Research Article

Citation date: 2019/11/06

Identification of Age Voiceprint Using Machine Learning Algorithms

Linus A Xavier

How to cite:

L. A. Xavier, "Identification of Age Voiceprint Using Machine Learning Algorithms," *Res. Rev. Sci. Technol.*, vol. 1, no. 1, pp. 1–16, 2019.

Article history:

Received: 2019/08/21
Available online: 2019/11/06

Abstract

The voice is considered a biometric trait since we can extract information from the speech signal that allows us to identify the person speaking in a specific recording. Fingerprints, iris, DNA, or speech can be used in biometric systems, with speech being the most intuitive, basic, and easy to create characteristic. Speech-based services are widely used in the banking and mobile sectors, although these services do not employ voice recognition to identify consumers. As a result, the possibility of using these services under a fake name is always there. To reduce the possibility of fraudulent identification, voice-based recognition systems must be designed. In this research, Mel Frequency Cepstral Coefficients (MFCC) characteristics were retrieved from the gathered voice samples to train five different machine learning algorithms, namely, the decision tree, random forest (RF), support vector machines (SVM), closest neighbor (k-NN), and multi-layer sensor (MLP). Accuracy, precision, recall, specificity, and F1 score were used as classification performance metrics to compare these algorithms. According to the findings of the study, the MLP approach had a high classification accuracy of 91%. In addition, it seems that RF performs better than other measurements. This finding demonstrates how these categorization algorithms may assist voice-based biometric systems.

Keywords: Age Voiceprint, K-NN, MFCC, MLP, RF, SVM

Declarations

Competing interests:

The author declares no competing interests.

1. Introduction

The ability to offer simple and safe authentication for customer service apps is becoming a need for current security standards as security concerns and incidents continue to expand across organizations in numerous sectors. Biometric technology, including as fingerprint scanners on computers, cameras with built-in facial recognition capabilities at airport terminals and stadiums, and voice-based authentication systems for account access on smartphones, have all seen widespread acceptance in the previous decade [1] [2].

The voice of the person whose identification must be recorded in the system is used to perform voice biometric recognition. For authentication purposes, this input is saved as a print. The spoken statement is separated into several frequencies to creates the input print [3]. At this step, behavioral characteristics are found that work together to create the voice print.

Iris and fingerprint recognition are similar to voice recognition. They are all one-of-a-kind and cannot be replicated. These prints are saved in a database for further verification [4]. Meanwhile, a system that is not reliant on text focuses and matches unexpected speech with previously recorded voice data.

There are two types of voice authentication. The first one is text-independent recognition [5]. In this situation, the system does not preserve any pre-recorded audio to compare with the input. It is a voice identification method that does not need the input of any previous speech data into the biometric system. Because it allows for free communication, it is significantly more practical. The second is the text-dependent recognition [5]. This demands the delivery of a previously given phrase that has been saved in the system, creating a speech content constraint.

There are few distinct advantages of voice biometrics. First, it Increases security and decreased Fraud [6]. The necessity for robust, multi-factor authentication has grown as the number of fraud assaults has expanded across sectors. Unlike PINs and security questions, which may be readily hacked, voice biometrics assures that the person on the other end of the line is who they claim to be [7]. Voice biometrics is an excellent way for verifying callers in contact center since it reduces the possibility of social engineering, which happens often with agents.

The second benefit of voice biometric systems that is often neglected is that they have the ability to greatly enhance customer experiences. Callers no longer need to offer passcodes, PINs, or answers to challenge questions to authenticate their identity using speech biometrics devices [7]. This makes speech biometrics perfect for omnichannel and multichannel implementations, since a customer's voiceprint may be used across all of company's support channels after they have been registered [8]. This seamless experience makes the process simpler and more efficient for client. In fact, depending on the cause for the call, voice biometrics have been shown to shorten the time it takes to authenticate a caller's identification from a few seconds to several minutes [9]. This advantage allows to increase not just call personalization but also customer experiences considerably.

The third benefit is that it reduces costs. Voice biometrics solutions have been shown to save millions of dollars in agent time by lowering the processes and time necessary in the verification process [10]. Furthermore, when adopting voice biometrics for verification, firms may minimize Average Handle Time by 30+ seconds each call on average [11] [12]. This result to not only enhanced security process efficiency, but also significant cost savings due to the reduced time agents must spend verifying consumers.

