World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:5, No:12, 2011

Feature Selection Approaches with Missing Values Handling for Data Mining - A Case Study of Heart Failure Dataset

N.Poolsawad, C.Kambhampati, and J. G. F. Cleland

***Abstract*—**In this paper, we investigated the characteristic of a clinical dataseton the feature selection and classification measurements which deal with missing values problem.And also posed the appropriated techniques to achieve the aim of the activity; in this research aims to find features that have high effect to mortality and mortality time frame. We quantify the complexity of a clinical dataset. According to the complexity of the dataset, we proposed the data mining processto cope their complexity; missing values, high dimensionality, and the prediction problem by using the methods of missing value replacement, feature selection, and classification.The experimental results will extend to develop the prediction model for cardiology.

***Keywords*—**feature selection, missing values, classification, clinical dataset, heart failure.

I. Introduction

RECENTLY,data mining has evolving area in information technology. Hundreds of novel mining algorithms and new applications such as medicine have been proposed play for a role to improve the quality of healthcare. The aim of data mining is to extract knowledge from data. The information and knowledge mined from the large quantities must be meaningful enough to lead to some advantages. The information and knowledge mined from the large quantities must be meaningful enough to lead to some advantages. Clinical datasets posed a unique challenge to data mining algorithms for classification because of their various systematic and human errors, their high dimensionality, multiple classes, noisy data and missing values [1].Currently large amounts of clinical data are available; however accurate models for predicting survivability of patients with heart failure are not extensively available. Thus effective planning for the treatment and medical care for patients with heart failure has proven to be elusive. Identifying good and robust predictive models has proven to be a difficult problem due to the nature of the clinical data that is available.

N. Poolsawad is with DRIS,Department of Computer Science, University of Hull, Cottingham Road, Hull, United Kingdom and is funded by the National Science and Technology Development Agency, Ministry of Science and Technology, Royal Thai Government (e-mail: N.Poolsawad@2008.hull.ac.uk).

C. Kambhampati is with DRIS,Department of Computer Science, University of Hull, Cottingham Road, Hull, United Kingdom (e-mail: C.Kambhampati @hull.ac.uk).

J. G. F. Cleland is with Department of Cardiology, HYMS, University of Hull, Cottingham Road, Hull, United Kingdom(e-mail: J.G.Cleland@hull.ac.uk).

This data is often extremely complex; in that there are extremely large numbers of variables, unbalanced classes in which one class is represented by large number of samples while the other is presented by a few numbers, a great deal of missing data and non-normally distributed data.

In this paper, the problem of high dimensionality in clinical data sets is not only investigated, but the properties of the various feature selection schemes are investigated vis à vis the data sets. The choice of technique is dependent on the nature of the final solution that is required. This paper provides a comprehensive evaluation of a set of diverse machine learning scheme on a clinical datasets. The paper aims to investigate and select the suitable techniques for clinical dataset. In this case we use the heart failure dataset, we will find to select the techniques and relevant and significant features to develop the prognostic model for the decision support system to be practically useful for stratifying patient-risk they need to be based on predictors and able to predict mortality event in a clinically relevant time frame.We set out to find out whether or not the problems encountered by feature selection are in clinical dataset.

In the first part, the data mining process of clinical dataset for finding the potential feature to be predictors for prognostic model for designing the treatment for patient due to heart failure, these processes compose of pre-processing, feature selection, classification, and evaluation. Later, feature selection techniques are surveyed and discussed. Three feature selection methods are looked into, t-Test [2], entropy ranking [3, 4] and nonlinear gain analysis [5], these methods use afeature importance measure according to its discriminative capability. The rationale for this selection is that the three techniques use different properties of the data to select feature. The t-Test method utilizes data distribution as a key property for selecting variables, the entropy method not only uses the distribution it also includes a measure for density of data and the and develops a measure for the degree of order in the data, whilst the last method is a wrapper technique, which enables the lack of balance in data to be overcome. Next, the experiments that implemented by along with feature selection techniques, missing value replacement method and classification. The results are discussed in the context of the characteristics and problems with the clinical datasets. The results are thus used the problems associated with clinical datasets by establishing a relationship between the complexity, the set of features being selected and data distribution. In doing so, we also identify the relationship between thefeature

selection techniques and data distribution. In particular, we attempt to establish procedures based on different subsets of features (variables) that are selected, and then tested on their ability to discriminate the classes present. The set of features that is the appropriate one is the one with the highest performance of classification and achieves the aim of the research.

II. Clinical Dataset

Clinical dataset in this paper, we are study on the heart failure dataset. It has diverse clinical features and numerous clinical subsets. There is no widely accepted characterization and definition of heart failure, probably because of the complexity of the syndrome [6]. High-risk candidates for heart failure need to be targeted for evaluation and treatment in a cost-effective manner [7].The dataset called, LIFELAB is used for this purpose and is a large cardiological database. LIFELAB is a prospective cohort study consisting of patients who were recruited from a community-based outpatient clinical based in England (the University of Hull Medical Centre, UK). This dataset presents the incidence, prevalence and persistence of heart failure, and the dataset routinely collected clinical data could be used for research purposes. This dataset contains both longitudinal and horizontal data across generations. This data set is composed of both 463 variables, which are continuous and categorical, collected for 2,032 patients. The variables consist of physiologic and symptomatic variables, e.g. blood testing, data of death, electrocardiograms (ECG), quality of life, drugs and history baseline. However, LIFELAB is a large clinical dataset reveals that many problems for data mining process. The challenge to apply data mining to clinical dataset is to convert data into an appropriate form for the activity's aim achievement. From our investigation we can split the challenges into the topics are as follows:

A. Incomplete, errors and noisy data

Raw clinical data in data storage can be incomplete, errors and noisy data. Inconsistent data can exist for instance the variable have to specific value, but another might enter as free text. Commonly problem of outliers due to entry errors is found. The variables are related on this problem, there was then manually inspected to remove obviously irrelevant variables.

