Final Year Project Report

User authentication from physiological signals for wearable computing

Junzhe Sun

A thesis submitted in part fulfilment of the degree of
BE. (The Internet of Things Engineering)

Supervisor: Dr. Deepu John

Beijing Dublin International College
University College Dublin

August 24, 2018

Abstract

The electrocardiogram (ECG) signal has been shown to contain relevant information for human identification. It is an emerging biometric modality. In this report, I implemented and simulated some papers' methods in MATLAB by some databases QT Database[1], MIT-BIH Arrhythmia Database[2], ECG-ID Database[3]. Chan's method and Lourenco's method are discussed in [4, 5], Recognize accuracy is 92.68% and 75.61% respectively for 41 subjects in MIT Database; 84% and 75.38% for 65 signals in ecg-id Database; 63.64% and 86.36% for 22 subjects in QT database. In addition, a method using machine-learning is also finished which is Ye's method [6], it is tested by MIT database, gives 83.33% accuracy; also, tested by QT database, accuracy is 94.44%. Both tested for 18 subjects.

Contents

1 Introduction

To tackle the risks and costs associated with personal health care, wearable biomedical sensors that monitor patient's physiological data is expected to be deployed ubiquitously soon. However, one of the fundamental problems, i.e., the issue of patient data security and user authentication, is not yet adequately addressed in the case of wearable sensors. Currently, anyone with physical access to the wearable sensor can use it to acquire his physiological data. In case, this access is unauthorized or accidental, there is no way to notify the user/health care manager about this security breach. Also, continuous authentication is a seriously consider, i.e., a person log in to a payment website in a Internet cafes, for some reason he/she need to leave a few minutes, the person close the website without log out, hacker can easily reopen the website, for convenient log in time out in some website is too long to hacker reuse the website. Without continuous authentication, server is hardly recognized the hacker.

To implement this, physiological data is a desirable choice especially ECG, which is unforgeable, and variety between different people is much bigger than variety between two different heart beat period from same person.

In this report author implement a method which using machine learning[6] to classify subjects, test it with two different database[1, 2]. Also, improve some reference code which implement two different method without machine learning[5, 4], fix some problems and use more data in QT database and test it by another two database[2, 3].

The following section will be organized as: Section 2 is Literature Review; Section 3 will discuss Implementation details and some problem during implementation period, also show some simulation result; Section 4 will do a conclusion; Improving attempt with unaccepted result and potential future work will be shown at the last section.

2 Literature Review

Biel et al.[7] It is the earliest for proved that the feasibility of using ECG to identify human beings. The authors reduce the dimension by analyzing the correlation matrix, select the feature set by experiment, and finally use multivariate analysis to classify. By using the features selected by experience, to achieve a 100% identification of the database 20, but this method lacks automatic identification, it limits the range of applications.

Irvine et al.[8] This paper introduces the use of heart rate variability (HRV) as a recognition system. The ECG signal is first preprocessing by the bandpass filter and the local maximum is determined to determine the position of the R-wave, using the minimum radius of curvature to find the beginning of the P-wave and the end of the T-wave. Wilks' Lambda method for feature selection, linear discriminant analysis for classification. The system accurately identifies all samples on 29 people and identifies 81% of the heartbeat.

Agrafioti et al.[9] This paper discusses some benefit and drawbacks of using ECG to authentication. For advantages:

- Versatility, inherent and can be collected from any living; persistent, electrocardiogram in the long term is stable;
- Unique, individual differences between the large;
- Tolerance of the attack, to avoid the forgery and replay attacks;
- Continuous certification, for monitoring Application, ECG provides continuous authentication of identity;
- Data minimization, ECG is usually used for health diagnosis and other scenes, without additional redundant signal collection can be identified.

Challenges:

- Time-dependent, ECG may be associated with physical activity or mood changes;

- Privacy effects, ECG signal leakage may cause catastrophic privacy issues.
- Collection periods, ECG collection will be much slower than other biometrics such as fingerprint or iris which is available for capturing at any time. Every heartbeat cost around one second to form so that it will cost longer waiting time especially for using long ECG period for getting features.

What's more, it discussed the significant variability among individuals is from physiological features such as heart mass, orientation, the conductivity of various areas and the activation order of the heart.

Israel et. al.[10] Author propose a more extensive set of ECG descriptors that more completely characterize the trace of a heartbeat. As a biometric, heartbeat data are difficult to disguise, reducing the likelihood of successfully applying falsified credentials into an authentication system. Also, author designed a optimum filter to reduce noise in ECG and tested features for proving ECG is invariance from individual's anxiety which can just influence heart rate.

Taylor et. al.[11] This paper aim to document the duration of fetal cardiac time intervals in uncomplicated singleton pregnancies using a novel non-invasive fetal electrocardiography (fECG) system and to demonstrate this technique's ability to acquire recordings in twin and triplet pregnancies. it can separate fetal signals were obtained in 78% of fetuses (91/116) in twins. As time goes on, the twins' body size, dietary habits and differences in personality make the difference in twins ECG become larger than that in infancy, which means that it is easier to distinguish twins by ECG.

Odinaka et. al.[12] In this paper, review most of the techniques that have been applied to the use of the electrocardiogram for biometric recognition. Summary those papers conclusion and compare some authentication result. Depend on the result, author found most of the algorithms that have been proposed for ECG-based biometrics perform well when the training and testing data come from the same session, which provide a motivation for improving the method discuss in this report.

3 Implementation details

3.1 Basic information

Before introducing some method, some basic information which is used in those function should be covered.

ECG

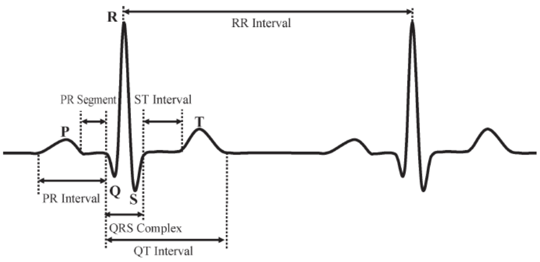

Figure 1: ECG basic information[9]

As Fig 1 P wave is beginning of each ECG segment, QRS complex will following p wave, R peak is the highest point of the segment, t wave will represent end of ECG segment. Also, some useful features as RR interval, which is the distance between two R peaks,

Wavelet transform

The wavelet transform provides the time-frequency representation[13], The wavelet transform of a given signal is to expand the signal by a wavelet function cluster, which is expressed as a series of linear combinations of wavelet functions with different scales and different time shifts, where the coefficients of each term are called wavelet coefficients. wavelet coefficients are a good feature for ECG, as ECG is non-periodic continuous signal.

