Contents lists available at *Dergipark*

Journal of Scientific Reports-A

journal homepage: *https://dergipark.org.tr/tr/pub/jsr-a*

E-ISSN: 2687-6167 ***Number 57, June 2024***

RESEARCH ARTICLE

Receive Date: 19.01.2024 *Accepted Date: 12.02.2024*

Unified voice analysis: speaker recognition, age group and gender estimation using spectral features and machine learning classifiers

Kaya Akgün[a], Şerif Ali Sadık[b]*

[a]*Kütahya Dumlupınar University, Faculty of Engineering, Computer Engineering Dept., 43100, Kütahya, Türkiye, ORCID: 0000-0001-5678-9643*

[b]*Kütahya Dumlupınar University, Faculty of Engineering, Software Engineering Dept., 43100, Kütahya, Türkiye, ORCID: 0000-0003-2883-1431*

Abstract

Predicting speaker's personal traits from voice data has been a subject of attention in many fields such as forensic cases, automatic voice response systems, and biomedical applications. Within the scope of this study, gender and age group prediction was made with the voice data recorded from 24 volunteers. Mel-frequency cepstral coefficients (MFCC) were extracted from the audio data as hybrid time/frequency domain features, and fundamental frequencies and formants were extracted as frequency domain features. These obtained features were fused in a feature pool and age group and gender estimation studies were carried out with 4 different machine learning algorithms. According to the results obtained, the age groups of the participants could be classified with 93% accuracy and the genders with 99% accuracy with the Support Vector Machines algorithm. Also, speaker recognition task was successfully completed with 93% accuracy with the Support Vector Machines.

Keywords: Gender estimation; Age group estimation; Speaker recognition; Support vector machine; k-nearest neighbors; gradient boosting; Classification and regression tree

* Corresponding author. Tel.:+90 274 443 4330; fax: +90 274 265 20 13.
E-mail address: serifali.sadik@dpu.edu.tr
http://dx.doi.org/10.1016/j.cviu.2017.00.000

.

1. Introduction

With the advancements in machine learning (ML) studies, the use of ML algorithms in applications such as classification, recognition, prediction and forecasting are increasing in recent years[1-5]. The most important parameters for ML applications are data and their features. Machines can be trained with data from past events, just like in the human learning process. Thus, in newly encountered events, the machine can now perform tasks such as prediction or classification, based on the features of the training data. Nowadays, digital data such as text, image, video or sound can be used as ML training data for the solution of many different problems [6].

Among these data types, human beings have been interacting with sound data for a very long time. Sound, which has been our communication tool since the first ages of history, has also enabled us to comprehend many events that exist around us. When the sounds are examined in terms of their sources, they can be examined in three groups. While the natural and artificial sounds we hear from our environment constitute the first two groups, the human voice can be classified as the third group [7]. Examples of natural sounds include thunder, wind, waves, and animal sounds. Artificial sounds, on the other hand, are mostly human-induced examples such as traffic sounds, machine noises and instrument sounds. Finally, when we say human voice, examples such as speaking, singing, coughing and sneezing can be given. In addition, other human-induced sounds such as heart sound and respiratory sound have been the subject of many studies[7-10]. Various studies have been carried out on the recognition, tagging and classification of all these sounds, and the determination of the events that may cause them. With the popularization of ML methods in the recent past, the focus of these studies has been ML applications using audio data [11-14].

Recognition or classification of natural sounds has been widely used in acoustic event classification (AEC) studies. In particular, studies to classify the sounds made by other living things with which we share our world are important for reasons such as analyzing biodiversity and getting to know the ecosystem better. In a study conducted in 2007, support vector machines (SVM) algorithm was used to classify the sounds of bird species given in two different datasets. With the mixed model prepared using different spectral and temporal features and the extracted Mel-frequency cepstral coefficients (MFCC) features the classification accuracy up to 98% was achieved [16]. In another study, classification of frog sounds using threshold-crossing rate, spectral centroid and signal bandwidth as distinctive features was carried out. The k nearest neighbor (kNN) and SVM algorithms were used as classifiers and their performances were compared. As a result of the comparative analysis, 89.05% and 90.30% classification accuracy were obtained by using kNN and SVM algorithms, respectively [17]. The classification of the sounds of bats, which also use sound waves for different purposes, has undoubtedly been a remarkable subject in AEC studies. In a study conducted in 2010, the performance of four different algorithms (SVM, artificial neural networks – ANN, discriminant function analysis – DFA and random forests - RF) were comparatively analyzed to classify the echolocation sounds of bats. In 5 different classification tasks defined, classification process could be performed with accuracies ranging from 84% to 96% [18]. In a study in 2015, syllable features such as frequency modulation, energy modulation, duration of syllable, dominant frequency, oscillation rate were used to classify frog sounds. With the kNN classifier algorithm, classification success was achieved as 90.5% [19]. In another study, Ribeiro et al. used RF and SVM algorithms for classification of fish sounds and obtained 96.9% classification accuracy [20]. Ribeiro et al. used the SVM algorithm to classify the sounds of tomato-pollinating bees and achieved a classification success of 73.39% [21]. In a recent study, the performances of RF, SVM and kNN algorithms for the classification of different animal sounds were compared using three different datasets. The comparative analysis showed that 99% classification success was achieved with SVM and kNN algorithms [22].

In addition to the studies on the classification of natural sounds, the environmental sound classification (ESC) has also been one of the highlights in this area of research. The focus of these studies are mostly publicly available datasets such as ESC-10, ESC-50 [23] and UrbanSound8k [24]. In a comparative study, Mushtaq et al. applied their proposed data augmentation method to all 3 datasets mentioned above and performed ESC with transfer learning method with deep networks. Researchers have achieved classification accuracy of 99% in the ESC-10 and UrbanSound8k datasets, and 97% in the ESC-50 dataset [25].

