Classifying Human Gender by Learning the Acoustic Features of Voice Samples

Sandeep Kumar
University *School of ICT*
GBU, Greater Noida
sk136398@gmail.com

Rahul Mishra
Department of computer science
C-DAC, Noida
rahulmishra@cdac.in

Bhawana Tyagi
Banasthali Vidyapith, Rajasthan
bhawana1988@gmail.com

***Abstract*:**

Human beings have various capability that is helpful for acquiring different types of knowledge form environment or surroundings. In which of them to find human gender by their voice is an easy task to the human because human as a growing phase of life acquires knowledge form surrounding and listening data. But when it comes to the computer, then it becomes a difficult task to identify the gender of the human being by their voice. To obtain this task from the machine or robots we need to train that particular machine by providing relevant features, methodologies, and diverse training dataset. In this research paper, we proposed a model by combining the Naïve Bayes and deep learning methods for gender identification. In this work, we consider the acoustic features of voice to classify the males and females speech. To persuade this work, we collect around 2000 voice sample form different sentence acoustic from classmates and different website voice samples and used that voice samples as an input dataset form our proposed model. The duration of voice in dataset is greater than 3 seconds to less than 8 seconds. For the model training and testing we used dataset in the ratio of one-third (66% samples for training and 33% samples for testing). Dataset have 50-50% portion for both gender voices. The proposed model is based on two classifier name as Neural Network classifier and Naïve Bayes classifier. These two classifier gives prominent result with this dataset. The accuracy of naïve Bayes is 98% and neural Network gives 99%.

Keywords: *speech signal processing, turnR library, humans gender recognition, Neural Network, Naïve Bayes, voice features, machine learning*

Number: 10.14704/nq.2022.20.7.NQ33199 Neuro Quantology 2022; 20(7):1600-1606

.

I. INTRODUCTION

In the field of Natural Langage proceing, either text or voice is an important roles in this digital advanced world. Nowadays various research work on the filed of voice based gender identification. We can determine the gender of a person by looking at the acoustic characteristics of speech data, such as pitch, energy signal, and fundamental frequency, among other things. Machine learning and deep learning provide motivated, satisfying, and beneficial outcomes for categorization and prediction issues in all major current study domains. Although we all understand, voice is still the most common means to communicate information. It gives a solid approach for conveying a message and may express gender, age, mood, word sequencing, and quasi speech. Gender Identification Technique (GIT) is a technique for predicting gender, or determining whether a speech is uttered by a woman or a man.

The proposed system process the speech signals to get acoustic attributes of speech and analyze the input. Later compares the input with the trained data and classifies the input by applying some computation method and provides the contiguous similar production. Human-computer interaction, voice recognition, speaker recognition, gender-specific advertisement, classifying audio by tools or equipment and defining and decreasing search area, and automatic greeting can all benefit from accurate categorization of all the above features. This can aid personal assistant such as Google Assistant, Alexa, and Siri in providing a female generic or male generic response to a question. The acoustic properties of voice are discovered via signal processing technologies. To eliminate unwanted signals from speech data samples and obtain characteristics from speech signal. The various type of speech signal features like sound pitch, Mel Frequency Cepstural Coefficient (MFCC), energy of signals, basic called as fundamental frequency, etc. The voice signal must be pre-processed with signal processing techniques. The suggested model is developed using a machine learning method based on these speech qualities.

Here the Figure 1.a and Figure 1.b shown below, represents the signal accourding to time space which was composed by Both female and male speakers have a sampling rate of

16kHz and a 16 bit Pulse Code Modulation Resolution. Woman's voice signals have a greater amplitude and frequency than men voice signal [1].