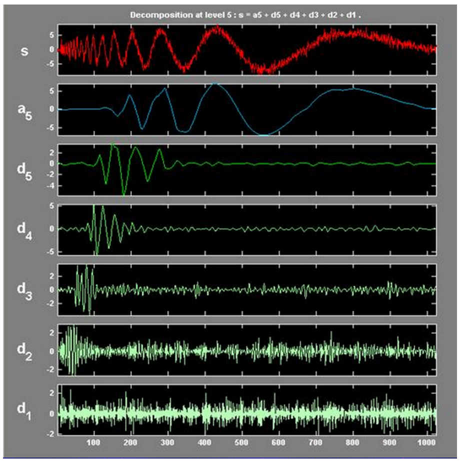

Figure 2: DWT result[13]

Database details

Databases below are used to test method in this report, all of them are two lead databases, which means there are two channel signal, one is filtered signal the other is raw signal.

1. QT Database[1]: Contains 10 fifteen-minute excerpts of two channel ECG, includes ECG which were chosen to represent a wide variety of QRS and ST-T morphologies. All samples in this database sampled at 250Hz. 105 subjects' ECG in gathered, 22 of them are used for test in this report.

2. MIT Database[2]: The MIT-BIH Arrhythmia Database was the first generally available set of standard test material, contains 48 half-hour excerpts of two-channel ECG signal. In the current work, the database is used to study the performance of the proposed subject identification algorithm in the scenario of short term ECG with arrhythmias. Sample rate for this database is 360Hz.

3. ECG-ID Database[3]: This database has experimental studies involved 90 volunteers. All signals sampled at 500 Hz

3.2 Ye's method

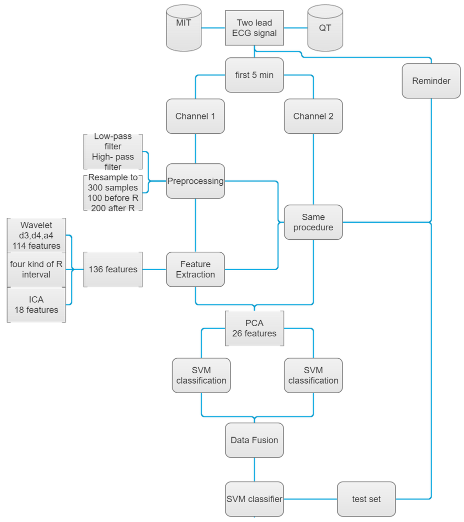

Figure 3: Flow chart for Ye's method

Overview

This method is discussed in [6], First reason for choosing this method is that this method can not only recognize ECG provided by healthy person which is also achieved in other papers, but also deal with the ECG signal with cardiac arrhythmias. Second reason for choosing this method is, after read review paper[12] and compare some accuracy from many papers, Ye's method is in the top part which gives 99.6% reorganization rates. One more reason is this paper choose ICA to do the classification which is used to distinguish twins infant[14]. When testing the method, because it requires two channels ECG signal, ECG-ID database and QT database are satisfied, so I use those databases to simulate. For first 5 min signal, detect several heart beats from position of R peak, start of p wave and end of t wave. In this report 27 heart beat segments are detected, do following procedure to each segment, first preprocessing procedure is done by using filter [0.5Hz,40Hz] to reduce noise and save ECG power as more as possible. Then resample is done by interpolation method, which is provide by MATLAB. After that, using Daubechies wavelet transform to get coefficients d3, d4 and a4 (total 114 features), four kinds of R interval: previous R interval; next R interval; local R interval (average 10 R intervals nearby); average R interval. In this step, 118 features are gotten. Next, fastica method[15] is used to detect some independent component, 18 features are gathered. To reduce dimension, make features more suitable for SVM, PCA is used by pca method which is provide by MATLAB, 26 features is selected. Using 26 features for 27 heart beat in different subject to train SVM classifier, which is provide by well-known library[16], do same procedure to channel two, those features create another classifier. Do same process to reminder of the signal, result for those signals play a role as test set. In data fusion part, if test result for each heart beat segment is not same for two lead, drop the segment to improve recognize rate.

ICA

ICA is independent component analysis, it is important to blind signal separation which is the separation of a set of source signals from a set of mixed signals. Behavior is like distinguishing some special sound in the cocktail party. ECG is a mixed signal, ICA can analysis some component to represent the whole signal. It is useful for reorganization.

$$x = A \cdot s \tag{1}$$

Each component of x is linearly related to the component of s, and both A and s are unknown, x is known and s needs to be derived from x. This process is also called blind signal separation. When using ICA to implement Ye's method, each ECG signal is divided to 18 features when using 18 subjects in total. An improving idea will be discussed in future work by finding lowest ICA features as match.

PCA

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation).[17] Idea of PCA is reduce the value which effect less to variance, so that the low dimension value keep the information as much as possible. It is used for reducing number of features so that it will be suitable for SVM classifier. The most classic approach is to select the first principal component instead of the original data for subsequent calculations, the variance of the first principal component should be the largest, because the greater the variance that the main component contains more information. In this implementation, Through the test, when only the first principal component is used to train the SVM classifier, the SVM classifier performs the best, reaching 83.33% accuracy, and when the 26 main components are trained, the classification result is reduced to 74.45%. This is the dimension curse of the SVM and will be mentioned in the next section

SVM

Support vector machines are supervised learning which means label need to be set manually. In linear mode, SVM performance well in binary classification. When classify multiple dimension features, a SVM can also use a kernel to effectively carry out non-linear classification, its input implicitly mapped to high-dimensional feature space. The kernel in this report is Gaussian radial basis function.

