**CHAPTER 4**

**PROPOSED SYSTEM DESIGN, IMPLEMENTATION, AND PERFORMANCE EVALUATION**

This chapter discusses system design and detailed work flow of the proposed speech based children gender classification system. And also the discussion about the experimental implementation and the performance evaluations of the proposed system are discussed. The proposed system is implemented using the Python programming language. The speech signals are digitized at a sample frequency rate 44.1 kHz. In this experiment, spoken sentences in Myanmar language are used as input and the system outputs the gender of the speaker.

**4.1** **System Design of the Proposed System**

The proposed system consists of four main components, namely, dataset preparation, preprocessing, feature extraction and classification. For dataset preparation, voice of children is needed and so children’s speech is recorded repeatedly in a quiet place. And then speech features are extracted from recording files to create a speech features dataset. The second stage is preprocessing of the speech signal. In machine learning and data mining, preprocessing makes input data easier to work with algorithms. In preprocessing, noise reduction and removal of silences and unvoiced regions are included. The next stage involves feature extraction. The extraction of the relevant and important information from the speech signals of the human voice is an important task to produce a latter classification performance. In this system, MFCC speech features are used for prediction. In the last stage, extracted features are evaluated using machine learning classifiers to predict children’s gender. Figure 4.1 shows the design of proposed children gender classification system using speech.

Children Speech

(6 to 11 age range)

Preprocessing

(Silence and Unvoiced Speech Removal)

Feature Extraction

(MFCCs)

Classification

Start

Predicted Output

(Male/Female)

End

Figure 4.1 System Flow Diagram

4.1.1 Dataset Preparation

The database used in this system consists of sentences read aloud by children both male and female in Myanmar language. The database was designed to create a training set of speech from children of KG to Grade V (age range 6 to 11 years). Recording specification used for dataset is shown in table 4.1.

Table 4.1 Recording Specifications

|  |  |
| --- | --- |
| File Type | .wav format |
| Duration | 2 or 3 second |
| Numbers of Channel | Mono (1 Channel) |
| Sampling Frequency | 44.1 kHz |
| Number of Bits | 16 bits |

Utterances from children are recorded with SONY Digital Stereo High Definition. These voice clips are preprocessed and evaluated. There are total of 1100 audio records. The female records contain 566 samples where male records have 534 samples.

4.1.2 Preprocessing

Only voiced region of speech contains most of the gender related information. Leading/trailing silence in the audio may not contain much information and thus not useful for the classification. There ar many silence and unvoiced regions in the recording files of children. Hence, removing this silence and unvoiced regions is done in preprocessing step. These regions are removed from the speech using librosa.effects.trim function which is a build in function of Python.

4.1.3 Feature Extraction

Features efﬁcient in discriminating female and male voice in children speech should be identiﬁed. The most commonly used acoustic features in gender classiﬁcation are MFCCs. They play a signiﬁcant role in applications such as speech recognition, speaker recognition, etc. MFCCs mimics human speech production and speech perception, by logarithmic perception of loudness. MFCCs which are short term spectral based features are extracted from children speech. MFCC features are frequently used by many researchers for speech recognition and in music/ speech classification problem.

4.1.4 Classification

The classiﬁcation task involves the implementation of various classiﬁers for gender identiﬁcation task. Classification is establishing a mathematical model that separates into male and female based on the features of children’s speech. Classification model is built on the training set and check the accuracy of the model by using it on the testing set. In this system, machine learning classification algorithms are compared using MFCC feature dataset. Train and test set accuracies are observed for five classification algorithms. The classiﬁers are chosen mainly based on the non-linear nature of data. Random Forest (RF), Artificial Neural Network (ANN), Logistic Regression (LR), Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB) are used to develop the gender prediction task.

**4.2 Implementation of the Proposed System**

Figure 4.2 illustrates the home page of the proposed children gender classification system using speech. In the system, two main parts are categorized for the gender classification: performance analysis and classification. In performance analysis, classification performance for each machine learning classifier can be analyzed using training and testing dataset divided by randomly. In next section, voice of a children can be selected from any folder and classify the speaker of this audio is boy or girl.

