Computers, Materials & Continua
DOI: 10.32604/cmc.2023.030567
Article

Age and Gender Classification Using Backpropagation and Bagging Algorithms

Ammar Almomani[1,2,*], Mohammed Alweshah[3], Waleed Alomoush[4], Mohammad Alauthman[5], Aseel Jabai[2], Anwar Abbass[2], Ghufran Hamad[2], Meral Abdalla[2] and Brij B. Gupta[1,6,7]

[1]Research and Innovation Department, Skyline University College, P. O. Box 1797, Sharjah, UAE
[2]IT-department- Al-Huson University College, Al-Balqa Applied University, P. O. Box 50, Irbid, Jordan
[3]Prince Abdullah Ben Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Al-Salt, Jordan
[4]School of Information Technology, Skyline University College, Sharjah P. O. Box 1797, United Arab Emirates
[5]Department of Information Security, Faculty of Information Technology, University of Petra, Amman, Jordan
[6]Department of Computer Science and Information Engineering, Asia University, Taizhong, Taiwan
[7]Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia
*Corresponding Author: Ammar Almomani. Email: ammarnav6@bau.edu.jo
Received: 29 March 2022; Accepted: 17 June 2022

Abstract: Voice classification is important in creating more intelligent systems that help with student exams, identifying criminals, and security systems. The main aim of the research is to develop a system able to predicate and classify gender, age, and accent. So, a new system called Classifying Voice Gender, Age, and Accent (CVGAA) is proposed. Backpropagation and bagging algorithms are designed to improve voice recognition systems that incorporate sensory voice features such as rhythm-based features used to train the device to distinguish between the two gender categories. It has high precision compared to other algorithms used in this problem, as the adaptive backpropagation algorithm had an accuracy of 98% and the Bagging algorithm had an accuracy of 98.10% in the gender identification data. Bagging has the best accuracy among all algorithms, with 55.39% accuracy in the voice common dataset and age classification and accent accuracy in a speech accent of 78.94%.

Keywords: Classify voice gender; accent; age; bagging algorithms; back propagation algorithms; AI classifiers

1 Introduction

The methods of communication between humans are branching out into many; speaking is one of them. According to the latest technology improvements in our world, speaking has become one of the ways to communicate with machines. Therefore, using the voice is not limited only to humans; there is more to talk to than humans. Voice classification systems are gaining popularity due to their wide range of applications used in various areas, ranging from security services, documentation, and retrieval of content-based information to criminal investigations. Gender discovery is gaining importance

due to the recent studies, which have shown that gender-based speech recognition models perform significantly better than gender-independent models. Over two billion individuals speak the English language. Thus, this language is the most significant. There are six large countries where most of the population are native speakers of the current language. Those countries are named the Anglosphere. Hence, the English language speakers in those countries differ in terms of accents. Estimating the gender and age of a speaker is a real challenge for machines because there are characteristics of humans, making the process harder. Human activities, such as smoking, may also cause some changes.

A method known as bagging predictors aims to create several iterations of a single predictor. To have an aggregated predictor, it seeks to use those versions of the models. When numerical results are projected, the sum is calculated by averaging the versions. When a class is projected, it uses a plurality vote to decide. Multiple nature versions are created by creating bootstrap replicas of the training data. They're created by utilizing them to teach themselves new sets of skills. Regression trees and classification and subset selection are used in linear regression tests, which are carried out on simulated and actual data sets to be worked on. Bagging has proved to be effective in improving accuracy based on the most recent testing results.

The accent was a sensitive point in the research, mainly because, although all of the samples we used spoke English as a native language, they did not speak it with the standard accent. The voice common and speech accent datasets were not very efficient due to a lack of features, and this problem is still open to most researchers. Because talking to a machine does not make contact as simple as it does with a human being, our work is focused on solving this problem to make the relationship closer to normal. We also chose more than gender to be classified, such as age and accent. We aim to give promising results to the classification problem in all research areas, especially neural network algorithms.

The research aims to build a new system called Classifying Voice Gender, Age, and Accent (CVGAA) to classify voices using machine learning based on an adaptive Back Propagation and Bagging algorithm. Moreover, we compare results using accuracy and other measurements to improve the quality of gender classification, determine the accent of speakers, and predict their age.

The remainder of this paper is as follows. Section 2 includes a Literature Review with related works, which provides relevant background on AI classifiers used in classifying voice, gender, and age. Section 3 outlines the new proposed system called Classifying Voice Gender, Accent and Age, which discusses how it works. Section 4 comprises the experiments and results. Finally, Section 5 includes the conclusions and implications for future work.

2 Literature Review

According to the previous features, there was a need for a system to identify the gender, accent, and age of the speakers to understand the speech and their needs. Many models and algorithms are used in this article to help people figure out what gender, accent, or age they are. We have talked about the majority of it below.

Sheik [1], whose work aims to build a gender-based classification system applied to the GMM (Gaussian Mixture Modelling) algorithm with Mel Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstral (SDC) (where the SDC fused model gave satisfactory results on the Vox forge dataset). Nevertheless, when they tested different data sets with different data languages, they were not large, and the accuracy was 80%. While Jiao et al. [2] investigated whether there are a major impact of speaking (Korean, Arabic, or Mandarin) languages as a native language on speaking English as a second language. He investigated the degree of linguistic familiarity.

Furthermore, estimating the speaker's age, they also found that smokers were older than non-smokers of the same age, probably due to the influence of smoking. Researchers discovered a ten-year miscalculation in acoustic characteristics and age estimates because the age of younger adults was overestimated, while older people underestimated their accuracy (71%). The genetic algorithm is also used to identify the gender of a speech by comparing various approaches, such as the combination of fuzzy logic and neural network [3].