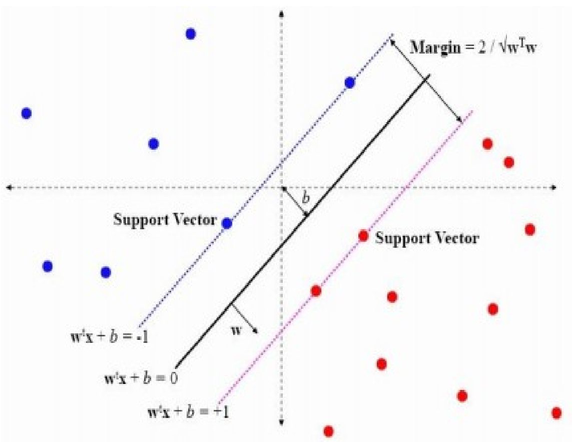

Figure 4: SVM model

As the Fig 4, optimal hyperplane is the black line in the middle, which separate different data. To improve performance, a maximum margin classifier is needed. As formula 5, under the corresponding constraints, maximize this value $1/||w||$ and create two dotted line both of which has same geometrical margin, the area between dotted line is called decision boundary. Those points on dotted line is Support Vector.

$$\max \frac{1}{||w||}, \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

Figure 5: Maximum margin classifier formula

Dimension curse is that when keep increasing dimension in features dimension without increase subjects, subjects will continue to spread out until most features in high dimension space are empty which will reduce the ability about predict. In this report, the first principal component is enough for training classifier.

In this report, a multi-class classification is used which is provided by libsvm. The idea of implementation this is doing binary classification for each label pairs, the result is considered to be a voting, in the end a point is considered to be in a class with the maximum number of votes. If the number is the same, the class appear first is chosen.[16]

Resample

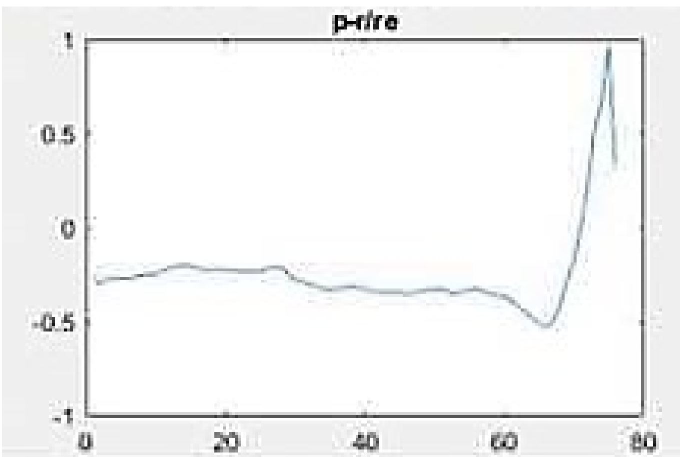

Figure 6: Resample result for one part

At the beginning, resample function provide by matlab is used to process data, however, both two-part data (before R peak and after R peak) are less than what point the method need. That means resample function need to use some interpolation method to create some points' value. As Fig 6 resample function will create a highest point which is higher than R peak, which should be the highest value of the whole signal.

After using interpolation method directly, the problem is solved. But interpolation method will lose some value at beginning or ending of the signal, I use the value next to the missing point to replace it, so that it can works well but will loss a little accuracy as Fig 7.

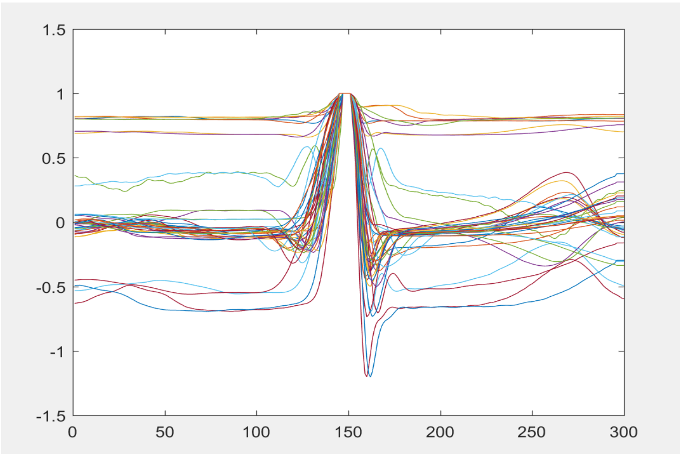

Figure 7: Result after interpolation

Simulation result

- MIT Database:

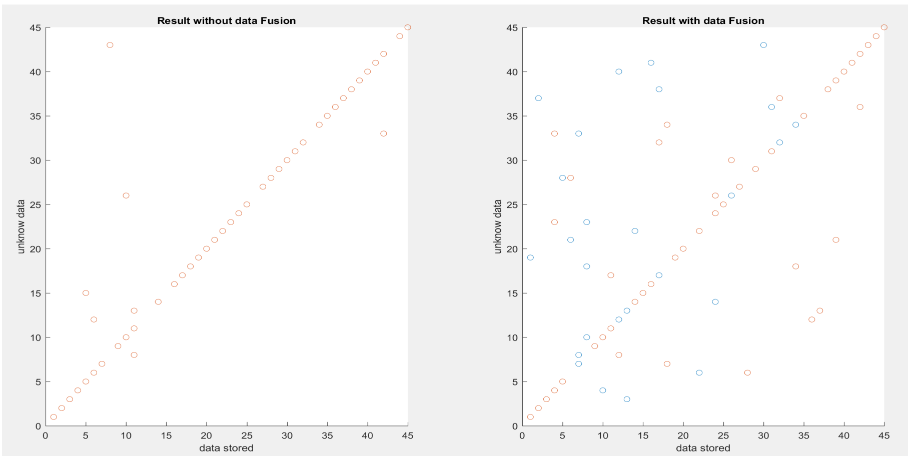

Figure 8: Result when testing MIT database

As Fig 8 shows, 45 subjects in MIT database is test. 77.78%(38/45) for simulation without data fusion. After finishing data fusion, result for two lead has distinct color, both return 83.33%(35/45) accuracy.