Undoubtedly, the subject of automatic speaker recognition (ASR) attracts the most attention in studies on sound. Since the larynx size, the anatomy of the vocal cords, the internal structure of the mouth or the anatomy of other organs that influence voice formation may be different in each person, the human voice has distinctive features. In addition, the human voice can convey characteristics such as the ethnicity, age and gender of the person, as well as having personal characteristics [26]. Studies on ASR, which has attracted the attention of researchers for about 50 years, have gained great momentum thanks to developments in digital signal processing and artificial intelligence [27]. Thanks to this automation, automatic voice or speaker recognition methods have been frequently used, especially in authentication, personalization, surveillance and forensic case applications [26].

Speaker recognition, which initially attracted the attention of researchers due to the need in forensic cases [28], continues to be widely studied today thanks to the ease of digital signal processing, extraction of qualified features, and feature selection. Looking at recent studies, Krishnamoorthy et al. obtained 78.20% accuracy by extracting MFCCs as features with limited data set and using Gaussian Mixture Model-Universal Background Model (GMM-UBM) as a classifier. In addition, they tried to overcome the limited data problem with the data augmentation method by adding white noise, and they achieved 80% accuracy with the data with noise added [29]. In another study focusing on the speaker recognition problem using MFCC features and a combination of Generalized Fuzzy Model (GFM) and HMM, a success rate of 93% was achieved [30]. In a study using neural networks (NN), MFCC features were extracted with two different methods and a 93.2% success rate was obtained using the second MFCC feature extraction method proposed in [31]. In 2018, a text-independent speaker recognition problem using a dataset of 24 volunteers was presented as a master's thesis. Feature vectors combining linear prediction cepstral coefficient (LPCC), MFCC, Higuchi fractal dimension (HFD), variance of fractal dimension (VFD), zero crossing rate (ZCR) and number of turns led to a recognition accuracy of 91.6%. [32]. The MFCC features were combined with the power normalization cepstral coefficient (PNCC) for speaker recognition and classification accuracies of 97. 52% and 85% were achieved with the extreme learning machine (ELM) using clean sounds and noisy sounds, respectively in [33]. Ayvaz et al. achieved a classification accuracy of 90.2% using a Multilayer Perceptron (MLP) network as the model for a speaker recognition study with MFCC features derived from the voice data of Turkish speakers. [34]. Another recent research focused on speaker identification using artificial intelligence algorithms and feature extraction methods, specifically MFCC and Multiband spectral entropy (MSE), from speech signals. Machine learning algorithms like k Nearest Neighbors, Random Forest, Deep Neural Networks, and Decision Trees were employed for classification. Experiments conducted on LibriSpeech and ELSDSR databases included speaker identification in a group of 20 participants, among men, among women, and by gender. Notably, using the ELSDSR database, the experiment for speaker recognition by gender achieved a precision of 93.99% [35]. Another paper aimed to boost speaker recognition by utilizing rich audio-visual data. It introduced a two-branch network to learn joint face and voice representations in a multimodal system. Extracted features from the network trained a speaker recognition classifier. Evaluation on VoxCeleb1 dataset yielded a notable 91% identification performance using solely audio features [36].

The brief literature review shows that studies on voice/speaker recognition have been carried out for many years, and with the popularity of artificial intelligence, it has still been a remarkable issue in recent years. In this study, MFCC features were extracted from audio files recorded with 12 male and 12 female participants. In addition, formant frequencies and fundamental frequency features are extracted, and all obtained features are fused. The obtained features were given as input values to train the classifier algorithms of SVM, kNN, Classification and Regression Tree (CART) and Gradient Boosting Classifier (GBC) to estimate the gender and age group and the speaker from voice data. Lastly the test results obtained with 4 different algorithms were compared. This study addresses a comprehensive analysis of speaker trait prediction, specifically focusing on gender and age group estimation using voice data. While prior research has touched upon aspects of speaker recognition, this study uniquely combines hybrid time/frequency domain features (MFCC) with fundamental frequencies and formants, offering a more robust and accurate approach. The integration of these diverse features into a unified pool, coupled with the application of four distinct machine learning algorithms, sets our study apart. Notably, the results showcase

better accuracy then reviewed literature, with 93% precision in age group classification, a 99% precision in gender estimation and 93% accuracy in speaker recognition task, using the Support Vector Machines algorithm.

The rest of the paper was organized as follows. In Section 2, the materials and methods which are the dataset, the feature extraction methods and the machine learning algorithms are explained briefly. In Section 3, the training, validation and test results obtained are given comparatively. In the last section, the results are discussed and ideas for future studies are made.

2. Material and methods

Fig.1 shows the block diagram of this study. Firstly, MFCCs, the fundamental frequencies and the formants were extracted as features from human voice data belong the dataset. The extracted features were fused together and one final feature vector was obtained. With the help of the feature vector obtained, 4 different ML algorithms were trained to predict the age group and gender of the individuals and the classification performances of the ML algorithms were evaluated with the test data excluded from the training procedure.

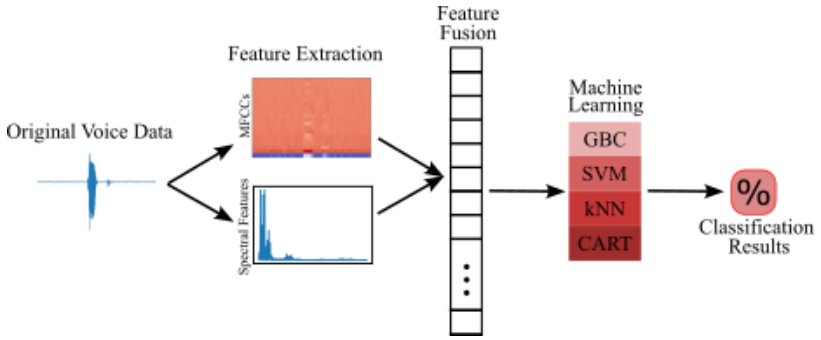

Fig.1. Block diagram of the classification problem in this study.

