Generalizing Speech Emotion Recognition: A Multimodal and Multidataset Approach

*ENGR 5510G: Foundations of Software Engineering

Vallika Kasibhatla
Master's of Engineering in Engineering Management
100928820
vallika.kasibhatla@ontariotechu.net

Sweta Patel
Master's of Engineering in Engineering Management
100915164
sweta.patel@ontariotechu.net

***Abstract*— In recent years, detecting emotions from speech has been given significant attention owing to its potential application in various fields like human-computer interaction, mental health assessment and customer service. Speech emotion detection involves the automation of recognizing various prominent emotions like happiness, disgust, sadness, surprise, and neutrality from the audio data, leveraging the use of Artificial Intelligence. The major challenge in this task is to capture the diversity of the audio data arising from speaker diversity, audio recording environments and audio quality, leading to a lack of generalization in the models, hindering their performance across different contexts and populations. This paper focuses on solving this problem by combining various datasets and coming up with a multimodal approach involving simple yet powerful traditional machine learning algorithms through ensemble learning, by using data augmentation and various feature extraction steps. There has been a comparison made between simple classification models and neural network models. The main focus has been on ensemble learning of SVM and RF, which resulted in the best accuracy of 81%. Thus, this research aims to improve the generalizability and reliability of the models, capturing a more comprehensive representation of every emotion.**

Keywords—Speech Emotion Recognition (SER), Ensemble Learning, Support Vector Machines, Random Forest, Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), MFCCs, Data Augmentation, Boosting.

I. Introduction

Emotion detection has become an emerging research subject, and finding software engineering projects using emotion detection has become attractive. There is an immense potential for ideas and creative opportunities in this field. As human emotions are challenging to analyse and guess straightforwardly, deep learning techniques have been used to achieve this goal. The issue's complexity resides in the idea that emotions are unpredictable and subtle for even humans to understand. However, if computers and AI can understand the sentiment of humans, then it can improve productivity and efficiency in day-to-day activities. The need for real-time feedback in these emotion-aware algorithms can benefit user experiences. Emotion detection could be done through various inputs such as text, speech, and facial expressions; these inputs have complex challenges to analyse and recognize. Much research has been done on these topics, especially for emotional detection from speech. The breakthrough of grasping the features using tone, pitch, speaker, amplitude, and speed rate of an audio wave helped produce sophisticated technology and algorithms to differentiate if a person is feeling sad, happy or angry. Each emotion has its widespread features that are unique from one another. Researchers found these features to lead to various real software projects for sentiment analysis for machine learning and natural language processing utilizing vocal qualities about voice.

Emotion detection technology has a wide range of real-world applications, each with its own specific use cases. For example, in the medical industry, this technology can provide remote patient monitoring, which can result in faster and more effective healthcare treatments. It can improve the customer experience in customer service by providing prompt and appropriate reactions to the user's emotions. Moreover, it may greatly ease daily living in the context of smart assistants (loot homes).

It might be difficult to infer a speaker's emotional state from their words for a variety of reasons. The primary problem is the difficulty of generalizing processing stages due to the diversity of sentences, tone, accents, speakers, and speaking styles, as well as the difficulty of distinguishing elements for certain emotions. Determining distinct emotional borders is further complicated by overlapping emotions within the same statement. While numerous SER research projects have used various speech characteristics and machine learning techniques, they frequently need to specify the techniques for feature extraction, emotion categorization, and data pre-processing [1]. As, there is more reliance on datasets for modelling, finding the data the model will be trained in is very important, the readily available datasets only focus on having speakers of English descent, can create a bias while training and testing. However, they have only used single datasets. Additionally, mixed emotions or rapid emotions switching aren't detected. As the dataset consists of a single speaker, it may not detect right if a test case has multiple speakers. Papers have highlighted that certain emotions such as fear, surprise, or disgust are complex to differentiate from anger and happiness due to taking only certain features such as tone or pitch. The feature extraction steps are created only for single datasets. The classification models are highly focused on neural networks, maybe this process is too complicated.

II. Problem Statement

This project focuses on incorporating ensemble learning and building a generic model to cater to speaker diversity, including differences in gender, age, accents, and language. Our aim is to provide a comprehensive representation of each emotion that is exposed to a wide variety of audio qualities, recording environments, and noise levels.

III. Literature review

Papers focus on emotion detection through speech. They address the range of emotions that can be identified and the varieties of methods and techniques adopted for that purpose.