- QT Database:

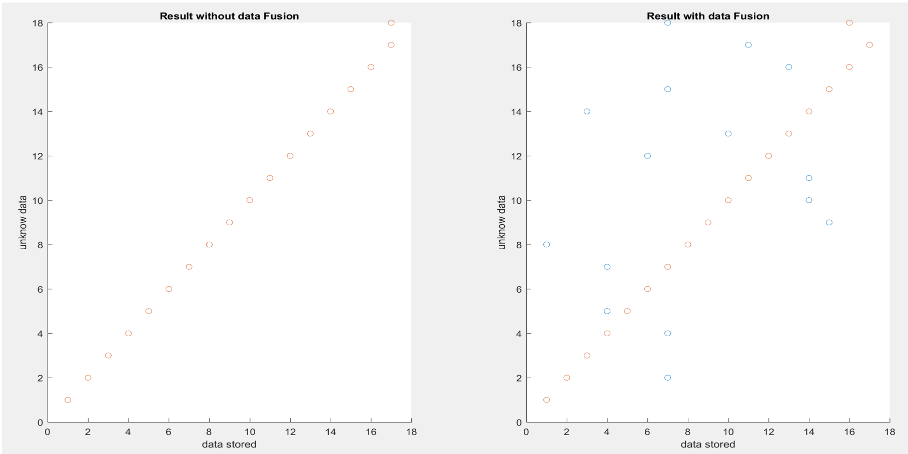

Figure 9: Result when testing QT database

In QT database, as Fig 9, test 18 subjects, 94.44%(17/18) for both of two method, without data fusion and with data fusion. However, it is obvious that in this database data fusion won't performance well, one of two lead return 16.67%(3/18). In my part, because of the original heart beat just have 27 segments, after data fusion it is reduced to 8 segments. For SVM classifier 8 features for 1 subject to train is far from enough. However, as the graph shows, the diagonal means match which has 16 point on it. That is the idle accuracy by combine result for two lead is 88.89%.

3.3 Chan's method

Overview

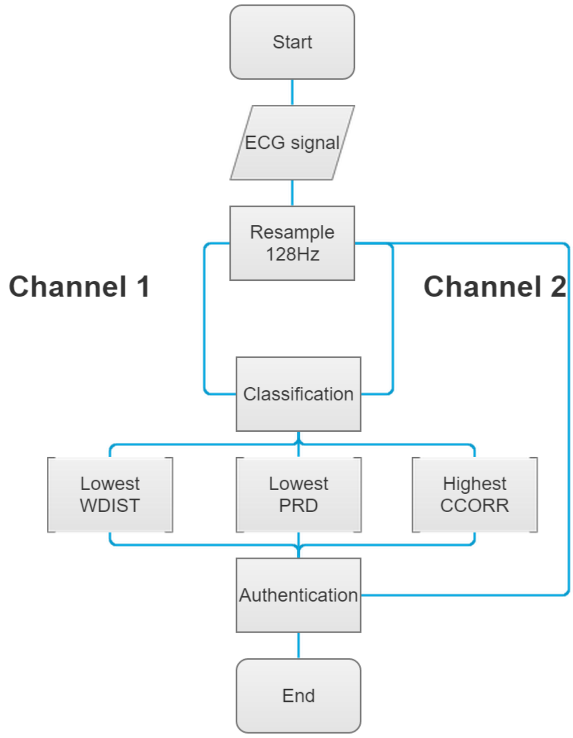

Figure 10: Flow chart for Chan's method

This method is discussed in [5], reason for choosing this paper is that, this paper provides three different method when classify subjects, main idea is using DWT which is useful. Also as a beginner, this paper is easy to understand and implement. After gather ECG signals, resample it to 128Hz. To implement this part, fining the position of R peak, 43 points on R left and 84 points on the right into a combination to get 128 point. As Fig 10, when doing classification there are three different method, all of them are calculate a distance between two signals. That means it need two signals as input, so different channel is chosen. When doing authentication, the procedure is as same as classification. When finishing this, it will show as Fig 11

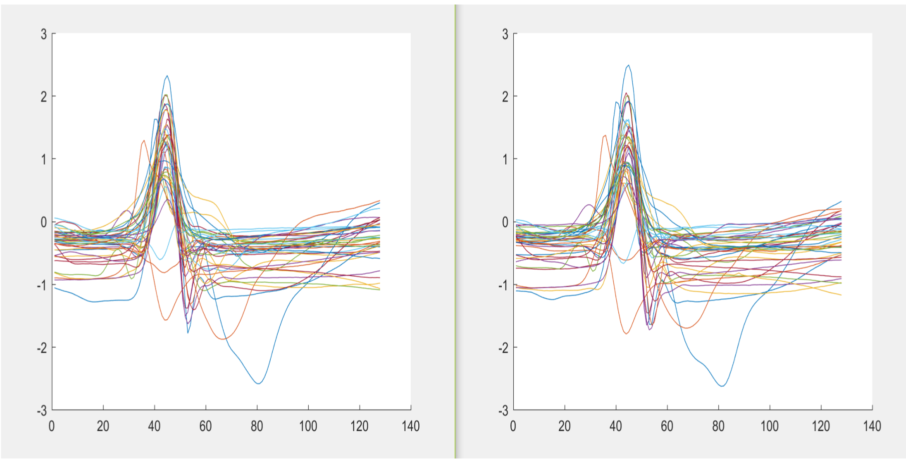

Figure 11: Model and test for Chan's method

Classification method

1. Novel Wavelet Distance measure(WDITS)

$$\mathrm{WDIST}_n = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{\left|\gamma_0^{p,q} - \gamma_n^{p,q}\right|}{\max\left(\left|\gamma_0^{p,q}\right|, \tau\right)}. \qquad (3)$$

Figure 12: Fomular for WDIST

The SAECG waveform can be comprised of multiple pulses, where the timing and shape of the pulses provide the distinguishing characteristics of the waveform[5]. In this report, calculate WDIST between two signal using the function provided by Chan. The lowest WDIST means match.

2. Percent Residual Difference(PRD)

$$\mathrm{PRD}_n = \sqrt{\frac{\sum_{i=1}^{M} \left(x_0(i) - x_n(i)\right)^2}{\sum_{j=1}^{M} \left(x_0(j) - \bar{x}_0\right)^2}} \times 100\% \qquad (1)$$

Figure 13: Fomular for PRD

PRD is a common measure to evaluate difference between signals, x_0 is unknown ECG signal, x_n is the enrolled model in which n means the number of people. The lowest value for PRD is chosen as match.

