Article citation info:
Bourouhou A, Jilbab A, Cherti M, Bourouhou Z, Nacir C. Discrimination between patients with CVDs and healthy people by voiceprint using the MFCC and pitch. Diagnostyka. 2021;22(4):9-16. https://doi.org/10.29354/diag/142560

DIAGNOSTYKA, 2021, Vol. 22, No. 4

ISSN 1641-6414
e-ISSN 2449-5220
DOI: 10.29354/diag/142560

DISCRIMINATION BETWEEN PATIENTS WITH CVDS AND HEALTHY PEOPLE BY VOICEPRINT USING THE MFCC AND PITCH

Abdelhamid BOUROUHOU [1], Abdelilah JILBAB [1], Mohammed CHERTI [2], Zaineb BOUROUHOU [2], Chafik NACIR [1]

[1] University Mohammed V, Ecole Normale Superieure de l'Enseignement Technique, Rabat, Morocco
E-mail : abdelhamid.bourouhou@um5s.net.ma
[2] University Mohammed V, Faculté de médecine et de pharmacie & CHU, Rabat, Morocco
E-mail : chertiy@gmail.com

Abstract

Heart diseases cause many deaths around the world every year, and his death rate makes the leader of the killer diseases. But early diagnosis can be helpful to decrease those several deaths and save lives. To ensure good diagnose, people must pass a series of clinical examinations and analyses, which make the diagnostic operation expensive and not accessible for everyone.

Speech analysis comes as a strong tool which can resolve the task and give back a new way to discriminate between healthy people and person with cardiovascular diseases. Our latest paper treated this task but using a dysphonia measurement to differentiate between people with cardiovascular disease and the healthy one, and we were able to reach 81.5% in prediction accuracy.

This time we choose to change the method to increase the accuracy by extracting the voiceprint using 13 Mel-Frequency Cepstral Coefficients and the pitch, extracted from the people's voices provided from a database which contain 75 subjects (35 has cardiovascular diseases, 40 are healthy), three records of sustained vowels (aaaaa…, ooooo… .. and iiiiiii….) has been collected from each one. We used the k-near-neighbor classifier to train a model and to classify the test entities. We were able to outperform the previous results, reaching 95.55% of prediction accuracy.

Keywords: cardiovascular diseases; speech analysis; voiceprint; MFCC; K-Near-Neighbor classifier;

1. INTORDUCTION

Heart diseases or cardiovascular diseases (CVDs) are some disorders which affect the heart and blood vessels, we can list the coronary heart disease, peripheral arterial disease, congenital heart disease, deep vein thrombosis and pulmonary embolism [1].

The CVDs can be murderess, and according to the World Heart Federation, it causes over 17 million deaths annually exceeding other diseases such as cancers, respiratory diseases, and diabetes, this portion of the deaths count 31% of global deaths. CVDs cost the world about 863 billion dollars and it's predicted that the CVDs reach 23 million deaths by 2030 which means more cost. So, we have to take action on CVDs or premature deaths will continue to rise, especially that 80% of those premature deaths can be avoided or delayed by early detection of CVDs.

In the aim to make CVDs detection and diagnosis more accurate, fast and accessible for everyone, several researches had launched to develop an automatic diagnosis by using different signal processing methods, and different source of information such as electrocardiogram "ECG" [3, 4], echography and phonocardiogram "PCG" [5, 6].

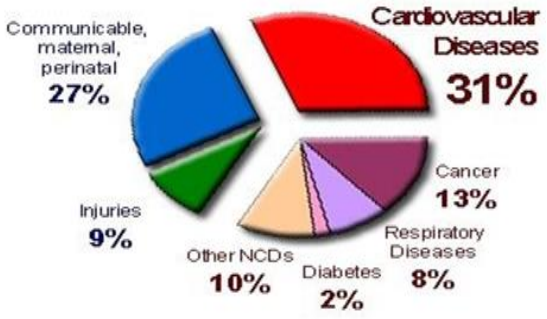

Fig. 1. Global causes of all deaths worldwide [2]

In our previous work [7], we tried to differentiate people with CVDs and healthy people by their voices using a dysphonia measurement. After pre-processing phase which consist on segmentation and filtering, we have extracted 26 sound features from all records and we construct models by training each chosen classifier. The validation phase has given impressive results. We have concluded that the CVDs influence the voice of the patients, and accuracy to discriminate people it was about 81.5%.

