



Electrocardiogram for Biometrics by using Adaptive Multilayer Generalized Learning Vector Quantization (AMGLVQ): Integrating Feature Extraction and Classification

Elly Matul Imah, Wisnu Jatmiko, and T. Basaruddin

‡Faculty of Computer Science Universitas Indonesia Depok, Indonesia

e-mail: elly.matul01@ui.ac.id e-mail: wisnuj@cs.ui.ac.id

Submitted: July 3, 2013

Accepted: Oct. 28, 2013

Published: Dec. 16, 2013

Abstract- Electrocardiogram (ECG) signal for human identity recognition is a new area on biometrics research. The ECG is a vital signal of human body, unique, robustness to attack, universality and permanence, difference to others traditional biometrics technic. This study also proposes Adaptive Multilayer Generalized Learning Vector Quantization (AMGLVQ), that integrating feature extraction and classification method. The experiments shown that AMGLVQ can improve the accuracy of classification better than SVM or back-propagation NN and also able to handle some problems of heartbeat classification: imbalanced data set, inconsistency between feature extraction and classification and detecting unknown data on testing phase.

Index terms: ECG biometrics, AMGLVQ, SVM, back-propagation-NN, feature extraction, classification, vector quantization.

I. INTRODUCTION

Biometric recognition increasingly popular and it has been extensively used for law enforcement. Biometric modality is characteristics of human body that are unique for every individual. Traditional biometrics such as face, fingerprint or iris have been used for technologies verification and system identification [1], [2]. Next generation of biometric technologies are have characteristic that robust from attacks. Characteristics of internal human body have been investigated such as DNA biometric, cognitive biometric, medical biometrics. Medical biometrics are new biometrics modalities that encompasses signal which are used in clinical diagnostics and it has been actively investigated in last decade because it robust, replay, obfuscation attacks and ensure low computational effort. The ECG signal is one of medical biometric.

Usually ECG use to gain insight on heart conditions, however in biometrics perspective it has been known that ECG has sufficient detail for identification. A part of that ECG biometric modality are universality, permanence, unique, robust to attack, liveness detection, continuous authentication, and data minimization. Some research in ECG biometric are [3–6], almost all of them used independent feature extraction and independent classification. Independent feature extraction and classification need much more time consuming and cause the inconsistency between feature extraction and classification. This study proposed a new model learning that integrating feature extraction and classification. This learning model is based on learning vector quantization and principal component, which is named Adaptive Multilayer Generalized Learning Vector Quantization (AMGLVQ). AMGLVQ is Competitive Based Learning, part artificial neural network (ANN). ANN still exist until now days, many researcher develop and modified many algorithm of ANN, such as [7–11].

The Learning Vector Quantization (LVQ) is prototype based learning. The prototypes are determined in a training process from training dataset and can be interpreted in a straightforward way as they capture essential features of the data in the very same space. LVQ is a simple way algorithm that can be applied to multiclass problem and the complexity of LVQ can be control during training phase according to the specific needs. Researchers proposed the combination of LVQ with other prototype-based learning schemes like SOM or Neural Gas to include neighborhood cooperation into the learning process [12–16]. Also, some techniques to realize fuzzy classification based on the general LVQ approach were proposed the last years [17], [18]. Recent research in vector quantization learning algorithm area are [15], [19], it is show that this learning area is interesting.

In previous study AMGLVQ has been done to Arrhythmias heart beat classification [20]. AMGLVQ was proposed to handle some problems in classification. The problems that have been handled by AMGLVQ are, imbalanced data set, inconsistency between feature extraction and classification and detecting unknown data on testing phase. Imbalanced data set is a classification problem especially for real-life classification problem, because almost learning algorithm is design for balanced data set. Imbalanced data set is a set of training data witch one of the classes is represented by small number than the other class. Many researchers that consent to handle imbalanced data set. There are two approaches for dealing this problem, with internal and external approach. Internal approach is handling imbalanced data set by using algorithm modification [21], [22]. External approach is handling it by using resampling: under sampling and oversampling [21], [23]. Recently some researchers use combining of internal and external approaches [24]. In our study, AMGLVQ also develop to handle the imbalanced data set by using combining external approach and internal approach.

In this study AMGLVQ is used to ECG biometric recognition and integrated with automatic heartbeats classification. ECG biometric recognition based on sinus rhythm of human heart beat. It is difficult to classify if the individual has abnormal heartbeat. So in this study we also classify type of arrhythmias heartbeat. Heart beat classification in an ECG signal is not trivial problem, because heartbeat data has a characteristics as follows: 1) a strong class imbalanced [25], 2) nonlinear separable. Because of that we used the MIT-BIH database to evaluate our proposed method. Standard automatic classifiers generally perform poorly because they are designed for balance data. A set of training data is said to be imbalanced if one of the classes is represented by a small number of cases compare to the other classes. And with the same data based, we used normal heart beat of 10 patient to classify the human identification.

This paper organized as follows. Section II describes system overview. The proposed method, multi codebooks adaptive multilayer generalized learning vector quantization is described in section III. Section IV is Data Environments that included pre-processing technique to reduce the baseline noise with cubic spline and then extract signal in beat basis. Section V contain the experiment result and discuss. The conclusions of this paper is describes in section VI.

II. SYSTEM OVERVIEW

This study is a part of a bigger research in our laboratory to develop Smart Cardio Device. The overall system of Smart Cardio Device is shown at figure 1. The device were divided into three main components, the first component is a circuit used to multiply and filter ECG signal. It was designed to make the ECG signals high, so that those signals could be executed in Beagle board. Second component is used to run classification process. The third component is mobile device as interfacing of our system.

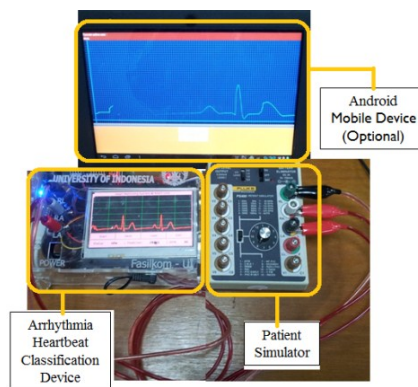


Figure 1. Overall System of Smart Cardio Device and human identification

The detail of system overview is shown at figure 2. In this system we built ECG module than executed the signal on Beagle board. We used Android mobile device for interfacing the signal and LCD touch-screen as user input interaction. To connecting the signal from ECG module to mobile device we used wireless. From Android mobile device, user can show the ECG signal and enter the input data to interaction to the system. From overall system of smart cardio device, this study is focus on classification algorithm. This device is not only able to detect the arrhythmias heartbeat but also able to recognize person who has the ECG signals.

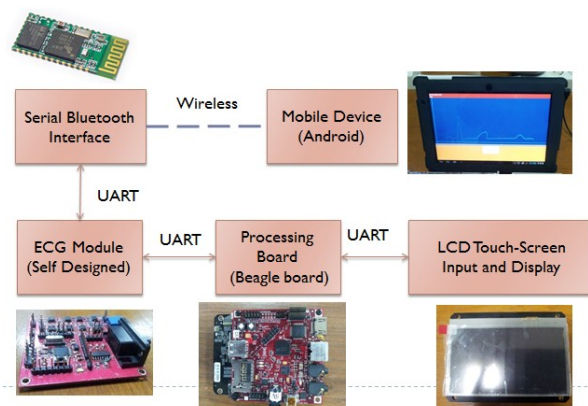


Figure 2. System overview of Smart Cardio Device

III. ADAPTIVE MULTILAYER GENERALIZED LEARNING VECTOR QUANTIZATION (AMGLVQ)

Feature extraction is a dimensionality reduction techniques and important step in classification. In application data, which the data is multidimensional and presented by a number of features, before classification process, the dimensionality of some feature space is often tried to be reduced. “Curse of dimensionality” is a condition that amount of computations required for pattern recognition and amount of data required for training system grows exponentially with the increase of dimensionality of the feature vectors [26].

In the conventional classification system, the process of feature extraction and classification done separately, this process took a lot of experiment to determine the best feature and to determine the optimum parameter of classification algorithm. Block diagram conventional system classification and new approach classification system can be seen on figure 3 and figure 4. This study proposes AMGLVQ, a new method that integrating feature extraction and classification, block diagram of our system can be seen on figure 5. Beside of that AMGLVQ also develop to handling imbalanced data set.

AMGLVQ is based on Artificial Neural Network (ANN). ANN in general designed for well-balanced data and does not work properly on imbalanced data sets. By modified GLVQ to AMGLVQ we would like handling the imbalanced data. AMGLVQ has adapted nodes in hidden layer. The adaptive layer have adaptive the number of nodes, are caused data selection that will be used on training phase. From all data sets, we would not use all the data, but we selected the best ones. We selected the data that has characteristic. To evaluate the data that would be process in training phase we used miss classification error of AMGLVQ.

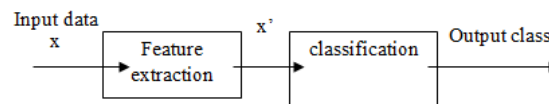


Figure 3. Block diagram of conventional classification system

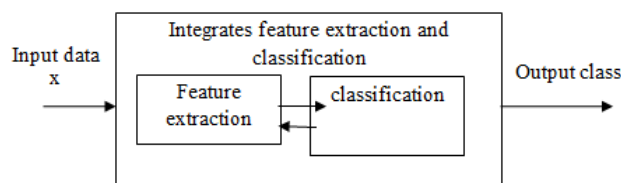


Figure 4. Block diagram of integrated feature extraction and classification

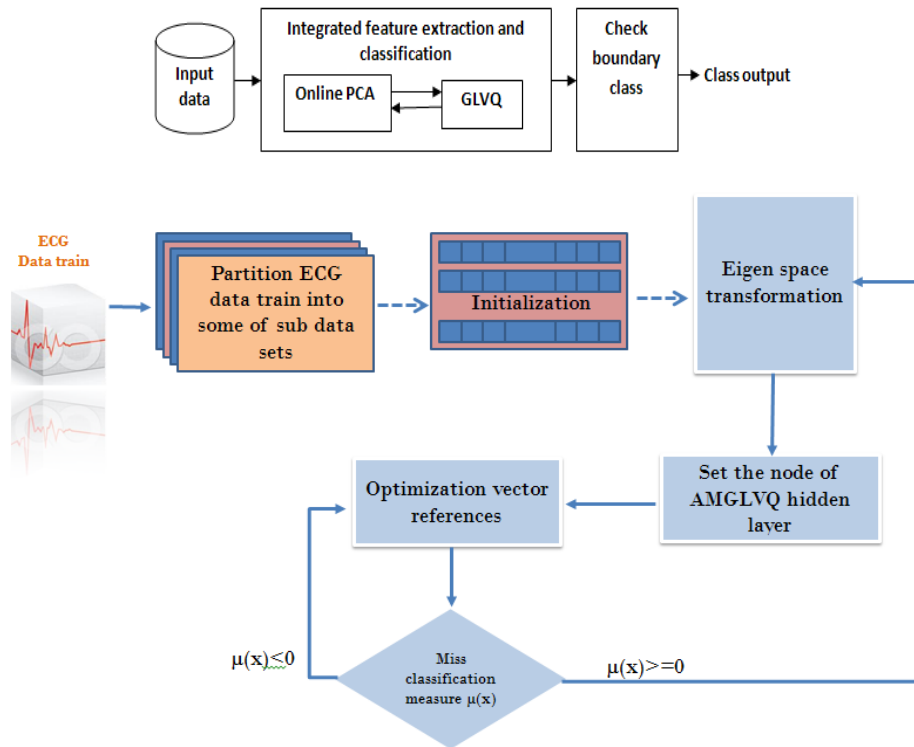


Figure 5. Block diagram of AMGLVQ

AMGLVQ is algorithm that integrated Principal Component Analysis (PCA) [23] and Generalized Learning Vector Quantization (GLVQ) [2]. It has two phases, first phase is Optimization Integrated Feature extraction-classification and also handling imbalanced data set, then second phase is finding boundary class for incremental class to solve the unclassifiable heartbeat. This algorithm has three layers on architecture neural network, first layer is input layer, then second layer is hidden layer that has adaptive number of neuron and third layer is output layer, for the detail see on figure 6.

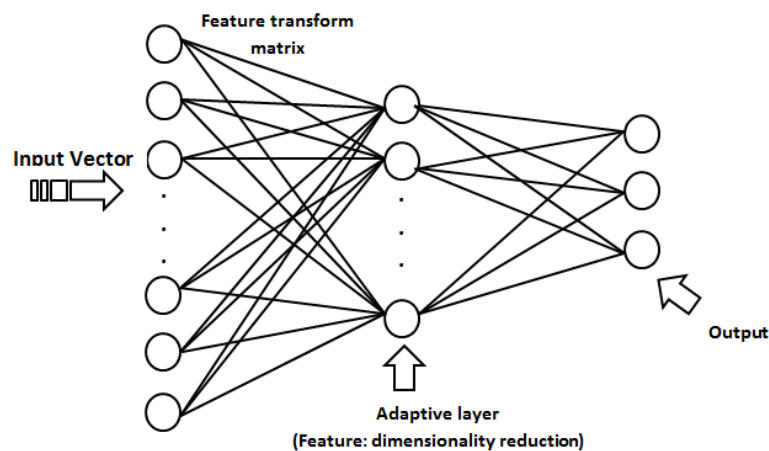


Figure 6. AMGLVQ Architecture

The architecture AMGLVQ show that feature reduction step is on the hidden layer. The number of nodes in hidden layer depends on number of feature that has been reduced by this algorithm. From input layer, full feature of data will be transformed to the feature space and then reduces the dimension of data. The adaptive of AMGLVQ is caused the number of node in hidden layer is adaptive depend on the number of feature and the number of feature depend on matrix transformation that is composed by selected data.

A. Optimization of Integrating Feature Extraction-Classification and Handling Imbalanced Data Set

First phase is the optimization of integrating feature extraction-classification and handling imbalanced data. First step on this phase is data selection to build the matrix which is used to find the principal component. The Matrix will be updated with adding the input vector that caused the miss classification error increasing, it is selecting data train process that able to handling imbalanced data set also.

Imbalanced data is a condition in which a class has a very small amount of data compared to other classes or the amount of data between classes has much different ratio. Imbalance of data is one of the current problems in the field of machine learning is applied to real problems. In classification problems, the standard classification algorithms often ignore data that has a very small amount, so the class with a very small data set will be hard to recognize. In practical applications the data rate is said to be very much imbalance is 1:100, 1:1000, 1:10000 or even could be more than that.

There are two approaches for dealing this problem, with internal and external approach. Internal approach is handling imbalanced data set by using algorithm approach. External approach is handling it by using resampling: under sampling and oversampling. Recently some researchers use combining of internal and external approach.

In this study, we developed an algorithm that integrates AMGLVQ as feature extraction and solve imbalanced data sets with a level approach to the incorporation of algorithms and data level. AMGLVQ besides integrating the feature extraction and classification, as well as handling the problems of imbalanced data sets. Procedures performed to address the imbalanced data sets, namely the under-sampling where the under-sampling done in the learning process takes place and performed by algorithms AMGLVQ. The process of under-sampling was done by the data selection, training data that exist in the data bank will be selected which are representative to be included in the learning process or ignored. The

selection process is done by the data evaluator miss classification error. For simple description see on figure 7, it is simplifying the figure 5.

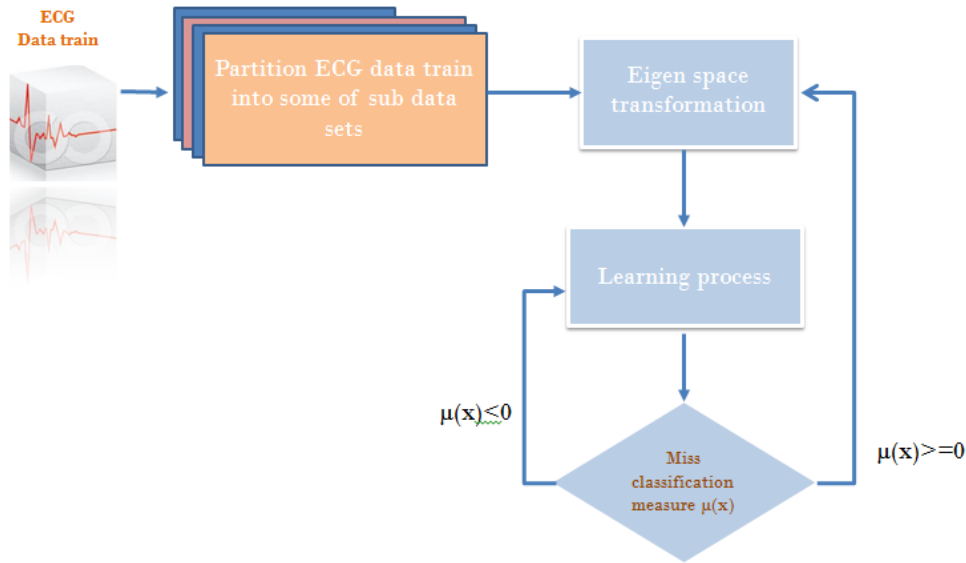


Figure 7. Block diagram process for imbalanced data set handling

If $X_1, X_2, \dots, X_N \in \mathbb{R}^N$ and $Y_1, Y_2, \dots, Y_M \in \mathbb{R}^N$ is data train from differences group, X_N from t -th group training data, and Y_N is $(t+1)$ -th group training data. Then we define Z as new matrix that is used to find the principal component, $Z = X \cup Y$, so we can write Z as follow:

$$Z = \{X_1, X_2, \dots, X_N, Y_1, Y_2, \dots, Y_M\} \in \mathbb{R}^N \quad (1)$$

Covariant Matrix Z (C_z) can be define as eq. 2.

$$C_z = \frac{1}{N+M} \sum_{j=1}^{N+M} Z_j Z_j^T$$

$$C_z = \frac{1}{N+M} \left(\sum_{j=1}^N X_j X_j^T + \sum_{j=1}^M Y_j Y_j^T \right) \quad (2)$$

We can find the principal component of Z with solve the eigenvalue and eigenvector problem. If γ is eigent value of C_z and u is eigenvector of C_z , so we can denote the eigenvalue problem as follow:

$$\gamma u = C_z u \quad (3)$$

$$\begin{aligned}
 u &= \frac{1}{(N+M)\gamma} \sum_{j=1}^{N+M} (Z_j u) \cdot Z_j \\
 &= \sum_{j=1}^{N+M} \xi_j Z_j = \sum_{j=1}^N \alpha_j X_j + \sum_{j=1}^M \omega_j Y_j
 \end{aligned} \quad (4)$$

If u is spanned by $Z = X \cup Y$ and number element of Z is k , where k is $(M+N)$, than:

$$\gamma Z_k u = Z_k C_z u \quad (5)$$

or

$$\gamma \left\{ Z_k \sum_{j=1}^{N+M} \xi_j Z_j \right\} = \frac{1}{N+M} \left\{ Z_k \left(\sum_{j=1}^{N+M} \xi_j Z_j Z_j^T \right) \left(\sum_{j=1}^{N+M} \xi_j Z_j \right) \right\} \quad (6)$$

We can write the left side of 6 as follow, so we get 7.

$$\begin{aligned}
 \gamma \left\{ Z_k \sum_{j=1}^{N+M} \xi_j Z_j \right\} &= \gamma \left\{ \left(\sum_{j=1}^N \alpha_j (Z_k X_j) \right) + \left(\sum_{j=1}^M \omega_j (Z_k Y_j) \right) \right\} \\
 &= \gamma \left\{ \begin{bmatrix} Z_1 X_1 & \dots & Z_1 X_N & Z_1 Y_1 & \dots & Z_1 Y_M \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{(N+M)} X_1 & \dots & Z_{(N+M)} X_N & Z_{(N+M)} Y_1 & \dots & Z_{(N+M)} Y_M \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \\ \omega_1 \\ \vdots \\ \omega_M \end{bmatrix} \right\} \\
 &= \gamma \left\{ \begin{bmatrix} X_1 X_1 & \dots & X_1 X_N & X_1 Y_1 & \dots & X_1 Y_M \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N X_1 & \dots & X_N X_N & X_N Y_1 & \dots & X_N Y_M \\ Y_1 X_1 & \dots & Y_1 X_N & Y_1 Y_1 & \dots & Y_1 Y_M \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_M X_1 & \dots & Y_M X_N & Y_M Y_1 & \dots & Y_M Y_M \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \\ \omega_1 \\ \vdots \\ \omega_M \end{bmatrix} \right\} \\
 &= \gamma \left\{ \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix} \begin{bmatrix} \alpha \\ \omega \end{bmatrix} \right\} \quad (7)
 \end{aligned}$$

And we can write down the right side of 6 as follow so we get 8.

$$\begin{aligned}
 \frac{1}{N+M} \left\{ Z_k \left(\sum_{j=1}^{N+M} \xi_j Z_j Z_j^T \right) \left(\sum_{j=1}^{N+M} \xi_j Z_j \right) \right\} &= \frac{1}{N+M} \left\{ \begin{bmatrix} XX^T XX^T + XY^T YX^T & XX^T XY^T + XY^T YY^T \\ YX^T XY^T + YY^T YX^T & YX^T XY^T + YY^T YY^T \end{bmatrix} \begin{bmatrix} \alpha \\ \omega \end{bmatrix} \right\} \\
 &= \frac{1}{N+M} \left\{ \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix} \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix} \begin{bmatrix} \alpha \\ \omega \end{bmatrix} \right\} \quad (8)
 \end{aligned}$$

note: $YX^T = (XY^T)^T$

So we can write 3 as 6 by substituted 7 and 8.

$$\gamma \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix} \xi = \frac{1}{N+M} \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix}^2 \xi$$

$$(N + M))\gamma\xi = \begin{bmatrix} XX^T & XY^T \\ YX^T & YY^T \end{bmatrix} \xi \quad (9)$$

Solve eq.7 to find the principal component of Z. We find T as principal component matrix that be used to transform data from data space to eigen-space. For the illustration see on figure 8.

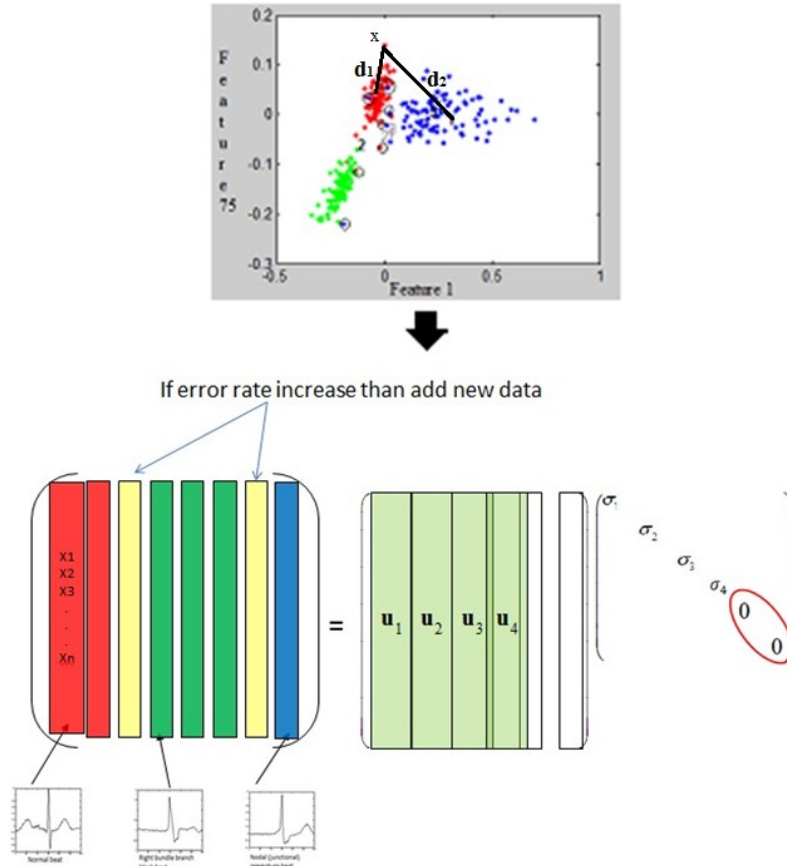


Figure 8. Illustration Data Selection and Transformation Matrix Update

Second step is optimization transformation matrix T and references vector of learning process. Let x is input data, w_1 the nearest reference vector that belongs to the same class of x, and likewise let w_2 be the nearest reference vector that belongs to a different class from x. Let us consider the relative distance $\mu(x)$ define as follows:

$$\mu(x) = \frac{d_1 - d_2}{d_1 + d_2} \quad (10)$$

Where d_1 , d_2 are the distance of x from w_1 and w_2 respectively, $\mu(x) \in [-1, +1]$ and if $\mu(x)$ is negative, x is classified correctly; otherwise, x is classified incorrectly. In order to improve error rates, $\mu(x)$ should decrease for all input vectors. $\mu(x)$ also defines as misclassification error.

AMGLVQ handling imbalanced data set by using under-sampling with selecting data. The evaluation measurement to selecting the relevant data is misclassification error, it is defined on equation 10. In previous study, data selection evaluation has used cost function [20] and it is not stable. If the misclassification error of the data ≥ 0 , it will be selected and added to the covariance matrix formation, because if the data has misclassification error ≥ 0 it is mean that the data is boundary of class that has the extreme and important inclusion in the learning process. For the illustration of d_1 , d_2 see on figure 8.

Thus, a criterion for learning is formulated as minimizing cost function S defined in eq.11.

$$S = \sum_{i=0}^N f(\mu(x_i)) \quad (11)$$

Where N is the number of input vector for training, and $f(\mu)$ is a monotonically increasing function, e.g. the logistic function $f(\mu) = 1/(1 + \exp(-\mu))$. For sigmoid functions, the classifier only learns from the training samples close to the decision boundary which carry most information. This algorithm adapted from A. Sato GLVQ. The GLVQ algorithm is in term of stochastic gradient descent. Input data in eigen space x' is defined as eq. 12.

$$x' = T^T x \quad (12)$$

So we can update the references vector by using eq. 13, 14. We only update w_1 the codebook that the nearest to the same class of x , and w_2 the nearest to a different class from x .

$$w_1 \leftarrow w_1 + \alpha \frac{\delta f}{\delta \varphi} \frac{4d_2}{(d_1 + d_2)^2} (T^T x - T^T w_1) \quad (13)$$

$$w_2 \leftarrow w_2 - \alpha \frac{\delta f}{\delta \varphi} \frac{4d_1}{(d_1 + d_2)^2} (T^T x - T^T w_2) \quad (14)$$

B. Handling Unknown Class on AMGLVQ

Third phase is finding the boundary class for incremental class on testing phase. To get the boundary class we use the threshold technical like on the heuristic clustering. We find the threshold of every class to be the boundary of class by using equation 15.

$$r = \frac{\sum_{i=1}^N d_n}{N} + 3\sigma \quad (15)$$

The testing data that out of boundary class output in testing phase, will be created new class as unclassifiable heartbeat, and this class is increment, depend on how far the differences of the data. For the illustration, see on figure 9.

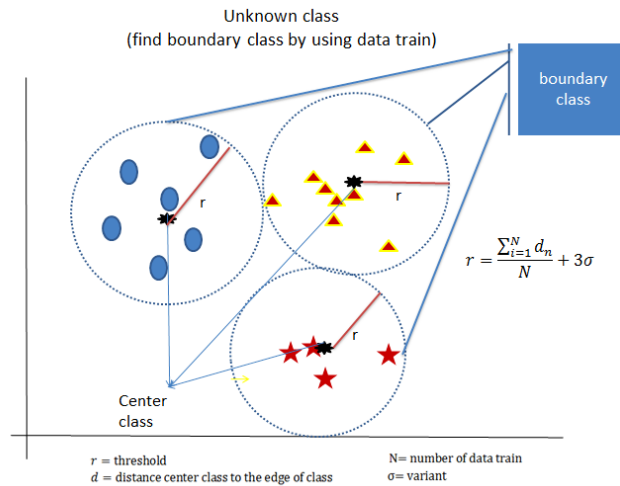


Figure 9. Illustration boundary class

Illustration figure 8, show that we have three classes, circle, star and triangle with each center and class boundary. Then one testing data that never we know in real testing data come, in illustration we draw with dot. The data test is out of the three classes' boundary, so we classify it as unknown data then we create a new class from this data.

IV. DATA SET ENVIRONMENT AND PREPROCESSING

A. MIT-BIH Arrhythmias dataset

In this study we used Iris data set and MIT-BIH arrhythmia database[27]. MIT-BIH Arrhythmia data set was used in this study includes many recordings of both common and life-threatening arrhythmias along with examples of normal sinus rhythm. The database contains 48 records, in which each record contains a pair of 30-min ECG leads signal, mostly MLII lead and lead V1/V2/V4/ V5. The frequency of the ECG data was 360Hz. For this research, we only used the MLII lead as our source data. The groups/classes that we want to classify in our study are;

1. Normal Beat (NOR),
2. Left Bundle Branch Block beat (LBBB),
3. Right Bundle Branch Block beat (RBBB),
4. Premature Ventricular Contraction beat(PVC),
5. Paced beat (P),
6. Atrial Premature beat (AP),
7. Fusion of Ventricular and Normal beat(fVN),

8. Fusion of Paced and Normal beat(fPN),
9. Nodal (Junctional) Escape beat (NE).
10. Aberrated Atrial Prematur beat(aAP),
11. Ventricular Escape (VE),
12. Nodal (junctional) Premature beat (NP).
13. Atrial Escape beat (AE)
14. Supra ventricular Premature beat (SP)

1. Data Complexity

The complexity of Arrhythmia data in term of overlapping data between classes is shown in figure 10 and figure 11. Minority class is omitted on the figure 10 and 11, since the occurrence is too small to be seen.

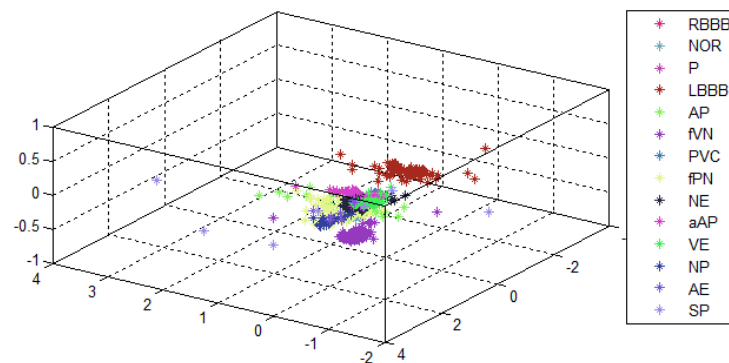


Figure 10. 3D Scatter plot Arrhythmias Heart beat

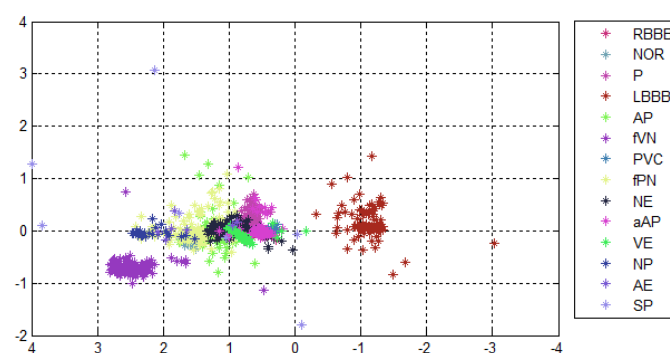


Figure 11. 2D Scatter plot Arrhythmias Heart beat

In this study, we transformed the continuous ECG signals into individual ECG beats that would be given into the classifier to label classes. In beat pattern recognition, it is important to initially suppress or normalize the features that actually not related to the classification but

will seriously degrade the classifier performance. To do so, series of ECG signals passed two preprocessing steps starting from reduction of baseline noise followed by beat extraction to extract the individual beat. Those entire two steps will be described in detail in the following section.

2. Baseline Noise Reduction

The first step of ECG data preprocessing was baseline noise reduction. Baseline noise or some of literatures call it baseline wander, is a condition where the generated ECG signal not lied in to iso-electric line, but shifted up or down for some degrees. It is happen because of the low frequency activity during the ECG record processing. Special treatment need to be done on the ECG signal to prevent miss- interpretation on the classification process.

There are some techniques to reduce low frequency noise, such as linear filtering as developed by J.A. Alste Van[28], polynomial fitting or cubic spline filtering as developed by Mayer [29] and Badilini [30]. By using linear filtering made distortion on the ECG signal, especially between PR segment and ST segment. On the other hand, nonlinear cubic spline interpolation can reduce noise with no significant influent to the ECG signal.

In this study, nonlinear cubic spline interpolation was employed to reduce the baseline noise. Cubic spline was generated exclusively from PR segment sample to estimate and remove noise from the baseline ECG. The baseline noise was estimated from the ECG using PR segment point by cubic splines method. Hence, QRS complex and PR segment need to be detected beforehand. Later on, baseline noise then reduced by simply subtracting the estimated one from the raw data. As can be seen from figure 12, on the top is the raw ECG signal in which some part are shifted away up and down and additional cubic spline illustration. On the bottom part is the resulted ECG after subtraction process.

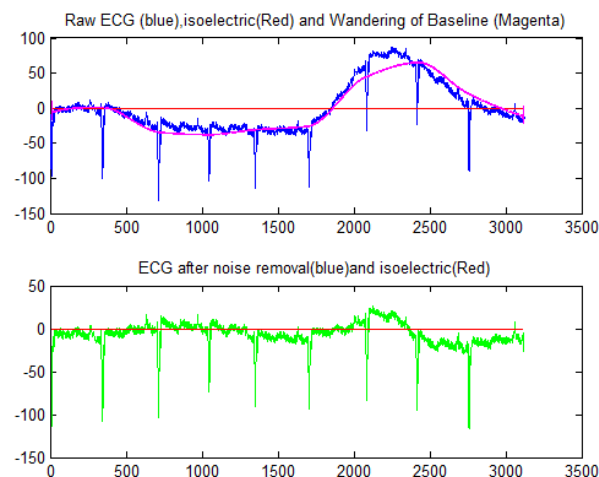


Figure 12. Baseline estimation and removal with cubic splines interpolation

3. Beat Extraction

The second step of ECG preprocessing was ECG beat extraction whereas the continuous ECG signals were segmented into individual ECG beats. To do so, we approximated the width of individual beat into 300 sample data point and the extracted beat was centered on R peak, by utilizing the annotation provided by the database to do the transformation. For each R peak, the beat was extracted by cutting of the continuous signal start at R-150 sample point position until R+149 positions. Hence, the extracted individual beat will contain 300 sample points in width. The illustration sees on figure 13.

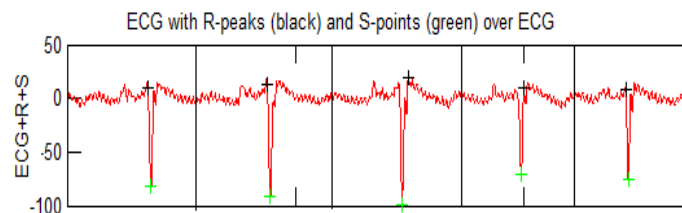


Figure 13. cutoff technique used in this transformation process

B. ECG Biometric Data Set

ECG biometric in this study is recognized normal heart beat from patient in MIT-BIH Arrhythmia databased. In Arrhythmias heat beat classification we only used a sample class, not all beat is used to classify, but in ECG biometric recognized we used 10 patients and classify all normal beat to identify the 10 person. As usually dataset, there is an outlier data, so in this study outlier data has removed by using jack knife method. Jack knife method that is used in this study uses Euclidean distance. For the illustration of outlier removal, see on figure 14.

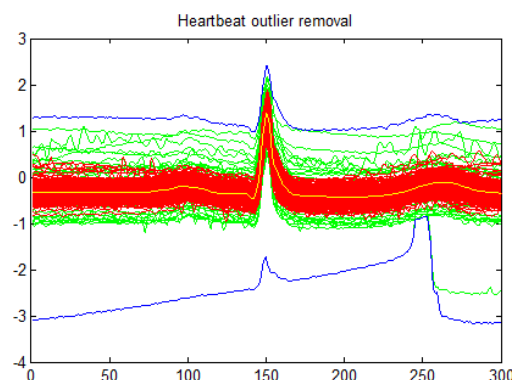


Figure 14. Heartbeat outlier removal

The deferent of number result of heartbeat outlier removal can be seen on table 1.

Table 1. Dataset After Outlier Removal		
Normal Beat of Patient number	Heartbeat with outlier	heartbeat without outlier data
100	2237	2237
101	1859	1831
103	2081	2042
105	2526	2297
106	1507	1505
108	1738	1489
112	2535	2533
113	1788	1744
114	1820	1789
115	1952	1922

C. Generate Synthetics Data set

Synthetics data set we generate from random value number. The synthetic data is generated from toolbox that develop by A. Rakotomamonjy at all [31]. With the same data set, we will compare the performance of multi codebook AMGLVQ to the other algorithm.

V. EXPERIMENT RESULT AND ANALYSIS

In this phase we will validate our method for the classification problem, we used iris and Arrhythmia heartbeats classification problem. We used the UCI database of iris dataset for iris classification problem for balance dataset scenario and MIT-BIH Arrhythmia database for Arrhythmia heartbeats classification problem for imbalanced dataset scenario. Iris and Arrhythmia heartbeats classification are non-linear classification problem. We would to prove that our method was able to apply in balanced dataset or imbalanced dataset. Other than confusion matrix; we used recall, precision, true positive rate, false positive rate and F-measure for evaluation measure of imbalanced dataset.

A. Evaluation measure

An imbalanced datasets, not only is the class distribution is skewed, the misclassification cost is often uneven too. The minority class example are often more important than the majority class examples[32]. Some relevant evaluation measurements are confusion matrix, recall, precision, true positive rate, false positive rate and F-measure. Confusion matrix is useful when accessing the performance without taking cost in to consideration. It is used as a basis for various measures, such as precision and recall. Precision is defined as 16 and recall as 17.

$$precision = \frac{TP}{TP + FP} \quad (16)$$

$$recall = \frac{TP}{TP + FN} \quad (17)$$

F-measure is a common evaluation metrics that combines precision and recall into a single value, usually with equal on both measures. F-measure can be defined as 18.

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (18)$$

True positive rate is the same as recall and the false positive rate equals to 19.

$$false\ positif\ rate = \frac{FP}{FP + TN} \quad (19)$$

In this study we use confusion matrix, recall, precision, F-measure, true positive rate and false positive rate to evaluation measurement.

B. Arrhythmia heartbeats classification

We used fourteen types of Arrhythmia heartbeats. We have groups of individual beats as follows; 200 beats NOR, 200 beats LBBB, 200 beats RBBB, 200 beats PVC, 200 beats P, 200 beats AP, 200 beats fVN, 200 beats fPN, 200 beats NE, 138 beats aAP, 106 beats VE, 47 beats NP, 15 beats AE, 2 beats SP contain 300 sample each as a raw feature for the signal. Then we apply decomposition on the raw feature (300 sample) using conventional PCA then classified them with SVM, back-propagation neural network, GLVQ and LVQ to compare with our methods result. The experiment has two scenario, first scenario we classified arrhythmia heartbeat without unknown class, then in second scenario we add unknown heartbeat type on testing phase.

1. Arrhythmia heartbeats classification without unknown category

In this scenario we evaluate our method AMGLVQ and compare the result to SVM, back-propagation neural network, GLVQ and LVQ. We used conventional PCA to get the best feature for SVM, back-propagation neural network, GLVQ and LVQ, where the feature extraction and classification step used conventional approach of pattern recognition. The experiment result is shown on figure 9. In this experiment PCA feature extraction phase outcome 60 features optimum from 300 full features. Then we used 60 features to classify the Arrhythmia heart beat by using LVQ, GLVQ, SVM and back-propagation-NN. Then the result we compare with AMGLVQ the method that we proposed. Accuracy of LVQ is 85.18%, GLVQ 87.94%, SVM 93.54 %, Back-propagation-NN 94.02 %, and AMGLVQ 95.16 %. Its show that our method leading the accuracy of heartbeats classification. Because with the integrating feature extraction phase to classification phase can be solve the inconsistency problem between optimization feature extraction and classification. For this problem we can see the evaluate measurement on table 2-4. Evaluation measurement of arrhythmias heartbeats classification is presented in table 2-4. AMGLVQ has true positive rate 0.928 and it is the highest than the others method. Not only true positive rate, AMGLVQ has higher value of precision, recall and also F-measure than SVM and back-propagation NN.

Table 2. Evaluation Measure of Backpropagation for Arrhythmia Heart beat Classification

Class	TP Rate	FP rate	Precision	Recall	F measure
'RBBB'	0.95	0.003	0.969	0.95	0.96
'NOR'	1	0.01	0.909	1	0.952
'P'	1	0.001	0.99	1	0.995
'LBBB'	0.96	0.008	0.923	0.96	0.941
'AP'	0.98	0.001	0.99	0.98	0.985
'fVN'	1	0.008	0.926	1	0.962
'PVC'	0.97	0.004	0.96	0.97	0.965
'fPN'	0.98	0.008	0.925	0.98	0.951
'NE'	0.98	0.012	0.899	0.98	0.938
'aAP'	0.406	0.005	0.848	0.406	0.549
'VE'	0.981	0.003	0.945	0.981	0.963
'NP'	0.913	0	1	0.913	0.955
'AE'	1	0.001	0.875	1	0.933
'SP'	0	0	0	0	0
Average	0.866	0.005	0.869	0.866	0.861

Table 3. Evaluation Measure of SVM for Arrhythmia Heart beat Classification

Class	TP Rate	FP rate	precision	Recall	F measure
'RBBB'	1	0	1	1	1
'NOR'	1	0.035	0.752	1	0.858
'P'	1	0	1	1	1
'LBBB'	1	0.005	0.952	1	0.976
'AP'	1	0.001	0.99	1	0.995
'fVN'	0.99	0.004	0.961	0.99	0.975
'PVC'	0.97	0.001	0.99	0.97	0.98
'fPN'	0.97	0.002	0.98	0.97	0.975
'NE'	0.99	0.014	0.884	0.99	0.934
'aAP'	0.435	0.002	0.938	0.435	0.594
'VE'	0.981	0.002	0.963	0.981	0.972
'NP'	0.174	0.003	0.571	0.174	0.267
'AE'	1	0.002	0.778	1	0.875
'SP'	0	0	0	0	0
Average	0.822	0.005	0.840	0.822	0.814

Table 4. Evaluation Measure of AMGLVQ for Arrhythmia Heart beat Classification

Class	TP Rate	FP rate	precision	Recall	F measure
'RBBB'	0.99	0.001	0.99	0.99	0.99
'NOR'	1	0.009	0.917	1	0.957
'P'	1	0.003	0.971	1	0.985
'LBBB'	1	0.006	0.943	1	0.971
'AP'	1	0.002	0.980	1	0.990
'fVN'	1	0	1	1	1
'PVC'	0.96	0.001	0.990	0.96	0.975
'fPN'	0.93	0.004	0.959	0.93	0.944
'NE'	1	0.016	0.870	1	0.930
'aAP'	0.522	0.002	0.947	0.522	0.673
'VE'	0.962	0.002	0.962	0.962	0.962
'NP'	0.913	0.003	0.875	0.913	0.894
'AE'	0.714	0.001	0.833	0.714	0.769
'SP'	1	0.002	0.333	1	0.5
Average	0.928	0.004	0.898	0.928	0.896

2. Arrhythmias beat classification with unknown category on testing phase

Experiments in stage I show that AMLVQ able to improve the accuracy of classification with the system that integrating feature extraction and classification. In stage two our experiment is focus on the ability of our algorithm to detecting the unclassifiable beat as new class on testing phase. We test our methods with adding Ventricular flutter wave on testing phase, which this heartbeat type never used as heartbeat type on training phase. We only train with fourteen heartbeat type, and test with fifteen heartbeat type. The Ventricular flutter wave heartbeat type we adding 50 beats as data test. The experiment result is presented on table 5 and figure 15.

Table 5. Accuracy of Classification with unknown heartbeat type

Method	Accuracy (%)	
	Scenario 1 (without unknown class)	Scenario 1 (with unknown class)
LVQ	85.18	81.32
GLVQ	87.94	83.95
Back- propagation- NN	94.02	89.76
SVM	93.54	89.30
AMGLVQ (proposed method)	95.16	95.04

Table 5 shows that AMGLVQ can classify Ventricular flutter wave as new type of heartbeat so the algorithm generate new class unknown. From 50 beat of Ventricular flutter wave, AMGLVQ can classify 47 beats correct as another fourteen class that have been train. We compare the result of this scenario with back-propagation NN and SVM, the result show that back-propagation and also SVM not be able to classify the Ventricular flutter wave heartbeat as new class. SVM and back-propagation NN classified them on the nearest class, so it makes the accuracy of SVM and back-propagation NN decreasing.

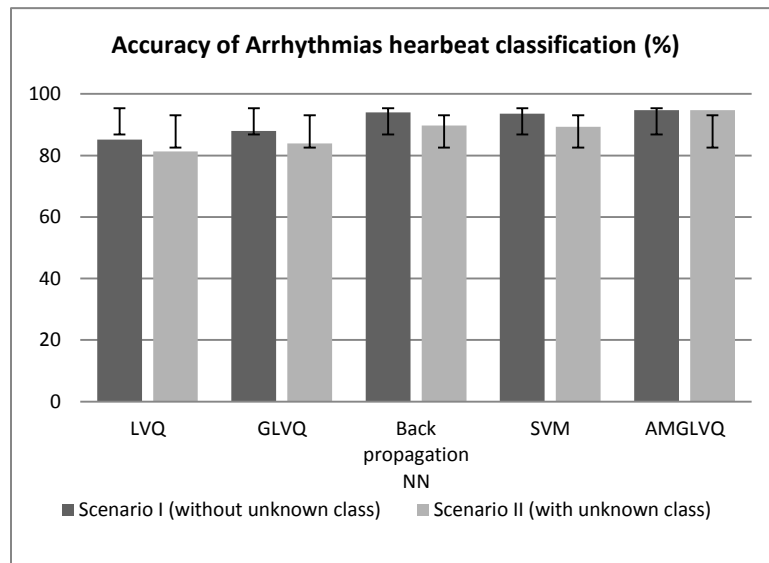


Figure 15. Experiment Result Arrhythmia Heartbeats Classification

Figure 15 also present that AMGLVQ is robust with unclassifiable beat, but not the other algorithms. The others classification algorithm decrease the accuracy after the data set is contain the unclassifiable beats.

C. ECG Biometric classification

This experiment was used normal heartbeat data set as view as on table 3. We compare the result of ECG biometric recognition by AMGLVQ to two others algorithms. We compare the result to Backpropagation Neural Network and RBF Neural Network. The number of heartbeats are 19389, than we spare in two part by cross validation with the ratio 50:50. Than 50% of data used to training proces and another 50% of data used to evaluation phase. The result show that AMGLVQ accuracy is better than two other algorithm, the accuracy of AMGLV is 97.3%, then the accuracy of Backpropagation NN is 88.93% and RBF Network is 95.03%. The detail of experiment result can be seen on table 6-8.

Table 6. Accuracy of Backpropagation for ECG Biometric

Class	TP Rate	FP rate	precision	Recall	F measure
100	0.9991	0.0036	0.9725	0.9991	0.9856
101	0.9967	0.0011	0.9892	0.9967	0.993
103	0	0	0	0	0.000
105	0.9713	0.0407	0.7625	0.9713	0.854
106	0.9973	0.0077	0.9158	0.9973	0.955
108	0.9974	0.0002	0.9974	0.9974	0.9974
112	0.9944	0.0005	0.9968	0.9944	0.996
113	0.9977	0.0236	0.8065	0.9977	0.892
114	0.9978	0.0447	0.6982	0.9978	0.822
115	1	0.001	0.992	1	0.996
Average	0.895	0.012	0.813	0.895	0.849

Table 7. Accuracy of RBF Network for ECG Biometric

Class	TP Rate	FP rate	precision	Recall	F measure
100	0.9517	0.0001	0.9990	0.9517	0.9748
101	0.9325	0.0082	0.9225	0.9325	0.9275
103	0.9187	0.0005	0.9958	0.9187	0.9557
105	0.9200	0.0032	0.9751	0.9200	0.9468
106	0.9388	0.0049	0.9413	0.9388	0.9401
108	0.9895	0.0129	0.8678	0.9895	0.9247
112	0.9871	0.0032	0.9784	0.9871	0.9827
113	0.9816	0.0029	0.9704	0.9816	0.9760
114	0.9561	0.0073	0.9316	0.9561	0.9437
115	0.9321	0.0117	0.8988	0.9321	0.9152
Average	0.9508	0.0055	0.9481	0.9508	0.9487

Table 8. Accuracy of AMGLVQ Network for ECG Biometric

Class	TP Rate	FP rate	precision	Recall	F measure
100	0.9632	0.0015	0.9886	0.9632	0.9758
101	0.9885	0.0060	0.9419	0.9885	0.9646
103	0.9665	0.0047	0.9598	0.9665	0.9631
105	0.9803	0.0059	0.9555	0.9803	0.9678
106	0.9811	0.0029	0.9655	0.9811	0.9732
108	0.9730	0.0004	0.9945	0.9730	0.9837
112	0.9907	0.0018	0.9884	0.9907	0.9895
113	0.9839	0.0025	0.9749	0.9839	0.9793
114	0.9590	0.0014	0.9863	0.9590	0.9725
115	0.9480	0.0023	0.9789	0.9480	0.9632
Average	0.973	0.003	0.973	0.973	0.973

From the table we can see that AMGLVQ has recall, precision, F-measure and TP rate higher than the other.

D. Generate Synthetic Data Sets

We generate random data sets that consist of three classes. The first one is three classes data set with ten data every class. The second, 300 data set with 100 data every classes. The result is shows that multi code book AMGLVQ show better performance than the SVM. We used SVMW because it's basic standard classifier recently. For the first synthetic data set generator, AMGLVQ accuracy is 93.33% with elapse time 0.0350 seconds, and the multiclass SVM accuracy is 90% with the elapse time 0.2215 seconds. The contour of topology data can be show at figure 16 and figure 17.

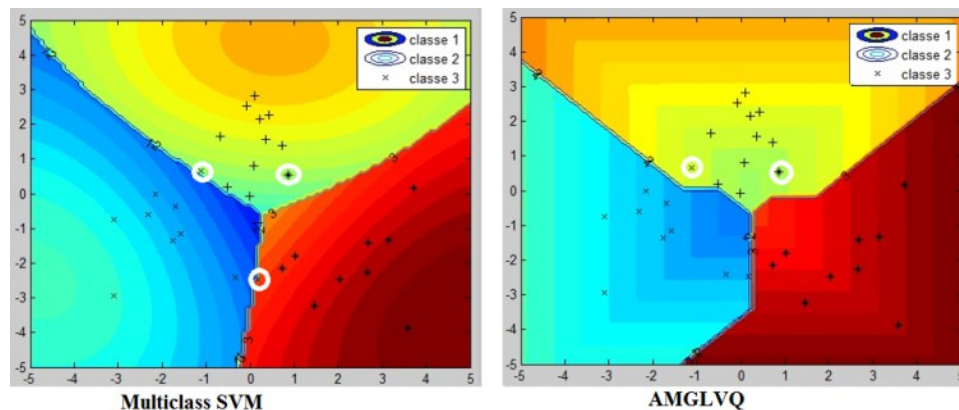


Figure 16. The contour of topology synthetics generate data I

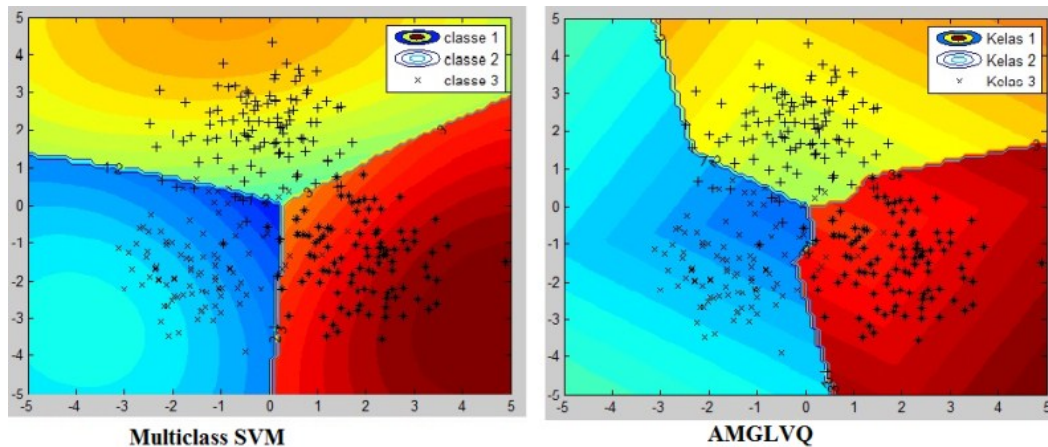


Figure 17. The contour of topology synthetics generate data II

For the Second synthetic data set generator, AMGLVQ accuracy is 93.67% with elapse time 0.1282 seconds, and the multiclass SVM accuracy is 92.33% with the elapse time 1.5697 seconds. The contour of topology data can be see at figure 16.

VI. CONCLUSIONS.

In this study we show that AMGLVQ show good performance for heartbeat classification, ECG biometric and generate synthetic data than the others classification algorithm, because AMGLVQ is able to handle inconsistency optimization on feature extraction and classification problem, imbalanced data set and also detecting uncategorical data test. Imbalanced data set handling by AMGLVQ because this algorithm using data selection in training phase, it is like under-sampling technic that embedded to algorithm process. The selecting data that representative in AMGLVQ uses miss classification error for evaluation criteria for selection.