3. Correlation Coefficient(CCORR)

$$\mathrm{CCORR}_n = \frac{\sum_{i=1}^{M} (x_0(i) - \bar{x}_0)(x_n(i) - \bar{x}_n)}{\sqrt{\sum_{j=1}^{M} (x_0(j) - \bar{x}_0)^2 \sum_{k=1}^{M} (x_n(j) - \bar{x}_n)^2}}. \quad (2)$$

Figure 14: Fomular for CCORR

CCORR using linear least squares fitting between two data, the highest CCORR is match.

Result

Three different database is used to test this method, MIT DB, ECG-ID DB and QT DB, the subjects of those database are 41,75,22 respectively. Result is shown as Table 1.

Database \ Method	WDIST	PRD	CCORR
MIT DB	92.68%	82.93%	75.61%
ECG-ID DB	84%	28%	9.3%
QT DB	100%	95.45%	95.45%

Table 1: Result for Chan's method

There is some strange result, in ECG-ID database, graph result is shown below, in Fig 16 because of that the first two persons has 20 different ECG signal, if the method identifies the heartbeat of the first or second person as any of his own 20 heartbeat data, it indicates that the identification of the method is inaccurate but also acceptable. So, the ideal accuracy is 84%, 45.67% and 61.34% respectively.

As result for QT DB, 100% is too ideal, this because the dataset is too small. But combine with some evidence in front of, can still prove that this method among the best of the three method.

