Emotion Recognition from Speech

A classification model utilizing the RAVDESS data set

Samantha Gonzales  
 Computer Science MS  
 University of Missouri-Kansas City  
 Kansas City, Missouri  
 sgg8xd@umkc.edu

Savannah Stewart  
Computer Science MS  
 University of Missouri-Kansas City  
 Kansas City, Missouri  
sfsg7f@umkc.edu

Bhagavan Bejjipurapu  
 Computer Science MS  
 University of Missouri-Kansas City  
 Kansas City, Missouri  
 bb8dv@umsystem.edu

Introduction

Emotion recognition of speech can come from a variety of factors, some of which are more physical than audible. However, there are factors within speech that may help to determine the emotion of the speaker without additional information such as context or facial features. In this report, these elements of speech are considered through the use of classification models, which attempt to accurately predict the emotion of the speaker from a set of predictors.

CCS CONCEPTS

• Classification Model • Emotion Recognition • Random Forest Classifier • XGBoost Classifier •

1 Related Work

With the rise of technology in our daily lives, the identification of emotion in human speech is becoming a more widely researched topic, with a focus on acoustic signals [4]. Classification models/algorithms are being utilized to identify and predict human emotions through speech in this growing attention to speech and emotion processing [4]. Y. R. Rochlani and A. B. Raut improve on previous results of these classification models through the use of acoustic features extracted from audio files provided in the RAVDESS dataset, which are pre-labeled for users by a variety of identifiers. By processing the audio files, Y.R. Rochlani and A.B. Raut extract the zero-crossing rate (ZCR), chroma short-term Fourier transformation (chroma STFT), mel-frequency cepstral coefficients (MFCC), root mean square (RMS), and mel-spectrogram for each audio file [4]. With these features, along with the features provided by the naming of the audio files in the data set. Y. R. Rochlani and A. B. Raut train and test commonly used classifier models (support vector machine, decision tree, and random forest) [4]. The accuracy results from Y. R. Rochlani and A B. Raut’s proposed system were exceptional, with the combination of their models having an accuracy of 88.71% [4].

S. Hamsa et al. approach achieving accuracy in emotion recognition differently by not only focusing on the extraction of features, but also the processing of audio files themselves. S. Hamsa processes the audio files by first using a time-frequency composition, which utilizes a WPT cochlear filter bank and modulation transformation, which mimics the cochlear model of the human ear [3]. The results of this time-frequency composition then undergoes speech segregation, which will maintain the variations in speech intensity while reducing fluctuations in frequency [3]. MFCC’s and dominant features are then extracted before being applied to train and test the random forest classification model [3]. In order to evaluate the model’s independence from text and speakers, S. Hamsa et al. utilized three different datasets; Ryerson audio-visual database of emotional speech and song (RAVDESS), speech under simulated and actual stress (SUSAS), and Arabic Emirati emphasized speech dataset (ESD). The average accuracy rates for the proposed model were 86.38%, 88.68%, and 89.45%, respectively [3].

M. M. Rezapour Mashhadi and K. Osei-Bonsu similarly focus on the processing of the audio files by augmenting each file to increase the diversity of data, with noise being added, timing being adjusted, and pitches being shifted [2]. A variety of features were then extracted from the augmented audio files, which served as predictors for a convolutional neural network (CNN) classification model with 1D convolutional layers, and a random forest classification model. Features extracted for M. M. Rezapour Mashhadi and K. Osei-Bonsu’s proposed model are as follows:

1. Zero crossing rate (ZCR)

2. Mel-frequency cepstral coefficients (MFCCs)

3. Roll-off frequency

4. Spectral contrast

5. Harmonic change in tonal centroid features

6. Chromagram from a waveform or power spectrogram

7. Root-mean-square

8. Mel-scaled spectrogram [2]

A combination of data sets were utilized to reduce bias within the classification models, and include: Ryerson audio-visual database of emotional speech and song (RAVDESS), Crema crowd-sourced emotional multimodal actors (CREMA), Toronto emotional speech set (TESS), and Surrey audio-visual expression emotion (SAVEE) [2]. The results for the proposed CNN model showed 60% accuracy, while the results for the proposed random forest model showed 69% accuracy [2].

