TLEFuzzyNet: Fuzzy Rank based ensemble of Transfer Learning models for Emotion Recognition from Human Speeches

KARAM KUMAR SAHOO 1, ISHAN DUTTA 1, Muhammad Fazal Ijaz 2 (Member, IEEE), Marcin Wozniak 3, PAWAN KUMAR SINGH 4 (Senior Member, IEEE)

1Department of Computer Science and Engineering, National Institute of Technology, Mahatma Gandhi Road, A-Zone, Durgapur, West Bengal-713209 India
2Department of Intelligent Mechatronics Engineering, Sejong University, Seoul 05006, Korea
3Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland
4Department of Information Technology, Jadavpur University, Kolkata 700106, India

Corresponding author: Muhammad Fazal Ijaz (email: fazal@sejong.ac.kr) and Marcin Wozniak (email: marcin.wozniak@polsl.pl)

This research was supported by Sejong University research fund.

ABSTRACT

Human speech is not only a verbose medium of communication but it also conveys emotions. The past decade has seen a lot of research going on with speech data which becomes especially important for human-computer interaction and also healthcare, security and entertainment. This paper proposes the TLEFuzzyNet model, a three-stage pipeline for emotion recognition from speech. The first stage includes feature extraction by data augmentation of speech signals and extraction of Mel spectrograms, followed by the use three pre-trained transfer learning CNN models namely, ResNet18, Inception_v3 and GoogleNet whose prediction scores are fed to the third stage. In the final stage, we assign Fuzzy Ranks using a modified Gompertz function which gives the final prediction scores after considering the individual scores from the three CNN models. We have used the Surrey Audio-Visual Expressed Emotion (SAVEE), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Berlin Database of Emotional Speech (EmoDB) datasets to evaluate the TLEFuzzyNet model which has achieved state-of-the-art performance and is hence a dependable framework for Speech emotion recognition (SER). All the codes are available using GitHub link: https://github.com/KaramSahoo/SpeechEmotionRecognitionFuzzy

INDEX TERMS

TLEFuzzyNet, Deep learning, Fuzzy rank-based ensemble, Speech emotion recognition, Spectrogram, Transfer learning, Gompertz function

I. INTRODUCTION

Speech emotion recognition (SER) has been gaining popularity within the research fraternity for the past few decades and has the potential in the domain of Human-Computer Interaction, along with Multimedia and Biomedical applications to name a few. Speech is one of the main forms of relaying information to our surroundings hence a detailed analysis of speech signals is necessary. Emotion is a piece of vital information that speech signals carry apart from the verbal corpus. Any human-computer interface must be able to capture the underlying emotion of human speech since the same sentence could carry different meanings depending on the emotion.

SER has a lot of potential in speech-enabled interfaces such as Artificial Intelligence enabled voice assistants which could keep track of emotions and changes in the pattern to predict any psychological changes or signs of mental stress and depression. This can be extended in the medical field for the detection of Autism and Parkinson's Disease as long as adjuvant treatment and diagnosis. Educational software and smart classrooms could make use of SER for student mental health detection. Automated vehicles could prevent accidents and ensure safer driving environments by analyzing the speech of the driver and judging whether the driver is sound to drive or not.

The entire process of SER has two indispensable phases
namely feature extraction and classification. For the former, we can divide the acoustic features into two broad categories. First is the temporal features such as the energy of the signal, zero-crossing rate, maximum amplitude, and minimum energy. Conversion of these temporal features to the frequency domain using Fourier Transforms gives us spectral features. Some of the spectral features include spectral centroid, Mel spectrograms, Mel Frequency Cepstral coefficients (MFCC), spectral flux, Shifted Delta Cepstral Coefficients (SDCC), spectral density, and chroma-stft. Since both Deep learning and image classification has attained great heights in the past few decades, our paper proposes to bring image classification using 2D convolutional neural networks (CNN) models to the field of SER. Hence, the choice of features is the Mel Spectrogram of the audio data.

In recent years, using ensemble models to fuse the prediction scores from different constituent models has been in practice. In our paper, we propose to build an ensemble model that assigns fuzzy ranks with the help of a modified Gompertz Function. The audio data used are from the Surrey Audio-Visual Expressed Emotion (SAVEE), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Berlin Database of Emotional Speech (EmoDB) datasets. Ensemble Learning assigns ranks to the individual model predictions hence providing superior results than individual models. Imbalance and Correlation problems are taken care of simultaneously by ensemble learning models. The Fuzzy rank approach is a fusion technique that predicts the final classification results by assigning adaptive weights to the multi-class confidence scores of the constituent models. The Gompertz function named after Benjamin Gompertz is a sigmoid function in the time domain which saturates to the slowest in the beginning and end of a given period. It was originally modeled for human mortality since mortality decreases exponentially with age after which it saturates asymptotically. In the TLEFuzzyNet model, we choose three pre-trained transfer learning CNN models: ResNet18, GoogleNet, and Inception_v3. The fuzzy ranks were given to the prediction scores of the three models on the test set using the Gompertz Function which provided better accuracy than the constituent models hence making this a novel approach.

