**CHAPTER 4**

**THE 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 digitised 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. Data preprocessing is an integral step in machine learning as the quality of data and the useful information that can be derived from it directly affects the ability of classification model to learn. 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 the good 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. The performance of five classifiers is compared in this study. 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. It was designed to create a training set of speech from children of KG to Grade V (age range 6 to 11 years). The database contains audio recordings of individuals from different grades. Firstly, 10 Myanmar sentences were prepared. They are:

* 1. မဂၤလာပါ။
  2. ေက်းဇူးတင္ပါတယ္။
  3. နာမည္ဘယ္လိုေခၚလဲ။
  4. ဘယ္သြားမလို႔လဲ။
  5. ေနေကာင္းလား။
  6. ဘယ္မှာေနပါသလဲ။
  7. စာေမးပြဲေျဖနိုင္လား။
  8. အေမကိုေခၚလာခဲ့ပါ။
  9. ဘာျဖစ္လို႔လဲ။
  10. ဘာ၀ါသနာပါလဲ။

Utterances from children are recorded with SONY Digital Stereo High Definition. These voice clips are preprocessed and evaluated. There are total of 1200 audio records. The female records contain 600 samples where male records have 600 samples. 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 ~ 32 bits |

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 is not useful for the classification. There are 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. The length of the analysis window used in MFCC feature extraction algorithm is 0.025s. The step between successive windows is 0.01s. Hamming window is used as the analysis window to apply to each frame. FFT size is 2048 and the value of pre emphasis filter coefficient is 0.97. The number of cepstrum to return is 13 that is number of MFCC feature points. After computing MFCC features using these parameters, a numpy array of size (number of frames by number of cepstrum) containing features. Each row holds 1 feature vector. MFCC features are frequently used by many researchers for speech recognition and in music/ speech classification problem.

4.1.4. Classification

A classification model attempts to draw some conclusion from observed values. When one or more inputs are given, a classification model will try to predict the value of one or more outcomes. 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 the accuracy of the model is checked 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’s speech based gender classification system. In the system, two main parts are categorized for the gender classification: performance analysis and classification. In classification section, voice of a children that is needed to classify can be selected from any folder and classify the speaker of this audio is boy or girl. In performance analysis, classification performance for each machine learning classifier can be analysed using training and testing dataset divided by randomly.

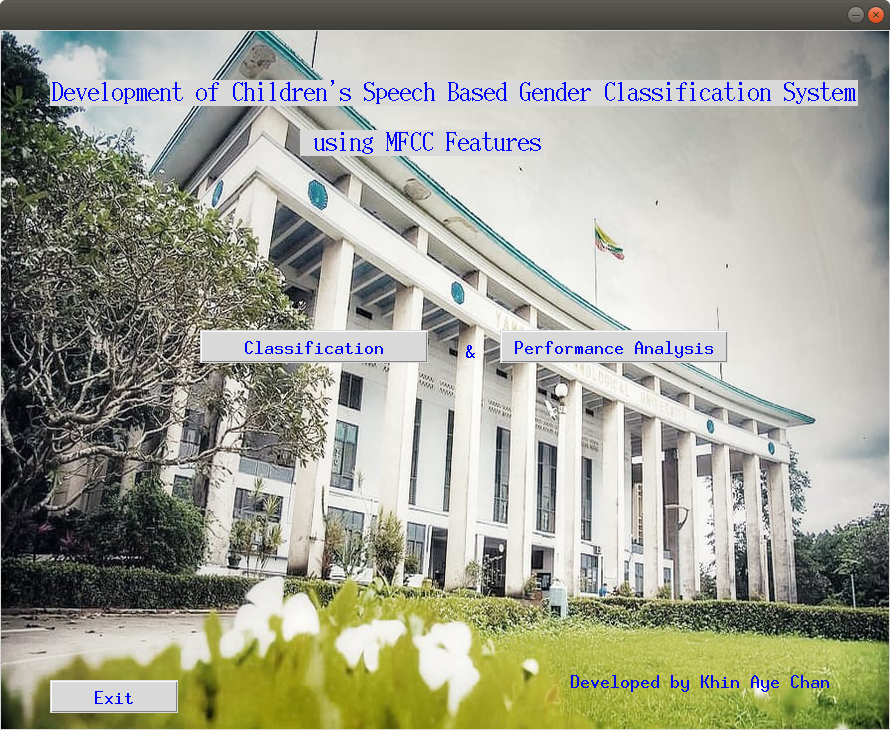


Figure 4.2. Start Form for Children’s Speech Based Gender Classification System

4.2.1. Real Time Classification

Real time classification of each audio file can be done in classification page as shown in Figure 4.3.