A. S. M. and A. U. M. similarly evaluate the CNN classification model’s ability to achieve accuracy in emotion recognition by utilizing the RAVDESS data set. Each audio file was processed through the use of pre-emphasis and silence removal, where pre-emphasis refers to the process of increasing the intensity of higher frequences while maintaining the intensity of lower frequences [1]. Mel-frequence cepstral coefficients (MFCC) were extracted as features for the CNN classification model, which utilized 1D convolutional layers to maximize efficacy [1]. The average accuracy of A. S. M. and A. U. M.’s proposed model was 92%, and ranked approximately 20% higher than reference models [1].

2 Approach

The discussed approach to the emotion classification model will include three different steps, as detailed in Section 2.1 below. The approach utilized standard methods of model creation given a data set, which will be discussed further in the provided sections. The data set utilized was the Ryerson audio-visual database of emotional speech and song (RAVDESS).

2.1  Methodology

*2.1.1 Pre-Processing.* To begin the creation of the emotion classification model the data provided by the RAVDESS data set was converted into a table format, which will be referred to as a ‘data frame’ throughout the remainder of this report.

*2.1.2 Feature Extraction.* After the creation of the data frame, features are extracted from the data using a variety of methods. The extracted features are then placed into a new data frame which will be utilized in training and testing the model.

*2.1.3 Data Preparation.* With the new data frame containing the extracted features and relevant data, values within the data frame are prepared as needed to allow best use by the computer/model.

*2.1.4 Classification Model.* A classification model is then built based off of the prepared data frame. This model is trained and tested on the data present in the data frame and will, ideally, correctly identify emotions given a set of predictors.

*2.1.5 Analysis.* Once the classification model has been trained, the difference in the predicted and true results are analyzed to determine if the model is effective. Based on this analysis, the model is adjusted.

*2.1.6 Data Visualization.* Data visualization is created throughout all previously mentioned steps to help ensure a balanced data set is being provided to the model, to evaluate predictors/features, and to assist in analysis.

2.2 Discussion

*2.2.1 Pre-Processing.* The RAVDESS data set for the emotion classification model provided a series of .wav files. These .wav files were recordings of actors saying one of two phrases, with each recording representing a specific emotion with a specific intensity. To assist with the identification of variables within each file, the names of the .wav files followed a pattern, with positions in the number sequence representing a value such as gender, phrase, emotion, or intensity. These values were extracted from the data set to the data frame by creating a Python dictionary, where keys were the number representation in the .wav file name and values were the values that were represented.

*2.2.2 Feature Extraction.* With information provided from the RAVESS data set placed into a data frame, the .wav files themselves were next evaluated. Through the use of the librosa library, which provides audio analysis in Python, the following features were extracted for each .wav file: mel-frequency cepstral coefficients (MFCC), mel spectrogram, tempo, brightness, and pitch.

Mel-frequency cepstral coefficients (MFCC) are commonly used within speech recognition technology to assist with the conversion between human and computer understanding of speech [5]. MFCC’s assist with this conversion by working to break down the signals found in speech and adjusting them to better represent human interpretations of sound, before then adjusting the signals back for computer processing [5].

Humans best interpret the differences in lower pitch than higher pitch. To account for this, a Mel Scale can be applied, which will apply a logarithmic scale to lower pitches and a linear scale to higher pitches to make up for the difference. A Mel Spectrogram contains the pitches from the .wav file that have already been adjusted to fit this scale [5]. To best account for human speech, 26 bins were evaluated in the creation of the Mel Spectrogram.

Tempo and pitch are key components in understanding emotion within speech, with certain emotions tending to have similar trends within tempo and pitch [6].

*2.2.3 Data Preparation.* Categorical information from the pre-processing stage of this approach was placed into the data frame. To allow the model to more accurately evaluate this data all categorical values were ‘one-hot encoded’, meaning that categories became numerically represented in the data frame.

*2.2.4 Classification Model.* Two classification models were created during this approach: random forest and XGBoost.

*2.2.4.1 Random Forest.*Random forest is a commonly used classification model that utilizes a set of decision trees (forest) to predict a classification based on predictors.

To begin, the x (predictors) and y (result) values are established through the use of the data frame. Predictors were established by removing, or dropping, columns from the data frame based on the level of information they provided to the model. *Filename*, *Modality*, and *Vocal Channel* were removed during this step, as all information has already been extracted from the .wav file name (gender, emotion, intensity, etc.) and *Modality* and *Vocal Channel* were the same for all rows. Emotion was also removed from the predictors, as it is what is being predicted by the model (y).