The paper by Fayek et al (.2017) [2] used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. They have performed feature extraction by calculating 40 log Mel-scale filter bank coefficients. Two different methods of classification have been conducted: feed-forward neural networks and recurrent neural networks. Under feed-forward neural networks, the researchers built a deep neural network (DNN), and convolutional neural networks (CNNs) were also evaluated. Then, under recurrent neural networks, they employed Long Short-Term Memory (LSTM). After considering all three models, it was found that ConvNet performed better than the other two models in terms of accuracy (64.78%) and unweighted average recall (UAR) (60.89%). In the paper by Hassan et al. (2019) [3], researchers go beyond the traditional feature extraction method, which results in high dimensionality and susceptibility to artifacts by designing a multimodal approach. Here, the researchers incorporate physiological signals that capture emotions from EDA(Electro-Dermal-Activity), PPG (Photoplethysmogram), and EMG (Electromyography) sensors. The DBN's hierarchical training approach effectively classified 'happy' emotion with 100% accuracy but only 53% for the 'neutral' emotion. This indicates that the non-linear separation of arousal and valence levels of neutral emotions makes it difficult to classify them. Still, this method showed better robustness and accuracy in recognizing emotions than earlier models. Kavitha et al. (2022) [4] proposed a new method of multilayer perceptron classifier for the RAVDESS dataset. MFCC, Mel spectrograms, Chroma characteristics, and Tonnetz were used to extract features that captured the audio data's pitch, tone, rhythm, and harmonic content. The model achieved 79% accuracy, outperforming an SVM classifier that only managed 72% accuracy.

The research article by Mande [5] applied classification to the TESS dataset. The extracted features also included MFCC, Tonnetz, Contrast, Chroma, and Mel spectrograms; thus, these extracted features captured important audio signal attributes. Traditional techniques in machine learning, such as K-Nearest Neighbours, Decision Trees, and Extremely Randomised Trees, were applied to emotion classification. The results have shown that the KNN-based system achieved 98% overall accuracy. Further, decision trees yielded a mean accuracy of 92%, and extra tree classifiers with 99% accuracy. Patel et al. [6] have proposed a new methodology in their research that combines traditional machine learning with advanced deep learning models and autoencoders. Mel-frequency cepstral Coefficients were extracted as features. The model's performance is evaluated using mean square error (MSE) to measure the reconstruction error. Two datasets, RAVEDESS and TESS, were tested. CNN models achieved an accuracy of 75-80% on RAVDESS and 94-96% on TESS after applying autoencoders, indicating significant improvements in accuracy. Guizzo et al. (2020) [7] address the usefulness of Convolutional Neural Networks (CNNs) for various audio-processing tasks such as transcription, source separation, audio denoising, and speech augmentation. The researchers have proposed a multi-timescale (MTS) convolution layer by evaluating four datasets to predict the emotions. This technique was evaluated with different scaling factors and incorporated into other CNN architectures, such as AlexNet, a complicated network, and a single convolutional network. This approach increased accuracy by an average of 3.78% across all datasets, with a maximum improvement of 8.04% on the RAVDESS dataset. The paper written by Chamishka et al. [8] was interested in capturing a speech to make a real-time categorical emotion prediction. The datasets used to create a model were IEMOCAP and MELD. Their model had three phases, with the first being feature extraction, Bag of Audio words (BoAW) and emotion extraction which was further subdivided into three major factors: the global context of the conversation (TGT), the speaker's state (Pt), and the emotion of the utterance (Et). These factors help understand the emotional variations and use three separate recurrent neural networks. This method reported a 60.87% weighted accuracy for the six basic emotions in the IEMOCAP dataset while resulting in statistics that weren't significant for MELD dataset. The paper by Badshah et al. [9] proposes a model that uses a CNN model on the dataset Berlinemodb over seven emotions. At first, the confusion matrix through the trained CNN resulted in a 50% below accuracy for Fear, Happy and Neutral emotions. After fine-tuning a pre-trained AlexNet model CNN, the values changed, creating a 50% below accuracy for Boredom, Disgust, Fear, and Happy. Overall, the trained CNN model achieved an accuracy of 84.3% for all the speakers on the test set.