**A. MOTIVATION AND CONTRIBUTIONS**

Since modern technology is all about automation with gestures and voice, SER holds the key to designing state of the art Human-Computer interaction interfaces. From voice enabled security devices and authentication systems to Automated Vehicle Environments, emotion can be analysed to prevent identity mismatch or accidents. The medical sciences field too can benefit by classification of emotion from patients speech for treating Parkinson’s disease and Autism to name a few. Key points regarding our proposed model are as follows:

1. A huge amount of data is required when it comes to building an end-to-end model. Due to the lack of enough labeled audio data we preferred to use three pre-trained transfer learning models as our constituent models for the ensemble pipeline which are: Inception_V3, GoogleNet and ResNet18.
2. Training the audio samples on a single model for classification may lead to the problem of imbalance. Ensemble approach gives an aggregate opinion to all the individual models thereby decreasing noise and giving better and unbiased prediction scores. This makes it a novel approach for SER.
3. A modified Gompertz function is employed to assign fuzzy ranks to the decision scores of the individual models. The prediction scores rarely go as low as zero and the Gompertz function saturates exponentially to an asymptote. This is different and more efficient than traditional ensemble pipelines since fuzzy ranks based fusion assigns adaptive priority weights to the prediction scores.
II. LITERATURE REVIEW

SER has seen a lot of research for the past few decades. A number of different machine-learning models have been proposed in the past which include hidden Markov models\textsuperscript{38}, Support Vector Machines\textsuperscript{41}, Decision Trees and Random Forests\textsuperscript{37}. Dellaert et al.\textsuperscript{9} used smoothing spline approximation on the contour of the pitch for feature extraction. Pattern Recognition techniques were employed for classification purposes. Machine Learning models were applied by Noroozi et al.\textsuperscript{36} such as Multi-class Support Vector Machine, Random Forest and Single-layered Adaptive boosting on SAVEE and Polish database. Noroozi et al.\textsuperscript{36} were able to obtain 75.71% and 87.91% testing accuracies on the respective databases using random forest models. By using a reduced feature size of 14 they were able to reduce computational time as well.

Nicholson et al.\textsuperscript{35} used a special type of neural network called the one-class-in-one neural network which predicted the best emotion by making use of 12 LPC and Delta LPC parameters which attained an accuracy of 50%. Deshmukh et al.\textsuperscript{11} worked with the Mel-frequency cepstral coefficients (MFCC) features and trained their dataset on a Support Vector Machine (SVM) achieving 80% accuracy. Slimi et al.\textsuperscript{45} generated Mel spectrograms of the audio data and resizened them to a size of 150 by 66. These resized spectrograms were flattened and fed to a single hidden layer neural network. The classifications were made by a softmax layer. Their proposed model achieved an accuracy of 81.82% after 1970 epochs of training on the EmoDB database. Badshah et al.\textsuperscript{1} used deep convolutional neural networks to feed their model with Mel spectrograms of the speech data. They conducted two experiments. In the first method they trained the CNN model from scratch and in the second approach they fine tuned a pre trained Alex-Net model which gave them a test accuracy of 84.3%. Etienne et al.\textsuperscript{13} tried to capture the long term dependencies of speech by using a CNN+LSTM architecture. The high level features were captured by the CNN model from spectrograms whereas the recurrent LSTM layers extracted relation among the temporal features.

Zehra et al.\textsuperscript{54} have used an ensemble approach by combining results from decision tree (J48), sequential minimal optimization (SMO), and random forest (RF). Their main objective was to design robotic systems that could analyze and detect emotions in not only in-corpus but also cross-corpus data. Khan and Roy\textsuperscript{25} have developed a model that is both gender dependent and independent, by training a Naive Bayes Classifier on both pitch and MFCC features using the EmoDB database and have achieved an accuracy of 95.20%.

III. DATASET

Since SER is a classification problem, the availability of sufficient data which is correctly labeled is of prime concern. Different individuals could perceive the same speech as different emotions hence it is very important to chose a dataset that is labelled by people from the same group of database creators. We have chosen to work with simulated databases where particular actors or speakers simulate the required emotion by reading out from a set of sentences. In our work we use the SSAVEE, EmoDB\textsuperscript{3} and RA\textsuperscript{3} VDESS\textsuperscript{3} dataset which contain accurately labelled and free of noise audio samples. All the three databases are accessible publicly.

A. SSAVEE

The SSAVEE dataset contains standard TIMIT sentences recorded by 4 male actors. All the audio samples have been recorded, processed and labelled using superior quality equipment in a visual and media laboratory. The dataset has 480 audio samples in the .wav format and divided into seven emotion classes: angry, fear, disgust, surprise, sad, happy and neutral.

B. EMOODB

EmoDB\textsuperscript{3} is a German Emotional Speech dataset of 535 utterances available in .wav format. It was created by the Institute of Communication Science, Technical University, Berlin, Germany. There are 5 male speakers and 5 female speakers. There are seven labelled emotion classes: anger, boredom, neutral, happiness, anxiety, sadness and disgust.
The sentences that the speakers are made to utter are everyday phrases and can be used with a variety of emotions.

C. RAVDESS

The RAVDESS is a facial and vocal expression database in North American English. 24 professional actors vocalize statements in a North American accent. There are 8 classes of emotions which are calm, happy, sad, angry, fearful, surprise, disgust and neutral. All the classes except neutral are uttered in two levels of emotional intensity. 247 participants reviewed the database with 72 participants retesting for better accuracy of the dataset.

IV. METHODOLOGY

The proposed framework has been categorized into the following subsections that include data augmentation, feature extraction, transfer learning and finally assigning the fuzzy ranks using the Gompertz function for the final prediction. A diagrammatic workflow of the entire TLEFuzzyNet model from feature extraction to final prediction has been given in Figure 1.