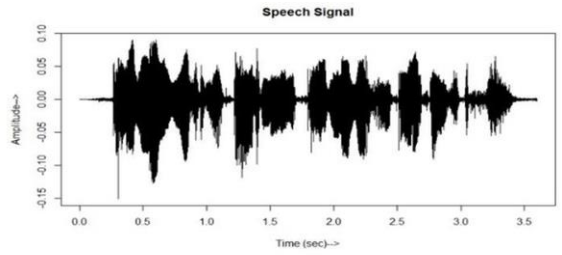

Fig. 1(a). 'Female' speech signal in the time domain

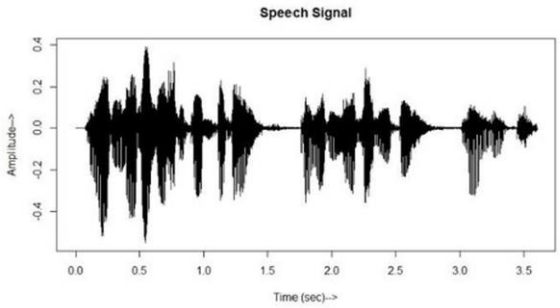

Fig. 1(b). 'Male' speech signal in the time domain

II. LITERATURE REVIEW

In 2016, Mamta Kumari et al proposed an approach to identify the gender, in which they mainly focus on the non-stationary behavior and pitch of speech signals. Authors proposed an algorithm that evaluates the highest pitchIn addition, they employed the FFT (Fast Fourier Transformation) technique to detect the speaker's voice and gender. The evaluation was carried out on 200 voice samples and resulted with a 96 percent accuracy [2]. On a total of 160 voice files, Esther, et al. provided a model that evaluated on three basic parameters such as MFCC, pitch or fundamental frequency, and signals energy. The Support Vector Machine (SVM) classifier are used and it provides 69 percent accuracy on this 160 voice files data [3].

Daryani et al. published a research paper on data mining and its applicability in 2018. They also analysed the properties of acoustic sound in order to use binary classification data mining algorithms to determine the gender of voice files. They utilised a database that contains 3168 samples of male and female voices. They employed a python programming language and 20 attributes. To build a model, they used six distinct strategies. The strategy was "Support Vector Machine (SVM), Random Forest, Logistic Regression, Adabooting, K- nearest neighbor (KNN) and classification and regression tree" [13]. All of these classifiers have an accuracy of over 90% [4]. Gender categorization has a variety of uses, including managing the business and, as a result, enhancing client demand, and allowing robots to comprehend gender, among others. Using acoustic aspects of voice data, P Gupta, S Goel, and A Purwar (2018) created a model that used machine learning methods to predict humans gender based on speech data with the existing machine learning classifier like Neural network with perceptron, Classification and regression tree (CART), and the Random Forest with the 97 percent accuracy [5].

III. THE DATASET

The formulas, in this study, voice samples were gathered in the form of .wav files from various individuals and stored in two distinct folders labelled male and female for the dataset. Some examples of speech sentences are depicted in figure 2 and the procedure to collect the voice from user is shown in figure 3.

1) Hello, my name is_____. I am recording my voice for testing. You are listening my voice.

2) My favorite pet is_____. and my favorite actor is _____.

3) My favorite actress is_____, and my favorite flower is_____.

4) My favorite food is _____and my favorite sweet is_____.

5) I have never visited , and I have never eaten_____.

Fig. 2. Sample Sentences

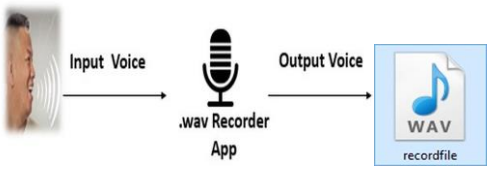

Fig. 3. Voice input collection in audio file format

Each participant in English has recorded a sentence twice. Each participant's voice is sampled ten times in this manner. Basically, we have collected 200

S. No.	Description	Acoustic parameters keyword
1.	Mean Frequency of acoustic signal (in kHz)	Meanfreq
2.	Standard deviation of acoustic signal frequency	Sd
3.	Median frequency of the acoustic signal (in kHz)	Median
4.	First quantile (in kHz)	Q25
5.	Third quantile (in kHz)	Q75
6.	Interquartile range(Q75-Q25) (in kHz)	IQR
7.	skewness shows the asymmetry of the voice frequency spectrum around the sample mean	Skew
8.	Kurtosis shows how much	Kurt
9.	Spectral entropy	Sp.entropy
10.	Spectral flatness	Sp.flatness
11.	Mode frequency	Mode
12.	Frequency centroid of the acoustic signal	Centroid
13.	Average/mean of fundamental frequency measured across the acoustic signal	Meanfun
14.	Minimum fundamental frequency measured across the acoustic signal	Minfun
15.	Maximum fundamental frequency measured across acoustic signal	Maxfun
16.	Average of dominant frequency measured across acoustic signal	Meandom
17.	Minimum of dominant frequency measured across acoustic signal	Mindom
18.	Maximum of dominant frequency that is measured across the acoustic signal	Maxdom
19.	Range of dominant frequency that is measured across the acoustic signal	dfrange
20.	Male or female	Label

voice samples with microphone and .wav file recorder, where these 200 voice samples obtained from 20 candidates in which 10 men, 10 women involved. And rest of 1800 voice data sample downloaded from the CMU ARCTIC website [6].