The paper by Qayyum et al. [10] presents a unique CNN-based speech-emotion recognition system that doesn't require any preprocessing of the input data stream. The model is developed using raw speech SAVEE dataset for training, classification, and testing purposes. There were various phases: Data Preprocessing, Recursive Feature Elimination, feature-based classification Methods, Model Architecture of Deep Convolutional Neural Network, and Training CNN model. The data preprocessing step was compared with the traditional feature extraction, such as MFCC and MSF. And directly using the raw audio with CNN. This training data resulted in an accuracy of 83.61% by using CNN directly. The paper written by Huang and Bao [11] used different methods that involve Mel Frequency Cepstral Coefficients (MFCC) and Short-time Fourier Transform (STFT) vocal extraction alongside deep learning approaches like CNN. They extracted features like pitch, MFCC, formants, etc., for vocal feature extraction and fed them into classifiers like SVM and HMM. Subsequently, they trained their model on deep neural networks like CNN and LSTM while also using deep network variants like DropConnect and ResNet. For the main testing and training datasets, the authors used the RAVDESS. They also used the Toronto Emotional Speech Set (TESS) dataset for additional training and testing. The highest accuracy for the SVM was 48.11%, and for the CNN, the accuracy was 85%. The most accurate emotions and their accuracy were anger, disgust, and calm, which were 86.8%, 78% and 72%, respectively. The paper by Kumbhar et al. [12] focuses on the MFCC feature and the LSTM (long-term memory) algorithm to analyze speech emotion recognition systems. Speech features used for emotion analysis are pitch, energy, formants, linear predictor coefficients (LPC), linear frequency cepstral coefficients (LFCC), MFCC, TEO, etc. The dataset cited in this paper was RAVDESS. They used MFCC to reduce the frequency information of speech signals into a small number of coefficients, which makes it easy to compute and extract features. Based on the receiver operating characteristics, there was a loss of 67.2%, but the average accuracy for all the emotions observed on test data was 84.81%, which can be improved by optimizing ML models. In the paper written by Nancy et al., [13], their main aim is to detect emotion exhibited by a speaker through their speech, facial expressions, or combining both. In the methodology for pre-

TABLE I. LITERATURE REVIEW

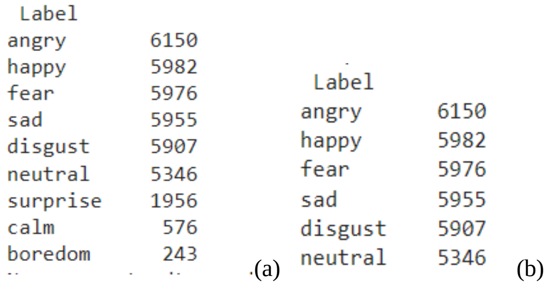
Ref	Datasets Used	Pre-Processing and Feature Extraction	Model	Results Of the Papers
[2]	IEMOCAP	MEL coefficients	DNN, CNN, LSTM	64.78%
[3]	EEG Signals	Feature fusion from Deep Belief Network and EDA, PPPG, EMG sensors.	Fine Gaussion SVM	'Happy' Emotion-100% 'Neutral' Emotion-53%
[4]	RAVEDESS	MFCCs, MEL Spectrograms, Chroma, Tonnetz	MLP, SVM	MLP-79% SVM-72%
[5]	TESS	MFCC, MEL Spectrograms, Chroma, Tonnetz	KNN, Decision Trees, Extra Tree Classifier	KNN-98% Extra Tree Classifier-99% Decision Trees-92%
[6]	RAVEDESS, TESS	MFCCs	CNN, AlexNet, ResNet50	RAVEDESS-80% with CNN TESS- 96% CNN
[7]	EMODB RAVEDESS TESS IEMOCAP	STFT, Feature extracted using Multi-Time Scale (MTS) versions of learned kernel	CNN	EMODB-70.97% RAVEDESS-55.85% TESS-53.05% IEMOCAP-55.01%
[8]	IEMOCAP, MELD	Low-Level Descriptors BoAW	RNN	IEMOCAP 60.87%
[9]	Berlinemodb	Spectrogram Generation, FFT	CNN, AlexNet model	84.3%
[10]	SAVEE	MFCC, MSF	CNN	83.61%
[11]	RAVDESS, TESS	MFCC, STFT	SVM, HMM, ELM, LSTM, CNN	CNN - 85%
[12]	RAVDESS	MFCC	LSTM	84.81%
[13]	EmoDB RED	VAD, MFCC, SBS	SVM, LDA	EmoDB:LDA: 78%, SVM: 80% RED:LDA: 71%, SVM: 73%

processing, voice activity detection was implemented, which detects speech segments while removing silences or non-speech fragments. Features such as energy, zero crossing rate, and Mel Frequency Cepstral Coefficients were extracted for short-term frames to analyze speech signals. Overall frames, a set of global statistics, such as mean, median, etc., were used to receive a feature vector for every utterance. The sequential backward selection (SBS) method combines k-fold cross-validation to select the most valuable features. They used either a pre-trained SVM model or a Linear Discriminant Analysis (LDA) classifier as a classification method to categorize the emotion detected through the feature vectors. This paper's datasets were the Berlin Database of Emotional Speech (EmoDB) and the RML Emotion Database (RED). Using the LDA classifier, there was an accuracy of 78% for EmoDB and 71% for RED. Using SVM, the accuracy was 80% and 73% for EmoDB and RED, respectively.