The present paper comes to increase the accuracy using people voices, more precisely we have proposed a new way to classify healthy people and people with CVDs, using their voices, this time we will not use a dysphonia measurement but we will try to extract 13 Mel-Frequency Cepstral Coefficients "MFCC", and the pitch of each voice record to construct a voiceprint from healthy people and from people with CVDs. We will train our classifier on a train dataset to get a classification model which will be the referee for our test dataset.

2. USED DATABASE

The used database in this work was collected and used previously in BOUROUHOU et al. [7], it's about 35 CVDs patients (17 women and 18 men), and 40 healthy people (19 women and 21 men). The ranged age of patients is between 30 and 81 (average 56, standard deviation 10.79), and the age healthy people ranges between 40 and 75 (average 55, standard deviation 6.64). The following table present the database:

Table 1. Database description

<table>
<tr><th colspan="6">Database</th></tr>
<tr><td rowspan="6">Total of people</td><td rowspan="6">75</td><td rowspan="2">People with CVDs</td><td rowspan="2">35</td><td>Age average</td><td>56</td></tr>
<tr><td>age standard deviation</td><td>11</td></tr>
<tr><td rowspan="2">Healthy people</td><td rowspan="2">40</td><td>Age average</td><td>55</td></tr>
<tr><td>age standard deviation</td><td>6.6</td></tr>
<tr><td>Total age average</td><td colspan="2">55.5</td></tr>
<tr><td>Total age standard deviation</td><td colspan="2">8.98</td></tr>
<tr><td rowspan="4">Gender of people</td><td rowspan="2">Women</td><td rowspan="2">36</td><td>Women with CVDs</td><td>17</td></tr>
<tr><td>Healthy women</td><td>19</td></tr>
<tr><td rowspan="2">Men</td><td rowspan="2">39</td><td>Men with CVDs</td><td>18</td></tr>
<tr><td>Healthy men</td><td>21</td></tr>
<tr><td>Total records</td><td colspan="5">225</td></tr>
</table>

Each people is required to pronounce a different sustained vowels in 3 separated records (/aaaaa…../, /oooo…./, /iiii…./), the duration of each recording is 5 seconds, and all records were done by a simple microphone from a smartphone, which we have settled his frequency at 44100Hz, and with noise attenuation at 20dB. We placed the microphone at 10 cm far from the pronouncer. All records are saved in WAV format, and in stereo-channel mode.

3. METHODOLOGY

In this part, we start by describing a machine learning approach to discriminate two kinds of people (CVDs people and healthy people), based on features extracted from voices previously recorded. Our extracted features will be the pitch of each segment of each record, and the Mel-Frequency Cepstral coefficients. These features serve as training sets for our classifier to build a model for each class, which means get the voiceprint of the CVDs people and voiceprint of the healthy people, and the test record will be compared against the voiceprints and the closest match is returned. The following diagram describe the approach used for CVDs people and healthy people classification:

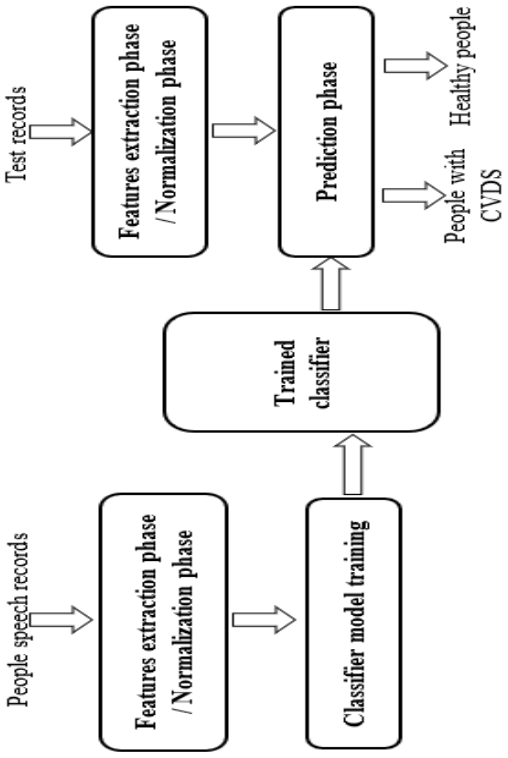

Fig. 2. Block Diagram of the used method

3.1. Features extraction

3.1.1. Pitch

In music information retrieval (MIR), speech coding or speech processing, the pitch detection represent one of a fundamental building block, also the pitch is used as an essential feature in machine learning system, to ensure the speech and speaker recognition [8][9][10]. The pitch is used as well in call centers in the aim to specify the gender of customers and his emotional state, another use of the speech pitch is to indicate and analyze pathologies and diagnose physical defects. In MIR, the pitch is used to categorize music, for query-by-humming systems, and as a primary feature in song identification systems. So, what is speech pitch?

The pitch is the fundamental period of the speech signal. It the perceptual correlate of the

fundamental frequency. It represents the vibration frequency of the vocal cords during the sound productions (like vowels, for example). It represents the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation.

We can define the speech pitch as the number of samples after which the waveform repeats itself is the pitch period in terms of the number of samples. If we know the sampling frequency, we can find the pitch period in seconds. Otherwise, in our algorithm, we calculate pitch by the default normalized autocorrelation approach, which requires a speech segment with at least two periods to find the pitch period. The Equation for Autocorrelation:

$$R_{xx}(x) = \sum_{i=1}^{2N+1} \frac{\big(a(i)a(i+k)\big)}{2N+1} \qquad (1)$$

Where:
N = window size
a = signal samples of the voiced segment
k = zero lag location

The main peak in the autocorrelation function is at the zero-lag location (k = 0). The location of the next peak gives an estimate of the period, and the height gives an indication of the periodicity of the signal.

3.1.2. MFCC

One of the most popular extracted features to resolve speech recognition tasks is the Mel Frequency Cepstral Coefficients "MFCC". Extracted from the speech signal, those coefficients give a good representation of the sound, it can be considered as a voiceprint of the speaker. The fig.3 shows the diagram which describe how to calculate the MFCC.

- **Fragmenting phase**

Pronounce a sustained vowel, can produce a voice signal which can be approximated as a stable signal only in a short duration 10 - 30 milliseconds [11], but still not stable for a long duration. From where we have to fragment each voice signal to a short fragment of 30ms. Another thing to mention is all fragments will be overlapped by 75%.

- **Windowing**

Windowing is used once we are interested in a signal of intentionally limited length. Indeed, a real signal can only have a limited duration in time. Moreover, a computation can be done only on a finite number of samples [12]. Our voice record represents a real signal, which means we have to window each frame, this phase is done by applying the Hamming window, using the following equation:

$$s_n' = \left[0.54 - 0.46 * \cos\frac{2\pi(n-1)}{(N-1)}\right] s_n \;;\; \{s_n, n = 1 \ldots . N\} \qquad (2)$$

The chosen windowing consists to get a portion of the signal which begins and ends at 0, to ensure a signal discontinuities reduction and gets the end smooth to be connected with the next beginning.

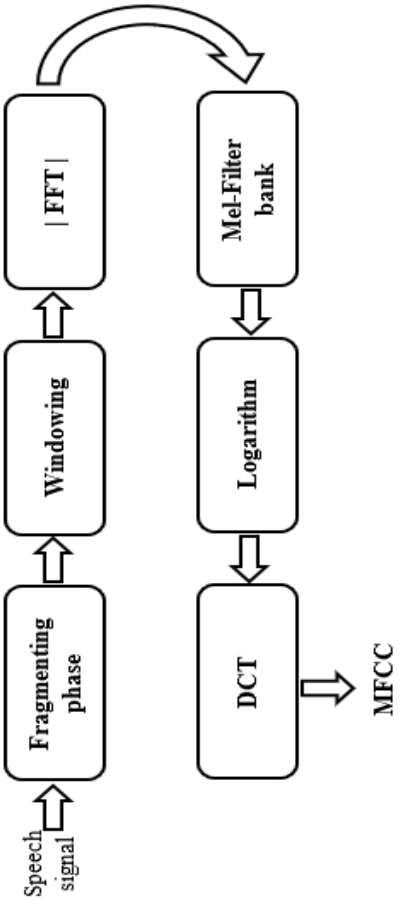

Fig. 3. Block diagram used to extract the MFCC

- **Fast Fourier Transform "FFT"**

The voice signal is a temporal signal, so to switch to the frequency domain we apply the FFT for each frame of N samples. The FFT allows us a fast implementation of Discrete Fourier Transform (DFT) [11], the DFT of a signal S_n of N samples is given by the following equation:

$$S_n = \sum_{k=0}^{N-1} s_k * e^{\frac{-2\pi jkn}{N}} \quad ; n = 0,1 \ldots ., N-1 \qquad (3)$$

Mel-Filter Bank

The Mel scale is a perceptual scale of pitches where we assign a perceptual pitch of 1000 Mel to 1000 Hz, in other words, the Mel scale is a linear function less than 1000Hz. Above 1000 Hz, the Mel-scale shows logarithmic progress (fig4) [13]. The Mel-filters bank are in triangular form and logarithmically spaced them thereafter (fig5). To convert hertz into Mel we used the most popular formula [14]:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

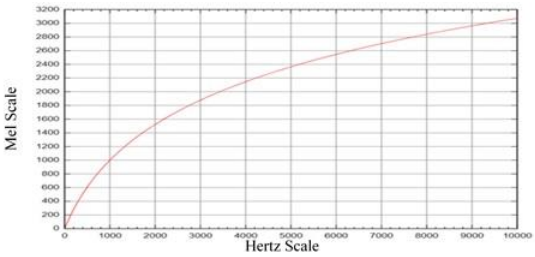

Fig. 4. The logarithmic progress for the Mel-scale

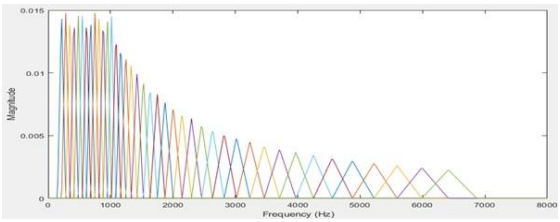

Fig. 5. Visualization of a typical Mel filter bank

- **Logarithm & DCT**

In the final step, we have to take the logs of amplitude at each Mel frequencies. Then, we have to apply the discrete cosine transform (DCT) for all the Mel log. We used the DCT defined in **[15][16],** for N filter bank channels as follow:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cdot \cos\left(\frac{i\pi}{N}(j - 0.5)\right) \quad (5)$$

3.2. Normalization

We have to mention that the extracted features (Pitch and MFCC) are not on the same scale, so we should normalize them. Consequently, we subtract the mean and we divide by the standard deviation of each feature extracted.

$$Feature'_n = \left(\frac{Feature_n - mean\ (Feature_n)}{std\ (Feature_n)}\right) \quad (6)$$

3.3. Training & testing classifier

In this phase, we train the K Near Neighbor (KNN), in the aim to build a model of classification. The KNN classifier is one of the simplest classifiers, the main idea resides in the K which represents the number of neighbors to take into consideration. Then the algorithm of classification calculates the distance between the K neighbors and the new entity to classify. The new entity must be the same class as of the majority class of K neighbors.

In our case, after many tries, the choice of the principal properties (K=number of neighbors and the type of the distance) of our classifier, was K=5 and a Euclidian distance to calculate, these proprieties were for maximum classification accuracy.

4. RESULTS & DISCUSSION

We dispose of a database which contains 225 records of the healthy and CVDs people, to train our classifier and test it, we choose to split randomly our database into two sets (training set and the test set), we take 80% of the total database to ensure the training phase, and 20% still to make the final blind test.