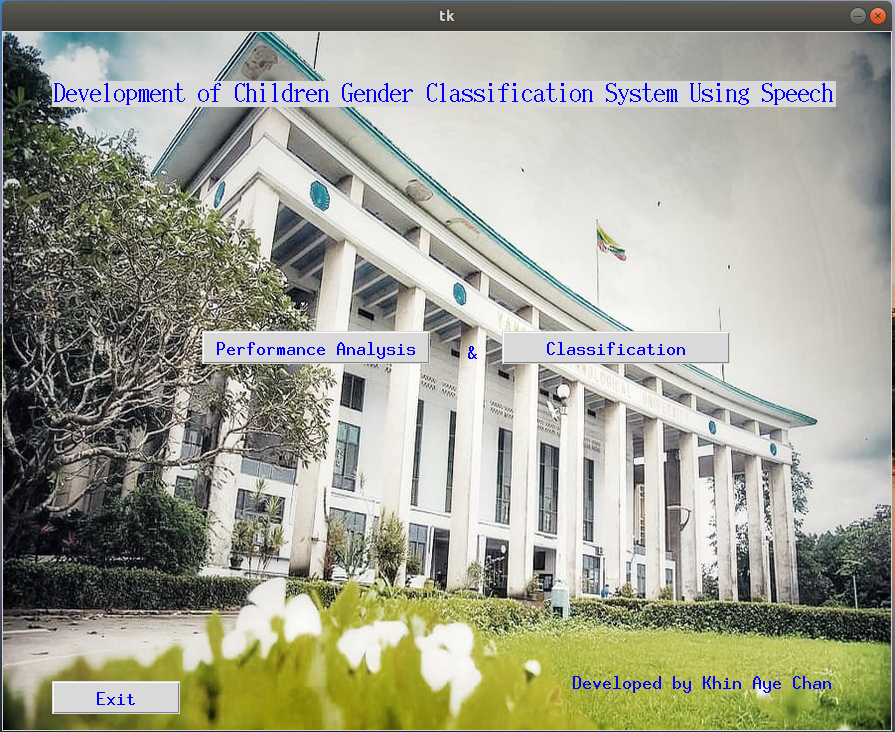


Figure 4.2 Start Form for Children Gender Classification System

4.2.1 Performance Analysis

Figure 4.3 shows performance analysis page for machine learning algorithms. In this experimental study of classification algorithms, performance of machine learning algorithms can be analyzed for gender classification using voice feature dataset. User can split feature dataset into random train and test subsets and choose a classification model among classifiers used in the system. Any classifier displayed in combo box can be chosen for classification. And performance measures: training accuracy, testing accuracy, precision, recall, F1-score, and support are used to evaluate performance of learning algorithms. When OK button is clicked, the system displays training and testing accuracy and other performance measures of selected classifier.