A. AUGMENTATION

The size of any dataset is a crucial deciding factor for any deep learning model. Small amount of data hinder the deep learning models to map the inputs to the ground label accurately. High variance in the predictions of test data is a major problem that inadequate data can give rise to. To overcome this problem we use augmentation to create multiple training data samples from the limited audio files in the SAVIEE and EmoDB databases. The technique we have employed in this time shifting whose equation is as follows:

\[ X_{\text{new}} = X[n \pm s] \]

where, \(X[n]\) is the audio signal and \(s\) refers to the number of samples shifted. In the present work, the original audio is sampled at 44100 Hz and the shifting is done by \(s\) samples both to the right and left.

B. FEATURE EXTRACTION

The choice of effective features from the speech data is very important to attain state-of-the art performance for TLE-FuzzyNet model. There is a cascade of important features pertaining to speech such as pitch, energy, Mel spectrograms, Mel-frequency cepstrum coefficients (MFCC), Linear Prediction Cepstrum Coefficients (LPCC), modulation spectral features (MSFs) to name a few. Since we are using Convolutional Neural Network (CNN) models to train our datasets, the Mel spectrogram is an ideal choice of features.

Spectrograms are generated with the help of Fourier Transforms on sound signals. The sound signal is divided into small time segments to which the Fourier Transform is applied individually. As a result we get a frequency versus time graph with the amplitude of the frequency denoted by the color of the spectrogram. However, humans do not perceive frequency linearly but rather logarithmically. The Mel scale solves this problem by mapping the perceived frequency to the measured frequency of a tone. Figure 2 plots the mel pitch against the frequency of the sound waves. We can clearly see from fig. 2 that with increasing frequencies the slope of the curve decreases thereby establishing that differentiating between higher frequencies is difficult compared to lower frequencies.
C. TRANSFER LEARNING

The non-availability of sufficient data makes it difficult to create Deep neural network models. Transfer learning solves this problem because here we can reuse models that have been pre-trained on huge datasets. The weights of the neural network from one model are used on another model which takes as input a totally new set of data. The last few layers of the pre-trained model are fine-tuned by re-training them on the new datasets which achieves better performance and lesser training time. In image classification problems, transfer learning has been used extensively because obtaining millions of images for a database is not always feasible. CNNs demand hardware memory and are compute-intensive hence making it difficult to train in scenarios with limited power supply. Therefore not having to re-train the entire transfer learning model saves training time and system resources.

The three databases are split into training, validation and testing sets in the ratio 8:1:1. The training and validation datasets are used to fine tune the transfer learning models for our mel spectrogram data. The testing dataset remains unseen by the model during training. The GoogLeNet, ResNet18 and Inception_v3 models are loaded from the Pytorch Model Zoo. They are pretrained on the ImageNet weights and fine tuned using the Adam\textsuperscript{24} optimizer. The confidence scores for the testing split dataset are generated for all the three models.
Figure 6: A training batch of 16 spectrograms from the RAVDESS dataset with the labels denoting the respective classes present in the database. The labels are mapped as following: 'f' for fear, 'd' for disgust, 'h' for happiness, 's' for sad, 'n' for neutral 'su' for surprised, 'a' for angry and 'c' for calm.

and stored as a csv file for using in the fuzzy rank based ensemble step.

The three transfer learning models are described below as follows:

1) Inception_v3

Inception_v3 is a popular 2D CNN model with great classification performance that uses transfer learning. It is an extended network of the GoogLeNet model that extensively uses Batchnorm in the activation layers. Input to this model are of sizes 299x299x3 which go through layers of different convolution layers for feature maps extraction. The inception blocks of Inception_v3 make it possible to compute on different filters of feature extraction by concatenating them into a single feature map (see figure 5 for further details). This architecture decreases the computational complexity by reducing the number of parameters in the model.

2) ResNet18

ResNet or Residual Network, introduced in 2015, is an architecture that uses residual mapping and is very effective against the “degradation problem” in deep networks. The residual learning approach enhances the optimization phase of the CNN model. Like most popular image recognition CNN models the ResNet-18 is pretrained on the ImageNet dataset. It takes as input images of size 3x224x224 which is lesser than the input size of inception_v3 model. The deeper the residual network the better performance it can achieve. There are different depth wise implementation of the Resnet model such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-110. The ResNet18 model (see Figure 3 for further details) used in our proposed TLEFuzzyNet framework is a perfect balance between computation complexity and accuracy.

3) GoogleNet

GoogLeNet is a 22 layer CNN model and one of the popular networks of the Inception architecture, developed by the researchers at Google. It takes as input images of shape 224x224. The network was specially designed to achieve computational efficiency and practicality so that inference can be run on any individual system, including those that do not have very high computational power. One method by which GoogLeNet achieves efficiency is through reduction of the input image whilst simultaneously retaining important spatial information. To prevent overfitting of the network a regularisation technique is used during training called the dropout layer (40%), just before the linear layer. It works by randomly reducing the amount of interconnecting neurons within a neural network. A pictorial view of the resulting network is depicted in Figure 4.

Each dataset is trained separately on the three models for a total 50 epochs on SAVIE, 10 epochs on RAVDESS and 25 epochs on the EmoDB dataset. The epochs are experimentally determined to prevent overfitting on sample data. A batch size of 16 is used with a learning rate of 0.001 using 6
the Adam optimizer. Figure 6 shows the training batch which is used to classify the emotions of the corresponding audio data.