The process is previously described in the methodology section, it consists to extract features from a training set, train the KNN classifier, built a model, and use this model to classify the entities from the test set. Then, we calculate some parameters (Accuracy, Specificity, and Sensitivity) to judge our classification algorithm, we have to calculate two other parameters which show the quality of binary classification (MCC and PE) [12]. The equations of all parameters are as follow:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (8)$$

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

$$MCC = \frac{TP.TN - FN.FP}{\sqrt{(FP+TN)(FN+TP)(FP+TP)(FN+TN)}} \quad (10)$$

$$PE = \frac{TP.TN - FN.FP}{(FN+TP)(FP+TN)} \quad (11)$$

4.1. Entries

In the training phase, the entries of our algorithm are randomly chosen, from the whole database. So, we took 84 CVDs people audio records and 96 healthy people records. The records englobe the three sustained vowels (/aaaaa…../, /oooo…./, /iiii…./) which people have been invited to pronounce.

The first sight of the audio record signal, we have the impression that the CVDs left a signature in the voice people because two people at the same age (49 years) pronounced the same vowel (/aaaaa…../) for the same duration, doesn't generate signals with same pace. The next figure presents a signal from a subject without CVDs (a) and CVDs patient (b), pronouncing /aaaaa…../ for 5 seconds (fig. 6).

We remark the difference between the signals, but we should indicate that the people without CVDs when pronouncing a sustained vowel /aaaaa…../), show nearly a signal at the same pace. Another thing to mention, the pace presented by the CVDs people signal, can be generated by a lot of causes, so to build a classification algorithm just reporting to the pace of signals can be misleading for classification, so we have to extract the voiceprint for the CVDs people and a voiceprint for the people without CVDs.

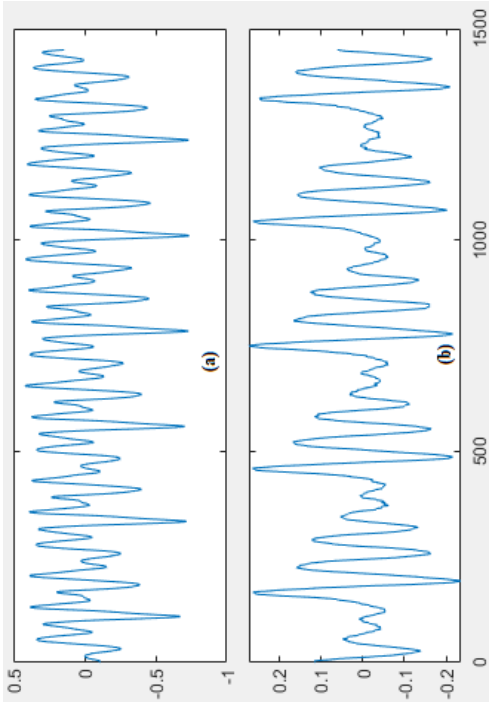

Fig. 6. (a) People without CVDs & (b) patient with CVDs pronounce the sustained vowel /aaa.../

4.2. FEATURES EXTRACTION

To extract the voiceprint in our case, we try to get 14 features (Pitch and 13 Mel Frequency Cepstral Coefficients) from each audio record, these features will help us to determine a voiceprint of our two classes (healthy people and patient with CVDs). We have 145 audio records to train our classifier, each record will be divided into portions of 30 milliseconds with an overlap of 75%.

The purpose of overlapping is primarily to reduce the effect of windowing. Most windowing functions (Blackman, Hamming, etc) are taper-shaped, which means that they drop to 0 (or close to 0) near the frame edges. This, of course, affects FFT results and we may lose some important information. So, to reduce this negative effect of the windowing, we use overlapping. The basic idea here is that we can average FFT results from overlapping frames and thus obtain a better frequency representation of our time-domain signal. The actual frequency resolution is still the same as without overlapping. As example, if our FFT length is 2048 samples, then an overlap of 1024 (50%) means that you have twice as many analysis (FFT) frames (as compared to the number of frames without any overlapping). 512 samples overlap (75%) means 4 times as many frames and so on. And the 75% give us the best frequency representation of our speech signal.

Then, we will extract the 14 features for every portion. Finally, we get a matrix with some hundred thousand features (14*353978) with which we will train our KNN classifier. The features before the normalization phase are not on the same scale when we talk about the pitch and the MFCC, thus it's important to normalize all features and we do that by applying the equation of normalization presented in the methodology section.