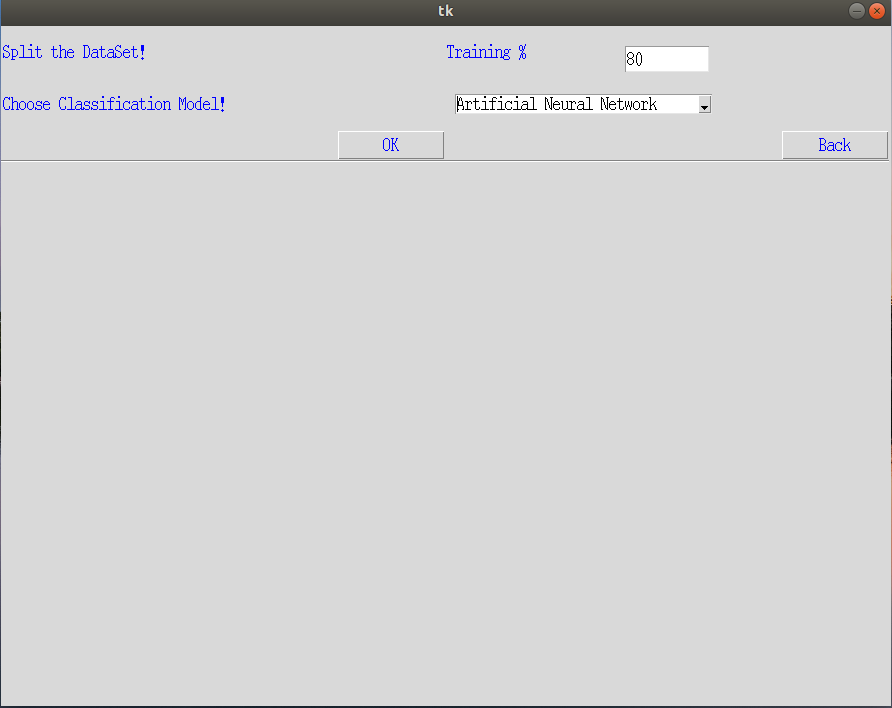


Figure 4.3 Form of Performance Analysis

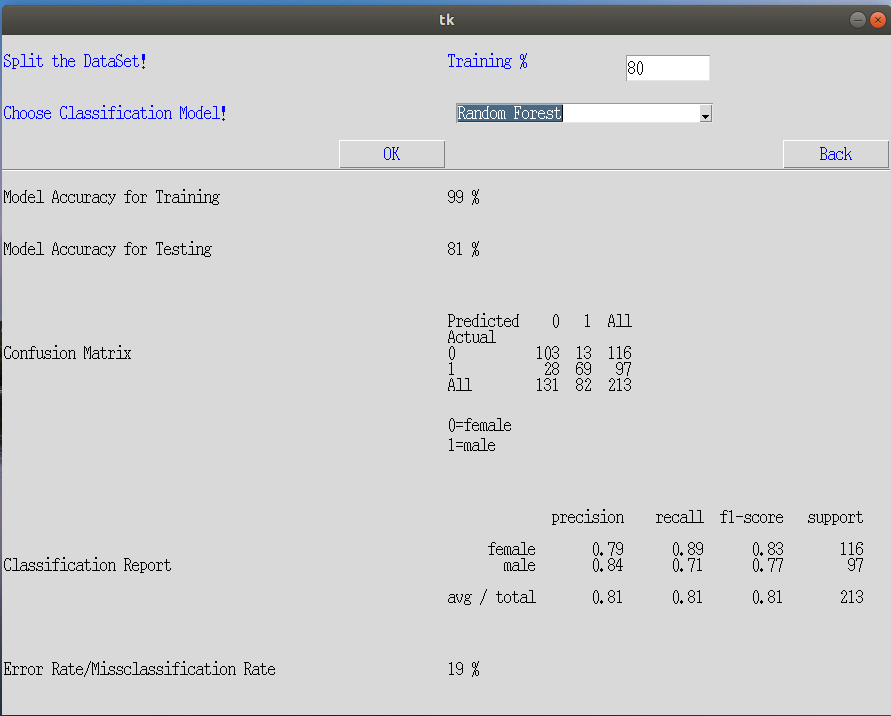


Figure 4.4 Classification Results of RF Classifier

Classification performance of RF and ANN can be seen in Figure 4.4 and Figure 4.5. 80% of features dataset is trained and testing is done on 20 % of dataset. Testing dataset includes the number of male is 116 and the number of female is 97. RF can predict male 69 and female 119 correctly and others are wrongly classified. Similarly ANN can make correct classification for male 63 and female 99. RF has 99% training accuracy and 81 % testing accuracy. ANN gets 84 % train accuracy and the accuracy on the test set is 76 % . Performance of other classification algorithms can study using MFCC feature dataset. SVM classification model gets 74 % testing accuracy, LR model achieves 77 % and GNB is 47 respectively. These accuracies can change depending on testing data set. But according to experimental results, RF has highest testing accuracy with over 80% and other classifiers generally get around 75%.



Figure 4.5 Classification Results of ANN

4.2.2 Real Time Classification

Real time classification of each audio file can be done in classification page as shown in Figure 4.6. User can browse a child’s voice from any folder and upload to the system by clicking browse button. After uploading audio, selected audio file can be listened when play button is clicked. MFCC features of this audio are extracted by using feature extraction button and machine learning algorithms classify this uploaded file whether a boy’s voice or a girl’s voice by clicking classify button.

