Comparison between K-Nearest Neighbor and Support Vector Machine Algorithms for PPG Biometric Identification

Aya Al Sidani, Ali Cherry, Houssein Hajj-Hassan, and Mohamad Hajj-Hassan
Lebanese International University, Bekaa, Lebanon
The International University of Beriut, Beirut, Lebanon
mohamad.hajjhassan@liu.edu.lb

***Abstract*—with the great evolution in information technology, data mining and its techniques became a need in almost every application. One of the data mining techniques is data classification, in this paper 2 classification techniques are compared, the K-nearest neighbor and the Support vector machine due to their several advantages. The classification techniques are implemented and tested on Photoplethysmography signals after extracting 40 features from the signals. This work was made in order to check the efficiency of using the Photoplethysmography signals as biometric identification technique and choose the best classification technique for this application.**

***Keywords*— Classification, K-nearest neighbor, Support vector machine, Feature extraction, Biometric identification**

I. *Introduction*

Data mining is the mechanism for finding a specific arrangement or template out of large data; it includes elimination, adjustment, classification, integration, analysis of data and much more... Classification is one of the data mining techniques; it is the process that estimates and forecasts the class or the category of a specific data [1].

Classification consists of 2 major phases, the first one is the learning or training phase where the classifier is built, the second phase is the classification phase where new data is classified into its appropriate class according to the training set. There are many classification techniques such as neural network, support vector machine, K-nearest neighbor, fuzzy logic… [1].

For each application, several classification techniques could be implemented to check the application's efficiency, for this application the K-NN and SVM classifiers were implemented on Photoplethysmographic signals to test its efficiency as a biometric technique [2].

The Photoplethysmographic signal is extracted using the SpO_2 sensor which is a non-invasive method used to measure oxygen saturation and give the PPG waveform by placing it on the finger of the individual. The SpO_2 sensor is made up of 2 LEDs that emit red and infrared signals, these signals pass through the blood vessels in the finger where only the unabsorbed ones reach the photodetector on the opposite side of the sensor, these unabsorbed signals are then processed where the oxygen saturation is calculated and the PPG signal is displayed.

The displayed photoplethysmogram waveform has both AC and DC components; however the AC components can be extracted easily from the DC components. The resulting signal consists of 2 peaks, the first one is the systolic peak followed by a dicrotic notch and then the diastolic peak as shown in Figure 1. From this signal, several features that are unique to each individual can be extracted which makes it a good source for biometric identification [2].

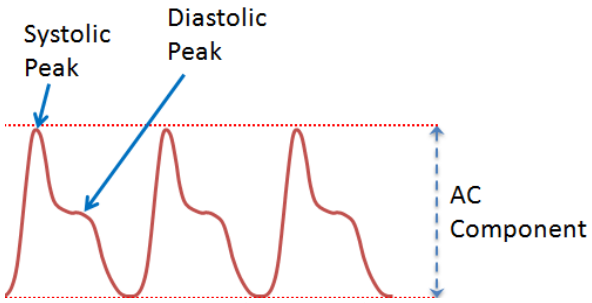

Fig. 1: The PPG signal

In this paper, only the classification step will be discussed in details since the aim is to compare the classification techniques and choose the ideal one among them to be used in the biometric identification technique.

II. *Related Works*

Classification techniques are classified into basic and advanced techniques. For our application, 2 advanced techniques were chosen, the K-nearest neighbor known as K-NN and the Support vector machine known as SVM. In this section, a summary of the classification methods is introduced and then some related works and applications on each method are discussed.

A. *K-Nereast Neighbor*

The K-nearest neighbor illustrated in Figure 2 depends on a distance metric where this distance is measured between the query points having unlabeled class and the saved training set, the k-closest samples which have the smallest distances are chosen. The most used distance is the Euclidean Distance represented by the below equation:

978-1-7281-2314-1/19/$31.00 ©2019 IEEE

$$E(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (1)$$

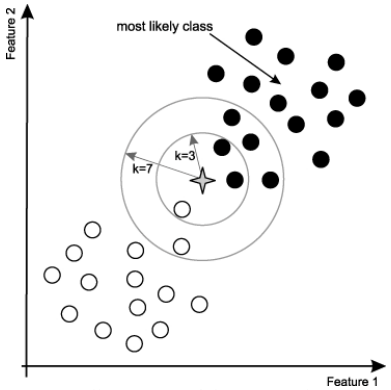

Fig. 2: Illustration of the K-NN algorithm

B. Support Vector Machine

The Support Vector Machine consists of 2 steps. The first step is displaying the input data or classes as vectors in a space having the dimensions equal to the number of features where the vectors found on the boundaries (the neighboring vectors) are called the support vectors. The second step is finding a hyperplane that separates these classes with a maximum margin between the support vectors; this margin is calculated by constructing 2 parallel hyperplanes on the opposite sides of the main hyperplane until they touch the support vectors [1].

SVM works on both linearly and non-linearly separated data where for the non-linearly separated data a kernel function is used, many kernel functions could be used but radial basis function (RBF) is the most common function used. SVM can work for multiple classes also, either by comparing several times one class versus one or by comparing one class versus all the other classes [1].

C. Related Work

There are several real-life applications based on SVM and K-NN classifiers. For example, the SVM is used in image classification, in bioinformatics and computational biology for classification of genes, and in handwriting recognition for confirming electronic signatures. One of the important applications of SVM is in cancer classification where SVM was used to classify 2 different types of Leukemia without prior knowledge and information of these classes [3].

K-NN is used in applications that depend on classifying similar objects such as pattern recognition applications; one of the important usages of K-NN classifier is to identify patterns in credit card operations to detect unusual patterns that point to doubtful behavior such as exchanging or returning the same object several times [4].

III. Methodology

The application consists of several steps. First, a dataset consisting of different signals was downloaded online and used, then 40 features were extracted from the signals after computing its first and second derivatives, and finally 2 classifications methods were tested. The feature extraction and classification processes were done using Matlab [2].

A. The Dataset Used

The first step to test the accuracy of the PPG as a biometric identification technique was by applying it on a downloaded dataset online [2]. The dataset was taken from the Vortal Study, National Clinical Trial no. 01472133 [5], it consists of 57 subjects; 41 subjects belong to young people having ages between 18 and 40 years that were divided into 2 groups according to the shape of the PPG signal, group (a) had a normal PPG signal as shown in Figure 3 whereas group (b) consisted of signals having a very smooth diastolic peak or none at all as shown in Figure 4. The remaining 16 subjects named group (c) belonged to old people having ages above 70 years where the signals were similar to group (b) where there was no diastolic peak as shown in Figure 5 since the PPG signal changes due to aging [2].

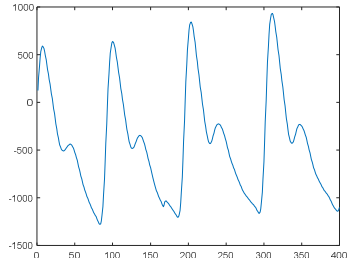

Fig. 3: The PPG signal of a sample belonging to group (a)

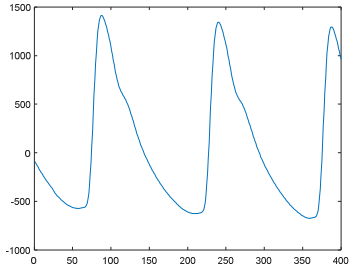

Fig. 4: The PPG signal of a sample belonging to group (b)

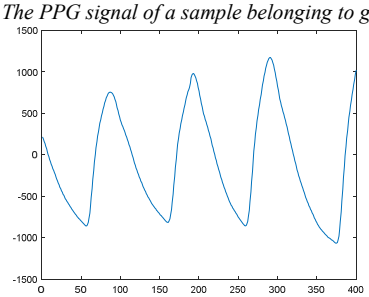

Fig. 5: The PPG signal of a sample belonging to group (c)

B. Feature Extraction

The first step in this part was computing the first and second derivatives of the signal and the second step was computing the 40 features from the signal and its derivatives. Some of these features are the systolic peak, diastolic peak, peak to peak interval, pulse interval, augmentation index and other features [6]. More details on the implementation of the feature extraction phase were described in our previous paper

"Biometric Identification Using Photoplethysmography Signal" [2].

C. Classification

For the K-NN classifier, the K was chosen to be 1, so that after comparing the input signal with all the samples saved in the database, only one sample, the one with the smallest distance, can be chosen to match the input signal since our main aim is biometric identification. To make our results more accurate, the distance was tuned as well in a way that allows the rejection of the sample if the distance calculated was found to be bigger than the indicated distance; thus making our system in this case function as a verification method rather than an identification method.

For the implementation of the SVM classification, since there were more than 2 classes (subjects), the algorithm was implemented to compare one subject versus all the other subjects and in a way such that from the saved samples a random portion of the subjects is taken as training set and another one as a testing set according to a percentage specified from the user where none of the training cases are in the test cases.

In order to measure the accuracy of the SVM classification, a confusion matrix is used. The confusion matrix calculates 4 parameters, the true positives (TP) which is the number of samples predicted to be true and are actually true, the true negatives (TN) which is the number of samples predicted to be false and are actually false, the false positives (FP) which is the number of samples predicted to be true but are actually false, and finally the false negatives (FN) which is the number of samples predicted to be false but are actually true, the sum of these 4 parameters should be equal to the number of tested samples. According to the confusion matrix, specific parameters are calculated such as sensitivity, error rate/ accuracy, specificity, and precision.

IV. Results and Discussion

A. Feature Extraction Results

The feature extraction method was applied to the 57 subjects found in the dataset mentioned above. The program functioned perfectly on group (a) and all features were extracted during an average of 10s, however in groups (b) and (c) the features weren't extracted due to lack of the diastolic peak which is an important factor in the PPG waveform [2].

B. Classification Results

After feature extraction, the classification was applied only on the 23 subjects of group (a) since in the remaining 2 groups the features weren't extracted.

1) K-NN results

In the K-NN classifier, the distance was chosen empirically to be 10000. In the first application, before tuning the distance, the method was used as identification technique; an input signal was being compared to the 23 samples and the results showed to which ID this input signal belonged. In the second application, where the algorithm worked as verification technique since the distance was tuned, the input signal was being compared to only 1 signal in the dataset; the output gave true/false answers based on the distance measured. In both verification and identification modes, the classifications were made correctly in all the trials thus giving high accuracy. Some examples of these results are shown in Table 1 below.

Table 1: K-NN classification results in identification and verification modes

Identification mode	Verification mode
Result is: 'ID #1'	d= 1.1618e+05 False
Result is: 'ID #17'	d= 6.5647e+04 False

2) SVM results

Several examples were tested on the samples of group (a). In the first example, the percentage of testing set was chosen to be 25%, so the group was divided into 18 samples for training and 5 samples for testing, 6 support vectors were found and the linear kernel function was chosen.

The results showed the classification of the 5 tested samples, and then the confusion matrix was calculated and displayed where FP, FN, and TP were all zeros whereas TN was found to be 5; this implies that the 5 testing samples were predicted to be false and are actually false. In this case, the accuracy and specificity were 100%, the error rate was zero, and the sensitivity and precision weren't calculated since FP, FN, and TP were zeros.

In the second example, the percentage of testing set was chosen to be 65%, so the group was divided into 9 samples for training and 14 samples for testing, 5 support vectors were found and similarly the linear kernel function was chosen.

The results are illustrated in Table 2, matrix C shows the classification of the 14 tested samples, and then the confusion matrix was calculated and displayed where in this case FN and TP were found to be zero whereas TN and FP were found to be 11 and 3 respectively, this implies that 11 testing samples were predicted to be false and are actually false, while 3 were predicted to be true but actually they should be false. In this case, the accuracy and the specificity were 79%, the error rate was 21%, the precision was 0 since TP was 0, and sensitivity wasn't calculated since FN, and TP were zeros.

Table 2: SVM classification results and measured parameters

Matrix C	false, false, false, true, false, true, true, false, false, false, false, false, false, false
Confusion Matrix	TN= 11 FP= 3 FN= 0 TP= 0
Accuracy	78.5714
Specificity	78.5714
Error Rate	21.4286
Precision	0
Sensitivity	NaN

C. Comparison between Classification Methods

K-NN and SVM classifiers were applied on a dataset of 23 samples based on 40 features extracted from each sample. Each classifier has many advantages and drawbacks that will be discussed in this section to choose finally the optimal classification technique to be used in the design.