The model was then split into a testing set and training set for boy x and y. The split between test and train was 80/20, with 80% of the available data being utilized in training the model, and 20% being used to test the accuracy of the model. A random state was set to ensure repeatability in testing and *stratify* was set to y, which ensures that each category in the result is being evenly represented in both the training and testing sets.

The model was then created. The ideal tuning parameter for each tree in a random forest model is the square root of the number of predictors. With 48 predictors, a tuning parameter of 6 was utilized. The model was then trained and tested, with a result of 61% prediction accuracy.

*2.2.4.2 XGBoost,* short for Extreme Gradient Boosting, is a powerful machine learning algorithm based on gradient boosting techniques. It builds models in a step-by-step manner by creating a series of decision trees, where each new tree focuses on correcting the mistakes made by the previous ones. This approach helps in building a strong predictive model from several weak learners. Additionally, XGBoost includes built-in mechanisms like regularization to help reduce the risk of overfitting and improve the model’s overall performance and efficiency. In this project, XGBoost was used alongside the random forest classifier for comparison. The same cleaned and feature-enriched dataset was used. Features such as Filename, Modality, Vocal Channel, and the target label Emotion were excluded from the predictor variables. Categorical data were transformed using one-hot encoding to make them suitable for the model. The dataset was split into training and testing sets with an 80/20 ratio, ensuring that each emotion class was evenly represented using stratification. Key model parameters like learning rate, tree depth, and the number of estimators were adjusted to optimize performance. XGBoost was chosen for its ability to handle complex feature interactions, imbalanced classes, and its robustness with large datasets.

2.3 Results

*2.3.1 Random Forest.* As mentioned previously, the Random Forest classifier has a result of 61% prediction accuracy. The model confused emotions that may be considered similar, such as happy and surprised, or sad and calm.