The CMU ARCTIC dataset have various language voice samples in the .wav file formats like American English male, American English female, Indian English male, Indian English female, Indian Hindi, Gujarati female voice and various other audio files available too.

We were able to acquire 2000 speech samples in this manner. There are 1000 male voice samples and 1000 female voice samples in total. There are a total of 20 attributes, with the class variable name 'label' being one of them. The duration of each voice sample varies from three to eight seconds.

IV. EXPERIMENTAL SETUP

Pre-processing: This is a process for cleaning raw data, such as removing silence and melodic voice samples, as well as voice files that are not in the three to eight second range. The voice recordings were pre-processed in R programming language using the seewave, warbleR, and tunR libraries with something like a range of frequency from 0Hz-280Hz.

Feature Extraction: Feature extraction is an important step. The main aim of this step is to reduce the dimensionality of the dataset. It reduces the features by eliminating those features that are not important and keep the features that are more important for that scenario. The obtained set is the reduced set of features that are used to condense the majority of the data from the initial collection of characteristics. The acoustic characteristics are derived from the voice frequencies and this voice frequencies are transferred from the speech signals in this manner. The acoustic features of voice dataset is discussed in table 1.

Table 1. Acoustic features of voice dataset

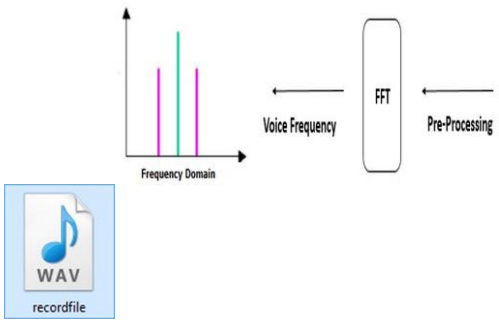

Fig. 4. FFT for domain conversion from time to frequency

Transformation: a mechanism called as FFT. It stand for Fast Fourier Transformation (FFT) which converts time domain signals in the form of frequency domain signals which provides more information instead of time domain signals. The Specan function is a method in the R programming language's warbleR package. Measure acoustic characteristics in groups of sound files, and this function return the values of the following 20 acoustic features/parameters (in double), as well as a two class variable named 'label' (male or female) in table 1 [7]. It shows, the sound features of each voice file from the dataset are calculated using different functions and that calculated values are stored automatically in

.csv file as a pre-trained dataset for model training and testing [15].

Classification: It is used to classify the result in some classes. It's a type of machine learning that refers to an algorithm that uses examples of labelled data to 'learn' how to categorize fresh observations. This technique is a common method for collecting characteristics from raw speech in order to predict a class or label. The machine learning classifier are trained using right label of datasets with all features, before the prediction algorithm can operate. This step is essential when using this technique to enter a training set. The Naive Bayes Classification strategy and the perceptron based Neural Network classification approach are the two primary approaches presented in our research study.

- **Naïve Bayes Classifier:**

It is a simple classifier that is based on conditional probability (Theorem of Bayes). The supervised dataset is used to train this classifier. It works better with the independent variable [13]. The naïve Bayes classifier is easy and less complex based on conditional probabilities instead of neural network, so it is good with less amount of training datasets. By smearing this formula, the decomposition of the conditional probability of naïve Bayes can represented as follows:

$$p(Ck \mid x) = \frac{p(Ck)\, p(x \mid Cx)}{p(x)} \dots (1)$$

Where x is the dependent variable and Ck is the independent variable, and k is one, two, three, and so on.

The above equation can be written as: Using this probability terminology,

$$Posterior = \frac{Prior\ X\ Likelihood}{Evidence} \dots (2)$$

In below Gaussian Naïve Bayes formula [8] where mean (**μ**) and variance (**σ^2**). Because the features value is Continuous value (**$-\infty < x < \infty$**).

$$p(x = v \mid Ck) = \frac{1}{\sqrt{2\pi{\sigma_k}^2}} e^{-\frac{(v-\mu k)^2}{2{\sigma_k}^2}} \dots (3)$$

