One Source to Detect them All: Gender, Age, and Emotion Detection from Voice

Syed Rohit Zaman*, Dipan Sadekeen*, M Aqib Alfaz*, Rifat Shahriyar†
Military Institute of Science and Technology*, Bangladesh University of Engineering and Technology†
Dhaka, Bangladesh
Email: rohitzaman00@gmail.com, udoy320@gmail.com, aqibalfaz@gmail.com, rifat@cse.buet.ac.bd

***Abstract*—Gender, age, and emotion detection from the speech are essential in machine and human interaction. Sometimes it is required to categorize audios by age and gender from speech. Sometimes it is required to predict age, gender, and emotion from audio clips for investigation purposes. Most telecommunication companies need to analyze audio calls to predict customer demography and recommend offers based on demographic segments. Several researchers have focused on detecting gender, age, and emotion from different types of sources. But according to the best of our knowledge, none of them use a single type of source to detect all of them. We have introduced a system to detect gender, age, and emotion from audio speech. In our system, all audio files were converted into 20 statistical features, and the converted numerical datasets were used to create the different prediction models to attain the objective. The different prediction models are Random Forest, CatBoost, Gradient Boosting, K-nearest neighbors (KNN), XGBoost, AdaBoost, Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine (SVM). All the prediction models were evaluated and compared based on their test accuracy. In predicting gender, CatBoost performs best among all predictive models with 96.4% test accuracy. On the other hand, Random Forest performs best for predicting age among all predicting models with 70.4% test accuracy. For emotion prediction, XGBoost performs best with 66.1% test accuracy. It was also analyzed among 20 features which features are most influential for the effective prediction models. We believe that our findings will be beneficial to future researchers in this area.**

I. Introduction

Speech is a medium of communication to exchange information from one speaker to one or more listeners. The speaker emits a voice signal in the form of pressure waves moving from the speaker's mouth to the audience's ears. In speech processing, speech is transformed into electrical signals, voltages, or currents, in which form typically speech signals can be analyzed. The frequency spectrum analysis [1] of the signal is called the analysis of the amplitude, frequency, and phase of the audio signals. From the frequency spectrum analysis, significant information can be extracted. Mean Frequency is one of them. It estimates the mean normalized frequency of the power spectrum of a time-domain audio signal. Another one is the Standard Deviation. It is the measure of the dispersion of a set of data from its mean. It measures the absolute variability of a distribution. Similarly, the median, first quartile, and third quartile can be extracted. The interquartile range is the difference between the first quartile and the third quartile. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. Then the mode is the value that appears most often in a set of data values. Similarly, some other properties also can be extracted from frequency spectrum analysis. The outcome of the frequency spectrum analysis can be used to form a dataset.

A supervised learning algorithm learns from labeled training data to predict unforeseen data results. One of the vital speech analysis applications is to predict gender, age, and emotion from speech. Detecting gender, age, and emotion from the speech will help a lot to enhance human-machine interaction. Audios can be classified by age and gender after forecasting properly. Age, gender, and emotion can significantly help the relevant investigation. These predictions can be beneficial to most telecommunication companies. The audio calls can be analyzed, and the generated prediction models for age, gender, and emotion can be used to predict customer demography. They can recommend offers based on that.

Several researchers have focused on detecting gender, age, and emotion from different types of sources. But according to the best of our knowledge, none of them use a single type of source to detect all of them. We have introduced a system to detect gender, age, and emotion from speech. The main contributions of our work are as follows:

1) A dataset generated from the Mozilla audio dataset [2] that can be used for predicting gender where feature columns were the 20 statistical properties that were extracted by the frequency spectrum analysis [1] of each audio with the help of R programming language and where each instance was labeled with gender.
2) A dataset generated from the Mozilla audio dataset that can be used for predicting age. Similar 20 statistical properties were extracted from each audio as used as feature columns with R programming language. In this case, the target variable was age.
3) A dataset generated from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [3] for predicting emotion. A similar process was followed to form the dataset. 20 statistical features were the feature columns, and the emotion was the target variable by applying frequency spectrum analysis with the help of the R programming language.
4) On each dataset, ten different machine learning algo-

rithms were applied for predicting gender, age, and emotion to get the best predictive model among all models with the help of Python's Scikit-learn library [4].