2. Methodology

We employ MFCC to do age group detection/identification in this study.

2.1 Mel Frequency Cepstral Coefficient (MFCC)

The most common and significant approach for extracting spectral information is Mel Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most widely used feature extraction methods in speech recognition [13]. They are based on the frequency domain and employ the Mel scale, which is based on the human ear scale. MFCCs that are well-designed as frequency domain characteristics are far more exact than time domain features [14].

Because human speech is not linear as a function of frequency, the pitch of a single frequency acoustic voice stream is mapped onto a "Mel" scale. The frequency spacing below 1 kHz is linear in Mel scale, whereas the frequency spacing beyond 1 kHz is logarithmic [15]. The below equation is used to compute the Mel frequencies that correlate to the Hertz frequencies.

$$fmel = 2595 * \log(1 + \frac{f}{700})$$

Figure 1 depicts the block design for Mel-Frequency Cepstral Coefficients (MFCC) calculations.

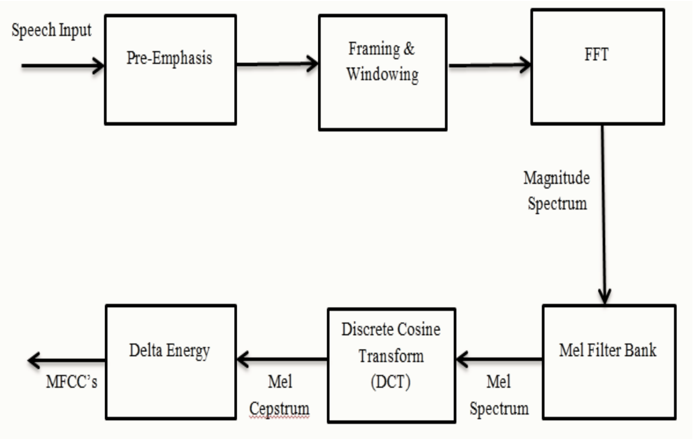


Fig. 1 Block Diagram for MFCC Computation

The inner blocks in Fig. 1 are detailed below:

a) Pre-Emphasis: Audio signals with a sample rate of 16 kHz are captured. Each word is saved as its own audio file. This stage comprises signal pre-emphasis to increase signal energy at high frequencies [15]. Equation below gives the difference equation of the Pre-emphasis filter.

$$H(z)=\frac{B(z)}{A(z)}=\frac{\left(b_0+b_1z^{-1}\right)}{1}=1-0.97z^{-1}$$

b) Framing and Windowing: In nature, the language (voice) signal is not stationary. The usage of stationary frame is employed to make it seem more professional. After pre-emphasis, the next stage is framing, in which the voice signal is divided into smaller frames that overlap each other [16]. Windowing is used to reduce discontinuities at the margins of frames after framing. The window approach employed in this study was the Hamming Window. The Hamming Window is determined by the following equation.

$$w(n)=\begin{cases}0.54\text{-}0.46cos\left[\dfrac{2\pi n}{N\text{-}1}\right] & 0\le n\le N-1\\ 0 & otherwise\end{cases}$$

N is the total number of samples in a single frame.

c) Fast Fourier Transform (FFT): The Discrete Fourier transform (DFT) of signal is calculated using the fast Fourier transform. This step is used to convert the signal from time domain to frequency domain. Equation below is used to compute the FFT [17].

$$x[k]=\sum_{n=0}^{N-1}x(n)e^{-j2\frac{\pi}{N}kn}$$

Where, N is the size of FFT.

d) Mel Filter Bank: The spectrum's power is turned into a Mel scale, which is the next stage. Mel's filter bank is made up of overlapping triangle filters.

e) Discrete Cosine Transform (DCT): After obtaining the logarithm of the Mel-filter bank output, the Discrete Cosine Transform (DCT) is used.

f) Delta Energy: It takes the base 10 logarithm of the output from the previous step in this step. Because the human ear's reaction to acoustic speech signal levels is not linear, the calculation of Log energy is required. At greater amplitudes, the human ear is less sensitive to differences in amplitude. The logarithmic function has the benefit of closely resembling the behavior of the human ear [18]. The equation below is used to compute energy

$$E = \sum_{t=t1}^{t=t2} x^2(t)$$

Finally, the Mel frequency cepstral coefficients are obtained.