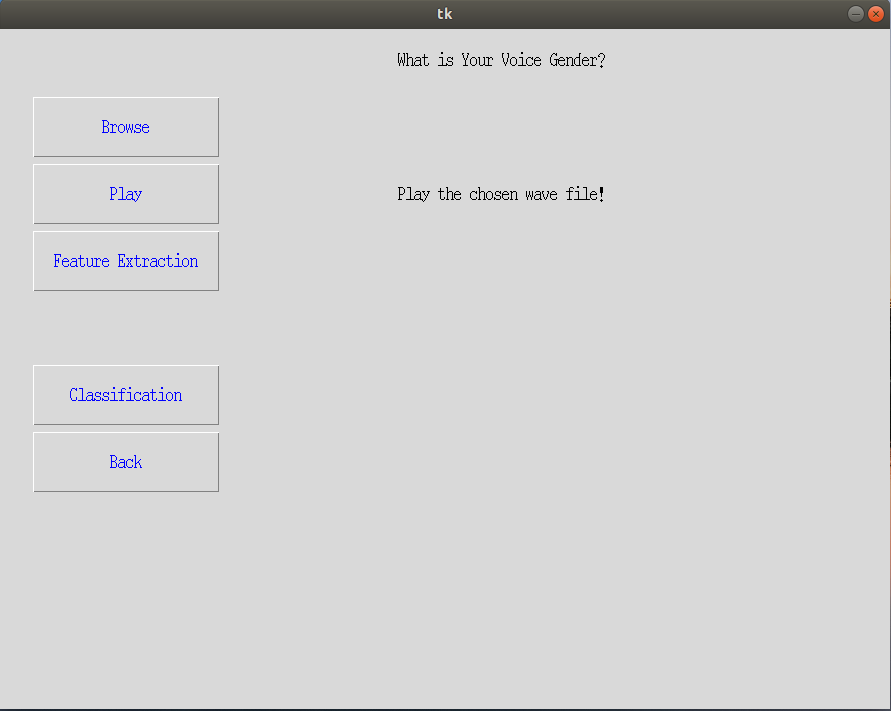


Figure 4.6 Classification Form

Firstly, browse button displays open file dialog box as shown in Figure 4.7. The open file dialog component allows users to browse the folders of their computer. The dialog box returns the path and name of the file the user selected in the dialog box.

