An ensemble deep learning approach to speech emotion recognition

Davis Munachimso Agughalam
School of computing
National college of Ireland
Dublin, Ireland
x19143354@student.ncirl.ie

Adelola Omonike Adebo
School of computing
National college of Ireland
Dublin, Ireland
x19134711@student.ncirl.ie

Adebola Olubunmi Abdullahi-Attah
School of computing
National college of Ireland
Dublin, Ireland
x19119283@student.ncirl.ie

Ridwan Dolapo Atanda
School of computing
National college of Ireland
Dublin, Ireland
x19142366@student.ncirl.ie

***Abstract*—Speech emotion recognition is one area of application in which deep neural networks have excelled. Most of the works done in this domain have used single learners. This work proposes a novel approach using an ensemble of binary classifiers to simplify the multiclass classification problem into a binary classification problem aiming to improve overall model performance. The binary classifiers are ensembled using a multilayer perceptron to obtain their final predictions on the multiclass classification problem. The effectiveness of this approach is validated using a benchmark dataset for speech emotion recognition. Based on the results obtained from the experiments, this approach outperforms state-of-the-art models with an accuracy of 98.8%.**

***Index Terms*—deep neural networks, speech, emotion, recognition, ensembles**

I. Introduction

Human-to-computer interactions are substantially growing across different sectors. These interactions sometimes include speech between humans and machines, as seen in speech recognition systems and digital assistants. Detecting human emotions can improve human-to-computer interactions and redefine user experience. Human emotions can be conveyed through verbal communication, for example, tone of voice or speech, or through non-verbal communication such as facial expressions or body language. Whereas facial expressions and body language can easily be misinterpreted, particularly for computer systems, verbal communication is a fast and efficient means of conveying emotions hence is a very active area of research [1]. An example of an area of application for speech emotion recognition systems is seen in healthcare where emotionally aware computer systems can be harnessed for patient emotional state monitoring and on-time development of appropriate forms of intervention.

Various theories on the categorization of human emotional states have been put forward by different researchers. While other researchers have categorized emotional states into diverse categories [2, 3, 4], the universal category of emotions discussed in [1] includes surprise, anger, happiness, fear, sadness and neutral. Selecting relevant speech features capable of portraying the inherent emotions is one of the main challenges faced in the area of speech emotion recognition research. Spectral features such as Mel's cepstral coefficient (MFCC) and log mel spectrograms have been found to yield satisfactory results. Different researchers have also investigated other acoustic and prosodic handcrafted features, such as pitch, using traditional machine learning classification techniques such as support vector machines (SVM). However, there is no consensus on the best features to identify speech emotions.

In recent years, deep learning approaches have become popular in various areas of application such as computer vision and natural language processing due to its ability to learn high-level features from raw data and even surpassing traditional machine learning approaches. Researchers have, therefore, begun to apply deep learning to speech emotion recognition research. Some of the adopted approaches include using the audio data as direct input into the deep learning model and empirically choosing the features to best represent the inherent emotions in the data. This approach, however, faces some bottlenecks as these features are not totally capable of representing all the information required by the model for optimum performance hence it is important to develop models able to extract high-level discriminatory features.

Convolutional Neural Networks (CNN) have been extensively deployed for computer vision tasks due to its ability to process spatial inputs and extract relevant features for proposed classification through a network of interconnected convolution and max pooling layers. For analyzing sequential data, long short-term memory (LSTM) networks have mostly been adopted. To extract more relevant and fully representative information to improve model performance, a spatial representation of the audio data can be used as input into a CNN-LSTM neural network where the former learns high-level features from low-level voice signals, and the latter learns the sequential dependencies that exist between the extracted high-level features.

This CNN-LSTM approach is used in work done by Zhao et. al. [5]. They used a 1D CNN-LSTM architecture to detect emotions from raw audio while using a 2D CNN-LSTM architecture to detect emotions from a 2D log mel spectrogram representation of the audio. Issa et. al. [6] proposed a different architecture using only 1D CNN for speech emotion classification. Asserting that a wider range of features will improve the ability of the model to generalise effectively, five features, MFCC, chromagram, log mel spectrogram, contrast and tonnetz were extracted from the audio files, stacked and used for the analysis. While their research was successful for categorizing emotion from speech, they only employed single learners and this is also the approach followed in the bulk of the research in this domain.

In this study, an ensemble approach using CNN-LSTM binary classifiers trained to specifically identify one of the emotion classes, essentially breaking down the multiclass classification problem into binary classifications is adopted. These binary classifiers are ensembled using a multilayer perceptron to make final predictions. To efficiently measure the performance of this approach, a baseline CNN-LSTM model trained on all classes as a multiclass classification problem is also created for direct comparison.