To implement a gender-based model, the gender must be correctly identified. Identifying the speaker's gender has been receiving attention for a long period. However, this process has been carried out by employing computer systems in recent years [4–6]. Various studies have shed light on gender recognition through processing sound files [7], Classifying the information related to the speaker's gender is very challenging while performing speech processing. A lot of research is being done on feature extraction and classifiers to improve classification accuracy. However, such accuracy isn't desired yet. The key issues in identifying the speaker's gender are represented in producing robust features and designing a good classifier [8]. The Deeper Long Short-Term Memory (LSTM) Network's structure was employed for predicting gender through analyzing an audio data set. The researcher was able to predict gender successfully by showing a good level of accuracy [9]. Several studies have been conducted to predict a gender by combining the face and voice [10]. Gender recognition is represented by predicting and processing gender-related information through the speaker's speech. It aims mainly to identify gender-based sound characteristics [11].

Singh et al. [12] designed and tested a proposed architecture on a common voice dataset. The proposed architecture consists of a cascade of Convolutional Neural networks (CNN) and Convolutional Recurrent Neural networks (CRNN). It is trained on the Mel-spectrogram of the audios. It targets the most popular English accents (i.e., Australian, Indian, US, Canadian and British accents). It shows an accurate rate of 78.48% when using CNN. It shows an accurate rate of 83.21% when using CRNN.

Parikh et al. [13] proposed a system to detect and convert speech that can conveniently differentiate one accent from another, which has achieved an accuracy of 68.67%. The main motivation was to solve the difficulty of Indians understanding foreign accents and of foreigners understanding the Indian accent. This study offers a novel architecture for identifying accents through employing a cascade of two deep-learning architectures.

In [14,15], they describe an experiment that used Gaussian mixture models (GMM) for automatic classification of the speaker's age and gender, using MFCC features and support vector machines (SVMs), and achieved a GMM super vector overall precision of about 75%.

Alkhawaldeh et al. [16] examined many machine learning algorithms. This model demonstrates that a neural network algorithm, such as SVM, has the highest accuracy in determining a voice signal's gender (male or female). Ramadhan et al. [17] employed the random forest algorithm. The latter algorithm has been used for classifying data by using parameter optimization. In their studies, the latter scholars achieved a performance rate of 96.7%.

Zvarevashe et al. [18] developed a method for recognizing gender based on voice. This method employs the feature selection method by using the Random Forest Recursive Feature Elimination (RF-RFE) algorithm with the Gradient Boosting Machines (GBMs) algorithm for gender classification. Based on the experimental results, GBMs outperform all the comparative algorithms in terms of classification accuracy. It's been proven that GBMs effectively recognise gender based on voice.

Asci et al. [19] focused on specific voice features without having their dynamic interaction investigated. The sampled voice records were processed by performing a machine learning analysis. This analysis was performed using a support vector machine algorithm. The latter scholars found that performing machine learning analysis effectively distinguishes between young people from old ones based on voice. They found that the latter analysis effectively distinguishes between males and females based on voice. They found that the statistical accuracy of the latter analysis was high.

Safavi et al. [20] focused on gender, speaker, and age-group recognition based on speech. Several classification methods were compared in terms of performance, including the Gaussian Mixture Model-Universal Background Model (GMM-UBM), GMM-Support, and Vector Machine (GMM-SVM), and vector-based approaches. For speaker recognition, the error rate decreases as age increases, as one might expect.

Zhong et al. [21] developed a decision tree binary classification algorithm. The latter algorithm can be used for identifying gender-based available speech data. The binary classification model of the decision tree is employed for predicting the gender of the structured speech data. The latter data must be classified and recognized.

Sánchez-Hevia et al. [22], used deep neural networks to improve the functionality of interactive voice response systems by combining gender identification with age group categorization of speech. These deep neural networks have lately shown the ability to successfully differentiate and classify different applications, such as speech feature extraction and selection challenges. To understand the relative performance of various neural network architectures and sizes, a comparison study of the various network topologies and sizes is offered. Mozilla's Common Voice dataset, an open and crowdsourced voice corpus, was used to train and assess the categorization framework. Systems have achieved gender and age group identification errors of around 2% and 20%, respectively.

Buyukyilmaz et al. [23] used a multilayer perceptron deep learning model to detect the gender of a voice-based on its acoustic features and speech. They used a dataset of 3168 recorded samples of human voices for their research. Their categorization model achieved an accuracy of 96.74 percent.

However, many algorithms are used to classify voices based on gender, age, or accent, like Linear Discriminate, K-Nearest Neighbour (KNN), Classification and Regression Trees, Decision Tree, etc. So, we discussed algorithms that were used before", which is about the algorithms used before and explaining their implementation, and "the proposed algorithms", which is about algorithms not used. We think it will be better for classifying voices than the other algorithms used before, as we will discuss in the proposed methodology.

3 Gender, Age, and Accent Classification System

Fig. 1 depicts my proposed Classifying Voice Gender, Age, and Accent (CVGAA) system design and how to classify the age, gender, and accent features of different people. This system is divided into a few stages after selecting 3 different datasets. The results of pre-processing datasets will be sent to the next stage, which will implement our algorithm. Dataset 1 for voice gender, Dataset 2 for voice common, and Dataset 3 for speech Accent archive. Bagging and Back-propagation, which is the first time used in this problem compared with other algorithms used to solve the same problem, which is Random forest, decision tree, and KNN classifiers algorithms that were used before in this area, finally evaluate the result of classification based on Weka tool [24]. Fig. 1 shows the general steps of Classifying Voice Gender, age, and accent (CVGAA), discussed below in detail.