B. Missing values

Clinical data values often are not collect for all data, but there will collect only the data that required for personalized analysis. So this problem will be the main issue that we are focusing on because it will lead to have a high misclassification value. Methods of data imputation [8] and missing value replacement are necessary to cope with this issue.

C. Diverse clinical features and their scales

The features appeared in the dataset approximately 400 features, it has many scales of measurement. Some variables are contains the integer, some variables are contained the decimal. Their some scales are wide range, and some ranges are small. The normalization will apply for solving this problem to manage the data elements in the data into the same scale for preparing to apply data mining.

D. Large and high dimensionality

From the issue of diverse of features, then the size of dataset is large and high dimensionality. When the dataset has too many features, the features should be reduced. But how to reduce the features or variables, which features should be removed and which features should be kept. Feature selection will be the efficient method to cope this issue. And also this technique will keep the meaningful of the features then we can use the selected features to be the predictor for prediction model.

E. The prediction problem

The goal of the data mining for health care system is to assist clinicians and improve the quality of prognosis and/or diagnosis.And especially can facilitate the timelines of medical problem. The target problems were extracted from the dataset using the data mining process is the prediction of the mortality and mortality time frame of patients due to heart failure. The machine learning of neural network and decision tree will apply to be classifier

III. Data Mining Process

A. Data Mining Procedure for Clinical Dataset

The procedure follows a four steps methodology 1) pre-processing the datasets to remove any redundant data and eliminating not useful variables for this work for example free text, remarks, etc. by manual removing. For missing values problem, we handle by using four different missing value replacement methods. And also normalization process to scale the data into the small range of data.2) Three feature selection techniques; t-Test, Entropy and NLGA. In this paper this step aims to select most relevant features and reduce the size of dataset. And then considering the appropriated features by evaluating classification performance measures, 3) Classifier; multilayer perceptron (back-propagation) and J48(decision tree) are used for classification the outcomes of mortality consists of dead/alive and period of dead, the results can be shown the performance of classificationfrom the different techniques of missing value replacement methods, feature selection and classifier.

B. Pre-processing

Data pre-processing is always the first step in the data mining process, this process is required before one can apply data mining to clinical data [9]. If this process without getting to know data carefully in advance, the classification task could be misleading and inaccuracy. First, the whole data sets were dealt with missing value replacement scheme. Then, normalizeddata into small scale before fed into a feature selection process sequentially.The Figure 1 shows the detail workflow that used in this paper.

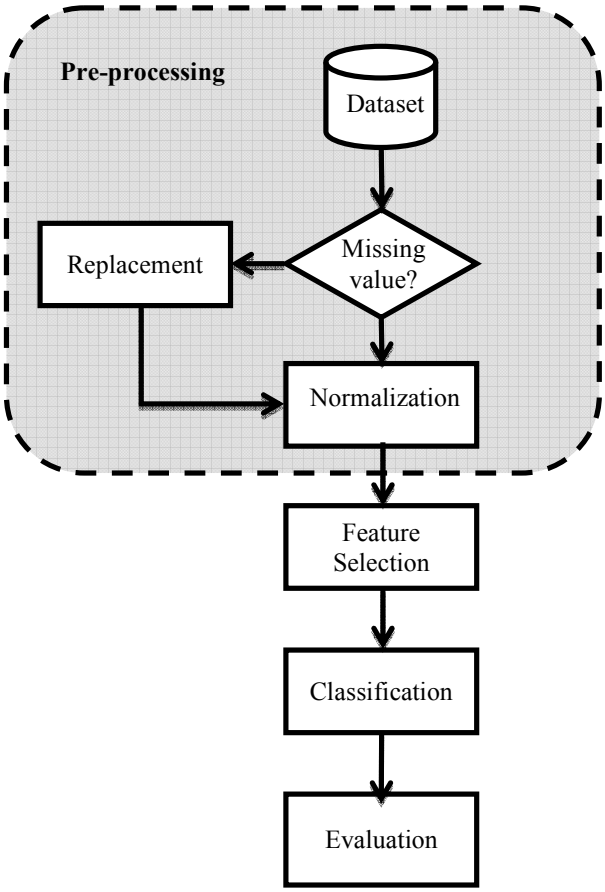

Fig. 1 The data mining procedure of clinical dataset

1) Missing value replacement

Since most data sets encountered in practice contain missing values and most learning schemes lack for ability to handle these data sets, we have replaced missing values with the missing values. Cleaning data is used before adopt the missing value replacement method, and also considering the percentage of missing values for each variable in case these variables appear missing more than 20%, these variables have been ignored. In this paper the missing values issue handled by replacement or imputation methods; mean imputation, expectation-maximization (EM) algorithm, k-nearest neighbor (k-NN) imputation, and artificial neural network (ANN) imputation have been applied to treat this.