5) A comprehensive evaluation of all the machine learning models in terms of accuracy, precision, recall, and F1 score.

Later sections of this paper are organized as follows. The literature review is briefly introduced in Section II. Section III briefly discusses the proposed system. Later on, in Section IV, the performance of the machine learning models are compared, and the relative features for the best predictive model are shown. Finally, the conclusion is stated in Section V.

II. Literature Review

This section briefly discussed the studies related to predicting gender, age, and emotion from speech.

A. Gender Detection

Shafran et al. [5] conducted a study to predict gender from speech using the Hidden Markov Model (HMM) and the Support vector machine (SVM). The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. Similarly, Přibil et al. [6] conducted a study to predict gender from speech using Gaussian Mixture Modelling (GMM). The predictive model's accuracy is 95% for children, 99% for males, and 98% for females. A study by Steve Jadav [7] used a statistical approach to analyze audio and detect gender using machine learning concluded with an accuracy of 97% [in SVM training 96.6% and testing 97%] by using a dataset from Kaggle [8]. Similarly, Harbet et al. [9] conducted a study to predict gender from speech using a set of neural networks as classifiers. The accuracy of the predictive model is 91.72%. Similarly, Djemili et al. [10] conducted a study to predict gender from speech using GMM, Multilayer Perceptron (MLP), Vector Quantization (VQ), and Learning Vector Quantization (LVQ). The accuracy of the predictive model is 96.4% using the IViE corpus dataset. Similarly, Bahari et al. [11] conducted a study to predict gender from speech using Weighted Supervised Nonnegative Matrix Factorization (WSNMF) and age with Generalized Regression Neural Network (GRNN). The accuracy of the predictive model 96% for gender using the Dutch database. Similarly, Ramdinmawii et al. [12] conducted a study to predict gender from Pitch's speech using Auto-Correlation, Signal Energy, Mel Frequency Cepstral Coefficients (MFCC), and SVM classifiers. The predictive model's accuracy is 69.23% for MFCC, 57.14% for Pitch, 55.81% for Energy using the TIMIT dataset. Wang et al. [13] conducted a study to predict gender from speech using Deep Neural Networks. The predictive model's accuracy for gender with age is 90.56% to 92.72% using 17,408 real-traffic Mandarin utterances collected from a Microsoft spoken dialogue system.

B. Age Detection

Shafran et al. [5] conducted a study to predict age from speech using HMM and SVM. The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. Similarly, Přibil et al. [6] conducted a study to predict age from speech using GMM. Bahari al [11] conducted a study to predict age with GRNN. Wang et al. [13] conducted a study to predict age from speech using Deep Neural Networks. The predictive model's accuracy for gender with age is 90.56% to 92.72% using 17408 real-traffic Mandarin utterances collected from a Microsoft spoken dialogue system.

C. Emotion Detection

Shafran et al. [5] conducted a study to predict emotion from speech using HMM and SVM. The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. Grosbras et al. [14] conducted a study to predict Emotion from speech using MLP. The accuracy of the predictive model is 74% by using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Wang et al. [13] conducted a study to predict emotion from speech using Deep Neural Networks. The accuracy of the predictive model for emotion is 59.40% to 63.20% using 17408 real-traffic Mandarin utterances collected from a Microsoft spoken dialogue system.

Several studies have been conducted to predict gender, age, and emotion from speech using different approaches. The frequency spectrum analysis [1] of speech can predict gender, age, and emotion. There is no study in the literature to compare all the predictive models using supervised machine learning algorithms to measure comparative performance. In most studies, all the approaches followed for predicting gender, age, and emotion from the speech are not the same. Our study focused on following a single approach of predicting gender, age, and emotion using machine learning by extracting statistical features from frequency spectrum analysis of speech and getting the best predictive model among all models through a comprehensive evaluation.

III. Proposed System

The proposed system is described in the following phases: dataset, methodology, and predictive model creation.