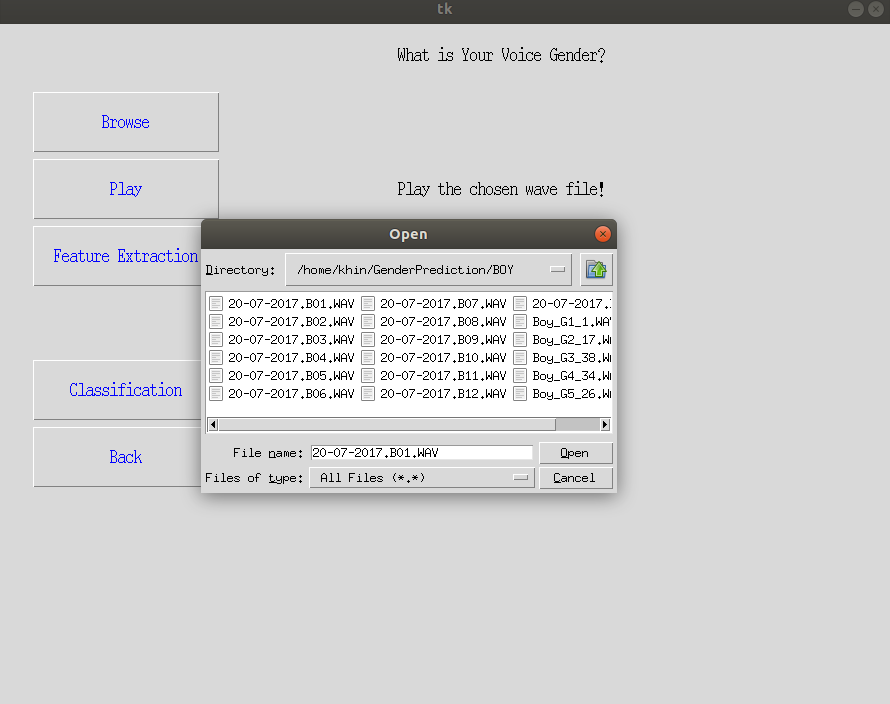


Figure 4.7 Browsing an Audio File from a Folder

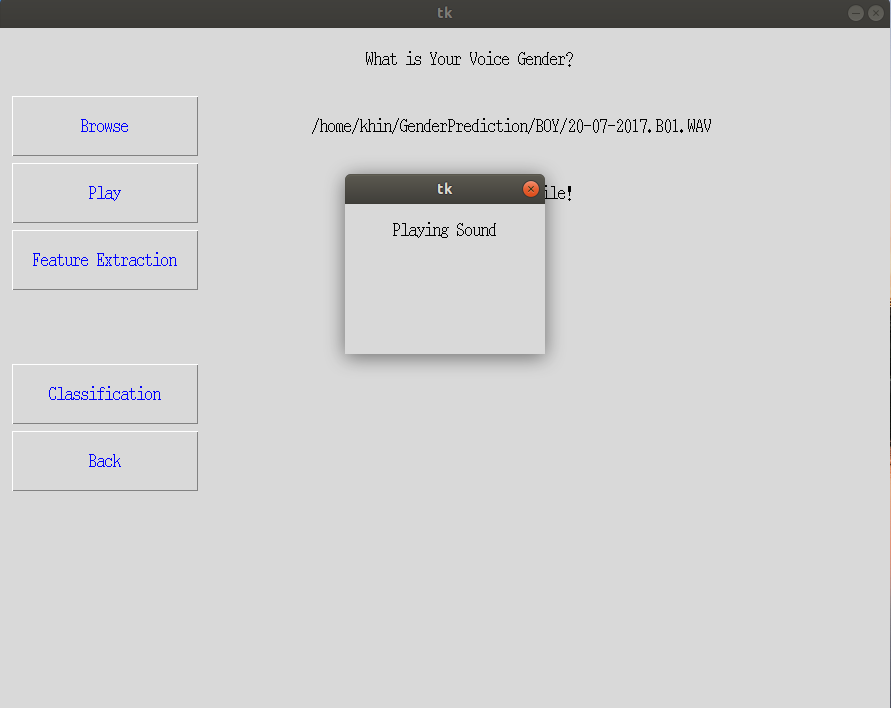


Figure 4.8 Screenshot of Playing Selected Audio File

By clicking play button, the sound of chosen audio file can be listened. Figure 4.8 shows screen shot of playing the wave file selected in a folder of the computer.