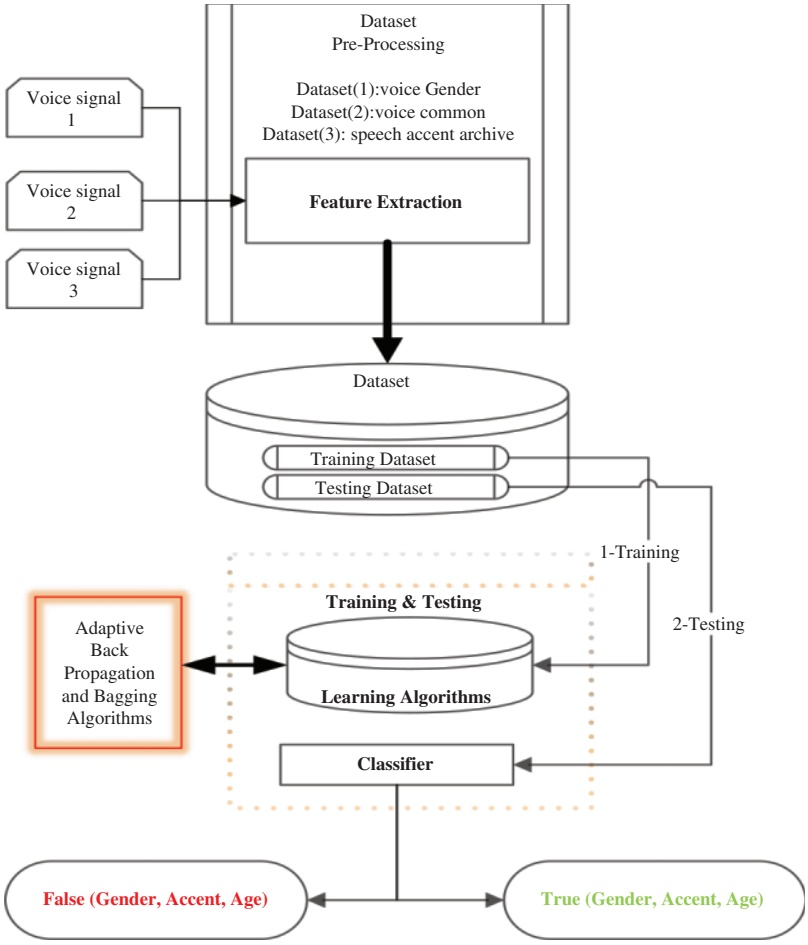

Figure 1: Flowchart scheme in clustering voice

3.1 Pre-Processing Dataset (Feature Extraction)

The statistical features employed in our study are grouped into several classes and have been demonstrated for training and testing data signals. The system will collect audio files and analyses them to extract their features. More than 35 features were extracted from 3 datasets. The system will be built as a vector of features representing a record of feature values used in the next stage based on AI classifiers. However, three datasets will be discussed in section 4 experiments.

3.2 Adaptive Backpropagation and Bagging Algorithm

This section will describe why and how we use the adaptive backpropagation and bagging algorithm to show how it logically and based on experiments has a good result in classifying voice gender, age, and accent, compared with other current classifiers in the same area of research.

The bagging procedure was first proposed by Leo [25]. When wise individuals make critical decisions, they usually consider the opinions of several experts rather than relying on their views. By data mining, a reliable decision-making approach involves combining the results of different models;

Several machine learning methods do that through learning an ensemble of models and employing them in combination: A scheme called "bagging is prominent among these models" [26].

Bagging predictors is a technique that seeks to generate multiple versions of a predictor. It aims to have those versions employed to have a aggregated predictor. The aggregation averages over the versions when having numerical outcomes predicted. It performs a plurality vote when having a class predicted. The versions of multiple nature are developed by designing bootstrap replicates of the learning set. They are developed by employing those as learning new sets.

Regarding tests, they are being performed on simulated and real data sets to use trees of regression and classification and subset selection in a linear regression. Based on the latter tests, B has been proven that Bagging can offer valuable gains in terms of accuracy [25]. Using a given training set the Bagging aims to neutralize the instability of the learning processes. The original training data is altered by removing several instances and replicating others instead of taking a sample of an independent actual training dataset. Instances are sampled randomly from the original data set for the creation of a new one of the same size for the replacement. This sampling technique replicates several instances and has other instances deleted inevitably. When using Bagging, the variance of a prediction is reduced since numerous models (or learners) are being combined that have been trained on various samples of the same data set, rather than just one. The procedure consists of the following steps:

1. dividing the original data into different sets,
2. Using various data sets to train classifiers,
3. For example, you might use the mean, median, or mode of all the models to provide a single answer number depending on the issue at hand.

However, more details can be shown in Fig. 3, which represents a flowchart of the bagging algorithm.

Back-Propagation NNs:

- The study algorithm has 2 phases in a backpropagation neural network.
- First, a pattern is provided for the training input to the network input layer.
- The input pattern is layer by layer until the output layer generates the output pattern. If that pattern differs from the desired output, the error must be calculated. Then, the error will propagate backward through the network from the output to the input layer. Regarding the weight, it is modified as the error gets propagated. Start with a random weight.
- Repeat until the sum of the squared errors is below 0.001 depending on initial weights, final convergence results may vary as shown in Fig. 2.

Fig. 2 represents the flowchart of the bagging algorithm, which included the flow diagram regarding how this algorithm is working with our proposed system starting from the entering dataset until the final prediction output. In our proposed, we have adaptive Back-propagation neural networks (BBNN)-Linear Regression: Fig. 3 represents Linear Regression using a graphical format (Bias b is not shown). The bias is a constant that helps the model so that it best fits the data in the given format. As shown in the diagram below, we have two inputs, as shown in the following diagram (x_1, x_2). The linear combination of the vector is represented by Z. The Z node can also be called a hidden unit because X & Y (for training) are visible, and Z is defined in the model [27].

