IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2021, accepted January 19, 2021, date of publication January 25, 2021, date of current version February 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054345

Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition

MAI EZZ-ELDIN[1,2], ASHRAF A. M. KHALAF[2], HESHAM F. A. HAMED[2,3], AND AZIZA I. HUSSEIN[4]

[1]Department of Electrical Engineering, Future High Institute of Engineering, Faiyum 63514, Egypt
[2]Faculty of Engineering, Minia University, Minia 61111, Egypt
[3]Faculty of Engineering, Egyptian-Russian University, Cairo 11829, Egypt
[4]Department of Electrical and Computer Engineering, Effat University, Jeddah 8482, Saudi Arabia

Corresponding author: Mai Ezz-Eldin (mai.ezzeldin.89@gmail.com)

ABSTRACT Robust automatic speech emotional-speech recognition architectures based on hybrid convolutional neural networks (CNN) and feedforward deep neural networks are proposed and named in this paper as: BFN, CNA, and HBN. BFN is a combination between bag-of-Audio-word (BoAW) and feedforward deep neural network, CNA based on CNN, finally, HBN is hybrid architecture between BFN and CNA. Overall accuracy is achieved by leveraging Mel-frequency cepstral coefficient features and bag-of-acoustic-words to feed the network, resulting in promising classification performance. In addition, the concatenated output from the proposed hybrid networks is fed into a softmax layer to produce a probability distribution over categorical classifications for speech recognition. The three proposed models are trained on eight emotional classes from the Ryerson Audio-Visual Database of Emotional Speech and Song audio (RAVDESS) dataset. Our proposed models achieved overall precision between 81.5% and 85.5% and overall accuracy between 80.6% and 84.5%, hence outperforming state-of-the-art models using the same dataset.

INDEX TERMS Bag-of-acoustic-words, convolutional neural network, feedforward deep neural network, hybrid features, Mel frequency cepstral coefficients, support vector machine.

I. INTRODUCTION

Accurate emotional recognition from speech and song files remain a challenging issue. In pattern recognition and artificial intelligence, recognizing an object or emotion from its characteristic attributes is an especially challenging task in fact.

Deep learning (DL) has shown substantial promise in many applications such as social network analysis [1], encryption and decryption [2], forensics [3] and automotive work [4]. Furthermore, studies such as [5] investigates the exponential stability analysis of Markovian neural networks (MNNs) that can be used to improve many engineering fields, such as communication systems, power systems, production systems, and network control systems. In addition, DL has led to significant advances in recognition research, including speech recognition, object recognition [6], and text recognition [7], [8]. Moreover, deep neural networks have been used as acoustic models because of their ability to learn high-level representations from raw features and classify data effectively [9]–[11]. Furthermore, emotional contents of a patient's speech are usually used medically as a diagnostic tool for various disorders [3].

However researchers have faced some limitations in speech emotion recognition [12], including the following. First, feature analysis has been studied much less in emotion recognition than in speech recognition with the consequence that there is no agreement among researchers regarding which features are best for feature extraction. In addition, the same mistakes have been repeated in recording for different emotional speech databases because of a lack of coordination among researchers and lack of benchmarking databases that can be shared among researchers.

Consequently, our paper focuses on three categories of emotional speech recognition. First, we examine feedforward neural networks containing sequences of two blocks, MFCC

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

and BoAW. Second, we examine DL methods using CNN with vectors produced using MFCC. Third, we investigate our proposed hybrid networks architectures.

The contributions of this paper span multiple dimensions. First, three new speech emotion recognition architectures are introduced based on feedforward networks with BoAW, CNN, and hybrid networks. Second, the performance of the proposed architectures is compared with those of several shallow models consisting of BoAW followed by one of various classifiers such as support vector machines (SVMs), k-nearest neighbor (KNN) and extreme gradient boosting (XGBoost), and also to the state-of-the-art. Third, all of the previously mentioned models are trained and evaluated using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and MFCC feature extraction.

The remainder of this paper is organized as follows. Section II explains prior work related to speech emotion recognition. The details of the proposed systems with the related background, including a description of the emotion recognition methodology, are discussed in Section III. Section V is devoted to the systems' implementation and performance evaluation, with comments on the metrics and the results obtained. Finally, conclusions are drawn and future work is discussed in Section VI.

II. PRIOR RESEARCH

In this section, the state-of-the-art for machine learning and DL techniques on speech emotion recognition are described. Recent research studies have shown that increasing the number of training classes has detrimental effects on the results because the speech features extracted from emotional classes, such as the calm and neutral classes, are closely related, thus slowing the learning performance of DL models [13]. Consequently, our goal in this section is to classify the recent research studies based on the number of training classes, two, four, five, six, seven, and eight.

A. CLASSIFICATION BASED ON TWO CLASSES

Classification based on two classes is discussed in [14] and [15]. In [14], CNN models are trained on the RAVDESS dataset using two different classes; Sad and Happy. The maximum recognition accuracy in [14] is 66.41%. The authors in [15] also have used RAVDESS dataset to train bi-directional long short-term memory (BLSTM) models based on the same two classes, Happy achieving up to 70.4% unweighted accuracy.

B. CLASSIFICATION BASED ON FOUR CLASSES

Most of the following research studies depend on different datasets to classify a four speech emotions set (FES): angry, happy, sad, and neutral. The authors in [16] used Berlin (286 speech samples) and Hindi (100 speech samples) datasets to classify the mentioned speech emotions. KNN usually employed for classification, providing 90% in the angry class in both datasets, while the minimum results achieved were 70-80% in the neutral class for the Hindi and Berlin datasets respectively. FES is also used in [17] for training models with data selected from the University of Michigan Song and Speech Emotion Dataset (UMSSED) and RAVDESS [18] datasets. Simple, simple task (ST), multi-task feature selection/learning (MTFS/MTFL), and group MTFS/MTFL (GMTFS/GMTFL) models were used for classification. The best accuracy based on a four-class emotion classification is 57.14%. In [19], FES is extracted from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset to train a CNN architecture using down sampling of input features maps in convolutional layers instead of a pooling layer. The recognition accuracy achieved 81.75% for the proposed Deep Stride CNN (DSCNN) architecture. However, the fear emotional speech class was used instead of the neutral emotional class in FES [20]. The emotional classes were created from the popular animation film, Finding Nemo, and the emotion classification using a radial basis function (RBF) kernel performed with an accuracy of up to 77.5%.

C. CLASSIFICATION BASED ON FIVE CLASSES

The authors in [21] have employed five different datasets (EMOVO [22], Surrey Audio-Visual Expressed Emotion (SAVEE) [23], Berlin emotional speech (EMO) [24], MOVIES [25] and Kids) with five common emotional classes (Happiness, Sadness, Anger, Fear and Neutral). The recognition accuracy is estimated to be between 43% and 83% based on Speeded-Up Robust Features (SURF) and BoW with an SVM classifier. In [20], the five classes were used based on two different datasets for training. Happiness, Sadness, Anger, and Neutral are the four common classes in both datasets, while the fifth emotional class is Fear for their own dataset and Surprise for the DES dataset [26]. A linear kernel, an RBF kernal and an SVM RBF kernel are used for emotional classification. The maximum overall accuracy values are 66.8% using five classes and 77.5% using four classes based on the *EFN* dataset and 67.6% using five classes based on the DES dataset.