2.1. Dataset

A publicly available human voice dataset [37] was used in this study. In the data set, there are 1056 waveform audio file format (wav) files for the pronunciation of 44 English words recorded with 24 volunteers. Of the volunteers, 12 are men and 12 are women. While 7 of the female volunteers are younger than 25 years old, 5 are individuals over 25 years old. With male volunteers, there is a 6:6 ratio between individuals older and younger than 25 years old. All 24 volunteers are individuals born and raised in Manitoba, Canada. The sampling frequency of recorded audio files is 44.1 kilo-samples per second. All audio files are 2 seconds long. Fig.2 shows the waveforms of the pronunciations of the word “Book” by two different individuals from the dataset used.

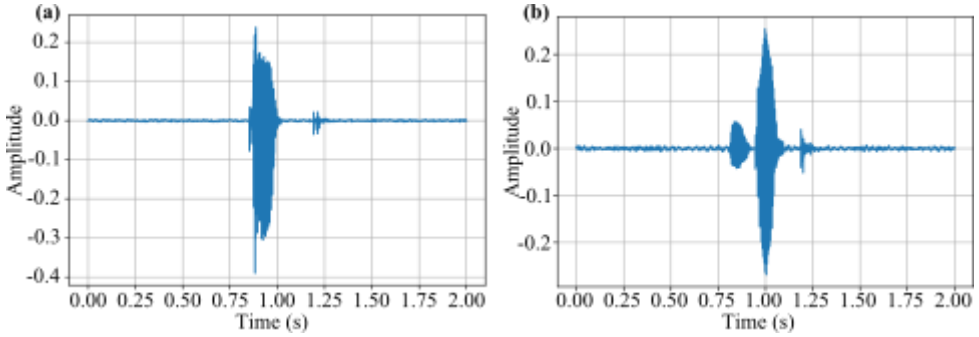

Fig.2. Waveform of the audio file recorded at the moment of pronunciation of the word "Book" (a) for a young male individual (b) for a young female individual.

2.2. Extracted features

The main purpose of machine learning is to train the machine with the help of data from past events and make it capable of making predictions for new or future events. However, raw data from past incidents can contain a lot of information, including some that may not be of use to the machine for its task. Therefore, with feature extraction more informative and distinctive features should be revealed from the raw data for a particular machine learning task. A feature chosen to suit the specific task can represent the raw data in a much more compact way [38]. Successful classification results of machine learning algorithms depend on the extraction of relevant features from the data. By extracting features that are relevant to the target labels, complex data can be analysed more easily, thus shortening the training process of the model. [39].

Research on sound analysis has focused primarily on extracting features from the time domain [40] over time, and as signal processing techniques have improved, features in the frequency domain have gained importance [41]. In recent studies, features that contain information from both time and frequency domains have become prominent [42].

In this study, a hybrid feature pool was created by extracting fundamental frequency and formant features that contain information from the frequency domain and MFCC features that can show changes in time-frequency domains.

2.2.1. Mel-frequency cepstral coefficients (MFCC)

MFCCs visualize the distribution of the energy of an audio signal in the frequency spectrum. MFCCs are extracted from the audio signal using 6 fundamental steps. Firstly, the audio signal is divided into frames, typically 20-30 milliseconds long. By dividing the signals into short "frames", reliable and stable time frames can be obtained where the raw data signal is long and shows frequent changes over time. On each frame, a window which are generally Hanning and Hamming windows, is applied to narrow the signal. Secondly, a power spectrum is computed for each frame. To calculate the power spectrum, Discrete Fourier Transform (DFT) is commonly used. Then the power spectrum is converted to a mel (short for melody) scale, which is a non-linear scale that is more closely aligned with human perception of pitch [43]. The mel scale can be obtained with equation (1).

$$mel = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Then the logarithm values of all mel filter banks is calculated. For the fifth step, a discrete cosine transform (DCT) is applied to the mel-scaled power spectrum to select most accelerative coefficients. And lastly, the first few

DCT coefficients are retained as the MFCCs. MFCCs are typically represented as a vector of 12-20 coefficients. The first few coefficients are the most important, as they capture the most significant information about the spectral envelope. MFCCs are invariant to changes in pitch and volume, which makes them well-suited for speech recognition and other tasks where the speaker or audio source may vary. They are also relatively robust to noise, which is another important property for many applications [44].

2.2.2. Frequency domain features

The distinctive features of audio data may be hidden in characteristic features in the frequency domain rather than changes in the time domain. Therefore, frequency domain features come to the fore in studies on audio signal processing. The most basic tools for extracting frequency domain features are Fourier transform and autocorrelation analysis [38]. In this study, fundamental frequency and formant features were extracted using these tools. The fundamental frequency (f_0), also known as the first harmonic, is the lowest frequency of a periodic waveform and can be determined by autocorrelation analysis using the periodic structure of the signal [45]. Formants are frequency components that result from changes in the shape and size of the human vocal tract. The first three formants (f_1, f_2, f_3) are the most important for speech recognition and are related to tongue height, tongue backing and lip roundness, respectively [46].

2.3. Machine learning algorithms

In this study, a hybrid feature space generated from human voice data is used for speaker recognition and classification of age and gender of speakers using machine learning algorithms. For this purpose, proven classifier algorithms used in human voice data classification studies in the literature were used and their performances were compared.