We can use linear regression to write the equation for predicting values as (this is shown using a blue arrow),

$$\hat{y} = z = b + x1w1 + x2w2 \tag{1}$$

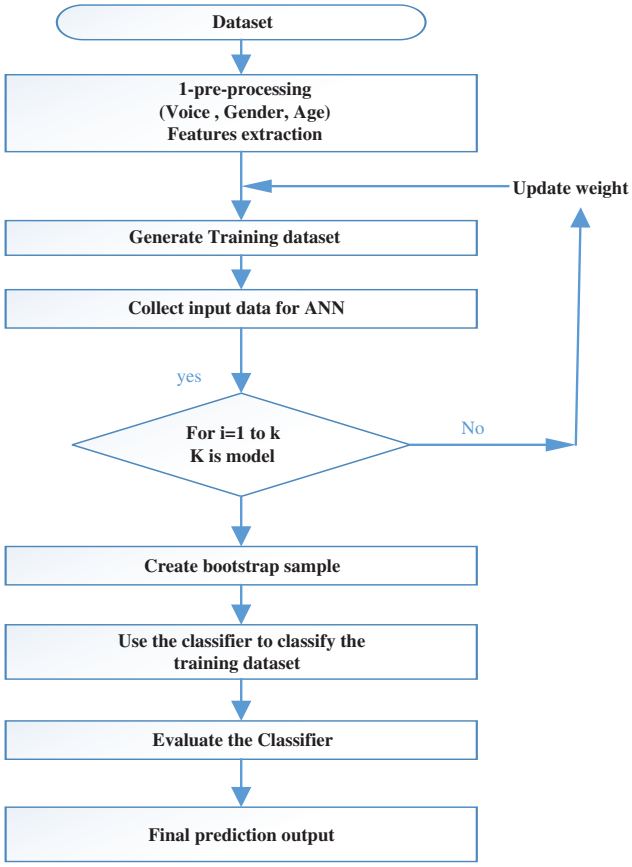

Figure 2: Flowchart of bagging algorithm

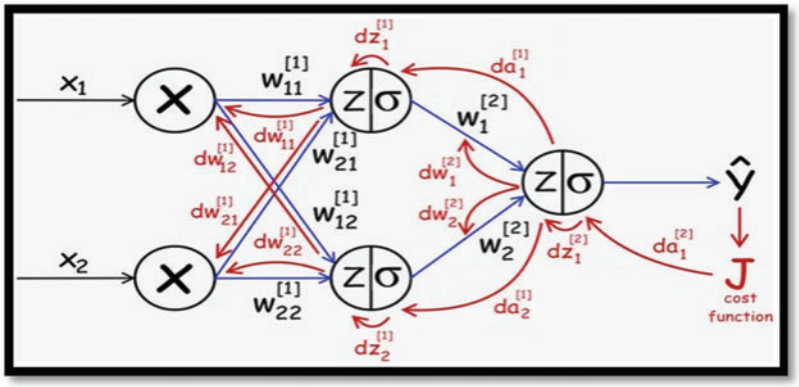

Figure 3: Back-propagation neural network architecture [27]

To find the most effective w, the J cost function must first be Take derivatives of the cost function J as regards w and b, then update w and b to a fraction (learning rate) of dw and db until convergence. We can use the following to write dw and db (using chain rule).

$$dJ/dW = dJ/dZdZ/dW \tag{2}$$

$$dJ/db = dJ/dZdZ/db \tag{3}$$

And for updating W and b there is the gradient descent equation as follow:

$$W =: W - \alpha(dJ/dW) \tag{4}$$

$$b =: b - \alpha(dJ/db) \tag{5}$$

In short, we predict y^ first, then use it to calculate the costs, and then use gradient descent to adjust the model parameters. This takes place in a loop, and we eventually learn about the best predictive parameters (w and b). The same is shown in Fig. 4.

Backpropagation:

The computational complexity is one of the major disadvantages of backpropagation. We have a complex equation to solve only for 2-layer Neural Network with 2 hidden devices in layer one. Imagine that a network of 100 layers and 1000 hidden units is computing complex in every layer. Dynamic programming can be used to solve this problem.

The high standard idea is to express the derivation of $dw^{[l]}$.

(where l is the current layer) using the already calculated values ($dA^{[l+1]}$,$dZ^{[l+1]}$ *etc*) of layer l+1. The backpropagation Algorithm is called in short.

- Initialize $W^{[1]}..W^{[L]},b^{[1]}...b^{[L]}$.
- Set $A^{[0]}=X$ (Input) , L=Total Layers·
- Loop epoch=1 to max iteration ·
  - Forward Propagation·
    - Loop l=1 to $L-1$·
      - $Z^{[l]}=W^{[l]}A^{[l-1]}+b^{[l]}$.
      - $A^{[l]}=g(b^{[l]})$·
      - Save $A^{[l]},W^{[l]}$ in memory for later use ·
  - $Z^{[L]}=W^{[L]}A^{[L-1]}+b^{[L]}$.
  - $A^{[L]}=\sigma(Z^{[L]})$·
  - Cost $J= -1/n(Ylog(A^{[2]})-(1-Y)log(1-A^{[2]}))$·
  - Backward Propagation·
    - $dA^{[L]}=-Y/A^{[L]}+(1-Y)/(1-A^{[L]})$·
    - $dZ^{[L]}=dA^{[L]}\sigma'(dA^{[L]})$·
    - $dW^{[L]}=dZ^{[L]}dA^{[L-1]}$.
    - $db^{[L]}=dZ^{[L]}$.
    - $dA^{[L-1]}=dZ^{[L]}W^{[L]}$.
    - Loop l=$L-1$ to 1·
      - $dZ^{[l]}=dA^{[l]}g'(dA^{[l]})$·
      - $dW^{[l]}=dZ^{[l]}dA^{[l-1]}$.
      - $db^{[l]}=dZ^{[l]}$.
      - $dA^{[l-1]}=dZ^{[l]}W^{[l]}$.
  - Update W and b·
    - Loop l=1 to L·
      - $W^{[l]}=W^{[l]}-\alpha.dW^{[l]}$.
      - $b^{[l]}=b^{[l]}-\alpha.db^{[1]}$

We derive the backpropagation algorithm for a 2-layer network and then for N-Layer Network generalized.

Firstly, we want to find dJ/dW[2] where J is the cost function and W[2] where all the weights are matrixed in the final layer. We can define the following using partial derivatives (follow the background (red color) in Fig. 4, if you are confused)

$$dJ/dW^{[2]} = dJ/dA^{[2]}\ dA^{[2]}/dZ^{[2]} dZ^{[2]} dW^{[2]} \tag{6}$$

We already know $Z^{[2]}$ from our forward propagation,

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]} \tag{7}$$

The derivative of the above $Z^{[2]}$

concerning $W^{[2]}$ will simply be $A^{[1]}$.

