International Journal of Automation and Computing 11(2), April 2014, 162-179
DOI: 10.1007/s11633-014-0778-5

Issues in the Mining of Heart Failure Datasets

Nongnuch Poolsawad[1] Lisa Moore[1] Chandrasekhar Kambhampati[1] John G. F. Cleland[2]

[1]Intelligent Systems Research Group (IS, Department of Computer Science), University of Hull, UK

[2]Hull York Medical School, Department of Cardiology, University of Hull, UK

Abstract: This paper investigates the characteristics of a clinical dataset using a combination of feature selection and classification methods to handle missing values and understand the underlying statistical characteristics of a typical clinical dataset. Typically, when a large clinical dataset is presented, it consists of challenges such as missing values, high dimensionality, and unbalanced classes. These pose an inherent problem when implementing feature selection and classification algorithms. With most clinical datasets, an initial exploration of the dataset is carried out, and those attributes with more than a certain percentage of missing values are eliminated from the dataset. Later, with the help of missing value imputation, feature selection and classification algorithms, prognostic and diagnostic models are developed. This paper has two main conclusions: 1) Despite the nature of clinical datasets, and their large size, methods for missing value imputation do not affect the final performance. What is crucial is that the dataset is an accurate representation of the clinical problem and those methods of imputing missing values are not critical for developing classifiers and prognostic/diagnostic models. 2) Supervised learning has proven to be more suitable for mining clinical data than unsupervised methods. It is also shown that non-parametric classifiers such as decision trees give better results when compared to parametric classifiers such as radial basis function networks (RBFNs).

Keywords: Heart failure, clinical dataset, classification, clustering, missing values, feature selection.

1 Introduction

Recently data mining has become an evolving area in information technology. Hundreds of novel mining algorithms and new applications in medicine have been proposed to play a role in improving the quality of healthcare systems. Data mining ties many technical areas, including machine learning, human-computer interaction, databases and statistical analysis. Clinical datasets pose a unique challenge for data mining algorithms and frameworks. These challenges are due to missing values, high dimensionality, unbalanced classes, and various systematic and human errors[1]. Data mining aims to automatically extract knowledge from large scale data. However, information and knowledge mined from the large quantity must be meaningful enough to lead to some advantages. As a result, effective planning of medical care and treatment of patients with heart failure has proved to be elusive.

With the advent of electronic health (patient) records (EHR/EPR)[2, 3], large amounts of clinical data have started to become available. However, good, robust, and accurate models for diagnosing and predicting the survivability of patients are not extensively available. Clinical datasets are often extremely complex due to the fact that there are large numbers of variables, and a great deal of missing data and non-normally distributed data. In addition, given the large number of data mining techniques, it can be difficult to decide which technique is required in order to get the correct results from a given dataset. This often means that if the underlying characteristics of the dataset change, the technique must also be changed.

The goal of data mining in health care systems is to assist clinicians in improving the quality of prognosis and diagnosis, and to generate timelines for the medical problem. The target problem was extracted from the dataset using a variety of data mining processes, which were also used to predict mortality and survival time of patients with heart failure. Machine learning techniques, such as supervised and unsupervised methods, were applied to compare the performance of prediction in clinical dataset. This paper looks into a large clinical dataset with a view to understand the underlying properties and the compromises necessary in the selection of methods for data mining. Thus this paper aims not only to explore and select suitable techniques to handle but also to analyse clinical datasets. The clinical dataset to be used is a large heart failure dataset (LIFELAB)[4, 5]. Over the years, a large number of results have been presented, specifically dealing with the issue of feature selection and the development of models for heart failure using data mining techniques[6−28]. A generic process applied here is: 1) missing values imputation, 2) feature selection, 3) classification and 4) clustering. There are a large number of techniques available for feature selection[29−31]. Three of these are selected: t-test[32], entropy ranking[33, 34], and nonlinear gain analysis (NLGA)[35]. All feature selection methods, indeed dimension reduction techniques, use a feature importance measure capability to select the most relevant features, therefore reducing the dimensionality of the problem. The rationale for this selection is that the three techniques use different properties of the data to select significant features or variables (Here, features and variables are interchangeably used). The t-test method utilizes data distribution as a key property for selecting variables. The entropy method not only uses the distribution, but also includes a measure of data density, and develops a measure for the degree of order in the data. NLGA considers higher weight variables to be more significant based on the artificial neural net input gain measurement approximation (ANNIGMA). ANNIGMA[35] uses neural networks for training large volumes of data and considers higher weight variables to be subset of significant features. The results indi-

Manuscript received October 29, 2012; revised July 18, 2013

cate non-parametric that classifiers, such as decision trees, show a better result when compared to parametric classifiers such as radial basis function networks (RBFN), multilayer perceptron (MLP), and k-means (because these assume that clinical data is normally distributed).

The paper is structured as follows: Section 2 provides some definitions, which are then used later in the paper. Section 3 describes a clinical dataset which has the typical characteristics of many clinical datasets. This section also outlines the embedded characteristics of the dataset, which will prove useful in the analysis of the results. In Section 4, several techniques for data mining are outlined. The category of techniques is dependent on the stage of the data mining process. Therefore, initially methods for imputing missing values are discussed, before moving on to feature selection and classification algorithms. Section 5 analyses the results in the context of the characteristics of the dataset, evaluating and validating the problems associated with the data by establishing a relationship between the complexities, the set of selected features, and the data distribution. The set of appropriate features are those with the highest classification. Section 6 discusses the results in relation and in comparison to previously established findings in literature. Finally, in Section 7 we draw some concluding remarks, summarize the analysed results and specify the further steps of the research as future works.

2 Preliminaries

Let $X_i \in X \subseteq \mathbf{R}^n; i = 1, \cdots, n$ be the clinical dataset, where n is the number of patient records, and m is the number of attributes (variables). Let $x_{ij} \in \mathbf{R}, i = 1, \cdots, n$ and $j = 1, \cdots, m$, be the i-th and j-th entry of the dataset under consideration. x_{ij} is defined as the value of the i-th variable for the j-th patient.

Issues associated with the dataset include high dimensionality, incomplete or missing values, and diverse clinical features and their magnitudes. However, many of the features present are irrelevant and redundant. The problem is determining a mapping from the high dimensional space to a lower dimensional space, i.e.:

$$v : \chi \longrightarrow \chi; \chi \in \mathbf{R}^k; k \ll n \tag{1}$$

For feature selection, the requirement is that X—since the main interest is to retain the labels associated with the variables. On the other hand, this is not required for feature extraction, since it employs latent variables. (See Fig. 1)

Definition 1. Subset of selected features (variables/attributes) is selected by dimensionality reduction techniques, the result is the matrix $\bar{X}_{n\times\bar{b}}$.

$$\bar{X}_{(n\times\bar{b})} \subset \bar{X}_{(n\times b)} \tag{2}$$

where $b \gg \bar{b}$, b is the number of the original features, $\bar{b}$ is the number of the selected features, $\bar{X}_{(n\times b)}$ is the data matrix that presents the significant features.

The process of reducing the dimension is essentially one of determining a projection, from the higher dimensional space to a lower dimensional one. Since most projection mappings employ local projections, it is imperative that the matrix A_{data} should not contain missing elements. As such, it is important to define missing data before designing an appropriate imputation method.

$$A_{\text{data}} = \begin{bmatrix} x_{11} & \cdots & x_{i1} \\ \vdots & \ddots & \vdots \\ x_{1j} & \cdots & x_{ij} \end{bmatrix} \tag{3}$$

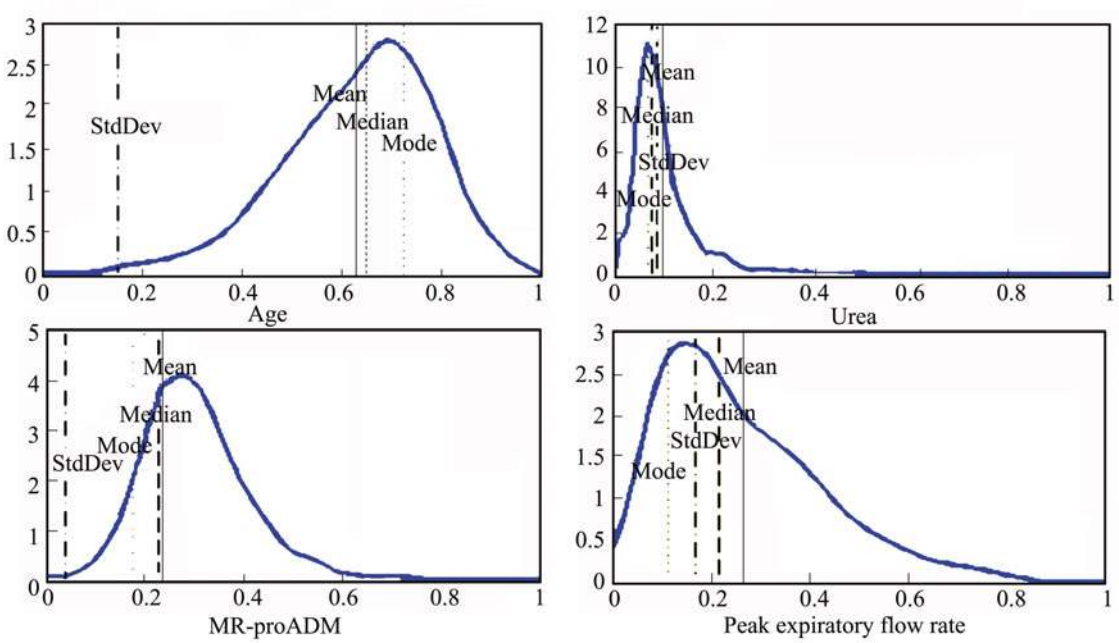

Fig. 1 Data distribution of variables in clinical dataset

Definition 2. Nullity values are defined as missing values, where values are absent or not recorded for a given attribute. The data matrix x is constructed by x_{ij}, where x_{ij} is null.

$$\text{nullity} = \{x_{ij} \in X : x_{ij} \in \varnothing\} \tag{4}$$

Find the numbers of missing value for each column (variable) $[N_1, N_2, N_3, \cdots, N_m]$.

$$[N_1, N_2, N_3, \cdots, N_m] = \text{count}_{j=1}^{m}(\text{nullity}(X)_{1,\cdots,n,j}) \tag{5}$$

(the nullity location of the dataset). The dataset

$$\bar{\chi}_{(n\times b)} = \text{find}_{(i=1)}^{n}(\text{nullity}(\chi_{(n\times b)})) \tag{6}$$

$$\bar{\chi}_{(n\times b)} = \begin{cases} 1, & \text{missing value} \\ 0, & \text{non missing value} \end{cases}$$

where $\bar{\chi}$ is the data matrix shows the location of missing value.