To discriminate between speech of people without CVDs (Healthy) and patients with CVDs, we proceed to calculate the average of each extracted features (from the 14 features) per each class of people, to get approximatively the voiceprints of our classes. The following figure shows the approximate voiceprints:

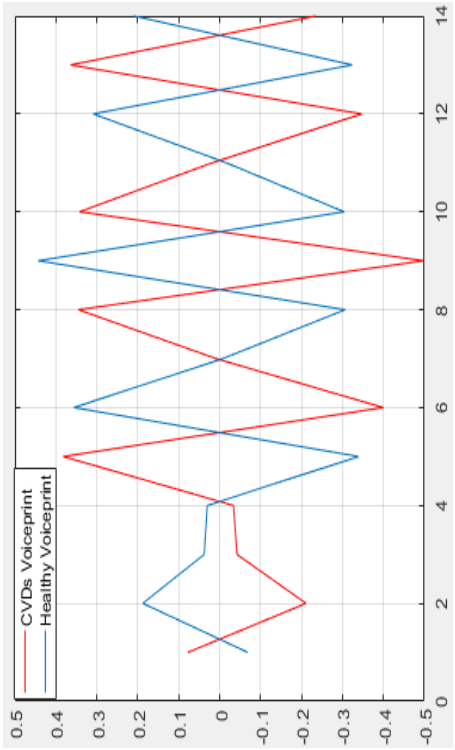

Fig. 7. Voiceprints of the patient with CVDs and of the healthy people

As shown, we can figure that the voiceprints of our two classes are different, which means we have a good opportunity to build a model able to discriminate between healthy people and patients with CVDs.

4.3. TRAINING & CLASSIFICATION TEST

We arrive to train our KNN classifier, so we have chosen 5 near neighbors and the Euclidean distance to search the neighbors, then we used the previously extracted features (Pitch and 13 Mel Frequency Cepstral Coefficients), as training matrix (features * number of observations = 14*353978) to train our classifier, and predefined function in Matlab (fitcknn) build the model.

The validation is ensured by the k-folds cross-validation, this type of validation divide the training set into k subsets, then it keeps one subset for the validation phase and trains the classifier by the k-1 subsets, and we repeat the routine k times until each subset may be used one time as a validation set. Each time we calculate the validation accuracy until

the k-folds cross-validation process finishes. Then the total validation accuracy must be the average of all validation accuracies calculated. The confusion matrix bellow shows the validation result:

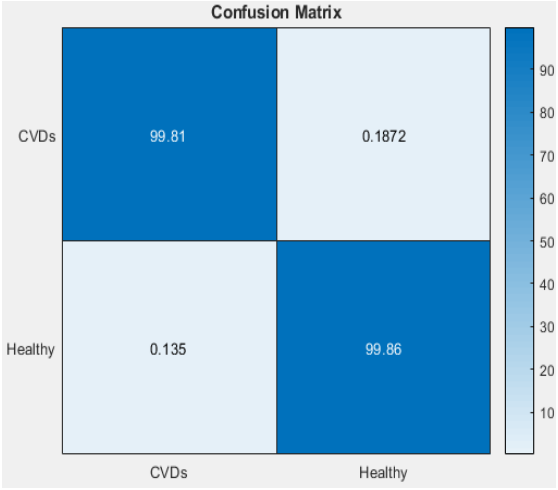

Fig. 8. The validation Confusion Matrix

As shown in the figure the validation accuracy is about 99.84%, this result was for the k-folds cross-validation process which uses the training set for validation. Let's see what would happen if we apply our algorithm on the test set (blind test).

The final test was done on 45 audio records (21 CVDs patients and 24 healthy people), the main idea consists in to classify all fragments from every record, and the class which represents the majority will be the class of the record. The following confusion matrix shows the reached results:

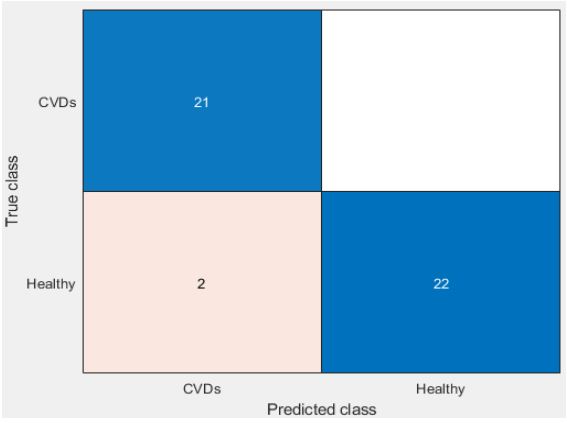

Fig. 9. The Confusion Matrix resulting from the blind test

To judge our results, we proceed to compute the parameters presented in the following table:

Table 2. Performance parameters result

Parameters	Results
Accuracy	**95.5%**
Sensitivity	100%
Specificity	91.67%
MCC	91.48%
PE	91.67%

From the computed parameters, our classifier reaches 95.5% of accuracy at a blind test, it also presents a good binary classification allowed to the results of MCC and PE. So, we can conclude that our algorithm of classification using the voiceprint is much better than the dysphonia measurement. We can also represent the ROC curve as shown in the figure below:

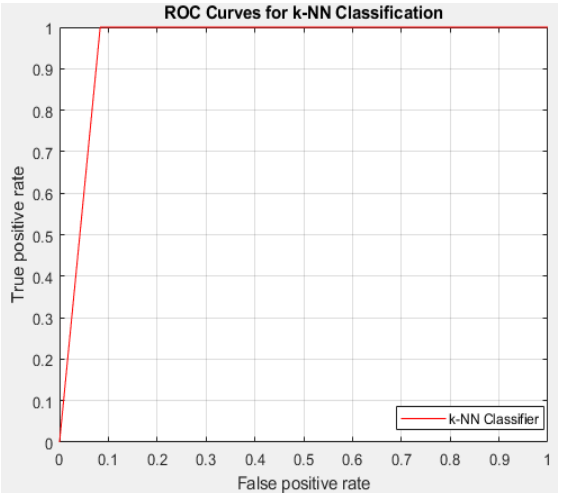

Fig. 10. The ROC curve resulting from the blind test

5. CONCLUSION

The cardiovascular diseases take many lives around the world, and his kill rate increases every year, but it's also possible to avoid 80% of the death by early diagnosis, so we have to make the diagnosis accessible to everyone, simple, fast, precise and inexpensive.

After our work in the classification of healthy people and patients with CVDs, using the features extracted by a dysphonia measurement, from audio records of different people pronounce sustained vowels /a/, /o/, and /i/. we have reached 81.5% of accuracy. We have chosen to ameliorate the process of discrimination and make it more precise.

In this order, we have tried this time to extract the voiceprint for each class of people, using the pitch and 13 MFC coefficients, extracted from audio records, then we train our KNN classifier, which gives us a good result at the final test by increasing the previous accuracy to 95.5%, so we conclude that the voiceprint is useful and much better to distinguish between our classes. But we affirm that the biggest challenge for our methodology will be the time of execution and the big size of data extracted from each audio record.

ACKNOWLEDGMENT

We would like to thank Dr. CHERTI, head of the cardiology B service, at the Ibn Sina hospital in Rabat, who offered us the opportunity to collect this database with which we worked. and we thank all

CVD patients who collaborated with us to record their voices.