2) Normalization

Normalization or scaling data to be in the same scale, this paper normalized data between 0 and 1. In order to prevent attributes with large numeric ranges dominate those with small numeric ranges, data instances are rescaled between 0 and 1 using min–max normalization procedure. The min-max normalization procedure performs a linear transformation of the original input range into a new specified range. The old minimum min_old is mapped to the new minimum min_new (i.e., 0) and max_old is mapped to max_new (i.e., 1), as shown in equation(1).

$$\mathrm{New}_{\mathrm{value}} = \frac{\mathrm{Original}_{\mathrm{value}} - \mathrm{Min}_{\mathrm{old}}}{\mathrm{Max}_{\mathrm{old}} - \mathrm{Min}_{\mathrm{old}}}(\mathrm{Max}_{\mathrm{new}} - \mathrm{Min}_{\mathrm{new}}) + \mathrm{Min}_{\mathrm{new}} \quad (1)$$

IV. Feature Selection

Feature selection (also known as subset selection) is a process that selects the most relevant attributes and tries to find the best subset of the input feature set. Feature selection attempts to reduce the number of dimensions considered in a task so as to improve performance on some dependent measure. A general feature selection algorithm is often composed of three components: an evaluation function, a performance function and a search algorithm. The evaluation function inputs a feature subsets and outputs numeric evaluation. The performance function gives the subsets that perform best of classifier. There are three categories of search algorithm, i.e., exponential, randomized and sequential. Feature selection has two models: one is a wrapper model and the other is a filter model. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subset. In wrappermethods, the learning algorithm itself is run with various subsets of features and the learner that performs best is chosen. In filter methods for feature selection, the data with the chosen subset of features is then presented to a learning algorithm. It separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm.

A key aspect needs to be considered when selective subset of features is the metrics of feature relevance and feature redundancy. An optimal subset of features should be obtained by using set of strong relevant features and weakly relevant features but non-redundant feature [10] and by a selected features that have a positive Z-score [11]. Different criterion, e.g., statistical correlation or mutual information, will lead to different inputs and algorithms, which in turn will give different subsets of features.

A. Nonlinear Gain Analysis (NLGA)

Nonlinear Gain Analysis (NLGA) is an approach of feature subset selection and is also known as Artificial Neural Net Input Gain Measurement Approximation (ANNIGMA) ranks features [5]. Neural networks are suitable for training large amount of data and it is an unsupervised learning, the variables that are higher weight is more important. The data flow of the NLGA is shown in Fig.2and Fig.3demonstrates the architecture of the neural network. NLGA consists of training process and calculating the ANNIGMA score as follows.

$$LG_{ik} = \sum_{j} \left| w_{ij} \times w_{jk} \right| \quad (2)$$

$$ANNIGMA_{ik} = \frac{LG_{ik}}{max(LG_{ik})} \times 100 \quad (3)$$

wherei, j, k are the input, hidden, and output layers node indicates, respectively. $LG_{ik}$is the local gain of all the other inputs. w_{ij}andw_{jk} are the weights between the layers.

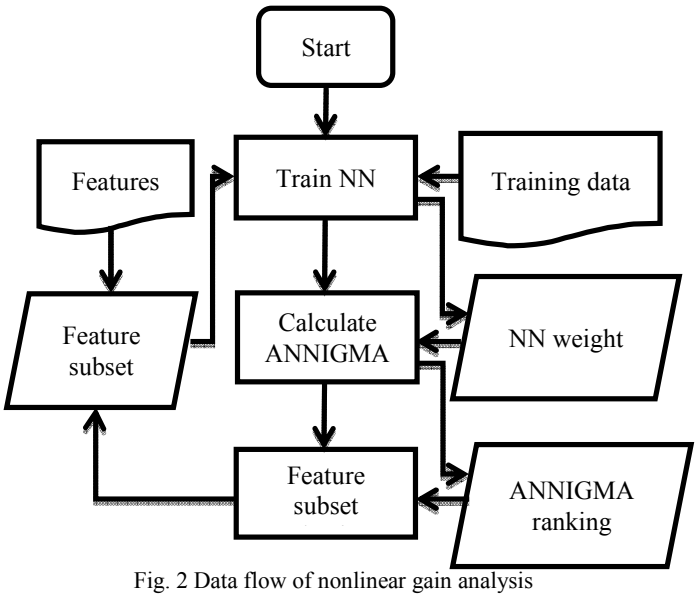

Fig. 2 Data flow of nonlinear gain analysis

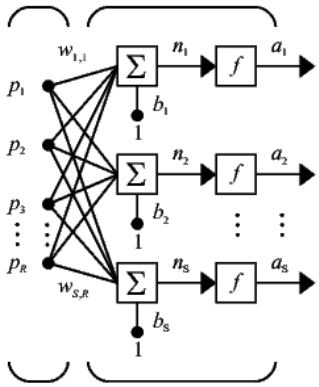

Fig. 3 Network architecture: A single layer of neurons

The training cycle takes a feature set as input. The process of calculate the ANNIGMA score is used for ranking of the features. The NLGA is the wrapper model that appropriated for the large volumes of data and many features, for these reason this method would be suitable to apply in the clinical dataset. And this approach can efficiently reduce the number of features and maintaining or even improving the accuracy. But it needs to improvement in speed because it needs to train the neural network in each points and needs to find the ANNIGMA scores ranks neural network features by relevance. It can be expensive application in real-world application.