- **Neural Network Classifier:**

Artificial Neural Network is the imitator of the biological neural network. The idea behind the generation of a neural network begins with the most basic form, a single perceptron [9]. A perceptron can have one or more inputs, activation function, bias, and output. There is a constant value of Bias that makes the model in such a way, which makes it best for the given set of data. By adjusting the weights on the inputs of the neuron and the value of the bias, the Neural Network adjusts the output. The single unit perceptron takes the inputs, multiplies the values of the input by some given weight, adds the value of the bias to it, and then finally passes the value to the activation function that produces an output. There are many numbers of possible activation functions, out of which we have to choose a logistic (sigmoid) function according to our needs, such as the **logistic function**, step function, a trigonometric function, hard limit function, etc. Figure 5 describes the conception of a perceptrons:

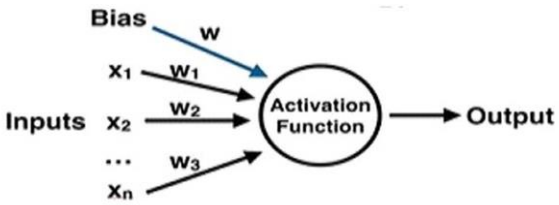

Fig. 5. Basic visualization of Neural Network model

To create a multi-layer neural network. The input layer takes the input from the feature vectors and an output layer will create the resulting outputs and the layers between the input layer and the output layer are known as hidden layers of Neural Network.

V. PROPOSED APPROACH

We propose supervised machine learning classification techniques for Gender Identification System by Voice in this part. Figure 6 depicts our proposed approach. To begin, collect speech samples and pass preprocessed data into a dataset from a different sources. After that, use the Specan method of the warbleR library to extract acoustic features from speech/voice files and cutting the value of each features. The value is recorded in a CSV file together with the class variable. The class variable determines whether a person is male or female. The CSV file is split into two sections, with both the train set including 66% of the dataset and the test set containing 33% of the dataset. The training model uses two classifiers: Nave Bayes and Neural Network classifiers, which are both trained in the model using a train set. The test set is handed to prediction, and the prediction is made using the trained model as a guide. The outcome of the forecast would then be male or female. The confusion matrix must be used to determine the model's classification performance on the test data set.

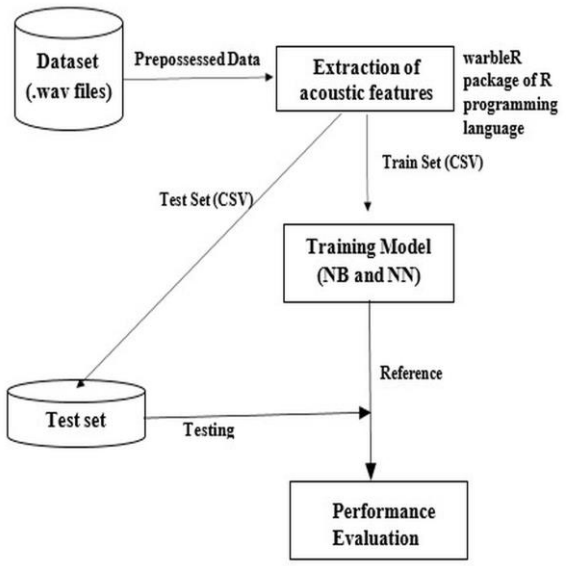

Fig. 6. The proposed approach model graph

VI. RESULTS

This was all accomplished in RStudio IDE. R language which is mostly similar to most popular language Python programming language. It is used in data analysis to better analyse the data. R language has various library that includes including caTools, neuralnet, caret, and e1071 [10].

The proposed model train on collected dataset with the 66 percent of split dataset portion and test on the rest remaining 33 percent of split dataset portion. So the whole dataset samples split in the samples of 1356 and 644 samples of 2000 voice samples of dataset.

Program Code for Naïve Bayes (NB):

```
Data_read<- read.csv("1kclassvoice.csv")
# this line of code Read row column values from csv
str(data_read)
# demonstrated the structure data values
set.seed(2220)
index<- sample(2, nrow(data_read),replace=T, prob=c(0.6,0.3)) train1 <- data_read[index==1,]
test1 <-data1[index==2,] set.seed(2032)
model <- naiveBayes(label~. , data=train1)
# Naive Bayes & model training
pred <- predict(model , test1)
#Testing on model
table(pred)
cm<- confusionMatrix(pred , test1$label)
```