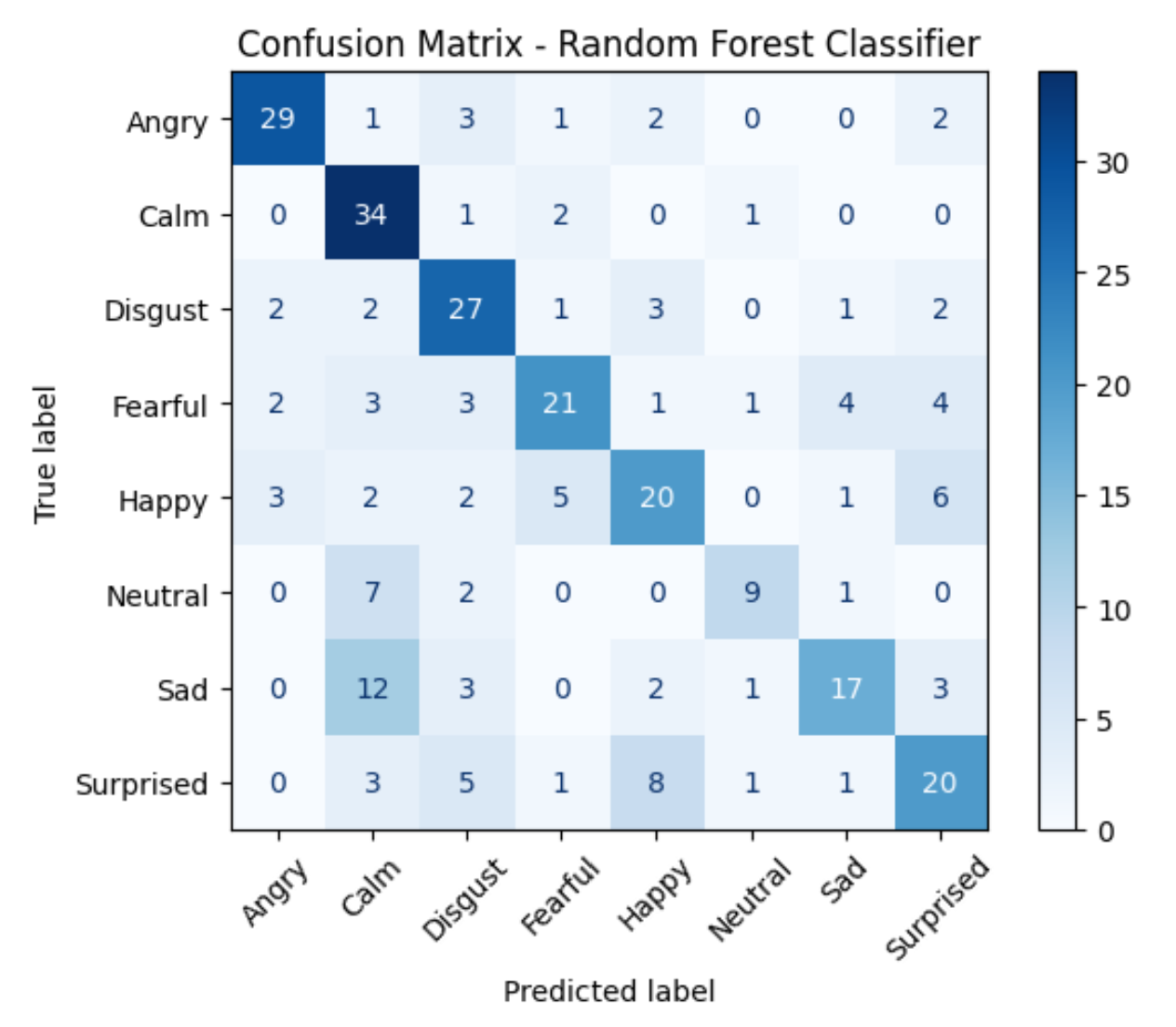


Figure 1: Confusion matrix for random forest classifier

*2.3.2 XGBoost* Compared to the random forest model, this model achieved better accuracy in predicting emotional states from the audio recordings. While both models occasionally misclassified emotions with overlapping characteristics—such as confusing happiness with surprise—XGBoost was generally more effective at distinguishing between emotions. This improvement can be credited to its gradient boosting strategy, which fine-tunes predictions over multiple iterations. As shown in Figure 2, XGBoost made particularly accurate predictions for emotions like anger and neutrality, suggesting a better grasp of the audio-based features like tempo, pitch, and MFCCs extracted during preprocessing.