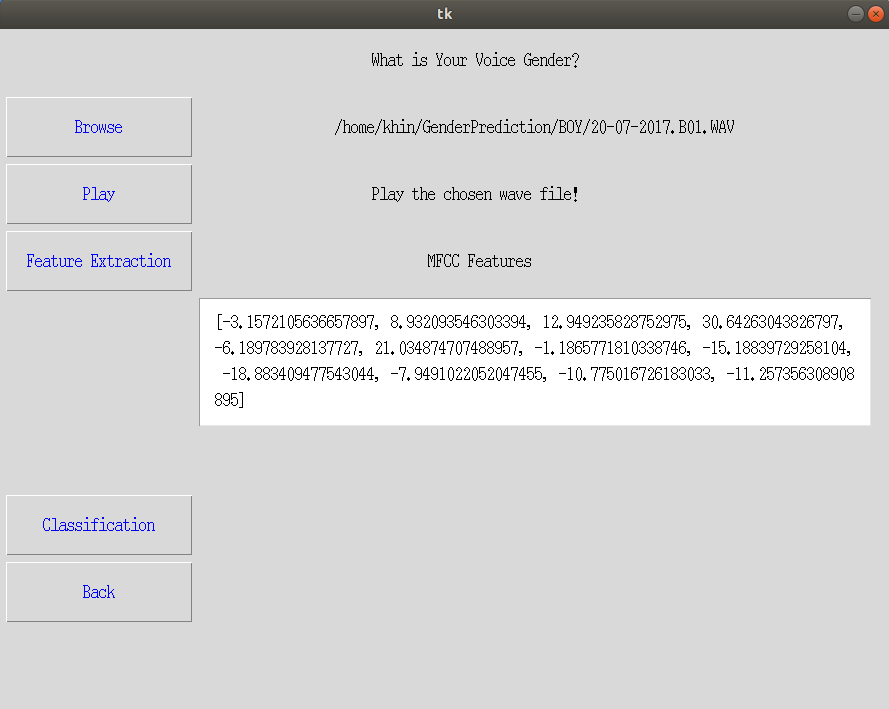


Figure 4.9 MFCC Features

Feature extraction button extracts MFCC features of wave files. Figure 4.9 is features points of selected wave file resulted by using MFCC feature extraction method. Figure 4.10 classification results of five machine learning algorithms applied in the proposed system after testing speech of a boy. Figure 4.11 shows classification results of five classifiers for testing speech of a girl.