B. t-Test Method

This method is used in genotype research [2, 4]. The statistical tools is t-Test, namely the Student's t-test [2, 4, 12] is often used to assess whether the means of two classes are statistically different from each other by calculating a ratio between the difference of two class means and the variability of the two classes.The *t*-Test has been used to rank features. These uses of t-Test are limited to two class problems. For multi-class problems, calculated a t-statistics value follows the equation [2, 4, 12] for each feature of each class by evaluating the difference between the mean of one class and the mean of all the classes, where the difference is standardized by the within-class standard deviation. From the investigation of the characteristics of dataset, the data distribution of each feature various into many tend of distributed consists of normal distribution and non-normal distribution. After applied the missing value replacement, the statistic value that shown in Table II, the mean and standard deviation are similar. This method is related to these two value so the selected features from this method will suitable for this dataset, but for some variables that were not tend to be normal distribution it may concerns.

C. Entropy Ranking

This method is used for selecting subset of features and has been used in many datasets such as gene, waveform, and echocardiogram data. Fayyad [3] presents the cut point selection by using class entropy of subset. For the remaining features, this method can automatically find out points in these features' value ranges such that the resulting expression intervals of every feature can be maximally distinguished. If every expression interval induced by the cut points of a feature contains only the same class of samples, then this partitioning by the cut points of this feature has an entropy value of zero. The class entropy of a subset S is defined as:

$$Ent(S) = -\sum_{i=1}^{k} \mathcal{P}(C_i, S)\, log\big(P(C_i, S)\big) \qquad (4)$$

when logarithm base is 2, Ent(S) measures the amount of information needed, in bit, to specify the classes in S. Sis a set of attribute and $P(C_i, S)$ be the proportion of examples in S that have class C_i. We sort the values of entropy in an ascending order and consider those features with lowest entropy values.Feature selection has been successfully applied to clinical dataset e.g., lymphoma, gene expression, cancers [2, 4,13, 14].Aha [15] claimed that feature selection consistently increased accuracy, reduced feature set size, and provided better accuracy of classification. Liu [4] said feature selection played an important role in classification. Effective in enhancing learning efficiently increasing productive accuracy, and reducing complexity of learning results learning can be achieved more efficiently and effectively with just relevant and non-redundant features.

V. Experiments

Aim of this research is to investigate the clinical dataset to explore the appropriate techniques for clinical dataset. These experiments are implemented and tested on a real clinical dataset. The proposed is to classify data to predict the death from the record of patient.

A. Dataset

The clinical dataset called "LIFELAB" is used to test the methods. It contains the patient's information due to heart failure. The dataset called, LIFELAB is used for this purpose and is a large cardiological database, which contains both longitudinal and horizontal data across generations. Datasets use for this paper split into two groups of dataset follow the classes of mortality as shown in Table I. The variables consist of physiologic and symptomatic variables, e.g. blood testing, data of death, electrocardiograms (ECG), quality of life, drugs and history baseline.

B. Missing Values Handling

The data in this dataset consists of both useful and unusable data, for the unusable data while exploring in this stage and shows the number of variables and the percentage due to each problem. The variables that are unusable would be removed excluding the problem of missing values because for this dataset the missing values is the main problems then we will need to handle a missing values in the next process.

TABLE I
CLASSES OF DATASETS

Group	Dataset	Class	Category	Frequency	%
1	Dead/Alive	2	Alive	698	66.41
			Dead	353	33.59
2	Dead months	6	6M	89	17.9
			12M	75	15.1
			18M	54	10.87
			24M	61	12.27
			36M	66	13.28
			>36M	152	30.58

1) Mean Imputation

This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. Mean imputation [8] makes only a trivial change in the correlation coefficient and no change in the regression coefficient. That should not be surprising. We have really added no new information to the data but we have increased the sample size. The effect of increasing the sample size is to increase the denominator for computing the standard error, thus reducing the standard error.

2) Expectation-Maximization (EM) Algorithm

EM imputations method is to estimate the covariance matrix and impute values,and also better than mean imputations because they preserve the relationship with other variables, which is vital if you go on to use something like Factor Analysis or Regression. They still underestimate standard error, however, so once again, this approach is only reasonable if the percentage of missing data are very small [16].It is an interactive procedure in which it uses other variables to impute a value (Expectation), then checks whether that is the value most likely (Maximization).

3) k-Nearest Neighbor (kNN) Imputation

Impute missing data using nearest-neighbor method [8]. This method the missing values of anvariable are imputed considering a given number of variables that are most similar to the instance of interest.The similarity of two instances is determined using a distance function.

4) Artificial Neural Network (ANN)

The method that using neural network for predicting the missing values. An ANN [17] is an interconnected assembly of nodes (neurons). The processing ability of the neural network is stored in the inter unit connection strengths, or weights, obtained by a process of learning from aset of training patterns.