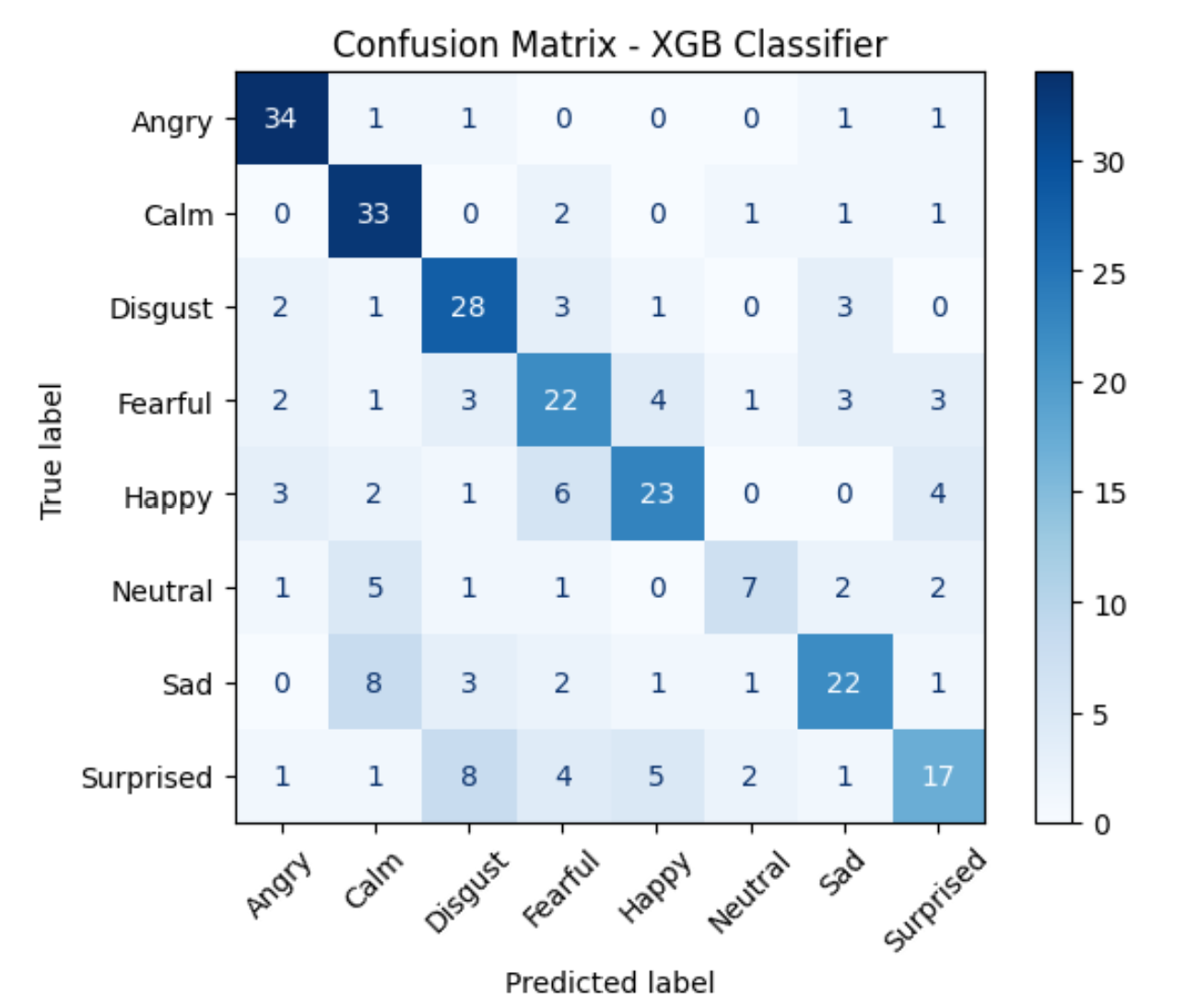


Figure 2: Confusion matrix for XGBoost classifier

Potential limitations may lie within exclusively using the RAVDESS data set, despite emotions and intensity being evenly distributed between genders. Emotion recognition in speech has been shown to differ between genders and ages, as well as within additional circumstances such as emotion perception [6]. As the RAVDESS data set has audio files recorded by actors, the emotions being portrayed are an interpretation of emotions instead of reality, which may impact the accuracy of an emotion’s representation [6]. By exclusively using the RAVDESS set, the predictors provided to both models may have been biased based on the actor’s location or background instead of realistic to the emotion as a whole. Because of this, the model may benefit from a wider range of predictors – either from another data set or from augmentation of the existing data in the RAVDESS data set. It should be noted that any augmentation of data may have enhanced the models’ accuracy, due to diversity in predictors leading to a more extensive and representative data set.

3 Conclusion

Based on the fine tuning of the classifier models, the MFCC and Mel Spectrogram predictors seem to have the most impact on the accuracy of the classification models. With the provided predictors, XGBoost appeared to have a higher accuracy than random forest. Both XGBoost and random forest models appeared to struggle with predicting neutral emotions, and frequently struggled with determining between calm, neutral, and sad.