IV. DATASETS

In this survey, we have come across various datasets that have been used in speech emotion recognition. The most used datasets are RAVEDESS and TESS [4-7,11,12]. RAVEDESS includes audio of 24 actors, with 12 males and 12 females speaking each sentence in a North American accent. TESS is one of the simplest datasets, consisting of audio of 2 Canadian actors, one male and one female. The emotions classified from these datasets are mainly anger, happiness, neutral and sadness. The other most commonly used dataset is EMODB [13]has recordings of 5 male and 5 female German actors. These datasets often identify anger, happiness, sadness, and neutral emotions. Subsequently, we have another set of datasets, SAVEE [10], and CREMA-D which are somewhat uncommon yet frequently used in multimodal speech recognition tasks. Similar to the other datasets previously covered, they contain the same set of emotions. Each dataset follows the same layout. There are multiple audio files in .wav format and these files are labelled with the respective emotion. The audio files are generally, the actors reading the same line in the set emotion.

The dataset used in this project is a combination of these five datasets, as the majority of the emotions are similar, the data can be observed to see how many entries are under each label. The datasets' emotions are angry, happy, fear, sad, disgust, neutral, surprise, calm and boredom. However, this combined dataset leads to unbalanced data. For example, "boredom" has only 243 entries, to balance there can be under sampling and oversampling. However, duplicating or removing entries can result in overfitting or underfitting. The data are audio files which are complex on their own, for better model building, in this project "surprise", "Calm" and "boredom" are removed, while only focusing on angry, happy, fear, sad and neutral emotions.



```
Label
angry       6150
happy       5982
fear        5976
sad         5955
disgust     5907
neutral     5346
surprise    1956
calm         576
boredom      243
```

(a)

```
Label
angry       6150
happy       5982
fear        5976
sad         5955
disgust     5907
neutral     5346
```

(b)

Fig. 1. (a) all emotions, (b) Data with main emotions

V. CLASSIFICATION MODELS

A. *Support Vector Machines (SVM)*

SVM is a popular machine-learning technique used for classification and regression problems. It aims to create the best line or decision boundary to separate n-dimensional space into classes. If the data is not linearly separable, a function known as the Radial Basis Function (RBF) kernel maps the data in a higher dimension.

B. *K-nearest Neigbors (KNN)*

It is a non-parametric algorithm used to find the k-nearest data points in the training set to a specified test data point and then classify the test data point depending on the class that occurs most frequently among its K-nearest neighbors. In addition, Euclidean distance in (1) and Cosine similarity (2) are used to calculate the distance between the test data point and each training data point.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (1)$$

$$cosine\_similarity(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\cdot\|\mathbf{b}\|} \qquad (2)$$

C. *Random Forest*

This ensemble learning technique builds various decision trees with the test data. It produces the mean prediction for regression or the mode of the classes for classification problems for each individual tree. Combining the strengths and essential features of several decision trees helps in generalization and improving accuracy.

D. *Extra Tree Classifier*

It is an Extremely Randomized Trees Classifier, an ensemble learning method combining multiple decision trees to produce a result. It is used for classification and regression. Compared to decision trees, it can handle large datasets and reduce overfitting.

E. *Gradient Boosting*

The effective and scalable machine learning technique known as XGBoost (Extreme Gradient Boosting) uses the gradient boosting framework. Iteratively, it creates a series of decision trees, with each tree fixing the mistakes of the preceding one. Additionally, handling large datasets, quick results, and better performance are all attributes of Gradient Boosting.

F. *Nueral Networks*

A class of machine learning techniques called neural networks is modelled after the composition and operations of the human brain. They are made up of layers of networked nodes or neurons. Every neuron modifies its input before sending the outcome to neurons in the layer above. Neural networks are frequently employed for tasks like image recognition, regression, and classification because they can discover intricate patterns in data.

1) CNN: It is a deep-learning model focusing on spatial hierarchies of data by using the convolutional, pooling, fully connected, dropout layers with activation functions.

2) MLP: It is a feedforward artificial neural network consisting of multiple layers of nodes with each focusing on other nodes. There are main opportunities, input, hidden, and output layers, with activation functions, weights and biases.