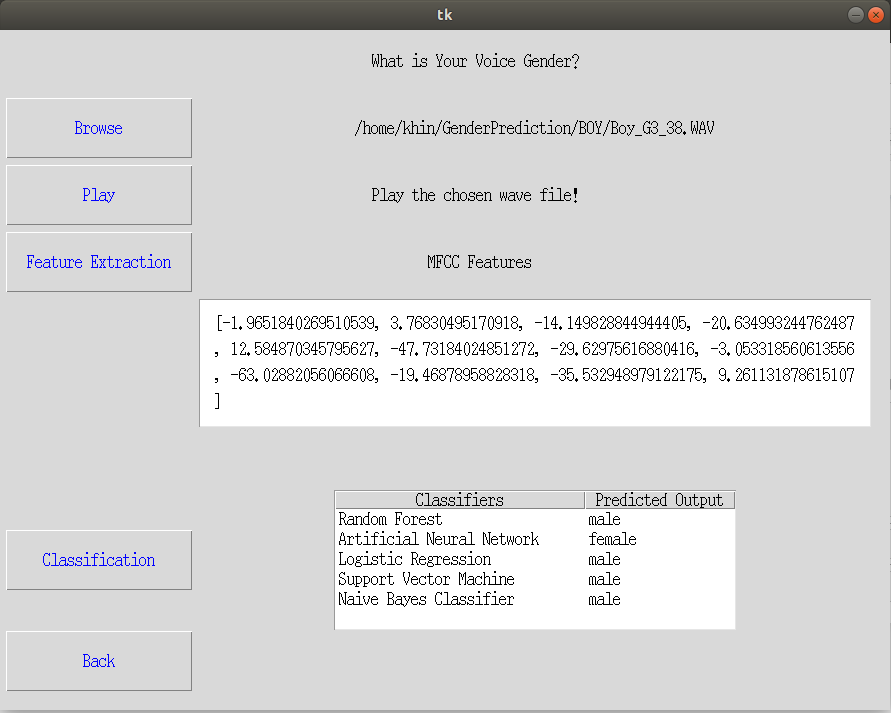


Figure 4.10 Results of Classifying a Boy’s Speech

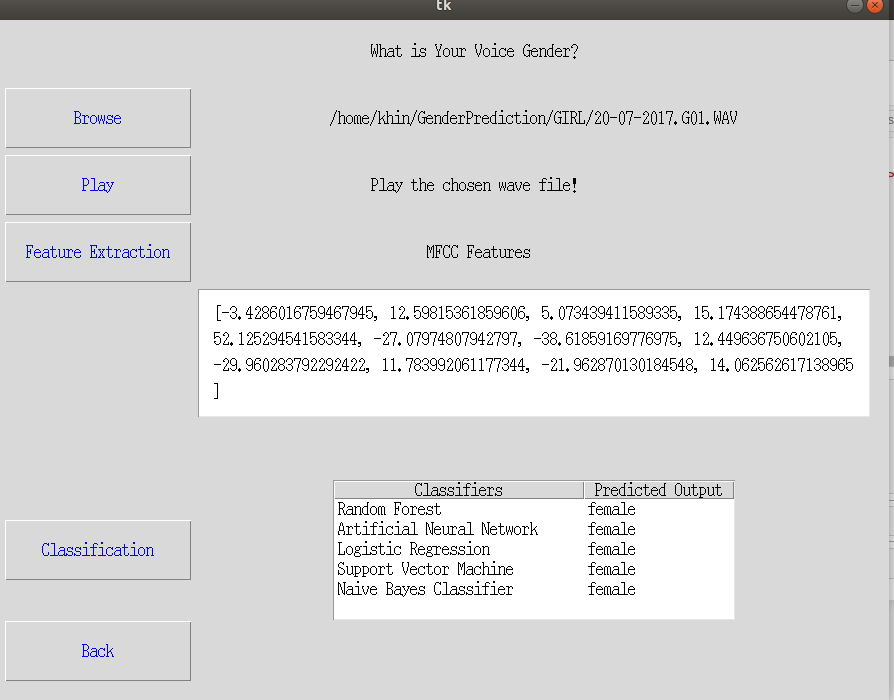
****.

Figure 4.11 Testing Results of a Girl’s Speech

**4.3 Comparison of Performance Evaluations of Different Classification Techniques**

In this system voice dataset contains 1100 audio records where recordings of female child contain 566 speech samples and male child recording have 534 speech samples. The performance of the proposed system is specified based on testing accuracy. Two testing is done to estimate the performance of the machine learning models: k-fold cross validation method and simple train test split. Moreover different performance metrics: confusion matrix, precision, recall, F1-score, support are used to evaluate performance of classifiers.

4.3.1 Simple Train Test Split

To do gender classification, a machine learning model is established and needed to train it using voice features dataset. After training, the model accuracy is checked on some test dataset. For this, a dataset which is different from the training set used earlier is required. In such cases, the obviously solution is to split the original dataset into two sets, one for training and the other for testing before start training classification model. Splitting test dataset and train dataset can make by any ratio and size of datasets can be declared as test size and train size.

In the proposed system, 90% of the instances are used for training and 10% for testing for simple train test split. Each classiﬁer is trained for feature dataset and observed the train and test set accuracies for five classification algorithms. The accuracy is percentage of total number of instances correctly identiﬁed. Table 4.2 shows the training accuracy and testing accuracy of classiﬁcation for different classifiers. RF is efﬁcient in building an accurate classiﬁer which can efﬁciently run on the small and large sized datasets of non-linear nature. Hence RF is observed achieving good accuracy compared to the other four classiﬁers. ANN is low accuracy compared to the RF. Though ANN is efﬁcient in modeling the non-linear data, small size of may affect the performance of ANN as they need large data for training. RF is efﬁcient in discriminating features non-linear in nature. It also works well with the small sized data. RF outperforms ANN with highest accuracy of 83% for feature dataset used in this system. LR has only 78% training accuracy and 76% testing accuracy, SVM achieves 79% and 77% and GNB obtains 75% and 74% respectively.

Table 4.2 Accuracy Results

|  |  |  |
| --- | --- | --- |
| Classifiers | Training Accuracy | Testing Accuracy |
| RF | 99 % | 83 % |
| ANN | 82 % | 79 % |
| SVM | 76 % | 74 % |
| LR | 76 % | 75 % |
| GNB | 76 % | 75 % |

Figure 4.12 and Figure 4.13 are bar charts which show correct and incorrect predictions made by five classifiers by using same ratio of training and testing data. Total number of testing dataset is 110 including 53 boys and 57 girls.