TABLE II
THE STATISTIC OF FERRITIN BEFORE AND AFTER MISSING VALUE HANDLING BY DIFFERENT METHODS

Variable name: Ferritin				
Missing values:	219 (20%)			
Missing value replacement method	Statistic			
	Mean	Std. Deviation	Unique	%Unique
Original	0.048	0.058	128	12
Mean imputation	0.048	0.051	128	12
EM Algorithm	0.114	0.05	345	33
kNN imputation	0.48	0.052	116	11
ANN imputation	0.131	0.058	345	33

TABLE III
THE DATA DISTRIBUTIONS OF FERRITIN GO ALONG WITH THE METHOD OF MISSING VALUE REPLACEMENT

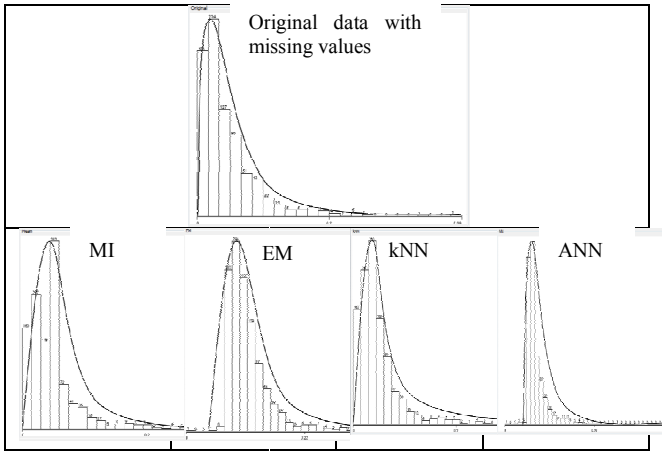

Table II shows the variable; Ferritin, it is with the most missing values approximately 20%. This table compared the statistical values between original data with missing values and the data that treated by different missing value imputation methods. Missing values problem is the major problem because it has effect to the reduction feature and classification processes so mean imputation was employed to solve this problem. Table III draws the data distribution from different

techniques. The mean imputation uses mean of data variable and replace for the missing values then after treat by using mean the standard deviation (σ) of variable will change but mean is not change. Also a compared σ between before and after handle missing value by mean imputation (no missing value), the σ decreased after applied the mean imputation when compared with variable with missing values. Obviously, considering the data distributions are different distributed, especially ANN imputation thatusing the observed data to be input and used target data. However the techniques of missing values handling that have various, and different techniques will give different results then this topic will be rise an issue to be the research of interest.

C. Classifiers

We implemented the classification schemes that provide the standard implementations in Wakaito Environment for Knowledge Acquisition (WEKA) [18].

1) Multilayer Perceptron (Back-propagation)

Multilayer perceptron is a feed-forward neural network based classifier that usesback-propagation to classify instances. All the nodes in this network are sigmoids, whichmeans that the activation function is a sigmoid.In a multilayer perceptron, there is an input layer with a node each for all theindependent variables, at least one hidden layer and an output layer with a node eachfor different classes of the target variable. In the paper, a feed-forward network consisting of input units, hidden neurons, and only 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 of the number of input attributes. All weights are randomly initialized to a number near 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, we update the weights on the arcs by a gradient descent method to minimize the squared error between the network values and the target values.

2) J48 (Decision Tree)

Generate a decision tree C4.5 algorithm for classification.Whenever a set of items (training set) is encountered, the algorithm identifiesthe attribute that discriminates the various instances most clearly. This is done using the standard equation of information gain. Among the possible values of this feature, if there is any value for which there isno ambiguity, that is, for which the data instances falling within its category havethe same value for the target variable, then that branch is terminated and theobtained target value is assigned to it.

D. Performance evaluation measures

Any single performance estimator suffers the risk of being fitted if we compare many classifiers based on the estimators. Thus, we carefully used five measures to evaluate the performance, which are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

whereTP 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 correctly classified examples (true positives)

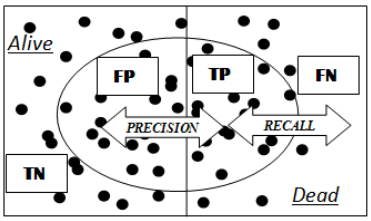

Fig. 4 A relationship between precision and recall values of classification

and the misclassified examples (false positives). Recall is a function of true positives and false negatives.

E. Experimental results

The experiments set up following the experiment procedure in Figure 1 and use the clinical dataset (called LIFLAB) which large, complex and high dimensions that showed in Table I. The results will present of the processes based on experiment procedure. First of all we will go to the pre-processing data for preparing the dataset, exploring the dataset to find out the characteristics of dataset. For the sake of research, variables with the meaningful were used for implementation so the cleaning process needs to be the first thing to do.

Table IV and V show the precision and recall values from different methods of missing value replacement and feature selection. There have shown the percentages of classification by classes of outcomes, Table IV presented the outcome of mortality of patient; dead/alive classes, the precision and recall values that had the best is the missing value imputation by using neural network because it used the observed data to be input and used target data so it has been given more accuracy than another techniques. On the contrary, when applied feature selection using entropy method the percentages of measurement from EM algorithm to fill missing values is highest. Because the EM algorithm uses the Kullback-Leibler (KL) [19], also known as relative entropy, divergence defines a distance measure between probability distribution same as entropy ranking for feature selection. Hsu*et al*. [5] claims that the NLGA (the wrapper model of feature selection) shows the effective for decision tree, and while considering the results from Table IV and Table V are shown consistent. The results of classification that presented in Table V are similar to the results that appeared in Table IV.