REFERENCES

1. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. WHO. Global atlas on cardiovascular disease prevention and control Geneva 2011.
3. Rawther NN, Cheriyan J. Detection and classification of cardiac arrhythmias based on ECG and PCG using temporal and wavelet features. IJARCCE. 2015; 4.
4. Bouguila Z, Moukadem A, Dieterlen A, Ahmed Benyahia A, Hajjam A, Talha S, Andres E. Autonomous cardiac diagnostic based on synchronized ECG and PCG signal. 7th International Joint Conference on Biomedical Engineering Systems and Technologies—ESEO, Angers. 2014.
5. Ghassemian H, Kenari R. Early detection of pediatric heart disease by automated spectral analysis of phonocardiogram in children. J. Inf. Syst. Telecommun. 2015;3(2):66–75. https://doi.org/10.7508/jist.2015.02.001.
6. Nabih-Ali M, El-Dahshan E-SA, Yahia AS. Heart diseases diagnosis using intelligent algorithm based on PCGsignal analysis. Circuits Syst. 2017; 8(7): 184–190.
7. Bourouhou A, Jilbab A, Nacir C, Hammouch A. Classification of Cardiovascular disease using dysphonia measurement in speech. Diagnostyka. 2021;22(1):31-37. https://doi.org/10.29354/diag/132586
8. Carey MJ, Parris ES, Lloyd-Thomas H, Bennett S. Robust prosodic features for speaker identification. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, 1996;3: 1800-1803. https://doi.org/10.1109/ICSLP.1996.607979.
9. Jhanwar N, Raina AK. Pitch correlogram clustering for fast speaker identification. EURASIP J. Adv. Signal Process. 2004:37280. https://doi.org/10.1155/S1110865704408026
10. Atal BS. Automatic speaker recognition based on pitch contours. The Journal of the Acoustical Society of America. 1972; 52(6B): 1687–1697
11. Kumar ChS, Mallikarjuna PR. Design of an automatic speaker recognition system using MFCC, vector quantization and LBG algorithm. International Journal on Computer Scienceand Engineering. 2011; 3(8): 2942–2954.
12. Yang ZR, et al. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics. 2005; 21(16):3369–3376.
13. https://www.pngwave.com/png-clip-art-vijgd
14. https://en.wikipedia.org/wiki/Mel_scale
15. Benba A, Jilbab A, Hammouch A. Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ. In The 2014 international conference on circuits, systems and signal processing, 23–25 September 2014. Saint Petersburg: Saint Petersburg State Polytechnic University 2014..
16. Young S, Evermann G, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P. The HTK book (for HTK version 3.4). Cambridge: Cambridge University Engineering Department. 2006.
17. Bourouhou A, Jilbab A, Nacir C; Hammouch A. Detection and localization algorithm of the S1 and S2 heart sounds. 2017 International Conference on Electrical and Information Technologies (ICEIT), Rabat. 2017:1-4. https://doi.org/10.1109/EITech.2017.8255217.
18. Bourouhou A, Jilbab A, Nacir C, Hammouch A. Comparison of classification methods to detect the Parkinson disease. 2016 International Conference on Electrical and Information Technologies (ICEIT), Tangiers. 2016:421-424. https://doi.org/10.1109/EITech.2016.7519634.
19. Bourouhou A, Jilbab A, Nacir C, Hammouch A. Heart sounds classification for a medical diagnostic assistance. International Journal of Online and Biomedical Engineering (iJOE). 2019;15(11):88–103.
20. Benba A, Jilbab A. Hammouch A. Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people. Int J Speech Technol. 2016;19:449–456. https://doi.org/10.1007/s10772-016-9338-4

Received 2021-02-06
Accepted 2021-09-24
Available online 2021-10-01

Abdelhamid BOUROUHOU was born in Rabat, Morocco on December 26th, 1989. Received the Master degree in Electrical Engineering from ENSET, Rabat Mohammed V University, Morocco, in 2014 he is a research student of Sciences and Technologies of the Engineer in ENSIAS, Research Laboratory in Electrical Engineering LRGE, Research Team in Computer and Telecommunication ERIT at ENSET, Mohammed V University, Rabat, Morocco. His interests are in sounds classification for medical diagnostic assistance.

Abdelilah JILBAB Professor at ENSET Rabat, Morocco; he graduated in electronic and industrial computer aggregation in 1995. Since 2003, he is a member of the laboratory LRIT (Unit associated with the CNRST, FSR, Mohammed V University, Rabat, Morocco). He acquired his PhD in Computer and Telecommunication from Mohammed V-Agdal University, Rabat, Morocco in 2009. His domains of interest include signal processing and embedded systems.

Mohammed CHERTI
Pr en cardiologie, Professor of Cardiology at the Faculty of Medicine and Pharmacy De rabat.

Head of service Cardiologie B.
CHU IBNSINA-Rabat

Zaineb BOUROUHOU, Doctorate in general medicine obtained in 2020, currently resident doctor in cardiology at the Rabat CHU.

NACIR Chafik
Teacher Researcher in Mathematics.
Former Head of the Department of Mathematics and Computer Science.
Former member of the Scientific Commission ENSET of Rabat Morocco.