$A^{[0]}$ here is nothing but our input X; however, if you have more than 2 hidden layers, it will just be the activation output of the previous later.

Full code description of Back-propagation algorithms shown below based on [27,28].

N-Layer Neural Network Algorithm

By generalizing the equations, we derived on our 2-layer network, we are now defining the complete N-Layer Neural Network Algorithm

4 Experiments and Results

We used Mark et al. [24] to learn from the information in the data. Various data mining and state-of-the-art machine learning techniques are implemented in the Waikato Environment for Knowledge Analysis (Weka). Weka may be downloaded for free from the internet. Accompanying new literature on data mining describes and documents all of the methods included in it in detail. Weka class libraries-written programs may be launched on any computer with an internet connection, regardless of platform. Then users will be able to utilize machine learning techniques on their data independent of the computer platform they are now using [24]. We use a device with an Intel (R) Core (TM) i3-4005U CPU and 4.00 GB of RAM and a 64-bit operating system and Windows 10.

The model's core is bagging with ANN that uses backpropagation to classify instances having the following parameters: number of iterations is set to 600, the learning rate is 0.1 and momentum is set to 0.1. an activation function is sigmoid.

We use the following matrix to evaluate the proposed model

- **Precision** measures the percentage of relevant samples in a given group, which are true positives (tp) and false positives (fp).

$$Precision = tp/(tp + fp) \tag{8}$$

- **Recall** determines the percentage of relevant samples that have been retrieved from the total samples amount.

$$Recall = tp/(tp + fn) \tag{9}$$

- Using the **F-measure**, precision and recall can be measured separately.

$$F\ measure = 2 \times precision\ \times\ recall\ precision + recall \tag{10}$$

4.1 Voice Gender Experiment

This database was created to identify a voice as male or female, based onvoice and speech features. The dataset contains 3300 recorded voice samples, but we chose 950 because there is a lot of missing information for another record, collected from both female and male speakers. The voice samples are pre-processed through performing an acoustic analysis by employing the sound wave and tuner packages, with an analyzed frequency that is within the range of 0–280 Hz (human vocal range) [29]. However, 70% of the dataset was used in the learning phase, while 30% was used in the testing phase for the classification process.

4.1.1 Feature's List

The following feature of each voice shows in Tab. 1:

Table 1: Voice gender experiment features

#	Feature's description
F1	Freq. mean (kHz)
F2	Freq. standard deviation.
F3	Freq. median (kHz)
F4	1st Quantal (kHz)
F5	3ed Quantal (kHz)
F6	Inter-Quantal range (kHz)
F7	Freq. skewness.
F8	Freq. kurtosis.
F9	Entropy of spectral
F10	The flatness of the spectral
F11	Freq. mode.
F12	Freq. centroid.
F13	Peak Freq.
F14	the average of fundamental Freq. measured across the acoustic signal
F15	Foundational minimum Freq. Measured using an acoustic signal.
F16	Maximum Basic Freq. Measured using an acoustic signal
F17	Dominant Freq's average. Measured using an acoustic signal
F18	Dominant Freq minimum. Measured using an acoustic signal
F19	Dominant Freq's maximum. Measured using an acoustic signal
F20	Dominant Freq's range. Measured using an acoustic signal
F21	Indices of modulation. Calculated as the absolute accumulated difference between adjacent measures of basic Freq frequencies.
Class	Gender

Fig. 4 shows the Weka diagram for the data features in the backpropagation algorithm to train on a dataset

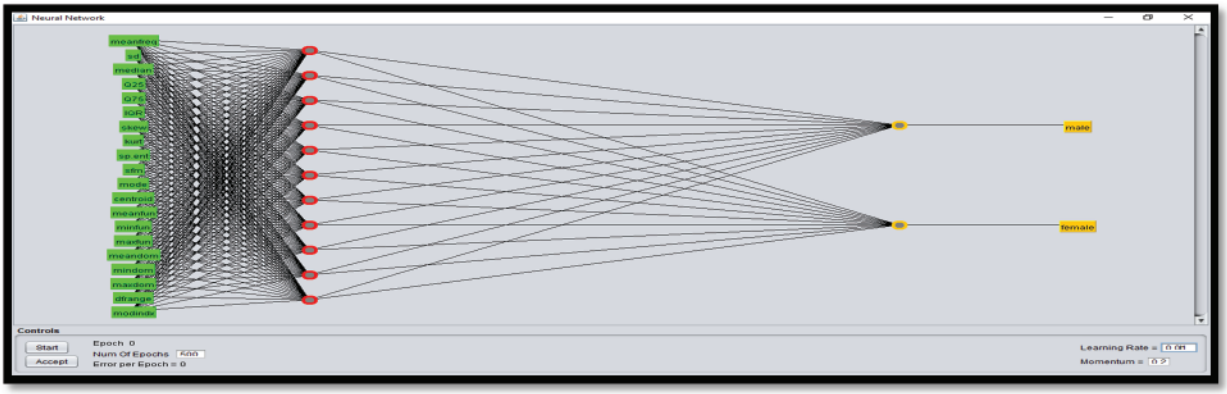

Figure 4: Back-propagation training model

- Tab. 2 shows the results of the voice gender dataset. We have implemented several algorithms that were previously used to compare their results with the results of the algorithms that we selected, and the results were as follows:
- Backpropagation NN and Bagging Backpropagation NN and Bagging results were very good, and most were better or equal to the algorithms used previously, as the accuracy of Backpropagation NN was 98 and Bagging was 98.10 in the gender identification data.
- The mean fundamental frequency serves as an indicator of the gender of the one speaking, with a threshold of 140 Hz. It separates the male classification from the female one.
- Decision Tree was the worst algorithm in voice classification, with an accuracy of 96.42% and a precision of 0.95%.