2.2 Algorithm selection

The goal of this study is to use Decision Trees, Random Forest, Support Vector Machines, k-Nearest Neighbor, and Multilayer Perceptron algorithms on the data set obtained after implementing the minimum-maximum normalization technique on the raw data set to apply different classification techniques based on machine learning. The dataset came from 5 samples from each of the 9 elderly and 11 young samples. The following are the algorithms' descriptions:

Decision Trees: It is represented as class tags at the level of the tree's leaves, as well as actions on features with branches leading to these leaves and expanding from the start. In terms of comprehension and interpretation, the method is straightforward. It may be used to process both numerical and class data.

Random Forest: The decision tree's low depth prevents categorization from taking place, while its excessive depth makes classification difficult [19]. In RF, increasing the number of trees reduces tree depth. In RF, bagging may be used to find proper categories among trees. In classification issues, tagging is done using vote computations [20]. The use of an adequate number of trees and accurate voting amongst trees may help a random decision tree operate well.

Support Vectors Machines (SVMs): Its a machine learning theory-based nonparametric classification method. For dual classifications, SVMs were created. SVM's working concept is based on estimating the most suitable decision function that can distinguish different classes from each other, or, in other words, establishing the hyper-plane that can effectively separate two classes [21] [22].

The categorization step in the k-Nearest Neighborhood method is made by evaluating relationships between data. The linear decomposition approach is used to operate this system on the coordinate plane. In the original k-NN technique, the item to be categorized is allocated to the class that includes the majority of the object's closest k neighbors [23].

Artificial neural networks, or multilayer perceptrons, are learning algorithms based on the modeling of human brain cells. They are made up of three layers: an input layer, one or more hidden layers, and an output layer. Forward propagation and backward propagation transitions exist in MLPs. The output and error value of the network are determined at the forward propagation step. The link weight values between the layers are changed at the back propagation step to reduce the computed error value.

2.3 Assessments of the selected ML models

A confusion matrix is a table that allows to see how well a classification system performs. Each column represented the number of anticipated classifications done by a classification model, whereas each row included information related with actual classifications.

Let TP denote the number of true positives (positive samples correctly classified as positive), FN the number of false negatives (positive samples incorrectly classified as negative), FP the number of false positives (negative samples incorrectly classified as positive), and TN the number of true negatives (negative samples incorrectly classified as negative) (negative samples correctly classified as negative) [24]. Table 1 shows the confusion matrix for a binary classifier. As can be observed, such a form makes it easy to evaluate prediction mistakes visually since they are plainly placed beyond the table's diagonal.

Table 1. Confusion matrix

		Predicted	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

A confusion matrix is often used to construct a variety of commonly used statistical measures, including [25]:

• Accuracy (Ac), which is described as the percentage of accurate classifications across all samples evaluated and may be calculated using the equation:

$$Ac = \frac{TP + TN}{TP + FP + TN + FN}$$

• Precision (Pr), which may be calculated using Equation below and is defined as the ratio of genuine positives to all expected positives:

$$Pr = \frac{TP}{TP + FP}$$

• Sensitivity (Se), also known as True Positive Rate (TPR) and Recall. It is defined as the fraction of true positives accurately recognized as such, which may be calculated using below. The proportion of patients who are appropriately recognized as having the illness is one example.

$$Se = \frac{TP}{TP + FN}$$

• True Negative Rate (TNR), which is also known as Specificity (Sp). It is defined as the fraction of real negatives accurately classified as negatives. For instance, the proportion of healthy people who are appropriately labeled as such.

$$Sp = \frac{TN}{TN + FP}$$

• The F-score, often known as the F measure, is the harmonic mean of accuracy and recall. As can be seen from the equation below, precision and recall are equally weighted, and the highest value of 1 is reached when precision and recall are equivalent.

$$F = 2 \times \frac{Precision \ \times Recall}{Precision + Recall}$$

Furthermore, among the major performance measurement metrics used in the assessment of model performance in Machine Learning methods are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Square Error (MSE) [26]. Equations below are used to calculate MSE, RMSE, and MAE, respectively.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid e_i \mid$$

The letter 'e' stands for error in these formulas. Because MSE, RMSE, and MAE are all error measures, lower values imply better performance [27]. If the RMSE is equal to zero, for example, it can be assumed that the performance result is better. The model prediction success of the five methods mentioned in this research was measured using the RMSE, MAE, and MSE criteria. Close to 0 RMSE, MAE, and MSE values indicate that no major errors were made.