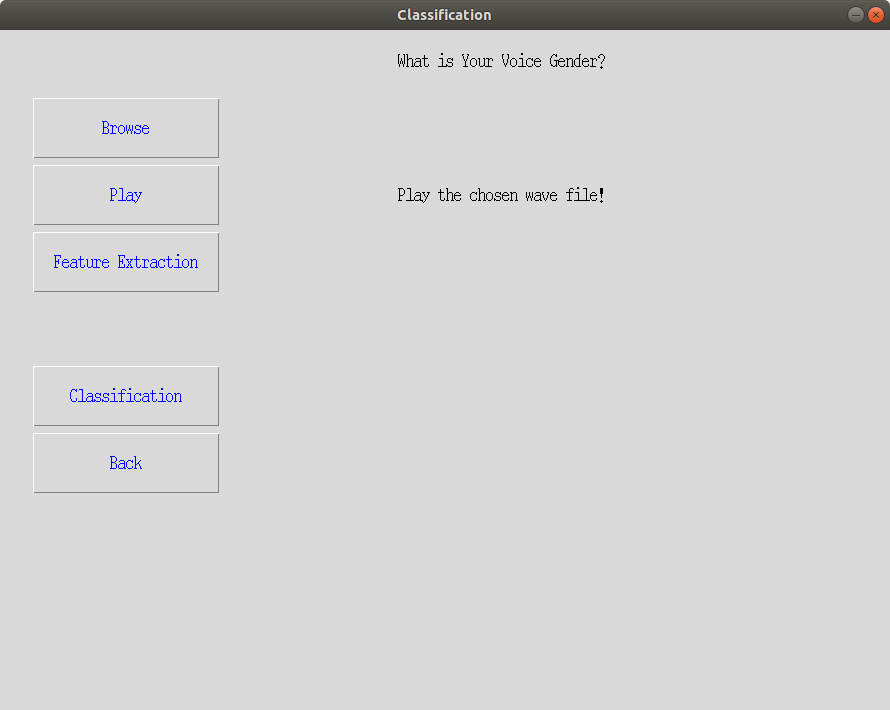


Figure 4.3. Classification Form

User can browse a child’s voice from any folder and upload to the system. After uploading audio, selected audio file can also be listened. MFCC features of this audio are extracted by using feature extraction algorithm and machine learning algorithms classify the gender of the speaker of this uploaded file. Firstly, browse button displays open file dialog box as shown in Figure 4.4. 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.