The incomplete, erroneous and noisy data are corrected by imputation. The dataset $\Psi_{(n\times m)}$ is the matrix of clinical dataset consists of n records of patient and m variables of attributes. Let $x_{ij} \in \mathbf{R}, i = 1, \cdots, n$ and $j = 1, \cdots, m$, be the i-th and j-th entry of the dataset under consideration. x_{ij} is defined as the record for each patient.

3 Mining issues in clinical dataset

This study focuses on a heart failure dataset consisting of continuous data, which contains diverse clinical features and numerous subsets, as well as both longitudinal and horizontal data across several generations. The dataset also importantly presents the incidence, prevalence and persistence of heart failure. High-risk patients with heart failure were targeted for evaluation and treatment in a cost-effective manner[26, 36]. The dataset in this paper is a large cardiological database called LIFELAB: A prospective cohort study consisting of 463 variables which are both continuous and categorical, and 2032 patients who were recruited from a community-based outpatient clinic based in the University of Hull Medical Centre, UK. Variables with missing values greater than 20% were excluded to minimize problems during the data mining process. As a result, the number of variables and patients were substantially reduced to 60 variables and 1051 patients. This indicates that the data consisted of multiple missing values that either needed replacement or elimination to allow appropriate analysis and algorithmic implementation. The challenges and complexities in large clinical datasets are discussed in the following outlined topics.

3.1 Incomplete, erroneous and noisy data

There is a wealth of clinical and health records generated every day and kept in storage. This raw clinical data is usually incomplete, containing missing values due to different systematic ways through which the real world data is collected by healthcare practitioners. Clinical datasets almost inevitably contain missing values and misclassified values. Methods of data imputation[37, 38] and missing value replacement are employed to cope with these issues. Inconsistent data can also exist, e.g., when data collection is done improperly or mistakes are made in data entry; the data may also contain error and noise. Commonly, outliers due to entry errors are also found and these were manually inspected to remove irrelevant variables.

3.2 Diverse clinical features and their scales

There are approximately 400 features in the dataset, comprised of many scales of measurement. Some variables consist of integer and decimal values and some scales have a wide range while some have a small range. Normalisation will be applied to solve these problems so that the data elements are within the same scale and manageable for sequential data mining processes.

3.3 Large dimensionality

Large dimensionality is indicated by too many features. Feature selection efficiently copes with this issue. The technique selects meaningful features which can be used in predictive modelling.

The data exploration reveals that the data distribution affects the mining process, including feature selection, classification and clustering analysis. Fig. 1 shows an example of the distribution of variables in the clinical dataset. In theory, the data should be normally distributed. However, it can be seen that this is not the case. It can be seen from Tables 2 and 3 that imputing missing values showed no significant changes and, as a result, the transformation procedure was unable to improve the precision.

4 Data mining processes in heart failure dataset

The mining process that is implemented in this paper can be represented as a four-stage process. The stages are 1) missing values imputation, 2) dimension reduction using feature selection techniques, 3) classification/clustering, and 4) evaluation. In this section, each of these four stages is discussed and the methods are outlined. The data mining framework for handling complexities is outlined in Fig. 2.

4.1 Missing value imputation

Data pre-processing is undoubtedly the first step in any form of data analysis and mining of data if the right results are to be obtained[36, 37]. At this stage, any redundant data, irrelevant variables and variables with more than 30% missing data are manually removed[38, 39].

Most datasets encountered contain missing values. Depending on their robustness, machine learning schemes have the ability to handle such datasets. The imputation methods used in this paper are mean imputation, expectation-maximization (EM) algorithm, k-nearest neighbour (k-NN) imputation, and artificial neural network (ANN) imputation[40]. After the application of each of the imputation methods, the data was normalized in order to ensure that all the variables were within the same range

so that both data integrity and high performance could be obtained during the mining process.

Step 0. Give data matrix $X_{(n\times m)}$, let $x_{ij} \in \mathbf{R}$, $i = 1, \cdots, n$ and $j = 1, \cdots, m$ be the i-th and j-th entry of the dataset under consideration.
Step 1. Find the missing values, let matrix $N_{(1\times m)}$, $= N_1, N_2, N_3, \cdots, N_m$, where $N_{(1\times m)}$ is the number of null values of each column.
  1.1 Find the numbers of missing value for each column (variable) $[N_1, N_2, N_3, \cdots, N_m]$ by using (5).
  1.2 Select the column with missing values less than 20%.
Step 2. Apply missing value imputation techniques: Calculate the imputation values by using mean imputation, EM algorithm, k-NN imputation, and ANN imputation. Let matrix $MI_{(1\times b)}$ has the imputation values.
  2.2 Find the missing values for each variable and impute the missing values by using (6).
  2.3 Replace the missing values by imputation values $MI_{(1\times b)}$.
Step 3. Normalise the data between 0 and 1.
Step 4. Apply feature selection techniques to select subset of feature. Select the significant variables by using NLGA, t-test method, and entropy ranking.
Step 5. Apply classification or clustering
  5.1 Classification: supervised learning technique (target variables required)
    5.1.1 Apply the classification techniques by using multilayer perceptron (MLP), J48 (decision tree), RBFN, support vector machines (SVMs) and random forest.
    5.1.2 Get the output from classification methods which classified into the classes of target.
    5.1.3 Validate the performance of classification results by comparing between target and output values: the confusion matrix.
  5.2 Clustering: unsupervised learning technique (no target variables)
    5.2.1 Apply the clustering method by using k-means and hierarchical clustering
    5.2.2 Determine the membership matrix.
    5.2.3 Validate the performance of classification results by comparing between target and output values.
Step 6. Evaluate the performance of prediction by using confusion matrix by using (16) and (17).

Fig. 2 The framework for handling complexities in clinical dataset

4.1.1 Mean imputation

A popular method is to use the mean of the data for imputation. Here missing data for a given feature (attribute/variable) is replaced using the mean of all known values of that attribute. However, mean imputation makes only a trivial change in the correlation coefficient and there is no change in the regression coefficient[40, 41].

4.1.2 Expectation-maximization (EM) imputation

Expectation-maximization uses other variables of the dataset to impute a value (expectation) and then checks whether that is the value most likely (maximization) to occur. Here the covariance matrix is estimated, and values to be imputed are generated using this covariance data. This method preserves the relationship with other variables, and is important where factor analysis or regression analysis is applied. As result, EM imputation is one of the most accurate methods of imputation. However, this is a reasonable approach only if the percentage of missing data is very small[42].

4.1.3 *k*-nearest neighbour imputation

Often, in large data sets it is possible to find two or more records which are similar, but one of them has a particular attribute missing. It is perfectly feasible to use the value from the closest record in similarity to replace the missing value. k-NN imputes missing data by applying this nearest-neighbour strategy[40]. Missing values of a variable are imputed by considering a number of records that are most similar to the instance of interest. In order to determine the similarity of records, a distance function (e.g., Euclidean distance) can be used as a measure.

4.1.4 Artificial neural imputation

ANN is an interconnected assembly of nodes (or neurons)[43, 44] where information or relationships are stored in the interconnections between them in the form of weights. In order to obtain these weights, the ANN has to learn or be trained using a training dataset. This approach can be seen as an extension of the EM approach, where instead of covariance, a nonlinear mapping is obtained to determine the missing values.

Table 1 The statistic of variables before and after missing value handling by different methods

Variable	Statistic	Missing value imputation				
		Original	EM	k-NN	Mean	ANN
Glucose	Missing (%)	4.19				
	Mean	0.088	0.088	0.088	0.088	0.089
	SD	0.060	0.059	0.059	0.059	0.060
	#Data	886	925	924	929	933
Haemoglobin	Missing (%)	0.95				
	Mean	0.577	0.577	0.457	0.577	0.577
	SD	0.131	0.131	0.107	0.131	0.131
	#Data	709	716	745	719	715
MCV	Missing (%)	20.74				
	Mean	0.795	0.795	0.811	0.795	0.788
	SD	0.066	0.061	0.068	0.059	0.063
	#Data	706	892	830	900	897
Iron	Missing (%)	13.51				
	Mean	0.262	0.329	0.258	0.262	0.327
	SD	0.127	0.112	0.119	0.118	0.105
	#Data	671	759	786	751	759
Vitamin B12	Missing (%)	7.04				
	Mean	0.094	0.094	0.094	0.094	0.093
	SD	0.062	0.060	0.060	0.060	0.068
	#Data	863	925	927	929	955
Red cell folate	Missing (%)	8.75				
	Mean	0.229	0.231	0.229	0.229	0.073
	SD	0.141	0.137	0.135	0.135	0.046
	#Data	767	840	840	842	937

These methods were used to impute missing values in the dataset described in Section 3. Table 1 shows some of the variables with approximately 1% to 20% missing values and the results obtained by imputing the missing values. The results shown in Table 1 compare the statistical properties of the data with no imputation and after imputation. It can be seen that with some methods the values of the standard deviation (σ) and mean (μ) have changed. In Table 2, #data indicates the number of data points within the normal distribution range, i.e., data points within the range of $[\mu - \sigma, \mu + \sigma]$. It can be seen that missing value imputation methods (EM, k-NN, Mean and ANN) show an increase in the number of data points under the distribution curve. In addition, the table show the effect of imputation methods on the same variable. For example Tables 1 and 2 shows that the imputation method based on k-NN produces the better results for Haemoglobin and Iron, whilst the ANN based method shows the most accurate results for Glucose, vitamin B12 and red cell folate, and that mean imputation is suitable for mean corpuscular volume (MCV). Each of these methods has a specific way of imputing the missing value, and the primary nature of the distribution is either retained by the imputation method or is fundamentally changed. Indeed, this can be seen from Table 2, where the distributions before and after imputation are shown.

Table 2 Data distribution of different variables of the original data and missing value replacement data

Variable	Original	EM	k-NN	Mean	ANN
Glucose					
Haemoglobin					
MCV					
Iron					
Vitamin B12					
Red cell folate					

4.2 Feature selection

Feature selection, also known as subset selection, is a process that selects the most relevant attributes (features). This process not only determines the most relevant features, it also reduces the dimensionality of the problem (Fig. 3). Thus reducing the complexity and processing time, while at the same time improving performance. In general, a feature selection algorithm is often composed of three components: a performance function, a search algorithm and an evaluation function. The performance function provides the optimal subsets appropriate for classification. The search algorithm performs the search of an appropriate subset of features. The evaluation function inputs a feature subset and outputs a numeric evaluation.

Feature selection has been successfully applied to the following datasets: lymphoma, gene expression, cancer[31, 33, 45]. Poolsawad et al.[39] state that feature selection consistently increases accuracy, reduces feature set size, and provides better accuracy for classification. Further, Liu et al.[34] also state that feature selection plays an important role in classification, and is effective in enhancing learning efficiently, increases productive accuracy, and reduces complexity of learning results. In addition, learning is efficiently achieved with just relevant and non-redundant features.