D. FUZZY RANK ENSEMBLE

In literature, the traditional ensemble technique gives equal priority to the classification scores of all constituent models, and pre-computed weights are used for the classifiers. The main issue with such an ensemble is the generation of static weights which are difficult to modify in the phase where we classify the test samples. However, in the proposed fuzzy-rank-based ensemble approach, each base classifier’s predictions scores are taken into account for every individual test case separately. This way, enhanced and more accurate scores for classification can be obtained using this ensemble method. This is a dynamic process and there is no need to change any weights for different test datasets.

The Gompertz function describes time series with the slowest growth in the beginning and end of a time period. Originally used to describe the mortality rate concerning growing age, it is now used extensively in the field of biology. The Gompertz function can explain the growth of a population, cancerous tumor, colony of bacteria, as well as the number of affected people during an epidemic. The equation to understand the function is:

\[ f(t) = ae^{-b - et} \]

where, \( a \) is an asymptote, \( b \) sets the displacement along the \( x \)-axis, \( c \) is used for \( y \)-scaling and \( e \) is Euler’s Number. Figures 7-9 shows the graphs of the Gompertz function with varying values for \( a, b \) and \( c \) respectively.

In our proposed method, we use a redesigned version of the Gompertz Function. Considering \( N \) to be the number of constituent models, we have \( N \) number of prediction scores for each image in the test split of the database. As discussed above, we have made use of three transfer learning CNN models therefore \( N=3 \). If \( L \) is the number of labels in the dataset then:

\[ \sum_{l=1}^{L} S_{l}^{(n)} = 1; \forall n, n = 1, 2, 3, \ldots, N \]  

(3)

The prediction scores denoted by \( S \) in equation 3 of each class for each given sample data of the are taken into account while generating the fuzzy ranks. For the \( l^{th} \) class the fuzzy ranks given on account of the \( n^{th} \) constituent model are given by the following formula:

\[ R_{l}^{(n)} = (1 - \epsilon^{-2 \times \epsilon^{(n)}}) \]  

(4)

\[ \forall l, n; n = 1, 2, \ldots, N; l = 1, 2, \ldots, L \]

Corresponding to each class in the dataset, there can be \( k \) top classes which in our proposed method we have chosen as 2. The eqs. (5) to (6) are used to calculate the Fuzzy Ranks \( FRS_{l} \) and complement of confidence factor sum \( CCFS_{l} \) for the class \( l \). If the label \( l \) does not fall under the top \( k \) classes a penalty value of \( P^{R} \) and \( P^{CF} \) is imposed on the corresponding class. The final predicted class for the data instance \( X \) is calculated by multiplying the \( FRS_{l} \) and \( CCFS_{l} \) and finding the minimum value among all the classes as shown in eq. (7).

\[ FRS_{l} = \sum_{i=1}^{N} \begin{cases} R_{l}^{(i)}, & \text{if } R_{l}^{(i)} \in K^{(i)} \\ P^{R}, & \text{otherwise} \end{cases} \]  

(5)

\[ CCFS_{l} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} CF_{l}^{(i)}, & \text{if } R_{l}^{(i)} \in K^{(i)} \\ P^{CF}, & \text{otherwise} \end{cases} \]  

(6)

\[ \text{class}(X) = \min \{ FRS_{l} \times CCFS_{l} \} \forall l = 1, 2, \ldots, L \]  

(7)

V. RESULTS AND DISCUSSION

In this segment, we have provided tabular data for the results we have acquired after working on the three aforementioned datasets. A detailed explanation of the evaluation metrics, performance of the CNN Transfer learning models, and the final ensemble model is provided. We have compared our work with previous researches and proved that our method has attained state-of-the-art performance for the SER problem using an ensemble approach with deep learning 2D CNN models.

A. EVALUATION METRICS

To assess the performance of TLEFuzzyNet model, we have considered F1-Score, Precision, Recall and Accuracy as our evaluation metrics. A majority of past researches have used Accuracy as the standard metric for evaluation of performance. As a result we will be providing a comparative study between the TLEFuzzyNet model and previous models for SER problem.

The mentioned evaluation metrics can be calculated using basic parameters such as True Positives, True Negatives, False Positives and False Negatives. The corresponding formulas are as follows:

Accuracy:

\[ \text{Accuracy}_i = \frac{\sum_i M_{i}}{\sum_i \sum_j M_{ij}} \]  

(8)

Precision:

\[ \text{Precision}_i = \frac{\sum_i M_{i}}{\sum_j M_{ij}} \]  

(9)

Recall:

\[ \text{Recall}_i = \frac{\sum_i M_{ii}}{\sum_j M_{ij}} \]  

(10)

F1 Score:

\[ \text{F1 Score}_i = \frac{2}{\frac{1}{\text{Precision}_i} + \frac{1}{\text{Recall}_i}} \]  

(11)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
The reason we have used Precision, Recall and F1 scores is because of the unsymmetrical distribution of samples in our database. In the upcoming sections we give a comparative study between previous SER model which includes deep learning as well as machine learning classifiers.