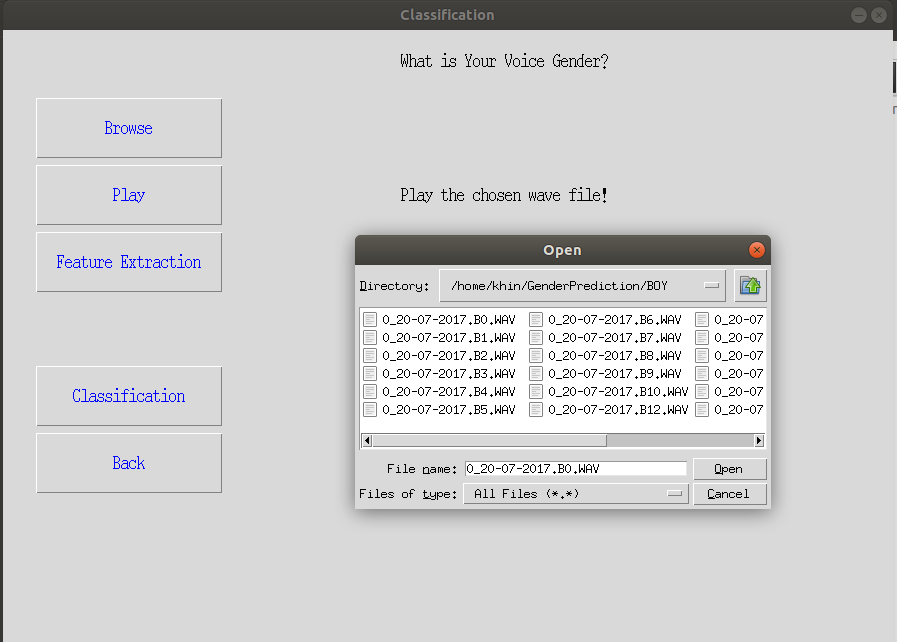


Figure 4.4. Browsing an Audio File from a Folder

The sound of chosen audio file can be listened. Figure 4.5 shows the screen of playing the wave file selected in a folder of the computer. After that, MFCC feature extraction method extracts speech features from the selected audio file.



Figure 4.5. Playing Selected Audio File

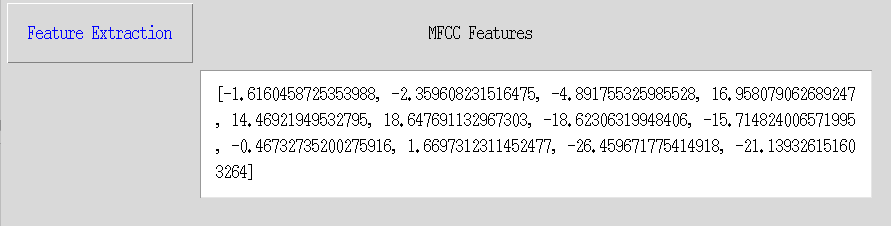


Figure 4.6. MFCC Features

Figure 4.6 is MFCC features points of selected wave file resulted by MFCC method. And then classification can be done based on these extracted features. Figure 4.7 shows classification results of five machine learning algorithms applied in the proposed system after testing speech of a boy.

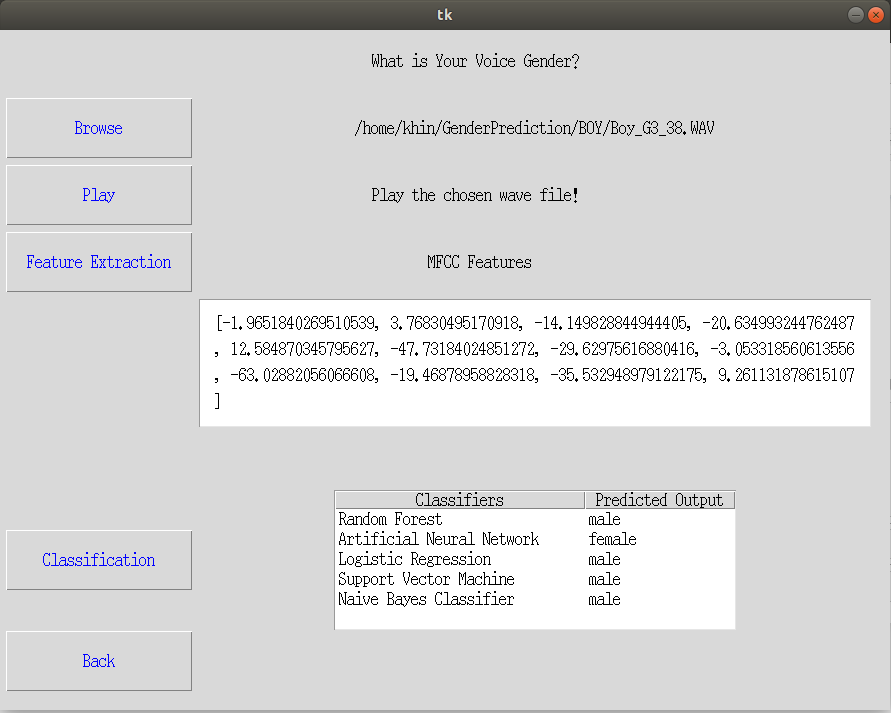


Figure 4.7. Results of Classifying a Boy’s Speech

4.2.2. Performance Analysis

Figure 4.8 shows performance analysis page for machine learning algorithms used for gender classification. In this experimental study of classification algorithms, performance of machine learning algorithms can be analyzed for gender classification using MFCC 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. The system displays training and testing accuracy and other performance measures of selected classifier.