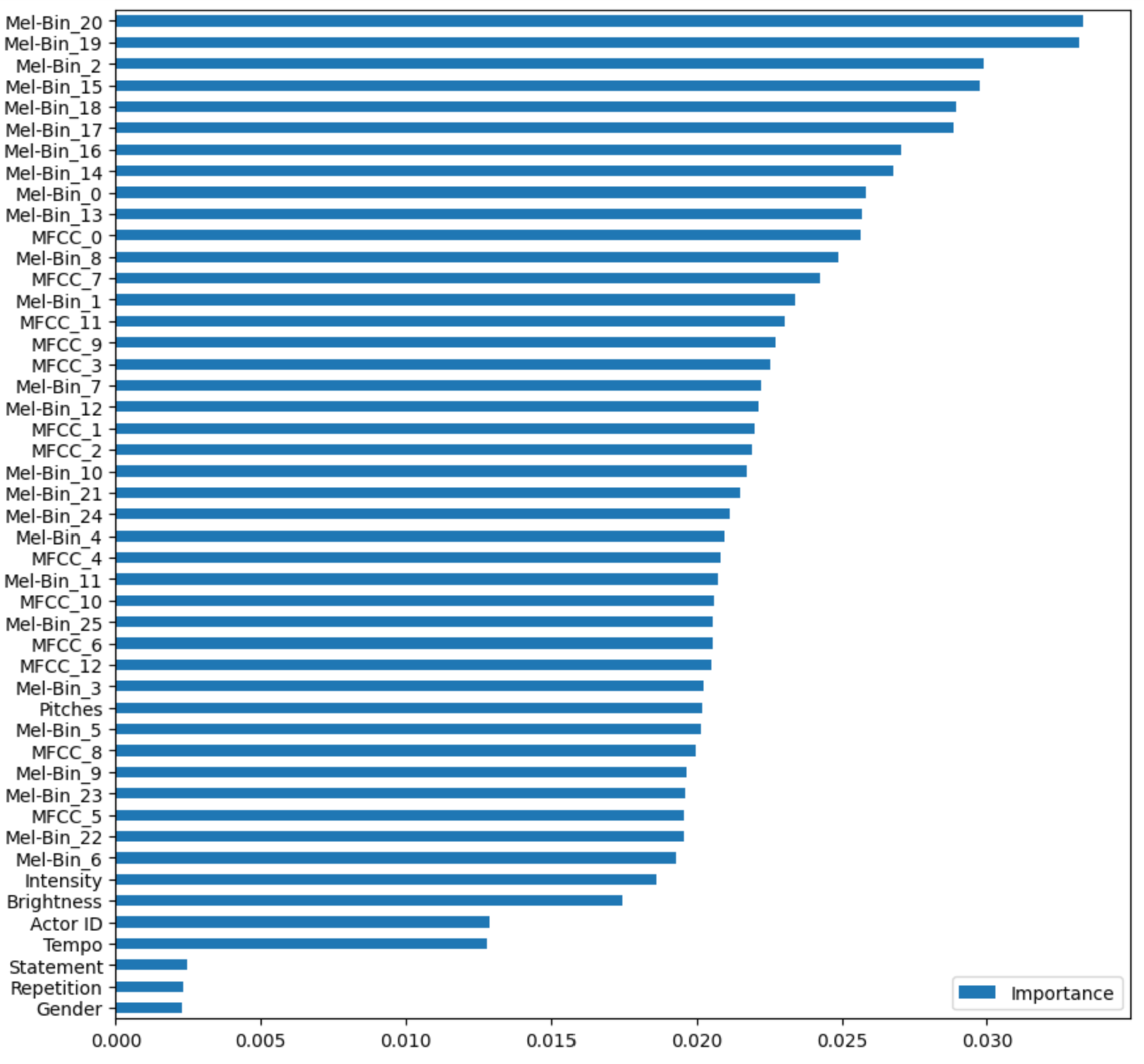


Figure 3: Predictor importance for random forest classifier

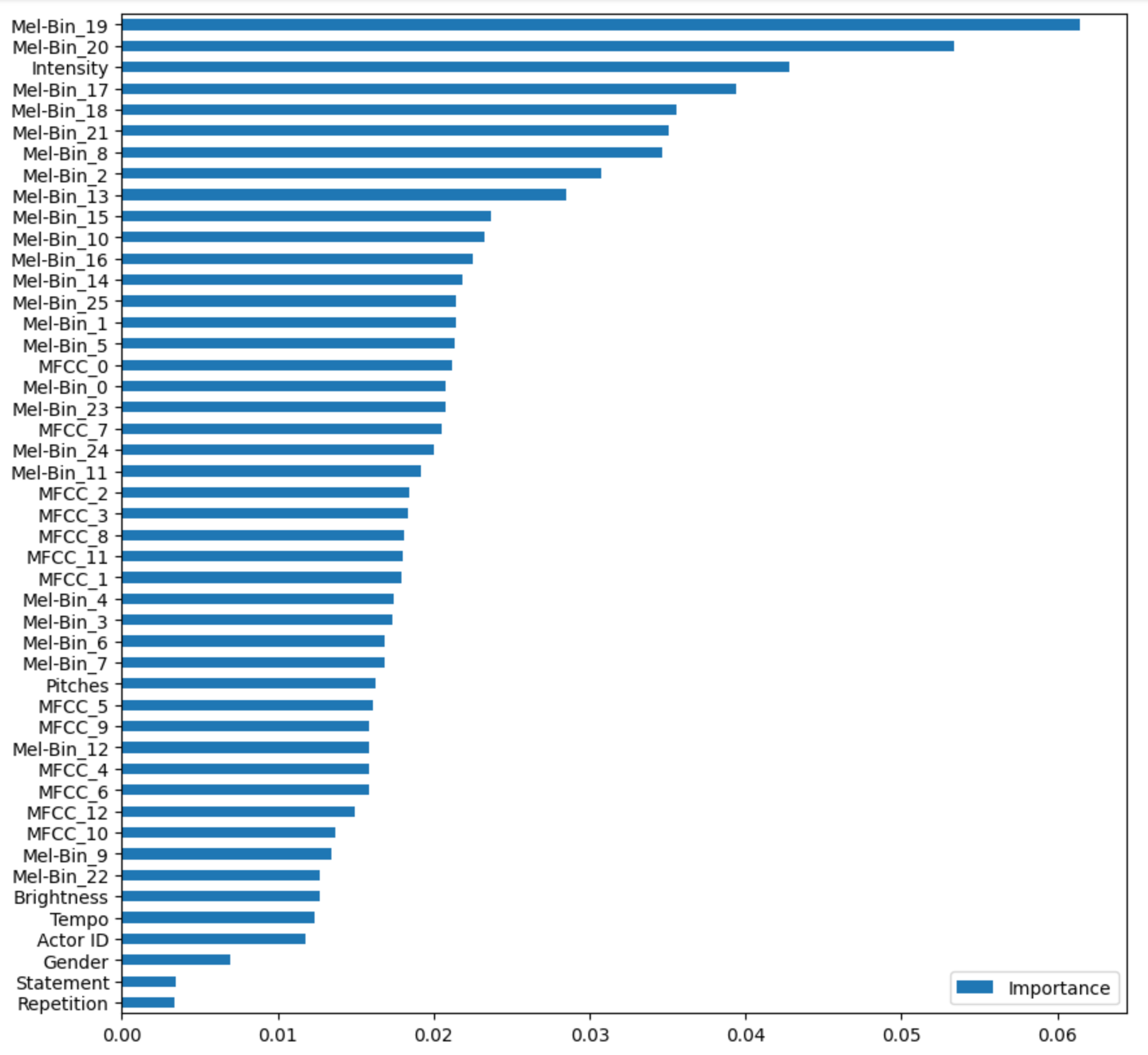


Figure 4: Predictor importance for XGBoost classifier

4 Future Work

Future work regarding these classification models is to improve overall accuracy. Both the random forest and XGBoost models have been best adjusted to fit the current predictors, and therefore most likely are unable to be improved.

The predictors, however, have potential to be reevaluated to assist in improving model accuracy. Further research into the following topics would most likely lead to higher model accuracy in both models mentioned:

1. Speech recognition in artificial intelligence

2. Emotion recognition in human audio processing

3. Sound processing

4. Signal processing

More extensive knowledge regarding current methods of speech recognition in artificial intelligence, as well as sound and signal processing, would assist in determining, mathematically, what features would be most likely to accurately represent discrepancies in speech patterns. Further knowledge regarding emotion recognition in human audio processing would allow a more detailed and nuanced approach to characterizing emotions in speech.

REFERENCES