The remaining sections in this paper are organized as follows. The work is related to prior research in section 2 while the proposed approach is explained in detail in section 3. The experimental findings are discussed in section 4 and the work is concluded in section 5.

II. Related Work

Several developments in speech emotion recognition (SER) research have been studied and this offers insights into the major challenges that have been at the core of work in this area. These challenges include selecting the best set of features that are capable of fully representing the underlying emotions present in the data and choosing an optimum algorithm or architecture that would better leverage these features and their combinations to identify emotions from speech[7]. Traditional machine learning techniques and deep learning techniques have both been employed for SER tasks and these approaches are reviewed as follows.

A. Traditional Machine learning techniques for SER

Support vector machines (SVM) have been largely employed for identifying emotion from speech due to their ability to identify patterns easily and efficiently. Majorly, the features used in recent research on SER using SVM models are suprasegmental features such as pitch and spectral-related features such as MFCC, which are both categorized as personalized features. This was the focus of the research done by [8]. However, the application of this approach is limited, as the training of the SER was carried out within a specific corpus. Speakers generally have different sound and voice characteristics. Such unique characteristics vary among different individuals leading to a disparity in the distribution of emotions between different speakers. To overcome this challenge, [9] research focused on the discovery of both personalized (MFCC) and non-personalized features derived from the first speech signal derivative. The research argued that both personalized and non-personalized features are essential for detecting the emotional state of the speaker as they provide in-depth information on the meaning of the frames and represent shifts in the speaker's emotions. This approach can be applied in call centres to monitor caller emotions and in autonomous vehicles.

In work done by [10] using the SVM radial kernel model, the researcher found out that the model was able to better distinguish emotions in audio recordings in a quiet and serene environment with up to 91% accuracy, compared to a noisy background with 66% accuracy. The authors also demonstrate that gender has a significant effect on the detection of emotions. However, the pre-processing, which is an essential part of the analysis that helps to improve on the quantisation signal to noise ratio, allowing the speech signal to be stable, short and pure was ignored. Improving on the above work [11,12 and 13] implemented SVM classifier with pre-processing approaches, such as pre-weighting to achieve high-frequency energy signal and enable pre-emphasis obtained from FIR high-pass filters. Research on both Gaussian nonlinear proximal kernel SVM and optimal GA SVM have also shown excellent results.

[14] suggested a hybrid approach using 40 linear SVM active learners to categorize four classes of emotions (happiness, sadness, anger, and neutral). They also adapted forward feature selection to boost classification efficiency by requiring all classifiers to select features consecutively until all features have been distributed in order to achieve the number of intended features to be expressed. The Active Learning approach selects features based on three criteria, vote entropy, uncertainty and random sampling. Results show that the method is suitable in the target domain for feature selection on a small dataset, generates diversity among the classifiers and reduces duplication of features.

Though conventional SVM methods have proven to be efficient in classifying speech emotions, however, due to the ability of deep neural networks to perform satisfactorily without the need for manual extraction of features, it has become highly recommended for emotion detection and has preference over traditional methods.

B. Deep learning for SER

Over the last years, deep learning techniques have contributed to significant breakthroughs in SER. Deep Retinal Convolutional Neural Networks (DRCNN) for SER, proposed by Niu et. al. [15] and Badshah et. al. [16] and stacked CNN and LSTM proposed by Zhao et. al. [5], have demonstrated state-of-the-art performances, which suggests that deep learning can effectively learn high-level abstractions from low-level features. In Han et. al.[17], Deep Neural Networks (DNN) was used to train segmented signals from each segment of the computed probabilities of the emotional state in which utterance level attributes are produced and input into an extreme learning machine (ELM) model. This technique helps to identify emotions at different speech rates. While there was

an increase in accuracy over traditional models, the accuracy was still relatively low.

Kerkeni et. al.[18] conducted a comparative study of SER systems on different classifiers such as the Recurrent Neural Network (RNN), SVM and Multivariate Linear Regression (MLR). Using MFCC and modulation spectral (MS) as input features for classifiers, the study showed that the RNN classifier without feature selection and speaker's normalization surpassed the other classifiers with 94% accuracy. Issa et. al.[6] also investigated the combination of five features, namely MFCC, Mel-scale spectrogram, Chromagram, Spectral and Tonnetz, as inputs for 1D CNN and evaluated on three datasets. The approach achieved state-of-the-art performance on two of the three datasets with 95.71% accuracy on the EMO-DB dataset and 86.1% accuracy on the RAVDESS dataset.