TABLE IV
THE CLASSIFICATION RESULTS FROM DIFFERENT TYPE OF MISSING VALUE REPLACEMENT METHODS AND FEATURE SELECTION TECHNIQUES BY DEAD AND ALIVE CLASSES. BOLD ENTRIES IN EVERY METHOD REPRESENT THE BEST PRECISION AND RECALL

Missing value replacement	Class	t-Test				Entropy				NLGA			
		MLP		Decision Tree		MLP		Decision Tree		MLP		Decision Tree	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
EM Algorithm	Dead	81.9	58.9	87.7	78.8	**72.5**	**51.6**	**93.2**	**77.3**	77.5	55.5	93.1	84.7
	Alive	81.8	93.4	89.8	94.4	78.6	90.1	89.4	97.1	80.3	91.8	92.6	96.8
k-NN imputation	Dead	76.1	61.2	95.9	79	70.5	52.7	86.5	75.9	77.2	56.7	79.9	76.8
	Alive	82.1	90.3	90.3	98.3	78.8	88.8	88.5	94	80.7	91.5	88.5	90.3
Mean imputation	Dead	81.6	57.8	93.1	84.1	71.1	49.6	87.3	81.6	74.6	55	79.2	68
	Alive	81.4	93.4	92.3	96.8	77.9	89.8	91	94	79.9	90.5	84.9	91
ANN imputation	Dead	**77.8**	**62.6**	**96.2**	**85.6**	71.3	54.1	91.6	83	**76.5**	**46.2**	**98**	**71.1**
	Alive	82.8	91	93.1	98.3	79.3	89	91.8	96.1	77.3	92.8	87.2	99.3

TABLE V
THE CLASSIFICATION RESULTS FROM DIFFERENT TYPE OF MISSING VALUE REPLACEMENT METHODS AND FEATURE SELECTION TECHNIQUES BY MORTALITY CLASSES OF MONTHS. BOLD ENTRIES IN EVERY METHOD REPRESENT THE BEST PRECISION AND RECALL

Missing value replacement	Class	t-Test				Entropy				NLGA			
		MLP		Decision Tree		MLP		Decision Tree		MLP		Decision Tree	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
EM Algorithm	**6M**	76.5	43.8	87.2	84.3	**53.9**	**46.1**	**88.6**	**87.6**	71	49.4	92.8	86.5
	12M	61.9	34.7	84	90.7	29.8	37.3	85.2	92	42.2	61.3	88	88
	18M	83.3	1.85	85.1	74.1	40.8	37	86.4	70.4	51.7	27.8	87.3	88.9
	24M	42.6	32.8	90.6	78.7	75	9.8	82.5	77	50	16.4	89.1	80.3
	36M	34.6	42.4	77.6	89.4	36	13.6	86.2	84.8	30.9	31.8	88.9	84.8
	>36M	49.6	86.2	91.6	92.8	48.6	78.3	87.7	93.4	57.4	78.9	88	96.1
k-NN imputation	**6M**	73.6	59.6	88.4	85.4	59.3	53.9	87.9	89.9	55.3	47.2	86.9	82
	12M	59.7	53.3	86.3	92	39.8	44	87.2	90.7	49	32	88.4	81.3
	18M	55.6	18.5	86.7	72.2	48.3	25.9	84.9	83.3	52.6	18.5	89.6	79.6
	24M	44.2	31.1	79.7	83.6	90	14.8	82	82	100	16.4	82.5	77
	36M	70	21.2	79.7	83.3	39	34.8	79.7	83.3	33.9	31.8	74	86.4
	>36M	49.1	89.5	92.2	92.8	50.2	77.6	93.1	88.8	46.3	85.5	86.4	92.1
Mean imputation	**6M**	57.3	57.3	86.2	91	63.5	52.8	87.6	87.6	**85.7**	**47.2**	86.7	87.6
	12M	41.9	41.3	86.3	84	43.6	22.7	76.5	86.7	52.9	36	84	90.7
	18M	55.6	27.8	89.8	81.5	37.5	27.8	87.5	64.8	53.8	25.9	86.3	81.5
	24M	55.6	24.6	85.7	78.7	100	8.2	77.8	80.3	45	29.5	87	77
	36M	29.1	37.9	87.1	81.8	34.5	28.8	84.6	83.3	47.2	37.9	87	90.9
	>36M	59.3	75.7	88.3	94.7	46.5	86.8	89.7	91.4	47.5	86.8	92.8	92.8
ANN imputation	**6M**	**82.6**	**64**	**91.9**	**88.8**	82	56.2	86.9	82	52.7	66.3	**96**	**80.9**
	12M	60	48	84.8	89.3	58.2	42.7	87.5	84	83.8	41.3	87.3	82.7
	18M	62.5	27.8	88.1	68.5	77.8	25.9	91.1	75.9	42.9	22.2	90.9	74.1
	24M	54.2	21.3	87.7	82	82.4	23	91.1	83.6	67.9	31.1	79.7	83.6
	36M	40.7	50	84.5	90.9	37.9	33.3	80.6	81.8	37.8	47	85.3	87.9
	>36M	54	84.9	89.5	95.4	46.5	88.2	82.7	94.1	53.8	74.3	84	96.7

When considering the classification results from Table IV and V, the classification performance from Table IV which mortality (dead/alive) class gave the better precision and recall. Because of the number of classes, the dataset which the multiple output classes will lead to imbalanced datasets and the distribution will not even. It reveals to pose the significant challenge in term of classification accuracy. A comparison the number of classes between two groups, the first group is mortality has two classes and the second group has six classes of mortality months.TableVI shows the results of classification for the five techniques of dimensionality reduction and compares with the classification of the dataset

without dimensionality reduction. These experimental results present the performance of classification by using feature selection and feature extraction to reduce the dimension. The evidence shows that *t*-Test is the feature reduction that gave highest precision of classification and can improve the performance from the dataset without dimensionality reduction. Although the sensitivity is lower than the dataset without dimensionality reduction but it gave high sensitivity as well.The variables that selected from t-Test are significant useful for developing the model for predicting heart failure because it selected Triglycerides, Potassium,Urea/ Uric acid, Creatinine, Nt-proBNP, and sodium are strong associations with mortality of heart failure [20, 21].

TABLE VI
THE SELECTED FEATURES BY USING ANN IMPUTATION AND NLGA S

No.	Outcome	
	Mortality (Dead/Alive)	**Mortality time frame (Dead Month)**
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 Phophatase	White Cell Count
9	ALT	Platelets
10	Total Protein	Total Protein
11	Albumin	Bilirubin
12	**Triglycerides**	Alkaline Phophatase
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**

From the selected features that have shown in Table VI, we can be displayed the sample of decision tree along outcome periods of mortality; 6, 12, 18, 24, 36 and more than 36 months. The decision tree Fig.5 is used the method of neural network for filling missing data and NLGA for selecting feature because it has shown highest percentages of classification measurement.The most important is which technique can use to design and create the appropriate predictive model, need to think about selected features, precision, and sensitivity that will get. However, the results are presented that feature selection is more appropriate to be a tool for developing the model for predicting. Because of the decision support system needs the meaningful and significant feature to make a decision to create effective model, the extracted feature could not be useful for our proposed. Obviously, from the experimental results, the feature selection and missing value handling gain the potential performance of classification.

VI. Conclusion

The aim of this paper is to investigate the dimensionality reduction based on feature selection. We attempt to understand and find the suitable techniques for developing the model for predicting heart failure. The features that got from feature selecting process are selected by picking them up from optimal criteria. However, both techniques have no question with dimensionality reduction, there are efficient techniques. The feature selection technique using three techniques: *t*-Test, entropy, and nonlinear gain analysis. The effect of these complexity measures on classification accuracy is evaluated using two diverse machine learning algorithms: multilayer perceptron (back-propagation) and J48 (decision tree). Using this methodology, we have performed experiments on two groups of dataset by their outcomes are a class of mortality and mortality time frame.

Results present the metrics of accuracy, precision, and recall. Obviously, the results claim that feature selection is sufficient method for improving the classification accuracy. Yu [10] argues that in theory more features should provide more power, but in practice only significant features will be more efficient which corresponding in the experimental results. In theory, data would be distributed following the normal distribution but in the real world situation it would not be. Feature selection techniques will depend on the nature of data and type of distribution of data. In the pre-processing process can give the story behind the data and can give you to make a choice of feature selection techniques that appropriated with the dataset that is used. Pre-processing process tend to understand the nature of data for example the measurement to describe the group of data by using mean and standard deviation, missing values handling might be change the distribution of data and could tend to be normally distribution. From the experimental results and experiences, it suggests that do not mention the best, find the suitable technique for instance *t*-Test will select the features that be normally distribution because this technique is used mean and standard deviation to find the significant feature in contrast entropy is related with density of data, it will find the maximum distance between the target classes. However, if *t*-Test is applied the selected feature will be normally distribution on the other hand if entropy is used the selected feature will be high density. The key factor is an understanding the nature dataset to choose the suitable techniques. The important outcomes of extensive study will help in choosing the suitable missing value handling method, feature selection techniques, and classification scheme for a particular nature of clinical datasets. In additional, the selected features will use for a predictor in the prognosis model for decision support system for the next stage.

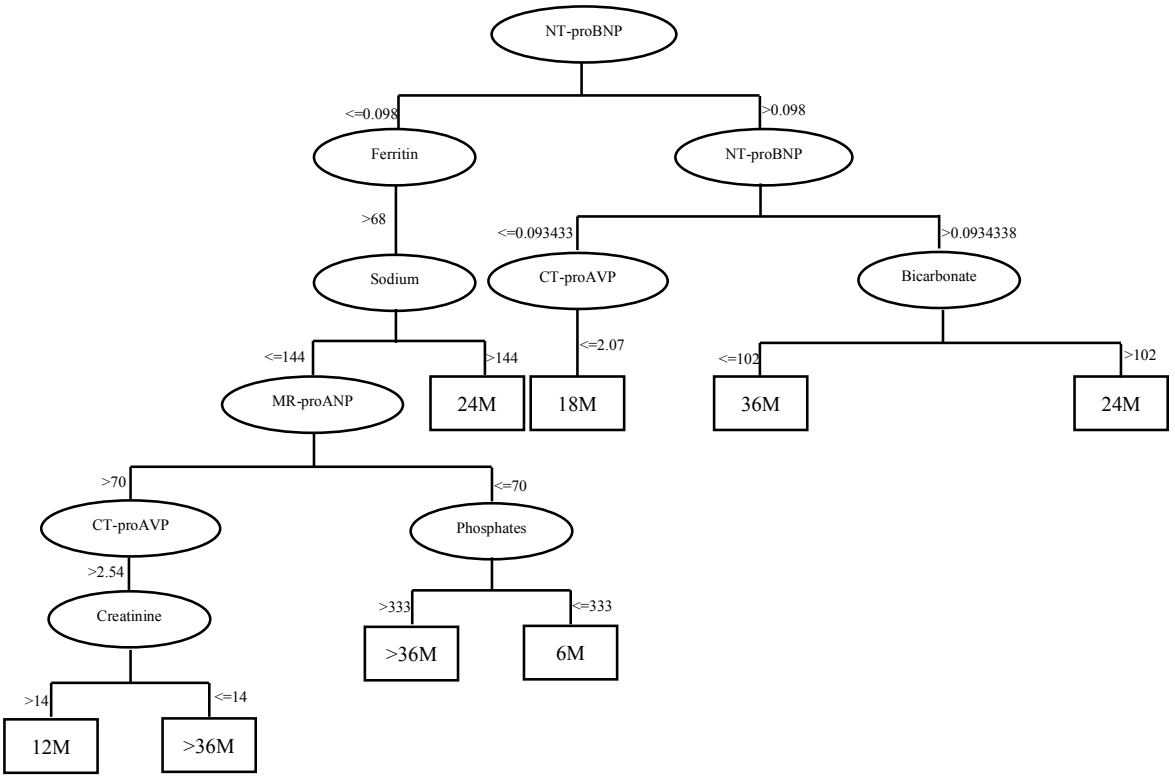

Fig. 5 Decision tree for predicting the mortality months

ACKNOWLEDGEMENT

The author of this paper is funded by the National Science and Technology Development Agency, Ministry of Science and Technology, Royal Thai Government. The heart failure dataset is supported by Cardiology Department, University of Hull, UK.

REFERENCES

[1] A. K. Tanwani, M. J. Afridi, M. Z. Shafiq, M. Farooq: Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. EvoBIO 2009: 128-139

[2] N. Zhou, L. Wang, "A Modified T-test Feature Selection Method and Its Application on the HapMap Genotype Data," Genomics, Proteomics & Bioinformatics, 5(3–4), pp. 242–249, 2007.

[3] U. Fayyad, K. Irani, "Multi-interval discretization of continuous-valued attributes for classication learning,"In: 13th International Joint Conference on Artificial Intelligencepp. 1022–1029, 1993.

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

[5] C.-N.Hsu, H.-J.Huang, D. Schuschel, "The ANNIGMA-wrapper approach to fast feature selection for Neural Nets," IEEE Transactions Systems, Man and Cybernetics, Part B, 2002, pp. 1–6.

[6] Heart Failure Society of, A. (2010). "Section 2: Conceptualization and Working Definition of Heart Failure." Journal of cardiac failure 16(6): e34-e37.

[7] W. B. Kannel, R. B. D'Agostino, H. Silbershatz, *et al*. "Profile for estimating risk of heart failure," Arch Intern Med 1999;159:1197–204.

[8] E. Acuna, C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," In: Banks, D., House, L., McMorris, F.R., Arabie, P., Gaul, W. (Eds.), Classification, Clustering and Data Mining Applications, Springer, Berlin, Heidelberg. pp. 639-648.

[9] J.-H. Lin, P. J. Haug, "Data Preparation Framework for Preprocessing Clinical Data in Data Mining," AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium, 2006, 489-493.

[10] L.Yu, H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," Machine Learning Research, 5, pp. 1205–1224, 2004.

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

[12] R. J.Harris, "A Primer of Multivariate Statistics, 3rd ed., New Jersey : Lawrence Erlbaum Associates, 2001.

[13] S.Li,C.Liao,J. T.Kwok, "Gene Feature Extraction Using *t*-Test Statistics and Kernel Partial Least Squares," ICONIP, 3, pp. 11–20, 2006.

[14] L.Wang, F.Chu, W.Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 40–53, 2007.

[15] D. W. Aha, R. L.Bankert, "A Comparative Evaluation of Sequential Feature Selection Algorithms," In: Fifth International Workshop onArtificial Intelligence and Statistics, pp. 1–7, 1995.

[16] Analysis Factor, "EM Imputation and Missing Data: Is Mean Imputation Really so Terrible?," [Online], 15 April 2009, (URL http://www.analysisfactor.com/statchat/tag/spss-missing-values-analysis/)(Accessed 30August 2011).

[17] 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, 24,1, 121-129, 2011.

[18] The University of Waikato, "WEKA: The Waikato Environment for Knowledge Acquisition," [Online],(URL http://www.cs.waikato.ac.nz/ml/weka/)(Accessed 30August 2011).

[19] F. Coetzee, "Correcting the Kullback-Leibler distance for feature selection", presented at Pattern Recognition Letters, 2005, pp.1675-1683.

[20] A.-N. Yahya, M. G. Kevin, Z. Jufen, G.F. C. John, L. C. Andrew, "Red cell distribution width: an inexpensive and powerful prognostic marker in heart failure,"European Journal Heart Failure,vol. 11,pp. 1155–1162, 2009.

[21] Atherotech Diagnotics Lab, "Atherotech Panels," [Online], (URL http://www.atherotech.com/athdiagtests/atherotechpanels.asp), (Accessed 13 June 2011).