B. PERFORMANCE OF CONSTITUENT MODELS

Each of the three constituent models has been loaded from the Pytorch Model Zoo and is pretrained on the ImageNet dataset. The entire model weights have been freezed except the classification layers. The classification layers initially had an output layer with softmax activation of size (1, 1000) which are fine-tuned to (1, num_of_classes). Each model has been trained for exactly 50 epochs on SAVVEE, 25 epochs for EmoDB, and 10 epochs for RAVDESS dataset after which the best validation accuracy has been taken into consideration. Adam optimizer has been used for gradient descent with the learning rate as 0.001 and $\beta$ values (0.9, 0.99). The training process has been experimented with different learning rates, batch sizes, and number of epochs and the final values have been experimentally chosen for TLEFuzzyNet model.

The Inception_V3 model achieved 98.76% accuracy, ResNet18 model achieved 98.77% accuracy and GoogLeNet model achieved 99.08% accuracy for the EmoDB dataset. The ensemble model after applying ranks to the classification scores predicts the labels from the corresponding testing datasets with 99.38% accuracy which is greater than that of the constituent models.

For the RAVDESS dataset, the Inception_V3, ResNet18, and GoogLeNet models achieved an individual classification accuracy of 92.07%, 95.29%, and 97.24% respectively. However, the ensemble model was able to assign Fuzzy ranks in a manner to correctly classify each instance from the corresponding testing dataset with a final accuracy of 99.66%. These results connote the ability of the ensemble approach to minimize the errors of each CNN model and generate more accurate classification scores.
Table 2: Performance of the TLEFuzzyNet model on the three benchmark SER datasets.

| Sl No. | Researcher | Method | Feature set used | Accuracy  |
|--------|------------|--------|------------------|-----------|
| 1      | Hajarol pavadi and Demirel [+] (2019) | 3D CNN based K-Means clustering algorithm | Spectrogram conversion using STFT | 81.05%    |
| 2      | Dey et al. [12] (2020) | Gradient boosting on features selected using optimization algorithms | LPC and LPCC | 97.31%    |
| 3      | Chattopadhyay et al. [8] (2020) | Manta Ray Optimization of concatenated features | MFCC and LPC | 97.49%    |
| 4      | TLEFuzzyNet model | Fuzzy rank based ensemble approach using Gompertz Function | Mel Spectrogram | 98.57%    |

Table 3: Comparison of TLEFuzzynet performance with state-of-the-art works for SAVEE dataset.

| Sl No. | Researcher | Method | Feature set used | Accuracy  |
|--------|------------|--------|------------------|-----------|
| 1      | Nantasri et al. [34] (2020) | Light-Weight Artificial Neural Network | Average Values of MFCCs and Their Derivatives | 87.80%    |
| 2      | Meng et al. [22] (2019) | Dilated CNN with residual block and Bi-LSTM based on the attention mechanism | 3-D Log-Mel spectrums | 90.78%    |
| 3      | Dey et al. [14] (2020) | Gradient boosting on features selected using optimization algorithms | LPC and LPCC | 98.46%    |
| 4      | TLEFuzzyNet model | Fuzzy rank based ensemble approach using Gompertz Function | Mel Spectrogram | 99.38%    |

Table 4: Comparison of TLEFuzzynet performance with state-of-the-art works for EmoDB dataset.

| Sl No. | Researcher | Method | Feature set used | Accuracy  |
|--------|------------|--------|------------------|-----------|
| 1      | Yadav and Vishwakarma [22] (2020) | CNN and Bi-LSTM | Deep neural network feature vectors and MFCCs | 73%       |
| 2      | Mustaqeem et al. [33] (2020) | Sequence segment selection and Bi-LSTM | Spectrograms and CNN feature vectors | 77.02%    |
| 3      | Kanwal and Asghar [21] (2021) | Feature optimization using clustering-based genetic algorithm | INTERSPEECH 2010 Challenge feature set | 82.50%    |
| 4      | TLEFuzzyNet model | Fuzzy rank based ensemble approach using Gompertz Function | Mel Spectrogram | 99.66%    |

Table 5: Comparison of TLEFuzzynet performance with state-of-the-art works for RAVDESS database.

| Sl No. | Method | Feature set used | Accuracy  |
|--------|--------|------------------|-----------|
| 1      | CNN and Bi-LSTM | Deep neural network feature vectors and MFCCs | 73%       |
| 2      | Sequence segment selection and Bi-LSTM | Spectrograms and CNN feature vectors | 77.02%    |
| 3      | Feature optimization using clustering-based genetic algorithm | INTERSPEECH 2010 Challenge feature set | 82.50%    |
| 4      | Fuzzy rank based ensemble approach using Gompertz Function | Mel Spectrogram | 99.66%    |

Figure 10 shows the learning curves for the SAVEE dataset using the Inception_v3, GoogLeNet and Resnet18 models respectively. We can infer from the graphs that the model does not learn anything new and reaches a maximum accuracy around the 10th epoch. Similarly, for EmoDB dataset (shown in Figure 11) and RAVDESS dataset (illustrated in Figure 12), the training of the model reaches a point where the accuracy halts to improve which is again around the 10th epoch. All the learning curves have been plotted using the TensorBoard library in Python.

**C. PERFORMANCE OF ENSEMBLE MODEL**

The ensemble model assigns fuzzy ranks to the classification scores given by the three CNN Transfer learning models mentioned in IV-C. Classification scores from the previous transfer learning phase are stored for each sample in the testing dataset. In this phase we assign fuzzy ranks for the top \( k \) classes and give penalties to the other class predictions as discussed in IV-D. The ensemble model generates the final prediction scores for the respective number of classes for each database. Table 2 gives us the accuracy, recall, precision, and F1 score along with the overall accuracy of the ensemble. The class column refers to the different emotion classes for each dataset. For the SAVEE dataset classes 1-7 are anger, disgust, fear, happiness, neutral, sad and surprised in order. For the EMODB dataset classes 1-7 are fear, disgust, happiness, boredom, neutral, sad and anger in order. For the RAVDESS dataset classes 1-8 are neutral, calm, happiness,
sad, anger, fear, surprise and disgust in order.