G. *Ensemble Learning*

Ensemble learning involves the aggregation of multiple classification models to obtain a combined model that exploits the strengths of the base classifiers to produce a high-quality performance for pattern recognition by outperforming every individual model in it. The most common ways to implement ensemble learning are by bagging, boosting and stacking. In bagging, the ensemble is made up of multiple classifiers on top of bootstrap replications of the training data. Boosting involves a sequence of multiple models, each focusing on correcting the errors of the previous classifiers by weighted sampling. Stacking involves two levels of learning, one is base learning, and the other is meta-learning. The base learners are trained on the training data, whose outputs create a new dataset for a metaclassifier to classify the new instances.

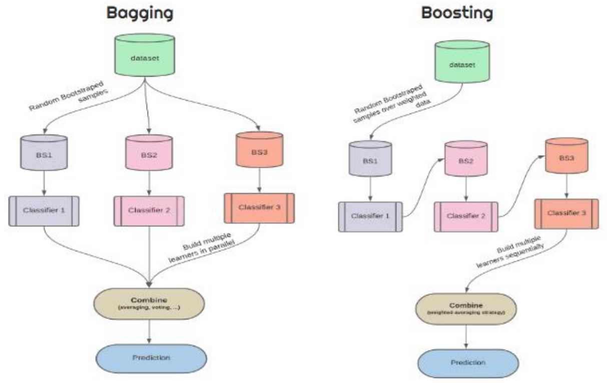

Bagging
dataset
Random Bootstraped samples
BS1
BS2
BS3
Classifier 1
Classifier 2
Classifier 3
Build multiple learners in parallel
Combine (averaging, voting, ...)
Prediction
Boosting
dataset
Random Bootstraped samples over weighted data
BS1
BS2
BS3
Classifier 1
Classifier 2
Classifier 3
Build multiple learners sequentially
Combine (weighted averaging strategy)
Prediction

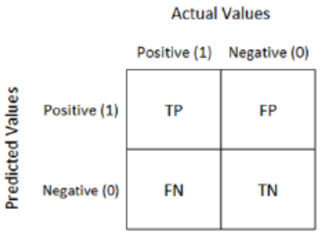
Fig.2. Bagging and Boosting

VI. BAGGING AND BOOSTING IN ENSEMBLE LEARNING.

A. *Evaluation Metrics*

These metrics can be employed to gauge the performance of the various models in our project. In the initial phase, the data is divided into training data (80%) and testing data (20%). The models are constructed using the training data and subsequently assessed using the testing data.

1) Confusion matrix: This tool evaluates the performance of a classification model with two values, 1 being positive and the other negative. Evaluation metrics such as accuracy, recall, precision, and F1-score can be calculated using Figure 3.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 3: Confusion Matrix

2) Accuracy: Equation (3) evaluates the number or percentage of correct predictions the built model makes.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

3) *Recall*: Equation (4) shows the fraction of data points correctly identified as belonging to the positive class over all those positive.

$$Recall=\frac{TP}{TP+FN} \quad (4)$$

4) *Precision*: Equation (5) shows the fraction of data points that were correctly identified as belonging to the positive class over all those predicted as positive.

$$Precision=\frac{TP}{TP+FP} \quad (5)$$

5) *F1-Score*: This evaluation metric considers the model's recall and precision, as shown in Equation 6.

$$F1\text{-}Score=2*\frac{FPR*Recall}{Precision+Recall} \quad (6)$$

VII. Data Augmentation and Feature Extraction

The main aim is to increase the generalizability of emotion detection from speech and find the apt feature extraction steps. The research mainly focused on feature extraction but not on data augmentation. We use data augmentation because it helps mimic real-life situations. For example, in the environment, there is background noise, and variations in accent, pitch, speed, and speaking style can affect the quality of speech.

Augmentation simulates these variations, helping models generalize better across different speakers and making them more robust to similar conditions during inference. Additionally, it can be used to extract other less prominent features invariant to noise, distortions, and speaker variations, leading to more reliable emotion detection. A few common Data augmentation techniques are:

1) *Noise Addition:* Noise is added to audio data to make it more generalizable to prevent overfitting.

2) *Time Stretching:* By changing the speed of the audio without altering its pitch.

3) *Shifting:* Understanding variations in the temporal alignment of audio events.

4) *Pitch Shifting:* To alter the pitch of the audio without changing its tempo.

This project mainly focuses on Noise addition and Time stretching for data augmentation. The audio is now replicated three times, one focuses on original data, one with noise addition and the other with time stretching. This changes the audio file and each file is analyzed for feature extraction of Zero Crossing Rate (ZCR), Chroma Feature, Mel-Frequency Cepstral Coefficients (MFCC), Root Mean Square Energy (RMSE), Mel-Spectrogram, Spectral Rolloff, Centroid, Contrast, Bandwidth and Tonnetz. These will result in numerical features with the labels. These feature extraction techniques are needed to understand and differentiate the audio note from other audio data. Irrespective of the emotion, these elements give a generalized analysis of audio.