AMGLVQ handling uncategorical data, by generate new class if the testing phase detected new type of heartbeat, different with the conventional algorithm. AMGLVQ is able to generate new class, because this algorithm adopted concept boundary class of heuristic clustering. With the other data set, we also can show that our methods is has good performance.

ACKNOWLEDGEMENT

This work are supported by part of Grant of Joint Research for Foreign Affairs and International Publication No. 1495/E5.2/PL/2011 by the Ministry of Education Republic of Indonesia and supported by part of "Sistem Inovasi Nasional" (SINAS) No. 1347/H2.R12/HKP.05.00 Perjanjian/2012 by Ministry Research and Technology Republic of Indonesia.

REFERENCES

- [1] A. H. Sahoolizadeh, B. Z. Heidari, and C. H. Dehghani, "A New Face Recognition Method using PCA, LDA and Neural Network," *International Journal of Computer Science and Engineering*, vol. 2, no. 4, pp. 218–223, 2008.
- [2] D. Zhang, G. L. G. Lu, W. L. W. Li, L. Z. L. Zhang, and N. L. N. Luo, "Palmprint Recognition Using 3-D Information," *IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews*, vol. 39, no. 5, pp. 505-519, 2009.
- [3] F. Agraftoti, F. M. Bui, and D. Hatzinakos, "Medical biometrics in mobile health monitoring," *Security and Communication Networks*, vol. 4, no. 5, pp. 525-539, 2011.
- [4] Y. Wang, F. Agraftoti, D. Hatzinakos, and K. N. Plataniotis, "Analysis of Human Electrocardiogram for Biometric Recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1-12, 2008.
- [5] Y. Wang, K. N. Plataniotis, D. Hatzinakos, T. Edward, S. R. Sr, and C. Engineering, "Integrating Analytic And Appearance Attributes For Human Identification From Ecg Signals," in *IEEE Biometrics Symposium*, 2006.
- [6] N. Belgacem, "ECG Based Human Authentication using Wavelets and Random Forests," *International Journal on Cryptography and Information Security*, vol. 2, no. 2, pp. 1-11, Jun. 2012.
- [7] R. D. Labati, A. Genovese, V. Piuri, and F. Scotti, "Wildfire Smoke Detection using Computational Intelligence Techniques Enhanced with Synthetic Smoke Plume Generation," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. PP , Issue, pp. 1 - 10, 2012.
- [8] A. Bouchachia and N. Nedjah, "Adaptive Incremental Learning In Neural Networks," *Neurocomputing*, vol. 74, no. 11, pp. 1783-1784, May 2011.

- [9] H. Atoui, J. Fayn, and P. Rubel, "A novel neural-network model for deriving standard 12-lead ECGs from serial three-lead ECGs: application to self-care.," *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 14, no. 3, pp. 883-90, May 2010.
- [10] B. M. Wilamowski and H. Yu, "Neural network learning without backpropagation.," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 21, no. 11, pp. 1793-803, Nov. 2010.
- [11] X. Liu and J. Cao, "Robust State Estimation For Neural Networks With Discontinuous Activations," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 40, no. 6, pp. 1425-37, Dec. 2010.
- [12] T. Kohonen, "Learning-Vector Quantization and the Self-Organizing Map," in *Theory and Applications of Neural Networks*, 1992, pp. 235-242.
- [13] A. Sato and K. Yamada, "Generalized Learning Vector Quantization," in *Advances in Neural Information Processing Systems 8 Proceedings of the 1995 Conference*, 1996, vol. 7, pp. 423-429.
- [14] F.-michael Schleif, T. Villmann, B. Hammer, P. Schneider, and M. Biehl, "Generalized derivative based kernelized Learning Vector Quantization," *Springer, Intelligent Data Engineering and Automated Learning – IDEAL 2010 Lecture Notes in Computer Science*, vol. 6283, pp. pp 21-28, 2010.
- [15] K. Marika, B. Hammer, M. Biehl, and T. Villmann, "Functional relevance learning in generalized learning vector quantization," *Neurocomputing* vol. 90, pp. 85-95, 2012.
- [16] B. Hammer, M. Strickert, and T. Villmann, "On The Generalization Ability of GRLVQ Networks," *Neural Processing Letters*, pp. 109-120, 2005.
- [17] B. Kusumoputro, H. Budiarto, and W. Jatmiko, "Fuzzy-neuro LVQ and Its Comparison with Fuzzy Algorithm LVQ In Artificial Odor Discrimination System," *ISA Transactions*, pp. 395-407, 2002.
- [18] I. M. A. Setiawan, E. M. Imah, and W. Jatmiko, "Arrhythmia Classification using Fuzzy-Neuro Generalized Learning Vector Quantization," in *IEEE International Conference on Advanced Computer Science and Information System 2011 (ICACSIS 2011)*, 2011, pp. 978-979.
- [19] E. Mwebaze et al., "Neurocomputing Divergence-based classification in learning vector quantization," *Neurocomputing*, vol. 74, no. 9, pp. 1429-1435, 2011.
- [20] E. M. Imah, W. Jatmiko, and T. Basaruddin, "Adaptive Multilayer Generalized Learning Vector Quantization (AMGLVQ) As New Algorithm With Integrating Feature Extraction

And Classification For Arrhythmia Heartbeats Classification,” in *IEEE international Conference on System Man and Cybernetics, Seoul 2012*, 2012.

- [21] S. García and F. Herrera, “Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy,” *Evolutionary computation*, vol. 17, no. 3, pp. 275-306, Jan. 2009.
- [22] B. X. Wang and N. Japkowicz, “Boosting support vector machines for imbalanced data sets,” *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1-20, 2009.
- [23] N. Japkowicz, “Learning from Imbalanced Data Sets: A Comparison of Various Strategies,” in *Proc. Am. Assoc. for Artificial Intelligence (AAAI) Workshop*, 2000, vol. 68.
- [24] C. Vivaracho-pascual and A. Simon-hurtado, “Improving ANN performance for imbalanced data sets by means of the NTIL technique,” *IEEE International Join Conference on Neural Networks (IJCNN)*, 2010.
- [25] G. de Lannoy, D. Francois, J. Delbeke, and M. Verleysen, “Weighted conditional random fields for supervised interpatient heartbeat classification,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 1, pp. 241-7, Jan. 2012.
- [26] X. Wang and K. K. Paliwal, “Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition,” *Pattern Recognition*, vol. 36, no. 10, pp. 2429-2439, Oct. 2003.
- [27] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet,” *Circulation*, vol. 101, no. 23, p. 215, 2000.
- [28] J. A. Van Alsté, W. Van Eck, and O. E. Herrmann, “ECG baseline wander reduction using linear phase filters,” *Computers and biomedical research an international journal*, vol. 19, no. 5, pp. 417-427, 1986.
- [29] C. R. Meyer, “Electrocardiogram Using Cubic Baseline Noise Estimation and Removal Splines and State-Space Computation Techniques” *Computers and Biomedical Research*, 1977.
- [30] F. Badilini, A. J. Moss, and E. L. Titlebaum, “Cubic Spline Baseline Estimation In Ambulatory Ecg Recordings,” *Engineering In Medicine And Biology*, vol. 13, no. 2, pp. 584-585, 1991.
- [31] A. Rakotomamonjy, R. Flamary, and F. Yger, “Learning with infinitely many features,” *Machine Learning*, vol. 91, no. 1, pp. 43-66, DApril 2013.
- [32] J. Weng, Cheng G, Poon, “A New Evaluation Measure for Imbalanced Datasets,” in *Seventh Austraasian Data Mining Conference (AusDM 2008)*, 2008.