In order to visualize the classification performance of our ensemble model, we make use of the receiver operating characteristic curve or ROC curve. The ROC curve can be used for multi-class classification though conventionally it is employed for binary classification. A method known as One vs All where the ground truth class is treated as one label and other classes treated as a collective label. The ROC curve is the measure of the models to distinguish between classes. The area under the ROC curve is proportional to the accuracy with which a class is classified correctly. The True Positive Rate is drafted against the False Positive Rate in the ROC curve.

The model achieves perfect accuracy is the area under ROC curve is 1. It can separate between the multiple classes with 100% accuracy. A ROC curve with almost 0 area under the curve will provide the wrong prediction for each sample data, hence being a poor model. The ROC curves for our ensemble model on the RAVDESS, EmoDB and SAVEE dataset are given in figs. 13 to 15. We can see in fig. 13 that the area under the ROC curve has nearly approached 1 which is implied by the accuracy of 99.66% for the RAVDESS dataset.

\[
TPR = \frac{TP}{TP + FP} \quad (12)
\]
\[
FPR = \frac{FP}{TN + FP} \quad (13)
\]

D. COMPARISON WITH OTHER FRAMEWORKS

In this section, we present an explicate analysis between the performance of past researches and frameworks with our proposed TLEFuzzyNet model. The datasets we have used are open source and widely used for research in the field of speech processing and speech emotion recognition. Since research in the domain of SER has been evolving since the last few decades there is a cascade of deep learning and machine learning models to compare TLEFuzzyNet model with. Tables 3 -5 give a tabular comparison of TLEFuzzyNet model with other state of the art researches.

The datasets we have worked with are popular SER-datasets. As mentioned in Table 1, it can seen that both EmoDB and SAVEE are relatively small datasets compared to RAVDESS dataset. Additionally, EmoDB is an example of imbalanced dataset with relatively more data corresponding to anger and least for disgust. As described in section IV-A we have implemented an additional augmentation phase by which we increase the data by time shifting of the original speeches. This aids in counteracting the problem of small dataset sizes. Since RAVDESS is a comparatively larger dataset, the models are able to capture more relevant features from the Mel spectrograms and map the input to the output (emotion) with lesser bias towards wrong emotion outputs. Despite the difference in dataset sizes, the prediction accuracy for the three datasets is numerically comparable. This is due to the dynamic nature of assigning ranks to the classification score of each test sample. As highlighted in Tables 3-4 the previous state of the art models were able to achieve accuracy >95% for EmoDB and SAVEE owing to the small dataset sizes. The models were able to fit the small data with greater accuracy. Nevertheless our proposed TLEFuzzyNet model beats the previous state-of-the-art models,
though the margin of improvement seems to be less due to the already high performance of previous models.

In case of RA VDESS dataset, the previous frameworks mentioned in Table 5 are able to classify different emotions with decent accuracy in the range 70%-85%. This might be due to the comparatively large size of the dataset. Our framework does not entirely depend on the transfer learning phase for classification, but the classification scores are fed to the fuzzy-ranked ensemble model which ultimately provides the final classification results after assigning ranks to the top-k (which is 2 in our proposed paper) predicted classes for each test sample for each constituent model. As a result of which we gain superior performance compared to the previous benchmark models that were trained on the RA VDESS dataset.

VI. CONCLUSION

This paper proposed an ensemble learning based framework for SER using transfer learning 2D CNN models. It was found that models pre-trained on huge image datasets can extract essential features from Mel spectrograms of audio data hence converting the task of speech processing and recognition into a computer vision task. TLEFuzzyNet model combined transfer learning, CNNs and fuzzy rank based ensemble approach by making use of the Gompertz function. Since the datasets used were not of very large scale hence transfer learning was a good choice for training the deep convolutional neural networks. The dynamic assignment of ranks to the classifiers makes it possible to make predictions without having to initialize a new set of weights for the entire ensemble phase of the framework for newer datasets. Errors of each individual CNN classifier is compensated by the fuzzy ranking algorithm. The experimental results depict that TLEFuzzyNet model has achieved state-of-the-art accuracy of 98.57%, 99.38% and 99.66% on all the three benchmark datasets namely, SAVEE, EmoDB, and RA VDESS respectively. There is a promising application of transfer learning and ensemble approaches for SER.

There are few areas where TLEFuzzyNet model can be improved which are as follows:

1) In our CNN models, we have used traditional Mel spectrograms. This greatly increases computation due to the input image sizes of (224, 224) for the transfer learning models. In future we can use smaller feature sets such as MFCC or feature vectors from neural network architectures of smaller size.

2) The generalization of the framework can be improved by using better data augmentation techniques such as voice conversion using generative model and speed perturbation.

References

[1] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 International Conference on Platform Technology and Service (PlatCon), pages 1–5, 2017.

[2] S. Basu, J. Chakraborty, and M. Aftabuddin. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pages 333–336. IEEE, 2017.