Another area of research in SER is focused on the use of multi-modal features to identify human emotions. Tripathi et. al.[19] used a combination of speech features, i.e. MFCC and Spectrogram and speech transcripts, as input features for the 2D CNN model. Speech transcripts are extracted from the audio file text sequences. The IEMOCAP dataset was used for this experiment. The 2D CNN models were trained independently of each feature and combined to achieve higher classification accuracy. The findings show that the MFCC plus text transcripts performed best with 76.1% accuracy. A similar approach was used in [20], where the authors trained RNNs with combined MFCC and speech transcripts. Their approach was also evaluated on the IEMOCAP data set and achieved 71.8% accuracy. The above works demonstrate multi-modality as a promising direction for SER research.

Other works explored the combined learning approach to SER tasks. The authors in Latif et. al.[4] proposed CNN as a feature extraction block to harness multiple high-level acoustic and spectral features and jointly train with LSTM classification, which is motivated by the fact that LSTM performs better when fed with discriminatory representations. The CNN layer captures both long-term and short-term interactions from speech signals. It outputs the most salient feature by locally aggregating the feature map of each layer, which are then fed into the LSTM classification block. The approach achieved a state-of-the-art result. [21] improved on the approach taken by [4]. The authors extracted Mel-spectrogram features from the audio signal and fed into a CNN, then fused the output from this block into an Extreme Learning machine and finally classified five basic emotions (Happy, Relaxed, Disgust, Sad and Neutral) using SVM. The experiment performed significantly increases emotion accuracy than the research done in [4].

In Zhang et al.[22] SER was investigated on the acted emotional speech dynamic database (AESDD) using spatial-temporal recurrent neural network (STRNN) model. They captured the co-occurring variations in human emotions by using a multi STRNN and bidirectional STRNN. The extracted vector of the frame of each speech signal serves as an input to the bidirectional STRNN, and the bidirectional STRNN is further used to learn discriminative features that eventually classify emotion into six states. Their experimental result was argued to outperform state of the art models. Similar works that have used ensemble methods to improve the performance of SER are [23] and [24].

Motivated by the above works, our research takes a different approach by breaking down the problem of multiclass classification into a binary classification for each of the emotion classes. A CNN-LSTM network is proposed as a binary classifier to classify the emotion categories using a log mel spectrogram of the audio as input. Then the binary classifiers are ensembled using a multilayer perceptron to make a final prediction. This is expected to improve classification accuracy.

III. Methodology

Ensemble approaches to deep learning tasks often outperform single learners as different diverse models can be combined to complement each other's weaknesses thereby creating more robust classifiers [25]. This study employs the ensemble approach to classify speech emotions into seven categories sad, angry, happy, disgust, surprise, neutral and fear using an ensemble of CNN-LSTM binary classifiers trained to identify each of the emotions. The binary models are then ensembled using a multilayer perceptron for the final multiclass classification. A baseline model using a CNN-LSTM network on the direct multiclass classification is also created to effectively measure the performance of this approach with direct comparison. Fig. 1 shows the proposed approach. The Knowledge discovery in databases (KDD) approach was followed to achieve the objective of this study.

A. Data Collection

The Toronto Emotional Speech Set (TESS)[1] is used to validate the feasibility of our proposed approach. It consists of 200 words spoken by two female actors who convey seven different emotions, namely anger, fear, sadness, neutral, pleasant surprise and happiness. The data set contains a total of 2800 speech recordings and is a standard dataset for SER research.

B. Data Exploration, Pre-processing and Transformation

At this stage, the audio recordings are processed into a suitable format to be understood by deep neural networks and explored. The waveforms for two emotions fear and neutral are shown in Figs. 2 and 3 respectively. It can be seen that the neutral emotion waveform has a more even spread compared to the fear emotion where there are sections that are clearly larger than the rest. The difference in their waveforms suggest that emotions indeed can be conveyed in speech as different emotions have different waveforms. Their log mel spectrograms are also shown in Fig. 4 and 5 and it is seen that there is a difference for the two emotions. Feature extraction is an important step in the machine learning project workflow. This is especially the case in SER research as extracting features that can fully convey the underlying speech emotion is important to the success of the model. The log mel spectrogram is a widely used feature in SER research and is the

