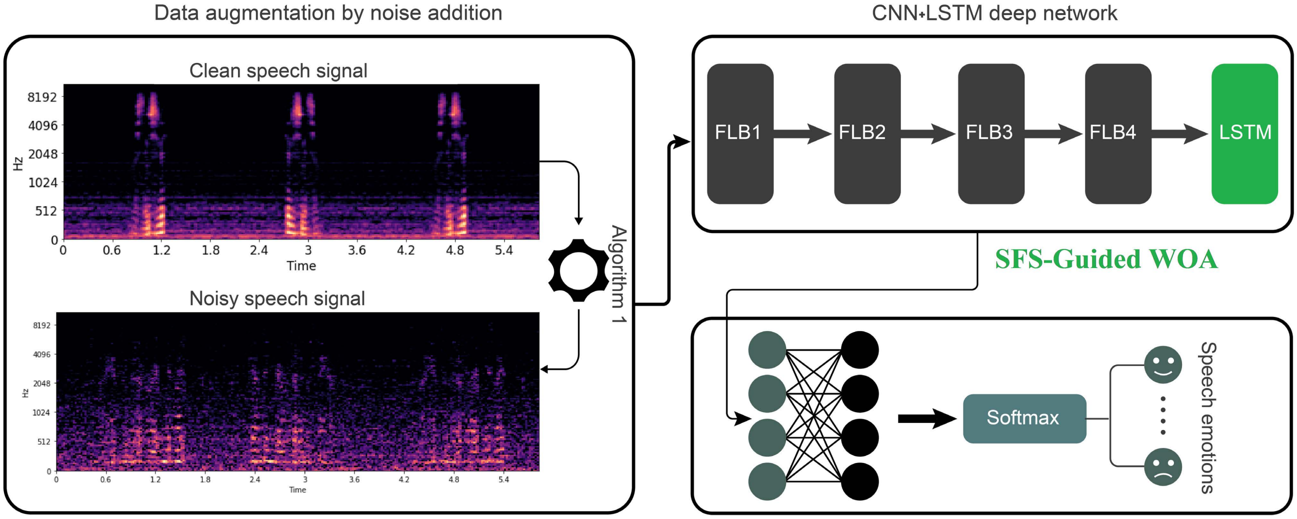


FIGURE 1. Overall architecture of the proposed speech emotion recognition.

features that represent emotions accurately. Based on the existing methods of feature extraction, speech features can be categorized as either studied features or handcrafted features. Deep neural networks, such as CNNs, offer a simple way to extract features that can achieve exceptional performance [45], [46].

To extract accurate emotional features, four neural network layers were stacked together to form a local feature-learning block (LFLB). These layers include a convolutional layer for performing the convolution of the speech features with a kernel mask, a normalization layer called batch normalization (BN) [47], an exponential linear unit (ELU) [48], and finally, a max-pooling layer. Four blocks of LFLB are then stacked to form the general architecture of the proposed approach, as shown in Figure 2. The LFLB's main layers are the convolution and max-pooling layers [49]–[51]. The function of the learning kernel is performed by the convolution layer. The BN layer increases the efficiency and reliability of deep networks by normalizing the activation of the convolutional layer in each batch. The batch normalization transition keeps the standard deviation of activation near the value of one and the mean activation near the value of zero [52].

The ELU layer controls the BN layer's performance. ELU has negative value, which resets the mean of the activation layer, allowing the learning rate to become much faster and thus boosting the recognition accuracy accordingly. The features can show noise and vibration resistance by using a pooling layer. Nonlinear functions, such as max-pooling, are the most widely used functions that can help in dividing the input into non-overlapping regions along with their max-values [53].

In this research, log-Mel spectrogram is used to extract local and global features that are then learned using a combination of LSTM and LFLB. The central layer of the LFLB is the convolution layer, which is designed to process a grid of values. It will learn sequence features based on the neighboring inputs. In particular, each feature element is formed in terms of a small number of these neighboring inputs. On the other hand, the learned features are based on the previous outputs. High-level features can be learned by LSTM and CNN in conjunction and provide both long-term and local contextual information.

The result $z(i,j)$ can be measured by the convolution of $x(i,j)$ with kernel $w(i,j)$, which has a size of $a \times b$. In contrast,

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m) \le k \le f(m+1) \\ 0 & k \ge f(m+1) \end{cases} \tag{3}$$

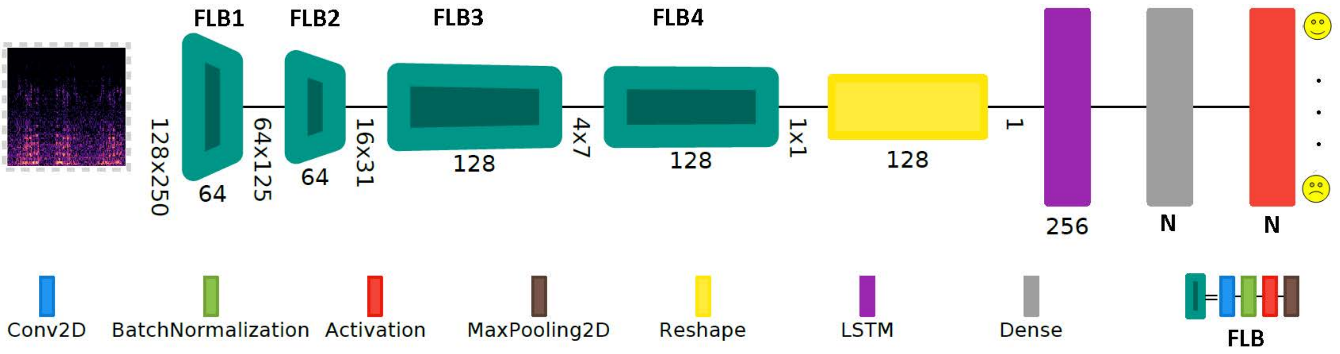


FIGURE 2. **The architecture of the proposed CNN+LSTM deep neural network. Each FLB is composed of Conv2D, Batch normalization, activation and max pooling 2D layers.**

the input to the convolution layer is $x(i, j)$. In the conducted experiments, the initialization of the 2D convolution kernel $w(i, j)$ is chosen arbitrarily.

$$\begin{aligned} z(i,j) &= x(i,j) * w(i,j) \\ &= \sum_{s=-a}^{a} \sum_{s=-b}^{b} x(s,t).w(i-s, j-t) \end{aligned} \tag{4}$$

The BN layer is fed with the convolved features from the previous layer, which are then normalized in each batch. The BN layer uses a transformation to keep the convolved features' variance equals to one and the mean equals to zero. This operation can be interpreted as follows:

$$z_i^l = \sigma(BN(b_i^l + \sum_j z_j^{l-1} * w_{ij}^l)) \tag{5}$$

where z_j^{l-1} and z_i^l refer to the l^{th} layer at which we obtain the i^{th} output and the j^{th} input feature at the $(l-1)^{th}$ layer; the convolution kernel between the j^{th} and i^{th} features is denoted by w_{ij}^l.

The normalization of the features learned by the convolution layer is denoted by the function $BN(\cdot)$. In addition, the network activation function is denoted by $\sigma(\cdot)$ and is defined as:

$$\sigma(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases} \tag{6}$$

where e is Euler's sum and $\alpha > 0$. The nonlinear down-sampling operation is performed by the pooling layer, which decreases the feature resolution. The following are the characteristics provided by the max-pooling layer.

$$z_l^l = \underset{\forall p \in \Omega_k}{max}\, z_p^l \tag{7}$$

where Ω_k represents the k^{th} pooling region. The l^{th} max-pooling layer output and input feature are denoted by z_k^l and z_p^l at index k and p, respectively.