3. Results

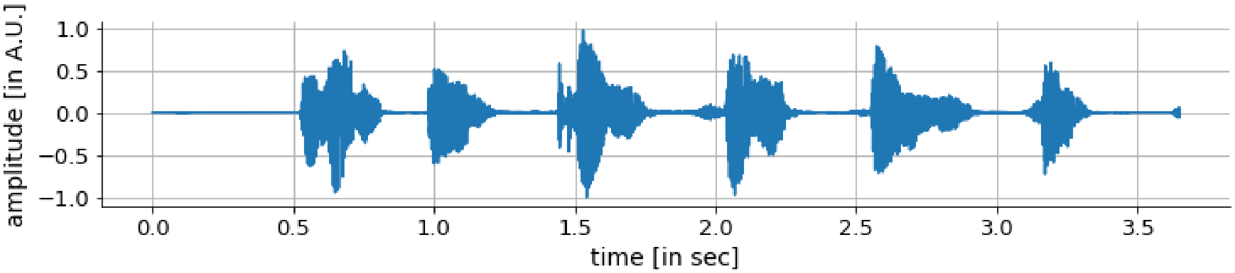


Figure 1. An original audio signal of a sample.

Once we have the filter bank, we can use it to transform using the magnitude FFT X. The number of mel filters (mel filter num) in the filterbank will be N in this case. In addition, we log the output values. This is due to the fact that we can hear both subtle and loud noises. The log transformation will assist us enhance low energy values while attenuating large amplitude values a little.

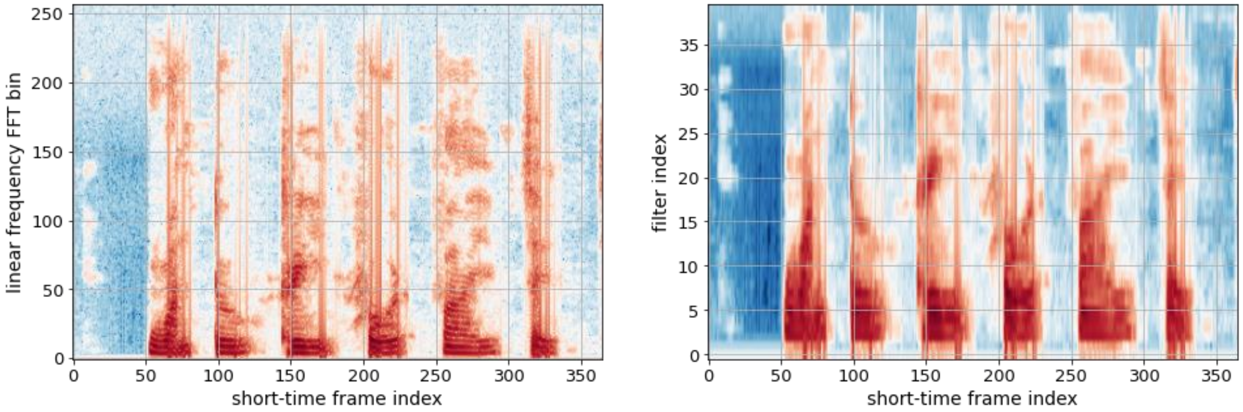
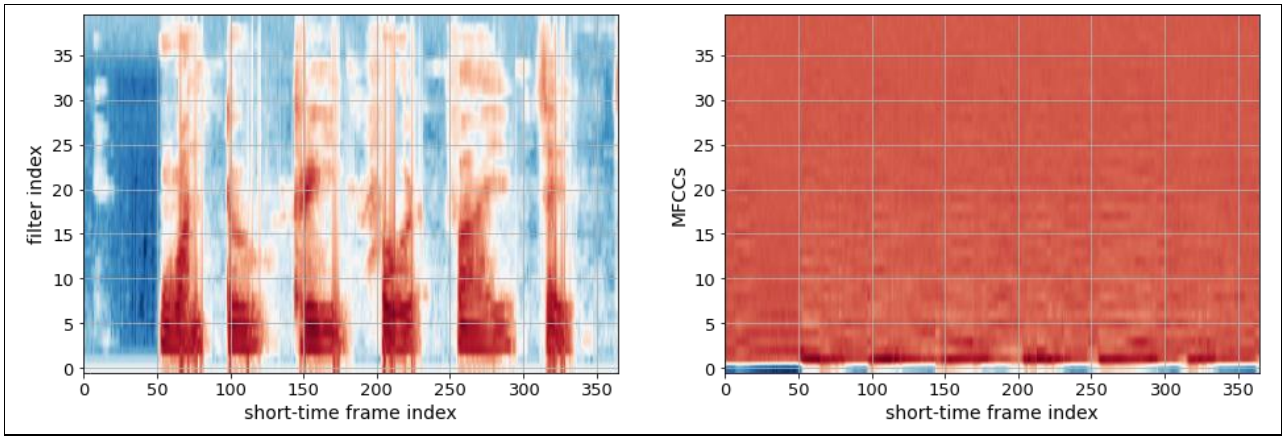