[1]https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess

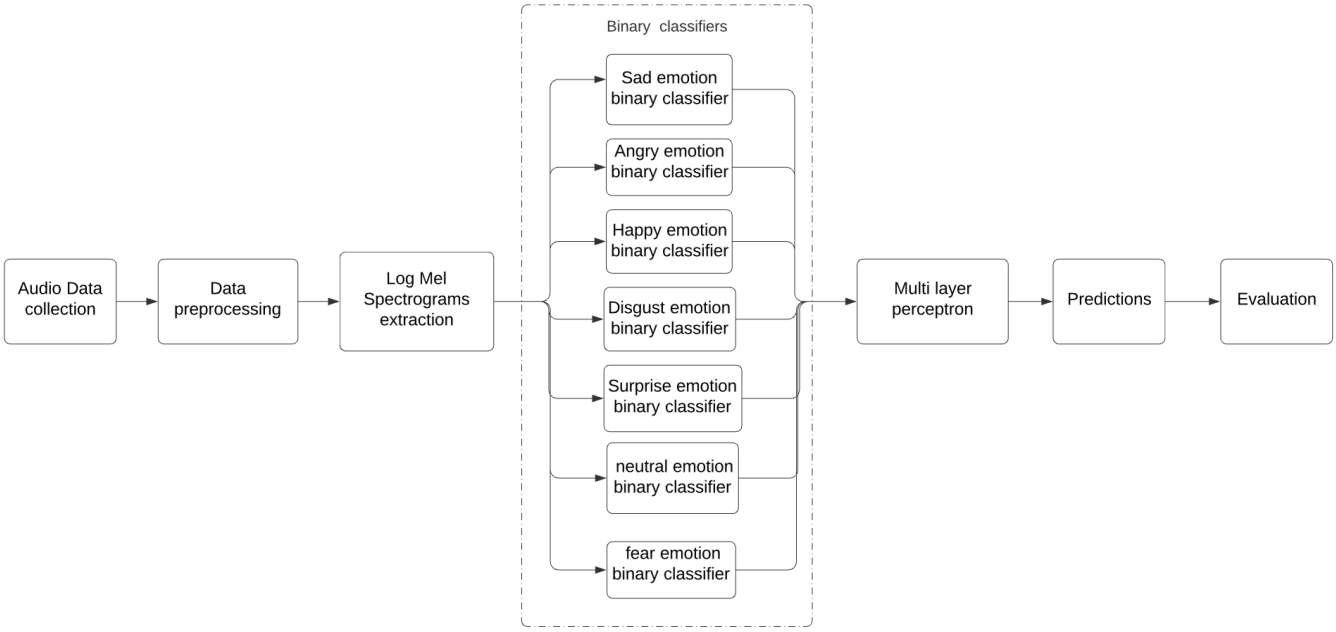

Fig. 1. Proposed approach framework

adopted feature in this study as it mimics the human reception to sound frequency to an extent[6]. The librosa python library is used to extract the log mel spectrograms from the audio recordings.

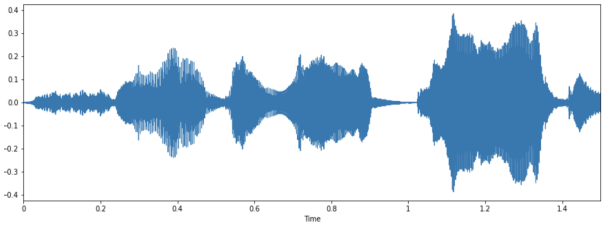

Fig. 2. Fear emotion wave form

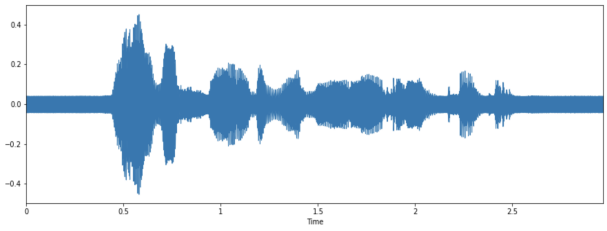

Fig. 3. Neutral emotion wave form

For CNNs, the input data needs to be of the same shape. Since the audio recordings have different lengths, this posed a challenge. Different pre-processing approaches were used to overcome this challenge and their impacts on the results were also evaluated. In the first approach, only recordings with length above 2 seconds were filtered from the dataset and their

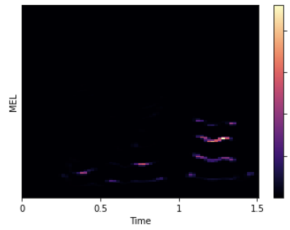

Fig. 4. Log Mel spectrogram for fear emotion

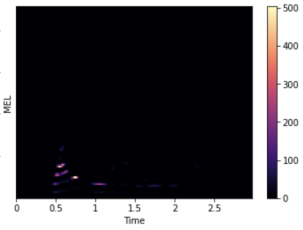

Fig. 5. Log Mel spectrogram for neutral emotion

log mel spectrograms extracted for only 2 seconds. Though this solves the problem of shape uniformity, some classes of emotions in the dataset were severely underrepresented and this led to high class imbalance. A second approach was used to address the problem of class under-representation. The log mel spectrograms were extracted for the full length of each of the recordings and the mean values were taken across each time step. This introduced missing values in the data as not all the recordings were of the same length. The missing values were filled with zero to obtain a uniform input shape for all the recordings and further analysis was done using the log mel spectrogram extracted from the full length of the audio recordings.