1) *Zero Crossing Rate (ZCR):* It helps to identify the noisiness or percussive elements of the audio.

2) *Chroma Feature:* Captures the harmonic and melodic structures.

3) *Mel-frequency Cepstral Coefficients (MFCC):* Represent the short-term power spectrum of sound and are widely used in speech and audio processing to capture the timbral texture of the audio.

4) *Root Mean Square Energy (RMSE):* Measures the signal's loudness.

5) *Mel-Spectrogram:* It captures the overall spectral characteristics.

6) *Spectral Rolloff:* Describes the shape of the audio spectrum and differentiates harmonic and percussive sounds.

7) *Centroid:* It is useful for different types of audio textures.

8) *Spectral Contrast:* It helps to find the difference between harmonic and noisy sounds.

9) *Spectral Bandwidth:* It helps to understand the frequency dispersion of the signal.

10) *Tonnetz:* Understand the tonal relations of pitches to reflect human perception of harmony.

VIII. Methodology

A. Use of simple classification models on the combined dataset.

In this section, we present the performance of various traditional classification models like Support vector machine (SVM), Random forests (RF), Gradient Boosting, K-nearest neighbours and Extra tree classifiers, which are frequently used for speech emotion recognition tasks. All these models are evaluated using standard metrics such as accuracy, precision, recall and F1 scores. We found out that out of all the algorithms, SVM with appropriate hyperparameter tuning and Random Forest gave the highest accuracies of 75.35% and 72.18%, respectively. Moreover, on further analysis of their results through confusion matrices, we found that SVM performed better in classifying emotions like disgust, happiness, neutrality and fear, whereas Random Forest performed better in classifying anger and sadness. The following table shows a summary of the different models used along with their respective accuracies.

TABLE II. SIMPLE CLASSIFICATION MODEL COMPARISIONS

	Model	*Hyperparameters*	*Results*
1.	SVM	Kernel="rbf", C=10, gamma=100	75.35%
2.	RF	n_estimators=100	72.18%
3.	Gradient Boosting	learning_rate=0.05	56.92%
4.	KNN	K=1	75%
5.	Extra Tree Classifier	n_estimators=100	69.21%

B. Further Analysis

Owing to the analysis, we thought of enhancing the performances of the individual models, SVM and Random Forest, by combining their results through ensemble learning to better classify all emotions. We implemented two approaches for ensemble learning, bagging, and boosting. Considering the strength of excelling in the high-dimensional space of SVM and robustness to overfitting of Random Forest, along with their superior performances as individuals, we thought of using them as the base classifiers for ensemble learning.

1) Boosted approach of SVM and Random Forest

In this method, we started by training the Random Forest model on the training dataset. Then we evaluated the model by performing predictions on the training dataset and identified misclassified samples. In the next step, we increased the weights of those misclassified samples by assigning higher weights, emphasizing their importance for their subsequent training by the SVM classifier. Then both SVM and Random Forest models are used to perform predictions on the test data. Finally, in the last step, we combined the probabilities of the predicted classes from both models by taking a simple average, assuming their equal importance, to make the final decision.

2) Deep learning approaches

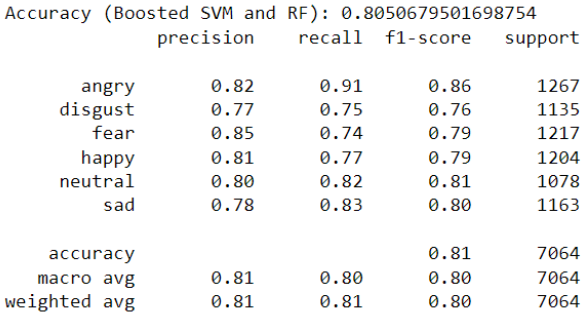
Deep learning has been significantly adopted for speech emotion recognition tasks owing to its automated feature extraction and robustness to data variability. Hence, we built two deep learning models, namely the Multilayer Perceptron (MLP) and the Convolutional Neural Network (CNN). The idea was to evaluate the efficiency of deep learning to capture the versatility of the audio data.

a. MLP: The model consists of an input layer, followed by seven dense layers with 2024, 1024, 512, 256, 128, 64, and 32 neurons, each using the ReLU activation function. In order to avoid overfitting, dropout layers with a 0.1 rate are added between the dense layers. The output layer consists of SoftMax activation to classify the six different emotions. The model is compiled with the Adam optimizer and sparse categorical cross-entropy and trained with early stooping to monitor for validation loss.

b. CNN: It begins with an input layer for data shaped (None, 178, 1). The first layer is a dense layer projecting the input to 256 dimensions. This is followed by a 1D convolutional layer with 16 filters and an activation layer. Another convolutional layer with 16 filters is the activation output. A second convolutional layer with 16 filters and a dense layer output. The outputs are then added, creating a residual connection.