Table 2: Result of voice gender dataset results

Algorithm	Accuracy		Detailed accuracy by class				Confusion matrix		
Decision Tree	Correctly Classified	916	Class	Precision	Recall	F-Measure	Classified as	A	B
		96.42%							
	Incorrectly Classified	34	Male	0.955	0.975	0.965	A: Female	462	12
		3.57%							
			Female	0.974	0.954	0.964	B: Male	22	454
			Weighted Avg.	0.964	0.964	0.964			
Back propagation	Correctly Classified	931	Class	Precision	Recall	F-Measure	Classified as	A	B
		98%							
	Incorrectly Classified	19	Male	0.971	0.981	0.980	A: Female	475	9
		2%							
			Female	0.981	0.979	0.980	B: Male	10	456
			Weighted Avg.	0.980	0.980	0.980			

(Continued)

Table 2: Continued

Algorithm	Accuracy		Detailed accuracy by class				Confusion matrix		
Bagging	Correctly Classified	932 98.10%	Class	Precision	Recall	F-Measure	Classified as	A	B
	Incorrectly Classified	18 1.89%	Male	0.983	0.979	0.981	A: Female	474	10
			Female	0.979	0.983	0.981	B: Male	8	458
			Weighted Avg.	0.901	0.895	0.887			
Random-Forest	Correctly Classified	932 98.10%	Class	Precision	Recall	F-Measure	Classified as	A	B
	Incorrectly Classified	18 1.89%	Male	0.983	0.979	0.981	A: Female	476	8
			Female	0.979	0.983	0.981	B: Male	10	456
			Weighted Avg.	0.981	0.981	0.981			
KNN	Correctly Classified	932 98.10%	Class	Precision	Recall	F-Measure	Classified as	A	B
	Incorrectly Classified	18 1.89%	Male	0.983	0.979	0.981	A: Female	474	10
			Female	0.979	0.983	0.981	B: Male	8	458
			Weighted Avg.	0.901	0.895	0.887			
DNN	Correctly Classified	917 96.5%	Class	Precision	Recall	F-Measure	Classified as	A	B
	Incorrectly Classified	32 3.4%	Male	0.956	0.976	0.966	A: Female	463	11
			Female	0.975	0.955	0.965	B: Male	22	454
			Weighted Avg.	0.966	0.966	0.966			

4.2 Voice Common Experiment

Regarding the common voice, it is a corpus of speech data that is read by a user on the Common Voice website [30]. It is based on text from several public sources (e.g., user-submitted blog posts, known movies, and books). Its main goal is represented by automatically allowing the processes of testing and training the systems used to recognise speech [30]. However, 70% of the dataset is used in the learning phase while 30% is used in the testing phase for the classification process, and the data is split into various parts to achieve convenience:

Dev-for experimentation and development

Train-used for training in speech recognition

Test-for word error rate testing

4.2.1 Features

In this experiment, we have selected 7 features. Each subset's audio clips are saved as mp3 files in the same directories as their related CSV files. Thus, all the pieces of audio data obtained from the

valid train set will remain within the folder "cv-valid-train" alongside the "cv-valid train.csv" metadata file [30].

Tab. 3 shows that Bagging has the highest accuracy of any algorithm, with 55.39% when used on 1,224 samples for age classification, followed by Random-Forest, Accuracy = 54.5, KNN Accuracy = 54, and Back-propagation NN algorithms Accuracy = 46.5. However, when it comes to precision measurement, the highest precision was found in the twentieth, the fifties, sixties, and forties classes, respectively.

Table 3: Result of voice common dataset and age classification results

Algorithm	Accuracy		Detailed Accuracy by Class				Confusion Matrix				
Decision Tree	Correctly Classified	674 55.0654%	Class	Precision	Recall	F-Measure	Classified as	A	B	C	D
	Incorrectly Classified	550 44.9346%	The twenties	1.000	1.000	1.000	A: Twenties	74	0	0	0
			Fifties	0.581	0.251	0.351	B: Fifties	0	54	127	34
			Forties	0.510	0.920	0.656	C: Forties	0	18	472	23
			Sixties	0.565	0.175	0.268	D: Sixties	0	21	327	74
			Weighted Avg.	0.571	0.551	0.489					
Back-propagation NN	Correctly Classified	654 53.4314%	Class	Precision	Recall	F-Measure	Classified as	A	B	C	D
	Incorrectly Classified	570 46.5686%	The twenties	1.000	1.000	1.000	A: Twenties	74	0	0	0
			Fifties	0.714	0.093	0.165	B: Fifties	0	20	127	68
			Forties	0.510	0.920	0.656	C: Forties	0	1	472	40
			Sixties	0.449	0.209	0.285	D: Sixties	0	7	327	88
			Weighted Avg.	0.554	0.534	0.463					
Bagging	Correctly Classified	678 55.3922%	Class	Precision	Recall	F-Measure	Classified as	A	B	C	D
	Incorrectly Classified	546 44.6078%	The twenties	1.000	1.000	1.000	A: Twenties	74	0	0	0
			The fifties	0.759	0.205	0.322	B: Fifties	0	44	122	49
			Forties	0.510	0.903	0.652	C: Forties	0	12	463	38
			Sixties	0.527	0.230	0.320	D: Sixties	0	2	323	97
			Weighted Avg.	0.589	0.554	0.501					
Random-Forest	Correctly Classified	667 54.4935%	Class	Precision	Recall	F-Measure	Classified as	A	B	C	D
	Incorrectly Classified	557 45.5065%	The twenties	1.000	1.000	1.000	A: Twenties	74	0	0	0
			Fifties	0.714	0.186	0.295	B: Fifties	0	40	122	53

(Continued)

Table 3: Continued

Algorithm	Accuracy		Detailed Accuracy by Class				Confusion Matrix				
				0.506	0.850	0.635					
			Forties				C: Forties	0	14	436	63
				0.502	0.227	0.357					
			Sixties				D: Sixties	0	2	303	117
			Weighted Avg.	0.571	0.545	0.501					
KNN	Correctly Classified	670	Class	Precision	Recall	F-Measure	Classified as	A	B	C	D
		54.7386%	The twenties	1.000	1.000	1.000					
	Incorrectly Classified	554					A: Twenties	74	0	0	0
		45.2614%		0.729	0.200	0.314					
			Fifties				B: Fifties	0	43	120	52
				0.507	0.873	0.642					
			Forties				C: Forties	0	14	448	51
				0.505	0.249	0.333					
			Sixties				D: Sixties	0	2	315	105
			Weighted Avg.	0.575	0.547	0.500					

- ***Speech Accent Archive Experiment:***

Everyone who speaks a language speaks to it with an accent. The speech accent archive was established to exhibit a big set of speech accents uniformly from various language backgrounds. Native speakers of English and non-native ones read the same paragraph written in English. Their reading voices will not be recorded.