LSTM is used to learn long-term information, however the stacks of LFLB are used to learn the local information. Outside a cell with a self-recurring relationship, LSTM can add or remove data on a block state based on three components: an input, output, and gates. Softmax is used to render predictions that include both local and global context information using the features tested. The fully connected layer is used to generalize these functions into the output space.

When designing a deep architecture, selecting a collection of hyperparameters is crucial. To improve the efficiency of the deep network, optimization of the hyperparameter is performed on a separate data collection. Random search and grid search have been successfully used in several deep learning applications to accelerate deep design training. As Bayesian optimization is suggested, it has been shown that it produces improved outcomes with fewer studies [54]. The Bayesian optimization approach is used to select the hyperparameters for the proposed deep network.

Bayesian optimization is a sequential architecture approach that effectively reduces the objective function. The hyperparameters in our experiments are optimized using Hyperopt, a Python library. Hyperopt determines a minimizable objective function and uses it as a random function [55]. Over the objective function, a prior is often applied. The prior is modified based on the gathered function evaluations to form the posterior distribution over the objective function. Using the posterior distribution, an acquisition mechanism is established. The hyperparameters are then iteratively chosen. The options distribution ('rmsprop', 'sgd', 'adam', 'adagrad') is followed to select an appropriate optimization algorithm. The best model is returned after practicing with the optimized hyperparameters [56].

D. HYPERPARAMETERS OPTIMIZATION

As the proposed CNN+LSTM consists of a set of hyperparameters, the significant parameters in this set are the learning rate and label smoothing regularization factor. The learning rate affects network performance and directly determines the convergence speed along with the model accuracy. On the other hand, the smoothing regularization factor affects the intensity of the disturbance applied to the correct labels and thus affects the correctness of the input labels to the model. In this research, both of these parameters are optimized to determine their optimal values to improve the trained model accuracy. The optimization of these parameters is performed in terms of the recently published SFS-guided WOA.

The basic idea of SFS-Guided WOA algorithm is based on the behavior of whales, which trap their prey using bubbles that push them up to the surface in the form of a spiral loop. In this SFS-Guided WOA, there is a whale that looks for the optimal values of the parameters, and this whale is guided by three other random whales [57]. This strategy is useful in improving the exploration and exploitation features of this optimization task. The representation of these whales is described by the following equation.

$$\begin{aligned}\vec{W}(t+1) = \vec{w_1} * \vec{W}_{rand1} \\ + \vec{z} * \vec{w_2} * (\vec{W}_{rand2} - \vec{W}_{rand3}) \\ + (1 - \vec{z}) * \vec{w_3} * (\vec{W}(t) - \vec{W}_{rand1}) \quad (8)\end{aligned}$$

where $\vec{W}_{rand1}$, $\vec{W}_{rand2}$, and $\vec{W}_{rand3}$ represent the three random whales, where each random whale represents a potential solution. $\vec{W}(t)$ and $\vec{W}(t+1)$ indicate the current and updated solutions at iteration number t. The $\vec{w_1}$, $\vec{w_2}$ and $\vec{w_3}$ parameters are three random variables with values of [0, 0.5], [0, 1], and [0, 1], respectively. To smoothly change between exploitation and exploration, the value of $\vec{z}$ decreases exponentially using the following equation for Max_{iter} indicates the maximum number of iterations.

$$\vec{z} = 1 - \left(\frac{t}{Max_{iter}}\right)^2 \quad (9)$$

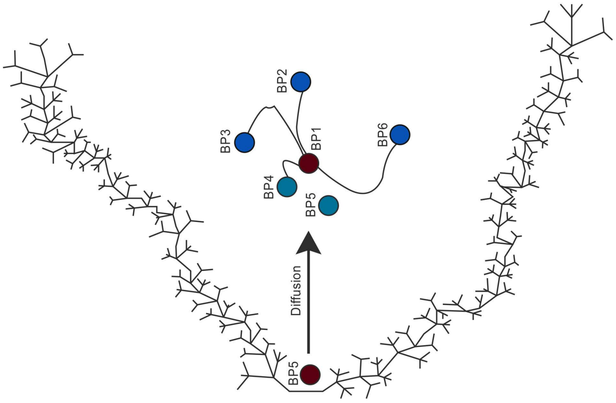


FIGURE 3. The diffusion around the best solution in a random fractal sample.

Statistical fractal search employed in this algorithm depends on diffusion-limited aggregation (DLA), which generates the objects' fractal shape. The SFS technique uses diffusion and two kinds of updating processes. Figure 3 depicts a graphical form of the SFS diffusion process. For the best-solution BP, a list of solutions BP1, BP2, BP3, BP4, and BP5 are listed around this best solution. Algorithm 2 presents the full process of SFS-guided WOA. For more details about this optimization algorithm, please refer to [58].

IV. EXPERIMENTAL RESULTS

To recognize the embedded emotions in the input speech, the speech utterance is segmented if it was longer than 8 seconds, and is padded to 8-second length otherwise. The FFT with window length of 2048, and hop length of 512 are used in the process of computing the log-Mel spectrogram. Consequently, the log-Mel spectrogram is estimated with 251 frames and 128 Mel frequency bins [59]. The 128 × 251 matrices are employed in the conducted experiments to provide a feedback to the CNN+LSTM network. The resulting 2D log Mel-spectrogram patches are fed to the CNN+LSTM network to learn the high-level contextual information.

A. EXPERIMENTAL PLATFORM

The platform used in running the conducted experiments has a set of parameters presented in Table 2. The main factor for accelerating the training process is the utilization of the available GPU and memory. These resources allow running the experiments with a batch of size >= 16, which enables completing the model training process in a relatively short time.

TABLE 2. Specifications of the experimental platform.

Hardware environment	CPU GPU RAM	Intel Core i7 GeForce RTX2070 Super 16 GB
Software environment	Platform	Ubuntu 20.04 TensorFlow 1.15 CUDA9.0 + Cudnn7.1 Spider + Python3.7

B. EXPERIMENTAL DATASETS

In this research, four datasets were included in the conducted experiments. These datasets are introduced in the following.

- **RAVDESS**: This dataset is composed of audio clips of songs and speech. The clips are recorded by 24 speakers; 12 women and 12 men. The emotional expressions included in the speech clips are surprise, fear, anger, sadness, happiness, calm, and disgust. On the other hand, the expressions included in the song clips are fear, anger, calm, sadness, and happiness. Each sentence is recorded twice by each speaker. The number of song clips is 1,012 and the number of speech clips is 1,440.
- **Emo-DB**: This dataset is recorded by the Department of Technical Acoustics, the Technical University of Berlin. The recording is performed in anechoic chamber. The number of utterances included in this dataset is 500. The emotions included in the recorded utterances are disgust, boredom, fearfulness, anxiety, anger, happiness, and neutral. The speakers of the utterances are aged between 20 and 30 years.
- **SAVEE**: This dataset is collected by the University of Surrey. The emotions included in the recordings of this

Algorithm 2 :SFS-Guided-WOA [57]

```
Initialize the population $\vec{W}_i (i = 1, 2, \ldots, n)$ with fitness function $F_n$, where $n$ is the population size and $Max_{iter}$ represents the maximum number of iterations.
Initialize the WOA parameters $\vec{a}, \vec{A}, \vec{C}, l, \vec{r_1}, \vec{r_2}, \vec{r_3}$
Initialize the guided-WOA parameters $\vec{w_1}, \vec{w_2}, \vec{w_3}$
Set $t = 1$
Convert the solutions to binary solutions [0 or 1].
Calculate the fitness function $F_n$ for each agent $\vec{W}_i$
Find the best solution $\vec{W^*}$
while $t \leq Max_{iter}$ do
  for $(i = 1 : i < n + 1)$ do
    if $(\vec{r_3} < 0.5)$ then
      if $(|\vec{A}| < 1)$ then
        Update the position of current search agent as in the following equation.
        $\vec{W}(t+1) = \vec{W}^*(t) - \vec{A}.\vec{D},$
        $\vec{D} = |\vec{C}.\vec{W}^*(t) - \vec{W}(t)|$
      else
        Select the three random search agents $\vec{W}_{rand1}$, $\vec{W}_{rand2}$, and $\vec{W}_{rand3}$ from the current solutions.
        Update the $(\vec{z})$ parameter by the following exponential form.
        $\vec{z} = 1 - \left(\frac{t}{Max_{iter}}\right)^2$
        Update position of current search agent as in the following equation.
        $\vec{W}(t+1) = \vec{w_1} * \vec{W}_{rand1} + \vec{z} * \vec{w_2} * (\vec{W}_{rand2} - \vec{W}_{rand3}) + (1 - \vec{z}) * \vec{w_3} * (\vec{W}(t) - \vec{W}_{rand1})$
      end if
    else
      Update the position of current search agent as in the following equation.
      $\vec{W}(t+1) = \vec{D}'.e^{bl}.cos(2\pi l) + \vec{W}^*(t)$
    end if
  end for
  for $(i = 1 : i < n + 1)$ do
    Calculate the following equation to update solutions based on SFS algorithm.
    $\vec{W_i'^*} = Gaussian(\mu_{\vec{W^*}}, \sigma) + (\eta \times \vec{W^*} - \eta' \times \vec{P_i})$
  end for
  Update $\vec{a}, \vec{A}, \vec{C}, l, \vec{r_3}$
  Convert the updated solution to binary using sigmoid function.
  Calculate the fitness function $F_n$ for each agent $\vec{W}_i$
  Find the best solution $\vec{W^*}$ from the updated solutions.
  Set $t = t + 1$
end while
Return $\vec{W^*}$
```

dataset are neutrality, surprise, sadness, happiness, disgust, anger,and fear. These recordings are collected by four men aged between 27 and 31 with native English-speaking.

- **IEMOCAP**: This dataset is multispeaker, and multimodal. The number of recording hours of this dataset is 12. The collected data includes text transcriptions, motion capture of faces, speech, and video. The dataset is composed of frustrating, exciting, happiness, anger, sadness, neutrality.

To used these datasets in the conducted experiments, each dataset is split into 80% for training/validation and 20% for testing. In addition, to allocate a subset of this dataset for validation, the training/validation part is split further into 80% for training and 20% for validation. The input to the proposed model is the log Mel-spectrograms of the input speech utterance. The log Mel-spectrograms are calculated for 3 seconds of the input speech utterance. Utterances that are less than 3 seconds are extended to 3 seconds by a zero-padding operation, otherwise they are split into 3 seconds chuncks.

C. MODEL TRAINING AND TESTING

The experimental data were arbitrarily divided into two groups, with the training group receiving 80% of the data and the study set receiving 20%. Experiments of

TABLE 3. Summary of the architecture of the proposed CNN+LSTM deep network.

Layer	Output Shape	# Params
Conv2D	(128, 251, 64)	640
Batch_normalization	(128, 251, 64)	256
Activation	(128, 251, 64)	0
MaxPooling2D	(64, 125, 64)	0
Conv2D	(64, 125, 64)	36,928
Batch_normalization	(64, 125, 64)	256
Activation	(64, 125, 64)	0
MaxPooling2D	(16, 31, 64)	0
Conv2D	(16, 31, 128)	73,856
Batch_normalization	(16, 31, 128)	512
Activation	(16, 31, 128)	0
MaxPooling2D	(4, 7, 128)	0
Conv2D	(4, 7, 128)	147,584
Batch_normalization	(4, 7, 128)	512
Activation	(4, 7, 128)	0
MaxPooling2D	(1, 1, 128)	0
Reshape	(1, 128)	0
LSTM	(256)	39,4240
Dense	(6)	1542
Activation	(6)	0
Total number of params		656,326
Number of trainable params		655,558
Number of nontrainable params		768

comparable findings demonstrate that the CNN+LSTM network is capable of accurately detecting speech emotions. On average precision, the constructed CNN+LSTM network performs satisfactorily compared to other well-established function representations and methods.

In the conducted experiments, only the most accurate and well-fit models are taken into consideration. The validity accuracy of the learned model is an important predictor of its generalization. The best predictive model will be available when the validation accuracy hits its limit during CNN+LSTM network preparation. As a result, the recorded model not only suits the experimental results well but also performs well in terms of predicting SER.

The CNN+LSTM deep network architecture is summarized in Table 3. Four local function learning blocks are depicted in the table. Convolutional layers, batch normalization, activation, and max-pooling layers are all used in each learning block. The table also shows the form of each layer. An LSTM layer is applied after our local function learning blocks to learn the global feature from the input spectrogram.

To verify the generalization ability of the developed CNN+LSTM network, the performance is recorded for the training and verification sets. Five-fold cross-validation was used to evaluate the true generalization error of the network. Figure 4 depicts the progress of the loss values during the training of the network. As shown in the figure, the model could learn the significant features necessary for classifying speech emotions accurately. The loss values become close to zero after starting from epoch number 60.

In the literature, many methods have been proposed to reduce the likelihood or the amount of overfitting in studies. Bad predictions for untrained sample data are caused in part by overfitting. When a model is overfitted, it memorizes the training data instead of learning to predict better. The phenomenon of overfitting can be caused by a variety of factors. Overfitting can occur because of the complexity of the deep network or because the network is overtrained. Therefore, model selection, early stopping, batch normalization, regularization, and cross-validation are adopted to overcome overfitting [60]–[63].

Early stopping, as shown in Figure 4, will prevent overtraining and increase the prediction ability of the model. Performance monitoring can be used to track training accuracy and validation accuracy. The number of epochs with no change in the display is the patience. The network would have superior predictive efficiency, while the validity accuracy does not increase in testing.

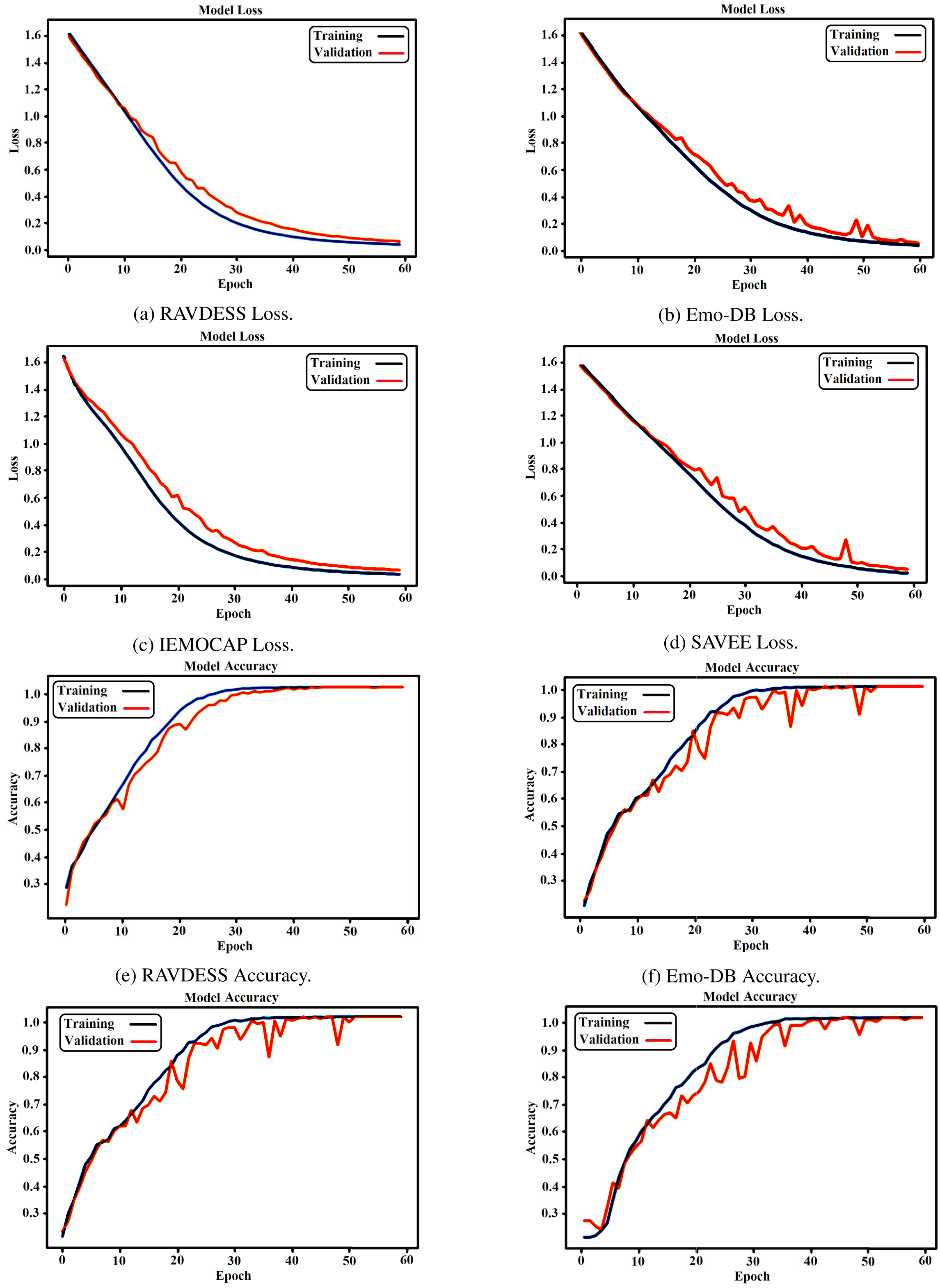
In addition, the accuracy of the trained model is recorded during the training process. Figure 4 shows the progress of the accuracy during the training epochs. As shown in the figure, the progress of the accuracy of the trained model moves smoothly for the selected learning rate. The model accuracy of the training sets achieves the best performance after the 60th iteration for the four datasets. In addition, the progress of the validation accuracy stabilizes after reaching the 60th iteration, which means that the model learns the training data accurately and is ready to generalize for the test set.

The accuracy increased significantly as a result of the use of data augmentation during the training period. As a result, the average recognition accuracy of the correctly identified emotions in the test sample was 99.2%, which is higher than all current competing approaches. The other rival method with close recognition precision was introduced in [26], which is based on a CNN+LSTM deep network but does not use data augmentation, making it less resilient to input speech emotions. This comparison clearly shows that the suggested solution outperforms the competition in regard to understanding speech emotions.

The confusion matrix of the recognition of speech emotions in the test set is shown in Figure 5. The test set is usually the final judge of the effectiveness of the developed approach. As shown in this figure, the proposed approach can successfully recognize almost all the speech emotions in the test set with very high accuracy. This reflects the efficiency of the proposed deep network along with the notion of data augmentation and parameter optimization, which positively affects the overall recognition accuracy.

D. COMPARISON WITH EXISTING SYSTEMS

The proposed approach is compared with a set of competing approaches in the literature to validate the superiority of the proposed approach. Table 4 presents the classification accuracy achieved by each approach, including the proposed approach. As shown in the table, the performance of the proposed approach outperforms the performance of the other approaches applied to the RAVDESS dataset, where the maximum accuracy achieved was 97.36.46%, but the proposed approach could achieve an accuracy of 99.47%. A similar interesting performance of the proposed approach



(a) RAVDESS Loss.

(b) Emo-DB Loss.

(c) IEMOCAP Loss.

(d) SAVEE Loss.

(e) RAVDESS Accuracy.

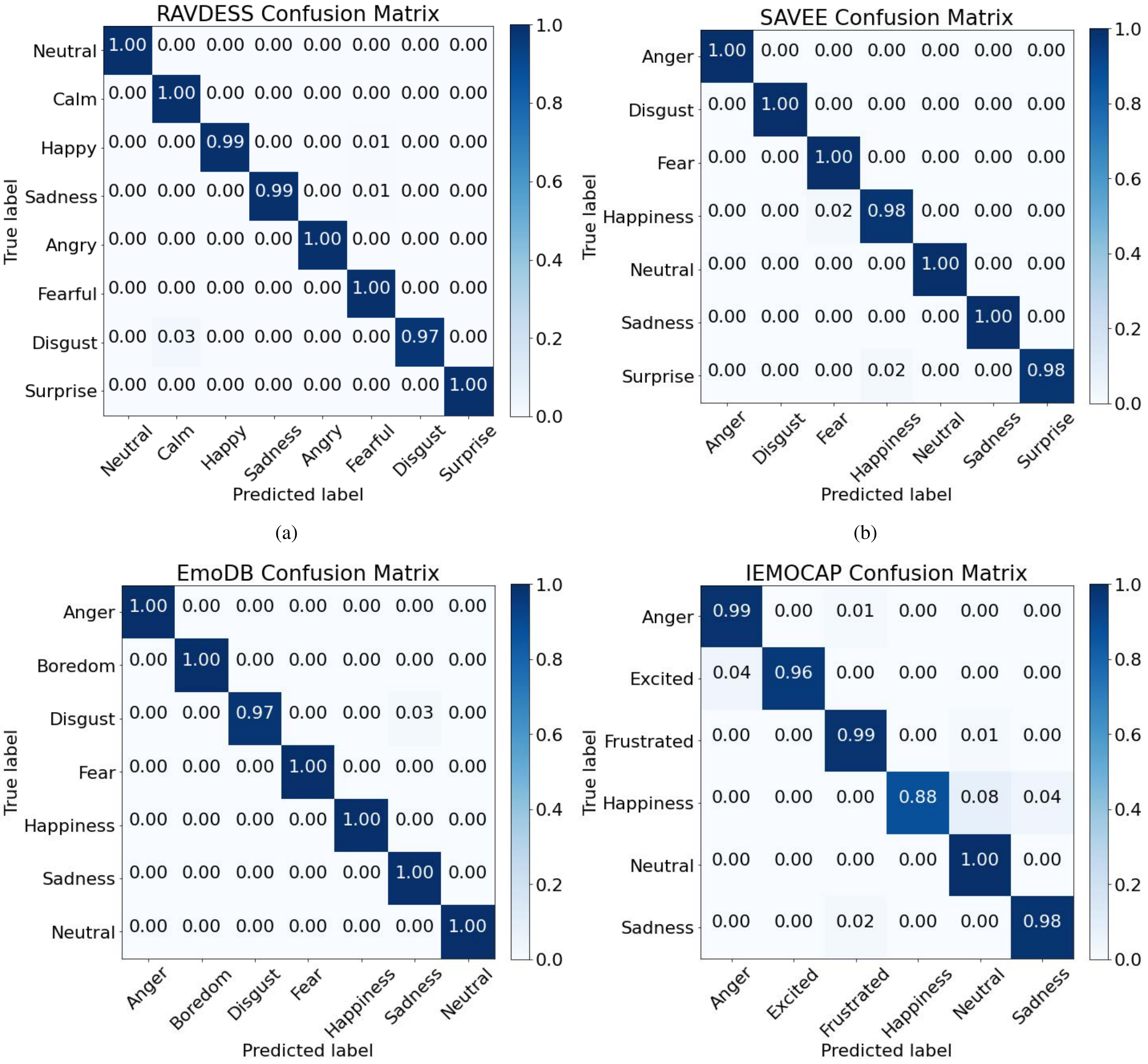
(f) Emo-DB Accuracy.

(g) IEMOCAP Accuracy.

(h) SAVEE Accuracy.

FIGURE 4. **Progress of the loss and accuracy values using four speech emotion datasets during the training process.**

is achieved when it is compared with the performance of other approaches applied to the SAVEE and Emo-DB datasets. On the other hand, the existing approaches could not achieve an accuracy of more than 89.16% when applied to the



(a)

(b)

(c)

(d)

FIGURE 5. Confusion matrix of testing the four datasets (a) RAVDESS, (b) SAVEE, (c) Emo-DB, and (d) IEMOCAP.

IEMOCAP dataset. However, the proposed approach could achieve higher accuracy, which reached 98.13% on the same dataset.

E. STATISTICAL ANALYSIS

Another perspective of evaluation of the proposed approach is presented in this section in terms of in-depth statistical analysis of the achieved results based on the employed datasets and in comparison with the other competing approaches.

Figure 6 presents multiple measures of the statistical analysis of the achieved results. One of these measures is heteroscedasticity, which measures the residual between the predicted emotion and the absolute residual of the recognition values, considering that the sum and mean of the residuals are equal to zero, as shown in Figure 6a. To achieve the ideal case, the residual values should be distributed uniformly around the horizontal axis, which is clearly shown in the figure. In addition, the heteroscedasticity plot is shown in Figure 6b. Homoscedasticity describes whether the error term is the same across the values of independent variables. Figure 6c also shows the quantile-quantile (QQ) plot and probability plot. Since the distributions of points in the QQ plot are well fitted on the predetermined line, the actual and

TABLE 4. Comparison between the performance of the proposed approach and the other competing approaches.

Dataset	Reference	Feature	Classification accuracy (%)
RAVDESS	CNN + LSTM [32]	Mel Filter bank	65.67
	DCNN + CFS + ML [34]	Log-Mel spectrogram	81.30
	RBFN+BiLSTM [36]	Spectrogram	77.02
	StarGAN+DCNN [43]	Log-Mel spectrogram	97.36
	TLCNN-RAM [64]	Log-Mel spectrogram	94.78
	Convolution-LSTM [65]	Log-Mel spectrogram	92.33
	Proposed	Log-Mel spectrogram	**99.47**
SAVEE	CNN + LSTM [32]	Mel Filter bank	72.66
	DCNN + CFS + ML [34]	Log-Mel spectrogram	83.80
	StarGAN+DCNN [43]	Log-Mel spectrogram	92.97
	TLCNN-RAM [64]	Log-Mel spectrogram	89.02
	Convolution-LSTM [65]	Log-Mel spectrogram	85.46
	Proposed	Log-Mel spectrogram	**99.50**
Emo-DB	ADRNN [30]	Log-Mel spectrogram	85.61
	CNN + LSTM [32]	Mel Filter bank	69.72
	DCNN + CFS + ML [34]	Log-Mel spectrogram	82.10
	RBFN+BiLSTM [36]	Spectrogram	85.57
	StarGAN+DCNN [43]	Log-Mel spectrogram	91.06
	TLCNN-RAM [64]	Log-Mel spectrogram	80.71
	Convolution-LSTM [65]	Log-Mel spectrogram	83.46
	Proposed	Log-Mel spectrogram	**99.76**
IEMOCAP	CNN+LSTM [26]	Log-Mel spectrogram	89.16
	ADRNN [30]	Log-Mel spectrogram	74.96
	CNN+BLSTM+SVMs [31]	Log-Mel spectrogram	62.31
	DCNN + CFS + ML [34]	Log-Mel spectrogram	83.80
	RBFN+BiLSTM [36]	Spectrogram	72.25
	CNN+LSTM [42]	MFCC	79.52
	DBN [46]	Mel filer banks	73.78
	Proposed	Log-Mel spectrogram	**98.13**

TABLE 5. Theoretical and actual means of the achieved accuracy by the proposed and the other competing approaches.

	ADRNN	TLCNN+RAM	Convolution-LSTM	DCNN+CFS+ML	CNN+LSTM	Proposed
Theoretical mean	0	0	0	0	0	0
Actual mean	90.21	94.72	92.27	95.03	92.24	99.47
Number of values	18	18	18	18	18	18

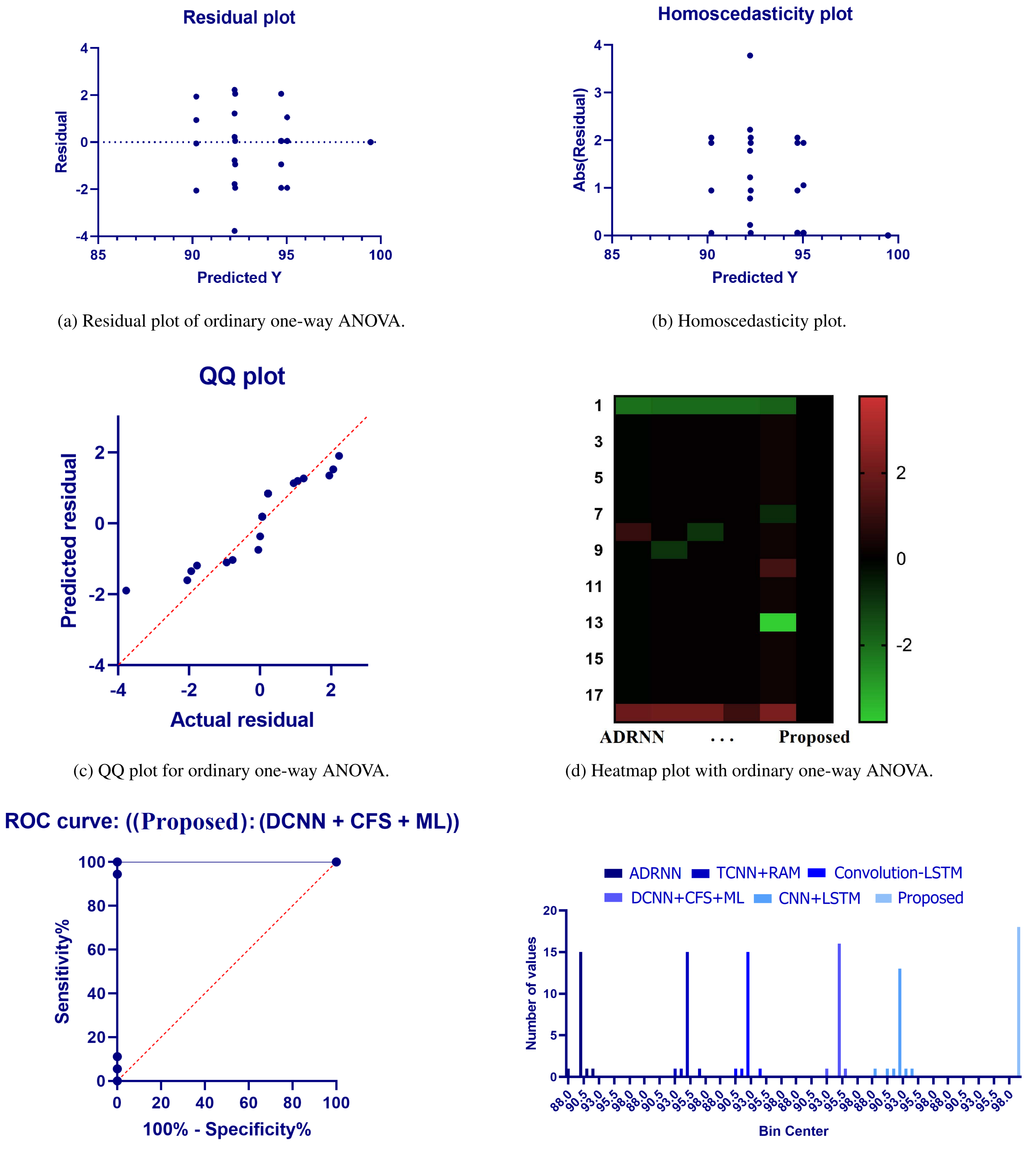
TABLE 6. ANOVA test for the proposed approach and the other competing approaches.

	SS	DF	MS	F(DFn, DFd)	P value
Treatment (between columns)	935.9	5	187.2	F(5,102)= 335.6	P<0.0001
Residual (within columns)	56.89	102	0.5577	-	-
Total	992.8	107	-	-	-

predicted residuals are considered to be linearly related. This confirms the performance of the proposed CNN+LSTM approach. Figure 6d presents the heatmap plot with ordinary one-way ANOVA.

Figure 7 shows the curve of the receiver operating characteristic (ROC) for the proposed optimized CNN+LSTM model and the other competing models. This figure indicates that the proposed model distinguishes data with a large area under the curve (AUC) with a value of approximately 1.0. In addition, a histogram of the achieved accuracies using the proposed and other approaches is presented in Figure 6f. In this figure, the presented models were evaluated using multiple test sets, and then the accuracy was binned and counted to make this plot. As shown in the figure, the proposed approach could achieve a stable performance while varying the test sets from the four datasets. The achieved performance resides in the scale of accuracy $> 98\%$. However, the highest accuracy achieved by the other approaches is within the range of 88% to 95%.

Moreover, Figure 7 shows the ranges of accuracy for each of the presented approaches, including the proposed approach. As shown in the figure, the competing



(a) Residual plot of ordinary one-way ANOVA.

(b) Homoscedasticity plot.

(c) QQ plot for ordinary one-way ANOVA.

(d) Heatmap plot with ordinary one-way ANOVA.

(e) ROC curve.

(f) Histogram of the achieved accuracy.

FIGURE 6. **Statistical analysis of the achieved results.**

approaches expose a variation in the performance, whereas the performance of the proposed approach exposes a stable performance in classifying the speech emotions. These values are represented in terms of the average accuracy for each run in the test set. The theoretical and actual means of the achieved accuracy are also shown in Table 5. The mean values in this table are calculated in terms of the four employed datasets.

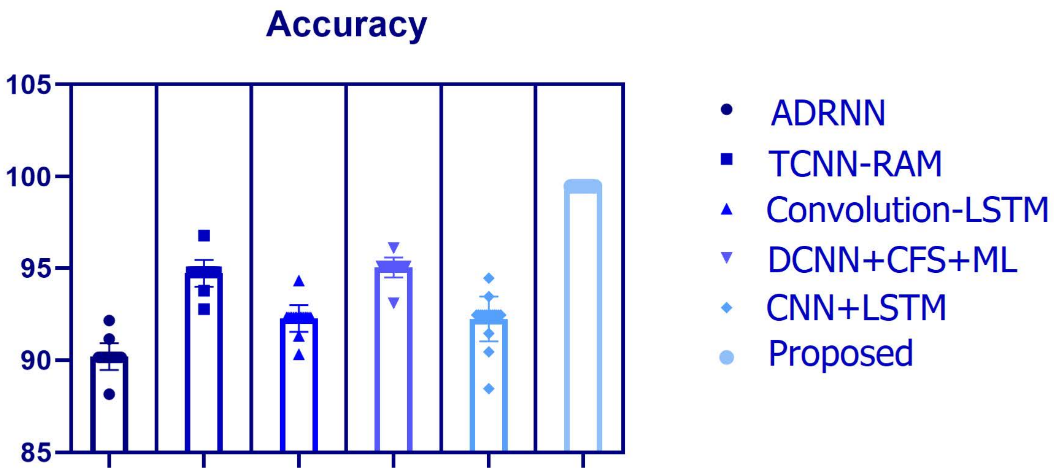


FIGURE 7. Stability of the accuracy achieved by the proposed and the other competing approaches.

TABLE 7. Two-tailed t-test of the performance achieved by the proposed model and the other competing models over 18 runs.

	ADRNN	TLCNN+RAM	Convolution-LSTM	DCNN + CFS + ML	Proposed
t, df	t=527.6, df=17	t=554.0, df=17	t=539.7, df=17	t=747.6, df=17	t=322.0, df=17
P value (two tailed)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
P value summary	****	****	****	****	****
Significant (alpha=0.05)?	Yes	Yes	Yes	Yes	Yes
Discrepancy	90.21	94.72	92.27	95.03	92.24
SD of discrepancy	0.7254	0.7254	0.7254	0.5393	1.215
SEM of discrepancy	0.171	0.171	0.171	0.1271	0.2865
95% confidence interval	89.84 to 90.57	94.36 to 95.09	91.91 to 92.64	94.77 to 95.30	91.63 to 92.84
R squared (partial eta squared)	0.9999	0.9999	0.9999	1.0000	0.9998

TABLE 8. Comparison between the performance of the proposed optimized model with and without data augmentation.

Dataset	Without augmentation	With augmentation
RAVDESS	94.11%	99.47%
SAVEE	92.59%	99.50%
Emo-DB	93.83%	99.76%
IEMOCAP	91.55%	98.13%

TABLE 9. Comparison between the performance of the softmax classifier and two other classifiers.

Classifier	RAVDESS	SAVEE	Emo-DB	IEMOCAP
Softmax	99.50%	99.47%	99.76	98.13
K-NN	93.11%	94.50%	92.32%	90.81%
SVM	97.41%	98.43%	98.91%	96.06%

The ANOVA test results for SER based on the proposed approach compared to other competing approaches are shown in Table 6. The two tailed t-test of SER results based on the proposed approach compared to other approaches is also shown in Table 7. These results confirm the superiority of the proposed approach with parameter optimization using the guided whale optimization algorithm and indicate the statistical significance of the proposed approach for the SER tested problem compared to the other competing approaches. Table 10 shows the proposed algorithm's descriptive statistics compared to other deep learning techniques over 18 runs.

Finally, a set of experiments are conducted to evaluate the effectiveness of the data augmentation and its impact on the achieved results. In this set of experiments, we tested the proposed approach without employing the proposed data augmentation and the results were recorded. Table 8 present the findings of this evaluation. As shown in the

TABLE 10. Statistical analysis of the achieved results based on the proposed approach and the other competing approaches over 18 runs.

	ADRNN	TLCNN-RAM	Convolution-LSTM	DCNN + CFS + ML	CNN+LSTM	Proposed
Number of values	18	18	18	18	18	18
Minimum	88.15	92.78	90.33	93.09	88.46	99.47
25% Percentile	90.15	94.78	92.33	95.09	92.46	99.47
Median	90.15	94.78	92.33	95.09	92.46	99.47
75% Percentile	90.15	94.78	92.33	95.09	92.46	99.47
Maximum	92.15	96.78	94.33	96.09	94.46	99.47
Range	4	4	4	3	6	0
10% Percentile	89.95	93.68	91.23	94.89	90.26	99.47
90% Percentile	91.25	94.98	92.53	95.19	93.56	99.47
Mean	90.21	94.72	92.27	95.03	92.24	99.47
Std. Deviation	0.7254	0.7254	0.7254	0.5393	1.215	0
Std. Error of Mean	0.171	0.171	0.171	0.1271	0.2865	0
Coefficient of variation	0.8041%	0.7658%	0.7861%	0.5675%	1.318%	0.000%
Geometric mean	90.2	94.72	92.27	95.03	92.23	99.47
Geometric SD factor	1.008	1.008	1.008	1.006	1.013	1
Harmonic mean	90.2	94.72	92.27	95.03	92.22	99.47
Quadratic mean	90.21	94.73	92.28	95.04	92.25	99.47
Skewness	-0.08563	0.08563	0.08563	-2.604	-1.735	-
Kurtosis	6.363	6.363	6.363	11.78	5.634	-
Sum	1624	1705	1661	1711	1660	1790

table, the proposed algorithm of data augmentation has a significant impact of the achieved results and thus recommended. In addition, the adopted softmax classifier is compared with two other classifiers to show its effectiveness. Table 9 presents the comparison results. As shown in the table, the other classifiers included in the experiments are K-NN with K equals to the number emotion categories in the dataset, and SVM with a kernel function of type (radial basis function). The presented results show the effectiveness of the adopted softmax classifier in the proposed approach.

V. CONCLUSION

A new approach for recognizing emotions embedded in a speech signal is proposed in this paper. The proposed approach is based on utilizing deep learning through developing cascaded layers of feature learning blocks with long short-term memory layer. The feature learning blocks are composed of four layers, namely, convolutional, batch normalization, activation, and max pooling. These layers are used to extract high level features from the log Mel-spectrum of the given speech samples. The log-Mel spectrograms are used to extract the local correlations and contextual information of the spoken utterances. To improve the performance of the proposed deep network, two hyperparameters were optimized using the whale optimization algorithm which is guided by the stochastic fractal search method. These hyperparameters are the learning rate and the label smoothing regularization factor. The evaluation of the proposed approach is performed in terms of four speech emotion datasets, namely, IEMOCAP, Emo-DB, RAVDESS, and SAVEE. To train the proposed model using these datasets, a new data augmentation algorithm is proposed to increase the number of training samples and to boost the generalization capability of the model. Experimental results show the effectiveness of the proposed approach in recognizing speech emotions of the adopted four datasets accurately. In addition, a comparison with the other competing approaches is performed to show the superiority of the proposed model. Moreover, a statistical analysis is performed to emphasize the stability of the performance of the proposed approach in recognizing speech emotions.

REFERENCES

[1] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole, and U. R. Acharya, "Automated emotion recognition: Current trends and future perspectives," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106646.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[3] T.-W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.

[4] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[5] A. Dey, S. Chattopadhyay, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar, "A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition," *IEEE Access*, vol. 8, pp. 200953–200970, 2020.

[6] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE Access*, vol. 9, pp. 51231–51241, 2021.

[7] M. Ezz-Eldin, A. A. M. Khalaf, H. F. A. Hamed, and A. I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," *IEEE Access*, vol. 9, pp. 19999–20011, 2021.

[8] 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.

[9] L. Yang, K. Xie, C. Wen, and J.-B. He, "Speech emotion analysis of netizens based on bidirectional LSTM and PGCDBN," *IEEE Access*, vol. 9, pp. 59860–59872, 2021.

[10] S. Zhong, B. Yu, and H. Zhang, "Exploration of an independent training framework for speech emotion recognition," *IEEE Access*, vol. 8, pp. 222533–222543, 2020.

[11] J.-J. Jiang, W.-X. Wei, W.-L. Shao, Y.-F. Liang, and Y.-Y. Qu, "Research on large-scale bi-level particle swarm optimization algorithm," *IEEE Access*, vol. 9, pp. 56364–56375, 2021.

[12] A. Ibrahim, S. Mirjalili, M. El-Said, S. S. M. Ghoneim, M. M. Al-Harthi, T. F. Ibrahim, and E.-S.-M. El-Kenawy, "Wind speed ensemble forecasting based on deep learning using adaptive dynamic optimization algorithm," *IEEE Access*, vol. 9, pp. 125787–125804, 2021.

[13] Q. Al-Tashi, S. J. A. Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary optimization using hybrid grey wolf optimization for feature selection," *IEEE Access*, vol. 7, pp. 39496–39508, 2019.

[14] A. E. Takieldeen, E.-S. M. El-kenawy, M. Hadwan, and R. M. Zaki, "Dipper throated optimization algorithm for unconstrained function and feature selection," *Comput., Mater. Continua*, vol. 72, no. 1, pp. 1465–1481, 2022.

[15] 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, 2008.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, vol. 5, 2005, pp. 1517–1520.

[17] R. S. Livingstone and A. F. 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, pp. 1–35, May 2018.

[18] P. Jackson and S. U. Haq, "Surrey audio-visual expressed emotion (SAVEE) database," Univ. Surrey, Guildford, U.K., Tech. Rep., Apr. 2014.

[19] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, Aug. 2017.

[20] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2746–2750.

[21] 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.

[22] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Process.*, vol. 12, no. 6, pp. 713–721, 2018.

[23] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech*, Sep. 2018.

[24] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," in *Proc. Interspeech*, Sep. 2018.

[25] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Interspeech*, Sep. 2018.

[26] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, 2019.

[27] T.-W. Sun and A.-Y.-A. Wu, "Sparse autoencoder with attention mechanism for speech emotion recognition," in *Proc. IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Mar. 2019, pp. 146–149.

[28] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors*, vol. 19, no. 12, p. 2730, Jun. 2019.

[29] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *Proc. 29th Int. Conf. Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2019, pp. 1–6.

[30] 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.

[31] Z.-T. Liu, P. Xiao, D.-Y. Li, and M. Hao, *Speaker-Independent Speech Emotion Recognition Based on CNN-BLSTM and Multiple SVMs*. Aug. 2019, pp. 481–491.

[32] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 1656–1660.

[33] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75798–75809, 2019.

[34] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, Oct. 2020.

[35] S. Sonawane and N. Kulkarni, "Speech emotion recognition based on MFCC and convolutional neural network," *Int. J. Adv. Sci. Res. Eng. Trends*, Jul. 2020.

[36] Mustaqeem, 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.

[37] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.

[38] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Continuous speech emotion recognition with convolutional neural networks," *J. Audio Eng. Soc.*, vol. 68, nos. 1–2, pp. 14–24, Feb. 2020.

[39] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[40] 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.

[41] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106547.

[42] J. Liu and H. Wang, "A speech emotion recognition framework for better discrimination of confusions," in *Proc. Interspeech*, Aug. 2021, pp. 4483–4487.

[43] L. Li, K. Xie, X. Guo, C. Wen, and J. He, "Emotion recognition from speech with StarGAN and dense-DCNN," *IET Signal Process.*, vol. 16, no. 1, pp. 62–79, Feb. 2022.

[44] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.

[45] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.

[46] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2016, *arXiv:1511.07289*.

[49] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[50] B. E. Boser, E. Sackinger, J. Bromley, Y. LeCun, R. E. Howard, and L. D. Jackel, "An analog neural network processor and its application to high-speed character recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 1, Jul. 1991, pp. 415–420.

[51] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 4, Jul. 2003, pp. 2758–2763.

[52] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," 2013, *arXiv:1304.1018*.

[53] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.

[54] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2546–2554.

[55] J. Bergstra, D. Yamins, and D. Cox, "Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms," in *Proc. 12th Python Sci. Conf.*, Jan. 2013, pp. 13–19.

[56] 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, 2008.

[57] E.-S. M. El-kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid, and S. E. Hussein, "Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images," *IEEE Access*, vol. 8, pp. 179317–179335, 2020.

[58] E.-S. M. El-Kenawy, M. M. Eid, M. Saber, and A. Ibrahim, "MbGWO-SFS: Modified binary grey wolf optimizer based on stochastic fractal search for feature selection," *IEEE Access*, vol. 8, pp. 107635–107649, 2020.

[59] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[60] K. Aho, D. Derryberry, and T. Peterson, "Model selection for ecologists: The worldviews of AIC and BIC," *Ecology*, vol. 95, no. 3, pp. 631–636, 2014.

[61] J. Loughrey and P. Cunningham, "Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search," Dept. Comput. Sci., Trinity College Dublin, Dublin, Ireland, Tech. Rep., Jan. 2005.

[62] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, 1998.

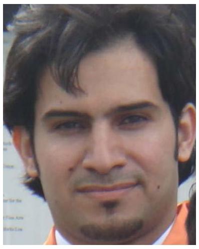
[63] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, 1995.

[64] C. Sun, Y. Yang, C. Wen, K. Xie, and F. Wen, "Voiceprint identification for limited dataset using the deep migration hybrid model based on transfer learning," *Sensors*, vol. 18, no. 7, p. 2399, Jul. 2018.

[65] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, Aug. 2017.

ABDELAZIZ A. ABDELHAMID received the M.Sc. degree in computer science from the Faculty of Computer and Information Sciences, Ain Shams University, and the Ph.D. degree in computer engineering from the Faculty of Engineering, The University of Auckland, New Zealand. He is currently an Assistant Professor with the Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University. He is currently working as an Assistant Professor with the Department of Computer Science, College of Computing and Information Technology, Shaqra University. His research interests include speech and image processing, and machine learning-based intelligent systems.

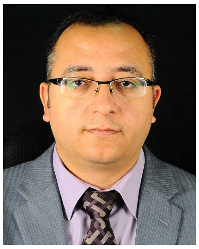
El-SAYED M. El-KENAWY (Senior Member, IEEE) is currently an Assistant Professor at the Delta Higher Institute for Engineering and Technology (DHIET), Mansoura, Egypt. He has published more than 35 papers with more than 1200 citations and an H-index of 22. He has launched and pioneered independent research programs. He motivates and inspires his students in different ways by providing a thorough understanding of various computer concepts. He explains complex concepts in an easy-to-understand manner. His research interests include artificial intelligence, machine learning, optimization, deep learning, digital marketing, and data science. He is a Reviewer of *Computers, Materials and Continua* journal and IEEE ACCESS.



BANDAR ALOTAIBI (Member, IEEE) received the B.Sc. degree (Hons.) in computer science (information security and assurance) from the University of Findlay, Findlay, OH, USA, the M.Sc. degree in information security and assurance from Robert Morris University, Coraopolis, PA, USA, and the Ph.D. degree in computer science and engineering from the University of Bridgeport, Bridgeport, CT, USA. He is currently an Associate Professor with the Department of Information Technology, University of Tabuk. His research interests include computer vision, network security, mobile communications, computer forensics, wireless sensor networks, and quantum computing.

GHADA M. AMER is currently the Vice Dean of Postgraduate Studies and the Research Faculty of Engineering, Benha University; the President of the Centre for Strategic Studies of Science and Technology; and the VP at the Arab Science and Technology Foundation. She holds a few more positions within her profession, like the Director of Innovation and Entrepreneurship Centre at Benha University, the CEO and the Co-Founder of the ASTF Innovation Laboratory, the Ex-Head of the Department of Electrical Engineering, Benha University, and the CEO of the Global Awqaf Research Centre. She believes in the importance of research and development and innovation for the community. She developed and led more than 20 projects and programs to support scientific development and entrepreneurship. She has published about 42 papers in international journals. She raised more than U.S. $2 million to support research, innovation, and entrepreneurship activities to create jobs and support the Arab community. She helped to establish 142 startups on innovative ideas from the region. Since 2009, she has been a Volunteer with the Arab Science and Technology Foundation and later joined as the Volunteer Manager of Women Programs. She was elected as a member of the Board of Directors, in 2011, then the VP of the Foundation, since 2012. In January 2014, she was named as one of the "Top 20 Influential Muslim Women Scientists in the World" by Muslim-Science Magazine. She is called the "Personality of the Year" from Muslim Science Magazine, U.K., in 2015. She also ranked first place for the 50 most prominent leaders in entrepreneurship of the Arab woman issued by the Sayidaty Magazine, in 2014. In 2016 and 2017, she was named one of the " 500 Most Influential Muslims," in 2016, in the field of science and technology, by The Royal Islamic Strategic Studies Centre. In 2016, she was selected to be one of the Rolex Enterprise Awards for Innovation Jury members. She is an Active Advocate of Socio-Economic Development Based on RDI within her country and the region.

MAHMOUD Y. ABDELKADER received the bachelor's degree in scientific computing from the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt. He is currently pursuing the Diploma degree with the EPITA School of Engineering and Computer Science, Paris, France. He is also working as an AI-Pro Collaborator in the role of machine learning engineer at Information Technology Institute (ITI), Cairo. In this position, he developed many machine learning-based intelligent systems. His research interests include computer vision, speech processing, 3D visualization, deep learning, and data science.



ABDELHAMEED IBRAHIM (Member, IEEE) received the bachelor's and master's degrees in engineering from the Department of Computer Engineering and Systems, in 2001 and 2005, respectively, and the Ph.D. degree in engineering from the Faculty of Engineering, Chiba University, Japan, in 2011. He was with the Faculty of Engineering, Mansoura University, Egypt, from 2001 to 2007, where he is currently an Associate Professor of computer engineering. He has published more than 50 publications with over 1500 citations and an H-index of 22. His research interests include machine learning, optimization, swarm intelligence, and pattern recognition. He serves as a Reviewer for the *Journal of Electronic Imaging*, IEEE Access, *Computer Standards and Interfaces*, *Optical Engineering*, IEEE Journal of Biomedical and Health Informatics|, *Biomedical Signal Processing and Control*, *IET Image Processing*, *Multimedia Tools and Applications*, *Frontiers of Information Technology and Electronic Engineering*, *Journal of Healthcare Engineering*, *Sensors*, *Applied Sciences*, *Entropy*, and *Healthcare* journals.

MARWA METWALLY EID received the Ph.D. degree in electronics and communications engineering from the Faculty of Engineering, Mansoura University, Egypt, in 2015. She worked as an Assistant Professor at the Delta Higher Institute for Engineering and Technology, from 2011 to 2021. She has been an Assistant Professor at the Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt, since 2022. Her current research interests include image processing, encryption, wireless communication systems, and field-programmable gate array (FPGA) applications.