[3] F. Burkhardt, A. Paeschke, M. Rolfs, W. Sendmeier, and B. Weiss. A database of german emotional speech. volume 5, pages 1517–1520, 09 2005.

[4] L. Cai, J. Dong, and M. Wei. Multi-modal emotion recognition from speech and facial expression based on deep learning. In 2020 Chinese Automation Congress (CAC), pages 5726–5729. IEEE, 2020.

[5] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkov, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4):377–390, 2014.

[6] S. Chakrabortty, Y. Deng, and G. Cauwenberghs. Robust speech feature extraction by growth transformation in reproducing kernel hilbert space. IEEE transactions on audio, speech, and language processing, 15(6):1842–1849, 2007.

[7] J. Chaloupka, P. Cerva, J. Silovský, J. Žiďánský, and J. Nouza. Modification of the speech feature extraction module for the improvement of the system for automatic lectures transcription. In Proceedings ELMAR-2012, pages 223–226. IEEE, 2012.

[8] S. Chattopadhyay, A. Dey, and H. Basak. Optimizing speech emotion recognition using manta-ray based feature selection. arXiv preprint arXiv:2009.08909, 2020.

[9] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP ’96, volume 3, pages 1970–1973 vol.3, 1996.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[11] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni. Speech based emotion recognition using machine learning. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pages 812–817, 2019.

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

[13] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch. Cnn-lstm architecture for speech emotion recognition with data augmentation. Workshop on Speech, Music and Mind (SMM 2018), Sep 2018. URL: http://dx.doi.org/10.21437/SMM-2018-5.

[14] N. Hajarolavadi and H. Demirel. 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy, 21(5):479, 2019.

[15] K. Han, D. Yu, and I. Tsachev. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech 2014, 2014.

[16] H. Hu, M.-X. Xu, and W. Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07, volume 4, pages IV–413. IEEE, 2007.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. PMLR, 2015.

[18] A. Iqbal and K. Barua. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pages 1–5. IEEE, 2019.

[19] N. T. Ira and M. O. Rahman. An efficient speech emotion recognition using ensemble method of supervised classifiers. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), pages 1–5, 2020.

[20] D. Issa, M. F. Demirci, and A. Yazici. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59:101894, 2020.

[21] S. Kanwal and S. Asghar. Speech emotion recognition using clustering based ga-optimized feature set. IEEE Access, 9:125830–125842, 2021.

[22] L. Kerkeni, Y. Serrestou, M. Mbarik, K. Raoof, M. A. Mahjoub, and C. Cleder. Automatic speech emotion recognition using machine
learning. In *Social media and machine learning*, InteConOpen, 2019.

[23] A. Khan and U. K. Roy. Emotion recognition using prosodic and spectral features of speech and naive bayes classifier. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WISPNET)*, pages 1017–1021, 2017.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] J. L. O. W. Kist, V. S. Asirvadam, and M. F. Hassan. Enhanced ensemble models for predictive modeling: A conceptual framework. In *2011 IEEE International Conference on Signal Processing & Its Applications (CSPA)*, pages 24–28. IEEE, 2021.

[26] D. Kumar and V. Kukreja. N-cnn based transfer learning method for classification of powdery mildew wheat disease. In *2017 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 707–710. IEEE, 2021.

[27] R. Kundu, H. Basak, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar. Fuzzy rank-based fusion of cnn models using gompertz function for screening covid-19 ct-scans. *Scientific reports*, 11(1):1–12, 2021.

[28] B. V. Le and S. Lee. Adaptive hierarchical emotion recognition from speech signal for human-robot communication. In *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 807–810. IEEE, 2014.

[29] M. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju. Speech based human emotion recognition using mfcc. In *2017 international conference on wireless communications, signal processing and networking (WISPNET)*, pages 2257–2260. IEEE, 2017.

[30] 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*, 13(5):e0196391, 2018.

[31] R. Lotfidereshgi and P. Gournay. Biologically inspired speech emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5135–5139. IEEE, 2017.

[32] H. Meng, T. Yan, F. Yuan, and H. Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019.

[33] Mustaqaem, M. Sajjad, and S. Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep bilstmr. *IEEE Access*, 8:79861–79875, 2020.

[34] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerittayanun, A. Rugchatjaroen, S. Usanavasin, and T. Shinozaki. A light-weight artificial neural network for speech emotion recognition using average values of mfccs and their derivatives. In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 41–44, 2020.

[35] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. In *ICONIP’99. ANZIS’99 ANNES’99 ACNN’99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378)*, volume 2, pages 495–501 vol.2, 1999.

[36] F. Noroozi, D. Kaminska, T. Sapinski, and G. Anbarjafari. Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost. *Journal of the Audio Engineering Society*, 65(7/8):562–572, 2017.

[37] F. Noroozi, T. Sapinski, D. Kaminska, and G. Anbarjafari. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246, 2017.

[38] T. L. Nwe, S. W. Foo, and L. C. De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[39] Y. Pan, M. Xu, L. Liu, and P. Jia. Emotion-detecting based model selection for emotional speech recognition. In *The Proceedings of the MultiConference on "Computational Engineering in Systems Applications"*, volume 2, pages 2169–2172. IEEE, 2006.

[40] H. Park, T. Takiguchi, and Y. Ariki. Integration of phoneme-subspaces using lda for speech emotion recognition. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pages 148–151. IEEE, 2008.

[41] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-biometric network learning. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–577, 2004.

[42] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797, 2010.