This dataset contains 2140 speech samples. We have selected 171 samples in the English language because of other samples in another language, each from various talkers reading the same reading passage. The speakers were chosen from 177 countries. They have 214 different native languages. Each of the speakers will be speaking in English. However, 70% of the dataset was used in the learning phase, while 30% was used in the testing phase for the classification process.

This dataset contains the following files [31]:

- speakers_all.csv: demographic information on every speaker.
- We have adopted the English language in this data and for this was the size of the data was 579

Features:

Fig. 5 shows the Features extracted from the Speech Accent Archive, which consists of 6 main features as follows.

Tab. 4 shows the accent accuracy in a speech accent dataset, compares it with the bagging algorithm. The Bagging gives the best accuracy among the algorithms with 78.94%, where Total Number of Instances 171 and Ignored Class Unknown Instances 36. In contrast, the other algorithms have the lowest results, following KNN with a percentage of 63.74%, with Total Number of Instances 171 and Ignored Class Unknown Instances 62, Decision Tree with a percentage of 69.59%, Total Number of Instances 171, and Ignored Class Unknown Instances 52, Random Forest with a percentage of 77%, Total Number of Instances 171 and Ignored Class Unknown Instances 22. When compared to other machine learning algorithms, our suggested model had the highest classification accuracy, however there is still some ambiguity in the Common Voice dataset between age groups forty and sixty. The

accuracy of the predictions was around 51%, with the majority of the predictions being confounded by these ages. There is still a lot of work to be done to clear up the uncertainty between the 40 and 60-year-old age groups.

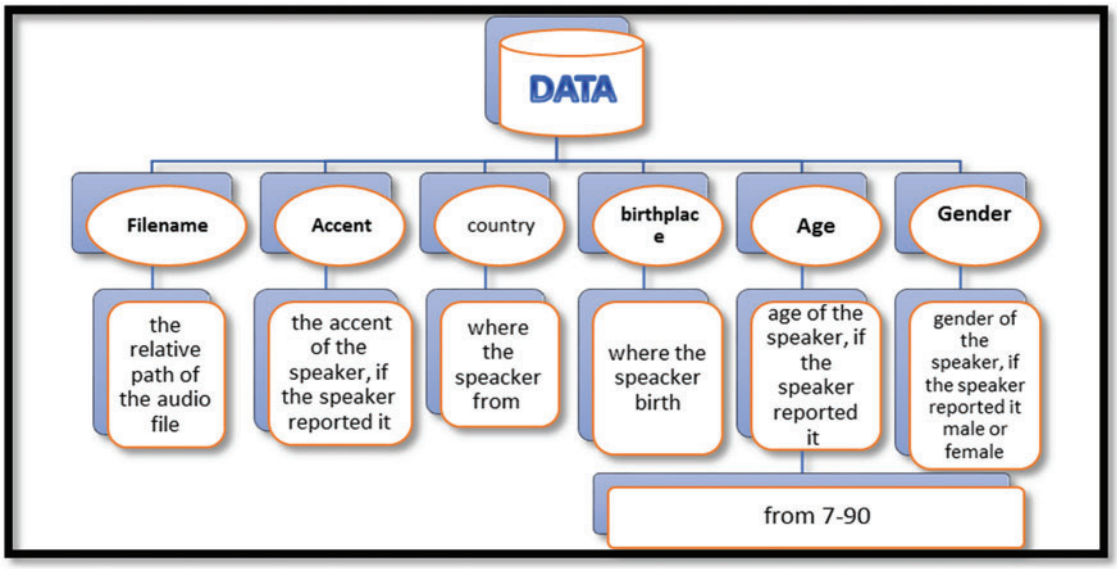

Figure 5: Features extracted from the speech accent archive [31]

Table 4: The accent accuracy in a speech accent dataset and compare it with the bagging algorithm.

ACCENT classification results		
Algorithm		Accuracy
Bagging	Correctly Classified	135 (78.9474%)
	Incorrectly Classified	36 (21.0526%)
KNN	Correctly Classified	109 (63.7427%)
	Incorrectly Classified	62 (36.2573%)
Decision Tree	Correctly Classified	119 (69.5906 %)
	Incorrectly Classified	52 (30.4094 %)
Random Forest	Correctly Classified	132 (77.193 %)
	Incorrectly Classified	39 (22.807 %)
Deep learning	Correctly Classified	125 (73.0994%)
	Incorrectly Classified	46 (26.9005%)

5 Conclusion

In this work, we proposed Classifying Voice Gender, age, and accent (CVGAA), by applying Back-propagation and Bagging Algorithms on datasets based on gender, age, and accent as classes after tremendous studies to make voice recognition clearer and more understood. The results we achieved vary according to the differences in the working mechanism of the algorithms. Some of them were unexpectedly very good. Still, on the other hand, we got the opposite. We used three known datasets used in three experiments to prove our objectives. The first dataset was used in the Voice Gender Experiment, which was designed to identify a voice as male or female based on voice and speech features. The accuracy of Backpropagation NN was 98 and Bagging was 98.10 in the gender identification data. The second dataset used in the Voice Common Experiment, Common Voice, is a corpus of speech data read by users on the Common Voice website. The results show that Bagging has the highest accuracy of all algorithms, gaining 55.39% when used on 1,224 samples for age classification. The third dataset used with the Speech Accent Archive Experiment was to know people based on language backgrounds. The classification result shows that Bagging gives the best accuracy between the algorithms with a percentage of 78.94%. A future study could investigate newer architectures, like wav2vec2.0, for obtaining embeddings from raw waveforms. The proposed methodology might easily be extended to other languages utilizing the CommonVoice or the recently published MLS datasets.