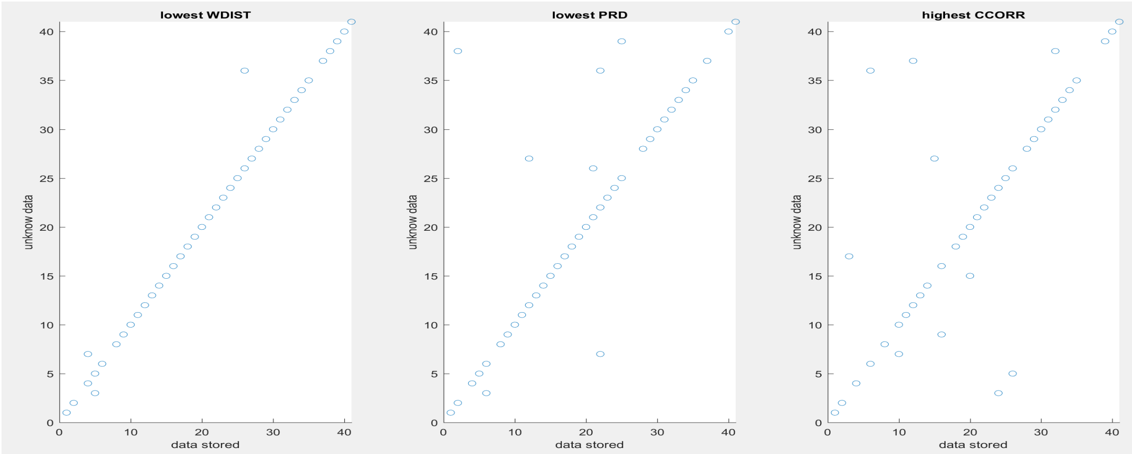

Figure 15: MIT Result

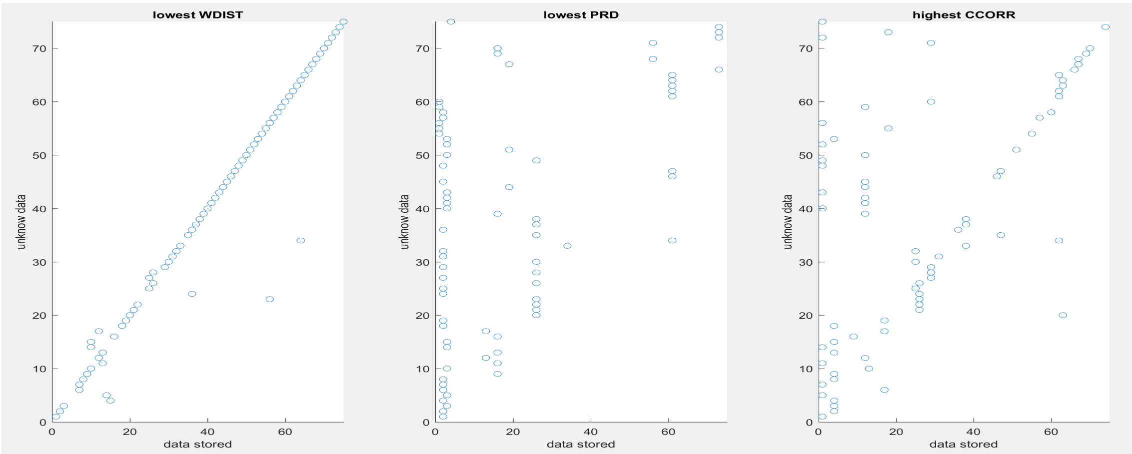

Figure 16: ECG-ID Result

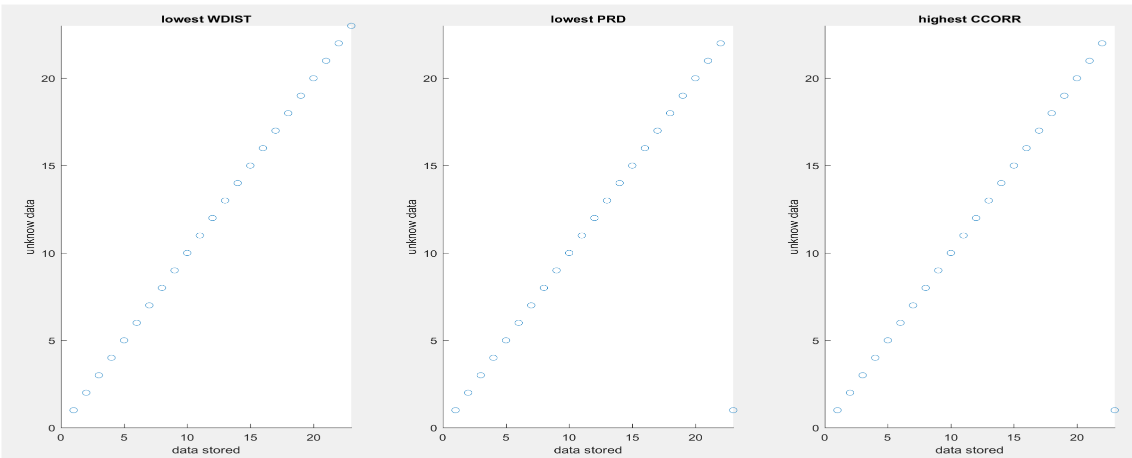

Figure 17: QT Result

3.4 Lourenco's method

Overview

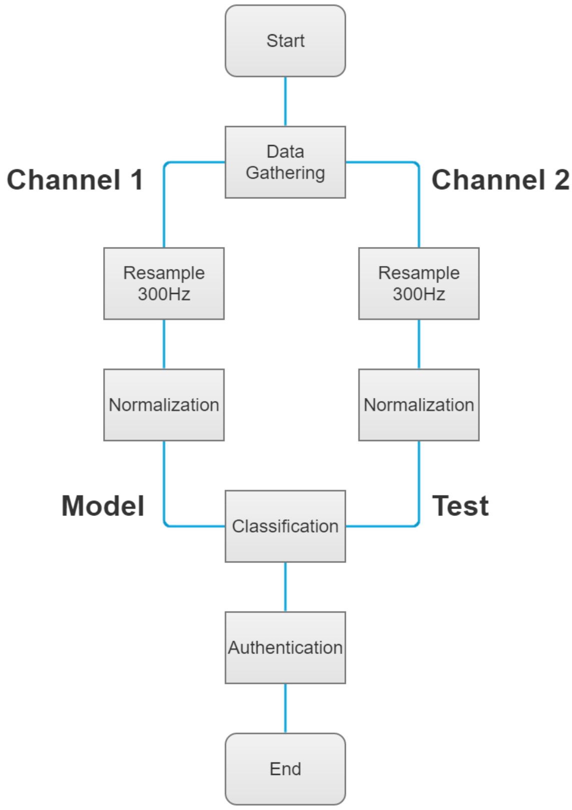

Figure 18: Flow chart for Lourenco's method

This method is discussed in [4], because the Euclidean distance is a common way in several detecting areas and the complexity is low, so this method is chosen. After get data from database, it detects position for R peak, start of p wave and end of t wave to rebuild a whole ECG signal. In resample period, 150 samples for both segments which is from start of t wave to R peak and R peak to end of t wave. After that, a strict normalization is used which will be discussed below. Then Euclidean distance is used to classification different subjects. The last step is authentication which step a threshold to result of Euclidean distance, FRR (False Rejection Rate) will be shown below.

Normalization

Because Euclidean distance is an uncomplicated way so it need some assist by preprocessing data before, so that the accuracy can be improved. In this method, ECG signal is non-periodic, which means R peak position in different segment can be different, that will affect a lot when calculate Euclidean distance. To avoid this, normalization is important. It can move features potion to same time line and limit feature's range in the same area.

- Time Normalization: To implement this function, using annotations file, which have the position for each feature, to find R peak's position. Then using interpolation(resample) to make this segment to 150 point, do same thing to segment after R peak and splice them to 300 points signal. Then each feature is at same position. Result as Fig 19.

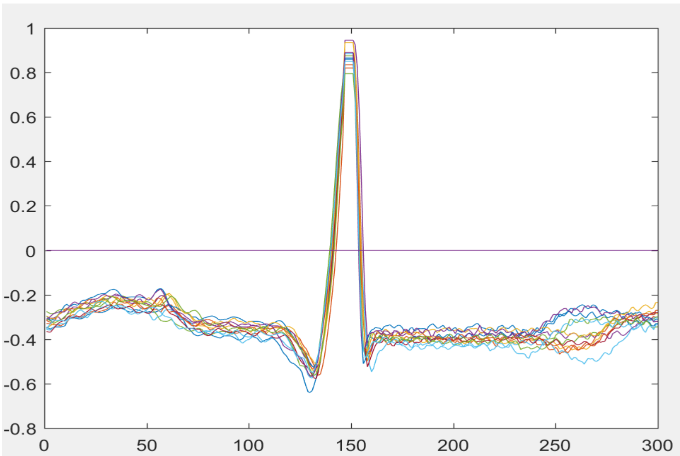

Figure 19: After time normalization

- Amplitude Normalization: After time normalization, R peaks are at same position but has different values, amplitude normalize is designed to fix it. First calculate average for all R peaks as norm factor; then divide all samples by the factor; finally get average from all heart beat segments provide by same person. Doing same procedure to reminder subject will finished amplitude normalization. Result as Fig 21.

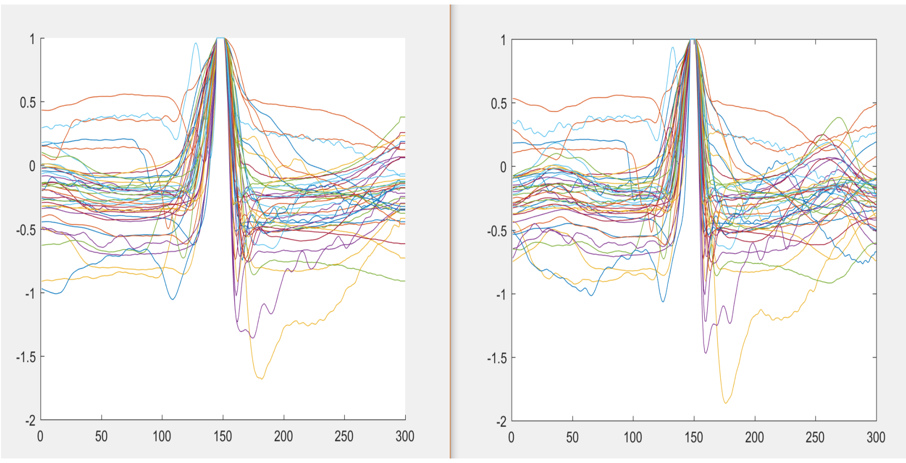

Figure 20: Model after normalization

Euclidean distance

Euclidean distance is the "ordinary" straight-line distance between two points in Euclidean space.

$$d = \sqrt{(x_0(i) - x_n(i))^2}$$

Figure 21: Formula for Euclidean distance

Euclidean distance in the application of digital image processing is very broad, especially for the skeleton of the image extraction. In this report, as the formula upper, calculate Euclidean distance for every point and get the sum. Any two signals have minimum Euclidean distance means match.