[1] A.S.M and A.U.M, “Enhancing Speech Emotion Recognition Through Advanced Feature Extraction and Deep Learning: A Study Using the RAVDESS Dataset,” 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-7, doi: 10.1109/ICITEICS61368.2024.10624849.

[2] M. M. Rezapour Mashhadi and K. Osei-Bonsu, “Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest,” PloS one, vol. 18, no. 11, p.e0291500, 2023, doi: https://doi.org/10.1371/journal.pone.0291500.

[3] S. Hamsa, I. Shahin, Y. Iraqi and N.Werghi, “Emotion Recognition from Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier,” in IEEE Access, vol. 8, pp. 96994-97006, 2020, doi: 10.1109/ACCESS.2020.2991811.

[4] Y. R. Rochlani and A. B. Raut, “Machine Learning Approach for Detection f Speech Emotions for RAVDESS Audio Dataset,” 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024, pp. 1-7, doi: 10.1109/ICAECT60202.2024.10468810.

[5] “Mel-frequency Cepstral Coefficients (MFCC) for Speech Recognition.” GeeksforGeeks. https://www.geeksforgeeks.org/mel-frequency-cepstral-coefficients-mfcc-for-speech-recognition/ (accessed May 6, 2025)

[6] “Emotional prosody.” Wikipedia: The Free Encyclopedia. https://en.wikipedia.org/wiki/Emotional\_prosody (accessed March 31, 2025)

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00