The K-NN can work with linear and non-linear data since it is flexible and adjustable, it requires only 1 parameter to be modulated (k) and a distance metric to be chosen. Also, it is efficient when the training data is huge, doesn't require a training phase and doesn't demand a complicated implementation. Most importantly, the K-NN classifier has relatively high accuracy. On the other hand, it requires high memory since it accumulates approximately all the training data and may be slow if the training set is large since it calculates the distance to all the points in the training set so it is computationally expensive and is sensitive to outliers.

The SVM can also work with linear and non-linear data using the Kernel trick. However, it works with less training samples than the K-NN since it only uses the support vectors instead of all the training set, it has low rate of over-fitting and it is strong and robust in avoiding noise and error. Also, the SVM is known for giving highly accurate results. On the other hand, it is also considered computationally expensive since it requires a lot of time for the training phase and may be slow if the training set is huge; also it needs adjustment when working with more than 2 classes since it is a binary classifier and finally the SVM and its parameters are considered hard to be tuned and implemented.

To sum up, the K-NN and SVM classifiers are similar in many ways and choosing the best among them depends in the first place on the type and number of the dataset used. Since our main concern in a biometric technique is the accuracy, Table 3 is made to compare the results of the 2 algorithms in terms of accuracy and it showed that for the used dataset although the SVM classifier showed high accurate results in most cases, the K-NN classifier provided maximum accuracy in all the trials done.

Table 3: Comparison between classifier's accuracy

Number of Trials	K-NN Classifier Accuracy (%)	SVM Classifier Accuracy (%)
1	100	100
2	100	79
3	100	57
4	100	92
5	100	100
6	100	77
7	100	100
8	100	75
9	100	82
10	100	85

V. Conclusion and Future Work

A. Conclusion

As a conclusion, 2 classification methods were used to check the possibility of using the PPG signal as a biometric technique after extracting 40 features from the signal. The K-NN classifier functioned as both verification and identification modes; it doesn't require a training phase and can be implemented easily, this classifier showed the highest accurate results. On the other hand, the SVM classifier that functioned in a verification manner, requires a training phase and a harder implementation, but is robust to noise and error, this classifier showed moderate accuracy. As mentioned before, the choice of the optimal classification method depends a lot on the size and nature of the data used and the dataset used is considered small to give a final decision.

B. Future Work

There are several steps that could be done to make sure that the chosen classification technique is the ideal one for the application:

1. Find a bigger dataset to test the computational time for each of the classification methods and the memory that will be used up. Also, to have more true results in terms of precision and accuracy.
2. Implement a prototype for the system design in which the signals are taken in real-time using SpO_2 sensor.
3. Test the feature extraction and classification algorithms on the real-time extracted PPG signals.
4. Test other classification techniques such as the neural networks or fuzzy logic classification techniques.

References

[1] J. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques, 3rd ed.: Morgan Kaufmann, 2012.

[2] A. Al Sidani, B. Ibrahim, A. Cherry, M. Hajj-Hasssan, "Biometric Identification Using Photoplethysmography Signal," The Third International Conference on Electrical and Biomedical Engineering, Clean Energy and Green Computing, pp. 12-15, 2018.

[3] S. Huang, N. Cai, P. P. Pacheco, S. Narandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Pubmed Central, pp. 41-51, Jan-Feb 2018.

[4] J. P. Mueller, L. Massaron, Deep Learning For Dummies, John Wiley & Sons, May 2019.

[5] P. H. Charlton, T. Bonnici, L. Tarassenko, J. Alastruey, D. A. Clifton, R. Beale, and P. J Watkinson, "Extraction of respiratory signals from the electrocardiogram and photoplethysmogram: technical and physiological determinants," Institute of Physics and Engineering in Medicine, pp. 669–690, March 2017.

[6] K. Polat, M.Recep Bozkurt A. Reşit Kavsaoğlu, "A novel feature ranking algorithm for biometric recognition with PPG signals," Computers in Biology and Medicine, pp. 1-14, March 2014.