A. Dataset

This study used the audio dataset of Mozilla [2] to build prediction models for gender and age. The data set contains 64000 audio files in MP3 format with 61528 (Common voice corpus 5.1) different voices. The dataset also includes a CSV file containing filename, accent, age, gender, upvotes, and downvotes. Among those, only 6247 data were taken. The CSV file was filtered by counting upvotes up to 2 and downvotes up to 0 and removing data with any missing attributes. Among all the columns of the CSV file, only filename, gender, and age were chosen. There were 2 choices for gender and 9 choices for age. The gender choices were male and female. The age choices were teens, twenties, thirties, forties, fifties, sixties, seventies, eighties, and nineties. We categorized the age column into 3 categories. The category young includes

teens and twenties, the category matured includes thirties, forties, fifties, and the category old included sixties, seventies, eighties, nineties. As mentioned before, the filtered CSV had three columns filename, gender, and age. The CSV was then converted into two CSV files, one for gender and another for age. The filtered audio dataset had 4659 male and 1588 female audio files. All the audio files converted into WAV format required for frequency spectrum analysis [1]. Similarly, to build the prediction models for emotion, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [3] dataset was used, which includes 7356 files. There are 24 professional actors in the database (12 female, 12 male), vocalizing two lexically matched statements in a North American neutral accent. The speech includes expressions of calm, happiness, sadness, angry, fear, surprise, and disgust. The song includes feelings of calm, happiness, sadness, angry, and fear. We only used speech for this study. The speech file contains 1440 audio files. Each filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.wav), where the third one from the left side denotes the emotion. If emotion is 01, then it means neutral, 02 means calm, 03 means happy, 04 means sad, 05 means angry, 06 means fearful, 07 means disgust, and 08 means surprised. A CSV file was created with two columns, filename and emotion. We have only used 3 choices for emotion: happy, sad, and angry. So, in a nutshell, we have three CSV files, one for gender (columns: filename, gender), one for age (columns: filename, age), and one for emotion (columns: filename, emotion).

B. Methodology

The WAV format audios were analyzed with Frequency Spectrum Analysis (FSA) [1] process using the R programming language. The process enables the extraction of statistical features such as mean, median, etc. 20 statistical features were extracted through the FSA process. They are: mean frequency (*meanfreq*), standard deviation of frequency (*sd*), median frequency (*median*), first quartile (*Q25*), third quartile (*Q75*), interquartile range (*IQR*), skewness (*skew*), kurtosis (*kurt*), spectral entropy (*sp.ent*), spectral flatness or tonality coefficient (*sfm*), mode frequency (*mode*), spectral centroid (*centroid*), average fundamental frequency (*meanfun*), minimum fundamental frequency (*minfun*), maximum fundamental frequency (*maxfun*), average dominant frequency (*meandom*), minimum dominant frequency (*mindom*), maximum dominant frequency (*maxdom*), range of dominant frequency (*dfrange*), and modulation index (*modindx*).

The above statistical data with a filename was then stored in a CSV file. By applying this process, three CSV files were generated. From the Mozilla audio dataset, two CSV were generated. One is for gender, and another one is for age. Similarly, from the RAVDESS audio dataset, another CSV file was generated having 21 attributes. So, for gender, age, and emotion, there were a total of six CSV files. For the final CSV containing gender's statistical information, a join operation was performed between two CSV files. As filename is the common column of both CSV files, the final CSV file is created using the join operation. A similar approach was followed for the final CSV containing age's and emotion's statistical information. Each dataset has no categorical column. But there were some missing values on the CSV files. Their mean replaced the missing numerical values. On each dataset, there was a class imbalance problem. This problem was handled by using an oversampling technique named SMOTE [15]. There were some outliers in the dataset, and that was removed as well.

C. Predictive Model Creation

We have used different machine learning algorithms to predict gender, age, and emotion. They are Random Forest, CatBoost, Gradient Boosting, K-nearest neighbors (KNN), XGBoost, AdaBoost, Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine (SVM). We have used the previously mentioned dataset containing the frequency spectrum analysis [1] of the speeches. All the datasets were split into train-tests, and the ratio was 80% and 20%. The train set was used for generating the models, and the models are evaluated by both the training set and test set. Then 10-fold cross-validation was also applied for each predictive model creation. Then each model's performance was evaluated and compared. We have used Scikit-learn [4] for model creation, an open-source machine learning and data mining library of Python.

IV. Results

In this section, we present all the experimental results with a comparative evaluation.

A. Confusion Matrix

Confusion matrices of the predictive model of gender, age, and emotion are reported for both train and test set here. The confusion matrix helps to see correctly and incorrectly classified instances of the models. For example, among 5348 training instances, 2655 and 2693 instances are correctly classified as female and male, respectively, whereas, among 1338 testing instances, 663 and 627 instances are correctly classified as female and male, but 25 instances are classified as male where the actual class was female, and 23 instances are classified as female where the actual class was male. Among 4759 training instances, 1564, 1598, and 1597 are correctly classified as matured, old, and young, respectively, whereas, among 1190 testing instances, 240, 349, and 253 instances are correctly classified as matured, old, and young. 214, 234, and 222 instances are correctly classified as angry, happy, and sad among 753 training instances. Here, 19 and 13 instances are classified as happy and sad though they are actually an angry class. For happy emotion, 16 emotions were wrongly classified (9 angry and 7 sad). As for the sad emotion, 13 and 22 instances are wrongly classified as angry and happy. For test instances of emotion, among 189 instances, 43, 47, and 35 instances are correctly classified as angry, happy, and sad emotions. 17 and 8 instances were classified as happy and sad where their actual classes are angry. For happy instances, 9

and 8 instances are classified wrongly as angry and sad, and 9 and 13 instances were wrongly classified as angry and happy where their actual classes are sad.

In this study, accuracy, precision, recall, and F1 score of train set and test set are measured on each model to predict gender, age, and emotion. All the data are sorted in descending order in all tables based on test accuracy.

B. Gender Prediction

In Table I, 100% training accuracy is achieved by CatBoost, Random Forest, KNN, and Decision Tree models. The above four classifiers also achieve 100% precision, recall, and F1 score for the train set. Gradient Boosting achieves 99.8% of accuracy, precision, recall, F1 score. The best test set accuracy, precision, recall, and F1 score of 96.4% in all parameters is achieved by CatBoost. The lowest is measured for Naive Bayes, whose accuracy, recall, precision, and F1 score are 89.7%. In the rest of the models, the parameters are found between 93.1% to 95.9%.

The result for tenfold cross-validation for gender detection is shown in Table II. The highest training accuracy of 96.2% is achieved by CatBoost, whereas Naive Bayes achieves the lowest training accuracy of 87.4%. CatBoost obtains the best value for precision, recall, and F1 score of 96.3%, 96.2%, and 96.2%, respectively. Naive Bayes obtains the lowest value for precision, recall, and F1 score of 87.4% for all parameters. CatBoost achieves the highest test accuracy, precision, and F1 score of 95.4% and test recall of 95.3%. The worst in all parameters are for Naive Bayes, test accuracy and F1 score of 89.5% and test precision and recall of 89.6%.

C. Age Prediction

In Table III, accuracy, precision, recall, and F1 score of train set and test set are shown on each model to predict age. The models find the best accuracy, precision, recall, and F1 score of 100% in all parameters: Random Forest, KNN, Decision Tree by using train set data. SVM achieves the lowest accuracy, precision, and recall of 43.5%, 43%, and 43.4%, respectively. The lowest F1 score is achieved using Naive Bayes, which is 42.1%. The highest accuracy and recall is achieved by Random Forest of 70.4% and 70.8%, respectively. In the cases precision and F1 score, the highest value of 70.1% and 70.3% respectively is gained by both Random Forest and CatBoost models. SVM finds all the lowest parameters though the test sets data gave higher percentages than train set data. SVM obtains 43.7%, 43.9%, 44%, and 42.9% accuracy, precision, recall, and F1 score, respectively.

The result for tenfold cross-validation for age detection is shown in Table IV. The highest training accuracy of 68.5% is achieved by Random Forest, whereas SVM achieves the lowest accuracy of 43.1%. Random Forest achieves the highest train precision of 68.1%, and SVM achieves the lowest precision of 42.7%. CatBoost achieves the highest recall and F1 score of 67.9% and 67.8% for training. The lowest recall and F1 score of 43% and 41.3% for training are achived by SVM and Naïve Bayes. The highest test accuracy and F1 score of 61.7% and test recall and precision of 61.8% are all achieved by CatBoost, and SVM obtained worst in all parameters: test accuracy 43.2%, precision 43.5%, recall 43.3% and F1 score 42.8%.

D. Emotion Prediction

In Table V, accuracy, precision, recall, and F1 score of train set and test set are shown on each emotion detection model. The models find the best accuracy, precision, recall, and F1 score of 100% in all parameters: CatBoost, Gradient Boosting, Random Forest, KNN, and Decision Tree by using train set data. Naive Bayes gives the lowest accuracy, precision, recall, and F1 score of 49.8%, 51.5%, 49.7%, and 48.8%, respectively. XGBoost obtained the highest accuracy, precision, recall, and F1 score of 66.1%, 66.7%, 66%, and 66%, respectively. All lowest parameters are achieved by the Naive Bayes, which is also the case for test sets data of 47.1%, 49.2%, 47.2%, and 45.5%, respectively.

The result for tenfold cross-validation for emotion detection is shown in Table VI. CatBoost achieves the highest train accuracy of 63.6%, whereas Naive Bayes achieves the lowest accuracy of 49.3%. CatBoost also achieves the highest values of train precision, recall, and F1 score of 64.2%, 63.5%, and 63.4%, respectively. On the other hand, Naive Bayes achieves the lowest values of train precision, recall, and F1 score of 51.1%. 49.1% and 48%, respectively. XGBoost achieves the highest test accuracy, precision, recall, and F1 score of 58.7%, 60.8%, 58.8%, and 57.9%, respectively The worst in all parameters are obtained by KNN, test accuracy 38.5%, precision 37.1, recall 37.6%, and F1 score 35.65%.

E. Relative Feature Importance

The relative feature importance of the best predictive models of gender, age, and emotion is reported here[1]. The best predictive model for gender, age, and emotion detection is CatBoost, Random Forest, and XGBoost, respectively. The most important feature for gender detection is *meanfun*, which the relative feature importance defines. The three most important features are *meanfun*, *maxfun*, and *Q25*. *mindom* is the least important feature for gender detection. In detecting age, the most important features are *sfm*, *Q25*, *meanfun*, which hold nearly the same relative importance, and the least important feature is *mindom*. Lastly, for emotion detection, the most important feature is *Q25*. *Mode* and *meandom* hold about the same relative importance, and these are also important to detect emotion. The least important feature for this case is *centroid*.

V. Conclusion

Several studies have focused on detecting gender, age, and emotion from different types of sources. But according to the best of our knowledge, none of them use a single type of source to detect all of them. We have introduced a system to detect gender, age, and emotion from speech. New datasets are generated from audio clips by frequency spectrum

[1] http://rifatshahriyar.github.io/figures.pdf

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
CatBoost	1	1	1	1	0.964	0.964	0.964	0.964
Gradient Boosting	0.998	0.998	0.998	0.998	0.959	0.959	0.959	0.959
Random Forest	1	1	1	1	0.958	0.958	0.958	0.958
KNN	1	1	1	1	0.957	0.957	0.956	0.957
XGBoost	0.964	0.964	0.964	0.964	0.955	0.955	0.955	0.955
ANN	0.949	0.949	0.949	0.949	0.95	0.95	0.95	0.95
AdaBoost	0.984	0.984	0.985	0.984	0.943	0.943	0.943	0.943
Decision Tree	1	1	1	1	0.939	0.939	0.94	0.939
SVM	0.935	0.935	0.935	0.935	0.931	0.931	0.931	0.931
Naive Bayes	0.875	0.875	0.875	0.875	0.897	0.897	0.897	0.897

TABLE I: Accuracy, Precision, Recall, and F1 score of different classifier for gender detection (both train and test, train-test split (80%-20%), sorted by test accuracy)

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
CatBoost	0.962	0.963	0.962	0.962	0.954	0.954	0.953	0.954
Random Forest	0.956	0.957	0.957	0.958	0.951	0.948	0.945	0.950
XGBoost	0.953	0.954	0.953	0.953	0.942	0.943	0.942	0.942
ANN	0.939	0.940	0.939	0.939	0.942	0.943	0.941	0.942
Gradient Boosting	0.958	0.958	0.958	0.958	0.941	0.942	0.941	0.941
SVM	0.933	0.933	0.933	0.932	0.936	0.936	0.936	0.936
AdaBoost	0.943	0.943	0.943	0.943	0.933	0.934	0.933	0.933
KNN	0.946	0.948	0.946	0.946	0.931	0.933	0.930	0.930
Decision Tree	0.928	0.923	0.925	0.924	0.919	0.919	0.921	0.920
Naive Bayes	0.874	0.874	0.874	0.874	0.895	0.896	0.896	0.895

TABLE II: Accuracy, Precision, Recall, and F1 score of different classifier for gender detection (both train and test, 10-fold cross validation, sorted by test accuracy)

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
Random Forest	1	1	1	1	0.704	0.701	0.708	0.703
CatBoost	0.992	0.992	0.992	0.992	0.703	0.701	0.706	0.703
Gradient Boosting	0.942	0.943	0.942	0.942	0.665	0.662	0.668	0.664
KNN	1	1	1	1	0.66	0.648	0.665	0.649
XGBoost	0.727	0.728	0.726	0.727	0.619	0.619	0.622	0.62
AdaBoost	0.685	0.686	0.684	0.685	0.591	0.593	0.594	0.592
Decision Tree	1	1	1	1	0.582	0.575	0.585	0.579
ANN	0.58	0.571	0.579	0.571	0.576	0.574	0.58	0.57
Naive Bayes	0.443	0.446	0.441	0.421	0.441	0.461	0.448	0.421
SVM	0.435	0.43	0.434	0.422	0.437	0.439	0.44	0.429

TABLE III: Accuracy, Precision, Recall, and F1 score of different classifier for age detection (both train and test, train-test split (80%-20%), sorted by test accuracy)

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
CatBoost	0.68	0.678	0.679	0.678	0.617	0.618	0.616	0.617
Random Forest	0.685	0.681	0.677	0.675	0.605	0.607	0.60	0.608
Gradient Boosting	0.636	0.634	0.635	0.634	0.59	0.593	0.591	0.591
XGBoost	0.615	0.616	0.614	0.614	0.569	0.569	0.57	0.567
KNN	0.630	0.613	0.629	0.615	0.536	0.534	0.539	0.531
AdaBoost	0.564	0.568	0.563	0.564	0.529	0.534	0.530	0.531
ANN	0.536	0.527	0.535	0.527	0.517	0.517	0.519	0.513
Decision Tree	0.575	0.567	0.572	0.57	0.489	0.492	0.473	0.503
Naive Bayes	0.436	0.437	0.435	0.413	0.436	0.443	0.440	0.430
SVM	0.431	0.427	0.430	0.418	0.432	0.435	0.433	0.428

TABLE IV: Accuracy, Precision, Recall, and F1 score of different classifier for age detection (both train and test, 10-fold cross validation, sorted by test accuracy)

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
XGBoost	0.89	0.892	0.89	0.89	0.661	0.667	0.66	0.66
CatBoost	1	1	1	1	0.64	0.64	0.641	0.64
Gradient Boosting	1	1	1	1	0.64	0.642	0.64	0.64
ANN	0.683	0.681	0.681	0.681	0.64	0.641	0.645	0.64
Random Forest	1	1	1	1	0.624	0.629	0.626	0.623
AdaBoost	0.85	0.849	0.85	0.849	0.587	0.588	0.591	0.587
SVM	0.612	0.611	0.611	0.611	0.582	0.583	0.583	0.583
KNN	1	1	1	1	0.582	0.589	0.586	0.582
Decision Tree	1	1	1	1	0.524	0.524	0.524	0.523
Naive Bayes	0.498	0.515	0.497	0.488	0.471	0.492	0.472	0.455

TABLE V: Accuracy, Precision, Recall, and F1 score of different classifier for emotion detection (both train and test, train-test split (80%-20%), sorted by test accuracy)

Model	Accuracy(train)	Precision(train)	Recall(train)	F1_score(train)	Accuracy(test)	Precision(test)	Recall(test)	F1_score(test)
XGBoost	0.606	0.608	0.604	0.602	0.587	0.608	0.588	0.579
ANN	0.584	0.594	0.583	0.581	0.571	0.571	0.568	0.561
SVM	0.587	0.592	0.586	0.585	0.561	0.606	0.554	0.549
Gradient Boosting	0.596	0.599	0.595	0.593	0.561	0.596	0.564	0.556
AdaBoost	0.587	0.588	0.586	0.583	0.556	0.564	0.551	0.539
CatBoost	0.636	0.642	0.635	0.632	0.518	0.526	0.513	0.508
Random Forest	0.619	0.632	0.61	0.625	0.507	0.593	0.566	0.540
Decision Tree	0.51	0.518	0.50	0.522	0.455	0.483	0.406	0.399
Naive Bayes	0.436	0.437	0.435	0.413	0.436	0.443	0.44	0.43
KNN	0.542	0.543	0.541	0.538	0.385	0.371	0.376	0.356

TABLE VI: Accuracy, Precision, Recall, and F1 score of different classifier for emotion detection (both train and test, 10-fold cross validation, sorted by test accuracy)

analysis (FSA) to build predictive models for gender, age, and emotion detection. Ten different machine learning algorithms are applied for predicting gender, age, and emotion. A comprehensive evaluation of all the machine learning models in terms of accuracy, precision, recall, and F1 score is carried out to find the best predictive model. CatBoost performs best for predicting gender, Random Forest performs best in detecting age, and for predicting emotion, XGBoost performs best. CatBoost shows 95.4% cross-validation accuracy, 100% train accuracy, and 96.4% test accuracy for gender. On the other hand, Random Forest shows 60.5 % cross-validation accuracy, 100% train accuracy, and 70.4% test accuracy for age. Similarly, XGBoost shows 58.7% cross-validation accuracy, 89% train accuracy, and 66.1% test accuracy for emotion. The relative feature importance of each model is also generated. We believe that our findings will be beneficial to future researchers in this area. As for the limitation, if the audio clips consisted of voices from more than one person simultaneously, it will not be able to find out the age, gender, or emotion of them within the provided accuracy. The dataset used to detect emotion is small in size and we plan to collect more data in the future to improve the accuracy of the model for emotion detection. We left the work to detect age, gender, and emotion from audio voices of multiple people for the future.

REFERENCES

[1] S. A. Fulop, *Speech spectrum analysis*. Springer Science & Business Media, 2011.

[2] Voice.mozilla.org, "Gender recognition by voice," https://commonvoice.mozilla.org/en/datasets, 2021, accessed: 2021-01-15.

[3] Smartlaboratory.org, "Ravdess — smart lab," https://smartlaboratory.org/ravdess/, 2021, accessed: 2021-01-15.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[5] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 2003, pp. 31–36.

[6] J. Přibil, A. Přibilová, and J. Matoušek, "Gmm-based speaker gender and age classification after voice conversion," in *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016, pp. 1–5.

[7] S. Jadav, "Voice-based gender identification using machine learning," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1–4.

[8] Kaggle.com, "Gender recognition by voice," https://www.kaggle.com/primaryobjects/voicegender, 2021, accessed: 2021-01-15.

[9] H. Harb and Liming Chen, "Gender identification using a general audio classifier," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, vol. 2, 2003, pp. II–733.

[10] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal based gender identification system using four classifiers," in *2012 International Conference on Multimedia Computing and Systems*, 2012, pp. 184–187.

[11] M. H. Bahari and H. Van Hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011, pp. 1–6.

[12] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in *2016 International Conference on Signal Processing and Communication (ICSC)*, 2016, pp. 244–249.

[13] Z. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5150–5154.

[14] M.-H. Grosbras, P. D. Ross, and P. Belin, "Categorical emotion recognition from voice improves during childhood and adolescence," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, p. 321–357, Jun. 2002.