D. CLASSIFICATION BASED ON SIX CLASSES

Happiness, sadness, anger, fear, neutral, and disgust were identified as various emotional classes using the Berlin dataset [27] in [28] and [29]. Different classifiers, such as SVM, NB, and KNN, are used with MFCC features [28], while KNN and Gaussian mixture model classifiers using three features, MFCC, pitch, and energy, are used in [29]. The maximum accuracy was estimated up to 87.7% in both research studies. Happiness, sadness, anger, fear, neutral, and calm are the classes that were chosen for emotional recognition based on the RAVDESS dataset [30]. Two types of vocal communication are presented; speech and song. In addition, three shared emotion recognition models for speech and song are introduced, a simple model (single classifier for recognition if two domains), a single-task hierarchical model (domain classification, then emotion classification), and a multi-task hierarchical model (domain classification,

then independent emotion classifiers). The highest accuracy is 53.8% using MFCC features. In contrast, the authors in [31] depended on the CASIA Chinese speech emotion recognition dataset to extract six emotional classes for training, including happiness, sadness, anger, fear, neutral, and surprise. Deep Belief Networks (DBN) was introduced as a new classifier and its performance is compared to those of the shallow classifiers, Back Propagation (BP) and SVM. The average recognition rates achieved were 92.5% and 90% for DBN and BP, respectively. In addition, in [32], restricted Boltzmann machines (RBMs) and DBN were used together with audio files from one female Spanish speaker from the emotional speech dataset [33] as part of large project, INTERFACE, with the big six classes, joy, sadness, anger, fear, disgust and surprise along with neutral. Those authors used two kinds of features for extraction, MFCC and prosodic features with RBM and DBN, providing a maximum classification error rate of 40.82%

E. CLASSIFICATION BASED ON SEVEN CLASSES

Average scores in [20] and [34] were evaluated based on different emotional classes; happiness, sadness, anger, fear, neutral, boredom and disgust from the Berlin dataset. The SVM classifier was used with different features (MFCC, total energy, $F0$) [20]. The best overall accuracy is up to 63.5%, while in [34], the training is conducted using two different approaches. The first approach depends on training the CNN model from scratch and evaluating the prediction performance on test audio files, while in the second approach, a transfer learning model is explored to utilize the learning from the pre-trained model by initializing the weights of the CNN model before training, thereby achieving better performance than with the first approach. The maximum overall accuracy is greater than 84.3%. Moreover, M. Khan *et al.* [35] develop their own dataset (350 samples) in English containing seven emotional classes; happiness, sadness, anger, fear, neutral, surprise and disgust. Two classifiers KNN and SVM were used with an average accuracy of 91.71% and 76.57%, respectively.

F. CLASSIFICATION BASED ON EIGHT CLASSES

The authors in [15], [19], [36], [37] selected happiness, sadness, anger, fear, neutral, surprise, calm and disgust from the RAVDESS dataset as the different categories of emotional speech. In [19], a softmax classifier was used for the classification of emotions in speech. In addition, a DSCNN model is trained on two different types of spectrograms; raw and clean. The overall accuracy is up to 79.5%. Zeng *et al.* [36] presented a multi-task model using various deep neural networks, including gated residual networks as a classification technique. The model achieved an overall accuracy of approximately 64.48%. Anjali *et al.* [37] depended on integration of MFCC, spectral centroids, and MFCC derivatives of spectral features with a bagged ensemble algorithm of SVM for recognizing speech emotion with an overall accuracy achieved of 75.69%. The authors in [15] recognized emotional speech data by using a hybrid architecture consisting of BLSTM, CNN, and Capsule networks which together classify the extracted representations. The overall accuracy was 69.4%. On the other hand, [38] relied on only male speech signals from the RAVDESS dataset with multiple features that selected based on a continuous wavelet transform and prosodic coefficients using a non-linear SVM classifier. The maximum accuracy equals is 60.05%.

In emotional recognition, there are two separate hyper-classes, which are high arousal and low arousal. High arousal contains anger, happiness, and anxiety/fear, while low arousal containing neutral, boredom, disgust and sadness. The two hyper classes have common properties, such as happiness/anger and neutral/sadness, sharing similar acoustic properties in a speaker. Generally, the most important issues that relate to prosody are pitch, intensity contour, and timing of utterances. The crucial aspects of angry and happy speech are characterized by energy values with wider ranges, longer utterance duration, higher pitch, and shorter inter-word silence. These aspects show the characteristics of exaggerated or hyper-articulated speech. The classification of disgust as low arousal can be challenged, but according to the literature, disgust is a low arousal emotion.

III. PROPOSED RECOGNITION SYSTEM

After the acoustic signal is received and MFCC is extracted, speech recognition models are used to detect emotional classes. New speech recognition architectures based on DL algorithms are introduced in this paper. The first architecture BFN presented in subsection III-A depends on using Bag-of-Acoustic Words (BoAW) and feedforward neural network (FFN) to obtain the emotional class. The second architecture is CNA and the third is HBN. These are introduced in subsections III-B and III-C, respectively. The CNA architecture uses MFCC feature extraction with CNN to extract the emotional classes. Finally, the HBN architecture combines the BFN and CNA architectures and then concatenates them in a fully connected layer to classify the output vector to obtain the emotional classes.

A. FIRST PROPOSED ARCHITECTURE (BFN)

The acoustic signal is fed to a BFN, which consists of an MFCC feature extractor, BoAW, and FFN to extract the emotional class from the input acoustic signal as shown in Figure 1.

1) MFCC FEATURE EXTRACTOR

MFCC is considered to be one of the acoustic low-level descriptors (LLDs) extracted from audio signals, because the audio signals do not follow a linear scale [39]. MFCC was used to represent a two dimensional short-term power spectrum of sound. The physical frequency scale $f(Hz)$ listened by human ears is implemented by a mel scale f_{mel} which simulates the frequency perceived by the human auditory

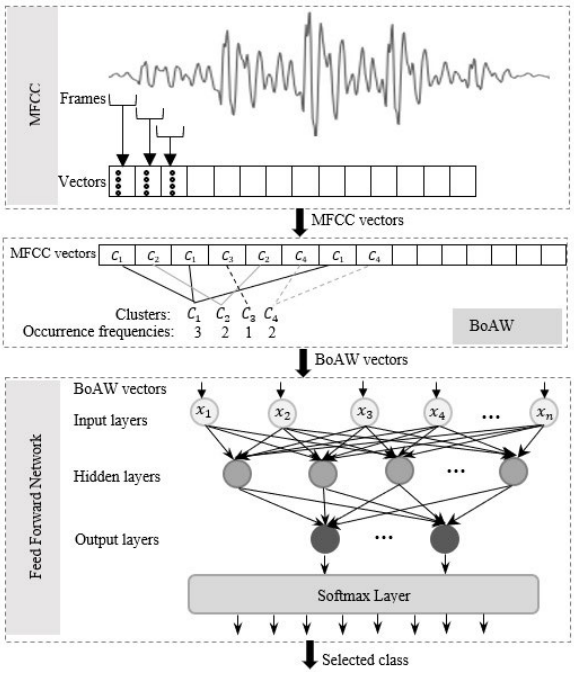

FIGURE 1. First proposed architecture (BFN).

system. The mel scale function was represented as (1) [40].

$$f_{mel} = 2595\, log_{10}(1 + \frac{f}{700}) \tag{1}$$

where f is the physical frequency in Hz, and f_{mel} is the perceived mel scale frequency. Then, the speech signal is split into multiple intervals (windows) and a short time Fourier transform (STFT) is applied to each interval to generate the input power spectrum $P(\omega)$ which is given by (2),

$$P(\omega) = |STFT(w(n) * s(n))|^2 \tag{2}$$

where $s(n)$ is the input speech, and $w(n)$ is the weighting window. f_j is the mel log frequency that calculates from equation (3) [41]:

$$f_j = log_{10}(\sum_{k=0}^{N-1} |x(k)|H_j(k)) \tag{3}$$

where $j = 1, 2, \ldots\ldots\ldots, M$, ($M$ is the number of triangle filters). $H_j(k)$ is the value of the j-triangle filter for the acoustic frequency of k.