Snippet Output--

A performance evaluation metric name as confusion metric. The confusion metrix is described the performance of classification models. Here the accuracy (true positive + true negative)/total no. of samples so (321+315)/644 = 0.98. The confusion matrix is depicted in figure 7.

```
> cm<- confusionMatrix(pred,test$label)
> cm$table
          Reference
Prediction female male
    female    321    0
    male        8  315
>
```

Fig. 7. Naive Bayes Classifiers Confusion matrix

Program Code for Neural Network (NN):

```
Data_read<- read.csv("1kclassvoice.csv")
# this line of code Read row column values from .csv file
label <- ifelse(data_read$label=="male", 0, 1)
#change male/female to 0/1
Data_read$label=NULL
Data_read = cbind(data_read,label)
#data partition, training set and testing set in the ratio of Neural Network classifier using logistic function and one hidden layer with three nodes (numbers 66 and 33).
library(neuralnet)
nn=neuralnet(label~.,data=train, hidden=c(3),act.fct = "logistic", linear.output = FALSE)
plot(nn) #The plotted figure 6(a) of Neural Network
predicted <- predict(nn,test[1:20]) # To Test that trained model
predicted<-sapply(predicted,round,digits=0)
table(test$label,predicted)
```

Snippet Output:

The confusion matrix is described the performance of classification models. Here the accuracy (true positive + true negative)/total no. of samples so (312+329)/644 = 0.99. The confusion matrix is depicted in figure 8. Weight and the bias of the neural network is depicted in figure 9.

```
> #testing
> pred <- predict(nn, test)
> predicted <- sapply(pred,round,digits=0)
> table(test$label,predicted)
   predicted
      0   1
  0 312   3
  1   0 329
>
> c=((312+329)/644)*100
> accuracy="%"
> paste(c,accuracy)
[1] "99.5341614906832 %"
>
>
```

Fig. 8. Neural Network Confusion matrix

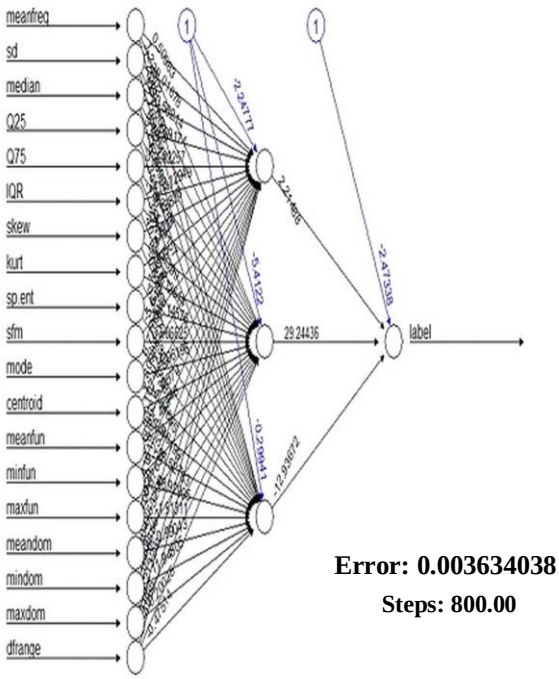

Fig. 9. Neural Network with weight and bais

The performance metric in the form of accuracy of the Nave Bayes & Neural Network models is matched, and the optimal model is the one that performs the best with our dataset that is neural network. The goal of combining the two methods is to get the optimal humans gender recognition. The ideal approach can aid in the development of future applications. The aforementioned two approaches, Nave Bayes and Neural Network Model, are compared.

The proof of outlier, It the Statistics textbook we use "Stats: Modeling the World" says that John Tukey, the creator of the +/-1.5IQR outlier rule said when asked that he settled on 1.5 because 1 was too small and 2 was too big. So we had checked and replaced the outlier value by the mean of feature value in the dataset.

Table 2: Accuracy of classifiers used.

Approach	Dataset	Accuracy
Naïve Bayes Classifier	On the collected dataset this model uses 66% for model training. On the collected dataset this model uses 33% for model testing.	Accuracy of this model gives 98.30% on training data. Accuracy of this model gives 98.69% on test data.
Neural Network Classifier	On the collected dataset this model uses 66% for model training.	Accuracy of this model gives 100% on training data.
	On the collected dataset this model uses 33% for model testing.	Accuracy of this model gives 99.53% on test data.

The Artificial Neural Classifier has the best accuracy, according to the results came from the model. The best model is the neural network classifier instead of naïve Bayes classifier. User can test your voice on a live demo model via model deployment link [11]. Comparison of the previous gender identification system`s result is shown in table 3.