Classification performance of RF can be seen in Figure 4.8. In this testing, 70% of features dataset is trained and testing is done on 30 % of dataset. Testing dataset includes the number of male is 180 and the number of female which is 180. RF classifier can predict male 140 and female 157 correctly and others are wrongly classified. RF has 99% training accuracy and 82% of testing accuracy. Performance of other classification algorithms can study using different ratio of the train set and test set based on MFCC feature dataset. Accuracies of classification models can change, depending on the ratio of training and testing data set.

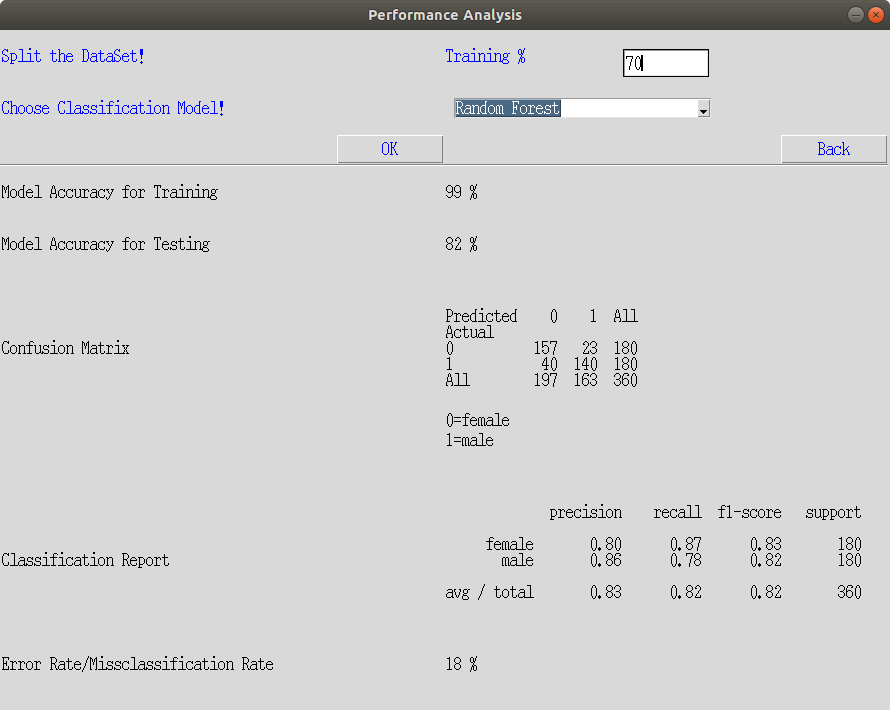


Figure 4.8. Classification Results of RF Classifier

**4.3. Comparison of Performance Evaluations of Different Classifiers**

In this system, voice dataset contains 1200 audio records where recordings of female child contain 600 speech samples and male child recordings have 600 speech samples. The performance of the proposed system is specified, based on classification results for testing data set. Two tests are done to estimate the performance of the machine learning models: simple train-test split and k-fold cross validation method. 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 using voice features dataset. After training, the model accuracy is checked on some test dataset. Splitting test dataset and train dataset can make by any ratio and size of datasets can be declared as test size and train size. 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 beginning, all of the stated classifier models were trained and tested for 70% and 30% of the total dataset respectively. Then in different ratios of the experiment, it was changed to 80% with 20% and 90% and 10%. Confusion matrices resulting after using these data set ratios can be seen in the following tables. Table 4.2 and table 4.3 are confusion matrices for data ratio 70:30. Table 4.4 and table 4.5 display classification results of 80:20. Confusion matrices for 90:10 can be seen in table 4.6 and table 4.7.