Finally, mel frequency coefficients are obtained by applying a discrete cosine transform (DCT) on the list of mel log frequency sub-bands to generate the spectrum (4),

$$c_i = \sum_{j=1}^{f_{sb}} f_j cos(\frac{\pi i}{f_{sb}}(j - 0.5)) \quad 0 <= i <= n_{mfc}, \tag{4}$$

where f_{sb} is the frequency subbands, and n_{mfc} number of mel frequency coefficients. MFCC are the amplitudes of the resulting spectrum. Generally, researchers take 12-13 mel frequency coefficients into consideration as features when training models.

2) BAG-OF-ACOUSTIC-WORD

BoAW is one of the most popular representation methods for emotional speech recognition [19], [21], [36], [42]. The input audio signals have various numbers of extracted MFCC vectors based on the length of the audio signal. However, the classifiers required a fixed-length vector to represent the input audio signal. Consequently, BoAW is used to resolve this issue by implementing a fixed length vector from variable length audio signal through the use of a clustering algorithm. BoAW clusters divide all the input MFCC vectors based on kmeans++ clustering [43] to generate a codebook (dictionary). The codebook size represents the number of audio words. During training, MFCC vectors of the training set were extracted and used for training kmeans++. After the codebook is generated, the acoustic input signal is tested by extracting MFCC vectors and then quantizing based on Euclidean distance to the closest codebook. Then, acoustic words were aggregated by computing the occurrence frequencies of each cluster as features for constructing a histogram.

3) FEEDFORWARD NETWORK

Each neuron input into FFN structure was connected to each neuron in the next layer. Feedforward networks require dealing with fixed-size input which is not deal with sequential data of variable length. A feedforward network consists of three types of layers in which each layer computes a vector. The input layer has a number of nodes equal to the BoAW vectors' dimensionality. The output layer contained eight nodes in the softmax layer [9] for classification. Hidden layers are constructed from one or more layers and represented with nonlinear functions. The DL models were trained using the ADAptive Moment estimation (Adam) optimizer [44] with a learning rate of 0.001 to update the models' parameters during the backpropagation process. Adam is an adaptive gradient algorithm that adapts the learning rate by dividing it by the root mean square of multiple gradients to enhance learning. Adam is a combination of momentum and RMSprop [45], as shown in (5). The gradient update value in equation 5 is split into two parts. First, the gradient component $\hat{m}$, the exponential moving average of gradients, is shown in (6). The second part was the learning rate component $\hat{v}$, which is calculated by dividing the learning rate α by the square root of v, as shown in (7). Given that L is the loss function, m and v are initialized to *zero*, and the parameter values $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^8$ are used. The input vectors of hidden layers are multiplied by the weights and are used with a nonlinear function to generate the output to the next layer. During training, the input BoAW vectors b_i are propagated from the input layer to the output layer using linear and nonlinear activation function. The input layer uses a linear activation function while a rectified linear unit (ReLU) and softmax nonlinear activation function are used for the hidden and output layers, respectively.

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}} - \epsilon}\hat{m}_t \tag{5}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{6}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{7}$$

where

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial L}{\partial w_t}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left[\frac{\partial L}{\partial w_t}\right]^2$$

$$Linearity : y_k = W_k x_k + b_k \tag{8}$$

$$ReLU : y_{ki} = ReLU(W_k^T b_i) \tag{9}$$

$$Softmax : y_{ki} = \frac{exp(x_{ki})}{\sum_j exp(x_{ki})} \tag{10}$$

$$Crossentropy - loss : L(\hat{y}, y) = \sum_i y_i log(\hat{y}_i) \tag{11}$$

where i is the index ($i = 0, 1, 2 \ldots$), $\hat{y}$ is the predicted value, y_i is the output vector, x_i is the input vector, W_k is a vector containing the weights related to output k, and b_k is the bias vector.

B. SECOND ARCHITECTURE (CNA)

Most audio studies convert the audio to spectrograms (image) to apply CNN in DL models. However, we purpose to use raw MFCC features directly that improves the results, as described below and shown in Figure 2.

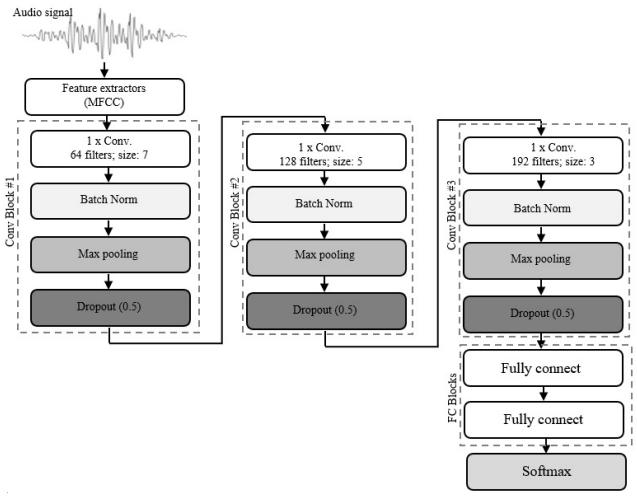

FIGURE 2. Second proposed architecture (CNA).

1) CONV1D LAYER

The input matrix in our experiments is $\in \mathbb{R}^{f \times d}$ where f is the number of frames, and d is the dimension of the MFCC representation. For the CNN layer, the input layer is convolved with N weighting filters each of which has size $S \times S$. These weights are learned the model's during learning process. In our proposed architecture, we choose to have three layers with 64 filters of size 7×7, 128 filters of size 5×5 and 192 filters of size 3×3, respectively. This architecture is inspired by the architecture of various well-known models, such as AlexNet [46] and VGG [47] which have achieved very high accuracy in image classification tasks. Generally, increasing the number of filters gradually assists in capturing more information, which enables us to represent the data in a higher space representation. In addition, bigger filter sizes represent more global, high-level, and representative information while smaller filter sizes collect as much local information as possible.

2) MAXPOOLING1D LAYER

The max pooling layer, is a sample-based discretization process which generally is inserted periodically between successive Conv layers. It aims to calculate the maximum, or largest, value in each batch of each feature map reducing its dimensionality without changing the depth dimension. This is done by applying a max filter to no overlapping sub-regions of the initial input. The principal function of the max pooling layer is to overcome the overfitting problem. Pooling layers with filter size 2×2 are the most common form of max pooling. Generally, it is applied with a stride of two down samples for every depth slice in the input doubly along both width and height.

3) BATCH NORMALIZATION LAYER

Generally, in deep neural network architectures, after updating the weights of each mini-batch, the input distribution in the deep network layers might change. In that case, a problem referred to as the "internal covariate shift" problem [48] arises. This problem occurs because the inputs pass through various adjustments in intermediate layers leading to change the values to be too high or too low while reaching distant layers. The rule of batch normalization is to solve this problem making sure to stabilize the input provided to the later layers to be between zero and one. Normalizing a batch can be performed using (12) where x_i is the i^{th} value in the batch, μ_β is the mean of the batch, and σ_β^2 is the variance of the batch.

$$\hat{x_i} = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2}} \tag{12}$$

4) DROPOUT(0.5) LAYER

This is used to prevent the model from memorizing or overfitting the training data by dropping $n\%$ randomly from the weights between layers in the DL model. The value of the dropout could be any value between 0%, and 100% with, 50% being the optimal value for a wide range of networks and tasks, as described in [49]. Dropout of $n\%$ of weights randomly in each batch assists in preventing overfitting and speeds up the training process. In the testing phase, all the weights are used without any dropout.

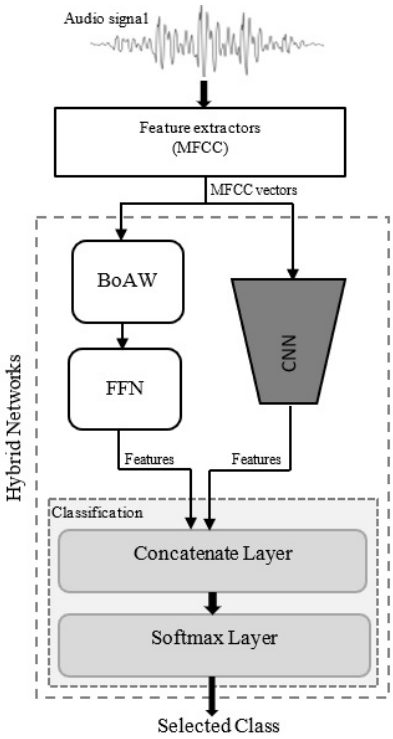

FIGURE 3. Third proposed architecture (HBN).