Figure 4.12 Bar Chart for Correct Predictions

Figure 4.13 Bar Chart for Incorrect Predictions

4.3.2 K-Fold Cross-Validation

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. This procedure is implemented by randomly dividing the set of observations into k groups, or folds, of approximately equal size. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:

* Take the group as a hold out or test data set
* Take the remaining groups as a training data set
* Fit a model on the training set and evaluate it on the test set
* Retain the evaluation score and discard the model

1. Summarize the skill of the model using the sample of model evaluation score

Table 4.3 Cross Validation Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifiers | RF | ANN | SVM | LR | GNB |
| 10-fold Cross Validation Score | 80% | 77% | 76% | 81% | 72% |
| 80% | 76% | 75% | 70% | 77% |
| 88% | 75% | 74% | 70% | 75% |
| 85% | 80% | 79% | 76% | 80% |
| 79% | 77% | 75% | 72% | 72% |
| 85% | 76% | 74% | 82% | 73% |
| 75% | 75% | 72% | 80% | 72% |
| 81% | 74% | 80% | 77% | 70% |
| 85% | 78% | 80% | 75% | 79% |
| 89% | 72% | 81% | 78% | 70% |
| Average | 83% | 76% | 77% | 76% | 74% |

Classifiers models are trained and tested for 90% and 10% of the total dataset

respectively. Table 4.3 describes average testing accuracy of classifiers by K-fold cross validation. The table shows that RF classifier has highest average accuracy among five classification algorithms.

4.3.3 Performance Metrics

The performance metrics chosen to evaluate machine learning models are very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. Metrics used for evaluation of children gender classification are precision, recall, F1-score and support.

4.3.3.1 Confusion matrix

The confusion matrix is a table with two dimensions (“Actual” and “Predicted”), and sets of “classes” in both dimensions. Actual classifications are columns and predicted ones are rows. The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on confusion matrix and the numbers inside it.

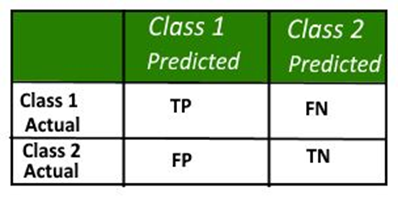


Figure 4.14 Sample Confusion Matrix

* True Positives (TP): True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True).
* True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False).
* False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one.
* False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one.

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made. Classification accuracy can be calculated based on confusion matrix as shown in Equation 4.1.

|  |  |
| --- | --- |
|  | Equation 4.1 |

4.3.3.2 Precision

**Precision is the amount of positive predictions that were correct.** Precision is the number of correct positive results divided by the number of positive results predicted by the classifier.

|  |  |
| --- | --- |
|  | Equation 4.2 |

4.3.3.3 Recall

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In other words, recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

|  |  |
| --- | --- |
|  | Equation 4.3 |

4.3.3.4 F1-Score

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It shows how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

|  |  |
| --- | --- |
|  | Equation 4.4 |

The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as Equation 4.4.

Table 4.4 Performance Matrices

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifiers** | **Gender** | **Precision** | **Recall** | **F1-score** | **Support** |
| RF | Female | 81% | 88% | 84% | 57 |
| Male | 85% | 77% | 81% | 53 |
| ANN | Female | 76% | 86% | 81% | 57 |
| Male | 81% | 69% | 74% | 53 |
| LR | Female | 77% | 74% | 75% | 57 |
| Male | 76% | 79% | 77% | 53 |
| SVM | Female | 72% | 82% | 77% | 57 |
| Male | 83% | 73% | 78% | 53 |
| GNB | Female | 74% | 75% | 74% | 57 |
| Male | 74% | 72% | 73% | 53 |

Table 4.4 gives performance measures of each classifiers calculated on feature dataset divided into 90% training and 10% testing.

**4.4 Summary**

This chapter presents the proposed system design and detailed work flows of children gender classification system. Feature extraction algorithms and machine learning algorithms are used. To get good efficient features , MFCC feature extraction algorithm is applied. Experimental results are analyzed using five different machine learning algorithms: RF, ANN, LR, SVM and GNB. Implementation of gender classification is explained step by step in this chapter. Usage of performance measures of machine learning classifiers is also considered. The conclusion, further extensions, and limitations are discussed in next chapter.