Table 4.2. Confusion Matrix of RF, ANN, SVM for Data Ratio 70:30

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RF | | ANN | | SVM | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 149 | 31 | 147 | 33 | 142 | 38 |
| Actual: female | 26 | 154 | 37 | 143 | 47 | 133 |

Table 4.3. Confusion Matrix of LR and GNB for Data Ratio 70:30

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR | | GNB | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 146 | 34 | 137 | 43 |
| Actual: female | 53 | 127 | 43 | 137 |

Table 4.4. Confusion Matrix of RF, ANN, SVM for Data Ratio 80:20

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RF | | ANN | | SVM | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 90 | 30 | 91 | 29 | 91 | 29 |
| Actual: female | 15 | 105 | 18 | 102 | 35 | 85 |

Table 4.5. Confusion Matrix of LR and GNB for Data Ratio 80:20

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR | | GNB | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 86 | 34 | 90 | 30 |
| Actual: female | 42 | 78 | 29 | 91 |

Table 4.6. Confusion Matrix of RF, ANN, SVM for Data Ratio 90:10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RF | | ANN | | SVM | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 46 | 14 | 39 | 21 | 46 | 14 |
| Actual: female | 5 | 55 | 8 | 52 | 11 | 49 |

Table 4.7. Confusion Matrix of LR and GNB for Data Ratio 90:10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR | | GNB | |
| Predicted: male | Predicted: female | Predicted: male | Predicted: female |
| Actual: male | 42 | 18 | 45 | 15 |
| Actual: female | 50 | 10 | 13 | 47 |

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. Classification accuracy of a machine learning classifier can be calculated by using values of confusion matrix. Table 4.8 shows training and testing accuracy results of five classifier tested on different ratios.

Table 4.8. Accuracy Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Ratio** | **70:30** | | **80:20** | | **90:10** | |
| Classifiers | Train Accuracy | Test Accuracy | Train Accuracy | Test Accuracy | Train Accuracy | Test Accuracy |
| RF | 99% | 84% | 99% | 81% | 99% | 84% |
| ANN | 83% | 81% | 80% | 80% | 82% | 76% |
| SVM | 76% | 76% | 76% | 73% | 76% | 79% |
| LR | 75% | 76% | 77% | 68% | 77% | 76% |
| GNB | 77% | 76% | 77% | 75% | 76% | 76% |

According to resulting accuracies, it can be seen that train-test ratio 70: 30 achieves highest accuracy for testing. RF classifier has better performance than other classifiers in this system. 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 overall highest accuracy of 84% for feature dataset used in this system. Moreover, RF classifier achieves 83 % for male prediction and 86 % for female prediction. LR has only 75% training accuracy and 76% testing accuracy, SVM achieves 76% and 76% and GNB obtains 77% and 76% respectively.

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:
4. Take the group as a hold out or test data set,
5. Take the remaining groups as a training data set,
6. Fit a model on the training set and evaluate it on the test set,
7. Retain the evaluation score and discard the model.
8. Summarize the skill of the model using the sample of model evaluation score.

Table 4.9. Cross Validation Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifiers | RF | ANN | SVM | LR | GNB |
| 10-fold Cross Validation Score | 80% | 77% | 75% | 73% | 75% |
| 80% | 76% | 70% | 69% | 71% |
| 88% | 75% | 77% | 76% | 78% |
| 85% | 80% | 74% | 75% | 77% |
| 79% | 77% | 78% | 79% | 75% |
| 85% | 76% | 76% | 77% | 78% |
| 75% | 75% | 71% | 74% | 68% |
| 81% | 74% | 79% | 80% | 81% |
| 85% | 78% | 77% | 78% | 77% |
| 89% | 72% | 68% | 68% | 82% |
| Average | 83% | 76% | 75% | 75% | 76% |

Classifiers models are trained and tested for 70% and 30% of the total dataset respectively. Table 4.9 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.

Table 4.10. Performance Matrices

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifiers** | **Gender** | **Precision** | **Recall** | **F1-score** | **Support** |
| RF | Female | 83% | 86% | 84% | 180 |
| Male | 85% | 83% | 84% | 180 |

Table 4.10. Continued;

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ANN | | Female | | 81% | | 79% | | 80% | | 180 |
| Male | | 80% | | 82% | | 81% | | 180 |
| LR | | Female | | 77% | | 77% | | 77% | | 180 |
| Male | | 77% | | 77% | | 77% | | 180 |
| SVM | | Female | | 78% | | 74% | | 76% | | 180 |
| Male | | 75% | | 79% | | 77% | | 180 |
| GNB | | Female | | 76% | | 78% | | 77% | | 180 |
|  | Male | | 78% | | 75% | | 76% | | 180 | |

Table 4.10 gives performance measures of each classifiers calculated on feature dataset divided into 70% training and 30% 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.