C. THIRD ARCHITECTURE (HBN)

Combining DL features and handcrafted features shows enhanced results in multiple studies [50]–[52]. The HBN structure is proposed as a combination of the FBN and CNA structures as shown in Figure3 to obtain the benefits of the different features extracted from each structure. Merging the two types of audio features results in a new type of features, called hybrid features, which have stronger discrimination ability than single audio features. These features contain information from the BoW vectors that depends on the histogram of the cluster centers reflecting the frequency of frames. In additions, the hybrid features include features from the DL layers that extract features from the raw MFCC features. Then, the features are combined by a concatenating layer and passed through a fully connected layer to weight each feature and learn new representations for the audio input. Finally, we use a softmax layer to classify the hybrid features into different emotional classes. Our experimental results show that our proposed method outperforms previous emotion classification methods by yielding higher accuracy and precision on the same audio dataset.

D. EVALUATION METRICS

Binary-class classification is considered with only two classes applied, while multi-class classification is assigned for applications with k classes. Multi-class classification includes our case in which the proposed models are trained on eight different emotional classes (k = 8). Moreover, various metrics are used to evaluate the trained model. In our paper, a confusion matrix, accuracy, precision, recall, and Receiver

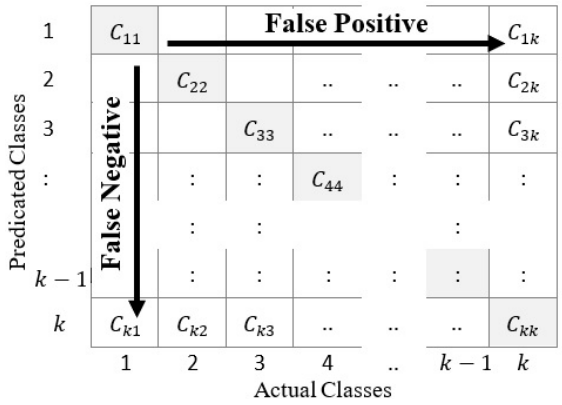

FIGURE 4. Confusion matrix for multi-class classification.

operating characteristics (ROC) are used as evaluation metrics.

1) CONFUSION MATRIX

A confusion matrix provides valuable information for the predicted classes as compared to the actual classes that present the classifier's performance. This matrix contains four categories as follows. True Positives (*TP*) and True Negatives (*TN*) are implemented when the predicated and actual emotional classes are positive and negative, respectively. The prediction does not match the actual emotional classes in two cases: False Positives (*FP*) and False Negatives (*FN*). In FN, the actual class is positive, and the predicted class is negative. In FP, the actual class is negative, and the predicted class is positive. Assume that the confusion matrix is denoted by C^k, where k is the number of class labels. As shown in Figure 4, *TPs* represented on the diagonal of the matrix (grayed cells). *TP*, *FN*, and *FP* for each class are provided by [53]

$$TP_C(i) = C_{ij}|_{i=j} \tag{13}$$

$$FN_C(i) = \sum_{j=1, j\neq i}^{k} C_{ij} \tag{14}$$

$$FP_C(i) = \sum_{i=1, j\neq i}^{k} C_{ij} \tag{15}$$

2) ACCURACY, PRECISION AND RECALL

The Accuracy (*ACC*) measure for the performance of the classifier is defined as the ratio of the correctly classified classes (diagonal) to the total number of predictions, as in (16)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\sum_{i=1}^{k} C_{ii}}{\sum_{i,j=1}^{k} C_{ij}} \tag{16}$$

The accuracy metric does not provide a good reflection of performance for two reasons. First, accuracy is sensitive to imbalanced data. Second, the performance of two classifiers can be completely different although the accuracy is the same in both, depending on the number of correct and incorrect

decisions [54]. Consequently, precision and recall aspects are used for evaluating the classifiers. They are used to evaluate the performance of the model determined from the confusion matrix. They also focus on true positives for one class (i). Precision is a measure of exactness that determines the false positives in the dataset skewing the overall accuracy. Recall is a measure of the goodness of a match, which measures the effectiveness of a classifier to identify positive labels [55]. Precision (P_i) and Recall (R_i) are given by

$$P_i = \frac{TP}{TP + FP} = \frac{C_{ii}}{\sum_{j=1}^{k} C_{ji}} \tag{17}$$

$$R_i = \frac{TP}{TP + FN} = \frac{C_{ii}}{\sum_{j=1}^{k} C_{ij}} \tag{18}$$

Recall equation relays on TP and FN which are located in the same column of the confusion matrix. Therefore, the classification performance with imbalanced data can be evaluated by recall [54].

On Average:

$$P_{avg} = \frac{\sum_{i=1}^{k} P_i}{k} \tag{19}$$

$$R_{avg} = \frac{\sum_{i=1}^{k} R_i}{k} \tag{20}$$

3) RECEIVER OPERATING CHARACTERISTICS (ROC)

The ROC metric is used to evaluate the output quality of a classifier. The ROC curve implements the fraction of correct predictions for the positive class (number of false positives) on the x-axis versus the fraction of errors for the negative class (number of true positives) on the y-axis. Each classifier represents the FP, TP pairs by single points on the ROC. On the ROC curve, the upper left corner point (0, 1) represents a perfect or ideal classifier, because it indicates correct classification for positive and negative samples. The lower left corner point (0, 0) represents correct classification for all negative samples only. In contrast, the upper right corner point (1, 1) indicates a correct classification for all positive classifications only. The line (L_{ROC}) between those two points indicates the performance of the model. Consequently, good classifiers appear in the upper left triangle of the ROC curve above the line L_{ROC}.

Algorithm 1 Proposed Model Using Different Classifiers

Require: $D_{Data} = [D_1, D_2, \ldots .D_N]$. N is number of audio files of database.

Require: $CLASSES = [L1, L2, \ldots Lm]$. L_i is the label of the class and m is the number of classes.

for i *in range*$(1, N)$: **do**
  Split D_i into M frames.
  for j *in range*$(1, M)$ **do**
    decode frame-j into vector v with length 512 using MFCC feature extraction.
  end for
end for
Setting number of cluster equal to C and apply cluster algorithm on all extracted vectors v. Similar v are grouped together into one cluster based on Euclidean distance.
Build BoW using cluster centroids.
for i *in range*$(1, N)$: **do**
  applying the histogram on all vector of D_i to get only one vector with length C.
end for
Array_fold = Split audio data into K fold.
for y *in range*$(1, K)$: **do**
  Training_data = concatenate array_fold $[u]$ where U from $1 : K$ and u != y.
  Testing data = array_Fold $[y]$.
  Input_Training = is a matrix with size = $\Re^{len(tainingdata)*c}$.
  Output_Training = is a matrix with size $\Re^{len(tainingdata)*1}$.
  Input_Testing = is a matrix with size $\Re^{len(testingdata)*c}$
  Output_Testing = is a matrix with size $\Re^{len(testingdata)*1}$
  {# Using one of different classifiers} ▷
  Define model as one of SVM (Eq. 20,21), KNN (Eq. 22), Or XGBoost (Eqs. 23)
  model.train(input_training, output_training)
  Output_predict = model.predict(input_testing)
  using Output_predict and Output_test, Calculate Precision, Recall and f_1 score in Testing_data for each L_i where i *in range*$(1, m)$.
end for
Calculate the average of the metrics over all the K folds.

IV. MODIFIED SHALLOW MODELS

This work is an extension for our previous paper [56] that discussed (SVM, KNN, and XGBoost) with using MFCC feature extraction followed by BoW output vector as the input for each classifier. For more, clarification, these models are renamed in our paper to (MBSVM, MBKNN, and MBXGBoost) respectively to avoid the confusion as reported in table 4. Algorithm 1 illustrates the steps of our modified shallow models.

A. SUPPORT VECTOR MACHINE CLASSIFIERS

One of the most popular binary classification techniques that is used in speech emotion recognition is SVM [37], [38]. It is used to identify patterns and analyze the data for classification and regression analysis. A kernel function is used to transform the original set of features to higher dimensional feature space, which is necessary to obtain optimum classification in this new feature space. The goal of our proposed MBSVM is to find the optimal separating hyperplane that maximizes the margin of the training data which contains eight classes as mentioned previously. We used MFCC features (2D matrix) followed by BoW output vector (1D vector) as the input to the SVM. Here, the output is a vector that represents a probability of each class from the eight classes

in the RAVDESS dataset. The margin can be obtained by the following equation [57]:

$$Margin = \frac{2}{||W||} \tag{21}$$

where the minimization of a norm of a hyperplane normal weight vector is shown as following [57]:

$$||W|| = \sqrt{(W^T W)} \tag{22}$$

B. K-NEAREST NEIGHBOR CLASSIFIER

KNN is a non-parametric method that is frequently used due to its ease of interpretation and low calculation time. KNN can be used for both classification and regression predictive problems. In both cases, the input consists of the k closest training examples in the future space. The output of KNN depends on whether it is used for a regression or classification process. In the proposed KNN model, the input is MFCC feature that followed by the BoW output vector (1D vector). While the output is a vector that represents the distance between the tested label and the first neighbor from the training classes of dataset. The distance between the item and the first nearest neighbor can be calculated as follows [58].

$$d(p, q) = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2 \tag{23}$$

C. EXTREME GRADIENT BOOSTING CLASSIFIER

XGBoost is a variant of the gradient tree boosting proposed by Friedman [59]. Gradient tree boosting is a tree ensemble boosting method that combines a set of weak classifiers to create a strong classifier. The strong learner is trained iteratively starting with a base learner [60]. Both gradient boosting and XGBoost follow the same principal. The key differences between them lie in implementation details. XGBoost achieves better performance by controlling the complexity of the trees using different regularization techniques [26]. An initial model $F0$ is defined to predict the target variable y. While a new model h1 is fit to the residuals from the previous step. Now, $F0$ and $h1$ are combined to give $F1$, the boosted version of $F0$. It can be done for 'm' iterations to improve the performance of $F1$, until residuals have been minimized as much as possible. [60].

$$f_m(x) \ < -f_{m-1}(x) + h_m(x) \tag{24}$$

The proposed XGBoost is used to predict the residuals or errors of prior models and then added together to make the final prediction. The input of this model is the output vector of the BoW that uses a gradient descent algorithm to minimize the loss when adding new models.

On the other hand, BFN uses the same features extracted from MFCC followed by BoW. The features used as input to the feedforward network with three dense layers and followed by softmax layer (classification layer) with eight neurons for predicting the target eight classes. **CNA** as CNN can deal with 2D input features, we use raw MFCC features as input without using Bow. The MFCC used as input to the three stages of CNN. Each stage consists of (1D conv + Batch norm + Max. pooling + dropout 0.5). different filter sizes have been used in the different stages followed by 2 fully connected layers. Finally, we used softmax layer for classification. **HBN:** is the hybrid model that merge between these two features and hence improve the results.

V. EXPERIMENTAL SETUP AND SIMULATION RESULTS

The simulation and synthesis results of emotional speech class extraction for the three architectures based on BoAW, FFN, and CNN are presented in this section. The confusion matrix for multiple classes, accuracy, precision, and ROC for our three architectures and baseline algorithms are determined and demonstrated in V-C1, V-C2, and V-C3, respectively.

A. DATASET USED IN SIMULATION EXPERIMENTS

The proposed models and baseline models are trained over the public dataset RAVDESS [18]. One of the most important reasons of using RAVDESS dataset is that it contains eight classes that increase the classification challenge as mentioned previous. Our experiments are carried out, based on speech and song in the RAVDESS dataset, which contains eight different classes; (neutral, calm, happy, sad, angry, fearful, disgust and surprise) for the speech recordings. The song recordings contain only six emotional classes, which are the same as in the speech recordings but with two classes, disgust and surprised, omitted. The various classes are recorded by 24 professional actors; 12 males and 12 females. In addition, each professional actor records the same sentence with two different emotional intensity; normal and strong. The different classes of speech and song recordings into RAVDESS dataset are listed in Table 1.

TABLE 1. Speech and song recordings details of RAVDESS dataset.

	Dataset Parameters	Speech	Song
	No. files	1440	1012
	No. of professional actors	24	23
Classes	Happy	192	184
	Sad	192	184
	Angry	192	184
	Fear	192	184
	Neutral	96	92
	Calm	192	184
	Surprised	192	-
	disgust	192	-

In our experiments, acoustic recordings of RAVDESS dataset are split into subsets of 80%, and 20% recordings for training, and testing, respectively. To get more robust results and avoid the overfitting problem, the initial training dataset are split into multiple mini train-test splits. Therefore, the training subset from RAVDESS is divided into n subsets (folds) cross-validation paradigm which use iteratively $(n - 1)$ folds for training and the remaining one fold for test set. In our experiments, n equals 5.

B. PROPOSED ARCHITECTURES' PARAMETERS

All of the proposed architectures are implemented in Python. All of our experiments are run on Windows 10 Pro 64-bit

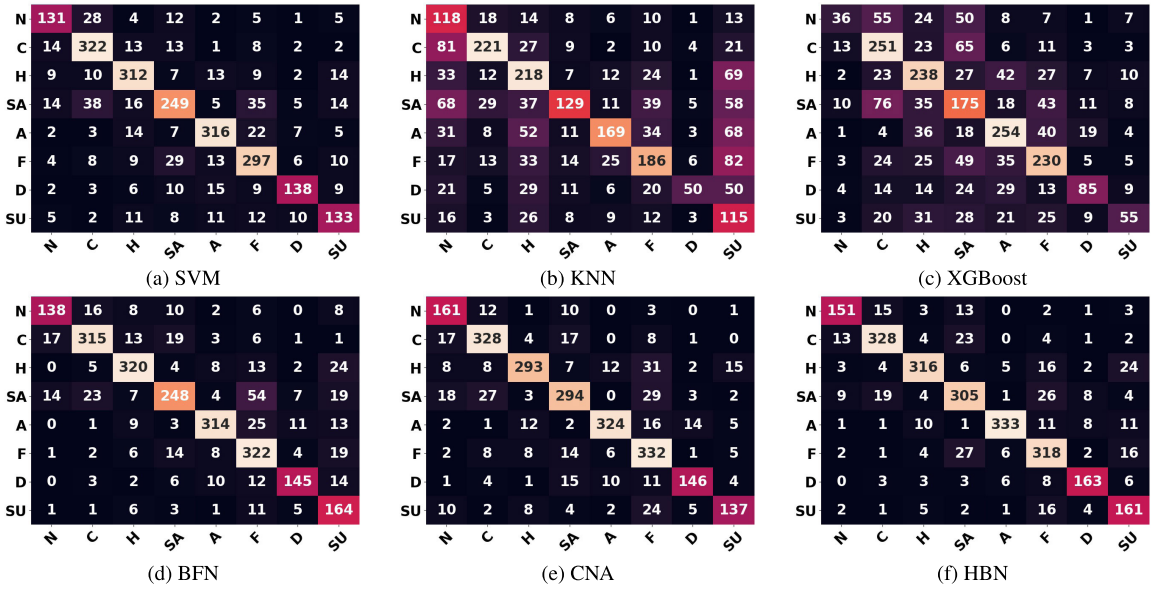

FIGURE 5. Confusion matrix of eight classes for the shallow models, (a) MBSVM, (b) MBKNN, (c) MBXGBoost, and our proposed architectures, (d) BFN, (e) CNA, and (F) Hybrid networks. **Abbreviation:-** N: Neutral, C: Calm, H: Happy, SA: Sad, A: Angry, F: Fearful, D: Disgust and SU: Surprised. (taking the same order in rows from left to right and in columns from top to down).

operating system, HP laptop with an Intel(R) Core(TM) i5-3210M CPU @2.50GHZ process and 4.00 GB RAM.

1) MFCC FEATURES

The length of the analysis window is set to be 25 ms, and the step between successive windows (*winstep*) is equals to 10 ms. As defaults, the number of cepstrum to return (*numcep*) is 13, and the number of filters in the filterbank (*nfilt*) is 26. The FFT size which is a non-equispaced fast Fourier transform (*nfft*) is 512. Finally, the zeros *cepstral* coefficient is replaced with the log of the total frame energy.

2) FEEDFORWARD NETWORK

The feedforward network in the proposed architectures consists of a fully connected layer with 2000 nodes, a dropout layer of 0.5, a fully connected layer with 500 nodes, a batch normalization layer, dropout layer 0.5, fully connected layer with 200 nodes, batch normalization layer, and a dropout layer of 0.5. The dropout layers are used to avoid overfitting and memorizing the training data. Batch normalization layer is used to normalize the input vector within the batch assisting in speeding the training.

The CNN network used in our models contains Conv1D with 64 filters with size $7 * 7$ per each, a batch normalization layer, a dropout layer of 0.5, Then, we use a Conv1D with 128 filters with size $5*5$ per each, a batch normalization layer, and a dropout layer of 0.5. Subsequently, we used Conv1D with 192 filters with size $3*3$ per each, a batch normalization layer, a dropout layer of 0.5, a flattened, fully connected layer with 200 nodes, a batch normalization layer, and a dropout layer of 0.5.

The concatenating layer consists of a fully connected layer with 200 nodes, a batch normalization layer, and a dropout layer of 0.5.

The softmax layer contains eight nodes for our eight emotional classifications.

C. EVALUATION

The multi-class classification problem is evaluated using standard evaluation metrics, such as a confusion matrix, precision, accuracy, and ROC curves.

1) CONFUSION MATRIX OF CLASSIFICATION RESULTS

As shown in Figure 5, the confusion matrix of classification results for the shallow models SVM, KNN, and XGBoost are shown in (a), (b) and (c), respectively, and those for our proposed architectures are demonstrated in (d), (e), and (f), respectively. The total numbers of correctly classified instances are 1898, 1147, 1324, 1966, 2015 and 2075 out of 1452 for SVM, KNN, XGBoost, BFN, CNA, and HBN, respectively. Consequently, our proposed models can predict correct classification better than the shallow models, reflecting the ability of our architectures to extract features to obtain correct predictions.

2) PRECISION AND RECALL

Precisions for the shallow and proposed models are listed in Table 2. The shallow models achieve precisions in the range of 69-85% for MBSVM, 31-71% for MBKNN, 40-62% for MBXGBoost, while the proposed models achieved 63-90% for BFN, 76-89% for CNA, and 66-92% for HBN. On average, the hybrid network models outperform the shallow models by up to 7.5%, 27.5%, and 31.5% for MBSVM,

TABLE 2. Precision (%) of our architectures and shallow models based on eight emotional classes.

	Classifier	Neutral	Calm	Happy	Sad	Angry	Fear	Disgust	Surprised	ALL
shallow architectures	MBSVM	73	78	81	75	85	75	82	69	80.1
	MBKNN	31	71	50	65	71	55	70	24	58
	MBXGBoost	51	54	56	40	62	58	62	54	54
Proposed architectures	BFN	81	86	87	81	90	72	84	63	81.5
	CNA	76	87	87	**83**	89	77	83	**78**	83.6
	HBN	**82**	**91**	**92**	81	**92**	**79**	**89**	66	**85.5**

TABLE 3. Recall of our architectures and shallow models based on eight emotional classes.

	Classifier	Neutral	Calm	Happy	Sad	Angry	Fear	Disgust	Surprised	ALL
shallow architectures	MBSVM	70	86	83	67	84	79	72	69	78
	MBKNN	63	59	58	34	45	49	26	60	49.4
	MBXGBoost	19	67	63	47	68	61	44	29	54
Proposed architectures	BFN	74	84	**85**	66	84	86	76	85	80.7
	CNA	**89**	**89**	80	76	**88**	86	81	73	83.2
	HBN	82	85	82	**81**	**88**	**87**	**84**	**86**	**84.2**

TABLE 4. Table 4 setups to compare between performance of different features/model architectures given the same dataset and number of classes.

Models		Feature	Classifiers	Accuracy		
				Speech	Song	All
Recent research studies		Spectrogram	CNN [19]	79.50	–	–
		Spectrogram	GResNets [36]	–	–	64.48
		MFCC, Spectral centroids, MFCC derivatives	bagged ensamble of SVM [37]	75.69	–	–
		eGeMAPS, Supervector, Log-spectrogram, F0, MFCC, Log-energy	BLSTM [15]	69.40	–	–
		CWT & prosodic	SVM [38]	60.05	–	–
shallow architectures		MFCC, BoW	MBSVM	79.36	88.48	80.10
			MBKNN	44.63	66.16	49.30
			MBXGBoost	50.56	48.58	54.12
Proposed architectures	BFN	MFCC, BoW	FFN	80.54	88.59	80.58
	CNA	MFCC	CNN	80.54	88.59	83.10
	HBN	MFCC, BoW	FFN, CNN	**83.00**	**90.00**	**84.50**

MBKNN, and MBXGBoost, respectively. High precision reflects the characteristic that the extraction of emotional classes by the HBN model detected events mostly correctly. Recall values for the shallow and proposed models are listed in Table 3. The models achieve recall in the ranges 67-86%, 26-63%, 19-68%, 66-86%, 73-89%, and 81-88% for MBSVM, MBKNN, MBXGBoost, BFN, CNA, and HBN respectively. The maximum average recall for all classes is 84.2% for HBN proposed architecture.

3) RECEIVER OPERATING CHARACTERISTIC (ROC)

ROC curves illustrated in Figure 6 were created by plotting the false positive rate (FPR) against the true positive rate (TPR) at various threshold settings using the MBSVM, MBKNN, MBXGBoost, BFN, CNA and HBN classifiers. The TPR, also known as sensitivity, measures the proportion of positives that are correctly identified. Similarly, the FPR, also known as specificity, measures the proportion of negatives that are correctly identified. The performance of each emotion can be measured by the area under the ROC curve, which is an indication of how each emotion is distinctively classified compared to others.

The comparison between the state-of-the-art accuracy (%) and the proposed model's accuracy (%) is illustrated in Table 4. In this table, we show only work using the RAVDESS dataset with eight classes (only speech, only song, or both). From Table 4, we can conclude that, in only speech RAVDESS dataset, the maximum accuracy of the state-of-the-art occurred when [19] used CNN and achieved 79.5 % accuracy with a training time equal to 14 min, while the proposed MBSVM with MFCC and BoW achieved 79.36 % but with a training time equal to 5 min.

On the other hand, when using both speech and song of the RAVDESS dataset, we find that [36] used GResNets and achieved 64.48 % accuracy while our MBSVM proposed model achieved 80.10 % overall accuracy. In addition, from the results in Table 4, we can see that only song has the highest result, because songs files contain hard and clear tones making it easier for the model to obtain information and predict the emotional classes. The proposed HBN model achieved the highest overall accuracy equal to 90% when using only song.