Figure 2. Log transformation

The MFCCs result is achieved by applying the DCT transform on the "X filtered log". For each short-time frame, the MFCC[0], which is the initial element of the vector obtained after DCT captures the spectral energy throughout the filterbank. The graph below demonstrates this. Furthermore, feature vectors for classification techniques benefit greatly from this compact structure.



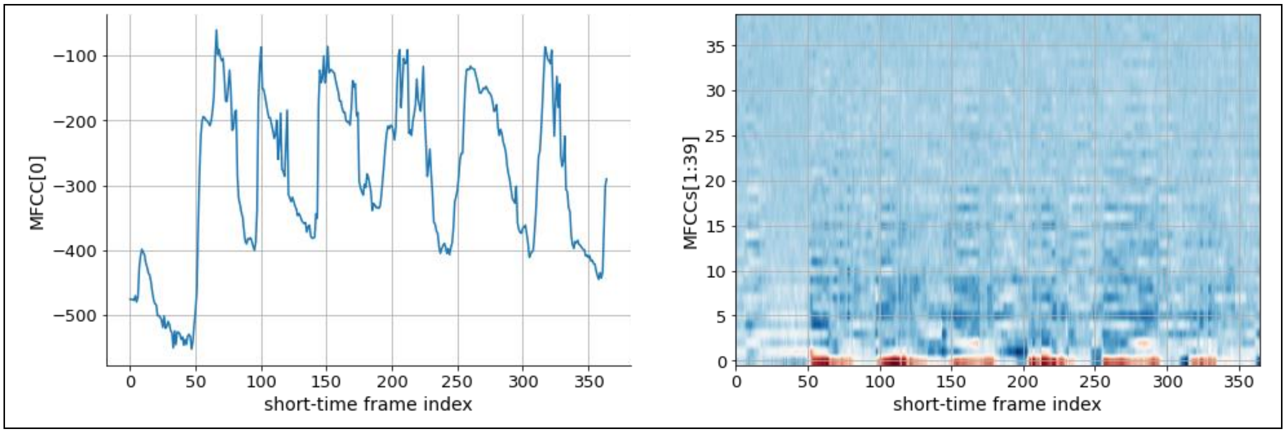


Figure 3 (a-d). DCT transformation

The techniques employed in this research were Decision Trees, Random Forest, Support Vector Machines, k-Nearest Neighbor, and Multilayer Perceptron. Classification algorithm performance was measured using parameters such as accuracy, sensitivity, specificity, precision, and F1-score. Table 2 summarizes the highest performing classifiers in age group identification after all of the findings have been examined. The MLP classifier was effective in classifying age groups with the greatest accuracy value of 91%, according to Table 2. These findings suggest that the MLP Algorithm outperforms alternative classification algorithms in this situation. The MLP method, as shown in Figure 1, provides the greatest accuracy, followed by SVM, DT, RF, and k-NN algorithms, in that order.

Table 2. Cross-validation test results for ML algorithms

Criteria	DT (%)	RF (%)	SVM (%)	k-NN (%)	MLP (%)
Accuracy	81.081	80.01	83.782	80.03	91.04
Recall	75.01	64.72	75.01	75.860	82.761
Specificity	88.243	100	94.122	83.87	96.77
Precision	88.2443	100	93.753	81.48	96.00
F1-Score	81.081	78.570	83.334	78.56	88.88

Figure 4 shows a bar graph of the cross-validation test results obtained as mathematical values in Table 2. The RF technique has the best specificity and accuracy rates, as demonstrated in Fig. 4.

Figure 4 shows the results of cross-validation tests for machine learning algorithms.

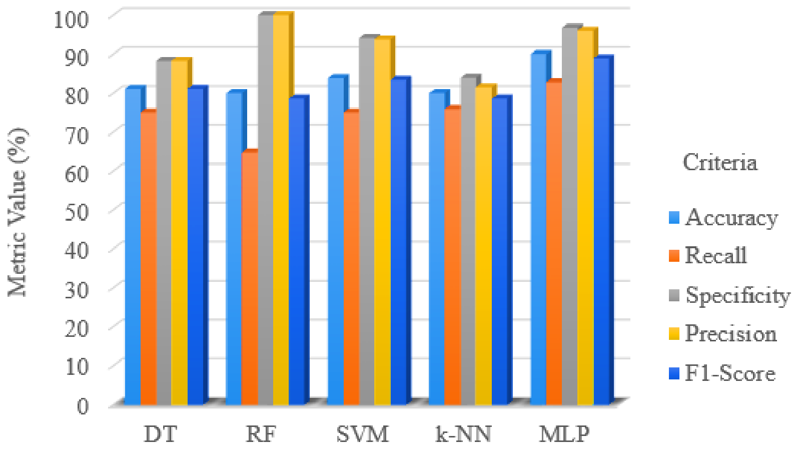


Table 3 shows the performance results of MAE, RMSE, and MSE values for the five distinct ML algorithms utilized in the research. In contrast to prior research, MSE, RMSE, and MAE error analyses were included to the study's performance assessment in order to create a more complete comparison to other algorithms. Given that lower MAE, RMSE, and MSE values suggest higher performance, it may be concluded that the performance results of MLP show considerably lower MAE (0.1), RMSE (0.3162), and MSE (0.1), implying that these performance results are superior. Other performance metrics are supported by these findings. The greatest number correlates to the highest mistake rate, while the lowest value reflects the best error rate. The RMSE computation is believed to be the most clear assessment criteria for assessing the difference between the models.

Table 3 compares the performance of ML algorithms using the MAE, RMSE, and MSE metrics.

Metrics	DT	RF	SVM	k-NN	MLP
MAE	0.1892	0.2	0.1622	0.2	0.1
RMSE	0.435	0.4472	0.4027	0.4472	0.3162
MSE	0.1892	0.2	0.1622	0.2	0.1

4. Conclusion

Gender voice is considered one of the most important aspects to discern from a given voice, a task that is fraught with difficulties. A series of strategies has been used to discover significant features to be used for developing a model from a training set in order to distinguish age groups from voice signals. This model is useful for detecting the age group of a vocal signal.

Due to the exponential rise in the smartphone user base and the incomparable convenience it provides, voice-based authentication is becoming more important among biometric authentication methods. Indeed, human speech may be caught effortlessly across long distances using just a conventional phone connection and no special reading gear. In addition, as compared to other biometric techniques, voice authentication gives the user more control over signal collection.

Speech, among other biometric signals, is notable for the quantity of data it contains about a speaker. Speech contains not just linguistic information, but also paralinguistic and extralinguistic data such as a speaker's age, gender, ethnicity, personality qualities, emotional state, and even information about his or her physical and mental health. All of this information is considered private and should be safeguarded at all times. In reality, we already live in a world where speech data and the information collected from it may be legally classified as personally identifiable information.