Table 3. Comparison of previous gender identification system`s result and proposed approach.

Ref	Accuracy	Dataset size	Algorithm
[2]	96%	200 samples	Peak detection algorithm
[3]	69%	160 samples	SVM classifier
[12]	96%	280 samples	SVM classifier and MATLAB software for simulation
[4]	Above 90%	3168 samples	SVM, LR, RF, CART, adaptive boosting and K-nearest neighbor
[5]	96%	950 samples	A stacked machine learning algorithm
Our work	99%	2000 samples	Naïve Bayes and Neural Network

VII. Conclusion and Future work

On the supervised dataset, this research work employed the most commonly used machine learning approaches for binary classification. In compared to other algorithms used in the classification area, because in various research model proof that the Naïve Bayes classifier and Neural Network classifier produce the desirable results on the binary type of problems. Both training and testing sets are created from the dataset. There are 2000 recorded samples in all, 1356 with a 'male' label and 644 with a 'female' label. The accuracy of the Nave Bayes model is up to 98.76 percent, while the Neural Network model is up to 99.53 percent. At the time of training and testing model accounts for 66 percent and 33 percent of the dataset, respectively, the model accuracy is achieved. More characteristics and a preprocessed dataset may use and analyzed using machine learning or advanced deep learning to obtain enhanced accuracy of up to 100%. We'll use this model to examine speech samples captured in diverse environments (noisy and vacuum) in the future [14]. Emotion detection and Speaker Recognition can both benefit from Gender Identification characteristics.

REFERENCES

[1] A Hannahs, S.J. & Davenport, Mike. Introducing phonetics & phonology. New York: Routledge, 2013.

[2] Kumari M. and Nilakshi Talukdar "A new gender detection algorithm considering the non-stationarity of speech signal" 2016 IEEE 2nd International conference on communication, control and Intelligent system (CCIS), pp. 141-146.

[3] Esther Ramdinmawii and V. K Mittal, "Gender Identification from Speech Signal by Examining the Speech Production Characteristics" 2016 IEEE International Conference on Signal Processing and Communication, pp.244-249. (http://ieeexplore.ieee.org/document/7980584/).

[4] Daryani MT, Khabiri H, and Yamini Z, "Application of Data Mining Techniques in the Analysis of Acoustic Sound Characteristics" 2018 Journal of Information Technology & Software Engineering Volume 8, Issue 3, 1000238.

[5] P Gupta, S Goel, and A Purwar, "A Stacked Technique for Gender Recognition through Voice" 2018 IEEE Eleventh International Conference on Contemporary Computing (IC3 2018), pp. 1-3.

[6] Festvox CMU_ARCTIC speech synthesis databases, http://festvox.org/cmu_arctic/ accessed on April 01, 2020, at 8:00 p.m.

[7] Becker K, Identifying the Gender of a Voice using Machine learning http://primaryobjects.com/ accessed on April 10, 2020, at 8:30 p.m.

[8] Harry Zhang, "The Optimality of Naive Bayes" 2004 Proceedings of the Seventeenth International Florida Artificial Intelligence Society Conference, FLAIRS 2004. Vol 2.

[9] Perceptron based algorithms-Neural Network classifier https://www.kdnuggets.com/ accessed on April 22, 2020, at 9:00 p.m.

[10] RStudio Documentation, https://docs.rstudio.com/ accessed on April 15, 2020, at 4:00 p.m.

[11] Live demo of proposed work, https://sandeep16064.shinyapps.io/webapp/ accessed on July 11, 2020, at 3:00 p.m.

[12] S Chaudhary and D.K Sharma, "Gender Identification Based on Voice Signal Characteristics", 2018 IEEE international conference on Advance in computing, Communication Control and Networking(ICACCCN2018), pp. 869-874.

[13] Susmita Ray, "A Quick Review of Machine Learning Algorithms", 2019 IEEE International Conference on Machine Learning, Big data, Cloud and Parallel Computing, India, pp. 35-40.

[14] Mohit Mishra, AK Shukla, "A Survey Paper on Gender Identification System using Speech Signal", 2017 International Journal of Engineering and advanced technology(IJEAT), pp. 165-167.

[15] Z, Kudakwashe, OO Oludayo, "Gender Voice Recognition using Random Forest Recursive Feature Elimination with Gradient Booting Machines", 2018 International Conference on Advances in Big data, Computing and Data Communication Systems (icABCD), Durban, pp. 1-6.