IX. RESULTS AND DISCUSSION

In this section, we cover the results and performance of the models built and tested. All the models were evaluated with key matrices: accuracy, precision, recall and F1-score. We inferred that simpler algorithms like Support Vector Machines (SVM) and Random Forest gave reasonable results individually. The boosted approach further enhanced the overall performance by leveraging the strengths of both models, reducing the number of misclassifications for each and every emotion. Below is the classification report for our model. As we can see, all the emotions were classified within a range, indicating no bias towards any class. Unlike in other papers, mellow emotions like neutral, fear, and sadness, which are difficult to distinguish, showed decent F1 scores.



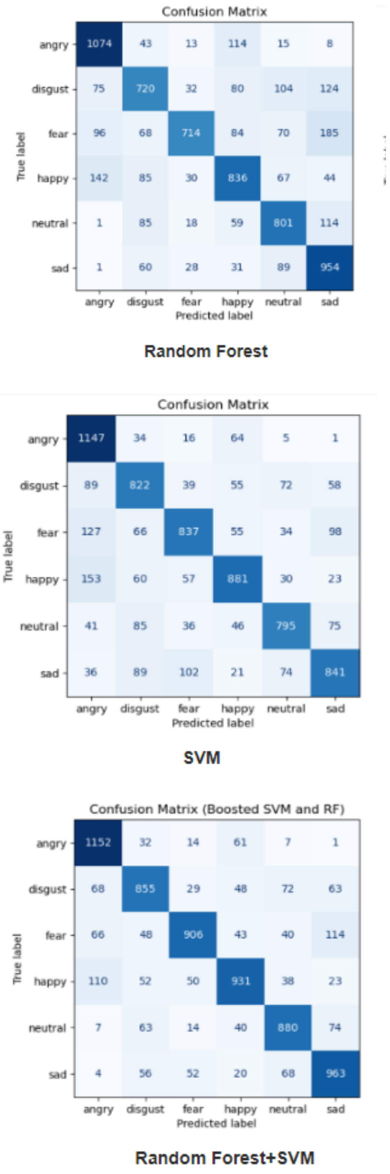
```
Accuracy (Boosted SVM and RF): 0.8050679501698754
              precision    recall  f1-score   support

       angry       0.82      0.91      0.86      1267
     disgust       0.77      0.75      0.76      1135
        fear       0.85      0.74      0.79      1217
       happy       0.81      0.77      0.79      1204
     neutral       0.80      0.82      0.81      1078
         sad       0.78      0.83      0.80      1163

    accuracy                           0.81      7064
   macro avg       0.81      0.80      0.80      7064
weighted avg       0.81      0.81      0.80      7064
```

Fig 4 Classification Report for SVM and Random Forest

Below are the confusion matrices for SVM, Random Forest and the boosted approach. The comprehensive feature extraction method capturing the acoustic, temporal, cepstral and time-domain aspects of the audio data was able to extract the required set of meaningful data, leading to the effectiveness of the proposed approach.



Random Forest

SVM

Random Forest+SVM

Fig 5: Confusion Matrix of the models

In table III, we saw that traditional models outperformed deep learning models like MLP and CNN. Thus, this signifies the potential of a simplified model architecture with careful hyperparameter tuning to maintain decent performance. Below is a summary of the models along with their respective accuracies with the comparison with the ensemble model.

TABLE III. MODEL COMPARISIONS

	Model	***Hyperparameters***	***Results***
1	CNN	Input Layer, Dense layer (512), convolutional layer (16), activation layer, and convolutional layer (16).	76%
2	MLP	8 hidden layers, dropout layers, activation function=" relu", output layer="SoftMax".	60.23%
3	Boosted SVM+RF	SVM- Kernel=" rbf", C=10, gamma=100 RF- n_estimators=100	81%

X. FUTURE WORKS

There could be more evaluations and trial and error to make models generalizable in the future. This can be done by finding more datasets in different languages; most datasets are in English and spoken with an American accent. This can be rectified by generating datasets with various accents, age groups and emotions. Additionally, more data augmentation techniques, such as pitch changing and time stretching, can be used. The feature extraction methods could be tailored or manipulated to understand each emotion specifically rather than just an audio note. The main focus was simple classification models with a basic comparison to neural network models. To enhance the performance of these models, more extensive hyperparameter tuning can be employed to develop better and more complex models. Apart from just basic emotions and producing one emotion as output, they could show a mix of emotions. Combining emotions can make more complex and realistic emotional states. If there is some probability of a combination of sadness and fear, the emotion could show embarrassment and if happiness is combined with surprise as delight. Additionally, the intensity can be shown as low, medium, and high levels. To achieve a holistic understanding of emotions, additional inputs such as facial expressions with audio can be integrated to detect accurate emotions.