2.3.1. K-nearest neighbors

The kNN algorithm tries to solve the classification problem by calculating the distances between the points in the sample space with methods such as Euclidean distance or Hamming distance and establishing a "neighborhood" relation between the sample points. To determine the class of a new data point, it finds the k closest points in the sample space and uses these neighbors to predict the label of the new point. The main criterion that determines the classification performance of the kNN algorithm is a user-defined hyperparameter called k. The value of k can be optimized according to the type of data and the distribution of labels in the data space. As the k value increases, the effect of noise or outliers in classification will decrease. As the k value becomes smaller, the size of the neighborhoods in the data space will decrease and the distribution of the labels in space will be scattered [47], [48].

In the kNN algorithm used in this study, k=5 was chosen. The weight function is determined as uniform. In a uniform weight distribution, all points in each neighborhood are weighted equally. Additionally, Euclidean distance was used as the distance metric between points.

2.3.2. Gradient boosting classifier

The gradient boosting classifier (GBC) is an ensemble learning algorithm, which combines the predictions of multiple weak learners to produce a more accurate prediction. The GBC works by iteratively training a weak learner on the residuals which are the errors made by the previous learner. The goal of the gradient boosting classifier is to minimize the loss function. The GBC minimizes the loss function by iteratively moving in the direction of the negative gradient of the loss function. The negative gradient of the loss function points in the direction of the steepest descent of the loss function. The learning rate controls the size of the steps that the GBC takes in the direction of the negative gradient of the loss function. A higher learning rate will result in the GBC taking larger steps and vice versa [49,50].

In the GBC algorithm used in this study, the number of estimators was selected as 100. The learning rate was set

to 0.1.

2.3.3. Support vector machine

The support vector machine (SVM) algorithm, which has been studied in many different fields in the literature, tries to obtain an optimal hyperplane to separate different classes in the sample space while classifying. While determining this hyperplane, it tries to maximize the distance between the data points closest to the plane, which also gives the algorithm its name as support vectors [51-53]. In the SVM used in this study, the regularization parameter was selected as 100. The kernel function was chosen as radial basis function.

2.3.4. Classification and regression tree

The classification and regression tree (CART) algorithm recursively splits the data into subsets based on a criterion called the Gini impurity and continues to split until only data belonging to one class remains in the subsets, called leaf nodes. The Gini impurity value represents the proportion of data in a node that belongs to the same class. The closer the impurity value is to 0, the more likely it is that the data in that node belongs to the same class, and the closer it is to 1, the more likely it is that the node is "impure". Thus, nodes with high impurity continue to be split [54,55].

3. Results

In this study, a data set consisting of audio data recorded while each of 24 volunteers was pronouncing 44 English words was used. First, MFCC, fundamental frequency and formant features were extracted from the data and a feature pool was created by combining these features. The resulting 1056 samples were divided into training and test data in a ratio of 8:2. Using the 5-fold cross-validation method with the training data obtained, kNN, GBC, SVM and CART algorithms were trained to classify the genders and age groups of the volunteers. The classification performances of the algorithms were tested with the remaining test data that had not been seen before by the classification algorithms. All ML algorithms mentioned above and used in this study was trained and tested with Python (v.3.9.7) [56].

3.1. Gender classification

Firstly, gender classification was performed from the voice data recorded from the volunteers. With the training data created, 4 classifier algorithms were trained with the 5-fold cross-validation method. The distribution of accuracies obtained with the training data as a result of 5-fold cross-validation is given in Fig.3. As can be seen from the figure, the highest accuracy was obtained with the SVM algorithm, while the CART algorithm gave the lowest accuracy. The mean value and standard deviations of the accuracies of the 5-fold cross-validation step are also given in Table 1.

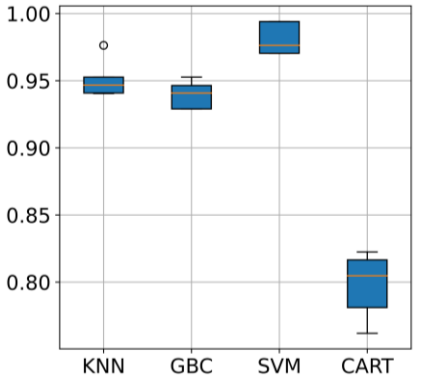

Fig.3. 5-fold cross-validation results of the gender classification.

Table 1. Mean and standard deviation values of the accuracies of the 5-fold cross-validation training process of the gender classification

Accuracy	kNN	GBC	SVM	CART
Mean Value	0.951	0.940	**0.981**	0.797
Standard Dev.	0.013	0.009	0.011	0.023

The confusion matrices of classifications belong to the gender prediction are given in Fig.4. One can see from the figure that, highest test accuracy was achieved with SVM algorithm with a true positive value of 102 and true negative value of 107.

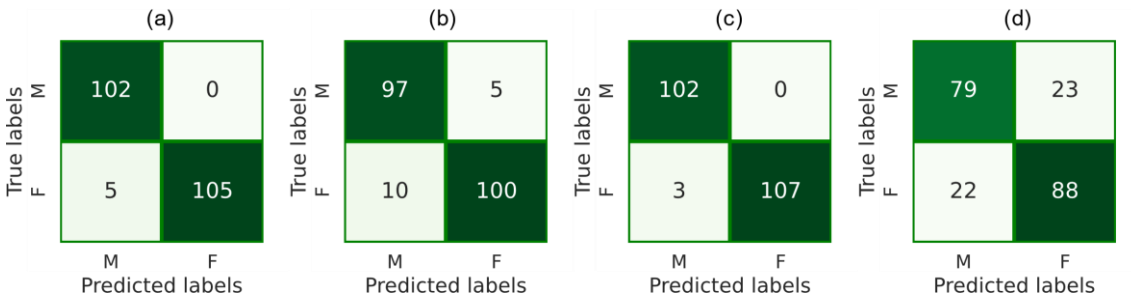

Fig.4. Confusion matrices of the gender classification test results (a) kNN (b) GBC (c) SVM (d) CART.

Table 2 gives the test results of the gender classification in terms of precision, recall, F1-score and accuracy. It can be seen from the table that the highest test accuracy of 0.99 was obtained with the SVM algorithm, which coincides with the complexity matrices.

Table 2. Evaluation metrics of the test process of the gender classification.

Classifier	*Precision*	*Recall*	*F1-Score*	*Accuracy*
kNN	0.98	0.98	0.98	0.98

GBC	0.93	0.93	0.93	0.93
SVM	**0.99**	**0.99**	**0.99**	**0.99**
CART	0.79	0.79	0.79	0.79

3.2. Age group classification

In the second part of the study, age group classification was performed from the voice data recorded from the volunteers. Data labels were divided into two groups according to the age of the participants: youth (18 - 25) and adults (25-50). As in the first part, 4 classifier algorithms were trained using the 5-fold cross-validation method with the training data. The distribution of training accuracies obtained as a result of cross-validation is given as a box plot in Fig.5.

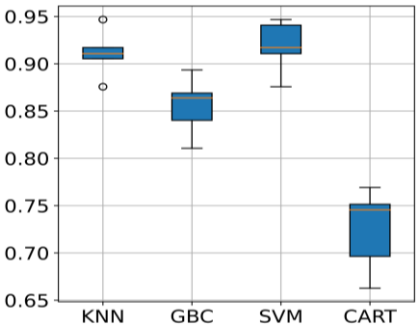

Fig.5. 5-fold cross-validation results of the age group classification.

As can be seen from the Fig.5, the algorithms that gave training accuracy above 90% were SVM and kNN, respectively. The mean accuracy values and standard deviations obtained as a result of 5-fold cross-validation are given in the Table 3. The table shows that 92% training accuracy was achieved with the SVM algorithm. The lowest accuracy was obtained as 73% with the CART algorithm.

Table 3. Mean and standard deviation values of the accuracies of the 5-fold cross-validation training process of the age group classification.

Accuracy	kNN	GBC	SVM	CART
Mean Value	0.911	0.855	**0.918**	0.725
Standard Dev.	0.023	0.028	0.025	0.039

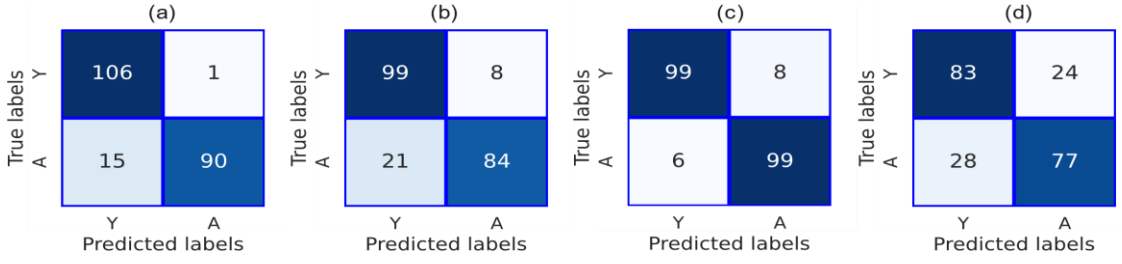

Fig.6. Confusion matrices of the age group classification test results (a) kNN (b) GBC (c) SVM (d) CART.

The confusion matrices of test results belong to the age group classification are given in Figure 6. One can see from the figure that, highest test accuracy was achieved with SVM algorithm with a true positive value of 99 and true negative value of 99. Among the labels in the confusion matrices, A indicates the adult group and Y indicates the youth group.

Table 4. Evaluation metrics of the test process of the gender classification.

Classifier	*Precision*	*Recall*	*F1-Score*	*Accuracy*
kNN	0.88	0.99	0.93	0.92
GBC	0.82	0.93	0.87	0.86
SVM	**0.94**	**0.93**	**0.93**	**0.93**
CART	0.75	0.78	0.76	0.75

Lastly, Table 4 gives the test results of the age group classification in terms of precision, recall, F1-score and accuracy. It can be seen from the table that the highest test accuracy of 0.93 was obtained with the SVM algorithm, which coincides with the complexity matrices.

3.3. Speaker recognition

The final objective of this study is speaker recognition from voice data. We classified the voices of 24 volunteers using four distinct classification algorithms (kNN, GBC, SVM, and CART). Similar to two-class classification scenarios, the four classifiers underwent sequential training with dedicated training data. We assessed the training performance of the algorithms using a 5-fold cross-validation approach.

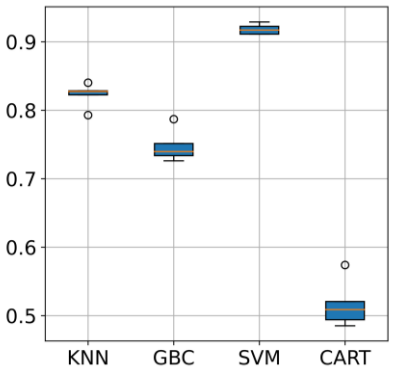

Fig.7. 5-fold cross-validation results of the speaker recognition.

Fig.7 displays a box plot illustrating the distribution of accuracies for each classifier algorithm during the cross-validation stages. The SVM algorithm exhibited the highest accuracy (>90%) in 5-fold cross-validation, while the kNN algorithm secured the second-highest accuracy, averaging over 80% in training performance. In line with two-class classification scenarios, the CART algorithm displayed the lowest accuracy.

Table 5. Mean and standard deviation values of the accuracies of the 5-fold cross-validation training process of the speaker recognition.

Accuracy	kNN	GBC	SVM	CART

Mean Value	0.822	0.748	**0.918**	0.517
Standard Dev.	0.015	0.021	0.006	0.031

Table 6 assesses the test results for speaker recognition based on precision, sensitivity, F1-Score, and Accuracy metrics. The SVM algorithm achieved the highest accuracy at 93%, while the CART algorithm yielded the lowest accuracy at 50%.

Table 6. Evaluation metrics of the test process of the speaker recognition.

Classifier	Precision	Recall	F1-Score	Accuracy
kNN	0.88	0.85	0.86	0.85
GBC	0.75	0.74	0.73	0.74
SVM	**0.94**	**0.93**	**0.93**	**0.93**
CART	0.52	0.50	0.50	0.50

Concluding the analysis, the confusion matrix of the classification performed using the SVM algorithm, known for providing the highest accuracy in speaker recognition, is presented In Fig.8. A detailed examination of the confusion matrix reveals the capability of accurately classifying the voice data of all 24 speakers.

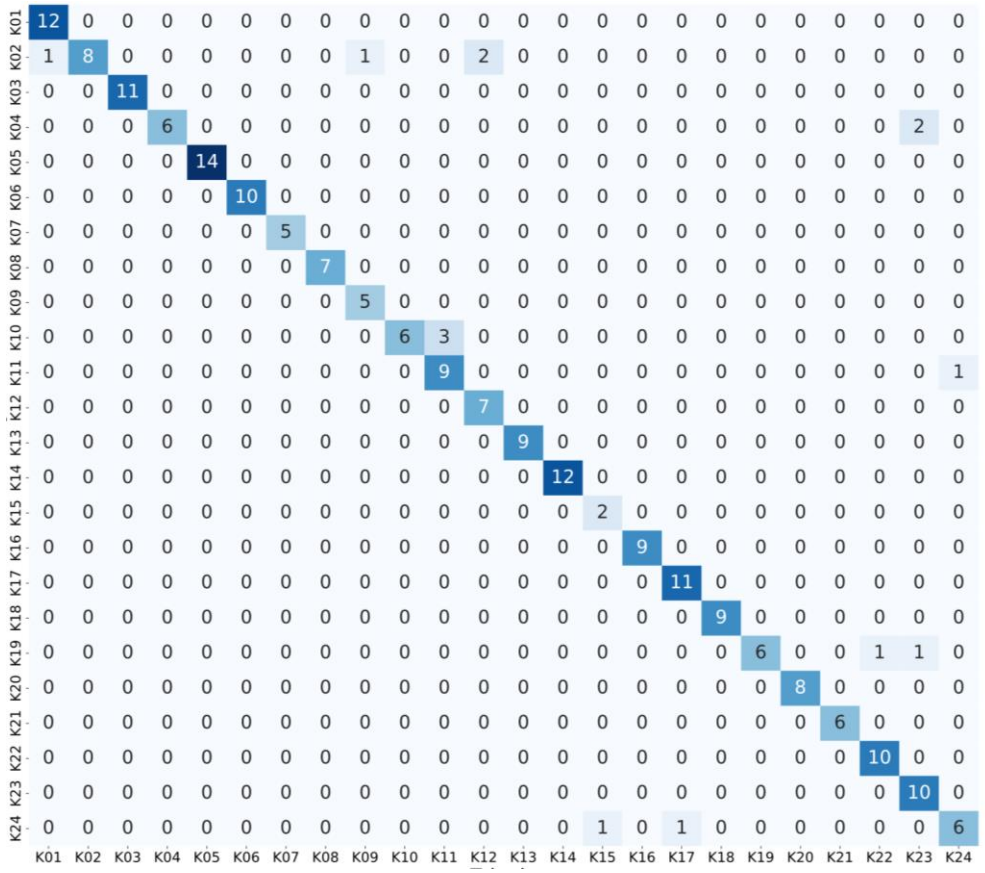

Fig.8. Confusion matrix of the speaker recognition test results for SVM classifier.

3.4. Comparison with the literature

As highlighted in the introduction, the identification of speakers from voice data or the determination of demographic characteristics, such as age and gender, holds significance in various applications like interactive voice response systems, forensic cases, e-marketing, and online banking. However, as indicated in the table, research in this domain remains limited, underscoring the importance of this study in addressing gaps within the existing literature related to Automatic Speaker Recognition (ASR) systems. A notable contribution of this research lies in demonstrating that speaker recognition can achieve high accuracy even with a restricted dataset, utilizing spectral features. Furthermore, the study undertakes a comparison by employing four widely used classification algorithms with distinct methodologies and principles: CART relies on decision trees, SVM seeks optimal hyperplanes, GBC forms an ensemble of weak learners, and kNN classifies based on the majority of nearest neighbors. Notably, the SVM classifier demonstrated the highest accuracy, precision, recall, and F1 score values among these algorithms. A comparative analysis with other studies in the literature is presented in the Table 7 for reference.

Table 7. Comparison of the speaker recognition results with existing literature works.

Reference	Method	Dataset	Features	Accuracy Score
Krishnamoorthy et al.,	GMM-UBM	100 speakers from TIMIT database	MFCC	78.20% with limited data

2011 [29]				80% with noise added data
Bhardwaj et al., 2013 [30]	HMM and GFM	VoxForge speech corpus A subset of NIST 2003 database	MFCC	90% - 93%
Soleymanpour & Marvi, 2017 [31]	NN	ELSDSR database that consists of 22 speakers	MFCC	91.9 % and 93.2 %
Sedigh, 2018 [32]	SVM	Manitoban voice dataset with 24 volunteeers	LPCC, MFCC, HFD, VFD ZCR, turns count	91.60%
Bharath & Rajesh Kumar, 2020 [33]	ELM	124 speakers from TIMIT, SITW-2016	MFCC, PNCC	97.66% highest
Ayvaz et al., 2022 [34]	MLP	Turkish voice dataset is collected from 15 people	MFCC	90.2%
Ramírez-Hernández el al., 2023 [35]	NN, kNN, DT, RF	ELSDSR database	MFCC, MSE	78.92% speaker recognition 93.99% gender classification
Shah et al., 2023 [36]	SVM	VoxCeleb1	MFCC	91% with only sound data
This study	kNN, GBC, SVM, CART	Manitoban voice dataset with 24 volunteeers	MFCC, Fundamental Freq., Formants	92% age group classification 93% speaker recognition 99% gender classification

4. Conclusion

The use of machine learning algorithms is becoming widespread in many application areas such as speaker recognition, speech-to-text, music analysis, and environmental sound classification by using audio data, and it can be predicted that these studies will expand further. In particular, detecting characteristics such as gender, age group, accent and emotion from the human voice has become a practical and reliable method in biomedical applications, forensic cases or interactive voice response systems. This study focuses on gender and age group prediction from voice data. In the dataset used in this study, voice data were recorded from each of 24 volunteers while pronouncing 44 English words. The MFCC feature, which contains frequency and time domain information, was extracted from the audio data, and the fundamental frequency and formant features were extracted from the frequency domain. A hybrid feature space was created by combining the extracted features. The data were labeled according to the gender and age groups of the volunteers, and classification was carried out with machine learning algorithms. In the study, k nearest neighbors, gradient boost classifier, support vector machines and classification and regression tree algorithms were used and their performances were comparatively analyzed. In the gender prediction study, the SVM algorithm gave the highest accuracy, as 99%. In addition, the highest accuracy in age group classification was achieved as 93% with the SVM algorithm. Finally, in the speaker recognition task from the voice data of 24 volunteers, the SVM algorithm managed to achieve high accuracy with 93%. In order to expand the scope of the study, it is planned to enlarge the dataset, collect data from volunteers with different ethnic identities or accents, and add noise to the data in future studies.

Acknowledgements

There is no conflict of interest with any person/institution in the prepared article. This study did not receive any specific funding or financial assistance from governmental, commercial, or non-profit organizations.

References

[1] A. Rana, A. Dumka, R. Singh, M. Rashid, N. Ahmad, and M. K. Panda, "An Efficient Machine Learning Approach for Diagnosing Parkinson's Disease by Utilizing Voice Features," *Electronics (Basel)*, vol. 11, no. 22, p. 3782, 2022.
[2] E. H. Houssein, A. Hammad, and A. A. Ali, "Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review," *Neural Comput Appl*, vol. 34, no. 15, pp. 12527–12557, 2022.
[3] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.
[4] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst Appl*, vol. 197, p. 116659, 2022.
[5] N. N. Arslan, D. Ozdemir, and H. Temurtas, "ECG heartbeats classification with dilated convolutional autoencoder," *Signal Image Video Process*, vol. 18, no. 1, pp. 417–426, 2024, doi: 10.1007/s11760-023-02737-2.
[6] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.
[7] S. Duan, J. Zhang, P. Roe, and M. Towsey, "A survey of tagging techniques for music, speech and environmental sound," *Artif Intell Rev*, vol. 42, no. 4, pp. 637–661, 2014, doi: 10.1007/s10462-012-9362-y.
[8] S. Jayalakshmy and G. F. Sudha, "GTCC-based BiLSTM deep-learning framework for respiratory sound classification using empirical mode decomposition," *Neural Comput Appl*, vol. 33, no. 24, pp. 17029–17040, 2021, doi: 10.1007/s00521-021-06295-x.
[9] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: A systematic review," *Biocybern Biomed Eng*, vol. 33, no. 3, pp. 129–135, 2013, doi: https://doi.org/10.1016/j.bbe.2013.07.001.
[10] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," in *2016 Computing in Cardiology Conference (CinC)*, 2016, pp. 565–568.
[11] M. Xiang *et al.*, "Research of heart sound classification using two-dimensional features," *Biomed Signal Process Control*, vol. 79, p. 104190, 2023, doi: https://doi.org/10.1016/j.bspc.2022.104190.
[12] S. Esmer, M. K. Uçar, İ. Çil, and M. R. Bozkurt, "Parkinson hastalığı teşhisi için makine öğrenmesi tabanlı yeni bir yöntem," *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, vol. 8, no. 3, pp. 1877–1893, 2020.
[13] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Sound Classification and Processing of Urban Environments: A Systematic Literature Review," *Sensors*, vol. 22, no. 22, p. 8608, 2022.
[14] Y. R. Pandeya, D. Kim, and J. Lee, "Domestic cat sound classification using learned features from deep neural nets," *Applied Sciences*, vol. 8, no. 10, p. 1949, 2018.
[15] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, IEEE, 2015, pp. 1–6.
[16] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP J Adv Signal Process*, vol. 2007, pp. 1–8, 2007.
[17] C.-J. Huang, Y.-J. Yang, D.-X. Yang, and Y.-J. Chen, "Frog classification using machine learning techniques," *Expert Syst Appl*, vol. 36, no. 2, Part 2, pp. 3737–3743, 2009, doi: https://doi.org/10.1016/j.eswa.2008.02.059.
[18] D. W. Armitage and H. K. Ober, "A comparison of supervised learning techniques in the classification of bat echolocation calls," *Ecol Inform*, vol. 5, no. 6, pp. 465–473, 2010, doi: https://doi.org/10.1016/j.ecoinf.2010.08.001.
[19] J. Xie, M. Towsey, A. Truskinger, P. Eichinski, J. Zhang, and P. Roe, "Acoustic classification of Australian anurans using syllable features," in *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2015, pp. 1–6. doi: 10.1109/ISSNIP.2015.7106924.
[20] M. Malfante, J. I. Mars, M. Dalla Mura, and C. Gervaise, "Automatic fish sounds classification," *J Acoust Soc Am*, vol. 143, no. 5, pp. 2834–2846, May 2018, doi: 10.1121/1.5036628.
[21] A. P. Ribeiro, N. F. F. da Silva, F. N. Mesquita, P. de C. S. Araújo, T. C. Rosa, and J. N. Mesquita-Neto, "Machine learning approach for automatic recognition of tomato-pollinating bees based on their buzzing-sounds," *PLoS Comput Biol*, vol. 17, no. 9, pp. e1009426-, Sep. 2021, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009426
[22] U. Haider *et al.*, "Bioacoustics Signal Classification Using Hybrid Feature Space with Machine Learning," in *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*, 2023, pp. 376–380. doi: 10.1109/ICCAE56788.2023.10111384.
[23] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.
[24] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.
[25] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Applied Acoustics*, vol. 172, p. 107581, 2021.
[26] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021, doi: https://doi.org/10.1016/j.compeleceng.2021.107005.
[27] S. Furui, "40 Years of Progress in Automatic Speaker Recognition," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1050–1059.
[28] N. Singh, A. Agrawal, and R. Khan, "The development of speaker recognition technology," *IJARET., no. May*, 2018.
[29] P. Krishnamoorthy, H. S. Jayanna, and S. R. M. Prasanna, "Speaker recognition under limited data condition by noise addition," *Expert Syst Appl*, vol. 38, no. 10, pp. 13487–13490, 2011, doi: https://doi.org/10.1016/j.eswa.2011.04.069.

[30] S. Bhardwaj, S. Srivastava, M. Hanmandlu, and J. R. P. Gupta, "GFM-Based Methods for Speaker Identification," *IEEE Trans Cybern*, vol. 43, no. 3, pp. 1047–1058, 2013, doi: 10.1109/TSMCB.2012.2223461.
[31] M. Soleymanpour and H. Marvi, "Text-independent speaker identification based on selection of the most similar feature vectors," *Int J Speech Technol*, vol. 20, no. 1, pp. 99–108, 2017, doi: 10.1007/s10772-016-9385-x.
[32] S. Sedigh, "Application of polyscale methods for speaker verification," Master Thesis, The University of Manitoba, Winnipeg, 2018.
[33] K. P. Bharath and M. Rajesh Kumar, "ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score," *Multimed Tools Appl*, vol. 79, no. 39, pp. 28859–28883, 2020, doi: 10.1007/s11042-020-09353-z.
[34] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. Bobomirzaevich, "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning," *CMC-Computers Materials & Continua*, vol. 71, no. 3, 2022.
[35] J. I. Ramírez-Hernández, A. Manzo-Martínez, F. Gaxiola, L. C. González-Gurrola, V. C. Álvarez-Oliva, and R. López-Santillán, "A Comparison Between MFCC and MSE Features for Text-Independent Speaker Recognition Using Machine Learning Algorithms," in *Fuzzy Logic and Neural Networks for Hybrid Intelligent System Design*, O. Castillo and P. Melin, Eds., Cham: Springer International Publishing, 2023, pp. 123–140. doi: 10.1007/978-3-031-22042-5_7.
[36] S. H. Shah, M. S. Saeed, S. Nawaz, and M. H. Yousaf, "Speaker Recognition in Realistic Scenario Using Multimodal Data," in *2023 3rd International Conference on Artificial Intelligence (ICAI)*, 2023, pp. 209–213. doi: 10.1109/ICAI58407.2023.10136626.
[37] S. Sedigh and W. Kinsner, "A Manitoban Speech Dataset," *IEEE DataPort, January*, 2018, doi: 10.21227/H2KM16.
[38] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020, doi: https://doi.org/10.1016/j.apacoust.2019.107020.
[39] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*, IEEE, 2014, pp. 372–378.
[40] K. N. Stevens, "Autocorrelation analysis of speech sounds," *J Acoust Soc Am*, vol. 22, no. 6, pp. 769–771, 1950.
[41] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.
[42] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun*, vol. 54, no. 4, pp. 543–565, 2012.
[43] P. Pedersen, "The mel scale," *Journal of Music Theory*, vol. 9, no. 2, pp. 295–308, 1965.
[44] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.
[45] M. Lahat, R. Niederjohn, and D. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans Acoust*, vol. 35, no. 6, pp. 741–750, 1987.
[46] I. V. Bele, "The speaker's formant," *Journal of Voice*, vol. 20, no. 4, pp. 555–578, 2006.
[47] G. Batista and D. F. Silva, "How k-nearest neighbor parameters affect its performance," in *Argentine symposium on artificial intelligence*, Citeseer, 2009, pp. 1–12.
[48] O. Kramer, "K-Nearest Neighbors," in *Dimensionality Reduction with Unsupervised Nearest Neighbors*, O. Kramer, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23. doi: 10.1007/978-3-642-38652-7_2.
[49] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif Intell Rev*, vol. 54, pp. 1937–1967, 2021.
[50] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, p. 21, 2013.
[51] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
[52] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, Elsevier, 2020, pp. 101–121.
[53] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.
[54] W. Loh, "Fifty years of classification and regression trees," *International Statistical Review*, vol. 82, no. 3, pp. 329–348, 2014.
[55] E. Şahin Sadık, H. M. Saraoğlu, S. Canbaz Kabay, M. Tosun, C. Keskinkılıç, and G. Akdağ, "Investigation of the effect of rosemary odor on mental workload using EEG: an artificial intelligence approach," *Signal Image Video Process*, vol. 16, no. 2, pp. 497–504, 2022.
[56] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.