[43] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrum predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[44] M. Smith, E. Aswathi, T. Deepa, C. Shameema, and S. Rajan. Emotion recognition from audio signals using support vector machine. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 139–144. IEEE, 2015.

[45] A. Slimi, M. Hamroun, M. Zrigui, and H. Nicolas. Emotion recognition from speech vectors and shallow neural networks. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia, MoMM ’20*, page 35–39, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450389242. URL https://doi.org/10.1145/3428690.3429153.

[46] M. H. Taneveer, H. Zhu, W. Ahmed, A. Thomas, B. M. Imran, and M. Salman. Mel-spectrogram and deep cnn based representation learning from bio-sonar implementation on uavs. In *2021 International Conference on Computer, Communication and Robotics (ICCCCR)*, pages 220–224. IEEE, 2021.

[47] K. M. Tjøeve and E. Tjørve. The use of gompertz models in growth analyses, and a new gompertz-model approach: An addition to the unified-controls family. *PloS one*, 12(6):e0178691, 2017.

[48] P. Tzirakis, J. Zhang, and B. W. Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[49] J. Unamaheswari and A. Akila. An enhanced human speech emotion recognition using hybrid of prnn and kmm. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 177–183. IEEE, 2019.

[50] J. Wang and Z. Han. Research on speech emotion recognition technology based on deep and shallow neural network. In *2019 Chinese Control Conference (CCC)*, pages 3555–3558. IEEE, 2019.

[51] J. Wang, S. Kim, and Y. Lee. Speech augmentation using wavenet in speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6770–6774, 2019.

[52] A. Yadav and D. K. Vishwakarma. A multilingual framework of cnn and bi- lstm for emotion classification. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2020.

[53] C. Yogesh, M. Hariraran, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat. A new hybrid pso assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, 69:149–158, 2017.

[54] W. Zehra, A. R. Javed, Z. Jailil, H. U. Khan, and T. R. Gadekallu. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, pages 1–10, 2021.

[55] J. Zhang, Y. Chen, and J. Zhang. Speech feature extraction of kpca based on kernel fuzzy k-means clustering. In *2011 International Conference on Computer Science and Service System (CSSS)*, pages 756–759. IEEE, 2011.

[56] J. Zhao, X. Mao, and L. Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.
KARAM KUMAR SAHOO (email: karamsahoo@gmail.com) is currently pursuing his Bachelor’s degree in Computer Science and Engineering from National Institute of Technology, Durgapur, India. His research interests include Deep Learning and Computer Vision.

ISHAN DUTTA (email: ishandutta1220@gmail.com) is currently pursuing his Bachelor’s degree in Computer Science and Engineering from National Institute of Technology, Durgapur, India. His research interests include Computer Vision, Deep Learning and Robotics.

MUHAMMAD FAZAL IJAZ (email: fazal@sejong.ac.kr) received his B.Eng. degree in Industrial Engineering and Management from University of the Punjab, Lahore, Pakistan, in 2011, and Dr. Eng. degree in Industrial and Systems Engineering from Dongguk University, Seoul, South Korea, in 2019. From 2019 to 2020, he worked as an Assistant Professor in Department of Industrial and Systems Engineering, Dongguk University, Seoul, South Korea. Currently, he is working as an Assistant Professor in Department of Intelligent Mechatronics Engineering, Sejong University, Seoul, Korea. He has published numerous research articles in several international peer-reviewed journals, including IEEE Access, Sensors, Symmetry, Journal of Food Engineering, Applied Sciences, Asia Pacific Journal of Marketing and Logistics, and Sustainability. His research interests include Machine learning, Blockchain, Healthcare Engineering, Internet of Things, Supply chain management, Big data, and Data mining.

PAWAN KUMAR SINGH (email: pksingh.it@jadavpuruniversity.in) received his B.Tech degree in Information Technology from West Bengal University of Technology in 2010. He received his M.Tech in Computer Science and Engineering and Ph.D. (Engineering) degrees from Jadavpur University (J.U.) in 2013 and 2019, respectively. M. Woźniak is an Associate Professor with the Faculty of Applied Mathematics, Silesian University of Technology. He is a Scientific Supervisor in editions of "The Diamond Grant" and "The Best of the Best" programs for highly talented students from the Polish Ministry of Science and Higher Education. He participated in various scientific projects (as Lead Investigator, Scientific Investigator, Manager, Participant and Advisor) at Polish, Italian and Lithuanian universities and projects with applied results at IT industry. He was a Visiting Researcher with universities in Italy, Sweden, and Germany. He has authored/coauthored over 200 research papers in international conferences and journals. His current research interests include neural networks with their applications together with various aspects of applied computational intelligence accelerated by evolutionary computation and federated learning models. In 2017 he was awarded by the Polish Ministry of Science and Higher Education with a scholarship for an outstanding young scientist and in 2021 he received award from the Polish Ministry of Science and Higher Education for research achievements. In 2020 M. Woźniak was presented among "TOP 2Dr". Woźniak was the Editorial Board member or an Editor for Sensors, IEEE ACCESS, Measurement, Frontiers in Human Neuroscience, PeerJ CS, International Journal of Distributed Sensor Networks, Computational Intelligence and Neuroscience, Journal of Universal Computer Science, etc., and a Session Chair at various international conferences and symposiums, including IEEE Symposium Series on Computational Intelligence, IEEE Congress on Evolutionary Computation, etc.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/