Kuwait J. Sci. **42 (2)** pp. 206-235, 2015

A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data

DIANA MOSES[*,**] AND DEISY C[*]

**Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai – 625015, India.*
*** itsdianamoses@gmail.com;*
Corresponding author: itsdianamoses@gmail.com

ABSTRACT

Remote cardiovascular disease (CVD) diagnosis from ECG plays an important role in health care domain. Data mining, the major step in the process of the extraction of knowledge using descriptive and predictive algorithms that aid in making proactive decisions, has also been used for CVD diagnosis. Recently, diverse techniques have been developed for analyzing the ECG signals. However, due to the diversity of techniques used, terminologies, performance measures used in different techniques makes analysis and comparing of results thwarting. The aim of this work is to essentially explore and present the analysis of different data mining algorithms proposed earlier in literature for CVD diagnosis, their advantages and limitations. This paper presents various techniques for CVD diagnosis using data mining from an ECG signal under four major phases – ECG Acquisition, ECG Compression, ECG Feature Extraction and ECG diagnosis. The primary aim of this paper is to categorize the various researches done in this regard to provide a glossary for interested researchers and to aid in identifying their potential research direction.

Keywords: Cardiovascular disease; data mining algorithms review; ECG analysis; telecardiology.

INTRODUCTION

The World Health Organization (WHO) rated Cardiovascular Disease (CVD) as the number one killer disease of human race with an estimate of 23.6 million deaths by 2030 (WHO). Mortality due to CVD in 2004 was 2.7 million in India and is projected to cross 4 million in 2030 (Patel *et al*., 2011). CVD diagnosis is crucial, where every second of delay in disease diagnosis and treatment initiation counts on the patient's life (Luca *et al*., 2004). ECG data is the key for diagnosing cardiac diseases. ECG, in many cases, is acquired for over 24 hours using the Holter monitoring system with data volume rising up to 50GB per patient per day (Pooyan *et al*., 2005). Manual screening of this huge amount of data becomes incomprehensible and requires automated computer aided systems. Furthermore, In India the availability of doctors for timely diagnosis and treatment initiation is only 1 doctor per 1,700 patients (Kumar UA, 2013). Hence computerised methods for CVD diagnosis is the need of the hour.

The first use of computers in ECG based CVD diagnosis was the work done by Steinberg *et al*. in 1962 (Steinberg *et al*., 1962). The system converted the analog ECG signal into a suitable format to store in magnetic tapes and provided this input to a general purpose computer. It analyzed voltage fluctuations from the baseline and calculated the amplitude of P, Q, R, S, T waves of the ECG signal and proposed that with statistical analysis cardiac diseases could be diagnosed from ECG using computers. But the system botched to calculate the width of the QRS waves. Consecutively Young & Huggins (1963) proposed a method to represent ECG signal as a series of numerical values to make it available for the statistical analysis. Using the proposed representation, in the late 1970's the Massachusetts Institute of Technology-Beth Israel Hospital (MIT BIH) provided a collection of ECG datasets in both signal and numerical formats (Moody & Mark, 2001). The availability of such a benchmark dataset promoted rapid research and development in the technology of automated CVD diagnosis.

Telecardiology is the electronic transmission of cardiac data from the patient site to a consulting site for the provision of health care services (Hailey *et al*., 2004). The remote monitoring of ECG allows the patient and doctor to be at distant places and still provides the patient with timely diagnosis and other health care services. The first step towards telecardiology dates back to as early as 1911 where ECG was transmitted over telephone communication lines for remote monitoring (Hailey *et al*., 2004). Subsequently, developments accommodating advancements in both healthcare and technology have been made especially in the past decade. These developments attract much research to address the overall accuracy and efficiency of the telecardiology system. Figure 1 shows the basic framework of telecardiology system that has evolved in the recent years (Lin *et al*., 2010; Ceylan *et al*., 2010; Sufi *et al*., 2009; Sufi & Khalil, 2011).

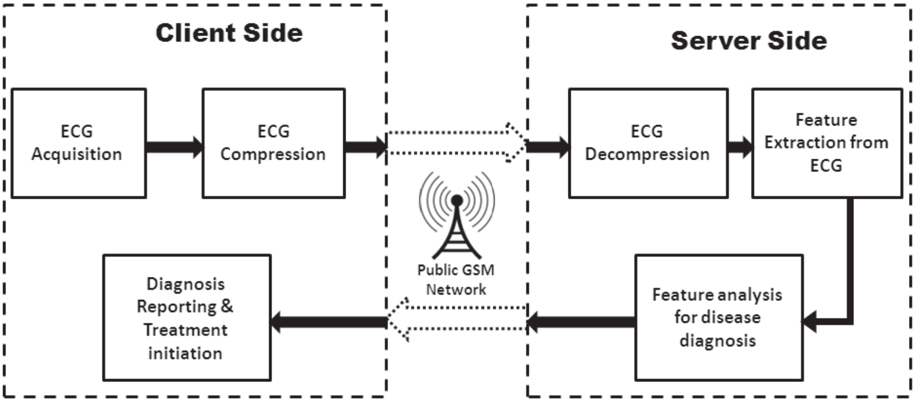
Client Side
Server Side
ECG Acquisition
ECG Compression
ECG Decompression
Feature Extraction from ECG
Public GSM Network
Diagnosis Reporting & Treatment initiation
Feature analysis for disease diagnosis

Fig. 1. General framework of a Telecardiology system

The major sections involved in computer assisted CVD diagnostic system includes ECG Acquisition, ECG compression and Transmission at the client side followed by decompression, extraction of feature values from ECG wave and diagnosis of the disease using data mining at the server side (Lin *et al*., 2010; Ceylan *et al*., 2010; Sufi *et al*., 2009; Sufi & Khalil, 2011). The accuracy of the data mining algorithm mainly depends on the feature extraction process. Both feature extraction and data mining being computationally expensive are employed on the server side of the telecardiology system. As the system covers multiple domains, work done by various researchers in this regard are categorized and presented under the previously mentioned sections. This paper examines the overview of ECG analysis systems, relevance and effectiveness of different data mining algorithms for CVD diagnosis. The review is intended to help biomedical researchers and decision-makers either commercial or public, to establish effective CVD diagnosis services.

The remaining part of the paper is organized under the following sections: ECG Acquisition - gives a brief introduction to different ECG Acquisition methods, ECG compression - explains the relevance of ECG compression in any CVD diagnosis system and provides a categorization of existing ECG compression algorithms, Feature Extraction - briefly discusses about various feature extraction methods, Data mining algorithms - advantages and disadvantages of different data mining algorithms specific to CVD diagnosis. Comparative discussions about the different data mining algorithms, different standards for presenting the results, perspectives and challenges of data mining algorithms in CVD diagnosis are presented in the end.

ECG ACQUISITION

The Electrocardiogram (ECG) is a device used to record the electrical impulses of the heart. It serves as the best measure to detect any abnormality affecting the heart. Figure 2 shows a sample ECG waveform depicting the key features that aid in diagnosis of a heart disease.

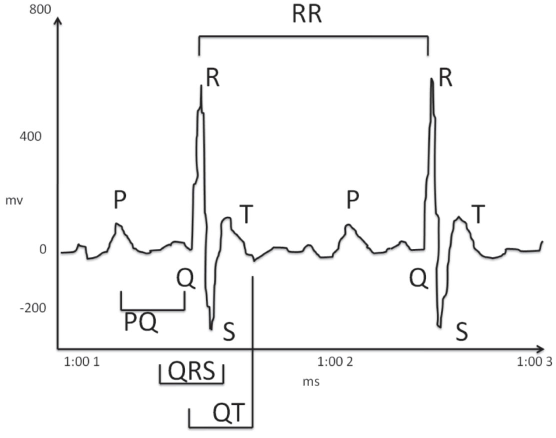

Fig. 2. An ideal ECG waveform and its features

ECG is essentially a signal and was used in analog processing systems until late 1930's. Later with the advancements in digital processing systems, various methods for digital representations are used of which three formats are used predominantly - SCP-ECG (Standard Communications Protocol for Computer-Assisted Electrocardiography), DICOM-ECG (Digital Imaging and Communications in Medicine), and HL7 aECG (Heath Level 7 - Annotated ECG) (Jumaa *et al*., 2008; Zheng *et al*., 2010). ECG may be recorded using single to 12 lead ECG devices. The leads are electrodes placed on the body surface of the patient to record the electrical activity of the heart. But normally a 3-lead ECG with lead-1(left ventricle, left atrium), lead-2(left and right ventricle) and lead-3 (opposite to lead 2) is sufficient for diagnosis purposes. Acquisition devices used in hospitals are bulky but provide extensive support for high precision 12- lead Holter monitoring, where ECG is recorded for extended periods of time (over 24 hours). Perceptibly, devices of such size with several wires require the patient to be immobilized, thwarting the patient's day to day activities. However, wearable health-monitoring systems provide real-time unobtrusive monitoring of patients' physiological parameters through the deployment of several on-body sensors are also available (Lin *et al*., 2010). Yoo *et al*. (2009) and has developed a compact planar-fashionable Circuit Board-Based Shirt that holds the ECG leads and other circuitry fabricated on a shirt. Although different ECG acquisition devices are available, researchers working on autonomous computer based diagnosis systems have worked (Sufi *et al*., 2009; Sufi & Khalil, 2011) mainly with the MIT's preprocessed datasets (Moody & Mark, 2001). This process allows the researchers to focus more on the crucial feature extraction and diagnosis methods than on the ECG de-noising and preprocessing methods.

ECG COMPRESSION

In a typical telecardiology system, ECG signal acquired from any of the devices mentioned above are transmitted from the patient to a remote server using different wireless protocols like SMS, MMS, HTTP and other custom socket routine, through mobile phone as a signal transmitter (Sufi *et al*., 2009; Sufi & Khalil, 2011). However, the manifestation of any cardiac abnormality occurs randomly in the ECG timeline. Therefore the monitoring of ECG is carried out for extensive periods of time (i.e., for over 24 h). Obviously the volume of the data to be transmitted and analyzed becomes enormous (Pooyan *et al*., 2005). Hence a method to compress the data without making significant change in the reconstructed signal is required. Many existing ECG signal compression methods have been proposed over the past four decades. Figure 3 shows how ECG signal compression techniques can be broadly classified into the major categories namely lossless time domain, lossy time domain and lossy frequency domain methods. While the ECG compression algorithms in the literature have dealt only with the categorization of lossy methods, we have included the lossless methods and frequency domain methods in the categorization.

Lossless methods involve no loss of information, which implies that the original data can be recovered exactly from the compressed data. In assurance for this quality, lossless methods offer low compression ratios than the lossy counterparts. Lossy compression techniques involve some loss of information, and original data generally cannot be recovered or reconstructed exactly. In return for accepting the distortion in the reconstruction, the lossy methods generally obtain much higher compression ratios than is possible with lossless compression of the order of 20:1. With the advancements in computational competence of the wireless devices and the affordability of communication facilities, systems are able to accommodate lossless methods for real-time compression and transmission as well.

Lossless time domain methods exploit the redundancy in the data to achieve compression (Health Informatics 2005). Lossless methods like Huffman coding and LZW (Lempel–Ziv–Welch) coding are used for ECG compression with compression ratios ranging from 2:1 to 4:1 (Health Informatics 2005, Welch, 1984). In the recent years much of research is done on lossless ECG compression methods. These methods involve a series of preprocessing steps followed by which is the actual compression (Sufi *et al*., 2009). The preprocessed signal is losslessly encoded using a symbol substitution method. Although information is lost during preprocessing these algorithms claim to be lossless, since the preprocessed signal is completely recoverable from the compressed data.

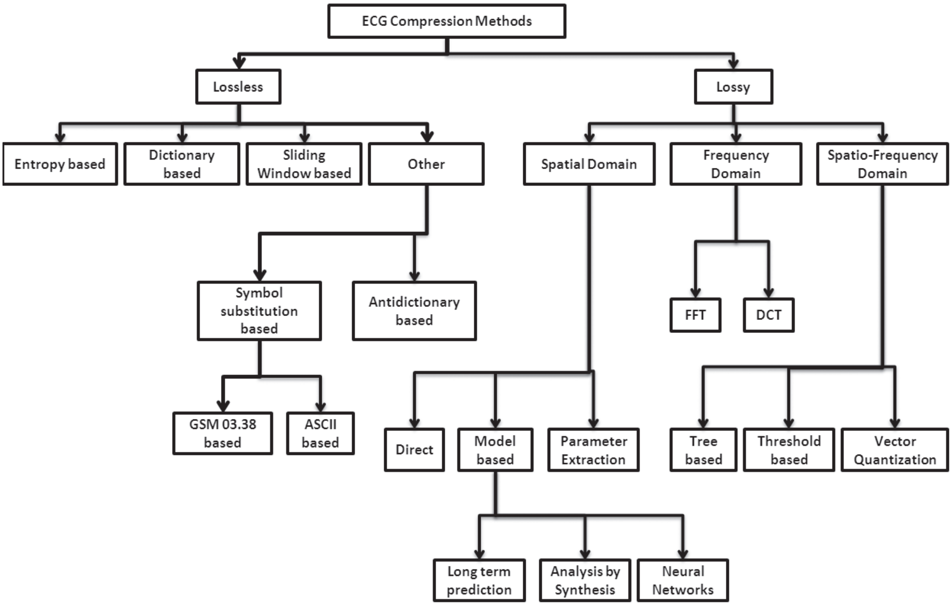

Fig. 3. Classification of ECG compression methods

Lossy methods on the other hand give up complete reconstruction and encode the significant portion of the signal, while discarding the rest (Benzid *et al*., 2008). Distortion measure is used to quantify the amount of data lost during the compression and specifies how much the recovered signal is close to the original signal. While lossless methods are employed only in the time-domain, lossy methods are classified into time domain and frequency/transformed domain. Lossy time domain methods include direct compression methods and parameter based extraction methods. Direct compression methods detect redundancies on direct analysis of the actual signal samples to achieve compression in the time domain. Examples of direct compression methods include Amplitude zone time epoch coding (AZTEC) (Cox *et al*., 1968), Co-ordinate reduction time encoding system (CORTES) (Abenstein & Tompkins, 1982) and the Fan algorithm (DiPersio & Barr, 1985). Parameter extraction methods extract and encode only the key features and parameters of the signal, hence is an irreversible process with which a particular characteristic or parameter of the signal is extracted and encoded. Examples include peak picking method (Jalaleddine *et al*., 1990), principal component analysis (Chawla, 2007) and syntactic method (Iwata *et al*., 1990). Lossy frequency domain methods transform the spatial data into frequency domain and analyze the energy distribution of the signal data. The available transformations include Fourier transform; Fourier descriptor (Reddy & Murthy, 1986), Karhunen–Loeve transform (KLT) (Womble *et al*., 1977), Walsh transform, Discrete cosine transform (DCT) (Batista *et al*., 2001) and Wavelet transform (Lu *et al*., 2000). The wavelets are predominantly used because of their capability to provide multi-level resolution including both the time and frequency domain (Pooyan *et al*., 2005; Goudarzi *et al*., 2005; Ku *et al*., 2010; Tohumoglu & Sezgin, 2007).

Considering the impact of the compression phase in a telemedicine environment and also the requirement to losslessly store and retrieve ECG recordings MIT launched the MIT-BIH ECG Compression test database (Moody *et al*., 1988). Challenges concerning successful ECG compression algorithm are:

- Efficient Storage – Lossless compression methods with high compression ratio
- Efficient transmission - When transmitted using messaging protocols, any compression scheme must use the limited character sets available in MMS and SMS protocols in order to avoid any data loss.
- Reduced decompression time – to mitigate the delay in diagnosis of abnormalities.

FEATURE EXTRACTION FROM ECG

The foremost step in automatic ECG diagnosis is the extraction of the key features of the state of the heart namely the P wave, Q wave, QRS complex, R-R interval and ST segment (Sufi *et al*., 2009). Features are numerical values representing the nature of ECG signals by the size, shape of its constituents, frequency of occurrence and other accompanying variations. Ranges of approaches employed for feature extraction depend on the time, memory and processing specifications. The first ever automated QRS detection algorithm was developed by Pan & Tompkins (1986). Preprocessing was accomplished using linear and nonlinear digital filters followed by peak analysis to produce event vectors. The vectors are processed by decision rules to locate QRS complexes. A time-averaged signal with the interval width equal to the averaging window was used for peak analysis. The QRS detection algorithm derived an optimized set of decision rules together with added T-wave discrimination. The QRS detection algorithm used integer arithmetic, making it particularly suitable for real-time implementation.

Extracted features may be represented in actual millisecond time scale or in other representative scales. The former are called morphological features, while the latter are called morphological descriptors. Feature extraction may be done in the spatial, frequency and time-frequency domain. While morphological features are extracted in the spatial domain, features extracted from other domains describe the features in a different scale, hence called morphological descriptors. Figure 4 shows different usage scenarios of various compression and feature extraction methods. Spatial methods include pattern matching, mathematical morphology, and artificial neural networks among others (Karpagachelvi *et al*., 2010). Olvera (2006) utilized a matched filter to detect different signal features on a human heart electrocardiogram signal. The detection of the ST segment was more difficult to extract using the matched filter, due to noise and amplitude variability. The more complex part was creating the revealing method to extract the feature of interest in each ECG signal. An approach using combinatorial model was proposed by Iliopoulos & Michalakopoulos (2010), where the data was indexed using suffix tree or suffix array. Patterns were searched in time, relative to the length of the pattern. He extended his work to detecting QRS complexes on low computational devices as well (Iliopoulos & Michalakopoulos, 2011). On the other hand, where computational costs have no bars, chaos theory (Babloyantz & Maurer, 1996; Yeragania & Rao, 2003), artificial neural networks (Maglaveras *et al* 1998; Moavenian & Khorrami, 2010) and support vector machines (SVM) (Lei *et al*., 2008; Mehta & Lingayat, 2008) are also used. The feature extraction using SVM is shown to have 99.75% accuracy (Mehta & Lingayat, 2008; Mehta & Lingayat, 2009). However due to the computational overhead and processing time involved, methods other than SVM with closer to such degree of accuracy are in demand (Mehta & Lingayat, 2009).

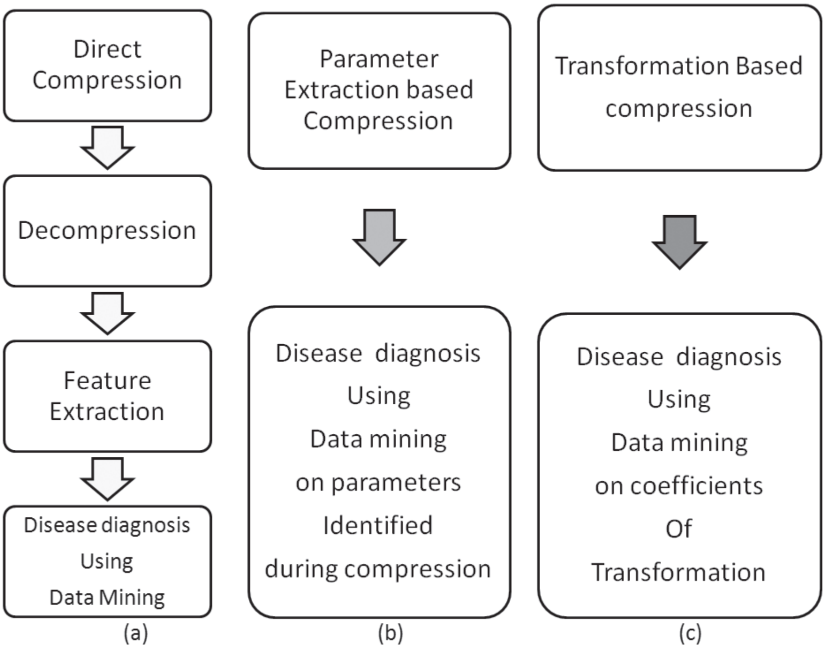
Direct Compression
Decompression
Feature Extraction
Disease diagnosis Using Data Mining
(a)
Parameter Extraction based Compression
Disease diagnosis Using Data mining on parameters Identified during compression
(b)
Transformation Based compression
Disease diagnosis Using Data mining on coefficients Of Transformation
(c)

Fig. 4. Different feature extraction scenarios (**a**) Direct compression in combination with any feature extraction method, (**b**) Parameter extraction based compression; (**c**) Transformation based compression.

The transformation methods, both frequency and time-frequency domain, are essentially employed for signal compression. Signals compressed using transformation methods and parameter extraction methods allow skipping the feature extraction process (Goudarzi *et al.*, 2005; Ku *et al.*, 2010; Tohumoglu & Sezgin, 2007). While the features are directly available as a result of parameter extraction based compression scheme, the transform coefficients are taken up for further processing after transformation based compression methods. This explains the key use of transformation based methods in remote monitoring systems. Transformation methods like Fourier transform provide the frequency domain resolution, while wavelets offer both time and frequency resolution (Karpagachelvi *et al.*, 2010). Wavelets are the most commonly preferred transform methods, as they facilitate analysis at multiple resolutions. Wavelet expansions represent the ECG as segments, each segment represented by a polynomial, whose coefficients are used to derive the features set (Addison, 2005). Wavelet is especially valuable because of its ability to elucidate simultaneously local spectral and temporal information from a signal in a more flexible way.

Another key advantage of wavelet techniques is the variety of wavelet functions available, thus allowing the most appropriate to be chosen for the signal under investigation. Castro *et al*. (2005) used wavelet transform and developed a method for choosing an optimal mother wavelet from a set of orthogonal and bi-orthogonal wavelet filter bank. The best correlation with the ECG signal was adopted by way of dividing the coefficients of each cycle into three segments that are related to P-wave, QRS complex, and T-wave. The summation of the values from these segments provided the feature vectors of single cycles.

Accuracy of disease diagnosis greatly depends on the efficient feature extraction method. A thorough assessment of feature extraction methods can be acquired from (Karpagachelvi *et al*., 2010). There are 22 spatial domain features that can be extracted from the ECG signal with only maximum of three feature combinations used for disease diagnosis reported in literature (Sufi *et al*., 2009; Sufi & Khalil, 2011). Although different efficient methods are available in literature, the choice of the method is based on resources available and other real-time considerations. An efficient method with lower computational requirements and higher accuracy is still required (Karpagachelvi *et al*., 2010).

DATA MINING ON EXTRACTED FEATURES

The process of ECG classification using the extracted feature set is accomplished by building a model for extracted feature set to describe and predict the state of health of the patient. Although the accuracy of the classification is mainly dependent on the extracted feature set, various methods are applied for building the classification model including artificial neural networks (Patra *et al*., 2005; Khashei *et al*., 2012), Fuzzy methods (Yeh *et al*., 2012; Shih *et al*., 2010), Support vector machines (SVM) (Moavenian & Khorrami, 2010; Yildiz *et al*., 2011), Hidden Markov Model (Dumont *et al*., 2008; Andreão *et al*., 2008) and data mining (Ros *et al*., 2004; Pecchia *et al*., 2011; Kumar *et al*., 2011). The choice of the classification method is based on the availability of computational resources, efficiency, real-time processing requirements, generalization to data not present in training set and ability to handle missing and erroneous data. Another aspect is understandability of the results, essentially because the end user of the process is health care personnel, who cannot be expected to be a technology expert and the exploitation of the results are mitigated by lower 'understandability quotient'. This is the key reason for the applicability of data mining methods in health care domain over neural networks and SVM based methods.

Data mining is a class of database applications that look for hidden patterns in a group of data that can be used to predict future behaviour. Many domains including retail, finance, healthcare, manufacturing transportation, and aerospace are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies, statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed (Minas *et al*., 2010; Alkoot, 2014). The knowledge gained is put to use in strategy planning and proactive decision making.

The main goals of data mining are either descriptive or predictive. Descriptive mining describes concepts or task-relevant data sets in concise, summarizing, informative, or discriminative form. A descriptive modelling technique, such as clustering, produces classes (or categories), which are not known in advance. To achieve this goal, some criteria that specify when two data items probably belong to the same class called similarity measure are used. Whereas, predictive mining is based on data and analysis, construction of models for the database, and predict the trend and properties of unknown data. A predictive modelling technique, such as classification, starts from a given classification of the data items, from which it derives conditions on the properties of the data objects, which allow predicting the membership to a specific class.

Generally, the data mining stage is computationally intensive and hence performed on the server side. Recently applicability of simple algorithms for this purpose is much analyzed to involve execution of these algorithms on low computational devices (LCDs), such as mobile phones and tablets. The LCDs being readily available when used for providing health care services could benefit both user and service provider by lowering both investments for high computational devices and time for the patient to get the access to one of these devices. Especially in the health care domain with numerous patients accessing resources simultaneously, the requirement for algorithms usable in LCDs is evident. For this reason algorithms other than support vector machines, genetic algorithms and artificial neural networks are required. This paper discusses descriptive data mining methods like k-means, Expectation Maximization and predictive methods such as decision trees and time series modelling that offer low computational complexity and can be implemented on LCDs.

K-nearest neighbour clustering

The k-nearest neighbour (k-NN) is a partitioning based clustering algorithm, where each cluster is a group of objects. k-NN aims to form spherical clusters surrounding the k selected means. The clusters satisfy constraints such as: (i) each cluster contains at least one object, and (ii) each object must belong to exactly one cluster. The clustering efficiency can be improved by moving objects from one cluster to another at every iteration and finally halts when the objects of one cluster are closely related and objects belonging to different cluster are far from each other.

To determine the distance between samples, different distance measures such as Euclidean and Mahalanobis measures are used (Given in Eq 1, 2).

Euclidean distance:

$$D(i,j)=\sqrt{\left((x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+.....+(x_{in}-x_{jn})^2\right)} \tag{1}$$

Where *D(i,j)* is the Euclidean distance between objects *i* and *j* represented as $i=(x_{i1}, x_{i2}, x_{i3}, ..., x_{in})$ and $j = (x_{j1}, x_{j2}, x_{j3}, ..., x_{jn})$ (Ros *et al*., 2004; Christov *et al*., 2006; Lanatá *et al*., 2011; Kiranyaz *et al*., 2011).

Mahalanobis distance:

$$D_i=\left(y-\overline{X}_i\right)' C\left(y-\overline{X}_i\right) \tag{2}$$

Where Di is the distance between the pattern vector y and the ith prototype, $i = 1,2,3,..., k$. *C* denotes the covariance matrix and $\overline{X_i}$ the mean vector of the ith class (Yeh *et al*., 2009).

Table 1. Use of k-NN in CVD Diagnosis

Author	Dataset	Feature extraction	Accuracy
Ros *et. al.* 2004	MIT ADB	Annotation file and fiducial points	92%
Christov *et al.* 2006	MIT ADB	Wavelet Packet	90.7%
Lanatá *et al.* 2011	MIT ADB	Bispectrum coefficients computed using FFT	92.1%
Kiranyaz *et al.* 2011	MIT ADB	Temporal features relating to heartbeat fiducial point intervals and morphology	99.04%
Yeh *et al.* 2012	MIT ADB	Difference operation method	94.30%
Mishra & Raghav 2010	MIT ADB	Local fractal dimension	88.64%
Kutlu & Kuntalp 2011	MIT ADB	Morphological features, FFT and higher order statistics of WPD	85.59%
Chen *et al.* 2013	PTB Dataset	Spectral energy	78.35%
Martis *et al.* 2013	MIT ADB	Higher order spectra	99.50%
Giri *et al.* 2013	Acquired Dataset	Linear discrimnants of DWT coefficients	87.5%
Acharya *et al.* 2013	Acquired Dataset	DWT, HOS and Texture based	78.45%
Wang *et al.* 2013	MIT Media Lab	Trend-based feature generation	97.78%

The two main reasons for adopting k-NN algorithm for ECG analysis: when there is large number of variables, k-means is time and computationally efficient than hierarchical clustering (if K is small) and produces tighter clusters than hierarchical clustering methods in this problem domain. These as justified by the work of Ros *et al*. (2004) used the MIT dataset and extracted 27 parameters depending on the ECG signal using the annotations and derived five other features from extracted values. Extracted features are then preprocessed to remove parameter redundancy using cross-correlation and supplied to k-NN classification process. The number of clusters was set to different values and results were compared for different number of clusters. A procedure to automatically detect the initial parameters was also proposed and a classification rate of up to 92% was obtained. Christov *et al*. (2006) compared two different feature extraction processes and clustered the extracted features using k-NN clustering. Morphological descriptors and time-frequency based descriptors were used. Although same clustering algorithm was employed on both the feature sets for the same MIT-BIH dataset, the clustering accuracy was higher for Time –frequency descriptors.

Lanatá *et al*. (2011) used features extracted from high order spectra and a nearest neighbour classification based on the Euclidean distance measure. The results were compared with a simple vector distance classifier (VDC) and indicated that it showed very small differences between VDC and classic K-NN in terms of sensitivity. Kiranyaz *et al*. (2011) adopted the Pan-Tompkins algorithm (Pan & Tompkins 1986) for feature extraction using the fiducial points annotated in the MIT-BIH arrhythmia database. A temporal segmentation phase, partitions the entire data into homogenous time segments that can be represented by minimal amount of key-beats. Then the two-pass exhaustive K-means phase using Euclidean distance was employed. It first extracts the key-beats and then the master key-beats among them. Kiranyaz *et al*. (2011) mentioned that the performance of the method depends on the initial mean values, as the method converges to the closest local optima and that the method is also dependent on the data distribution.

Yeh *et al*. (2009) used difference operation method (DOM) for the detection of QRS complex followed by range overlaps method for qualitative feature selection and the use of Mahalanobis distance measure for the k-nearest neighbour classification. Mishra & Raghav, (2010) used local fractal based feature extraction and employed two different distance measures for classifying the extracted features using the nearest neighbour classification. He compared Geometrical separability index (GSI) and Bhattacharya distance measures (Given in Eq 3, 4).

Geometrical separability index (GSI) or Thornton's separability index:

$$GSI = \frac{\sum_{i=1}^{N}\left[\left(f\left(x_i\right)+f\left(x_{iNN}\right)+1\right)\bmod 2\right]}{N} \tag{3}$$

Where x_i is a sample and x_{iNN} is its nearest neighbour, f is the binary decision classification function and N is the total number of samples.

Bhattacharya distance:

$$BD = \frac{1}{8}\left(M_1 - M_2\right)^T \left[\frac{\sum_1 - \sum_2}{2}\right]^{-1} \left(M_1 - M_2\right) + \frac{1}{2}\left(\ln\left(\frac{\left|\frac{\sum_1 - \sum_2}{2}\right|}{\sqrt{\left|\sum_1\right|\left|\sum_2\right|}}\right)\right) \tag{4}$$

Where BD is the distance between the Gaussian distributed data cluster 1and 2, M_1, M_2 are cluster means, $\sum_1$ and $\sum_2$ are covariance matrices of cluster one and two (Mishra & Raghav, 2010).

Kutlu & Kuntalp (2011) used a diverse set of features including higher order statistics, morphological features, Fourier transform coefficients, and higher order statistics of the wavelet package coefficients extracted from each different type of ECG beat. Genetic algorithm is used for finding the optimal combination of heart beat features, which give the best discrimination of a heart beat type from all other heart beats. The selected features are supplied to a nearest neighbour classifier with the Euclidean distance metric. The system detected 12 types of arrhythmia from the MIT-BIH arrhythmia database.

A low-power on-body classifier was designed by Chen *et al*. (2013) for remote monitoring of ECG signals. The spectral energy features per ECG beat are extracted and selected using Fisher's feature selection criterion. The samples are then clustered in the reduced feature space using SVM and kNN. The higher order spectral features were also extracted by Martis *et al*. (2013). The discriminative features selected using independent component analysis (ICA), and multiple classifiers were explored for validating the discriminative power of the extracted set of features. The study included classification and regression trees (CART), random forest, kNN and artificial neural networks (ANN). The study revealed that kNN outperformed the other algorithms for the particular set of selected features.

As the resolution of the extracted features is crucial, Giri *et al*. (2013) extracted the Discrete wavelet coefficients. Three different feature extraction methods viz. principal component analysis (PCA), independent component analysis (ICA) and linear discrimnant analysis (LDA) and four classifiers viz. SVM, kNN, gaussian mixture model (GMM), probabilistic neural networks (PNN) were employed. The n-to-n combination of these feature selection with classifiers were analyzed. kNN showed a robust performance, independent of the feature selection algorithm used. Discrete wavelet feature were extracted by Acharya *et al*. (2013) along with higher order spectral (HOS) coefficients and other texture based features. The extracted features were classified using six different classifiers such as decision tree, fuzzy sugeneo classifier, GMM, radial basis probabilistic neural networks (RBPNN), naïve bayes classifier and kNN.

Wang *et al*. (2013) applied kNN for estimating the driving stress from ECG signals where a 56 dimensional feature vector was generated by the trend-based feature generation. The dimensionality of the feature vector was reduced using PCA, LDA. The reduced feature vector was fed to kNN and the K-nearest neighbour classifier. An overall accuracy of 97.78% was achieved by this method. The other major areas in which kNN is used in CVD diagnosis include detection of QRS complexes (Saini *et al*., 2013) and for combining the decision of multiple classifiers, where kNN is used for consensus clustering (Abawajy *et al*., 2013).

In general, kNN is analytically tractable with a simple implementation, and has a high adaptive behaviour. Although the diagnosis accuracy depended greatly on the feature extraction, the classification accuracy was higher than the decision tree based classifiers ID3, C4.5 and CART (Refer Table 1). kNN comes with large storage requirements and is highly susceptible to the curse of dimensionality problem. Also, it is difficult to compare quality of the clusters produced and requires a cross-validation method to avoid different clusters outcomes because of different initial partitions. Precisely, the choice of distance metric (both non-weighted and weighted) and number of clusters, kNN can be easily moulded to build the exact diagnosis model.

Expectation maximization clustering (EM)

Expectation Maximization (EM) clustering is a parametric model based approach. It learns generative models from the statistical distribution of the data with each model corresponding to one particular cluster. It works in two steps; the expectation step and the maximization step. Initially it selects k random objects to represent the cluster centres and iteratively refines the parameters based on the cluster membership of the object calculated as in Eq 5.

$$p(x_i \in C_k) = p(C_k | x_i) = \frac{p(C_k) p(x_i | C_k)}{p(x_i)} \tag{5}$$

where $p(x_i|C_k)$= N(mk,E(xi)) is a Gaussian distribution with mean mk, expectation Ek.

$$m_k = \frac{1}{n} \sum_{1}^{n} \frac{x_i p(x_i \in C_k)}{\sum_j p(x_i \in C_j)} \tag{6}$$

The Maximization step uses the probability estimates and refines the model parameters based on the likelihood of the distributions of the data as given in Eq 6.

Martis *et al*. (2013) used the MIT-BIH and European ST-T Ischemia datasets and developed a simplified version of the Pan - Tompkins algorithm (Pan & Tompkins 1986) for QRS complex detection. Principal component analysis (PCA) was adopted for dimensionality reduction of 200 to 12 extracted features from the ECG signals. He compared the use of three different clustering techniques k-Means, Fuzzy c-Means, and expectation maximization clustering and explained that for a given feature set EM clustering provided the best performance of 94.29% of classification accuracy.

Sufi & Khalil (2011, *IEEE Trans Info Tech Biomed*) delved on the MIT-BIH dataset and proposed a novel lossless encoding algorithm designed exclusively for ECG compression and a method to extract features directly from the compressed ECG. The frequency of occurrence of each symbol contained in the alphabet used to encode the signal was taken as the feature set. These extracted features were then classified using EM clustering. The system claimed 38 times faster patient identification from compressed ECG. Sufi & Khalil (2011, *J Net Comp App*) expanded the work for heart disease diagnosis from compressed ECG using EM classifier on features extracted from compressed ECG and achieved a classification accuracy of 97%. Giri *et al*. (2013) extracted the Discrete wavelet coefficients for capturing both frequency and time information of the ECG signal. Three different feature extraction methods viz. principal component analysis (PCA) and Linear Discrimnant analysis (LDA) and four classifiers viz. SVM, kNN, Gaussian mixture model (GMM), Probabilistic neural networks (PNN) were employed. All combinations of feature selection methods with classifiers were analyzed. EM outperformed other algorithms for extracted HRV features selected using ICA. Acharya *et al*. (2013) extracted DWT, HOS and Texture based features. The Statistical p-value was used to select the discriminative features. The extracted features were classified using six different classifiers such as decision tree, fuzzy sugeneo classifier, EM, radial basis probabilistic neural networks (RBPNN), naïve bayes classifier and kNN. For the combination of feature extraction methods and selected feature set EM outperformed other algorithms.

Table 2. Use of EM in CVD diagnosis

Author	Dataset	Feature Extraction	Accuracy
Martis *et al.* 2009	MIT ADB and European ST-T Ischemia datasets	Principal component Analysis	94.29%
Sufi & Khalil 2011 (*IEEE Trans Info Tech Biomed*)	MIT ADB	Novel algorithm	>90%
Sufi & Khalil 2011 (*J Net Comp App*)	MIT ADB	Novel algorithm	97%
Giri *et al.* 2013	Acquired Dataset	DWT	96.8%
Acharya *et al* . 2013	Acquired Dataset	DWT, HOS, Texture based	100%

In general, EM produces fast, robust and unbiased clusters. The major advantages of EM clustering are its capability to handle high dimensional data and faster convergence with a good initialization (Cheng *et al*., 2010). Table 2 shows the usage of EM in CVD diagnosis methods. Martis *et al*. (2009) compared the use of kNN and EM for the same data set and same set of extracted features and proved that for a given extracted feature set EM clustering provides higher performance than kNN. EM clustering efficiently deals with missing or erroneous dataset, therefore used in the CVD diagnosis process where errors are indispensible in ECG acquisition and feature extraction processes (Xu & Wunsch, 2010).

Classification and regression trees (CART)

The CART methodology represents a unification of all tree-based classification methods. The basic idea behind CART is that "Inside every big tree is a small, perfect tree waiting to come out" - Dan Steinberg, 2004 CAS P.M. Seminar. CART is a nonparametric technique that produces classification or regression trees, depending on whether the dependent value is categorical or numeric, respectively (Breiman *et al*., 1984). The splitting criterion is sum of squared errors for Regression trees (numeric values) or Gini measure (Given in Eq 7) or Twiong rules (Given in Eq 8) for Classification (Categorical values). The main advantages of CART are its easy interpretability and its capability to classify both numerical and categorical data.

Gini criteria:

$$ArgMax \rightarrow P_l \sum_{j=1}^{k} p^2\left(\frac{j}{t_l}\right) + P_r \sum_{j=1}^{k} p^2\left(\frac{j}{t_r}\right) \tag{7}$$

Twiong criteria:

$$ArgMax \rightarrow \frac{P_l P_r}{4}\left[\sum_{j=1}^{k}\left\| p\left(\frac{j}{t_l}\right) - p\left(\frac{j}{t_r}\right)\right\|\right] \tag{8}$$

Where P_l, P_r are probability to get left and right nodes, *j* (1 . . .K) is the class index, k is the number of classes in a sample *p*(*j* | *t*) is the probability to have class *j* given at node *t*.

Pecchia *et al*. (2011, IEEE Trans Info Tech Biomed) investigated the significance of short-term heart rate variability (HRV) features in classifying CVD patients according to disease severity, by using CART. Since the study mainly aimed at distinctively identifying the key features that caused chronic heart failure in two RR interval databases, one with normal middle-aged subjects, the other with patients suffering from CHF were used. The data of normal subjects was retrieved from the normal sinus rhythm RR interval database (NSRDB) (Goldberger *et al*., 2000). The data for the CHF group was retrieved from the congestive heart failure RR interval database. The overall accuracy obtained was 89.7%. The limitation of the study specified by the author was the use of small dataset and the unbalanced ratio of the number of normal subjects on the total number of subjects in the dataset. Further, Pecchia *et al*. (2011, *IEEE Trans Biomed Eng*) revised the work to evolve a telemedicine platform with advanced functionalities for remote health monitoring of patients suffering from heart failure. The telemedicine platform used a two-stage model for early disease diagnosis. The model first detected heart failures if any and then severity was assessed, with CART algorithm applied at both stages.

Bukkapatnam *et al*. (2008) analyzed ECG recordings of atrial fibrillation, a common arrhythmia, collected from 60 different subjects posted on the AF termination challenge database (Moody, 2004). Multi-resolution analysis using wavelets and principal component analysis methods were adopted for feature extraction and dimensionality reduction of the extracted features. CART was employed for analyzing extracted features and decision rules learned using a training data set to classify different groups of objects. Bukkapatnam *et al*. (2012) investigated spatio-temporal patterns of the Vector Cardiogram (VCG) signals for the identification of various types of heart disease. The 3D cardiac VCG signal was initially partitioned into eight octants. Three feature groups: Conventional ECG features, VCG vector features and VCG octant features were extracted using the PhysioNet PTB database (Bousseljot *et al*., 1995). CART analysis was used to demonstrate that VCG octant features can distinguish diseased from healthy subjects and suggested that the approach provides an effective way for interpretation of ECG data and to influence the current clinical cardiac diagnostic practice.

Table 3. Use of CART in CVD diagnosis

Author	Dataset	Feature extraction	Accuracy
Pecchia *et al.* 2011(*IEEE Trans Info Tech Biomed)*	Normal sinus rhythm RR interval database(NSR DB) and congestive heart failure RR interval database	Principal component Analysis (PCA)	89.7%
Pecchia *et al.* 2011(*IEEE Trans Biomed Eng)*	NSR DB and congestive heart failure RR interval database	PCA	96.39%
Sathyadevi, 2011	UC-Irvine archive of machine learning datasets (UCI Machine Learning Repository)	Attribute valued dataset (AVD)	64.8% 71.4% 83.2%
Bukkapatnam *et al.* 2008	AF Termination Challenge Database	Wavelet analysis	90%
Bukkapatnam *et al.* 2012	PhysioNet PTB database(VCG)	Wavelet analysis	95.00%
Krivokapich *et al.* 1999	Data collected from 1183 patients	AVD	--
Roche *et al.* 2003	Holter ECG from 147 patients	Wavelet Analysis	90.1%
Glickman *et al.* 2012	3,575,178 patients	AVD	91.9%
Macek 2005	MIT-BIH arrhythmia Dataset and PTB Diagnostic ECG aatabase	Wavelet Analysis	--
Fayn, 2011	Acquired from 90 patients	Spatiotemporal CAVIAR serial ECG analysis method	93.8%

Krivokapich *et al*. (1999) adopted CART to analyze ECG Samples of 1,183 patients. The medical records included Dobutamine stress echocardiography (DSE) and ECG. CART analysis was used to identify features that best predicted future cardiac events. Roche *et al*. (2003) employed CART analysis on ECG features extracted using Wavelet analysis methods. A decision tree was constructed using all levels of wavelet coefficients with a classification accuracy of 90.1% was obtained.

Glickman *et al*. (2012) adopted CART for analyzing the attribute valued medical data from 3,575,178 patients and constructed a decision tree to identify, if the patient entering the Emergency department required a 12-lead ECG monitoring. Macek (2005) created an ensemble classifier for incremental learning from ECG. The method used combined perceptron decision tree architecture. The results obtained by the ensemble classifier was compared with C5.0 and confirmed the use of ensemble classifier was more efficient. Fayn (2011) developed a novel decision tree based classifier T-3C, Tree algorithm based on conditions combinations competition, and compared with the chi-squared automatic interaction detection (CHAID); and CART algorithms. The features are extracted using the spatiotemporal CAVIAR serial ECG analysis method. The classification was based on the features of a reference beat stored for the patient to segment the acquired beat sample.

Decision tree based classifiers comes with many advantages including detecting outliers and applicability, even when there is little or no domain knowledge. But a tree can be sub optimal in many situations, which is not acceptable in healthcare applications. The CART approach facilitates the interpretation with effective creation of clinical decision rules and an effective means to combine physiological knowledge base with ECG analysis systems. Table 3 shows different scenarios in which CART was applied for CVD diagnostics. To prevent "overfitting" problems, k-fold cross validation can be used to find a generalized, optimal, and simpler tree structure.

Time series analysis

Time series data is a stream of data points or recordings, each with a time stamp where data points are 'chronologically' ordered (Fu, 2011). In real world, all applications such as financial, healthcare, astronomical etc., the data is inherently temporal. Time series analysis includes the impact of the time factor on the predictive results. ECG being a series of data recorded against time, time series analysis can be applied to diagnose or predict different cardiac diseases.

Nikolopoulos *et al*. (2003) studied linear time series methods (Fourier transforms and Autocorrelation function) along with non-linear time series analysis methods, such as Discrete wavelet transforms and approximate entropy method, and found than non-linear methods are more immune to noise and produced higher classification accuracy even in the presence of noise. He indicated the requirement of further investigation of more robust methods and the combined use of methods. Further Ashkenazy *et al*. (2003)

explored on the magnitude and sign sub series and indicated that the magnitude sub series of the original time series exhibited non-linear behaviour and the sign sub series exhibited linear behaviour. He pointed out that the magnitude sub series of ECG time series data has more discriminative power with respect to cardiac disease diagnosis.

Ge *et al.* (2002) used the generalized linear model based algorithm on the extracted auto regression (AR) coefficients for ECG time series data. The AR coefficients were computed using Burg's algorithm and classified using a generalized linear model (GLM) based algorithm in various stages. The method gives reasonably high accuracies but requires further validation for clinical implementation. Kalpakis *et al*. (2001) suggested the use of Cepstral coefficients over discrete fourier transform (DFT), DFT of Auto-correlation function (ACF), discrete wavelet transform (DWT), PCA measures. Corduas & Piccolo (2008) further investigated on the different Autoregressive models and derived an efficient approximation of squared AR distance and applied it for the ECG time series data.

Table 4. Use of Time Series Analysis in CVD Diagnosis

Author	Dataset	Feature extraction	Classifier
Nikolopoulos *et al.* 2003	20 Samples collected from hospital in controlled environment (Patients selected by a cardiologist)	DWT, Approximate entropy method, Single Value Decomposition	Time series analysis
Ashkenazy *et al.* 2003	MIT-BIH NSRDB and BIDMC congestive heart failure dataset	Decomposition into magnitude and sign subseries	Time Series Analysis
Ge *et al.* 2002	MIT-BIH ADB and MIT-BIH NSRDB	Autoregressive coefficients	Generalized linear model (GLM) based algorithm
Kalpakis *et al.* 2001	MIT-BIH ADB	Linear predictive coding (LPC) – cepstral coefficients	ARIMA time–series Analysis- Novel distance measure
Corduas & Piccolo 2008	MIT-BIH ADB	Autoregressive coefficients	Time series analysis + Novel of squared AR distance measure
Fuchs *et al.* 2009	ECG200 dataset	Polynomial Least-Squares Approximations	Time series analysis + probabilistic modeling + Similarity measurement techniques
Fuchs *et al.* 2010	MIT-BIH ADB and European ST-T database	SwiftSeg (Polynomial least-squares approximations)	Time series analysis
Fisch *et al.* 2011	ECG dataset from Carnegie Mellon University	Piecewise polynomial representation	Time series analysis + Novel Segment Similarity measurement techniques

Fuchs *et al*. (2009) developed a piecewise probabilistic representation of time series data called SwiftMotif and also proposed a novel and very fast segmentation of time series data called SwiftSeg (Fuchs *et al*., 2010). Fisch *et al*. (2011) using these methods and derived a time series classification rule mining algorithm – SwiftRule. This algorithm unlike other time series algorithms that derived association rules in two passes, used a single pass for segmentation and subsequence matching. The classification rules thus derived a classification accuracy of 92% was achieved with on-line response times.

DISCUSSION

Medical diagnosis is a complex process that is too intricate to represent in an algorithmic model. Not only does medical diagnosing require the understanding of symptoms, drug-drug interactions, and patient history, the diagnosing process requires knowledge of diseases in general as well as the general population. Furthermore, it is imperative that the diagnosing systems provide reasoning for the medical diagnosis provided. Such a process would allow the physician to understand the reasons the system may have had for a specific decision that may have been made. The development of technology that has led to enhanced health care delivery, particularly the decision support aspect of the clinical diagnosis. But the acceptance of computer based diagnosis heavily deponds on physicians. It concerns the patient-physician relationship, the quality of patient care, the balance between clinical guidelines and decision support technology, and physician autonomy.

The integration of data mining techniques with cardiovascular disease diagnosis practices constitutes an important portion of an emerging and rapidly developing field of biomedical informatics. The diagnosis of cardiac disease diagnosis is gaining acceptance from the health care domain. This integration will greatly benefit both fields and aid in providing solutions to long-standing challenges.

Data characteristics

One major predicament concerning automatic ECG analysis is the immense variations in the morphologies of ECG waveforms of different patients and patient groups. An ECG analysis system, which operates well for a given training database often declines regrettably, when presented with a different ECG waveform. Such unpredictability in performance is a major barrier preventing use of clinical ECG processing systems. A remarkable perception is that all DM based CVD diagnosis applications reported in literature used preprocessed data sets. Although some studies reported the use of real time data, others typically used a preprocessed data set with at most 400 samples. Even though effectiveness of most DM methods depends on the size of the data set, this has not been reported as a concern in the applications based on small data.

Features extracted from ECG waveform

The difference in the accuracy of the classifier is mainly due to variation in the feature extraction process. Methods to exactly extract features from the ECG signal are heavily researched. Health care personals use time domain features from the ECG signal for disease diagnosis. Although time domain methods are available none of the methods have extracted more than three features from the ECG signal sufi & Khalil (2011). Time-frequency domain methods are preferred over spatial domain methods for efficient feature extraction process. But the methods do not exploit the domain knowledge for diagnostic purposes and also involve higher computational complexities. Time series analysis offer a hand in exploring the features from ECG signal as they natively represent the ECG signal accommodating its very time stamped nature. More efficient feature extraction methods provide a line for further research. Methods with higher accuracy, lower computation requirements, and lower time complexity are the need of the hour.

Data mining application to CVD diagnosis

It has been eminent that there is an increasing drift in the use of DM algorithms for CVD diagnosis. The major factors that impact the use of DM are scalability, dimensionality, robustness, number of clusters, arbitrary cluster shapes, good visualization and interpretation. Out of which interpretation of the results, reasoning for the predictive result and understandability of the results by health care personnel probe the use of DM in CVD diagnosis. Because of the need for a real time diagnosis, clustering based methods provide better results than other methods (Refer Table 5). Among clustering algorithms reviewed, EM shows high accuracy with fast execution times. Considering the accuracy and the chances for the solution becoming suboptimal by using CART further increases the use of clustering based systems. Extensions to the basic clustering algorithm provide better accuracy and therefore implementations based on fuzzy c-means, extended k-means and variations of EM are increasing in the healthcare domain. Also time series methods are on the rise and provide promising direction for further research and development. Time series methods could model the ECG in their natural form for more efficient disease diagnosis purposes, but the accuracy of diagnosis has not been reported thus far. Considerable progress has been made in CVD diagnosis throughout the years. However, no promising real-time method has been proposed yet. This is mainly due to the fact the real-time systems require incorporation of a variety of domain knowledge and the real-time hardware and software constraints. Hence probing for better methods to process the ECG signal with lower computational costs and lower time complexity are future direction of work.

Different standards for presenting diagnosis results

The MIT BIH Arrhythmia dataset (MIT-ADB) was used as the benchmark for presenting the results of different research methods. The MIT- ADB contains 48 half hour intercepts from 24-hour ECG Holter recording from 47 patients aged 23 to 89 (both men and women). The ECG was digitized 360 Hz with 11-bit resolution over a ±5 mV range. The MIT provides both beat annotations and rhythm annotations for these records. While the former provides annotation of every single beat categorized into 20 different beat types, the latter considers a sequence of beats as a rhythm and the rhythm annotations are categorized into 15 types. Most of the CVD diagnosis methods present a pattern recognition algorithm to classify individual beats and compare with the standard beat annotations provided by MIT. Few methods are presented for rhythm analysis and few other methods presented the diagnosis result as the type of arrhythmia present in ECG sample. To bridge these gaps, Association for the Advancement of Medical Instrumentation (AAMI) developed a standard involving five beat classes and mapped 16 different beat types from MIT-ADB into five of the AAMI beat classes. It is worth notable that none of the algorithms presented thus far in literature have attained 100% accuracy for classifying the five AAMI beat classes.

Table 5. Use of data mining in CVD diagnosis

Algorithm	Authors	Average Accuracy
K Nearest Neighbor	Ros *et. al.* (2004); Christov *et al.* (2006); Lanatá *et al.* (2011); Kiranyaz *et al.* (2011); Yeh *et al.* (2012); Mishra & Raghav (2010);Kutlu & Kuntalp (2011); Chen *et al.* (2013); Martis *et al.* (2013); Giri *et al.* (2013); Acharya *et al.* (2013); Wang *et al.* (2013)	90.32%
Expectation Maximization Clustering	Martis *et al.*(2009); Sufi & Khalil (2011, *IEEE Trans. Info.Tech.Biomed*) Sufi & Khalil (2011, *J.Net. Comp.App*); Giri *et al.* (2013); Acharya *et al* . (2013)	95.62%
CART	Pecchia *et al.* (2011,*IEEE Trans Info Tech Biomed*) Pecchia *et al.* (2011,*IEEE Trans Biomed Eng*) Sathyadevi (2011); Bukkapatnam *et al.* (2008) Bukkapatnam *et al.* (2012); Krivokapich *et al.* (1999); Roche *et al.* (2003); Glickman *et al.* (2012); Macek, (2005); Fayn (2011)	86.63%
Time Series analysis	Nikolopoulos *et al.* (2003); Ashkenazy *et al.* (2003); Ge *et al.* (2002); Kalpakis *et al.* (2001); Corduas & Piccolo (2008) Fuchs ***et al***. (2009); Fuchs *et al.* (2010); Fisch *et al.* (2011)	--

Other directions for future research

The feature extraction technique needs to be highly accurate and should ensure fast extraction of features from the ECG signal. Real data from Hospitals and other health care organizations needs to be collected and all the available techniques should be compared for the optimum accuracy and processing time in real-time environment. The analysis of ECG data collected in a geographic area could provide a great insight into the ECG patterns occurring in that area, thus offering more efficient diagnosis with much lower response times. This study can guide researchers and software producers in their effort to develop/further improve the methods and tools, by providing them with significant information on typical characteristics of ECG data collected, necessary/most preferred DM functions and methods, and expected results.

CONCLUSIONS

This paper provides a comprehensive overview of ECG processing and the use of low computationally feasible data mining algorithms used in CVD diagnosis proposed thus far in literature. Our survey was based on factors that will assume an increasing level of significance for future telecardiology systems and are also of importance to our research into low computational feature extraction and classification algorithms for ECG based CVD diagnosis. Different ECG acquisition methods, compression methods with a different dimension on categorization of existing algorithms, feature extraction methods and impact of different data mining methods in real time, resource-critical applications were presented. The different existing standards for presenting the results of ECG classification problem were also included. Based on the review criteria, a number of limitations in every stage of ECG processing have come to light. The study clearly indicates the importance and need for accurate feature extraction process. It is noteworthy that methods for incorporating expert knowledge into the classification process remain unexplored. Of the existing methods analyzed, EM clustering based on features extracted in spatial domain confirms more precise medical diagnosis. As a future direction of work, different spatial domain feature extraction methods have to be examined and validated with real-time data collected from health centre in multiple states of the patient including rest, emotional stress and physical exertion. When these methods are used in either time-critical or resource-critical environments, robustness in classification amidst acquisition discrepancies has to be validated.

Ethical issues (if any)

Ethical clearance obtained.

Conflict of interest

None Declared.

ACKNOWLEDGEMENTS

The authors thank Dr. Moses K. Daniel, Professor and Head, Dept of Medicine and Dr. Naina Mohammed, Assistant Professor, Dept of Cardiology, Government Rajaji Hospital, Madurai, India for their valuable support throughout the study. . This work is in part supported by All India Council for Technical Education (AICTE) under the Research Promotion Scheme – Reference No: 20/AICTE/RIFD/RPS/(POLICY-II)66/2012-13.

REFERENCES

Abawajy, J.H., Kelarev, A.V. & Chowdhury, M. 2013. Multistage approach for clustering and classification of ECG data, Computer Methods and Programs in Biomedicine, **112**:720–30.

Abenstein, J.P. & Tompkins, W.J. 1982. A new data reduction algorithm for real time ECG analysis. IEEE Transactions on Biomedical Engineering, BME **29**(1):43–8.

Acharya, U.R., Faust, O., Kadri, N A., Suri, J S. & Yu, W., 2013. Automated identification of normal and diabetes heart rate signals using nonlinear measures, Computers in Biology and Medicine, **1**;43(10):1523-9.

Addison, P.S. 2005. Wavelet transforms and the ECG: a review, Physiological Measurement, **26**:R155 199.

Alkoot, F.M. 2014. Mltimodal biometric authentication using adaptive decision boundaries, Kuwait Journal of Science, **41**(3):103-27.

Andreão, R.V., Muller, S.M., Boudy, J., Dorizzi, B., Bastos-Filho, T.F. & Sarcinelli-Filho, M. 2008. Incremental HMM training applied to ECG signal analysis. Computers in Biology and Medicine, **38**:659-67.

Ashkenazy, Y., Havlin, S., Ivanova, P.C., Peng, C.K., Schulte-Frohlinde, V H. & Stanley E. 2003. Magnitude and sign scaling in power-law correlated time series, Physica A: Statistical Mechanics and its Applications, **323**:19 – 41.

Babloyantz, A. & Maurer, P. 1996. A graphical representation of local correlations in time series - Assessment of cardiac dynamics, Physics Letters A, **221**:43-55.

Batista, L.V., Melcher, E.U. & Carvalho, L.C., 2001. Compression of ECG signals by optimized quantization of discrete cosine transform coefficients, Medical Engineering & Physics, **23**(2):127–34.

Benzid, R., Messaoudi, A. & Boussaad, A. 2008. Constrained ECG compression algorithm using the block-based discrete cosine transform, Digital Signal Processing, **18**:56–64.

Bousseljot, R., Kreiseler, D. & Schnabel, A. 1995. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet, Biomedizinische Technik, Band 40, Ergänzungsband, **40**:317–8.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., 1984. Classification and Regression Trees. Wadsworth Int. Group

Bukkapatnam, S., Komanduri, R., Yang, H., Rao, P., Lih, W.C. & Malshe, M. 2008. Classification of atrial fibrillation episodes from sparse electrocardiogram data, Journal of Electrocardiology, **41**:292–9.

Bukkapatnam, S., Yanga, H., Leb, T. & Komanduri, R. 2012. Identification of myocardial infarction MI) using spatio-temporal heart dynamics, Medical Engineering & Physics, **34**(4):485-97.

Castro, B., Kogan, D. & Geva, A.B. 2005. ECG feature extraction using optimal mother wavelet. The 21st IEEE Convention of the Electrical and Electronic Engineers in Israel, 346-50.

Ceylan, R., Özbay, Y. & Karlik, B. 2010. Telecardiology and Teletreatment System Design for Heart Failures Using Type-2 Fuzzy Clustering Neural Networks, International Journal of Artificial Intelligence and Expert Systems, 1(4):100-10.

Chawla, M.P.S. 2007. Parameterization and R-peak error estimations of ECG signals using independent component analysis, Computational and Mathematical Methods in Medicine, **8**(4):263–85.

Chen, T., Mazomenos, E.B., Maharatna, K., Dasmahapatra, S. & Niranjan, M. 2013. Design of a Low-Power On-Body ECG Classifier for Remote Cardiovascular Monitoring Systems, IEEE Journal of Emerging Selected Topics in Circuits and Systems, **3**(1):75-85.

Cheng, Z., Yu, P.S. & Bell, D. 2010. Introduction to the domain-driven data mining. Special section, IEEE Transactions on Knowledge and Data Engineering, **22**(6):53-4.

Christov, I., Gómez-Herrero, G., Krasteva, V., Jekova, I., Gotchev, A. & Egiazarian, K. 2006. Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification, Medical Engineering and Physics, **28**:876-87.

Corduas, M. & Piccolo, D. 2008. Time series clustering and classification by the autoregressive metric, Computational Statistics and Data Analysis, **52**:1860 – 72.

Cox, J.R., Nolle, F.M., Fozzard, H.A. & Oliver, G.C. 1968. AZTEC. A preprocessing program for real time ECG rhythm analysis, IEEE Transactions on Biomedical Engineering, BME-**15**(4):128–9.

DiPersio, D.A. & Barr, R.C. 1985. Evaluation of the FAN method of adaptive sampling on human electrocardiograms, Medical & Biological Engineering & Computing, 23(5):401–10.

Dumont, J., Hern´andez, A.I., Fleureau, J. & Carrault, G. 2008. Modeling temporal evolution of cardiac electrophysiological features using Hidden Semi-Markov Models, Annual Intl. Conf. of the IEEE Engg in Med. and Biology Society: Personalized Healthcare through Technology.

Fayn J. 2011. A Classification Tree Approach for Cardiac Ischemia Detection Using Spatiotemporal Information From Three Standard ECG Leads, IEEE Transactions on Biomedical Engineering, **58**(1):95-102.

Fisch, D., Gruber, T. & Sick, B. 2011. SwiftRule: Mining Comprehensible Classification Rules for Time Series Analysis, IEEE Transactions on Knowledge and Data Engineering, **23**(5):774-87.

Fu, T C . 2011. A review on time series data mining, Engineering Applications of Artificial Intelligence,e, **24**:164 –181.

Fuchs, E., Gruber, T., Nitschke, J. & Sick, B. 2009. On-Line Motif Detection in Time Series with SwiftMotif, Pattern Recognition, **42**(11):3015-31.

Fuchs, E., Gruber, T., Nitschke, J. & Sick, B. 2010. Online Segmentation of Time Series Based on Polynomial Least-Squares Approximations, IEEE Transactions on Pattern Analysis and Machine Intelligence, **32**(12):2232-45.

Ge, D., Srinivasan, N. & Krishnan, S.M. 2002. Cardiac arrhythmia classification using autoregressive modeling, BioMedical Engineering OnLine, doi: **10**.1186/1475-925X-1-5

Giri, U.D., Acharya, R., Martis, R.J., Sree, S.V., Lim, T.C., Ahamed V.I.T & Suri, J.S. 2013. Automated diagnosis of Coronary Artery Disease affected patients using LDA, PCA, ICA and Discrete Wavelet Transform, Knowledge Based Systems, **37**:274-82.

Glickman, S.W., Shofer, F.S., Wu, M.C., Scholer, M.J., Ndubuizu, A. & Peterson, E.D. 2012. Development and validation of a prioritization rule for obtaining an immediate 12-lead electrocardiogram in the emergency department to identify ST-elevation myocardial infarction, American Heart Journal, **163**(3):372-82.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.Ch., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C-K. & Stanley, H.E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, Circulation,

101(23):e215-e220

Goudarzi, M.M., Moradi, M.H. & Taheri, A. 2005. Efficient Method for ECG Compression Using Two Dimensional Multiwavelet Transform, Proceedings of Word Academy of Science, Engineering and Technology, **2**:10-4.

Hailey, D., Ohinmaa, A. & Roine, R. 2004. Evidence for the benefits of telecardiology applications: a systematic review, Edmonton, Alberta Heritage Foundation for Medical Research, **34**:1- 60.

Health Informatics. 2005. Standard Communication Protocol. Computer assisted Electrocardiography, British-Adopted European Standard BS EN 1064, 2005.

Iliopoulos, C.S. & Michalakopoulos, S, 2010. Combinatorial ECG Analysis for Mobile Devices. Proc. of MIR'10, Philadelphia, USA, March 2010,29-31.

Iliopoulos, C.S. & Michalakopoulos, S. 2011. A Combinatorial Model for ECG Interpretation, International Journal of Biological Sciences, **7**:10-14.

Iwata, A., Nagasaka, Y. & Suzumura, N. 1990. Data compression of the ECG using neural network for digital Holter monitor, IEEE Engineering in Medicine & Biology, **9**(3):53–7.

Jalaleddine, M.S., Chriswell, G.H., Strattan, R.D. & Coberly, W.A. 1990. ECG Data Compression Techniques-A Unified Approach, IEEE Transactions on Biomedical Engineering, 37(4):329-43. Saini, I., Singh, D. & Khosla, A. 2013. QRS detection using K-Nearest Neighbor algorithm KNN) and evaluation on standard ECG databases, Journal of Advanced Research, **4**:331–44.

Jumaa, H., Fayn, J. & Rubel, P., 2008. XML based mediation for automating the storage of SCP-ECG data into relational databases, Computers in Cardiology, **35**:445-8.

Kalpakis, K., Gada, D. & Puttagunda, V. 2001. Distance measures for effective clustering of ARIMA time series, Proc. IEEE Int. Conf. Data Mining, 273–80.

Karpagachelvi, S., Arthanari, M. & Sivakumar, M. 2010. ECG Feature Extraction Techniques - A Survey Approach, International Journal of Comput Science and Information Security, **8**(1):76-80.

Khashei, M., Hamadani, A Z. & Bijari, M. 2012. A novel hybrid classification model of artificial neural networks and multiple linear regression models, Expert Systems with Applications, **39**(3):2606–20.

Kiranyaz, S., Ince, T., Pulkkinen, J. & Gabbouj, M. 2011. Personalized long-term ECG classification: A systematic approach, Expert Systems with Applications, **38**:3220–6.

Krivokapich, J., Child, J S., Walter, D O. & Garfinkel, A. 1999. Prognostic Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients with Known or Suspected Coronary Artery Disease, Journal of the American College of Cardiology, **33**(3):708-16.

Ku, C.T., Hung, K.C., Wu, T.C. & Wang, H.S. 2010. Wavelet-Based ECG Data Compression System with Linear Quality Control Scheme, IEEE Transactions on Biomedical Engineering, **57**(6):1399-409.

Kumar, D.S., Sathyadevi, G. & Sivanesh, S. 2011. Decision Support System for Medical Diagnosis Using Data Mining, International Journal of Computer Science Issues, **8**(3):147-53.

Kumar UA, 2013: http://www.indianexpress.com/India-has-just-one-doctor-for-every-1700-peopleKutlu, Y. & Kuntalp, D. 2011. A multi-stage automatic arrhythmia recognition and classification system, Computers in Biology and Medicine, **41**:37–45.

Lanatá, A., Valenza, G., Mancuso, C. & Scilingo, E P. 2011. Robust multiple cardiac arrhythmia detection through bispectrum analysis, Expert Systems with Applications, **38**:6798–804.

Lei, W.K., Li, B.N., Dong, M.C. & Fu, B.B. 2008. An Application of Morphological Feature Extraction and Support Vector Machines in Computerized ECG interpretation, Sixth IEEE Mexican International Conference on Artificial Intelligence, Special Session

Lin, C.T., Chang, K.C., Lin, C.L., Chiang, C.C., Lu, S.W. & Chang, S.S. 2010. An Intelligent Telecardiology System Using a Wearable and Wireless ECG to Detect Atrial Fibrillation, IEEE Transactions on Information Technology in Biomedicine, **14**(3):726-33.

Lu, Z., Kim, D.Y. & Pearlman, W.A. 2000. Wavelet compression of ECG signals by set partitioning in hierarchical trees algorithm, IEEE Transactions on Biomedical Engineering, **47**(7):849–56.

Luca, G.D., Suryapranata, H., Ottervanger, J.P. & Antman, E.M. 2004. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: Every minute of delay counts, Circulation, **109**:1223–5.

Macek, J. 2005. Incremental Learning of Ensemble Classifiers on ECG Data, Proc. of the 18th IEEE Symposium on Computer-Based Medical Systems.

Maglaveras, N., Stamkopoulos, T., Diamantaras, K., Pappas, C. & Strintzis, M. 1998. ECG pattern recognition and classification using non-linear transformations and neural networks: A review, International Journal of Medical Informatics, **52**:191–208.

Martis, R.J., Chakraborty, C. & Ray, A.K. 2009. A two-stage mechanism for registration and classification of ECG using Gaussian mixture model, Pattern Recognition, **42**(11):2979-88.

Martis, R.J., Acharya U.R., Prasad, H., Chua, C.K., Lim, C.M. & Suri, J.S. 2013. Application of higher order statistics for atrial arrhythmia classification, Biomedical Signal Processing and Control, **8**:888– 900.

Mehta, S.S. & Lingayat, N.S. 2008. Support Vector Machine for Cardiac Beat Detection in Single Lead Electrocardiogram, IAENG International Journal of Applied Mathematics, **36**(2):1-7.

Mehta, S.S. & Lingayat, N.S. 2009. Identification of QRS complexes in 12-lead electrocardiogram, Expert Systems with Applications, **36**:820–8.

Minas, A.K., Moutiris, J.A., Hadjipanayi, D. & Pattichis, C.S. 2010. Assessment of the risk factors of coronary heart Events based on data mining with decision trees, IEEE Transactions on Information Technology in Biomedicine, **14**(3):559-66.

Mishra, A.K. & Raghav, S. 2010. Local fractal dimension based ECG arrhythmia classification, Biomedical Signal Processing and Control, **5**:114–23.

Moavenian, M. & Khorrami, H. 2010. A qualitative comparison of Artificial Neural Networks and Support Vector Machines in ECG arrhythmias classification, Expert Systems with Applications, **37**: 3088–93.

Moody, G.B., Mark, R.G. & Goldberger, A.L. 1988. Evaluation of the "TRIM" ECG data compressor, Computers Cardiology, **15**:167-70.

Moody, G.B. & Mark, R.G. 2001. The Impact of the MIT-BIH Arrhythmia Database, IEEE Engineering in Medicine & Biology, **20**(3):45-50.

Moody, G.B. 2004. Spontaneous Termination of Atrial Fibrillation: A Challenge from PhysioNet and Computers and Cardiology, Computers in Cardiology, **31**:101-4.

Nikolopoulos, S., Alexandridi, A., Nikolakeas, S. & Manis, G. 2003. Experimental analysis of heart rate variability of long-recording electrocardiograms in normal subjects and patients with coronary artery disease and normal left ventricular function. Journal of biomedical informatics, **36**(3): 202-17.

Olvera, F.E. 2006. Electrocardiogram Waveform Feature Extraction Using the Matched Filter, Statistical Signal Processing, II,1-6.

Pan, H.S., & Tompkins, W.J. 1986. Quantitative Investigation of QRS Detection rules using the MIT BIH Arrhythmia Database, IEEE Transactions on Biomedical Engineering, BME-**33**(12), 1157-65.

Patel, V., Chatterji, S., Chisholm, D., Ebrahim, S., Gopalakrishna, G., Mathers, C., Mohan, V., Prabhakaran, D., Ravindran, R.D. & Reddy, K.S. 2011. Chronic diseases and injuries in India, The Lancet, **377**(9763): 413-28.

Patra, D., Das, M.K. & Pradhan, S. 2005. Integration of FCM, PCA and Neural Networks for Classification of ECG Arrhythmias, IAENG International Journal of Computer Science, **36**(3):1-5.

Pecchia, L., Melillo, P., Sansone, M. & Bracale, M. 2011. Discrimination Power of Short-Term Heart Rate Variability Measures for CHF Assessment, IEEE Transactions on Information Technology in Biomedicine, **15**(1):40-6.

Pecchia, L., Melillo, P. & Bracale, M. 2011. Remote health monitoring of heart failure with data mining via CART method on HRV features, IEEE Transactions Biomedical Engineering, 58(3):800-4.

Pooyan, M., Taheri, A., Moazami-goudarzi, M. & Saboori. 2005. Wavelet Compression of ECG Signals Using SPIHT Algorithm, World Academy of Science, Engineering and Technology, **2**:705-8.

Reddy, B.R.S. & Murthy, I.S.N. 1986. ECG data compression using Fourier descriptors, IEEE Transactions on Biomedical Engineering, BME **33**(4):428–34.

Roche, F., Pichot, V., Sforza, E., Court-Fortune, I., Duverney, D. & Costes, F. 2003. Predicting sleep apnoea syndrome from heart period: a time-frequency wavelet analysis, European Respiratory Journal, **22**:937–42.

Ros, E., Mota, S., Fernández, F.J., Toro, F.J. & Bernier, J.L., 2004. ECG Characterization of paroxysmal atrial fibrillation: parameter extraction and automatic diagnosis algorithm, Computers in Biology and Medicine, **34**:679–96.

Sathyadevi, G. (2011, June). Application of CART algorithm in hepatitis disease diagnosis. In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on (pp. 1283-1287). IEEE.

Shih, D.H., Chiang, H.S., Lin, B. & Lin, S.B. 2010. An Embedded Mobile ECG Reasoning System for Elderly Patients, IEEE Transactions on Information Technology Biomedicine, **14**(3):854-65

Steinberg, C.A., Abraham, S. & Caceres, C.A. 1962. Pattern Recognition in the Clinical Electrocardiogram, IRE Transactions on Bio-Medical Electronics, **9**(1): 23-30.

Sufi, F., Fang, Q., Khalil, I. & Mahmoud, S.S. 2009. Novel Methods of Faster Cardiovascular Diagnosis in Wireless Telecardiology, IEEE Journal on Selected Areas in Communications, **27**(4):537-52.

Sufi, F. & Khalil, I. 2011. Diagnosis of Cardiovascular Abnormalities from Compressed ECG: A Data Mining-Based Approach, IEEE Transactions on Information Technology in Biomedicine, **15**(1):33-9.

Sufi, F. & Khalil, I. 2011. Faster person identification using compressed ECG in time critical wireless telecardiology applications, Journal of Network and Computer Applications, **34**:282–93.

Tohumoglu, G. & Sezgin, E. 2007. ECG signal compression by multi-iteration EZW coding for different wavelets and thresholds, Computers in Biology and Medicine, **37**:173-82.

UCI Machine Learning Repository. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html. Accessed on 24 February 2015.

Wang, J.S., Lin, C.W. & Yang, Y.T.C. 2013. A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition, Neurocomputing, **116**:136–43.

Welch, T.A. 1984. A technique for high-performance data compression, IEEE Computer, **17**(6):8–19.

Womble, M.E., Halliday, J.S., Mitter, S.K., Lancaster, M.C. & Triebwasser, J.H. 1977. Data compression for storing and transmitting ECGs/VCGs, Proc. IEEE May, **65**(5):702–6.

Xu, R. & Wunsch, D.C. 2010. Clustering Algorithms in Biomedical Research: A Review, IEEE Reviews in Biomedical Engineering, **3**:120-54.

Yeh, Y.C., Wang, W.J. & Chiou, C.W. 2009. Heartbeat Case Determination Using Fuzzy Logic Method on ECG Signals, International Journal of Fuzzy Systems, **11**(4):250-61.

Yeh, Y.C., Chiou, C.W. & Lin, H.J. 2012. Analyzing ECG for cardiac arrhythmia using cluster analysis, Expert Systems with Applications, **39**:1000–10.

Yeragania, V.K. & Rao, R. 2003. Effect of nortriptyline and paroxetine on measures of chaos of heart rate time series in patients with panic disorder, Journal of Psychosomatic Research, **55**:507– 13.

Yildiz, A., Akın, M. & Poyraz, M. 2011. An expert system for automated recognition of patients with obstructive sleep apnea using electrocardiogram recordings, Expert Systems with Applications, **38**:12880-90.

Yoo, J., Yan, L., Lee, S., Kim, H. & Yoo, H.J. 2009. A Wearable ECG Acquisition System with Compact

Planar-Fashionable Circuit Board-Based Shirt, IEEE Transactions on Information Technology in Biomedicine, **13**(6):897-902.

Young, T.Y. & Huggins, W.H. 1963. On the Representation of Electrocardiograms, IEEE Trans Bio-medical Electronics, **10**(3):86-95.

Zheng, H., Wang, H.Y., Black, N.D. & Winder, R.J. 2010. Data structures, coding and classification, Technology and Health Care, **18**(1):71-87.

Submitted **:** 04/06/2014

Revised **:** 28/10/2014

Accepted **:** 29/10/2014

استطلاع حول خوارزميات عن البيانات المستخدمة في تشخيص أمراض القلب بواسطة قراءات تخطيط القلب متعدد الأقطاب

*** ديانا موسى، **ديزي س.**

قسم علوم وهندسة الحاسوب – كلية ثياغاراجار للهندسة – مادوراي – 625015 – الهندس

خلاصة

يلعب تشخيص أمراض القلب عن بعد بواسطة تخطيط القلب دوراً هاماً في مجال الرعاية الصحية. التنقيب عن البيانات– والذي يعد خطوة كبيرة ضمن عملية استخلاص المعرفة باستخدام خوارزميات وصفية وتنبؤية تساعد بدورها في اتخاذ القرارات الاستباقية – يتم استخدامه أيضاً في تشخيص الأمراض القلبية والوعائية. لقد تم مؤخراً تطوير تقنيات متنوعة لتحليل نتائج تخطيط القلب. ولكن نظراً لتنوع التقنيات المستخدمة والمصطلحات ومقاييس الأداء المستخدمة في هذه التقنيات المختلفة، أصبحت عملية تحليل ومقارنة النتائج محبطة. الهدف من هذا البحث هو فحص وتقديم تحليل عن خوارزميات التنقيب عن البيانات المختلفة والمقترحة سابقاً على الساحة البحثية لتشخيص الأمراض القلبية والوعائية باستخدام التنقيب عن البيانات ضمن قراءات تخطيط القلب وفق أربع مراحل رئيسية وهي – الحصول على تخطيط القلب وضغط نتيجة التخطيط واستخلاص النتائج من التخطيط والتشخيص من خلال التخطيط. الهدف الأساسي من هذه الورقة هو استعراض وتصنيف مختلف البحوث التي أجريت في هذا الصدد لتقديم مصطلحات للباحثين المهتمين والمساعدة في تحديد الاتجاه المحتمل للبحوث الخاصة بهم.