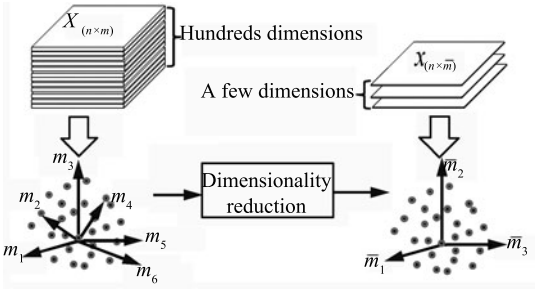

Fig. 3 The dimensionality reduction from a high dimension to a small dimension

There are two general forms of feature selection procedures: 1) a wrapper model and 2) A filter model[46].

The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. The learning algorithm is run with various subsets of features, and the learner that performs the best is chosen. In contrast, the filter model presents the data with the chosen subset of features to a learning algorithm. It separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm[14, 47]. In comparison to the wrapper model, the filter model is computationally efficient. However, the filter model is known to perform much worse than the wrapper model. A key aspect which needs to be considered when selecting a subset of features is the metrics used for determining the relevance or redundancy of a particular feature. An optimal subset of features should contain a set of robust and relevant features along with a set of weak features[46]. This allows for the selection of features with a positive Z-score[47]. It is possible to obtain different selection of subsets of features depending on the criterion used. Thus the subset obtained using a statistical correlation criterion would be different from when mutual information is used.

4.2.1 Nonlinear gain analysis

Nonlinear gain analysis (NLGA), also known as artificial neural net input gain measurement approximation (ANNIGMA), is a feature ranking procedure[34]. In this approach, a neural network is repeatedly trained. And after each training operation, a set of variables is eliminated based on their effectiveness and significance in predicting the required class or outcome. In the first step, all the features are used as inputs and the network is trained. Once the network has been trained, an ANNIGMA score is determined as

$$LG_{ik} = \Sigma_j |w_{ij} \times w_{jk}| \tag{7}$$

$$\text{ANNIGMA}_{ik} = \frac{LG_{ik}}{\max(LG_{ik})} \times 100 \tag{8}$$

where i, j, k are the input, hidden, and output layer nodes indicated, respectively. LG_{ik} is the local gain of all the other inputs, while w_{ij} and w_{jk} are the weights between the layers.

Features associated with low ANNIGMA scores are eliminated and another network is trained. This is carried out till such a point that the network performance starts to degrade. The NLGA is a wrapper model and appropriate for handling large datasets with a high dimension. This approach can reduce the dimensions while also maintaining the required accuracy. However, due to its high computational requirements, its application to extremely large data sets is limited.

4.2.2 *t*-test

Student′s t-test approach uses statistical tools to assess whether the means of two classes that are statistically different from each other by calculating a ratio between the difference of means and the variability of two classes. This method has been found to be efficient in a variety of application domains, for example in: 1) genotype research[31, 33, 47], where the problem is one of evaluating differential expressions of genes from two experimental conditions, and 2) the ranking of features for mass spectrometry[48−50] and microarray data[47, 51, 52]. The use of t-test is limited to two class challenges. For multi-class problems, the procedure requires the computing of a t-statistic value (following the equations in [32, 33, 47]) for each feature corresponding to each class by evaluating the difference between the mean of one class and all the other classes, where the difference is standardized by within-class standard deviation as

$$t(x_i) = \frac{(\bar{y}_1(x_i) - \bar{y}_2(x_i))}{\sqrt{\left(\frac{s_1^2(x_i)}{n_1} + \frac{s_2^2(x_i)}{n_2}\right)}} \tag{9}$$

where $t(x)$ is the t-statistics value for the number of features; and $\bar{y}_1, \bar{y}_2$ are means of classes 1 and 2, while s_1^2, s_2^2 are the within-class standard deviations of classes 1 and 2, n_1 and n_2 are the numbers of all the samples in classes 1 and 2, respectively.

4.2.3 Entropy ranking

While the NLGA approach selects features purely based on their contribution to the final result, and the t-test approach utilizes statistical properties to determine the required features, entropy based approaches not only take into account the statistical properties of the features, but also the compactness and density of the data. Entropy is a measure of the information conveyed by the probability distribution function of a particular variable/feature. Using this entropy, Fayyad[32] suggests a cut-off point selection procedure by using class entropy of subset. In general, if we are given a probability, $P(\cdot)$, then the information conveyed by this distribution, also called the entropy of P, is as

$$\text{Ent}(S) = -\sum_{i=1}^{k} P(C_i, S)\log(P(C_i, S)) \tag{10}$$

$$\text{Ent}(S) = -\sum_{i=1}^{k} \frac{C_i}{S}\log\frac{C_i}{S} \tag{11}$$

where $\text{Ent}(S)$ measures the amount of information required to specify the classes in a set of attributes S, and $P(C_i, S)$ is the proportion of examples in S consisting of class C in the i-th feature. The entropy values are sorted in an ascending order and consider those features with the lowest entropy values.

Table 3 shows the features selected using the ANN imputation and NLGA feature selection technique. The result compares the selected features in both outcomes — mortality (dead/alive) and mortality time frame, and it indicates that the variables highlighted appeared in both outcomes. This signifies that both applied techniques are capable of locating significant variables in the dataset.

4.3 Classifiers

The classifier algorithms employed in this paper are multilayer perceptron (back-propagation), J48 (decision tree) and radial-basis function (RBF) network. These classification techniques were implemented in Waikaito environment for knowledge acquisition (WEKA)[53].

Table 3 The selected features using ANN imputation and NLGA

No.	Outcome	
	Mortality (dead/alive)	Mortality time frame
1	Potassium	Sodium
2	Chloride	Bicarbonate
3	Urea	Urea
4	Creatinine	Creatinine
5	Calcium	MR-proANP
6	Phosphate	CT-proAVP
7	Bilirubin	Haemoglobin
8	Alkaline phosphatase	White cell count
9	ALT	Platelets
10	Total protein	Total protein
11	Albumin	Bilirubin
12	Triglycerides	Alkaline phosphatase
13	Haemoglobin	Adj calcium
14	Iron	Phosphate
15	Vitamin B12	Cholesterol
16	Ferritin	Uric acid
17	TSH	CT-proET1
18	MR-proANP	Red cell folate
19	CT-proET1	Ferritin
20	CT-proAVP	NT-proBNP

4.3.1 Multilayer perceptron (back-propagation)

Multilayer perceptrons (MLP) are feedforward neural networks, and are used for learning classification or unknown nonlinear functions[54]. In multilayer perceptron (see Fig. 4), there is an input layer with a node; each node represents an independent variable. There may be one or more intermediate hidden layers, and each node in the output layer corresponds to a different class of the target variable. In this paper, a feed-forward network consisting of input units, hidden neurons and one output neuron is optimized to classify the outcome. The number of input units is the same as the number of input attributes of the selected variables and the number of hidden neurons is half the number of input attributes. All weights are randomly initialized to a number close to zero and then updated by the back-propagation algorithm. The back-propagation algorithm contains two phases: forward phase and backward phase. In the forward phase, we compute the output values of each layer unit using the weights on the arcs. In the backward phase, the weights on the arcs are updated by a gradient descent method to minimize the squared error between the network values and the target values.

The architecture of multilayer perceptron showing the output y, which is a vector with n components determined on the terms of m components of an input vector; x and l components of the hidden layer. The mathematical representation is expressed as

$$y_i(x) = \sum_{j=1}^{l}\left[v_{ij}g\left(\sum_{k=1}^{m} w_{ij}x_k + b_{wj}\right) + b_{vi}\right], \quad i = 1, \cdots, n \tag{12}$$

where v_{ij} and w_{ij} are synaptic weights, x_k is the k-th element of the input vector, $g(\cdot)$ is an activation function, and b is the bias which has the effect of increasing or decreasing the net input of the activation function depending on whether it is positive or negative, respectively.

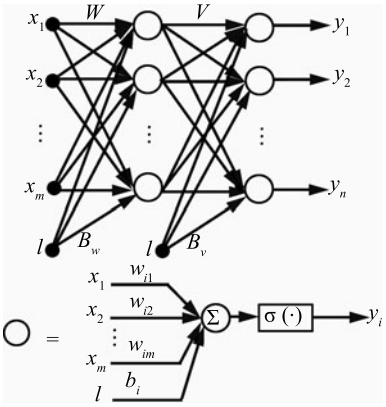

Fig. 4 A multilayer perceptron structure

In general, MLPs use a supervised training paradigm for determining the weights and to learn the classification problem. MLP learns how to transform input data into a desired response, so they are widely used for pattern classification[55, 56]. In terms of training itself, there are other training paradigms available for these networks, here back-propagation is used for illustration.

4.3.2 J48 (decision tree)

A decision tree partitions the input feature of a dataset into regions, where each assigned label is a value or an action to characterize its data points (Fig. 5). In this paper, a decision tree C4.5 algorithm is generated for classification. The algorithm identifies attributes that discriminates various instances clearly, when a set of items (training set) are encountered. This is performed using a standard equation of information gain. Among the possible values of this feature, if there is any value with no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then that branch is terminated and the obtained target value is assigned to it.

4.3.3 Radial basis function network

Radial basis function network (RBFN) is an artificial neural network model that uses RBF as an activation function. Fig. 6 presents the architecture of RBFN. It is composed of three layers: an input layer, a hidden layer and an output layer. Each hidden unit implements a radial activation function (a non-linear transfer function) and each output unit implements a weighted sum of hidden unit outputs.

The output of the i-th neuron in the output layer of the RBF network is determined as

$$y_i(x) = \sum_{j=1}^{M} w_{ij}\varphi(||x - c_j||), \quad i = 1, \cdots, m \tag{13}$$

where $\varphi(\cdot)$ is the basis function which is described using $x - c_j, c_j$ is the centre vector for hidden neuron j, w_{ij} is the weight between the node j of the hidden layer and the node i of the output layer, and m is the number of nodes in the output layer. The norm is typically taken to be the Euclidean distance and the basis function is taken to be

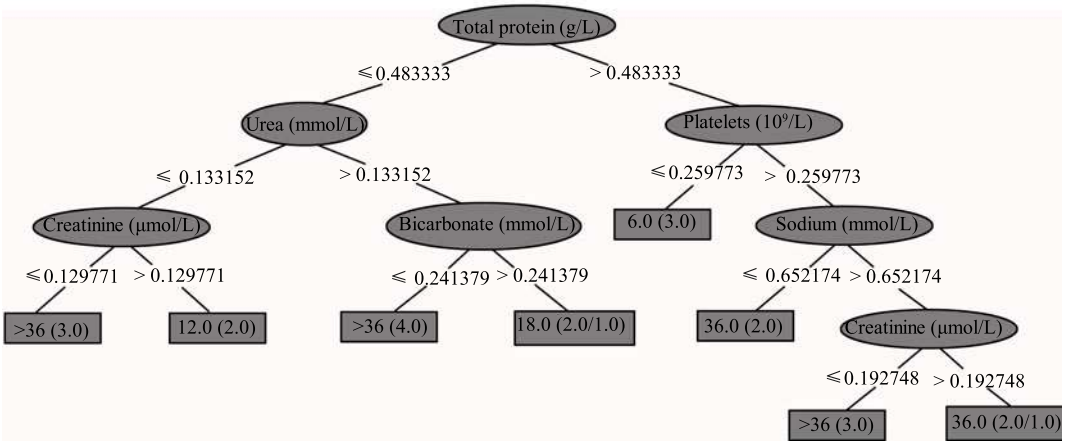

Fig. 5 Decision tree for predicting the survival months

Gaussian:

$$\varphi(||x - c_j||) = \mathrm{e}^{\left\{\frac{||x-c_j||^2}{2\sigma_j^2}\right\}} \tag{14}$$

where $\varphi(\cdot)$ is the width parameter of the j-th hidden unit in the hidden layer.

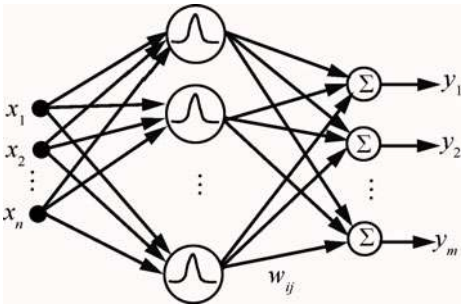

Fig. 6 A radial basis function network architecture

4.3.4 Support vector machines and random forests

Support vector machines (SVMs)[57] are supervised learning models. SVM′s are essentially a non-probabilistic binary linear classifier and is a model which uses a representation of the key example points which are mapped so that separate categories are divided by a gap that is as wide as possible. New data points are then mapped into the same space and a prediction is made depending on which side of the divide they fall.

The learning in an SVM is the construction of a hyperplane which is used for classification. An ideal or an optimal hyperplane can be defined as a linear decision function which provides the maximal margin between the vectors of the two classes (see Fig. 7). The support vectors define the margin of largest separation between the two classes. SVMs are a popular classification tool as they have excellent generalization properties. However, the training is slow and the algorithms are numerically complex[58]. This paper uses the SVM algorithm called sequential minimal optimization or SMO[58, 59].

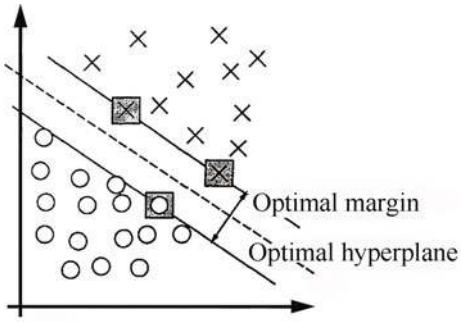

Fig. 7 A separable problem in a 2-dimensional space[57]

Random forests, as the name suggests, is a collection of trees: decision trees, in this case. Algorithms for classification using a random forests approach was developed by Breiman[60]. Here a combination of tree predictors are used, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The input class of the random forest for a given input is the mode of the classes predicted by individual trees.

4.4 Clustering

Clustering is a popular multivariate statistical technique embodied in many processes such as data mining, image processing, pattern recognition and classification[61]. The unsupervised method partitions inherent patterns into clusters, based on the order of similarity, thus discovering the structure of a given data. Data points in the same cluster are classified as similar between one another while those in different clusters are dissimilar. In this paper, we have applied two clustering algorithms known as k-means and hierarchical clustering.

Two major issues should be considered in practice: 1) deciding on the number of clusters to use for each clustering algorithm, and 2) defining the categorical attributes[61, 62]. In this study, the number of clusters will be fixed for both algorithms to ensure a fair and consistent analysis, and different categorical attribute are present in the dataset, each representing a different clin-

ical testing. It is important to bear in mind that defining categorical attributes can be a difficult task in cluster analysis[63]. For this reason, the following clustering algorithms are implemented to achieve the best possible clustering outcome based on their respective function.

4.4.1 *k*-means clustering

k-means clustering is a partition algorithm that organizes the number of objects into k partitions ($k \leqslant n$). Where each partition corresponds to a cluster, k and n represents the number of objects. The method assumes that k is fixed[64, 65] and the means in k-means signifies an aggregation of clusters which is usually referred as centroids, as depicted in Fig. 8, denoted as "+". The centroid based technique ensures objects within the same cluster are similar, and that dissimilar objects are assigned to different clusters. However, this is dependent on the distance between the object and the cluster mean — a new mean must be calculated for each cluster. The process is repeated until a criterion known as the "square-error criterion" is initiated as[66]

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{15}$$

where E is the sum of the square error for all objects (n) present in the datasets, p and m_i are multidimensional this is jointly represented as C_i, p represents a given object and the point in space, while m_i is the mean of clusters. As a result, the distance between each object to each cluster centre (centroid) marked as "+" is squared and summed. The criterion is an essential part of the k-means process because it compacts and effectively separates the resulting k clusters simultaneously.

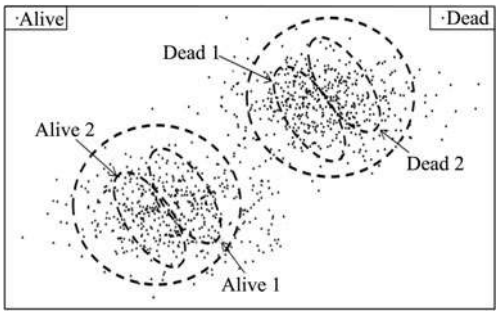

Fig. 8 Four clusters of the dataset are illustrated

Fig. 9 illustrates k number of clusters in this case, two clusters (A and B). Each object indicated by the bold black dots is distributed to a cluster based on the nearest cluster centre. This is further demonstrated by the dashed circles in A. Based on these objects in the cluster, the mean and distributions are recalculated and redistributed based on the nearest cluster centre and this forms the faded oval shapes shown in cluster B.

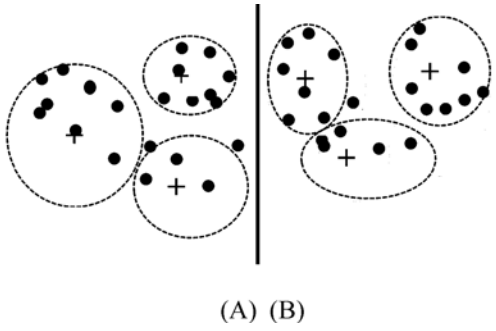

Fig. 9 A schematic clustering of a set of objects based on the k-means method. The mean or centroid of each cluster are represented by "+"

The structure is characterized by subsets $S_k \subset I$ and M-dimensional centroids $C_k = (c_{kv}), k = 1, \cdots, k$. Subsets S_k forms a partition $S = \{S_1, \cdots, S_k\}$ with a set of centroids $c = \{c_1, \cdots, c_k\}$[44, 67]. Where the M-dimensional centroid vectors (C_k) are cluster centroid that updates the S_k cluster list based on the "minimum distance rule". The rule classes entities to their nearest centroids, this is specifically achieved by computing the distances of each entity i.e., $I \in I$, to all centroids and then assigned to the nearest centroid.

Sridhar and Sowndarya[68] have shown k-means to produce reliable clustering results, as it is computationally easy and memory efficient. There are two types of k-means explained by Napoleon and Lakshmi[69], namely enhanced and bisecting k-means. However, neither are further discussed in this study. Moreover, studies conducted by Steinbach et al.[63] found bisecting k-means to be a better algorithm compared to the standard k-means. Fig. 10 shows three clusters of two distinctive dead and alive classes, alive patients which are represented by the triangulated symbol and the dead patients are represented by the black circles, alive 1 (right) cluster are patients predicted as alive with a few projected towards the dead groups. While Fig. 8 illustrates four clusters grouped into two classes of dead and alive, with dead 1 (left) cluster represented as dead patients.

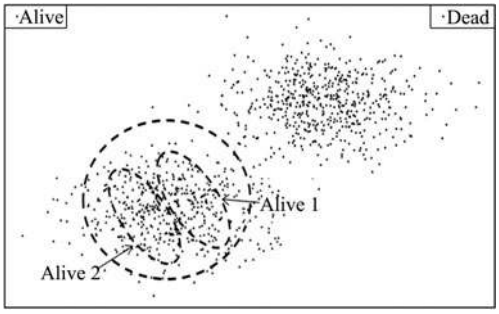

Fig. 10 k-means clustering indicating three clusters of the data

4.4.2 Hierarchical clustering

Hierarchical clustering is employed in this study to reveal similarities between the data attributes. The method par-

titions the data into a division of clusters and points during each stage of the process and then the clusters are combined in a different layer and thus building up a hierarchy of clusters, that resembles a tree diagram. This is presented through the use of a dendrogram.

Hierarchical clustering is generally classified as either agglomerative or divisive. The agglomerative method also known as the "bottom up" approach begins with each observation in their individual cluster and then sequentially merges into groups of larger clusters[44, 70]. The clusters are formed according to the minimum Euclidean distance (also known as a nearest neighbour clustering algorithm) between two objects from different clusters and their similarity are measured based on the closest pair of data points belonging to the different clusters. In contrast, the divisive approach is considered as the "top down" approach—the reverse of agglomerative hierarchical clustering—which begins with all the observations in one cluster and then divides into smaller clusters repeatedly until each observation is assigned to a cluster (Fig. 11). The clusters are divided based on the maximum Euclidean distance principle that considers the closest neighbouring objects in the cluster.

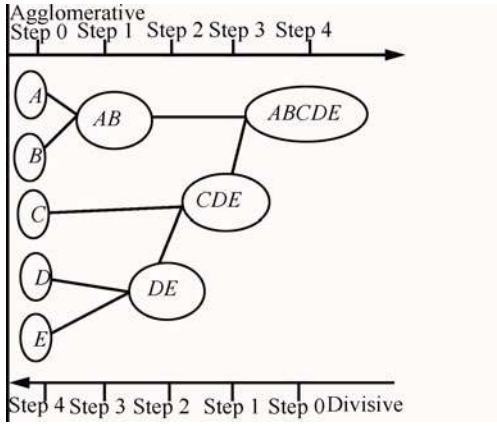

Fig. 11 Agglomerative and divisive hierarchical clustering on data objects (A, B, C, D, E)

Fig. 12 demonstrates the relationship and similarities between the variables; and a vertical axis is used to illustrate the similarity scale between clusters. As indicated by the dendrogram, urea and creatinine are the most similar followed by MR-proANP and CT-proET1. This signifies a clear relationship between the variables and correlation values shown in Table 4 which further supports their relation and similarity. Urea and creatinine are linked to CT-proAVP, ferritin while uric acid and red cell folate are also merged together to form one cluster with a similarity scale of approximately 50.

Table 4 Indicates correlation comparison

Test variables	Correlation	Similarity levels
Creatinine and Urea	0.8	90.7
MR-proANP and CT-proET1	0.6	79.9

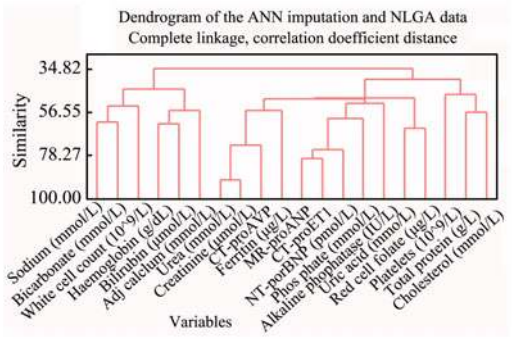

Fig. 12 Dendrogram used in hierarchical clustering to illustrate similarities

4.5 Performance evaluation measures

Performance measures are efficiency to evaluate the performance of classification. Many classifiers based on the performance measures are compared. Thus, we carefully used the measures to evaluate the performance, which are defined as

$$\text{"Precision"} = \frac{TP}{(TP + FP)} \tag{16}$$

$$\text{"Recall"} = \frac{TP}{(TP + FN)} \tag{17}$$

where TP is the number of true positives, FP is the number of the false positives, TN is the number of true negatives, and FN is the number of false negatives, respectively. Precision is a function of the correct classified examples (true positives) and the misclassified examples (false positives). Recall is a function of true positives and false negatives. Fig. 13 classifies the relationship between precision and recall values in the dead and alive categories.

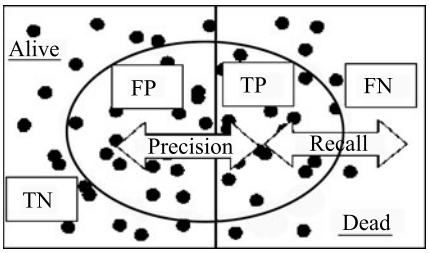

Fig. 13 A relationship between precision and recall values of classification

5 Experimental results

The experiments aim to assess the performance between supervised and unsupervised method for mining large clinical datasets by using different feature selection and missing value imputation methods. The dataset that used in the experiments is normalised to a range between 0 and 1. In most numerical procedures, such normalization is carried out in order to prevent some attributes with large numeric ranges dominating those with small numeric ranges.

The procedure that used in the experiments follows the framework proposed in Table 5. In all experiments, the data

is to be classified into two: mortality (dead or alive) and survival (6, 12, 18, 24, 36, or more than 36 months) (see Table 6). The dataset that is used in these experiments required the data mining process to analyse the data characteristics. The performance of classification (precision and recall) is used to evaluate the performance after applying the different methods for imputing the missing values and for selecting features.

It can be seen that the following combination produced the better results using the features shown in Table 4: 1) classification done by the decision tree (Fig. 14). 2) imputation carried out using a neural network and 3) an NLGA for selecting feature.

It can be seen in Tables 1 and 2 that all the imputation techniques, even though imputing different values, resulted in similar classification results (Tables 5 and 6). However,

Table 5 The classification results from different missing value replacement methods and feature selection (FS) techniques by dead and alive classes

FS	CSPA		Missing values imputation method							
			EM algorithm		k-NN imputation		Mean imputation		ANN imputation	
		Class	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
t-test	MLP	Precision	81.9	81.8	76.1	82.1	81.6	81.4	77.8	82.8
		Recall	58.9	93.4	61.2	90.3	57.8	93.4	62.6	91
	DT	Precision	87.7	89.8	95.9	90.3	93.1	92.3	96.2	93.1
		Recall	78.8	94.4	79	98.3	84.1	96.8	85.6	98.3
	RBFN	Precision	100	96.81	99.7	96.94	100	96.81	100	96.81
		Recall	93.48	100	93.77	99.86	93.48	100	93.48	100
	k-Means	Precision	61.54	76.86	63.35	77.27	61.51	77.11	63.08	77.07
		Recall	49.86	84.24	50.42	85.24	50.71	83.95	49.86	85.24
	SVM	Precision	68.5	73	68.9	72.9	68.7	73	68.7	73
		Recall	32.6	92.4	32	92.7	32.3	92.6	32.3	92.6
	Random forest	Precision	57.2	78.4	55.4	77.5	47.9	73.1	55.1	76.8
		Recall	57.2	78.4	55.5	77.4	45.3	75.1	53.5	77.9
Entropy	MLP	Precision	72.5	78.6	70.5	78.8	71.1	77.9	71.3	79.3
		Recall	51.6	90.1	52.7	88.8	49.6	89.8	54.1	89
	DT	Precision	93.2	89.4	86.5	88.5	87.3	91	91.6	91.8
		Recall	77.3	97.1	75.9	94	81.6	94	83	96.1
	RBFN	Precision	99.7	97.48	100	98.31	99.7	97.76	99.7	97.76
		Recall	94.9	99.86	96.6	100	95.47	99.86	95.47	99.86
	k-Means	Precision	62.59	76.84	65.24	75.43	62.59	76.84	66.38	75.86
		Recall	49.29	85.10	43.06	88.40	49.29	85.10	44.19	88.68
	SVM	Precision	69.6	72.9	71	73.2	70.4	73	70.8	72.8
		Recall	31.7	93	32.6	93.3	31.7	93.3	30.9	93.6
	Random forest	Precision	57.9	78.4	57.1	78.3	47.4	72.7	55.4	76.6
		Recall	56.9	79.1	56.9	78.4	43.9	75.4	52.4	78.7
NLGA	MLP	Precision	77.5	80.3	77.2	80.7	74.6	79.9	76.5	77.3
		Recall	55.5	91.8	56.7	91.5	55	90.5	46.2	92.8
	DT	Precision	93.1	92.6	79.9	88.5	79.2	84.9	98	87.2
		Recall	84.7	96.8	76.8	90.3	68	91	71.1	99.3
	RBFN	Precision	100	97.08	100	97.08	100	97.76	99.7	97.35
		Recall	94.05	100	94.05	100	95.47	100	94.62	99.86
	k-Means	Precision	47.80	74.70	58.33	76.86	58.52	76.89	54.90	77.38
		Recall	52.41	71.06	51.56	81.38	51.56	81.52	55.52	76.93
	SVM	Precision	73.2	71.7	71	72.9	68.8	71.9	68	72
		Recall	25.5	95.3	31.2	93.6	27.5	93.7	28.3	93.3
	Random forest	Precision	55.2	76.2	53.5	76.5	54.8	78.1	57.3	77.6
		Recall	51.3	78.9	53.5	76.5	58.1	75.8	54.7	79.4

Table 6 The classification results from different type of missing value imputation methods and feature selection techniques on mortality time frame outcome

				Missing values imputation method											
				EM algorithm						*k*-NN imputation					
	Class (months)			6	12	18	24	36	>36	6	12	18	24	36	>36
Feature selection & Classifier	*t*-test	MLP	Precision	76.5	61.9	83.3	42.6	34.6	49.6	73.6	59.7	55.6	44.2	70	49.1
			Recall	43.8	34.7	1.85	32.8	42.4	86.2	59.6	53.3	18.5	31.1	21.2	89.5
		DT	Precision	87.2	84	85.1	90.6	77.6	91.6	88.4	86.3	86.7	79.7	79.7	92.2
			Recall	84.3	90.7	74.1	78.7	89.4	92.8	85.4	92	72.2	83.6	83.3	92.8
		RBFN	Precision	50.7	37.3	52.2	35.3	29.4	40.1	41.6	36	48.5	28	31.7	46
			Recall	42.7	25.3	22.2	9.8	7.6	82.9	41.6	12	29.6	23	30.3	71.7
		KM	Precision	35.1	21.9	18.6	12.5	14.3	52.8	39.0	16.9	19.5	0	17.6	48.5
			Recall	38.5	36.8	44.4	28.6	2.1	47.8	30.8	34.2	41.7	0	6.4	47.8
	Entropy	MLP	Precision	53.9	29.8	40.8	75	36	48.6	59.3	39.8	48.3	90	39	50.2
			Recall	46.1	37.3	37	9.8	13.6	78.3	53.9	44	25.9	14.8	34.8	77.6
		DT	Precision	88.6	85.2	86.4	82.5	86.2	87.7	87.9	87.2	84.9	82	79.7	93.1
			Recall	87.6	92	70.4	77	84.8	93.4	89.9	90.7	83.3	82	83.3	88.8
		RBFN	Precision	42.4	35.3	28.1	45.5	25	39.7	60.4	42.1	37	40	14.3	35.8
			Recall	28.1	16	29.6	8.2	6.1	83.6	32.6	10.7	18.5	6.6	1.5	90.8
		KM	Precision	36.5	21.6	0	13.8	16.1	93.4	33.3	19.6	11.8	16.7	17.5	54.8
			Recall	36.5	21.6	0	34.1	18.2	84.1	15.4	23.7	5.6	29.5	23.9	50
	NLGA	MLP	Precision	71	42.2	51.7	50	30.9	57.4	55.3	49	52.6	100	33.9	46.3
			Recall	49.4	61.3	27.8	16.4	31.8	78.9	47.2	32	18.5	16.4	31.8	85.5
		DT	Precision	92.8	88	87.3	89.1	88.9	88	86.9	88.4	89.6	82.5	74	86.4
			Recall	86.5	88	88.9	80.3	84.8	96.1	82	81.3	79.6	77	86.4	92.1
		RBFN	Precision	57.6	27.3	40	31.3	45	49.1	53.6	38.3	47.4	33.3	29.7	41.6
			Recall	42.7	40	25.9	16.4	13.6	75.7	41.6	24	16.7	8.2	16.7	84.9
		KM	Precision	32.8	16.9	0	17.4	15.6	40.2	38.3	14.3	16.7	27.3	16.0	49.0
			Recall	38.5	28.9	0	9.1	29.8	31.6	34.6	2.6	25	6.8	27.7	55.1
				Missing values imputation method											
				Mean imputation						ANN imputation					
	Class (months)			6	12	18	24	36	>36	6	12	18	24	36	>36
Feature selection & Classifier	*t*-test	MLP	Precision	57.3	41.9	55.6	55.6	29.1	59.3	82.6	60	62.5	54.2	40.7	54
			Recall	57.3	41.3	27.8	24.6	37.9	75.7	64	48	27.8	21.3	50	84.9
		DT	Precision	86.2	86.3	89.8	85.7	87.1	88.3	91.9	84.8	88.1	87.7	84.5	89.5
			Recall	91	84	81.5	78.7	81.8	94.7	88.8	89.3	68.5	82	90.9	95.4
		RBFN	Precision	40.2	36.4	38.9	35	38.5	38.4	50	26.7	22.2	42.1	21.1	43.5
			Recall	37.1	16	13	11.5	7.6	83.6	38.2	16	22.2	13.1	6.1	83.6
		KM	Precision	34.5	22.6	18.4	12.5	14.3	51.2	35.6	19.4	19.0	12.5	14.3	52.5
			Recall	38.5	36.8	44.4	4.5	2.1	46.3	40.4	34.2	44.4	4.5	2.1	46.3
	Entropy	MLP	Precision	63.5	43.6	37.5	100	34.5	46.5	82	58.2	77.8	82.4	37.9	46.5
			Recall	52.8	22.7	27.8	8.2	28.8	86.8	56.2	42.7	25.9	23	33.3	88
		DT	Precision	87.6	76.5	87.5	77.8	84.6	89.7	86.9	87.5	91.1	91.1	80.6	82.7
			Recall	87.6	86.7	64.8	80.3	83.3	91.4	82	84	75.9	83.6	81.8	94.1
		RBFN	Precision	50.8	42.3	38.5	44.4	30.8	36	45.2	33.3	20	42.9	50	40
			Recall	33.7	14.7	18.5	6.6	6.1	86.2	37.1	33.3	3.7	4.9	1.5	86.8
		KM	Precision	28.3	18.2	10	20.5	8	55.5	31.6	20	15.2	13.2	14.3	52.8
			Recall	25	31.6	5.6	34	4.3	48.5	11.5	31.6	33.3	20.5	6.4	41.2
	NLGA	MLP	Precision	85.7	52.9	53.8	45	47.2	47.5	52.7	83.8	42.9	67.9	37.8	53.8
			Recall	47.2	36	25.9	29.5	37.9	86.8	66.3	41.3	22.2	31.1	47	74.3
		DT	Precision	86.7	84	86.3	87	87	92.8	96	87.3	90.9	79.7	85.3	84
			Recall	87.6	90.7	81.5	77	90.9	92.8	80.9	82.7	74.1	83.6	87.9	96.7
		RBFN	Precision	53.1	27.9	38.5	23.3	50	48.5	45.3	38.6	44.1	23.1	33.3	42.8
			Recall	29.2	45.3	27.8	11.5	18.2	74.3	38.2	22.7	27.8	9.8	10.6	83.6
		KM	Precision	31.9	11.1	13.5	11.8	16.9	50	27.6	19.1	0	18.7	21.9	53.2
			Recall	28.8	5.3	27.8	4.5	29.8	41.9	15.4	34.2	0	38.6	29.8	36.8

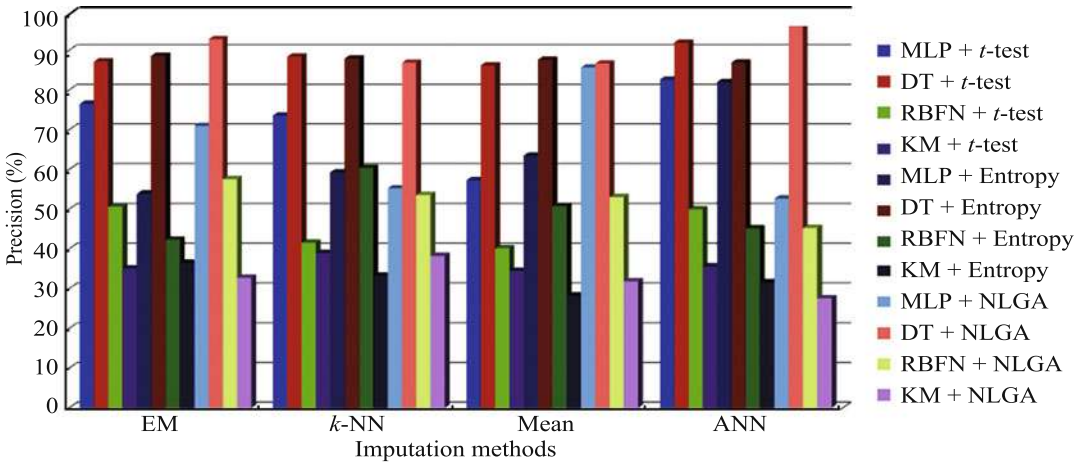

Fig. 14 The classification results from different missing value imputation methods and different feature selection (FS) techniques on 6 months class

the robust methods, for example EM algorithm, showed better results than others. The reason for this is that the EM algorithm determines maximum likelihood estimates. Tables 1 and 2 show that the statistics (mean and standard deviation) of variables and data distribution before and after applying imputation techniques. The means and standard deviations (Table 1) for EM algorithm are similar to original data. The similarity indicates, that this method provides greater flexibility in the shape of the distribution while maintaining about the same means and standard deviations (Table 2).

Tables 5 and 6 show the differences in the performances between the wrapper and filter approaches to feature selection. It can be seen that NLGA approach provided features which classified the data better than t-test and entropy (Tables 5 and 6). NLGA uses the efficiency of neural network to search for features which satisfies an error criterion. However, in general, wrapper approaches are more computationally intensive than the filter approaches (t-test and entropy). It can be seen from Fig. 14 that for the critical class of 6 month decision trees provide higher precision value than other classifiers.

Amongst the various approaches for classification, RBFN′s and decision tree′s (DT) had a slightly better performance than that of the other classifiers (Tables 5 and 6 and Fig. 14). The basic functions can be advantageous when the data has a multimodal distribution. It is typically trained using a maximum likelihood framework by maximizing the probability (minimizing the error), and hence the model performs a better approximation, and noisy interpolation.

Decision tree is a form of non-parametric multiple variable analysis. This method requires no information on the distribution of data. Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments and can generate rules that are easy to understand. Thus often clinical support systems are developed on the basis of these decision trees[71]. Internally, decision trees used information gain and entropy to select appropriate attributes at each node in order to create the branches.

6 Discussion

It is important to note that the issue of missing values in datasets is a major issue as it affects dimensionality reduction and classification[72]. This paper demonstrates four missing values imputation methods: 1) mean imputation, 2) EM algorithm imputation, 3) k-NN imputation and 4) ANN imputation. The primary reason carrying out imputation is to retain the size of the data rather than reduce it by eliminating record from the datasets. Tables 1 shows the statistical properties are mean and standard deviation, and Table 2 shows the data distribution before and after data imputation. The mean imputation techniques used the population mean of the data variable to replace the missing values, while k-NN calculates the population mean of k-nearest variables. Therefore, both methods produced similar results. The EM algorithm estimates values by using maximum likelihood technique. The EM algorithm results shown in Tables 1 and 2 fall in different distribution to the original distribution while this method can maintain the means and standard deviations. ANN imputation shows an increase in the number of data under the distribution curve. In addition, imputation techniques have shown that they are able to maintain the size of the datasets and also applicable for many data types including categorical and numerical data. It is important to note that imputing missing data with an inappropriate algorithm or technique can lead to biased, invalid or insignificant results. Hence it is vital to select an appropriate method specific for a particular dataset. A rule of thumb could be adopted to visualize the initial distribution of the data if the data is skewed or the data contains high percentages of missing values, then the single imputation method may not be appropriate.

Tables 5 and 6 show the results for various combinations of the imputation methods, feature selection methods and classification methods. It is important to note that the EM algorithm uses the Kullback-Leibler distance (KL)[48], which is also known as relative entropy. Relative entropy

defines a distance between two probability distributions, and thus imputes missing values. This process is similar to entropy ranking for feature selection. Results shown in Table 5 indicate that for only two classes, the precision and recall values are similar. However, unbalanced classes, i.e., the distributions of the two classes are not even, pose a challenge in terms of classification accuracy. This is a major issue with most clinical datasets where the observations are based on people with a particular ailment, and a good clinical system is always one where the number of alive patients far out weights the patients who succumb to the ailment. Table 6 shows the results when class of alive patients in further split into 6 classes of mortality months. Comparing the results from the two tables, it can be seen that, non-parametric classifier such as decision tree shows the most significant (precision and recall) results compared to parametric classifiers such as RBFN, MLP and k-means. The key point to note here is that the parametric methods are more suitable for data which is normally distributed. Further, considering one class (6 months) in Fig. 14, the decision tree classifier shows better performance on different feature selection methods and different imputations.

On further analysis of the results, it can be seen that the variables selected using the t-test reduction method, such as triglycerides, potassium, urea/uric acid, creatinine, NT-proBNP and sodium have strong associations with mortality of heart failure[73, 74]. Thus a conclusion can be drawn that this method provides the most suitable set of features. However, the results also indicate that all feature selection algorithms perform equally well; classification accuracy is improved in similar magnitudes. However, the clinical importance of the variables selected would result in a particular method being used. Yu and Liu[46] argue that in theory, more features should provide more power, argue that in theory, more features should provide more power, however, in practice an appropriate subset of features perform well as more features[45].

Feature selection depends on the nature of the distribution of data. The pre-processing step provides information on the data and a better understand of the nature of distribution of the data. This information allows for appropriate feature selection technique to be selected. The clustering algorithms employed in this study have shown that the dataset is structured in an unsupervised manner in order to simplify the process of information retrieval. This finding correlates with works by Bean and Kambhampati[62], where the authors exploited this notion by presenting knowledge extracted from real data in the form of a decision rule set with minimal ambiguity to support and aid in decision making. This was accomplished by employing clustering analysis and rough set theory, also explored the conceptual differences and similarities as well as the link between the two techniques[67].

It is well know that k-means[62] algorithm for clustering and classification has some issues, particularly as the results are dependent on the initial conditions. However, there are methods for selecting the correct initial conditions. In this paper, the method developed by Mirkin[67] has been employed. In this method, the number of clusters, k and number of centroids, $c_1, c_2, \cdots c_k$ are specified initially. Without this initialization, clustering can often produce misleading results as a result of inappropriate final centres and clusters. Mashor[75] suggests that k-means plays an important role in enhancing the performance of RBF, the algorithm determines the centres of the RBF. The location of the centres influences the performance of RBF networks. Obtaining accurate centres is important for RBF networks, for the activation function is dependent on the distance between the data and centres.

Hierarchical clustering suffers from a disadvantage that the quality of the dendrogram can be poor, for example once a merge (agglomerative) or split (divisive) decision has been completed, it is unfeasible to adjust or correct it. Agglomerative is known to perform remarkably slowly for large datasets due to the complexity of $O(n^3)$ where n is the number of objects[76].

7 Conclusions and future work

The methods illustrated in this paper have been applied to a heart failure dataset, and can be applied to various clinical datasets as these datasets present with similar issues. This paper has addressed some of the many challenges presented by clinical datasets. It has also showed how these can be handled using the current methods from statistics and data mining. The first challenge faced is that of missing values (Tables 1 and 2). There are several methods for handling this challenge. Often a preliminary exercise is to[37, 77] discard the variables with a large percentage of missing values, followed by imputing missing values (Tables 5 and 6). An alternative is to ignore missingness by analysing the incomplete data. Imputation techniques are essential if the original size of the dataset is to be retained, and if some useful information is to be extracted. In this paper, techniques for imputing missing values were outlined, these methods produce appropriate values for the missing data.

Table 1 shows the means and standard deviations from different types of imputation methods, these mean values are close to the expected mean value and are in confirmation with the law of large numbers[78]. When the sample size is small, imputation can have a dramatic effect than when the sample size is large.

In the framework (Fig. 1) provided in the paper, indeed in any data mining framework, after the initial pre-processing of the data, reduction of dimensions is almost a necessity. This paper outlined methods for reduction of dimensions. There are a wide variety of methods, which are broadly classified as feature extraction or feature selection. In most clinical applications, feature selection is more appropriate as it retains the variable labels and hence the final model is more meaningful. Features are selected based on a criterion, and often these are based around how effective the features are in performing the task of classification and prediction. In this paper, classification accuracy was selected as the criteria to assess the effectiveness of the feature selection methods. The classifier used were: Multilayer perceptron (back-propagation), J48 (decision tree), RBFN (neural network), SVM and random forest. From the results (Tables 5 and 6) it can be seen that both missing value imputation and feature selection do affect the result. However, the fundamental factor here is to understand the nature of the dataset in order to choose a suitable technique. An-

other issue that should be noted is the difference between supervised and unsupervised methods in mining of clinical datasets. These datasets have embedded within them numerous complexities and uncertainties in the form of class imbalances, missing values (which could be systematic). Supervised techniques show better results in the form of confusion matrix (precision and recall) than unsupervised techniques such as clustering (see Tables 5 and 6).

This paper has presented a framework for mining of clinical datasets. Currently research is being focused on ways to handle class imbalances within clinical datasets. Often in a clinical setting, the success of the clinic is judged on the number of patients who have recovered from illness and not the number that have succumbed to it. Thus real clinical datasets have a large imbalance, in that the class of live patients would far outweigh the number in the dead class. This imbalance affects imputation, feature selection and classification. Some preliminary results have been obtained and can be seen in [39, 40, 79].

References

[1] A. K. Tanwani, J. Afridi, M. Z. Shafiq, M. Farooq. Guidelines to select machine learning scheme for classification of biomedical datasets. In *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 128–139, 2009.

[2] A. K. Jha, C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, D. Blumenthal. Use of electronic health records in U. S. hospitals. *The New England Journal of Medicine*, vol. 360, no. 16, pp. 1628–1638, 2009.

[3] C. Safran, H. Goldberg. Electronic patient records and the impact of the internet. *International Journal of Medical Informatics*, vol. 60, no. 2, pp. 77–83, 2000.

[4] J. G. F. Cleland, K. Swedberg, F. Follath, M. Komajda, A. Cohen-Solal, J. C. Aguilar, R. Dietz, A. Gavazzi, R. Hobbs, J. Korewicki, H. C. Madeira, V. S. Moiseyev, I. Preda, W. H. van Gilst, J. Widimsky, N. Freemantle, J. Eastaugh, J. Mason, for the Study Group on Diagnosis of the Working Group on Heart Failure of the European Society of Cardiology, N. Freemantle, J. Eastaugh, J. Mason. The EuroHeart Failure survey programme — A survey on the quality of care among patients with heart failure in Europe, Part1: Patient characteristics and diagnosis. *European Heart Journal*, vol. 24, no. 5, pp. 442–463, 2003.

[5] U. R. Acharya, P. S. Bhat, S. S. Iyengar, A. Rao, S. Dua. Classification of heart rate data using artificial neural network and fuzzy equivalence relation. *Pattern Recognition*, vol. 36, no. 1, pp. 61–68, 2003.

[6] P. Shi, S. Ray, Q. F. Zhu, M. A. Kon. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics*, vol. 12, pp. 375, 2011.

[7] T. Mar, S. Zaunseder, J. P. Martinez, M. Llamedo, R. Poll. Optimization of ECG classification by means of feature selection. *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 8, pp. 2168–2177, 2011.

[8] M. Sugiyama, M. Kawanabe, P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, vol. 23, no. 1, pp. 44–59, 2010.

[9] P. Y. Wang, T. W. S. Chow. A new feature selection scheme using data distribution factor for transactional data. In *Proceedings of the European Symposium on Artificial Neural Networks*, ESANN, Bruges, Belgium, pp. 169–174, 2007.

[10] M. Dash, H. Liu, J. Yao. Dimensionality reduction of unsupervised data. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, Newport Beach, CA, USA, pp. 532–539, 1997.

[11] J. H. Chiang, S. H. Ho. A combination of rough-based feature selection and RBF neural network for classification using gene expression data. *IEEE Transactions on Nanotechnology*, vol. 7, no. 1, pp. 91–99, 2008.

[12] Z. G. Yan, Z. Z. Wang, H. B. Xie. The application of mutual information-based feature selection and fuzzy LS-SVM-based classifier in motion classification. *Computer Methods and Programs in Biomedicine*, vol. 90, no. 3, pp. 275–284, 2008.

[13] D. P Muni, B. R. Pal, J. Das. Genetic programming for simultaneous feature selection and classifier design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, pp. 106–117, 2006.

[14] E. Yom-Tov, G. F. Inbar. Feature selection for the classification of movements from single movement-related potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 3, pp. 170–177, 2002.

[15] R. Varshavsky, A. Gottlieb, D. Horn, M. Linial. Unsupervised feature selection under perturbations: Meeting the challenges of biological data. *Bioinformatics*, vol. 23, no. 24, pp. 3343–3349, 2007.

[16] J. C. Kelder, M. J. Cramer, J. Van Wijngaarden, R. Van Tooren, A. Mosterd, K. G. Moons, J. W. Lammers, M. R. Cowie, D. E. Grobbee, A. W. Hoes. The diagnostic value of physical examination and additional testing in primary care patients with suspected heart failure. *Circulation*, vol. 124, no. 25, pp. 1865–2873, 2011.

[17] J. C. Kelder, M. R. Cowie, T. A. McDonagh, S. M. Hardman, D. E. Grobbee, B. Cost, A. W. Hoes. Quantifying the added value of BNP in suspected heart failure in general practice: An individual patient data meta-analysis. *Heart*, vol. 97, no. 12, pp. 959–963, 2011.

[18] P. N. Peterson, J. S. Rumsfeld, L. Liang, N. M. Albert, A. F. Hernandez, E. D. Peterson, G. C. Fonarow, F. A. Masoudi. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 1, pp. 25–32, 2010.

[19] K. D. Min, M. Asakura, Y. L. Liao, K. Nakamaru, H. Okazaki, T. Takahashi, K. Fujimoto, S. Ito, A. Takahashi, H. Asanuma, S. Yamazaki, T. Minamino, S. Sanada, O. Sequchi, A. Nakano, Y. Ando, T. Otsuka, H. Furukawa, T. Isomura, S. Takashima, N. Mochizuki, M. Kitakaze. Identification of genes related to heart failure using global gene expression profiling of human failing myocardium. *Biochemical Biophysical Research Communications*, vol. 393, no. 1, pp. 55–60, 2010.

[20] R. A. Damarell, J. Tieman, R. M. Sladek, P. M. Davidson. Development of a heart failure filter for Medline: An objective approach using evidence-based clinical practice guidelines as an alternative to hand searching. *BMC Medical Research Methodology*, vol. 11, pp. 12, 2011

[21] D. S. Lee, L. Donovan, P. C. Austin, Y. Y. Gong, P. P. Liu, J. L. Rouleau, J. V. Tu. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Medical Care*, vol. 43, no. 2, pp. 182–188, 2005.

[22] D. S. Lee, P. C. Austin, J. L. Rouleau, P. P. Liu, D. Naimark, J. V. Tu. Predicting mortality among patients hospitalizeed for heart failure, derivation and validation of a clinical model. *Journal of the American Medical Association*, vol. 290, no. 19, pp. 2581–2587, 2003.

[23] I. Holme, T. R. Pedersen, K. Boman, K. Egstrup, E. Gerdts, Y. A. Kesäniemi, W. Malbecq, S. Ray, A. B. Rossebø, K. Wachtell, R. Willenheimer, C. Gohlke-Bärwolf. A risk score for predicting mortality in patients with asymptomatic mild to moderate aortic stenosis. *Heart*, vol. 98, no. 5, pp. 377–383, 2011.

[24] K. K. L. Ho, G. B. Moody, C. K. Peng, J. E. Mietus, M. G. Larson, D. Levy, A. L. Goldberger. Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation*, vol. 96, no. 3, pp. 842–48, 1997.

[25] G. C. Fonarow, W. T. Abraham, N. M. Albert, W. G. Stough, M. Gheorghiade, B. H. Greenberg, C. M. O′Connor, K. Pieper, J. L. Sun, C. Yancy, J. B. Young. Association between performance measures and clinical outcomes for patients hospitalized with heart failure. *Journal of the American Medical Association*, vol. 297, no. 1, pp. 61–70, 2007.

[26] J. Bohacik, D. N. Davis. Data mining applied to cardiovascular data. *Journal of Information Technologies*, vol. 3, no. 2, pp. 14–21, 2010.

[27] J. Bohacik, D. N. Davis. Alert rules for remote monitoring of cardiovascular patients. *Journal of Information Technologies*, vol. 5, no. 1, pp. 16–23, 2012.

[28] J. Bohacik, D. N. Davis. Estimation of cardiovascular patient risk with a Bayesian network. In *Proceedings of the 9th European Conference of Young Research and Scientific Workers*, University of Žilina, Žilina, Slovakia, pp. 37–40, 2011.

[29] A. Jain, D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[30] Y. Saeys, T. Abeel, Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 313–325, 2008.

[31] L. Yu, H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 856–863, AAAI, Washington DC, USA, 2003.

[32] N. Zhou, L. Wang. A modified T-test feature selection method and its application on the HapMap genotype data. *Genomics, Proteomics & Bioinformatics*, vol. 5, no. 3–4, pp. 242–249, 2007.

[33] U. M. Fayyad, K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1022–1029, 1993.

[34] H. Liu, J. Li, L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, vol. 13, pp. 51–60, 2002.

[35] C. N. Hsu, H. J. Huang, S. Dietrich. The ANNIGMA-wrapper approach to fast feature selection for neural nets. *IEEE Transactions Systems, Man, and Cybernetics, Part B*, vol. 32, no. 2, pp. 207–212, 2002.

[36] J. Bohácik, D. N. Davis, M. Benediković. Risk estimation of cardiovascular patients using Weka. In *Proceedings of the International Conference OSSConf 2012*, (The Society for Open Information Technologies — SOIT in Bratislava, Slovakia, Žilina, Slovakia), pp. 15–20, 2012.

[37] E. Acuña, C. Rodriguez. The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering, and Data Mining Applications*, D. Banks, L. House, F. R. McMorris, P. Arabie, W. Gaul, Eds., Berlin, Heidelberg: Springer, pp. 639–648, 2004.

[38] J. H. Lin, P. J. Haug. Data preparation framework for preprocessing clinical data in data mining. In *Proceedings of AMIA Annual Symposium*, AMIA, American, pp. 489–493, 2006.

[39] N. Poolsawad, C. Kambhampati, J. G. F. Cleland. Feature selection approaches with missing values handling for data mining — A case study of heart failure dataset. *World Academy of Science, Engineering and Technology*, vol. 60, pp. 828–837, 2011.

[40] N. Poolsawad, L. Moore, C. Kambhampati, J. G. F. Cleland. Handling missing values in data mining — A case study of heart failure dataset. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Chongqing, China, pp. 1934–2938, 2012.

[41] W. J. Frawley, G. Piatetsky-Shapiro, C. J. Matheus. Knowledge discovery in databases: An overview. *Artificial Intelligence Magazine*, vol. 13, no. 3, pp. 57–70, 2011.

[42] Analysis Factor. EM Imputation and Missing Data: Is Mean Imputation Really so Terrible? [Online], Available: http://www.analysisfactor.com/statchat/tag/spss-missing-values-analysis, 30 August 2011.

[43] E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, M. D. Cubiles-de-la-Vega. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.

[44] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*, 2nd ed., San Francisco: Morgan Kaufman Publishers, 2006.

[45] D. W. Aha, R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In *Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics*, pp. 1–7, 1995.

[46] L. Yu, H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[47] T. Jirapech-Umpai, S. Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, vol. 6, pp. 148, 2005.

[48] F. M. Coetzee. Correcting the Kullback-Leibler distance for feature selection. *Pattern Recognition Letters*, vol. 26, no. 11, pp. 1675–1683, 2005.

[49] B. L. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Y. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[50] I. Levner. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, vol. 6, pp. 68, 2005.

[51] J. Jäeger, R. Sengupta, W. L. Ruzzo. Improved gene selection for classifcation of Microarrays. *Pacific Symposium on Biocomputing*, vol. 8, pp. 53–64, 2003.

[52] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, S. Kasif. RankGene: Identification of diagnostic genes based on expression data. *Bioinformatics*, vol. 19, no. 12, pp. 1578–1579, 2003.

[53] The University of Waikato. WEKA: The Waikato Environment for Knowledge Acquisition. [Online], Available: http://www.cs.waikato.ac.nz/ml/weka, 30 August 2011.

[54] M. W. Gardner, S. R. Dorling. Artificial neural networks (the multilayer perceptron) — A review of applications in the atmospheric sciences. *Atmospheric Environment*, vol. 32, no. 14–15, pp. 2627–2636, 1998.

[55] L. Autio, M. Juhola, J. Laurikkala. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. *Computers in Biology and Medicine*, vol. 37, no. 3, pp. 388–397, 2007.

[56] A. Khemphila, V. Boonjing. Parkinsons disease classification using neural network and feature selection. *World Academy of Science, Engineering and Technology*, vol. 64, pp. 15–18, 2012.

[57] C. Cortes, V. Vapnik. Support-vector networks. *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[58] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods — Support Vector Learning*, B. Schoelkopf, C. Burges, A. Smola, Eds., Cambridge, MA, USA: MIT Press, pp. 185–208, 1998.

[59] T. Hastie, R. Tibshirani. Classification by pairwise coupling. *Advances in Neural Information Processing Systems*, Cambridge, MA, USA: MIT Press, pp. 507–513, 1998.

[60] L. Breiman. Random forests. *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[61] W. D. Kim, H. K. Lee, D. Lee. Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1263–1271, 2004.

[62] C. L. Bean, C. Kambhampati. Knowledge-oriented clustering for decision support. In *Proceedings of the International Joint Conference on Neural Networks*, IEEE, Portland, OR, USA, pp. 3244–3249, 2003.

[63] M. Steinbach, G. Karypis, V. Kumar. A comparison of document clustering techniques. In *Proceedings of KDD Workshop on Text Mining*, pp. 1–2, 2000.

[64] Z. X. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[65] T. Kanungo, M. D. Mount, S. N. Netanyahu, D. C. Piatko, R. Silverman, Y. A. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[66] K. Alsabti, S. Ranka, V. Singh. An efficient k-means clustering algorithm. In *Proceedings of IPPS/SPDP Workshop on High Performance Data Mining*, pp. 1–7, 1998.

[67] B. Mirkin. *Clustering for Data Mining: A Data Recovery Approach*, Florida: Chapman and Hull/CRC, 2005.

[68] A. Sridhar, S. Sowndarya. Efficiency of k-means clustering algorithm in mining outliers from large data sets. *International Journal on Computer Science and Engineering*, vol. 2, no. 9, pp. 3043–3045, 2010.

[69] D. Napoleon, G. P. Lakshmi. An efficient k-means clustering algorithm for reducing time complexity using uniform distribution data points. In *Proceedings of the Trendz in Information Sciences & Computing*, IEEE, Chennai, India, pp. 42–45, 2010.

[70] Y. Zhao, G. Karypis, U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, vol. 10, no. 2, pp. 141–168, 2005.

[71] J. S. J. Lee, J. N. Hwang, D. T. Davis, A. C. Nelson. Integration of neural networks and decision tree classifiers for automated cytology screening. In *Proceedings of the IJCNN-91-Seattle International Joint Conference on Neural Networks*, IEEE, Seattle, WA, USA, vol. 1, pp. 257–262, 1991.

[72] Y. Zhang, C. Kambhampati, D. N. Davis, K. Goode, J. G. F. Cleland. A comparative study of missing value imputation with multiclass classification for clinical heart failure data. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Sichuan, China, pp. 2840–2844, 2012.

[73] Y. Al-Najiar, K. M. Goode, J. Zhang, J. G. Cleland, A. L. Clark. Andrew. Red cell distribution width: An inexpensive and powerful prognostic marker in heart failure. *European Journal Heart Failure*, vol. 11, no. 12, pp. 1155–1162, 2009.

[74] Atherotech Diagnotics Lab. Atherotech Panels. [Online], Available: http://www.atherotech.com/athdiagtests/atherotechpanels.asp, 13 June 2011.

[75] M. Y. Mashor. Improving the performance of k-means clustering algorithm to position the centres of RBF network. *International Journal of the Computer, the Internet and Management*, vol. 6, no 2, 1998.

[76] J. Herrero, A. Valencia, J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, vol. 17, no. 2, pp. 126–136, 2000.

[77] W. R. Myers. Handling missing data in clinical trials: An overview. *Drug Information Journal*, vol. 34, no. 2, pp. 525–533, 2000.

[78] C. M. Grinstead, J. L. Snell. *Introduction to Probability*, Rhode Island: American Mathematical Society, 1998.

[79] M. M. Rahman, D. N. Davis. Machine learning-based missing value imputation method for clinical datasets. *IAENG Transactions on Engineering Technologies*, Netherlands: Springer, pp. 245–257, 2013.

Nongnuch Poolsawad received her B. Sc. degree in computer science from the University of the Thai Chamber of Commerce (UTCC), her M. Sc. degree in computer science at the Mahidol University, Thailand. In master degree, her research area is database security and encryption models. She is currently working toward her Ph. D. degree in the area of computer science at University of Hull, UK. She has been funded by National Metal and Materials Technology Center, National Science and Technology Development Agency. Her role is engineer in management information system section.

Currently, she belongs to Intelligent Systems Research Group, focuses on decision support and data mining in tele-health. Her current project is selecting significant variables in very large clinical datasets: The research aims to establish a novel feature selection technique for selecting the significant variables and provide the practical data mining framework to achieve the efficiency of classification by using data mining techniques instead of the specific knowledge from clinical experts.

Her research interests include data mining on big data, handling missing values, imbalanced classes handling techniques and data classification.

E-mail: N.Poolsawad@2008.hull.ac.uk (Corresponding author)

Lisa Moore received her B. Sc. degree in forensic biology from the University of Westminster, UK, her M. Sc. degree in Analytical Genetics at the University of Birmingham, UK. She is currently working toward her Ph. D. degree in the area of computer science at University of Hull, UK.

She has published a few papers in international journals and conferences. She is currently an IEEE student member. She is currently the postgraduate research representative at University of Hull and has participated in organizing and planning the department′s conference in 2012. She has gained work experience in the areas of biology, bioinformatics and contributed her computer science knowledge to undergraduates by taking on the role of a demonstrator.

Her research interests include pattern recognition, machine learning, reasoning under uncertainty, artificial intelligence, data mining, bioinformatics, very large scale integration and dealing with real-world clinical data for decision support systems.

E-mail: Lisa.Moore@2011.hull.ac.uk

Chandra Kambhampati is a reader in computer science. He has published 125 papers in international journals and conferences in architectures of neural networks, and their applications for complex control. He was an investigator on a number of EPSRC funded projects which investigated intelligent predictors for power systems, and neural network based control of nonlinear systems. His research offered both theoretical and practical advances to the management of power systems, and the intelligent control of nonlinear systems. In addition, he was involved with Predictive Control Ltd in the development of intelligent controllers. This work lead to the first UK based and marketed intelligent control solution for chemical processes and was incorporated into "Connoiseur".

His research interests include nonlinear control, modelling of learning systems and neurons. Currently his research in telehealth and medical informatics is sponsored by both the EU (FP-7 Network of Excellence - SemanticHealth Network, FP7 Integrated Project Braveheart) and by industry (Phillips Health care).

E-mail: C.Kambhampati@hull.ac.uk

John G. F. Cleland qualified in medicine in 1977 at University of Glasgow. After a period of postgraduate training and an introduction to research he was appointed from 1986–1994 first as a senior registrar and subsequently as senior lecturer in cardiology and honorary consultant cardiologist at St Mary′s Hospital, Paddington and the Hammersmith Hospital, London. In 1994 He was awarded a Senior Research Fellowship by the British Heart Foundation to transfer to the Medical Research Council′s Clinical Research Initiative in Heart Failure. He was appointed to the Foundation Chair of Cardiology at University of Hull in 1999.

He heads The Academic Unit of Cardiology that includes a reader, 3 senior lecturers and a team of basic and clinical scientists, technicians and research nurses dedicated to the above research programme.

His research interests include heart failure, extending from its epidemiology, detection and prevention, through the development and implementation of guidelines for the application of current knowledge, to large randomised trials to study new (and old) treatments heart failure. Particular current interests include the role of myocardial hibernation contributing to heart failure and its treatment (including beta-blockers and revascularisation), "diastolic" heart failure, vascular dysfunction, the potential deleterious effect of aspirin in heart failure, ventricular resynchronisation, telemonitoring, implantable haemodynamic monitoring devices, co-morbidities including diabetes, anaemia, atrial fibrillation and renal dysfunction and new interventions for acute decompensated heart failure. Active programmes for the assessment of heart failure and its optimal management using cardiac impedance, magnetic resonance, computer tomography and advanced electrophysiology are also in place.

E-mail: J.G.Cleland@hull.ac.uk