This study included five machine learning algorithms: decision trees, random forests (RF), support vector machines (SVM), nearest neighbor (k-NN), and multi-layer sensors (MLP). To evaluate these algorithms, accuracy, precision, recall, specificity, and F1 score were employed as classification performance criteria. The MLP technique showed a high classification accuracy of 91%, according to the study's results. Furthermore, it seems that RF outperforms other metrics. This research indicates how voice-based biometric systems might benefit from these categorization algorithms.

While speech biometrics provides a safe method of user authentication, it is not without risk. High-quality voice spoofing, or "deepfakes," is now possible because to breakthroughs in machine learning, recording technology, and synthetic speech, which may fool people and vocal biometrics systems into believing they are hearing a real person. These attacks may be used to obtain access to accounts that aren't supposed to be accessed. To combat speech spoofing, future research is nedded to advance the equipment that can differentiate between a live voice and a recorded or synthetic voice.

References

[1] R. Feldman, “Considerations on the emerging implementation of biometric technology,” *Hast. Comm. Ent. LJ*, vol. 25, p. 653, 2002.

[2] J. A. Unar, W. C. Seng, and A. Abbasi, “A review of biometric technology along with trends and prospects,” *Pattern Recognit.*, vol. 47, no. 8, pp. 2673–2688, 2014.

[3] A. Das, C. Galdi, H. Han, R. Ramachandra, J.-L. Dugelay, and A. Dantcheva, “Recent advances in biometric technology for mobile devices,” in *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, 2018, pp. 1–11.

[4] S. Liu and M. Silverman, “A practical guide to biometric security technology,” *IT Prof.*, vol. 3, no. 1, pp. 27–32, 2001.

[5] R. C. Johnson, T. E. Boult, and W. J. Scheirer, “Voice authentication using short phrases: Examining accuracy, security and privacy issues,” in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[6] R. Saini and N. Rana, “Comparison of various biometric methods,” *Int. J. Adv. Sci. Technol.*, vol. 2, no. 1, pp. 24–30, 2014.

[7] A. Saleema and S. M. Thampi, “Voice biometrics: the promising future of authentication in the internet of things,” in *Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science*, IGI Global, 2018, pp. 360–389.

[8] H. Aronowitz, R. Hoory, J. Pelecanos, and D. Nahamoo, “New developments in voice biometrics for user authentication,” 2011.

[9] L. J. Chetalam, “Enhancing Security of MPesa Transactions by Use of Voice Biometrics.” United States International University-Africa, 2018.

[10] C. Alver, “Voice Biometrics in Financial Services,” *J. Financ. Serv. Technol.*, vol. 1, no. 1, pp. 75–81, 2007.

[11] H. Lamm, “How best to use voice biometrics in the contact centre,” *Biometric*

Technol. Today, vol. 2016, no. 10, pp. 5–7, 2016.

[12] J. A. Markowitz, "Voice biometrics," *Commun. ACM*, vol. 43, no. 9, pp. 66–73, 2000.

[13] A. Jain and O. P. Sharma, "A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review 1," 2013.

[14] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2017, pp. 2257–2260.

[15] A. C. Kelly and C. Gobl, "A comparison of mel-frequency cepstral coefficient (MFCC) calculation techniques," *J. Comput.*, vol. 3, no. 10, pp. 62–66, 2011.

[16] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," in *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, 2012, pp. 248–251.

[17] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv Prepr. arXiv1003.4083*, 2010.

[18] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *Cognit. Comput.*, vol. 5, no. 4, pp. 533–544, 2013.

[19] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[20] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, 2018.

[21] S. Ding, Z. Zhu, and X. Zhang, "An overview on semi-supervised support vector machine," *Neural Comput. Appl.*, vol. 28, no. 5, pp. 969–978, 2017.

[22] J. Ahmad, M. Fiaz, S. Kwon, M. Sodanil, B. Vo, and S. W. Baik, "Gender identification using mfcc for telephone applications-a comparative study," *arXiv Prepr. arXiv1601.01577*, 2016.

[23] H. D. Nguyen, K. P. Tran, X. Zeng, L. Koehl, and G. Tartare, "Wearable sensor data based human activity recognition using machine learning: a new approach," *arXiv Prepr. arXiv1905.03809*, 2019.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[25] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, 2006, pp. 1015–1021.

[26] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE)

over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[27] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.