For the binary classifiers, the target variable for each classifier was transformed to contain only the class of emotion the model is trained for while the rest of the classes are represented as 'other'.

C. Model Architecture

The deep neural networks employed in this study are implemented using the keras library with TensorFlow backend.

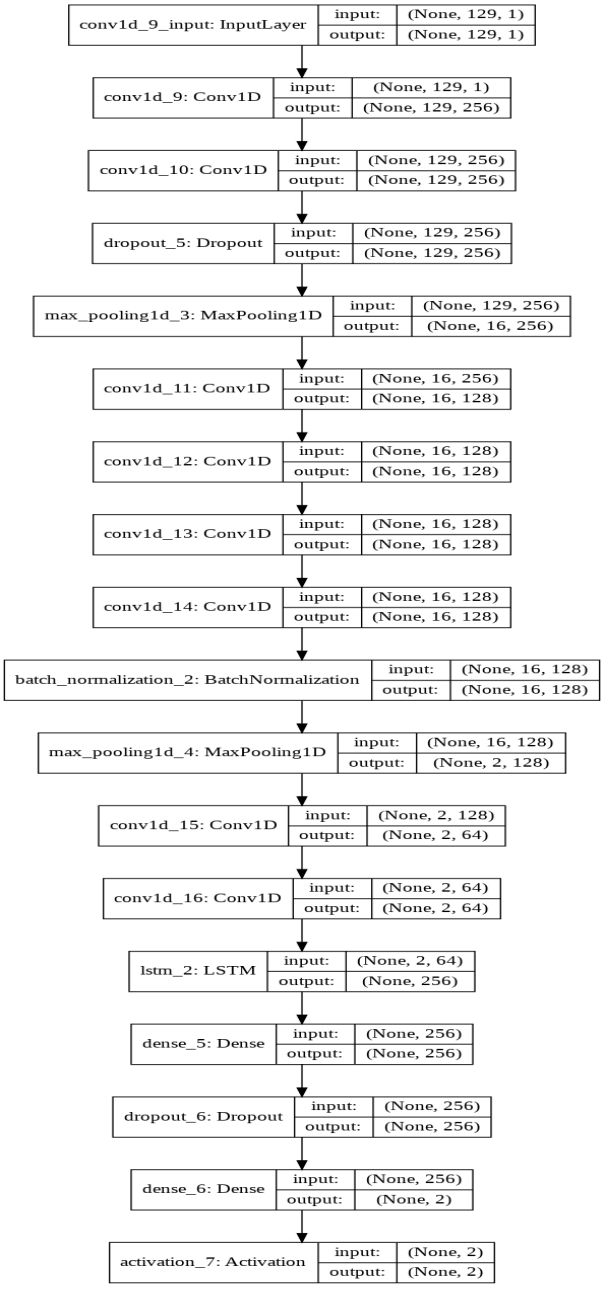

Fig. 6. CNN-LSTM model architecture

1) CNN-LSTM baseline model: In the baseline model, a CNN-LSTM architecture was used on the extracted log mel spectrograms for a multi-class classification. As considered from the literature reviewed, eight convolution layers were applied in total to facilitate the hierarchical decomposition of the input log mel spectrograms. The input convolution layer had an input shape of 129 x 1 and filter size of 256. The output from this layer is passed to two convolution layers with filter size of 256 followed by dropout with rate 0.25 and a maxpooling layer. Four more convolution layers were stacked with filter sizes of 128 before a batch normalization and maxpooling layer. The final two convolution layers followed with filter sizes of 64. An LSTM layer with 256 units was also introduced before the fully connected network. Other hyperparameters include kernel size: 8, optimizer: adam, loss: categorical crossentropy, metrics: accuracy and activation: reLu except the final output layer which used a softmax activation. Fig. 6 shows the structure of the CNN-LSTM baseline model.

2) CNN-LSTM Binary Classifier: For our proposed approach, we ensembled 7 CNN-LSTM binary classifiers. Each classifier is created to identify the emotion class for which it was trained from the audio file while denoting the other classes as 'other'. A similar architecture to the baseline model is used for each of the 7 classifiers with only minor modifications: the output layer consisted of only 2 neurons and the loss used is binary crossentropy. After binary classification, the model outputs were ensembled using a multilayer perceptron (MLP). The MLP model includes an input layer of 32 neurons and input dimension of 14 with activation function ReLU, followed by a dense layer with 32 neurons, activation function reLU, an output layer with 7 neurons, and softmax activation function. Other hyperparameters include, optimizer: adam, loss: categorical crossentropy and metrics: accuracy.

IV. Results and Discussion

To evaluate the performance of the models we adopt the widely used evaluation metric for SER, accuracy.

A. Experiment 1: 2D CNN-LSTM model with 2 seconds of data

In our first experiment, we used the first pre-processing method of extracting just 2 seconds of audio recordings to ensure input shape uniformity for the CNN model. The model had a 98% accuracy. Nevertheless, the fear emotion class was significantly under-represented, as most audio recordings were not up to 2 seconds. Furthermore, fear and angry classes were misclassified, as were surprise and joy. This could be explained by the fact that these feelings could be argued to be fundamentally identical and could easily be mistaken for each other, even by humans. Fig. 7 shows the confusion matrix for this model.

B. Experiment 2: 1D CNN-LSTM Baseline Model

For our baseline model, we used a 1D CNN-LSTM network for a multi-class classification of all emotion categories. This model achieved a 97.5% accuracy. Surprise and happy classes were both misclassified, and this may be the result of the theories suggested earlier. The surprise class was also misclassified with angry and disgust. While the baseline model with more misclassifications had a lower accuracy than the model from experiment 1, it is considered to be a better model because all the classes were fully represented leading to better generalization. The confusion matrix is shown in figure 8.

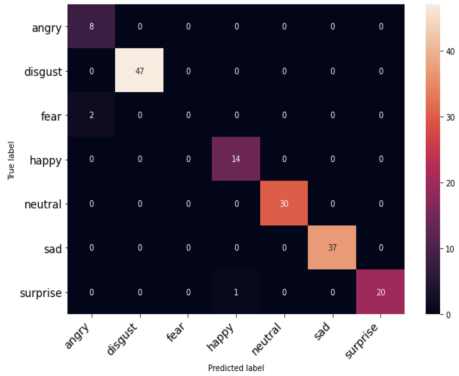

Fig. 7. Confusion matrix for model with 2 seconds of data

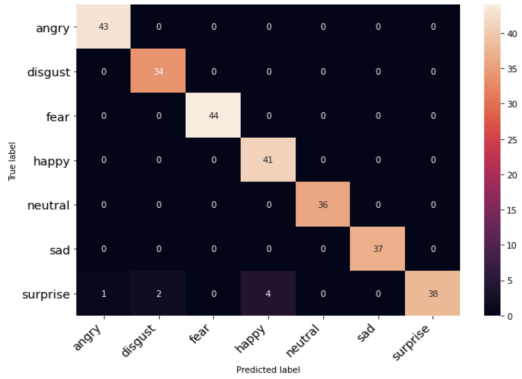

Fig. 8. Baseline model confusion matrix

C. Experiment 3: 1D CNN-LSTM binary classifiers

Table 1 shows the results of the binary classifiers trained for each of the emotion categories. The confusion matrix for each of the models are also shown in the figures below. The models performed well with high individual accuracy. The neutral model was perfect on the test dataset as it correctly classified all the classes. Fear and Angry also had high accuracy of 99.6%. The model for sad and disgust had an accuracy of 98.9% and 98.2% respectively while surprise and happy had 97.8% accuracy.

TABLE I
BINARY CLASSIFICATION RESULTS

Binary classifier	Accuracy%
Fear	99.6
Surprise	97.8
Sad	98.9
Angry	99.6
Disgust	98.2
Happy	97.8
Neutral	100

After ensembling the 7 binary classifiers for the multiclass classification using a multilayer perceptron, the combined accuracy was 98.8%. Though there was a drop in from the individual accuracy of the models, the approach still outperformed the baseline model with an accuracy of 97.5%.

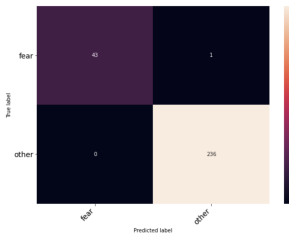

Fig. 9. Fear emotion classifier confusion matrix

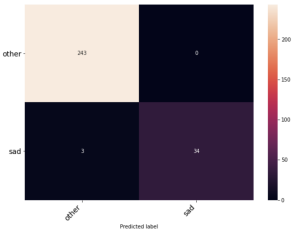

Fig. 10. Sad emotion classifier confusion matrix

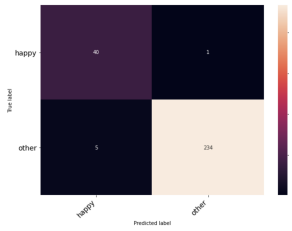

Fig. 11. Happy emotion classifier confusion matrix

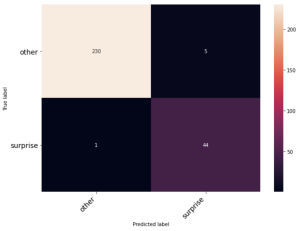

Fig. 12. Surprise emotion classifier confusion matrix

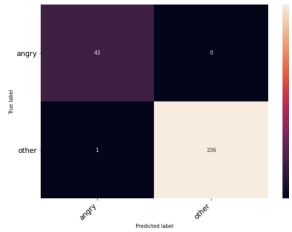

Fig. 13. Angry emotion classifier confusion matrix

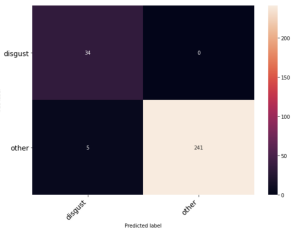

Fig. 14. Disgust emotion classifier confusion matrix

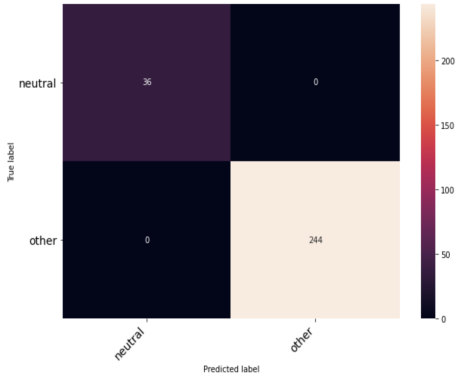

Fig. 15. Neutral emotion classifier confusion matrix

These results further validate our approach and hypothesis that ensemble techniques can outperform single learners and are efficient for speech emotion recognition tasks.

The results of our approach outperform the 95.71% accuracy reported in [6] using an ensemble on another benchmark dataset. One of the reasons for this improvement can be attributed to the method used to ensemble the binary classifiers. While [6] used a rule based decision system to ensemble the binary learners from best performing to least performing, we trained a multilayer perceptron on the outputs of the binary classifiers to better understand the intrinsic relationships between their outputs and thus putting forward a better result than a rule based system.

V. Conclusion

The field of machine learning and artificial intelligence is rapidly expanding and driving innovations in human to computer interaction (HCI) systems such as autonomous cars and facial recognition systems. Harvesting emotions from speech can be used to improve user experience in these HCI systems as the computer can learn to adapt to the current emotional state of the user and perform accordingly. Several researchers in this area have adopted deep neural networks over traditional machine learning methods as deep neural networks have to ability to learn features required for the emotion detection with little or no human effort. Though their works were often successful, they mostly only used single learners.

In this paper, we explore a different approach for speech emotion recognition by using an ensemble approach. CNN-LSTM binary classifiers were created for each of the seven emotion classes to be identified and their results ensembled using a multilayer perceptron for the final predictions. This approach was effected using the TESS dataset. To achieve this, three different experiments were performed. In the first experiment, we trained a 2D CNN-LSTM network using log mel spectrograms extracted for only 2 seconds of the recordings in other to obtain a uniform input for the CNN. Though this approach had a high accuracy, the fear emotion class was severely underrepresented thereby impacting negatively on the model's ability to generalize. To tackle this, a second experiment was performed. In this experiment, we trained a 1D CNN-LSTM network on the log mel spectrograms extracted for the full length of the recordings. The mean was taken at each time step and the shorter recordings were padded to obtain a uniform input for the CNN. This approach was set as the baseline model as it achieved satisfactory results. In the third experiment, which is the proposed approach, we trained seven binary classifiers using the same architecture as the baseline model and ensembled them using a multilayer perceptron. The results from the ensemble approach outperformed the other experiments with high accuracy and its ability to generalise well as the model was exposed to all the emotion classes to an equal extent. These results further validate the ability of ensemble learners to outperform single learners. The performance of the model could be further evaluated using a different dataset.

For future work, harnessing multimodal audio-visual data to explore speech emotion recognition could be investigated. Advanced data augmentation techniques using generative adversarial networks (GAN) to expand the amount of training data and further improve the performance of the model on unseen data can also be investigated.

References

[1] S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20,no. 6, pp.1576-1590, Oct. 2017. doi:10.1109/TMM.2017.2766843

[2] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM," *in IEEE Transactions on Affective Computing, China, October 2019*, pp. 190-202. doi:10.1109/TAFFC.2019.2947464

[3] K.Y. Huang, C.H. Wu, Q.B. Hong, M.H. Su, and Y.H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," *in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), China, May 2019*, pp. 5866-5870. doi:10.1109/ICASSP.2019.8682283

[4] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. and Epps, "Direct modelling of speech emotion from raw speech," *in Electrical Engineering and Systems Science - Audio and Speech Processing, New south wales, Australia, Jul 2019*, pp. 21-25.

[5] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312-323, Jun. 2020. doi: 10.1016/j.bspc.2018.08.035.

[6] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural network," *Biomedical Signal Processing and Control*, vol. 59, May 2020. doi: 10.1016/j.bspc.2020.101894.

[7] J. Hook, F. Noroozi, O. Toygar, and G. Anbarjafari, "Automatic speech based emotion recognition using paralinguistics features", *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol.67, no 3, pp. 479-486, Mar. 2020. doi: 10.24425/bpasts.2019.129647

[8] A. Rajasekhar and M. K. Hota, "A Study of Speech, Speaker and Emotion Recognition Using Mel Frequency Cepstrum Coefficients and Support Vector Machines," *in 2018 Int. Conf. on Communication and Signal Processing (ICCSP), Chennai, India, April 3-5, 2018*, pp. 0114-0118. doi:10.1109/ICCSP.2018.8524451

[9] W. Cao, J. Xu and Z. Liu, "Speaker-independent speech emotion recognition based on random forest feature selection algorithm," *in 2017 36th Chinese Control Conference (CCC), Dalian, China , July 26-28 ,2017*, pp. 10995-10998. doi:10.23919/ChiCC.2017.8029112

[10] M. Jain, S. Narayan, P. Balaji, A. Bhowmick and R.K, Muthu "Speech emotion recognition using support vector machine", *Electrical Engineering and Systems Science - Audio and Speech Processing*, arXiv preprint arXiv:2002.07590. Feb. 3, 2020.

[11] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," *in 2016 Int. Conf. on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, September 9-10, 2016*, pp. 1080-1084. doi:10.1109/ICACDOT.2016.7877753

[12] Y. Wang and H. Huo, "Speech Recognition Based on Genetic Algorithm Optimized Support Vector Machine," *in 2019 6th Int. Conf. on Systems and Informatics (ICSAI), Shanghai, China, November 2-4, 2019*, pp. 439-444. doi:10.1109/ICSAI48974.2019.9010502

[13] Z. Han and J. Wang, "Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine," *in 2017 Chinese Automation Congress (CAC), Jinan, China, October 20-22, 2017*, pp. 2513-2516. doi:10.1109/CAC.2017.8243198

[14] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," *in 2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, March 5-9, 2017*, pp. 5000-5004. doi: 10.1109/ICASSP.2017.7953108

[15] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, "A breathrough in speech emotion recognition using deep retinal convolution neural network", *Computer Science*, arXiv preprint arXiv:1707.09917. Jul. 12, 2017.

[16] 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), Busan, South Korea, Feb. 13-15, 2017, pp. 1-5. doi:10.1109/PlatCon.2017.7883728

[17] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", *in Fifteenth annual conference of the international speech communication association, Singapore, September 14-18, 2014*, pp. 223-227. [Online]. Available:https://www.isca-speech.org/archive/interspeech2014/i140223.html [Accessed on: Mar. 2, 2020]

[18] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M.A. Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning," *in Social Media and Machine Learning, IntechOpen, Mar. 2019*, pp. 247-250. doi: 10.5772/intechopen.84856.

[19] S. Tripathi, A. Kuamr, A. Ramesh, C. Singh, and P. Yenigalla, "Deep Learning based Emotion Recognition System using Speech features and transcriptions,"*Electrical Engineering and Systems Science - Audio and Speech Processing*, arXiv preprint arXiv:1906.05681, Jun. 11, 2019.

[20] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," *in 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, Dec. 2018*, pp. 112-118. doi: 10.1109/SLT.2018.8639583.

[21] M.S. Hossain, and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data" *Information Fusion*, vol. 49, pp. 69-78, Sept. 2019. doi:10.1016/j.inffus.2018.09.008

[22] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition" *IEEE transactions on cybernetics*, vol. 49, no 3, pp.839-847, Jan. 30, 2018. doi:10.1109/TCYB.2017.2788081

[23] D. Luo, Y. Zou and D. Huang, "Speech emotion recognition via ensembling neural networks," *in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, December 12-15, 2017*, pp. 1351-1355. doi:10.1109/APSIPA.2017.8282242

[24] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Continuous Speech Emotion Recognition with Convolutional Neural Networks" *Journal of the Audio Engineering Society*, vol. 68, no. 2, pp.14-24. Jan. 24, 2020. doi:10.17743/jaes.2019.0043

[25] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY: Manning Publications Co., 2017.