XI. THREATS TO VALIDITY

There are threats to the validity of this approach.

1) Internal threats- If the data used to train the model is not a good sample of the population as a whole, then there exists a selection bias that decreases the generalizability of the model. Overfitting is another serious problem, where the model learns noise along with the actual patterns in data and can not perform well on unseen test data. The labelling errors that come from the mistakes and subjective biases of the people who do annotation of data, intrinsically decrease the accuracy of the model. While data augmentation is done to introduce more data, may at times, introduce unrealistic variations that can mislead the model.

2) External Threats-The information in the training data may not capture all the variations of the real-world scenarios. The recording quality and speakers' health can also introduce noise and biases.

XII. CONCLUSION

In this paper, various approaches to speech-emotion recognition (SER) systems were discussed. Through the literature review, 12 pages were analyzed, which focused on different models and datasets. The main issue found through the research was the lack of focus on generalizability across gender, age and language. So, the approach was to improve this issue by using multi-model and multi-dataset steps. Additionally, performance analysis across various datasets, features, recognition rates, and classifiers was focused on. Five datasets were combined; data augmentation and feature extraction steps were used. A comparison was made between simple and neural network classification models. Through ensemble learning, combining the strengths of SVM and RF, an accuracy of 81% was achieved. Additionally, around 80% accuracy was observed in every emotion classification through the confusion matrix. However, in the future, more exploration will be conducted to find datasets in different languages, and other data augmentation steps and hyperparameter tuning can be introduced to improve generalizability.

REFERENCES

[1] Madanian, , S., Chen, , T., Adeleye, , O., & Templeton, J. M. (2023, August 14). Speech emotion recognition using machine learning - A systematic review. Intelligent Systems with Applications. https://www.sciencedirect.com/science/article/pii/S2667305323000911

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

[3] Hassan, M. M., Alam, Md. G. R., Uddin, Md. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. Information Fusion, 51, 10–18. https://doi.org/10.1016/j.inffus.2018.10.009

[4] Kavitha, M., Sasivardhan, B., Deepak, P. M., & Kalyani, M. (2022). Deep Learning based Audio Processing Speech Emotion Detection. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 1093–1098. https://doi.org/10.1109/ICECA55336.2022.10009064

[5] Mande, A. A. (2019). EMOTION DETECTION USING AUDIO DATA SAMPLES. International Journal of Advanced Research in Computer Science, 10(6), 13–20. https://doi.org/10.26483/ijarcs.v10i6.6489

[6] Patel, N., Patel, S., & Mankad, S. H. (2022). Impact of autoencoder based compact representation on emotion detection from audio. Journal of Ambient Intelligence and Humanized Computing, 13(2), 867–885.

[7] E. Guizzo, T. Weyde and J. B. Leveson, "Multi-Time-Scale Convolution for Emotion Recognition from Speech Audio Signals," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6489-6493, doi: 10.1109/ICASSP40776.2020.9053727.

[8] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., Silva, D. D., Chilamkurti, N., & Nanayakkara, V. (2022, June 22). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling - multimedia tools and applications. SpringerLink. https://link.springer.com/article/10.1007/s11042-022-13363-4?utm_source=xmol&utm_medium=affiliate&utm_content=meta&utm_campaign=DDCN_1_GL01_metadata

[9] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 International Conference on Platform Technology and Service (PlatCon), 2017, pp. 1-5, doi: 10.1109/PlatCon.2017.7883728.

[10] A. B. Abdul Qayyum, A. Arefeen and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 2019, pp. 122-125, doi: 10.1109/SPICSCON48833.2019.9065172.

[11] A. Huang, M. (. Puwei and). Bao, "Human Vocal Sentiment Analysis," in May 19, 2019, Available: https://arxiv.org/pdf/1905.08632.pdf.

[12] H. S. Kumbhar and S. U. Bhandari, "Speech Emotion Recognition using MFCC features and LSTM network," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2019, pp. 1-3, doi: 10.1109/ICCUBEA47591.2019.9129067

[13] N. Semwal, A. Kumar, and S. Narayanan, “Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models,” 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), 2017.