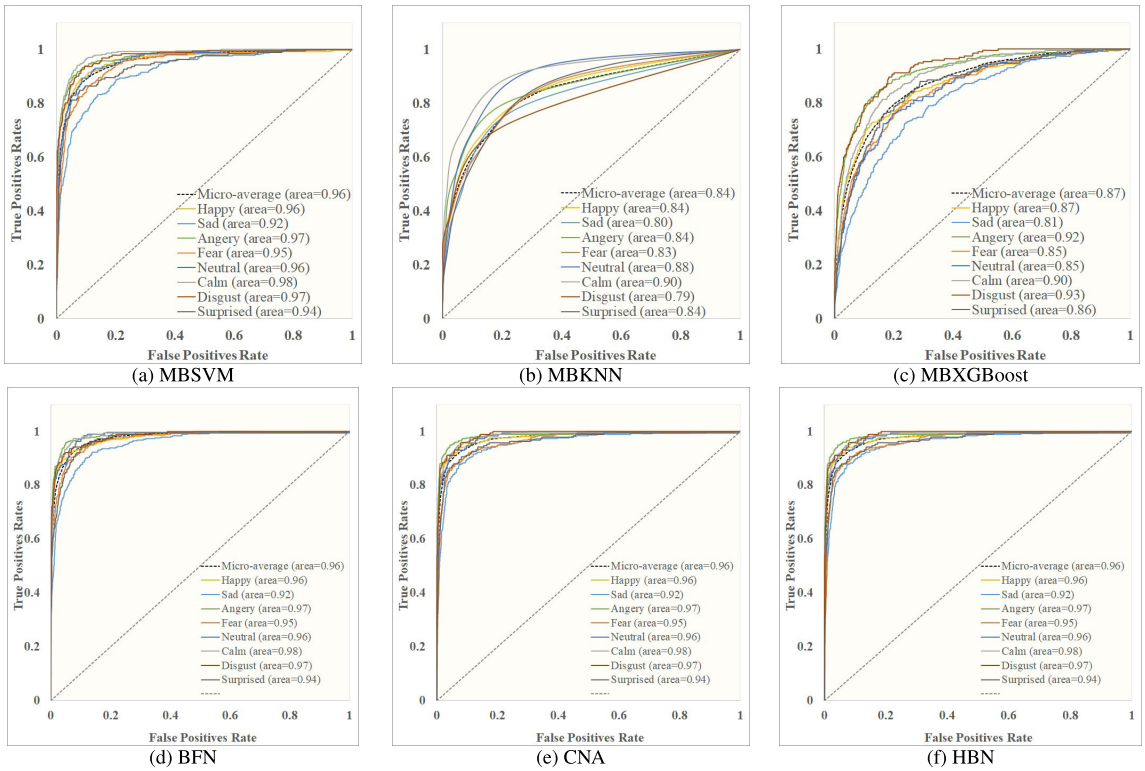

FIGURE 6. ROC results: Shows the receiver operating characteristic to multiple classes using MBSVM, MBKNN, MBXGBoost and our proposed architectures; (d) BFN, (e) CNA, and (F) Hybrid networks.

VI. CONCLUSION AND FUTURE WORK

In human computer interaction, automatic speech emotion recognition has emerged recently as an important research area. Emotion recognition in speech is a challenging problem because it is unclear which features are effective for speech emotion.

Three proposed architectures, BFN, CNA, and HBN, were presented for extracting emotional classes from acoustic signals based on DL techniques. The BFN architecture is built based on BoAW and a feedforward network, while CNA depends on training a CNN to classify the emotional classes. The HBN architecture is built based on a combination of FBN and CNA to enhance the benefits derived from concatenating different extracted features to create hybrid features. The proposed architectures were evaluated based on the RAVDESS audio dataset to classify eight emotions in speech and six emotions in song files. Table 4 illustrates the comparison between the proposed models and the state-of-the-art related published work using the same dataset (RAVDESS) and the eight classes for fair comparison. As mentioned in the dataset section that RAVDESS dataset is split into audio files and song files, so we compare our results with the state-of-the-art results based on using RAVDESS audio files, song files, or both together. All proposed models shown in table 4 depend basically on MFCC features output vector (2D matrix) only or MFCC followed by BoW output vector (1D vector). Each one of proposed model has its own architecture to evaluate the performance of the modes. Any work illustrated in literature review and not included in table 4 can be explained by the fact that they used different number of classes from RAVDESS dataset or used different dataset also with fewer number of classes. With increasing number of predicted classes, the result challenging is increased. All the state-of-the-art, the shallow models and the proposed architectures that have shown previously are based on the same dataset (RAVDESS) for fair comparison as mentioned before. The proposed models achieved significantly better performance in comparison to the shallow modified models; MBSVM, MBKNN, and MBXGBoost. The average precision for the proposed BFN, CNA, and HBN architectures were 81.5%, 83.6%, and 85.5%, respectively. Finally, the overall accuracy of the HBN architecture was up to 84.5%, thereby outperforming the state-of-the-art models. We expect that our future work to include applying our architectures to other datasets with other languages and with other DL algorithms. In addition, we expect that in our future work we will use different datasets and used more features such as prosody, pitch, and energy that will achieve more accurate results.

REFERENCES

[1] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[2] S. F. Yousif, "A new speech cryptosystem using DNA encoding, genetic and RSA algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 4550–4557, 2018.

[3] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000, doi: 10.1109/10.846676.
[4] O. Hosam, "Deep learning-based car seatbelt classifier resilient to weather conditions," *Int. J. Eng. Technol.*, vol. 9, no. 1, pp. 229–237, 2020.
[5] X. Li, F. Li, X. Zhang, C. Yang, and W. Gui, "Exponential stability analysis for delayed semi-Markovian recurrent neural networks: A homogeneous polynomial approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6374–6384, Dec. 2018, doi: 10.1109/TNNLS.2018.2830789.
[6] X. Jiang, A. Hadid, Y. Pang, E. Granger, and X. Feng, *Deep Learning in Object Detection and Recognition*. Singapore: Springer, Jan. 2019, doi: 10.1007/978-981-10-5152-4.
[7] P. Bhatt and I. Patel, "Optical character recognition using deep learning—A technical review," *Nat. J. Syst. Inf. Technol.*, vol. 11, p. 55, Jun. 2018.
[8] R. Vaidya, D. Trivedi, S. Satra, and M. Pimpale, "Handwritten character recognition using deep-learning," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 772–775.
[9] T. Masters, "Supervised feedforward networks," in *Deep Belief Nets in C++ and CUDA C*, vol. 1. Berkeley, CA, USA: Apress, 2018, pp. 9–89, doi: 10.1007/978-1-4842-3591-1.
[10] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.
[12] M. El 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, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.
[13] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *Int. J. Speech Technol.*, vol. 6, no. 3, pp. 311–320, 2003, doi: 10.1023/A:1023426522496.
[14] R. Jannat, I. Tynes, L. L. Lime, J. Adorno, and S. Canavan, "Ubiquitous emotion recognition using audio and video data," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervas. Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 956–959.
[15] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1701–1705, doi: 10.21437/Interspeech.2019-3068.
[16] A. Bombatkar, G. Bhoyar, K. Morjani, S. Gautam, and V. Gupta, "Emotion recognition using speech processing using k-nearest neighbor algorithm," *Int. J. Eng. Res. Appl.*, vol. 4, pp. 2248–9622, Apr. 2014.
[17] B. Zhang, E. M. Provost, and G. Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5805–5809, doi: 10.1109/ICASSP.2016.7472790.
[18] 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.
[19] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/s20010183.
[20] T. Danisman and A. Alpkocak, "Emotion classification of audio signals using ensemble of support vector machines," in *Proc. Int. Tutorial Res. Workshop Perception Interact. Technol. Speech-Based Syst.* Berlin, Germany: Springer, 2008, pp. 205–216, doi: 10.1007/978-3-540-69369-7_23.
[21] E. Spyrou, R. Nikopoulou, I. Vernikos, and P. Mylonas, "Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms," *Technologies*, vol. 7, no. 1, p. 20, Feb. 2019, doi: 10.3390/technologies7010020.
[22] G. Costantini, I. Iadarola, Paoloni, and M. Todisco, "EMOVO corpus: An Italian emotional speech database," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, Reykjavik, Iceland, 2014, pp. 3501–3504.
[23] P. Jackson and S. U. Haq. (Apr. 2011). *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/Download.html
[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.
[25] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 65–68, doi: 10.1109/ICASSP.2009.4959521.
[26] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, vol. 4, 1997, pp. 1695–1698.
[27] H. Muthusamy, K. Polat, and S. Yaacob, "Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0120344, doi: 10.1371/journal.pone.0120344.
[28] S. Emerich, E. Lupu, and A. Apatean, "Emotions recognition by speech and facial expressions analysis," in *Proc. 17th Eur. Signal Process. Conf.*, Glasgow, Scotland, Aug. 2009, pp. 1617–1621.
[29] C. Prakash, V. Gaikwad, R. R. Singh, and O. Prakash, "Analysis of emotion recognition system through speech signal using KNN & GMM classifier," *IOSR J. Electron. Commun. Eng.*, vol. 10, no. 2, pp. 55–61, 2015.
[30] B. Zhang, G. Essl, and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Xi'an, China, Sep. 2015, pp. 139–145, doi: 10.1109/ACII.2015.7344563.
[31] B. Chen, Q. Yin, and P. Guo, "A study of deep belief network based Chinese speech emotion recognition," in *Proc. 10th Int. Conf. Comput. Intell. Secur.*, Kunming, China, Nov. 2014, pp. 180–184, doi: 10.1109/CIS.2014.148.
[32] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Deep learning for emotional speech recognition," in *Proc. Mex. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 311–320. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-07491-7_32
[33] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: From speech database to TTS," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, NSW, Australia, Nov./Dec. 1998, pp. 923–926.
[34] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.
[35] M. Khan, T. Goskula, M. Nasiruddin, and R. Quazi, "Comparison between KNN and SVM method for speech emotion recognition," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 2, pp. 607–611, Feb. 2011.
[36] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.
[37] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886, doi: 10.1016/j.knosys.2019.104886.
[38] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *Proc. 10th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Gold Coast, QLD, Australia, Dec. 2016, pp. 1–8, doi: 10.1109/ICSPCS.2016.7843306.
[39] S. K. Modi, *Biometrics in Identity Management: Concepts to Applications*. Boston, MA, USA: Artech House, 2011.
[40] J. R. Deller, J. G. Proakis, and J. H. Hansen, "Signal processing and analysis," in *Discrete-Time Processing of Speech Signals*. Piscataway, NJ, USA: IEEE Press, 2000, chs. 1–22, p. 936. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?bknumber=5266102
[41] D. Anggraeni, W. Sanjaya, M. Solih, and M. Munawwaroh, "The implementation of speech recognition using mel-frequency cepstrum coefficients (MFCC) and support vector machine (SVM) method based on python to control robot arm," in *Proc. Annu. Appl. Sci. Eng. Conf.*, vol. 2, 2018, pp. 1–9.
[42] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 495–499, doi: 10.21437/Interspeech.2016-1124.
[43] Y. Li and H. Wu, "A clustering method based on K-Means algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, Dec. 2012, doi: 10.1016/j.phpro.2012.03.206.
[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.
[45] T. Dozat, "Incorporating Nesterov momentum into adam," in *Proc. ICLR Workshop*, 2016, pp. 1–4.

[46] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, vol. 25, Jan. 2012, pp. 1097–1105, doi: 10.1145/3065386.

[47] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Kuala Lumpur, Malaysia, Nov. 2015, pp. 730–734, doi: 10.1109/ACPR.2015.7486599.

[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 448–456.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[50] T. P. Anh and T.-D. Mai, "Combining deep feature and handcrafted features for material classification," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 219–224.

[51] D. T. Nguyen, T. D. Pham, N. R. Baek, and K. R. Park, "Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors," *Sensors*, vol. 18, no. 3, p. 699, Feb. 2018, doi: 10.3390/s18030699.

[52] 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.

[53] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.* Berlin, Germany: Springer, 2006, pp. 1015–1021.

[54] V. García, R. A. Mollineda, and J. S. Sanchez, "Theoretical analysis of a performance measure for imbalanced data," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 617–620.

[55] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[56] M. Ezz-Eldin, H. Hamed, and A. Khalaf, "Bag-of-words from image to speech a multi-classifier emotions recognition system," *Int. J. Eng. Technol.*, vol. 9, no. 3, pp. 770–778, 2020.

[57] V. Kecman, "Support vector machines—An introduction," in *Support Vector Machines: Theory and Applications*. Berlin, Germany: Springer, 2005, pp. 1–47, doi: 10.1007/10984697_1.

[58] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, pp. 218–218, Jun. 2016.

[59] J. H. Friedman, "1999 Reitz lecture," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[60] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

MAI EZZ-ELDIN received the B.Sc. degree in electrical engineering from the Faculty of Engineering, Misr University for Science and Technology, Egypt, in 2011, and the M.Sc. degree in electronics and communication engineering from Fayoum University, Faiyum, Egypt, in 2015. She is currently pursuing the Ph.D. degree with the Electronics and Communication Engineering Department, Faculty of Engineering, Minia University, Egypt. Since 2015, she has been a Lecturer Assistant with the Department of Electronics and Communication Engineering, Future High Institute of Engineering, Faiyum.

ASHRAF A. M. KHALAF received the B.Sc. and M.Sc. degrees in electrical engineering from Minia University, Egypt, in 1989 and 1994, respectively, the Doctor of Engineering degree in system science and engineering from the Graduate School of Natural Science and Technology, Kanazawa University, Japan, in 22 March 2000, and the Ph.D. degree, in Egypt. He is currently a Professor of DSP and the Head of the Electronics and Communication Engineering Department, Faculty of Engineering, Minia University.

HESHAM F. A. HAMED received the B.Sc. degree in electrical engineering and the M.Sc. and Ph.D. degrees in electronics and communication engineering from Minia University, Minia, Egypt, in 1989, 1993, and 1997, respectively. From 1989 to 1993, he worked as a Teaching Assistant with the Electrical Engineering Department, Minia University. From 1993 to 1995, he was a Visiting Scholar with Cairo University, Cairo, Egypt. From 1995 to 1997, he was a Visiting Scholar with the Texas A&M University, College Station, TX, USA, (with the group of VLSI). From 1997 to 2003, he was an Assistant Professor with the Electrical Engineering Department, Minia University. From 2003 to 2005, he was an Associate Professor with Minia University. From 2005 to 2007, he was a Visiting Researcher with Ohio University, Athens, OH, USA. He was a Professor and the Dean of the Faculty of Engineering, Minia University, till 2019. He is currently a Professor with the Department of Electrical Engineering, Russian University, Cairo. He has published more than 65 articles and one book chapter. His research interests include analog and mixed-mode circuit design, low voltage low power analog circuits, current mode circuits, nano-scale analog and digital integrated circuits design, and FPGA.

AZIZA I. HUSSEIN received the B.Sc. and M.Sc. degrees from Assiut University, Egypt, in 1983 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from Kansas State University, Manhattan, KS, USA, in 2001. In 2004, she joined Effat University, Saudi Arabia, and established the first Electrical and Computer Engineering Program for women in the country and taught related courses. She was the Head of the Department of Electrical and Computer Engineering, Effat University, from 2007 to 2010. She was the Head of Computer and Systems Engineering Department, Faculty of Engineering, Minia University, Egypt, from 2011 to 2016. She is currently the Head of the Department of Electrical and Computer Engineering, Effat University. Her research interests include microelectronics, analog/digital VLSI system design, RF circuit design, high-speed analog-to-digital converters design, and wireless communications.