Result

Database	Result
MIT DB	75.61%
ECG-ID DB	75.38%
QT DB	63.64%

- MIT DB: 41 subjects are used to test, result is shown as Fig 22, in which the dark diagonal is match.

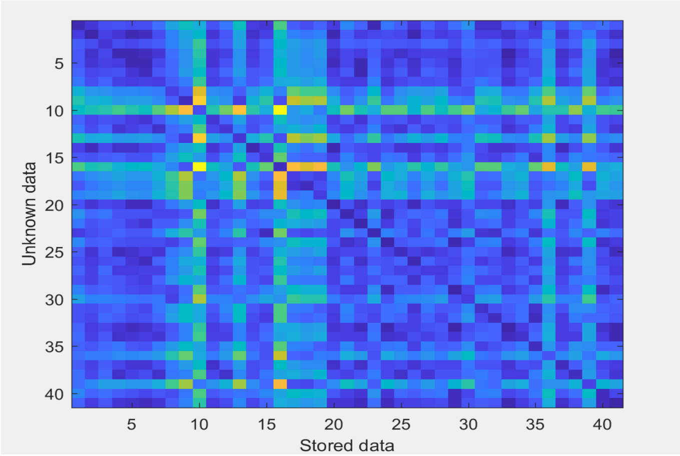

Figure 22: Result for MIT DB

- ECG-ID DB: This database is used for testing authentication function, result is in next part.
- QT DB: 22 subjects are tested, as Fig 9, also diagonal means match.

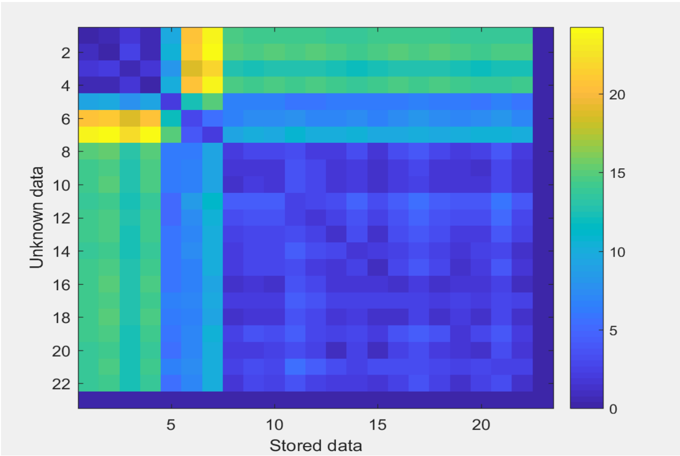

Figure 23: Result for QT DB

Authentication Result

In this part, an authentication simulation is discussed by using three different databases. As the threshold increase, the FRR (False Rejection Rate) rise.

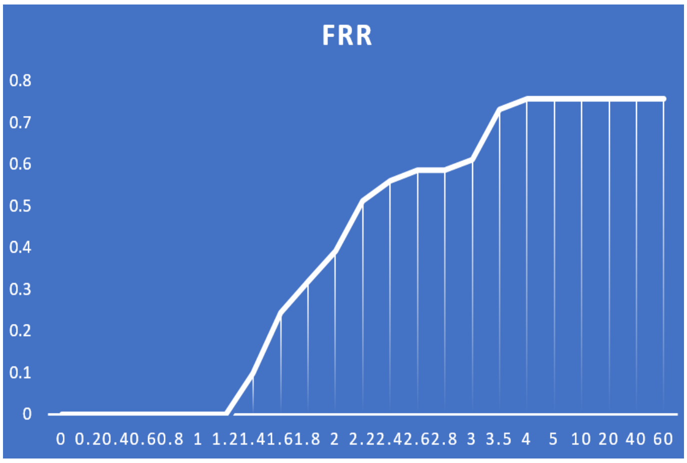

Figure 24: Result for MIT DB

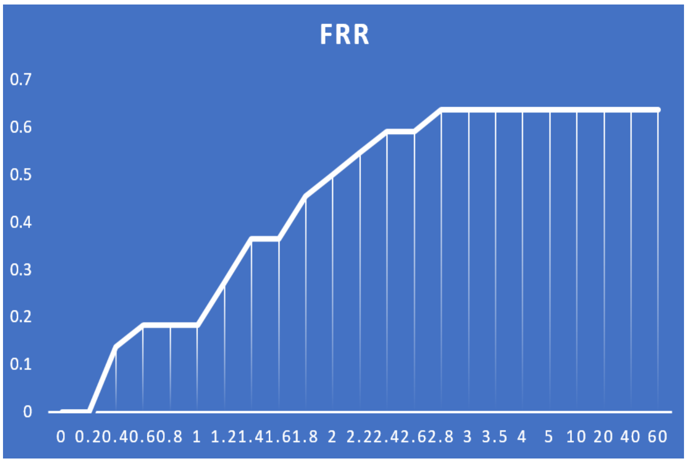

Figure 25: Result for QT DB

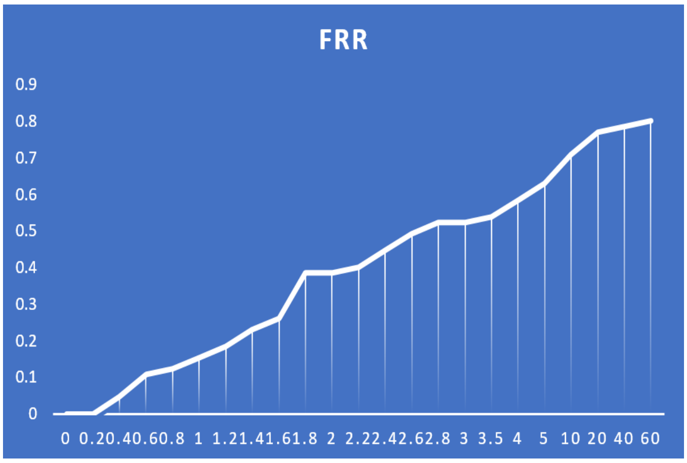

Figure 26: Result for ECG-ID DB

4 Conclusion

After simulate several methods, the most stable method is using SVM for classification, but its algorithm is the most complex, Euclidean distance classification is the simplest, but the noise tolerance is low, fault tolerance is poor. In the Chan's method, WDIST always shows the best performance.

