JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Bhanusree, Y., Kumar, S. S., & Rao, A. K. (2023). Time distributed attention layered convolution neural network with ensemble learning using random forest classifier for speech emotion recognition. *Journal of Information and Communication Technology, 22*(1), 49-76. https://doi.org/10.32890/jict2023.22.1.3

Time-Distributed Attention-Layered Convolution Neural Network with Ensemble Learning using Random Forest Classifier for Speech Emotion Recognition

*[1]Yalamanchili Bhanusree, [2]Samayamantula Srinivas Kumar, & [3]Anne Koteswara Rao
[1]Department of Computer Science Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, India
[2]Department of Electronics and Communications Engineering, Jawaharlal Nehru Technological University Kakinada, India
[3]Department of Computer Science Engineering, Kalasalingam Academy of Research and Education, India

*[1]bhanusree_y@vnrvjiet.in
[2]samay_ssk2@jntucek.ac.in
[3]k.r.anne@klu.ac.in
*Corresponding author

Received: 20/3/2022 Revised: 20/7/2022 Accepted: 5/9/2022 Published: 19/1/2023

ABSTRACT

Speech Emotion Detection (SER) is a field of identifying human emotions from human speech utterances. Human speech utterances are a combination of linguistic and non-linguistic information. Non-linguistic SER provides a generalized solution in human–computer interaction applications as it overcomes the language barrier. Machine learning and deep learning techniques were previously proposed for classifying emotions using handpicked features. To achieve effective and generalized SER, feature extraction can be performed using deep neural networks and ensemble learning for classification.

The proposed model employed a time-distributed attention-layered convolution neural network (TDACNN) for extracting spatiotemporal features at the first stage and a random forest (RF) classifier, which is an ensemble classifier for efficient and generalized classification of emotions, at the second stage. The proposed model was implemented on the RAVDESS and IEMOCAP data corpora and compared with the CNN-SVM and CNN-RF models for SER. The TDACNN-RF model exhibited test classification accuracies of 92.19 percent and 90.27 percent on the RAVDESS and IEMOCAP data corpora, respectively. The experimental results proved that the proposed model is efficient in extracting spatiotemporal features from time-series speech signals and can classify emotions with good accuracy. The class confusion among the emotions was reduced for both data corpora, proving that the model achieved generalization.

Keywords: Speech emotion recognition, ensemble classifiers, random forest, time-distributed layers, spatiotemporal features.

INTRODUCTION

Speech Emotion Recognition (SER) is one of the trending and attention-seeking research areas in the era of interactive voice command devices, robots, driverless cars, etc. (Zehra et al., 2021). The voice commands are converted to text effectively with the existing state-of-the-art technology, but Human-Computer Interaction (HCI) is successful if the human emotions are properly understood and an appropriate response is provided (Gudmalwar et al., 2019). Use of speech beyond facial expressions, body language, and biosignals is more advised as it carries truthful and deeper emotions and also, less equipment is required to capture in real-time applications. SER has applications in the field of security, clinical diagnosis, call centers, psychological/mental health applications, driver assistant systems, e-learning, etc.

SER has the major challenges in the extraction of appropriate features, classifying cross-corpus data, overcoming overfitting of the classifier, and handling spatial and temporal features of time-series nature (Mustaqeem & Kwon, 2020; Zehra et al., 2021; Zvarevashe & Olugbara, 2020b). These challenges can be addressed by properly designing two major stages; suitable feature extraction and selection, and a classifier that can provide a generalized solution to work with a multilingual corpus for different age groups and genders (Chen et al., 2020b).

The first stage of feature extraction and selection can be performed by handpicking features or using neural networks. Previous studies have used prosodic, spectral data, Mel Frequency Cepstral Coefficient (MFCC), Linear prediction cepstral coefficients (LPCC) and formant features (Gudmalwar et al., 2019; Kuchibhotla et al., 2014; Lalitha et al., 2019; Pawar & Kokate, 2021). Training neural networks to extract appropriate features rather than handpicking will solve most of the issues related to spatial-temporal features (Fayek et al., 2017). Extracting high-level features from a data set using deep convolution neural networks has shown good performance in several previous studies (Jiang et al., 2019; Mustaqeem et al., 2020).

When time-distributed layers are added to deep networks, temporal features are identified and extracted (Wei et al., 2020). Further, the attention layers added to this network can enhance the spatial features (Zhao et al., 2019). The Mel spectrogram, which is an image representation of speech, has been used by several researchers with the wide implementation of neural networks for SER applications (Lech et al., 2020). The use of Mel spectrograms in deep learning models for SER has been proven effective by several researchers (Atila & Şengür, 2021; Chen et al., 2018; Issa et al., 2020; Yao et al., 2020)

Speech emotion data, which are time-series information, contain important temporal features in every speech frame. The spectrogram, which is considered as the input to the proposed model, is split into five frames through a Hamming window and is handled to capture the features. Temporal knowledge from each frame can be handled by the time distribution of wrappers with equal weights (Lieskovská et al., 2021). The time-distributed structure learns the short and long-term features of time-series data. Attention layers added to the Convolution Neural Networks (CNNs) apply heavy weights to specific regions of the spectrogram to extract useful spatial features. Attention mechanisms were widely and effectively used in SER (Li et al., 2018; Mirsamadi & Barsoum, 2017; Neumann & Vu, 2017).

Decision trees (DT) are one of the most prominent classification methods. DTs create new decision-making structures after each iteration and split data into two subsets (Kumar et al., 2021). Several algorithms have been developed that construct more than one DT and are referred to as ensemble classifiers. A few ensemble classifiers are bootstrap aggregating trees (Lee et al., 2019), rotation forests (Alonso et al., 2006) and random forests (Breiman, 2001).

Random forests (RFs) are bagging ensemble models that consist of multiple DTs built on bootstrap samples. RF classifiers perform bagging of the unpruned DTs. In addition to the bagging process, RF classifiers perform randomized selection of features at every split. RF classifiers predict the emotion class of the test samples using a large set of tree classifiers. Therefore, RF classifiers are more efficient than regular-bagging ensemble classifiers (Zhou et al., 2002). The success of RF lies in the construction and splitting of DTs. To split a node, the best features are selected among the set of 'M' randomly chosen features from the 'N' number of features in the feature vector, where $M < N$.

Bagging methods use deterministic DTs, where the evaluation is based on all features, whereas RF classifiers evaluate a subset of features (Noroozi et al., 2017). RF classifiers work on multiple randomly generated DTs and have the advantages of ensembled methods, bagged trees, and decision trees. RF classifiers can interconnect the classes because the decision is based on majority voting throughout the iterations, which is not the case in non-ensemble models and deep neural networks. The features of this classifier improve the probability of achieving higher recognition rates. RF is used in multiple classification sectors such as SER, biomedical, health monitoring, image processing, and the Internet of Things (Agajanian et al., 2019; Albornoz & Milone, 2015; Huihui Qiao & Wang, 2018; Kong & Yu, 2018; Lucky & Suhartono, 2022; Sun et al., 2020). Related studies in the field of SER using ensemble learning algorithms are detailed in Table 1.

Table 1

Existing Models in the Field of SER

Title	Author	Model	Accuracy
CLSTM: Deep feature-based speech emotion recognition using the hierarchical convLSTM network.	Mustaqeem and Kwon (2020) deep learning, and machine learning are dominant sources to use in order to make a system smarter. Nowadays, the smart speech emotion recognition (SER	Hierarchical ConvLSTM Network	IEMOCAP-75% RAVDESS-80%

(continued)

Title	Author	Model	Accuracy
Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition	Zvarevashe and Olugbara (2020a)	Random decision forest	RAVDESS-99.5%
An ensemble model for multi-level speech emotion recognition	Zheng, Chunjun Wang, Chunli Jia, and Ning (2020)	Convolution recurrent neural network for feature extraction and ensemble model for classification	IEMOCAP-75%

This study explored the integration of Time-Distributed Attention-Layered CNN (TDACNN) and RF for SER. The major contributions of the proposed TDACNN-RF model are: extracting the desired features for emotion recognition from speech signals instead of handpicking limited features by implementing the TDACNN; RF was used as an ensemble classifier to classify emotions to achieve more generalization and avoid overfitting; and Mel spectrogram was used as the input to TDACNN to better handle the spatiotemporal information of time series speech.

The remainder of the study is organized as follows. The background for the proposed work is given in Section 2. The proposed TDACNN methodology with RF is presented in Section 3. The details of the experimental results and a comparison of the performance of the proposed method with various metrics are discussed in Section 4. Conclusions and future work are provided in Section 5.

Background

SER is a highly active research field with various generalization challenges, and state-of-the-art models are required for efficient emotion classification. Several studies in the field of SER have been conducted using deep learning algorithms and have achieved good accuracy (Wang et al., 2019; Zehra et al., 2021). Recently, with the use of deep neural networks for SER, Mel frequency spectrograms have been widely used as input features. Deep spectrum representations have been extracted from spectrogram images by several researchers

for most speech and audio applications (Amiriparian et al., 2017; Cummins et al., 2017). The use of a deep spectrum instead of handpicked spectral and prosody features in SER has resulted in improved performance in several studies (Mao et al., 2014).

Recent studies have used Deep Neural Networks (DNN) for feature extraction to capture the required features to solve relevant problems. CNNs have been used for extracting features from speech spectrograms in earlier studies (Ren et al., 2017). Further, RNNs and LSTM-RNNs are frequently used to extract deep spectrum features from the spectrum, particularly for SER tasks (Tzirakis et al., 2017). CNN and RNN, with a combination of LSTM models, have been proposed for extracting features from the Mel spectrogram instead of using manually handpicked features (Sainath et al., 2015). These models have proven their efficiency in effectively handling time-series data. Moreover, attention-based CNNs are used to extract salient spatial features from spectrograms for SER (Atila & Şengür, 2021). Time-distributed layers are added to CNNs with time wrappers to support deep neural networks to capture temporal features (Lieskovská et al., 2021). To summarize, there are several studies in the literature on using spectrograms as the input to networks such as CNNs, RNNs, and LSTMs, and their combinations for feature extraction. Time series applications, such as SER, should work with features that have both temporal and spatial features.

The selection of classifiers for SER applications plays a vital role in achieving generalization when classifying emotions in relation to age, gender, and language. Several researchers have used hidden Markov models (HMMs) (Tuncel & Baydogan, 2018), support vector machines (SVMs) (Sun et al., 2019), Gaussian mixture models (GMMs) (Bhavan et al., 2019), and artificial neural networks (ANNs) (Fayek et al., 2017). Recent studies have proposed models for never-seen languages to obtain a solution for generalized SER that can address multiple language problems (Chen et al., 2020b). Ensemble learning is implemented for SER using machine learning methods, such as SVM, RF, and Adaboost (Bhavan et al., 2019). Random decision forests and ensemble methods using hybrid acoustic features by agglutination of prosodic and spectral features have been proposed for more efficient SER (Zvarevashe & Olugbara, 2020a; Zvarevashe & Olugbara, 2020b). Cross-corpus multilingual emotion identification has been developed using an ensemble of SVMs, RFs, and DTs (Zehra et al., 2021). The RF algorithm is a bagging

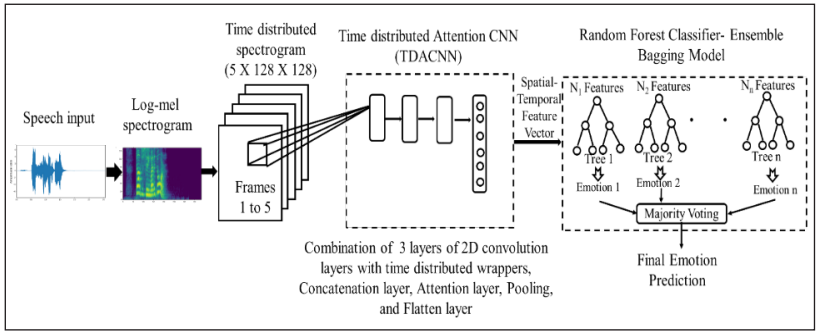
ensemble method based on classification trees that can handle a large number of features and has been proposed by previous researchers (Breiman, 2001). Compared to ANN, SVM, and logistic regression, RF performs well in speech applications (Kondo & Taira, 2018). RF, in combination with a two-layer fuzzy model, has been suggested for emotion classification in human-robot interaction (Chen et al., 2020a). In summary, a combination of deep learning techniques for proper feature extraction and the use of ensemble or hierarchical classifiers for generalized classification is crucial for achieving efficient SER applications.

Proposed Methodology

The proposed model for SER combined a deep learning network for feature extraction and an ensemble learning algorithm for classification. The TDACNN for spatial-temporal feature extraction and the bagging ensemble classifier RF for effective classification of emotions were implemented in this study. The proposed model is illustrated in Figure 1.

Figure 1

Time-distributed Attention Convolution Neural Network with Random Forest (TDACNN-RF)



Preprocessing

Log-Mel spectrograms were used as the input for the proposed model to extract features from the speech signal. A spectrogram is a 3-dimensional time-frequency image representation of the signal.

The time details of the signal were projected on the horizontal axis and the frequencies on the vertical axis. The amplitude or energy was represented by the intensity or color distribution of the spectrum. Spectrograms were produced by the application of the short-time Fourier transform (STFT). The speech samples were framed to a window length of 256 with a hop length of 128. A sample rate of 16 KHZ with the fast Fourier transform set to 512 was chosen, and a Hamming window was applied to avoid spectral leakage. The resulting signal was converted to Mel scale, which is a nonlinear transformation of the frequency scale based on pitch perception. Mel filter banks were applied for conversion to Mel scale, and logarithmic scale was applied to finally produce a log-Mel spectrogram as given in Equation 1. The log-Mel spectrograms were split into five frames with overlapping hops to fetch the model.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \quad (1)$$

where, m is the log Mel frequency and f is the frequency.

Feature Extraction by TDACNN

The time-series information of the log-Mel spectrograms was segmented into five frames using the sliding window and was fed to the subsequent time-distributed layers of the 2D convolution neural layers. The time-distributed wrappers of the 2D convolution layers were assigned the same weights. The features observed in each frame were treated equally by applying the same weights. All frames were provided with similar transformations, and the essential temporal features were extracted from each frame. Three layers of time-distributed 2D convolutional neural layers were stacked, and the individual features were extracted from each frame. These extracted features were concatenated using concatenation layers. The attention layer was stacked on top of the time-distributed layers to extract spatial features (Sun et al., 2020). The proposed TDACNN model is illustrated in Figure 2. The fully connected layer, along with the flattened layer, were used at the end of the network. The required spatiotemporal features for efficient SER were extracted using successive time-distributed and attention layers. The time-distributed layers turned the raw input time-series signal into shorter and easier chunks to learn the long temporal dependencies, and the attention layers focused on extracting the spatial features from the input.

CNNs are efficient in the extraction of spatial features, but they are not completely efficient in extracting temporal features. The time-distributed 2D convolution layers were concatenated to the CNN for efficient time-series compatibility. Moreover, to select salient emotional features from sections of speech input, the attention mechanism was incorporated with time-distributed 2D convolution layers (Mirsamadi & Barsoum, 2017; Chorowski & Bahdanau, 2015). The attention layers generate a weight vector that merges the frame-level features from each time step to an utterance-level feature vector. The attention weights α_i were determined using each vector entry x_i from the TD CNN layers, as expressed in Equation 2.

$$\propto_i = \frac{\exp(f(x_i))}{\sum_j \exp(f(x_j))}, \tag{2}$$

where $f(x)$ denotes the scoring function and is defined in Equation 3.

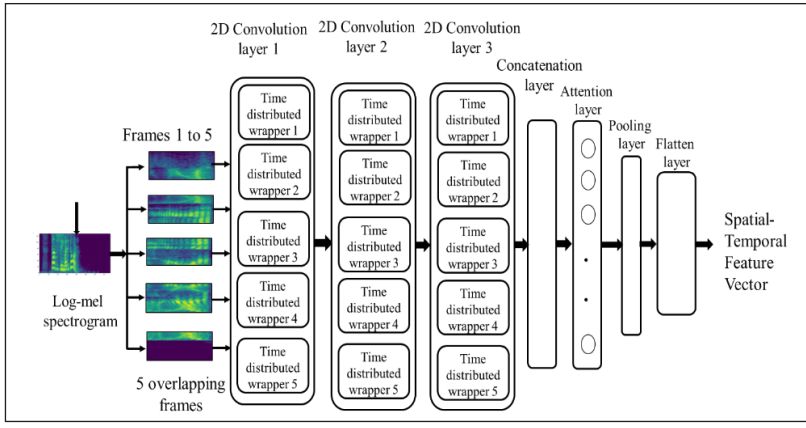
$$f(x) = W^T x, \tag{3}$$

where W denotes the trainable parameters as linear scoring.
The output of the attention layer is the weighted sum of the input sequences, as shown in Equation 4.

$$attn_x = \sum_i \propto_i x_i \tag{4}$$

Figure 2

The Architecture of TDACNN for Feature Extraction and Classification



The model has three layers of time-distributed CNNs with three max pool layers and RELU as the activation function. The model was

trained for feature extraction, and the output obtained from attention pooling is a feature vector *'C'* as given in Equation 5.

$$C = \sum_{i=1}^{L} \propto_i a_i, \tag{5}$$

where a_i denotes the inputs fed to the networks, $\propto_i$ denotes the attention weights, and L is a variable-length grid $L \stackrel{\text{def}}{=} F \times T$, where F and T denote the frequency and time domains of the spectrogram, respectively.

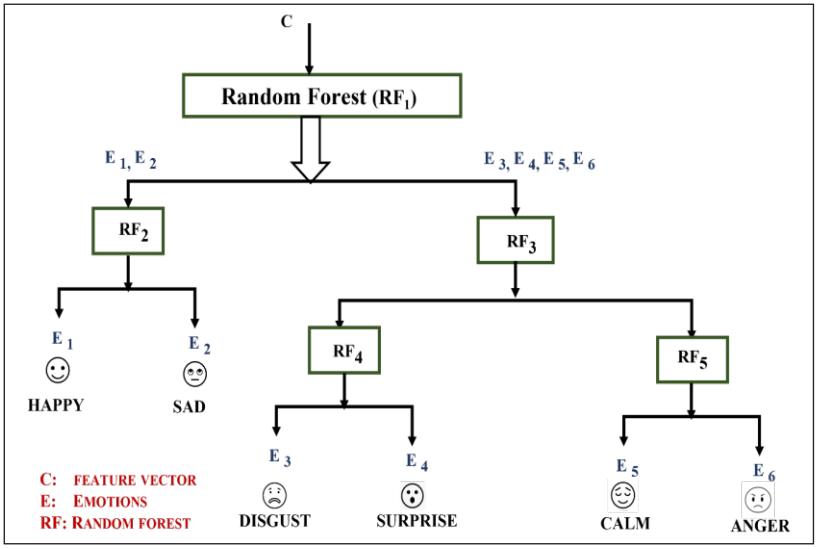
Bagging Ensemble Classifier—Random Forest Algorithm

RF classifiers are ensemble models based on the bagging principle. RF classifiers combine several binary DTs using bootstrap samples from the sample feature vector '*c*'. They randomly select a subset of explanatory variables at every node. RF is a highly efficient algorithm based on the model aggregation concept for classification problems. Generalization is one of the prime features that suit SER applications (Noroozi et al., 2018). As RF utilizes multiple randomly generated DTs, it includes the advantages of ensemble methods, bagging approach, and DTs. RF classifiers predict the class through majority voting based on predictions made through multiple DTs. Unlike bagging, the features are randomly selected at each split, which makes the RF algorithm more efficient and results in less overfitting (Breiman, 2001). The procedure for generating RF classifiers includes the following steps:

1. The original training data for a count of DTs 'k', where 'k' is a random subset of samples considered without replacements from the original data using the bootstrap method. Each subset is used to train the growing tree.
2. From '*N*' features, '*n*' random features ($n << N$) are selected. The nodes of the trees are split on these '*n*' variables using an optimized splitting criterion. The value of '*n*' is unchanged during the growth. This process is repeated until a complete tree is obtained.
3. Each tree grows to its maximum and does not undergo any cutting. Figure 3 shows the framework of RF classifier, and Algorithm 1 presents the pseudocode for the proposed algorithm.

Figure 3

Framework of Random Forest Classifier for Classifying Emotions Using Feature Vector 'C'



Algorithm 1: Random Forest

Algorithm 1: Random Forest

Input: C: *Feature vector*
S: *Speech Signals of data set*
K: *Subset samples for training*
N: *No. of features in feature vector* ***C***
n: *Randomly selected features*
i: *1 to N*
t: *test signal*
Output: Y: *Predicted emotion*

For ***i=1 to k****, do*

Select training feature vector samples of ***S*** *to make* i^{th} *training of* S_i *randomly and by replacement.*

Prepare the root node of S_i *to compare feature values.*

Prepare a decision tree based on S_i *and find the root node.*

Choose a feature vector for i^{th} *decision tree by splitting.*

Select features f_i *with a high probability.*

While testing signal ***t,*** *do Prepare child node of the* i^{th} *decision tree for the feature vector*

(continued)

*For **i=1 to m**,*

*Compare content of the nodes of **i**th decision tree with contents of feature vector*
Build tree to generate complete tree

end for loop

end while loop
Find emotion labels from every decision tree
*Perform majority voting from all the emotion labels to find **Y***

end for loop

The output of the RF classifier is an ensemble of the results provided by all the DTs, and the most accurate prediction is considered for the final output as given in Equation 6:

$$H(x) = argmax \sum_{i=1}^{M} I(h_i(x) = y), \quad (6)$$

where, $H(x)$ denotes the output of the RF ensemble, $h_i(x)$ denotes an indicatory function, denotes a single DT model, and denotes the target emotion.

Bootstrapping of the training subsets is performed randomly using the Gini coefficient as given in Equation 7. The smaller the Gini coefficient, the better the selection characteristics of the RF algorithm.

$$Gini(D) = \sum_{i=1}^{k} P_i(1 - P_i) = 1 - \sum_{i=1}^{k} P_i^2, \quad (7)$$

where, k denotes the subset samples in set D, and P_i denotes the probability of class *i*.

For a two-category node, assuming that the first sample probability is *P,* the Gini coefficient is as given in Equation 8.

$$Gini(D) = 2P(1 - P), \quad (8)$$

IMPLEMENTATION AND EXPERIMENTAL RESULTS

Data Corpus and Setup

The proposed model was evaluated using two data corpora: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2007). The

RAVDESS data corpus consists of emotional speech and songs from 24 actors (12 males and 12 females) with North American English accents. This data set is a class-balanced corpus with seven distinct classes: calm, happy, sad, angry, fearful, surprised, and disgusted. The IEMOCAP data set is a multimodal corpus of data generated by multiple speakers. This data corpus includes dyadic sessions, where improvisations and scripted scenarios in the English language are performed by actors. IEMOCAP is a class-imbalanced dataset with emotions—Neutral, Anger, Happiness, Sadness, Excitement, Disgust, Frustration, Fear, and Surprise. Among these 9 emotion samples, only 4 emotion samples (Neutral, Anger, Happy and Excitement, Sadness) that are balanced are considered here. These annotated emotions overlap each other, which is a challenge in the field of SER. The details of the data corpus are listed in Table 2.

Table 2

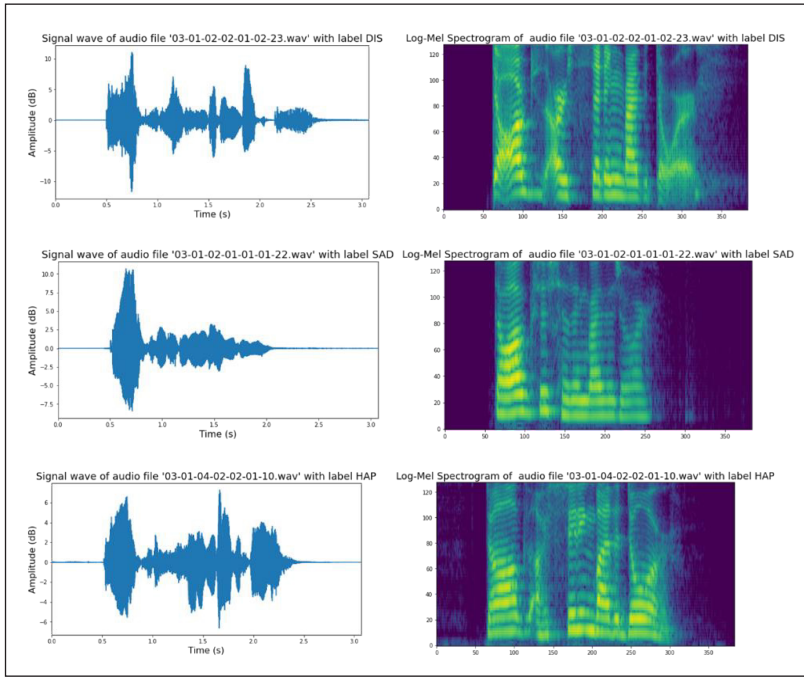
Details of the Data Corpus Used for Verifying the Performance of the Proposed Model

Data Corpus	Speech files considered	Train set	Test set	Emotions considered	Language
IEMOCAP	5531	4424	1107	4	English
RAVDESS	1344	1078	266	7	

The model was tested on an NVIDIA GeForce® GTX 1650 Ti (dedicated 4 GB GDDR6) GPU, Windows 10 machine with an Intel 8 core processor and 16 GB RAM. The process was performed on Keras using TensorFlow as the backend for the DNN and scikit-learn for the RF and SVM algorithms. Training and testing were performed on the data sets using 10-fold cross-validation. Speech samples of both the data corpora were split into training and test sets in the ratio of 80:20. The log-Mel spectrograms were extracted from speech samples by applying SFFT, Mel filter bank, and logarithmic operations. A sample log-Mel spectrogram of speech samples for disgust, sadness, and happiness in the RAVDESS data corpora is shown in Figure 4.

Figure 4

Sample Speech Signal and Corresponding Log-Mel Spectrogram for Disgusted, Sad and Happy Emotions from RAVDESS Data Corpora



To achieve the optimal model, hyperparameters, such as kernel size, pooling size, stride on the convolution layer, number of neurons, type of activation function, and number of convolution layers, were optimized. The log-Mel spectrogram was generated with a 40 Mel band to input into the TDACNN-RF. Three time-distributed 2D convolution layers with a kernel size of and a RELU activation function were used. Batch normalization was performed at each convolution layer to stabilize the learning process. An attention layer with a stride of (1,4) followed by the convolution layers, dropout layer, batch normalization, and RELU activation were used to prevent overfitting in feature extraction. The model was trained with a maximum epoch of 100 and an early stopping criterion. After the flattening layer, the feature vector '*C*' was fed to an RF classifier with labels. The RF classifier consisted of 150 estimators with a bootstrap sample size equal to the training sample sizes of IEMOCAP and RAVDESS.

Performance Measurements and Experiment Results

The performance of the proposed TDACNN-RF model on the two datasets, IEMOCAP and RAVDESS was compared with CNN-SVM and CNN-RF models. Performance metrics, such as classification accuracy, precision, recall, F1-score, and confusion matrix, were used for analyzing automatic classification applications.

Classification accuracy is defined as the percentage of correct predictions over the total predictions as given in Equation 9. Precision is defined as the fraction of relevant predictions over all retrieved predictions, as given in Equation 10. Recall/sensitivity is defined as the fraction of relevant predictions over all relevant predictions, as given in Equation 11. Although difficult to achieve, an ideal model has a precision and recall equal to 1. The F1-score provides a balance between precision and recall, as given in Equation 12. It is the harmonic mean between the precision and recall for every class. The confusion matrix is a consolidated representation of the class predictions.

$$Classification\ Accuracy = \frac{TP + TN}{TN + TP + FP + FN}, \tag{9}$$

$$Precision = \frac{TP}{TP + FP}, \tag{10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{12}$$

where *TP*, *TN*, *FP*, and *FN* denote True Positive, True Negative, False Positive, and False Negative.

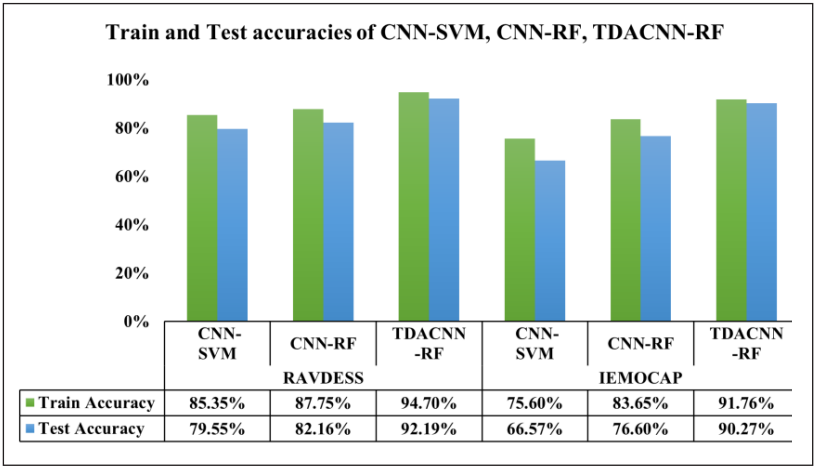
The classification accuracies achieved by the proposed TDACNN-RF, CNN-SVM, and CNN-RF models on both data corpora are shown in Figure 5. To compare the efficacy of the proposed model, CNN-RF and CNN-SVM were implemented with both the data corpora and log-Mel spectrograms as inputs. The CNN-RF and CNN-SVM models were implemented using a three-layer CNN with RELU activation function for feature extraction. Except for the time distribution wrappers and the attention model, the feature extraction process was similarly implemented in all three models. The CNN model was chosen for

performance comparison to verify the efficiency of time-distributed wrappers and attention layers in extracting features from time-series data.

The RF algorithm in the CNN-RF model selected log2(n)+1 features at each node, and the number of estimators was set to 180. The radial basis function (RBF) was used as the kernel function for the SVM algorithm in the CNN-SVM model. SVM was chosen along with CNN, as it is an efficient machine learning algorithm and can be compared with RF classifier performance. A comparison of the training and test accuracies of the three models is presented in Figure 5.

Figure 5

Performance Comparison of the Proposed TDACNN-RF Model with the CNN-SVM and CNN-RF Models



	RAVDESS			IEMOCAP		
	CNN-SVM	CNN-RF	TDACNN-RF	CNN-SVM	CNN-RF	TDACNN-RF
Train Accuracy	85.35%	87.75%	94.70%	75.60%	83.65%	91.76%
Test Accuracy	79.55%	82.16%	92.19%	66.57%	76.60%	90.27%

The test classification accuracies of the CNN-SVM, CNN-RF, and TDACNN-RF models were 79.55 percent, 82.16 percent, and 92.19 percent, respectively, on the RAVDESS data. Similarly, for the IEMOCAP data, the test accuracy of the proposed model was 90.27 percent, which was comparatively higher than the 66.57 percent and 76.6 percent test accuracies of the CNN-SVM and CNN-RF models, respectively. The CNN network used in the latter models for extracting features was not effectively equipped to extract spatial and temporal features when compared to TDACNN. The time-distributed wrappers and attention layers added to the CNN in the proposed

model efficiently extracted the temporal and spatial features of the time-series emotional speech compared to the CNN alone.

The RF classifier in the proposed model can identify indistinguishable emotions compared to SVM. The RF is a bagged ensemble classifier of a group of DTs, and each DT classifier votes to determine the optimal emotion at the end. This feature outperforms a well-performing SVM algorithm for classifying emotions. The class confusion in RF was also reduced, and the generalization property was proven.

The train and test accuracies of the proposed model, as shown in Figure 5, depicts strong evidence that the RF classifier is capable of classifying indistinguishable emotions when compared to SVM. RF being a bagged ensemble classifier, is a group of decision trees where each decision tree classifier votes to determine the optimal emotion. The features of the RF classifier, CNN-RF and TDACNN-RF, have overcome the performance of the SVM algorithm in classifying emotions. The class confusion in RF is also reduced and supports the generalization property.

All three models were trained to classify four emotions in the IEMOCAP data corpus and seven emotions in the RAVDESS data corpus. The confusion matrices for the three models with respect to the RAVDESS data corpus are listed in Tables 3 to 5. The CNN-SVM classifier exhibited good performance in classifying the sadness and surprise emotions of the RAVDESS data corpus. CNN-RF performed well in classifying disgust and surprise emotions. The proposed TDACNN-RF model showed consistency in classifying all six emotions of RAVDESS except happy emotions. The time-distributed wrappers and attention layers, along with the CNN model, can capture the required feature sets for training the RF classifier.

Table 3

Confusion Matrix for CNN-SVM Classifier Model on RAVDESS Data Set

Emotion	Ang (%)	Dis (%)	Fear (%)	Hap (%)	Calm (%)	Sad (%)	Sur (%)
Ang (%)	76.47	11.76	0	5.88	0	0	5.88
Dis (%)	11.9	71.43	0	7.14	0	2.38	7.14

(continued)

Emotion	Ang (%)	Dis (%)	Fear (%)	Hap (%)	Calm (%)	Sad (%)	Sur (%)
Fear (%)	0	0	78.05	7.32	0	7.32	7.32
Hap (%)	2.94	2.94	2.94	76.47	5.88	2.94	5.88
Calm (%)	0	0	0	9.52	76.19	7.14	7.14
Sad (%)	0	2.56	2.56	2.56	2.56	84.62	5.13
Sur (%)	0	2.70	0	0	0	2.70	94.59

* Ang=Angry, Dis= Disgust, Fear= Fearful, Hap=Happy, Sur=Surprise

Table 4

Confusion Matrix for CNN-RF Classifier Model on RAVDESS Data Set

Emotion	Ang (%)	Dis (%)	Fear (%)	Hap (%)	Calm (%)	Sad (%)	Sur (%)
Ang (%)	85.29	8.82	0	2.94	0	0	2.94
Dis (%)	4.76	88.1	0	0	0	7.14	0
Fear (%)	0	0	75.61	14.63	0	7.32	2.44
Hap (%)	2.94	5.88	8.82	67.65	2.94	0	11.76
Calm (%)	0	0	0	4.76	83.33	9.52	2.38
Sad (%)	0	7.69	0	2.56	5.13	84.62	0
Sur (%)	0	2.7	0	5.41	0	2.7	89.19

* Ang=Angry, Dis= Disgust, Fear= Fearful, Hap=Happy, Sur=Surprise

Table 5

Confusion Matrix for TDACNN-RF Classifier Model on RAVDESS Data Set

Emotion	Ang (%)	Dis (%)	Fear (%)	Hap (%)	Calm (%)	Sad (%)	Sur (%)
Ang (%)	94.12	2.94	0	0	0	0	2.94
Dis (%)	2.38	92.86	0	0	0	4.76	0
Fear (%)	0	0	95.12	2.44	0	2.44	0
Hap (%)	0	0	5.88	85.29	2.94	0	5.88
Calm (%)	0	0	0	0	92.86	7.14	0
Sad (%)	0	2.56	2.56	0	2.56	92.31	0
Sur (%)	2.70	0	2.70	0	0	0	94.59

* Ang=Angry, Dis= Disgust, Fear= Fearful, Hap=Happy, Sur=Surprise

The confusion matrices for the three models with respect to the IEMOCAP data corpus are listed in Tables 6–8. The CNN-SVM and

CNN-RF showed poor accuracies in classifying the happy and neutral emotions of the IEMOCAP dataset. The CNN layers used for feature extraction do not efficiently extract these features. The proposed TDACNN-RF model exhibited good performance in classifying all emotions. The TDACNN successfully extracted spatiotemporal features, and the RF classifier efficiently classified the emotions.

The proposed TDACNN-RF model exhibited good performance in classifying the anger and sadness emotions but performed poorly in happy and neutral emotion classification. However, the proposed model performed better than the other two models.

Table 6

Confusion Matrix for CNN-SVM Classifier Model on IEMOCAP Data Set

Emotion	Anger (%)	Happy (%)	Neutral (%)	Sadness (%)
Anger (%)	75.5	17.7	0	6.6
Happy (%)	23.4	55.3	3.19	18
Neutral (%)	9	9	63.63	18.18
Sadness (%)	8.9	9.9	10.89	70.29

Table 7

Confusion Matrix for CNN-RF Classifier Model on IEMOCAP Data Set

Emotion	Anger (%)	Happy (%)	Neutral (%)	Sadness (%)
Anger (%)	86.66	11.11	0	2.22
Happy (%)	9.57	72.34	1.06	17.02
Neutral (%)	0	4.54	63.63	31.81
Sadness (%)	2.97	16.83	2.97	77.22

Table 8

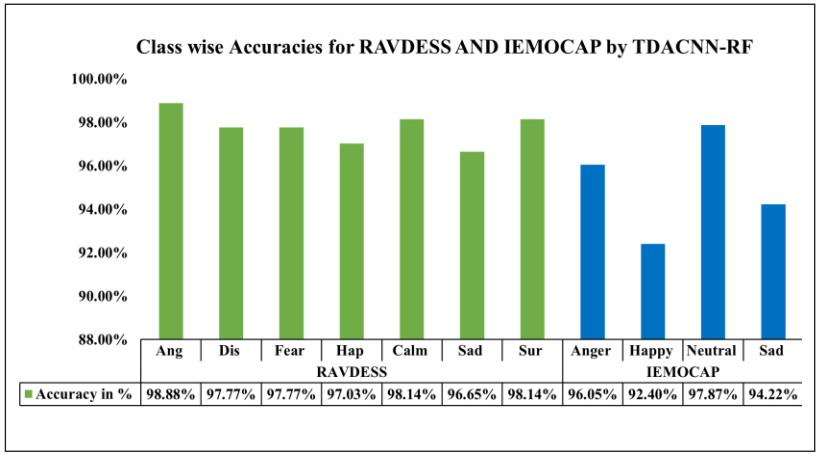
Confusion Matrix for TDACNN-RF Classifier Model on IEMOCAP Data Set

Emotion	Anger (%)	Happy (%)	Neutral (%)	Sadness (%)
Anger (%)	92.22	6.66	0	1.11
Happy (%)	4.2	88.29	0	7.44
Neutral (%)	2.27	4.54	86.36	6.81
Sadness (%)	1	5.9	1	92

The performance of the proposed TDACNN-RF model on each emotion in terms of class-wise accuracies on the RAVDESS and IEMOCAP data corpus is shown in Figure 6. The model achieved good class-wise accuracies for both data corpora. For the IEMOCAP data corpus, the proposed model exhibited the least accuracy (92.4 %) in classifying happiness emotions and higher accuracies (96.05 % and 97.87 %) in classifying anger and neutral emotions, respectively. For the RAVDESS data corpus, the proposed model exhibited the least accuracy (96.65 %) in classifying sad emotions and higher accuracies (98.88 %, 98.14 %, and 98.14 %) in classifying anger, calm, and surprise, respectively.

Figure 6

Class-wise Accuracies of TDACNN-RF on the Seven Emotions of the RAVDESS Data Set and Four Emotions of the IEMOCAP Data Set



The RF classifier ensemble process reduced confusion between emotions, as can be observed from the improvised class-wise accuracies. To further verify the efficacy of the proposed model, precision, recall and F1-scores of all three models for each emotion were compared for both data corpora and are summarized in Table 9. The results obtained by the proposed model are better than those obtained by the CNN-SVM, CNN-RF, and base models shown in Table 1. From Table 9, it is observed that the F1-scores of each class for the TDACNN-RF model show consistent improvement over the other two models.

Table 9

Precision, Recall and F1-score on RAVDESS and IEMOCAP Data Set for the Three Models

Emotion	CNN-SVM			CNN-RF			TDACNN-RF		
	Precision	Recall	F1 -score	Precision	Recall	F1 -score	Precision	Recall	F1 -score
				RAVDESS					
Angry	0.76	0.81	0.79	0.85	0.91	0.88	0.94	0.97	0.95
Disgust	0.71	0.81	0.76	0.88	0.80	0.84	0.93	0.93	0.93
Fear	0.78	0.94	0.85	0.76	0.91	0.83	0.93	0.93	0.93
Happy	0.76	0.67	0.71	0.68	0.66	0.67	0.85	0.91	0.88
Calm	0.76	0.91	0.83	0.83	0.92	0.88	0.93	0.95	0.94
Sad	0.85	0.79	0.81	0.85	0.75	0.80	0.92	0.86	0.89
Surprise	0.95	0.70	0.80	0.89	0.82	0.86	0.95	0.92	0.93
				IEMOCAP					
Anger	0.76	0.66	0.70	0.87	0.87	0.87	0.92	0.93	0.93
Happy	0.55	0.63	0.59	0.72	0.70	0.71	0.88	0.86	0.87
Neutral	0.64	0.67	0.65	0.64	0.88	0.74	0.86	0.97	0.92
Sadness	0.70	0.70	0.70	0.77	0.71	0.74	0.92	0.89	0.91

The results for both data corpora proved that a deep neural network with an appropriate balance for extracting spatial and temporal features improves the performance of the well-designed classifier. The time-distributed attention layers can extract the required spatiotemporal features, and the ensemble learner classifier can successfully generalize emotion classification from speech.

CONCLUSION AND FUTURE WORK

A TDACNN-RF model was proposed for an effective SER system. The experimental results show that the model is efficient in extracting temporal and spatial features. The time-distributed layers with wrappers assign equal weights to all features from the log-Mel spectrogram frames. The attention layers in the model effectively capture spatial features. This combination is better suited to handle spatiotemporal information, which is a common hindrance issue in time-series problems. The ensemble classifier RF classifies the emotions more efficiently without class confusion. The deep neural network for feature extraction integrated with the ensemble model for emotion recognition exhibits good performance in emotion classification on both the RAVDESS and IEMOCAP data corpora. The model shows good classification accuracy and performance measures (precision, recall, and F1-score) on both the data corpora, proving that generalization is achieved to a good extent. The results from the experiments are promising and provide a new direction to consider for various applications of SER in HCI.

In future work, the proposed model can be enhanced with rotational forest classifiers, deep neural network ensemble models, and CapsuleNets (Alonso et al., 2006; Ganaie et al., 2021; Patrick et al., 2022). Further, the proposed model can be applied to other speech-related applications and to time-series data.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Agajanian, S., Oluyemi, O., & Verkhivker, G. M. (2019). Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for classification and biomolecular modeling of cancer driver mutations. *Frontiers in molecular biosciences*. https://doi.org/10.3389/fmolb.2019.00044

Albornoz, E. M., & Milone, D. H. (2015). Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. *IEEE Transactions on Affective Computing*, 8(1), 43-53. https://doi.org/10.1109/TAFFC.2015.2503757

Alonso, C. J., Rodrı, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* March 2014. https://doi.org/10.1109/TPAMI.2006.211

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., & Schuller, B. (2017). Snore sound classification using image-based deep spectrum features. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3512–3516. https://doi.org/10.21437/Interspeech.2017-434

Atila, O., & Şengür, A. (2021). Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Applied Acoustics*, 182, 108260. https://doi.org/10.1016/J.APACOUST.2021.108260

Bhavan, A., Chauhan, P., Hitkul, & Shah, R. R. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184, 104886. https://doi.org/10.1016/j.knosys.2019.104886

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2007). IEMOCAP: Interactive emotional dyadic motion capture database. *Lang Resources & Evaluation,* 42, 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020a). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, 509, 150–163. https://doi.org/10.1016/j.ins.2019.09.005

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D Convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10), 1440–1444. https://doi.org/10.1109/LSP.2018.2860246

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, 28. https://doi.org/10.48550/arXiv.1506.07503

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., & Schuller, B. W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. MM 2017 - *Proceedings of the 2017, ACM Multimedia Conference*, 478–484. https://doi.org/10.1145/3123266.3123371

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92, 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

Fayek, H. M., Lech, M., & Cavedon, L., (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92(December), 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. (2021). Ensemble deep learning: A review. *arXiv*. https://doi.org/10.1016/j.engappai.2022.105151

Gudmalwar, A. P., Rama Rao, C. V., & Dutta, A. (2019). Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology*, 22(3), 521–531. https://doi.org/10.1007/s10772-018-09576-4

Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59. https://doi.org/10.1016/j.bspc.2020.101894

Jiang, W., Wang, Z., Jin, J. S., Han, X., & Li, C. (2019). Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors* (Switzerland), 19 (12), 1–15. https://doi.org/10.3390/s19122730

Kondo, K., & Taira, K. (2018). Estimation of binaural speech intelligibility using machine learning. *Applied Acoustics*, 129, 408–416. https://doi.org/10.1016/j.apacoust.2017.09.001

Kong, Y., & Yu, T. (2018). A Deep Neural Network model using Random Forest to extract feature representation for gene expression data classification. *Scientific Reports*, 8, 16477. https://doi.org/10.1038/s41598-018-34833-6

Kuchibhotla, S., Yalamanchili, B. S., Vankayalapati, H. D., & Anne, K. R. (2014). Speech emotion recognition using regularized discriminant analysis. *In Advances in Intelligent Systems and Computing*, 247, 363-369. https://doi.org/10.1007/978-3-319-02931-3_41

Kumar, S., Ratnoo, S., & Vashishtha, J. (2021). Hyper-heuristic evolutionary approach for constructing decision tree classifiers. *Journal of Information and Communication Technology*. 20, 2, 249–276. https://doi.org/10.32890/jict2021.20.2.5.

Lalitha, S., Tripathi, S., & Gupta, D. (2019). Enhanced speech emotion detection using deep neural networks. *International Journal of Speech Technology*, 22(3), 497–510. https://doi.org/10.1007/s10772-018-09572-8

Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2(May), 1–14. https://doi.org/10.3389/fcomp.2020.00014

Lee, T. H., Ullah, A., & Wang, R. (2020). Bootstrap aggregating and random forest. *In Macroeconomic forecasting in the era of big data. Advanced Studies in Theoretical and Applied Econometrics* (pp. 389-429). Springer.

Li, P., Song, Y., Mcloughlin, I., Guo, W., & Dai, L. (2018). An attention pooling based representation learning method for speech emotion recognition. *Interspeech*, September, 3087–3091.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* (Switzerland), 10 (10), 1163. https://doi.org/10.3390/electronics10101163

Livingstone, S. R., & Russo, F. A. (2018). 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*, 13 (5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lucky, H., & Suhartono, D. (2022). Botnet detection in IoT devices using random forest classifier with independent component analysis. *Journal of Information and Communication Technology Technology*, 1(1), 71–94. https://doi.org/10.32890/jict2022.21.2.3

Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia* 16(8), 2203–2213. https://doi.org/ 10.1109/TMM.2014.2360798

Mirsamadi, S., & Barsoum, E. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. *IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP) October. https://doi.org/10.1109/ICASSP.2017.7952552

Mustaqeem, & Kwon, S. (2020). CLSTM: Deep feature-based speech emotion recognition using the hierarchical convlstm network. *Mathematics*, 8(12), 1–19. https://doi.org/10.3390/math8122133

Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, 8, 79861–79875. https://doi.org/10.1109/ACCESS.2020.2990405

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input Features, signal length, and acted speech. *Interspeech*.

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Member, S. (2018). A study of language and classifier-independent feature analysis for vocal emotion recognition. November, 1–24. *arXiv* preprint arXiv:1811.08935.

Noroozi, F., Sapiński, T., & Kamińska, D. (2017). Vocal - based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2), 239–246. https://doi.org/10.1007/s10772-017-9396-2

Patrick, M. K., Adekoya, A. F., Mighty, A. A., & Edward, B. Y. (2022). Capsule networks–a survey. *Journal of King Saud University-computer and information sciences*, *34*(1), 1295-1310. https://doi.org/10.1016/j.jksuci.2019.09.014

Pawar, M. D., & Kokate, R. D. (2021). Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. *Multimedia tools and applications*. https://doi.org/10.1007/s11042-020-10329-2

Qiao, H., Wang, T., Wang, P., Qiao, S., & Zhang, L. (2018). A time-distributed spatiotemporal feature learning method for machine health monitoring with multi-sensor time series. *Sensors*, *18*(9), 2932. https://doi.org/10.3390/s18092932

Ren, Z., Pandit, V., Qian, K., & Yang, Z. (2017). Deep sequential image features for acoustic scene classification. *Detection and Classification of Acoustic Scenes and Events 2017,* 113-117.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. ICASSP, *IEEE International Conference on Acoustics, Speech and Signal Processing* -

Proceedings, 2015-August,4580–4584. https://doi.org/10.1109/ICASSP.2015.7178838

Sun, H., Zheng, X., Lu, X., & Wu, S. (2020). Spectral-spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5), 3232–3245. https://doi.org/10.1109/TGRS.2019.2951160

Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*,115, 29–37. https://doi.org/10.1016/J.SPECOM.2019.10.004

Tuncel, K. S., & Baydogan, M. G. (2018). Autoregressive forests for multivariate time series modeling. *Pattern Recognition*, 73, 202–215. https://doi.org/10.1016/j.patcog.2017.08.016

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8), 1301-1309. https://doi.org/10.1109/JSTSP.2017.2764438.

Wang, H., Zhang, Q., Wu, J., Pan, S., & Chen, Y. (2019). Time series feature learning with labeled and unlabeled data. Pattern Recognition, *Pattern Recognition*, 89, 55–66. https://doi.org/10.1016/J.PATCOG.2018.12.026

Wei, C., Chen, L. lan, Song, Z. zhen, Lou, X. guang, & Li, D. dong. (2020). EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomedical Signal Processing and Control*, 58, 101756. https://doi.org/10.1016/j.bspc.2019.101756

Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication*, 120(March), 11–19. https://doi.org/10.1016/j.specom.2020.03.005

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., & Gadekallu, T. R. (2021). Cross corpus multilingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4), 1845–1854. https://doi.org/10.1007/s40747-020-00250-4

Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019). Exploring Deep Spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, 7, 97515–97525. https://doi.org/10.1109/ACCESS.2019.2928625

Zheng, C., Wang, C., & Jia, N. (2020). An ensemble model for multi-level speech emotion recognition. *Applied Sciences* (Switzerland), 10(1), 205. https://doi.org/10.3390/app10010205

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1–2), 239–263. https://doi.org/10.1016/S0004-3702(02)00190-X

Zvarevashe, K., & Olugbara, O. (2020a). Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, 13(3), 70. https://doi.org/10.3390/a13030070

Zvarevashe, K., & Olugbara, O. (2020b). Recognition of cross-language acoustic emotional valence using stacked ensemble learning. *Algorithms*, 13(10), 246. https://doi.org/10.3390/a13100246.