Funding Statement: This work is supported by the Research and Innovation Department, Skyline University College, University City of Sharjah – P.O. Box 1797 - Sharjah, UAE. Grant numbers: 1-DRI- 000001- 2022-123, Dr.ammar Almomani https://www.skylineuniversity.ac.ae/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Sheikh, "Who is speaking? Male or female - a dissertation submitted to the university of manchester for the degree of master of science in the faculty of engineering and physical sciences," *University of Manchester*, vol. 1, no. 1, pp. 1–79, 2013.

[2] D. Jiao, V. Watson, S. G.-J. Wong, K. Gnevsheva and J. S. Nixon, "Age estimation in foreign-accented speech by non-native speakers of English," *Speech Communication*, vol. 106, no. 1, pp. 118–126, 2019.

[3] K. Adi, S. Pujiyanto, R. Gernowo, A. Pamungkas and A. B. Putranto, "Identifying the developmental phase of Plasmodium falciparum in malaria-infected red blood cells using adaptive color segmentation and back propagation neural network," *International Journal of Applied Engineering Research*, vol. 11, no. 15, pp. 8754–8759, 2016.

[4] M. Li, C.-S. Jung and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Eleventh Annual Conf. of the Int. Speech Communication Association*, Makuhari, Chiba, Japan, pp. 1–4, 2010.

[5] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 1–18, 2022.

[6] K. Chachadi and S. Nirmala, "Voice-basedgender recognition using neural network," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Jaipur, India: Springer, pp. 741–749, 2022.

[7] I. E. Livieris, E. Pintelas and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 492–503, 2019.

[8] Z. Qawaqneh, A. A. Mallouh and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, no. 1, pp. 5–14, 2017.
[9] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, vol. 156, no. 1, pp. 351–358, 2019.
[10] S. M. Huestegge and T. Raettig, "Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization," *Journal of Voice*, vol. 34, no. 3, pp. 487-e1, 2020.
[11] C. Chen, W. Wang, Y. He and J. Han, "A bilevel framework for joint optimization of session compensation and classification for speaker identification," *Digital Signal Processing*, vol. 89, no. 3, pp. 104–115, 2019.
[12] U. Singh, A. Gupta, D. Bisharad and W. Arif, "Foreign accent classification using deep neural nets," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6347–6352, 2020.
[13] P. Parikh, K. Velhal, S. Potdar, A. Sikligar and R. Karani, "English language accent classification and conversion using machine learning," in *Proc. of the Int. Conf. on Innovative Computing & Communications (ICICC)*, University of Valladolid, Spain, pp. 1–5, 2020.
[14] G. Dobry, R. M. Hecht, M. Avigal and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.
[15] J. Přibil, A. Přibilová and J. Matoušek, "GMM-based speaker age and gender classification in Czech and Slovak," *Journal of Electrical Engineering*, vol. 68, no. 1, pp. 3–12, 2017.
[16] R. S. Alkhawaldeh, "DGR: Gender recognition of human speech using one-dimensional conventional neural network," *Scientific Programming*, vol. 2019, no. 1, pp. 1–12, 2019.
[17] M. M. Ramadhan, I. S. Sitanggang, F. R. Nasution and A. Ghifari, "Parameter tuning in random forest based on grid search method for gender classification based on voice frequency," in *2017 Int. Conf. on Computer, Electronics and Communication Engineering (CECE 2017)*, Sanya, China, pp. 625–629, 2017.
[18] K. Zvarevashe and O. O. Olugbara, "Gender voice recognition using random forest recursive feature elimination with gradient boosting machines," in *2018 Int. Conf. on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, IEEE, Durban, South Africa, pp. 1–6, 2018.
[19] F. Asci, G. Costantini, L. Di, P. Zampogna, A. Ruoppolo *et al.,* "Machine-learning analysis of voice samples recorded through smartphones: The combined effect of ageing and gender," *Sensors*, vol. 20, no. 18, pp. 5022–5033, 2020.
[20] S. Safavi, M. Russell and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," *Computer Speech & Language*, vol. 50, no. 1, pp. 141–156, 2018.
[21] B. Zhong, "Gender recognition of speech based on decision tree model, in 3rd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2019)," *Atlantis Press*, vol. 90, no. 1, pp. 1–7, 2019.
[22] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso and M. Rosa-Zurera, "Age and gender recognition from speech using deep neural networks," in *In Workshop of Physical Agents*, Cham, Springer, pp. 332–344, 2020.
[23] M. Buyukyilmaz and A. O. Cibikdiken, "Voice gender recognition using deep learning," in *2016 Int. Conf. on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*, Paris, France, Atlantis Press, pp. 409–411, 2016.
[24] H. Mark and F. Eibe, "Geoffrey holmes, bernhard pfahringe," *Weka: The Workbench for Machine Learning*, vol. 11, no. 1, pp. 10–18, 2009.
[25] B. Leo, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[26] I. H. Witten, E. Frank and M. A. Hall, "Data mining: Science direct," *Practical Machine Learning Tools and Techniques (Morgan Kaufmann)*, vol. 2, no. 4, pp. 587–605, 2005.
[27] A. Jana, "Understand and implement the backpropagation algorithm from scratch in python," *A Developer Diary*, 2019. [Online]. Available: http://www.adeveloperdiary.com/data-science/machine-learning/,.
[28] R. Rojas, "The backpropagation algorithm," in *Neural Networks: Springer*, vol. 10, no. 5, pp. 149–182, 1996.

[29] K. Becker, "Gender recognition by voice, Identify a voice as male or female," 2016. [Online]. Available: https://www.kaggle.com/primaryobjects/voicegender,.
[30] M. Team, "Common Voice is Mozilla's initiative to help teach machines how real people speak," 2018. [Online]. Available: https://voice.mozilla.org/en,.
[31] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," *Telecommunication Systems*, vol. 68, no. 4, pp. 687–700, 2018.