5 Improving attempt and Future Work

5.1 Improving attempt

Motivation for improving

When testing methods, noisy signal testing the classification model which is created by the same time period as testing signal is not enough, since the test signal is substantially the same as the main part of the model signal, such as shape, R-wave intervals, and so on. The results will be very ideal but not accurate. I try to replace a test method by dividing a complete signal into two parts, the first part used to create a classification model or training classifier, the latter part used for the test. Unfortunately, after testing, the accuracy of several of the methods mentioned above is far less than expected, and the new results do not exceed the accuracy around 30% for each of those method above. This makes me start trying to improve these methods.

Improve Lourenco's method

Lourenco's method just calculate Euclidean distance to classify. When using original test method, the accuracy is acceptable. However, after updating the test method, the accuracy is reduced to around 25%. After analysis the model and testing data, no correlation was observed by naked eye. That means simple Euclidean distance calculation can not be qualified for classification work. So some improving method is used.

- Doing SVM classify with 4 features from ICA, 41 features from wavelet transform, SVM classifier is much reliable than Euclidean distance when dealing with complex method as ECG classification.

- Doing same procedure as Chan's method, PRD, CCORR, WDIST may perform better than Euclidean distance.

- Getting the first principle component from PCA which is designed as the PCA part in Ye's implementation.

However, those method give result even worth than Euclidean distance, as 7%, 17.68%, 6.8%, 8.1% and 1.4% respectively for 141 subjects. The problem is because of the ICA cannot get 141 features, so I just use 4 features instead, it need to be replaced by other feature detect method. The figure below give the evidence about why Euclidean distance is not suitable.

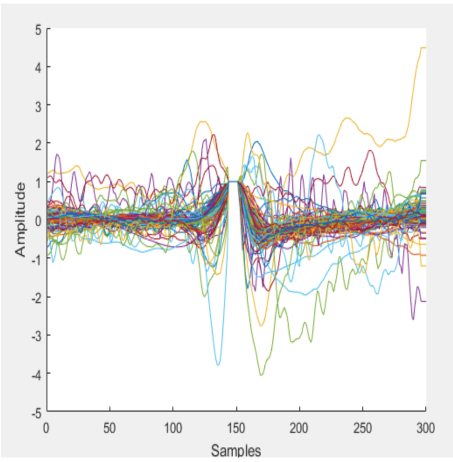

Figure 27: Model for first half signal

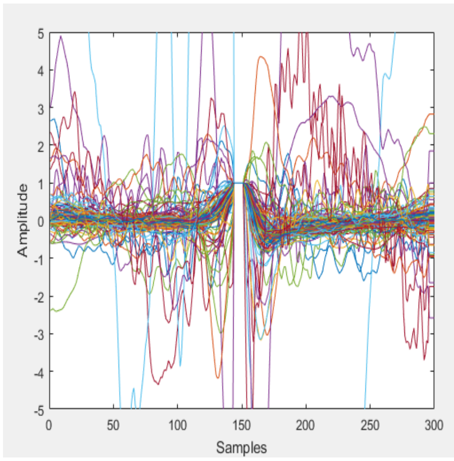

Figure 28: Test data for second half signal

5.2 Future Work

In future, ECG can be detected on figures or chest by some hardware, and gather those data to test if those method really works for authentication person. Also, because some library also has c version, develop an embedded system is also a good idea.

Bibliography

[1] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg," 1997, pp. 673–676.

[2] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[3] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[4] H. Louren, Andr?and Silva and A. Fred, "Unveiling the biometric potential of finger-based ecg signals," *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–8, 2011.

[5] A. D. C. Chan, M. M. Hamdy, A. Badre, and V. Badee, "Wavelet distance measure for person identification using electrocardiograms," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 2, pp. 248–253, 2008.

[6] C. Ye, M. T. Coimbra, and B. V. K. V. Kumar, "Investigation of human identification using two-lead electrocardiogram (ecg) signals," 2010, pp. 1–8.

[7] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *Instrumentation & Measurement IEEE Transactions on*, vol. 50, no. 3, pp. 808–812, 2001.

[8] N. Akhter, S. Tharewal, V. Kale, A. Bhalerao, and K. V. Kale, *Heart Based Biometrics and use of Heart Rate Variability in Human Identification Systems*, 2015.

[9] F. Agrafioti, F. M. Bui, and D. Hatzinakos, "Medical biometrics in mobile health monitoring," *Security & Communication Networks*, vol. 4, no. 5, p. 525–539, 2011.

[10] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "Ecg to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[11] M. J. Taylor, M. J. Smith, M. Thomas, A. R. Green, F. Cheng, S. Osekuafful, L. Y. Wee, N. M. Fisk, and H. M. Gardiner, "Non-invasive fetal electrocardiography in singleton and multiple pregnancies." *Bjog An International Journal of Obstetrics & Gynaecology*, vol. 110, no. 7, p. 668–678, 2003.

[12] I. Odinaka, P. H. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh, "Ecg biometric recognition: A comparative analysis," *IEEE Transactions on Information Forensics & Security*, vol. 7, no. 6, pp. 1812–1824, 2012.

[13] R. Polikar, "The wavelet tutorial," 1996. [Online]. Available: http://users.rowan.edu/~polikar/WAVELETS/WTpart1.html

[14] M. Keralapura, M. Pourfathi, and B. Sirkecimergen, "Impact of contrast functions in fast-ica on twin ecg separation," *Iaeng International Journal of Computer Science*, vol. 38, no. 1, 2011.

[15] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

[16] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[17] Wikipedia, "Principal component analysis." [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis