SUPPORT VECTOR MACHINE FOR HUMAN IDENTIFICATION BASED ON NON-FIDUCIAL FEATURES OF THE ECG

HATEM ZEHIR[1*], TOUFIK HAFS[1], SARA DAAS[1], AMINE NAIT-ALI[2]

[1]LERICA, Faculty of Technology, Badji Mokhtar-Annaba University, B.O. Box 12, Annaba, 23000 Algeria

[2]L.I.S.S.I., University of Paris 12, 61 Avenue du Général de Gaulle, 94010 Créteil, France

Abstract: The demand for reliable identification systems has grown recently. Using the mean frequency, median frequency, band power, and Welch power spectral density (PSD) of ECG data, we proposed a novel biometric approach in this study. ECG signals are more secure than other traditional biometric modalities because they are impossible to forge and duplicate. Three different support vector machine classifiers—linear SVM, quadratic SVM, and cubic SVM—are employed for the classification. The MIT-BIH arrhythmia database is used to evaluate the suggested method's precision. For the linear SVM, quadratic SVM, and cubic SVM, respectively, test accuracy of 93.6%, 96.4%, and 97.0% was obtained.

Keywords: biometrics, hidden biometrics, security, identification, ECG, machine learning, SVM

1. INTRODUCTION

Classical techniques used for identification and authentication, such as passwords, tokens, and ID cards are no longer up to the security standards as they can be easily stolen, falsified, and spoofed. This implies that a more robust technique is needed, namely biometrics, which can be defined as the science of processing human biological data for identification and authentication purposes.

Unlike traditional methods that are used for identification, biometric systems are more secure and efficient, for example, individuals are not required to memorize long and sophisticated passwords, or carry with them access keys or smart cards all the time. Biometrics overcomes all of these disadvantages, as they are fortified and immune against falsification, duplication, spoofing, etc.

Biometric modalities are classified into three categories, which are, physiological modalities such as iris and face, behavioural modalities such as keystroke and voice, and finally hidden biometric modalities such as the electrocardiogram (ECG) signal and electroencephalogram (EEG) signal.

Biometric modalities respond to most of these criteria:

- Collectability: easy for data acquisition. Non-expensive and non-intrusive;
- Acceptability: subjects are ready to use modality, with ease;
- Uniqueness: different between two subjects;
- Universality: measurable on every subject;

[*] Corresponding author, email: zehir.hatem@univ-annaba.org

doi.org/10.29081/jesr.v29i1.007

- Permanence: does not change over time;
- Circumvention: cannot be falsified;
- Aliveness Detection: only present on a living subject.

The first authors who proved that the electrical activity of the heart can be used as a biometric modality are Biel et al. [1], in their paper, they introduced a novel technique where they collected ECG signals from 20 subjects aged between 20 and 55 during rest using the 12 conventional leads. The signals were acquired by a SIEMENS Megacart, and the number of recordings for each individual varied from 4 to 10. For the features selection, the authors extracted 30 different fiducial features, including QRS wave duration (ms), QRS wave deflection (ms), QRS wave peak-to-peak amplitude (µV), and ST segment slope (90.90) deg. And, for the classification they used the soft independent modelling of class analogy (SIMCA) [2], their system showed that even the data extracted from a single lead only is sufficient to perform identification applications.

Since then, many researchers conducted their research in the field of ECG biometrics. Kim et Pyun [3] achieved a maximum classification accuracy of 100% and 99.8% when they tested their algorithm on the MIT-BIH Normal Sinus Rhythm (NSRDB) and the MIT BIH Arrhythmia (MITDB) databases respectively. To pre-process the signal, the authors used two types of filters, first, they applied a derivative filter and then a moving average filter, after denoising the signal, they normalized the amplitude according to the following equation:

$$y[n] = 2(x[n] - x_{median}) / (x_{\max} - x_{\min}) \tag{1}$$

After removing the noise from the signal, they segmented the signal, and because of the difference in sampling frequency between the two databases used, 288 samples for each segment are grouped for the NSRDB and 444 samples for the MITDB.

In another paper, [4] proposed a method based on single lead data collected from 269 subjects, the signals were collected from each individual during three separate days. The power line interference was removed using a notch filter with a cut-off frequency of 60 Hz, and then the signal was segmented into windows of 0.7 seconds around each R peak, as classification features the authors used time-frequency features, and they computed the spectrogram. When the used training and testing data were collected from separate days, they were able to achieve an equal error rate (EER) of 5.58% in verification, an accuracy of 76.9% in rank-1 recognition, and an accuracy of 93.5% in rank-15 recognition. When the data is from the same day, the authors achieved an accuracy of 99% and an equal error rate of 0.37%.

In a more recent research, [5] used three different parameters, Entropy, Cepstral Coefficient, and Zero Crossing Rate to extract the features of ECG signals of 54 subjects from the hysikalisch-Technische Bundesanstalt(PTB) diagnostic database. To pre-process the signal, the authors applied a filter with a bandpass of [2 Hz - 50 Hz], after denoising the contaminated signal, they segmented the signal into heartbeats by detecting the T peaks, then, the beats were normalized to have the same size. After the pre-processing and features extraction phase, the data was classified using k-nearest neighbours and support vector machine algorithms, the best-achieved accuracy was 95.4%.

As we can see from the previous researches, the electrocardiogram signal is a very reliable biometric trait, that achieves high accuracy for both identification and authentication, and biometric systems based on the electrical activity of the heart prove themselves as a prominent and emerging security tool, hence in our study we are interested in the ECG.

In this paper, we will focus on developing a biometric system based on the electrocardiogram (ECG) signal, it is a relatively new and robust biometric modality. It has been proven that the heartbeat differs from one person to another, and this is even among identical.

The ECG represents the electrical activity of the heart, it is mostly used for diagnostic purposes by medical staff, but in recent years researchers [1] have found that it can be used as a biometric modality with a very high accuracy rate.

With the technological development that has taken place in recent years, the matter of data acquisition of the ECG signal (Fig**. 1** 1) is no longer a complicated issue, as many low-powered, wearable devices used in everyday life already do this, such as smartwatches, and this is what makes its acceptance as a biometric modality very feasible.

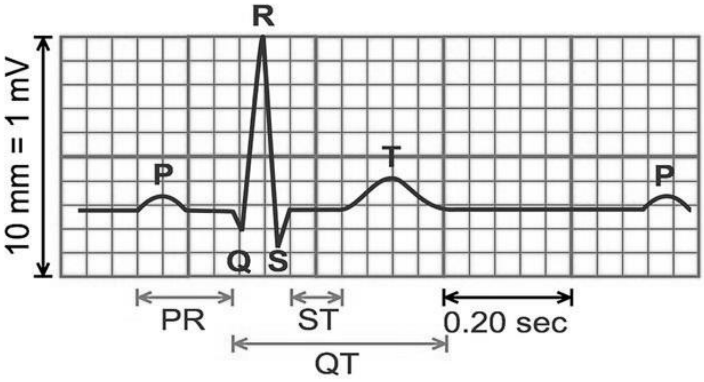

Fig. 1. The ECG signal components [6].

The contribution of our work includes:

- Classifying the features of short segments of ECG signals using SVM;
- The extracted features are the mean frequency, the median frequency, the band power and the welch spectrum;

To the best of our knowledge, we are this paper is the first to discuss so.

The rest of this paper is organised as follows: in section 2, the proposed system is detailed. Section 3 discusses the experimental findings, and finally, section 4 presents the conclusion and the prospects of this paper.

2. PROPOSED METHOD

In this section, our proposed method will be discussed in detail. We will start by describing the database used to evaluate the system, then, we will describe the pre-processing phase where we denoise the signal and segment it into different heartbeats, after that the features extracted from the signal are detailed. After that, the extracted features are classified using a support vector machine algorithm. Figure 2 shows an overview of the proposed system.

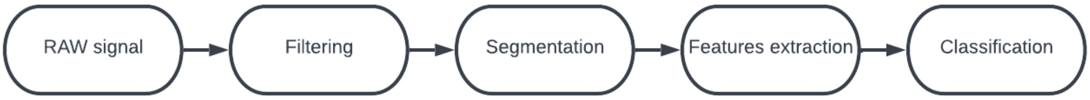

Fig. 2. General architecture of our proposed system.

2.1. Database

In order to evaluate our proposed system, we have used the well-known and reputable MIT-BIH Arrhythmia Database [7]. It contains data collected from 47 different subjects at Boston's Beth Israel Hospital. Of the 47 subjects from whom samples were collected, 22 were women and 25 were men, ranging in age from 23 to 89 years. The signals were acquired from two-channels using a Del Mar Avionics model 445. The database contains 48 recordings of 30 minutes each collected at a sampling frequency of 360 Hz and 11 bits resolution. Only one recording was collected from all forty-seven subjects, except for one subject from which two recordings were collected.

Data collection on this database began in 1975, that is, long before the start of using electrocardiogram signals as a biometric modality. Despite this, it was used in many state-of-art scientific papers [8-11] published in prestigious scientific journals concerned with the biometric field, as the signals of this database are proven that they can be used for identifying and authenticating individuals.

2.2. Pre-processing

The pre-processing phase is divided into two stages, the first one is filtering to remove the existing noise from the signal, and the second one is the R peaks detection in order to segment the signal.

2.2.1. Filtering

In practice, the ECG signal can be contaminated by various types of noises, these noises include but are not limited to [6, 12]*:*

- Baseline Wander: it is a low-frequency noise, it can be caused by breathing, the used electrodes, etc. The frequency of this type of noise ranges from 0.15 to 0.30 Hz;
- Power-Line Interference: It is considered among the most common types of noise, and it is caused by the electric current of the power line. It is a signal in the form of a sinusoidal function with a frequency of 50 or 60 Hz, depending on the region;
- Motion Artefacts: It is impossible for the subject to remain stationary during the signal acquisition process, and this inevitably results in noise generation;
- Muscle Noise: also known as Electromyography Noise, it is caused by the electromyographic signal (EMG) resulted when the muscles of the subject are contracted;
- Instrumentation noises: this type of noise is caused by external equipment such as the device used to collect the ECG signal.

This noise directly affects the process of recognizing people, which makes it more than necessary to get rid of it through filtering and signal processing techniques. To do so, we have used a 4th-order Butterworth filter with a bandpass of [1 Hz - 40 Hz].

A bandpass filter is a filter that allows only frequencies present in a certain range to pass, all the frequencies outside that range are eliminated.

Fig. **3** shows, the raw signal before filtering and the denoised signal.

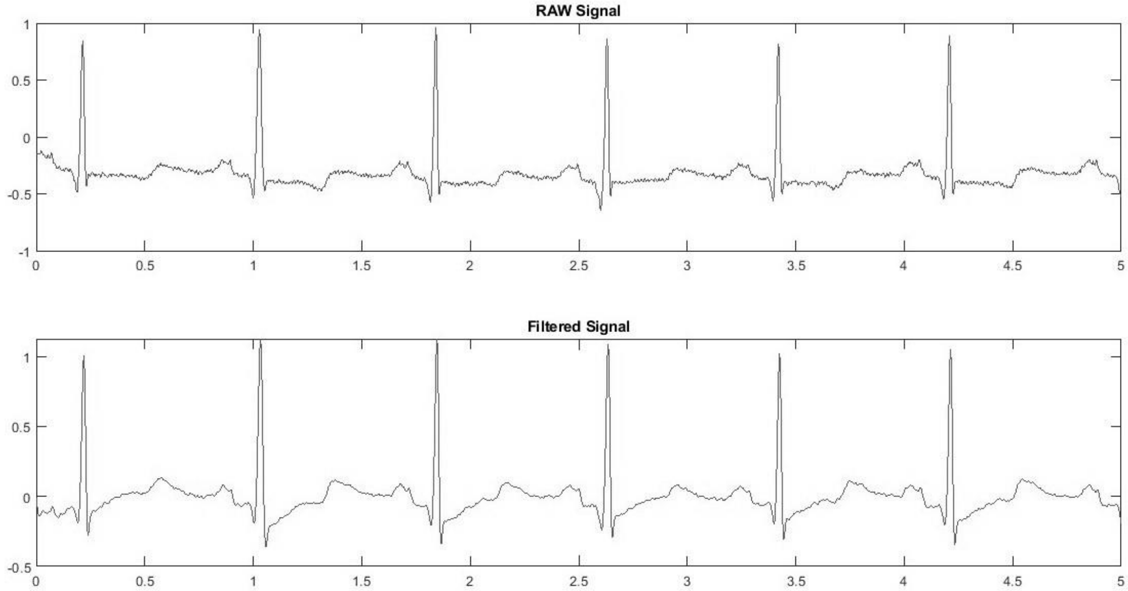

Fig. 3. A segment of the preprocessed signal in the filtering stage.

2.2.2. Segmentation

In order to segment our signals, we have implemented the same method that [13] used. First, we have to detect the R peaks of the whole signal, this is done by using the Pan-Tompkins algorithm [14] and we have used the MATLAB implementation introduced by [15], the main steps of the code implementation are shown in Figure 4 and can be described by the following steps:

- Pre-processing: in this phase, the signal is first bandpass filtered with [5 Hz - 15 Hz] cut-off frequencies, and then a derivative filter is applied before squaring the signal and averaging it;

- Decision Rule: this phase can be divided into:
    - The detection of fiducial points;
    - Adaptive thresholding;
    - Looking back for any missed peaks;
    - Elimination of the duplicate detection that lies in the range of 200ms;
    - Discriminate the T waves.

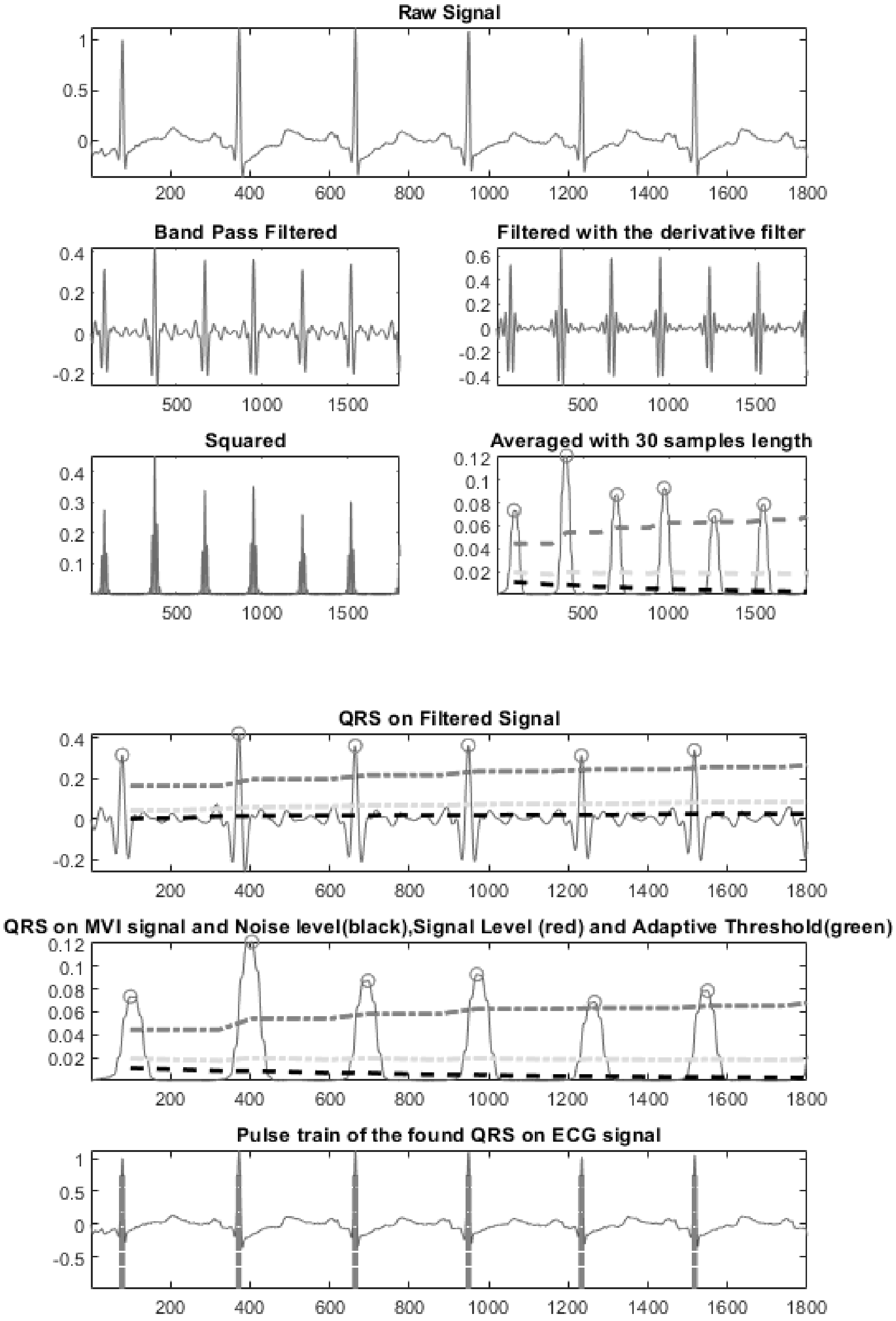

Fig. 4. Detecting R peaks using the Pan-Tompkins algorithm.

When the peaks are detected, we use the index of each one of them to create a heartbeat segment, 125 samples before and after each peak are concatenated. Taking into account the sampling frequency of our signals which is 360 samples per second and the 251 samples of each of our segments, we can deduce that the duration of each segment is equal to 697 ms, different heartbeats are presented in Figure 5.

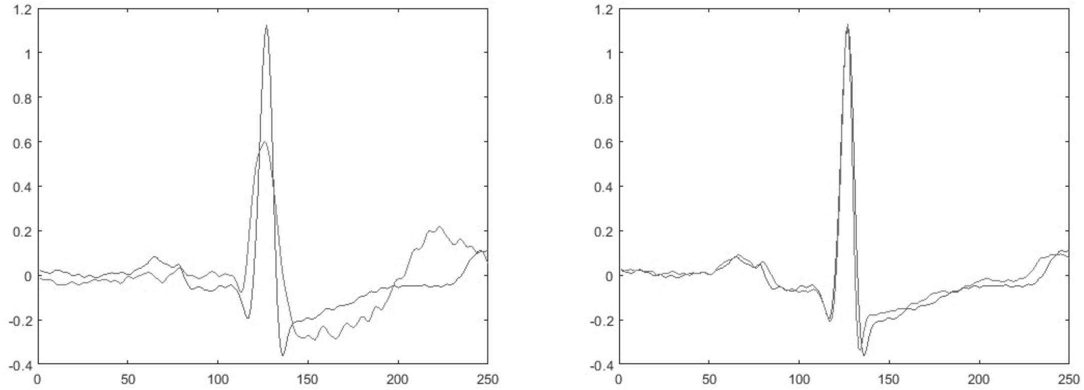

Fig. 5. Different heartbeat segments side by side. Left: two heartbeats of two different individuals. Rigt: two heartbeats of the same individual

2.3. Features extraction

In this section, we will discuss the extracted features of the signal. The most critical phase of any machine learning system is feature extraction. The original information is processed at this step in order to extract important information that may be utilized to best explain the raw data, with the objective of increasing accuracy and decreasing training time and memory usage.

Many feature extraction techniques are proposed in the literature. These techniques can be classified into two main categories:

- Fiducial features: the aim of this approach is to extract features related to the fiducial point of the ECG signal, i.e. the most prominent peaks, extracted features include: R wave duration, R wave amplitude, P wave onset, ST segment amplitude, etc.;
- Non-fiducial features: This technique makes it possible to extract information related to the morphology of the complete signal or just a part of it. These characteristics can be statistical, frequency domain or time-frequency domain. Tools used to extract these features may include but are not limited to: discrete wavelet transform (DWT), short-Time Fourier Transform (STFT), and empirical mode decomposition (EMD).

In this study, we are primarily concerned with the signal's non-fiducial properties, specifically the mean frequency, median frequency, band power, and welch power spectral density.

2.4. Classification

Now, after we extracted the features, we need to classify them using an appropriate algorithm, for this, we have used three different support vector machine classifiers.

2.4.1. Linear SVM

Support vector machine (SVM) classifiers are supervised machine learning algorithms. SVM was originally designed for binary classification problems, but its application has expanded to include multi-class classification also Because of its accuracy, SVM remains one of the most popular algorithms in terms of data categorization.

The primary goal of linear SVM classifiers is to maximize the separation between the support vector and the hyperplane, and this according to the following equation:

$$\overrightarrow{W} \times \overrightarrow{X} + b = 0 \tag{2}$$

The linear SVM's kernel equation is defined as follows:

$$k(X, X_i) = X \times X_i + c \tag{3}$$

2.4.2. Quadratic SVM

The quadratic SVM approach is non-linear. Its's kernel equation is defined as follows:

$$k(X, X_i) = (X \times X_i + c)^2 \tag{4}$$

2.4.3. Cubic SVM

The cubic SVM is also a non-linear method. The kernel equation of cubic SVM is defined by:

$$k(X, X_i) = (X \times X_i + c)^3 \tag{5}$$

We can deduce from (3), (4), and (5) that the general equation for polynomial kernel is given by:

$$k(X, X_i) = (X \times X_i + c)^n \tag{6}$$

3. RESULTS AND DISCUSSION

To evaluate the system that we have proposed, the MIT-BIH arrhythmia database was used, it includes the recording of ECG data including both healthy and ill people. Being one of the most popular ECG databases, it contains signals taken from 2 leads and 47 distinct people. Only lead I signals are selected in our method's testing since [1] has established the validity of single lead recordings for biometric identification.

In this study, we proposed a novel method for human identification from ECG signals. This method is based on the extraction of the signal's mean frequency, median frequency, band power, and welch power spectral density features, and it classifies the features using a support vector machine.

The recordings are first denoised with a band-pass Butterworth filter with cut-off frequencies of 1 Hz and 40 Hz. Next, we segment the filtered signals around each detected r-peak detected using the pan-Tompkins algorithm, with each segment lasting roughly 700 ms. The mean frequency, median frequency, band power, and welch power spectral density features of the signal are the retrieved features in the second stage, as was previously stated. The classification stage is the final step in our proposed system, and we employ three different support vector machine classifiers—linear SVM, quadratic SVM, and cubic SVM.

For the classification, we divide the data randomly into 80% training and 20% testing. As a validation, the k-fold cross-validation value was set to 5. Table 1 shows the results of the experiment.

Table 1. Classification results.

	Linear SVM	Quadratic SVM	Cubic SVM
Accuracy	93.6%	96.4%	97.0%

As shown in Table 1, all the classifiers produce acceptable results, with the cubic SVM producing the highest accurate results with an accuracy of 97.0%, outperforming several state-of-the-art ECG biometric systems [16–18] as can be seen in Table 2. The classifier with the lowest accuracy, 93.6%, was the linear SVM.

Table 2. Obtained results in comparison to some state-of-the-art methods.

Authors	Database	Performance
Zhang et al. [16]	ECG-ID MIT-BIH	92.53% 91.31%
Lee et al. [19]	CU-ECG	98.48%
Zihlmann [17]	ECG-ID MIT-BIH	90.70% 91.15%
Our proposed method	MIT-BIH	97.00%

As we can see in Table 1, the obtained results are 93.6%, 96.4%, and 97.0% for the linear SVM, quadratic SVM, and cubic SVM respectively. The higher the order of the polynomial kernel, the better the results. Hence, the degree of the polynomial directly affects the classification accuracy. And higher-order kernels allow a better separation between the categories as shown in Figure 6.

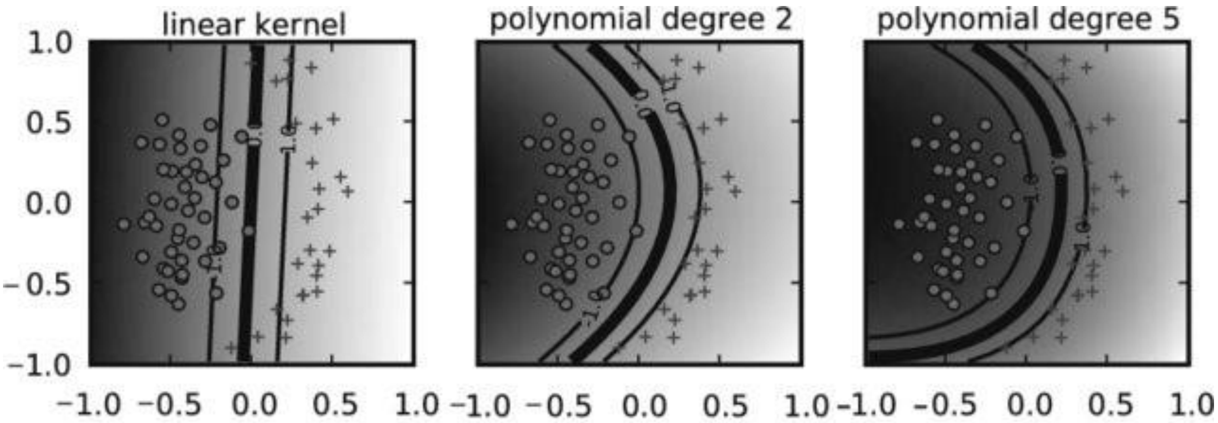

Fig. 6. The effect of the degree of a polynomial kernel. Higher degree polynomial kernels allow a more flexible decision boundary [20].

Therefore, this study shows that a kernel of a higher-order polynomial outperforms kernels with a lower order.

4. CONCLUSION

In this study, we introduced a new approach for analyzing ECG signals based on their mean frequency, median frequency, band power, and welch power spectral density (PSD). Later, those features were categorized using three different SVM - based classifiers: linear SVM, quadratic SVM, and cubic SVM.

With the cubic SVM classifier, the proposed approach achieves very good performance, with an accuracy rate of 97.0%. The linear SVM was the classifier with the lowest accuracy of 93.6%.

In our future works, we will concentrate on reducing dimensionality and improving overall accuracy by extracting only the most significant characteristics using techniques such as principal component analysis (PCA) and implementing our work in an embedded system.

REFERENCES

[1] Biel, L., Pettersson, O., Philipson, L., Wide, P., ECG analysis: a new approach in human identification, IEEE Transactions on Instrumentation and Measurement, vol. 50, no. 3, 2001, p. 808-812.

[2] Esbensen, K., Schonkopf, S., Midtgaard, T., Multivariate analysis in practice: training package, Computer-Aided Modelling, 1995.

[3] Kim B.H., Pyun J.-Y., ECG identification for personal authentication using LSTM-based deep recurrent neural networks, Sensors, vol. 20, 2020, art.no. 3069.

[4] Odinaka, I., Lai, P.H., Kaplan, A.D., O'Sullivan, J.A., Sirevaag, E.J., Kristjansson, S.D., Sheffield, A.K., Rohrbaugh, J.W., ECG biometrics: A robust short-time frequency analysis, 2010 IEEE International Workshop on Information Forensics and Security, 2010, p. 1-6.

[5] Hamza, S., Benayed, Y., Recognition of person using ECG signals based on single heartbeat, Intelligent Systems Design and Applications, 2022, p. 452-460.

[6] Musa, N., Gital, A.Y.U., Aljojo, N., Chiroma, H., Adewole, K.S., Mojeed, H.A., Faruk, N., Abdulkarim, A., Emmanuel, I., Folawiyo, Y.Y., Ogunmodede, J.A., A systematic review and meta-data analysis on the applications of deep learning in electrocardiogram, Journal of Ambient Intelligence and Humanized Computing, 2022, p. 1-74.

[7] Moody, G.B., Mark, R.G., The impact of the MIT-BIH arrhythmia database, IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 3, 2001, p. 45-50.

[8] Salloum, R., Kuo, C.C.J., ECG-based biometrics using recurrent neural networks, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, p. 2062-2066.
[9] Li, R., Yang, G., Wang, K., Huang, Y., Yuan, F., Yin, Y., Robust ECG biometrics using GNMF and sparse representation, Pattern Recognition Letters, vol. 129, 2020, p. 70-76.
[10] Jyotishi, D., Dandapat, S., An LSTM-based model for person identification using ECG signal, IEEE Sensors Letters, vol. 4, no. 8, 2020, p. 1-4.
[11] Boujnouni, I.E., Zili, H., Tali, A., Tali, T., Laaziz, Y., A wavelet-based capsule neural network for ECG biometric identification, Biomedical Signal Processing and Control, vol. 76, 2022, art. no. 103692.
[12] Uwaechia, A.N., Ramli, D.A., A Comprehensive survey on ECG Signals as new biometric modality for human authentication: recent advances and future challenges, IEEE Access, vol. 9, 2021, p. 97760-97802.
[13] Lynn, H.M., Pan, S.B., Kim, P., A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks, IEEE Access, vol. 7, 2019, p. 145395-145405.
[14] Pan, J., Tompkins, W.J., A real-time QRS detection algorithm, IEEE Transactions on Biomedical Engineering, no. 3, 1985, p. 230-236.
[15] Sedghamiz, H., Matlab implementation of Pan Tompkins ECG QRS detector, 2014, p. 3, https://www.researchgate.net/publication/313673153_Matlab_Implementation_of_Pan_Tompkins_ECG_QRS_detector (10.02.2023).
[16] Zhang, Q., Zhou, D., Zeng, X., HeartID: a multiresolution convolutional neural network for ECG-based biometric human identification in smart health applications, IEEE Access, vol. 5, 2017, p. 11805-11816.
[17] Zihlmann, M., Perekrestenko, D., Tschannen, M., Convolutional recurrent neural networks for electrocardiogram classification, Computing in Cardiology (CinC), 2017, p. 1-4.
[18] Yildirim, Ö., A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification, Computers in Biology and Medicine, vol. 96, 2018, 189-202.
[19] Lee, J.N., Kwak, K.C., ECG-based biometrics using a deep network based on independent component analysis, IEEE Access, vol. 10, 2022, p. 12913-12926.
[20] Ben-Hur, A., Weston, J., A user’s guide to support vector machines, Methods in Molecular Biology, vol. 609, 2010, p. 223-39.