VOL. 10, NO 20, NOVEMBER, 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A BENCHMARK OF CLASSIFICATION FRAMEWORK FOR NON-COMMUNICABLE DISEASE PREDICTION: A REVIEW

Daniel Hartono Sutanto and Mohd. Khanapi Abd. Ghani
Biomedical Computing and Engineering Technologies Applied Research Group, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
p031220009@student.utem.edu.my

ABSTRACT

Non-Communicable Disease (NCDs) or chronic disease is the high mortality rate in worldwide, such as diabetes mellitus, cardiovascular diseases and cancers. The accuracy of prediction model is required to enhance the quality of health care. In data mining, the classification algorithms have been applied to predict NCDs. Meanwhile, the benchmark of the classification algorithm for NCDs prediction is needed to analyze the optimal algorithm. The classification algorithms were used likely Decision Tree (DT), k-Nearest Neighbor (k-nn), Linear Discriminant Analysis (LDA), Linear Regression (LR), Naïve Bayes (NB), Neural Network (NN), Rule Induction (RI), and Support Vector Machine (SVM). In order to test the algorithms, this research used secondary data such as breast cancer, lung cancer, colon cancer, heart disease, and diabetes dataset. The research objective is benchmarking the optimal performance of classification algorithms using AUC. The optimal classifier for NCDs prediction showed by AUC Mean, such as NB (0.7938); LR (0.7569); NN (0.7436); k-nn (0.7386); SVM (0.6783), and there is no significant different both of them. DT and LDA has poor result of AUC Mean. The NCDs datasets have noisy data and irrelevant attribute. The outcome proved that NB, SVM and NN robust with noisy dataset, meanwhile irrelevant attribute problem can be handled with pre-processing technique for improving accuracy rate.

Keywords: classification, benchmark, non-communicable disease, chronic disease, Friedman test, Nemenyi test, AUC.

INTRODUCTION

Non-communicable Diseases (NCDs) are leading mortality rate and cause of death in worldwide. NCDs also known as chronic diseases are a long-lasting condition that can be controlled, but not be instantly cured. Top three main types of NCDs are diabetes mellitus, cardiovascular diseases and cancers [1]. On that point are some aspect affects the quality of health care. Firstly, inequity of diagnosis of NCDs due to discrepancy numbers between patients and doctors [2], [3], [4]. Secondly, most ASEAN countries had less than 1 physician per 1000 population [5]. Thirdly, less number of expert system development in most of ASEAN country [6]. Fourthly, too few people receiving proper diagnosis and treatment [7]. Founded on that fact, the lack of infrastructure and labor are affecting the poor quality of health care. Thus, expert systems are widely used in healthcare either predicting or diagnosing diseases and when the medical experts are unavailable [8].

In data mining, a method that is used to extract the hidden knowledge from large amounts of data, is commonly used [9]. To enhance non-communicable disease prediction model, data mining is the prediction technique to diagnose disease [10]. For data mining task, classification is the most widely used methods such as image and pattern recognition, medical diagnosis. Nevertheless, the prediction model using classification algorithm for non-communicable disease is needed to improve the quality of health care [11].

RELATED WORK

The classification algorithms were applied to predict the NCDs prediction model. The existing researches were developed using classification algorithm likely Decision Tree (DT), k-Nearest Neighbor (k-nn), Linier Discriminate Analysis (LDA), Linier Regression (LR), Naïve Bayes (NB), Neural Network (NN), Rule Induction (RI) and Support Vector Machine (SVM). The state-of-the-art researches of classification algorithm for non-communicable disease prediction during 2010-2015 (Table-1).

Table-1. A state-of-the-art of classification algorithm.

Contributor	DT	k-nn	LDA	LR	NB	NN	RI	SVM
[12]								
[13]								
[14]								
[15]								
[16]								
[17]								
[18]								
[19]								
[20]								
[21]								
[22]								
[23]								
[24]								
[25]								
[26]								
[27]								
[28]								
Total	1	4	1	2	1	6	1	6

The classification algorithms have been applied in Table-1, the researchers didn't used it for comparison task. In other hand, some of the researchers have been compared classification algorithms. Sarwar has developed comparative analysis of machine learning techniques in prognosis of type II diabetes, the classification algorithm used k-nn, NB, NN [29]. The result is ANN performed the best prediction with an accuracy of 96%. Ahmad compared NN and DT against decision diabetes mellitus, and the experiment showed that J48 algorithm in DT, resulted highest accuracy of 89.3% [30]. Upadhyaya compared of NN and LR classifiers for screening native American elders with diabetes, and the NN has better prediction rate compare than LR [23]. Kurt compared 3 classifiers for predicting coronary artery disease, the performed best resulted from NN with AUC = 0.783 [31].

In particular, we confirmed the reliability of scorecard benchmarks in the light of new findings related with the conceptual shortcomings of the AUC [27]. They have been compared 3 classification algorithms yet, however the literature review showed that at least 8 algorithms used in NCDs prediction model. Meanwhile, there is no benchmark performance to compare 8 classification algorithms for Non-Communicable Disease prediction. This research helps practitioners to stay abreast of technical advancements in Non-Communicable Prediction. Hence, the research questions of this study are as follows: which the classification algorithm has the optimal performance for Non-Communicable Disease prediction?

METHODOLOGY

A. Proposed benchmark of classification framework

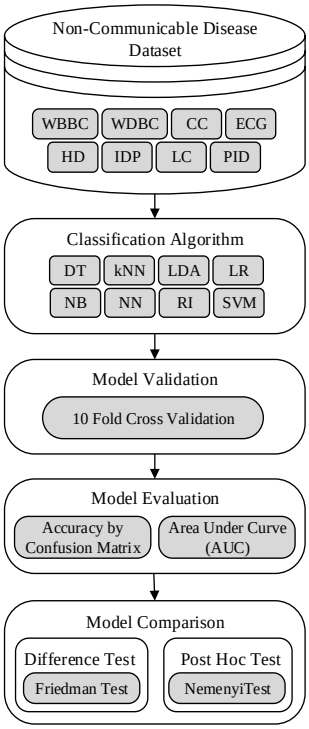

Figure-1. A benchmark of classification framework for NCDs prediction adopted from [32].

This research proposes benchmark data mining framework for NCDs dataset. The benchmark framework consists of NCDs dataset, classification algorithm, model validation, model evaluation and model benchmark. The benchmark framework maintained during the research study is shown in Figure-1.

B. Classification algorithms

The proposed classification algorithm framework aims to compare the performance of a wide range of classification models within the field of non-communicable disease prediction. For the purpose of this research, 8 classifiers have been selected likely DT, k-NN, LDA, LR, NB, NN, RI, and SVM. The selection aims for obtaining a balance between established classification algorithms used in NCDs prediction.

1. Decision tree (DT)

Decision tree is the most popular classification methods. Rules produced by decision tree are easy to interpret and understand, and hence, can help greatly in appreciating the underlying mechanisms that separate samples in different classes. Among many decision trees based classifiers, C4.5 is a well-proven and widely used algorithm. C4.5 uses the information gain ratio criterion to determine the most discriminatory feature at each step of its decision tree induction process. In each round of selection, the information gain ratio criterion chooses, from those features with an average-or-better information gain, the feature that maximizes the ratio of its gain divided by its entropy. C4.5 stops recursively building sub-trees when (1) an obtained data subset contains samples of only one class (then the leaf node is labelled by this class), or (2) there is no available feature (then the leaf node is labelled by the majority class), or (3) when the number of samples in the obtained subset is less than a specified threshold (then leaf node is labelled by the majority class) [33].

2. k-nearest neighbor (k-nn)

The k-nearest neighbor (k-nn) model is a well-known supervised learning algorithm for pattern recognition that first introduced by Fix and Hodges (1951), and is still one of the most popular nonparametric models for classification problems [34]. K-nn assumes that observations which are close together are likely to have the same classification. The probability that a point x belongs to a class can be estimated by the proportion of training points in a specified neighborhood of x that belong to that class [34]. The point may either be classified by majority vote or by a similar degree sum of the specified number (k) of nearest points. In majority voting, the number of points in the neighborhood belonging to each class is counted, and the class to which the highest proportion of points belongs is the most likely classification of x. The similarity degree sum calculates a similarity score for each class based on the K-nearest points and classifies x into the class with the highest similarity score. Due to its lower sensitivity to outliers, majority voting is more commonly used than the similarity degree sum [35]. In this method, majority voting is used for the data sets. In order to determine which points belong in the neighborhood, the distances from x to all points in the training set must be calculated. Any distance function that specifies which of two points is closer to the sample point could be employed [34]. The most common distance metric used in K-nearest neighbor is the Euclidean distance [36]. The Euclidean distance between each test point ft and training set point fs, each with n attributes, is calculated using the equation:

$$d = [(f_{t1} - f_{s1})^2 + (f_{t2} - f_{s2})^2 + \cdots + (f_{tn} - f_{sn})^2]^{1/2} \quad (1)$$

In general the following steps are performed for the K-nearest neighbor model [37]:

1) Chosen of k value.
2) Distance calculation.
3) Distance sort in ascending order.
4) Finding k class values.

5) Finding dominant class

One challenge to use the k-nn is to determine the optimal size of k, which behaves as a smoothing parameter. A small k will not be sufficient to accurately estimate the population proportions around the test point [38]. A larger k will result in less variance in probability estimates but the risk of introducing more bias [36]. K should be large enough to minimize the probability of a non-Bayes decision, but small enough that the points included give an accurate estimate of the true class. Enas and Choi (1986) found that the optimal value of k depends upon the sample size and covariance structures in each population, as well as the proportions for each population in the total sample [38]. For cases in which the differences in the covariance matrices and the difference between sample proportions were either small or both large, Enas and Choi (1986) found that the optimal k to be N3/8, where N is the number of samples in the training set. When there was a large difference between covariance matrices and a small difference between sample proportions, or vice versa, they determined N2/8 to be the optimal value of k. In additional, when the boundaries between classes cannot be described as hyper-linear or hyper-conic, K-nearest neighbor performs better than the linear and quadratic discriminant functions. They found that the linear discriminant performs slightly better than k-nn when population covariance matrices are equal, a condition that suggests a linear boundary. As the differences in the covariance matrices increases, k-nn performs increasingly better than the linear discriminant function [38].

However, despite of all the advantages cited for the k-nn models, they also have some disadvantages. k-nn model cannot work well if large differences are present in the number of samples in each class. k-nn provides poor information about the structure of the classes and of the relative importance of each variable in the classification. Furthermore, it does not allow a graphical representation of the results, and in the case of large number of samples, the computation can become excessively slow. In addition, k-nn model much higher memory and processing requirements than other methods. All prototypes in the training set must be stored in memory and used to calculate the Euclidean distance from every test sample.

3. Linear discriminant analysis (LDA)

The Linear Discriminant Analysis (LDA) [39], [40] is used as a class specific discriminative. This LDA method benefits supervised learning to find a set of base vectors. They are represented by w_k. These w_k vectors are a ratio of the between and within class scatters of the training sample set. They are maximized. The following generalized eigen value problem should be solved to find w_k base vectors,

$$W_{opt} = \arg max \frac{|W^T S_C W|}{|W^T S_V W|} = [W_1, W_2, \dots, W_L] \quad (2)$$

Herein, $\{w_k|1\ 6\ k\ 6\ L\}$ are the Linear Discriminant Analysis (LDA) subspace base vectors. L is the dimension of the subspace. Sc and Sv are the between and within class scatter matrices. These matrices can be given as follows:

$$S_c = \sum_{k=1}^{a} M_k\ (\mu_k - \mu)(\mu_k - \mu)^T \quad (3)$$

$$S_V = \sum_{k=1}^{a} \sum_{X_u \in X_k} (X_u - \mu_k)(X_u - \mu_k)^T \quad (4)$$

where, a is the number of classes and $X \in R^N$ is the data sample. X^k is the set of samples with class label k. μ_k is the mean for all the samples with the class label k. M_k is the number of samples in class k. The base vectors w_k sought in the Eq. (1) are the first L largest eigen values $\{w_k|1\ 6\ k\ 6\ L\}$, if Sv is non-singular. The base vectors w_k can be obtained by representation in LDA subspace by a simple linear projection $w^T x$ for a given test sample x as the LDA base vectors are orthogonal to each other.

4. Linear regression (LR)

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications [41]. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$of n statistical units, a linear regression model assumes that the relationship between the dependent variable ε_i and the p-vector of regressor xi is linear. This relationship is modeled through a disturbance term or error variable εi — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regression. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \dots, n, \quad (5)$$

where ^{T} denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β. Then these n equations are stacked together and written in vector form as $y = x\beta + \varepsilon$

5. Naïve bayes (NB)

Naïve Bayes (NB) is a simple learning algorithm that utilizes the Bayes rule together with a strong

assumption that the attributes are conditionally independent, given the class. While this independence assumption is often violated in practice, NB nonetheless often delivers competitive classification accuracy. Paired with its computational efficiency and many other desirable features, this leads to NB being widely used in exercise. NB is based on the Bayes rule

$$(y|x) = (y)P(x|y)P(x) \quad (6)$$

together with an assumption that the attributes are conditionally independent given the class. For attribute value data, this assumption entitles

$$(x|y) = \Pi(x_i|y)n_i = 1 \quad (7)$$

where x_i is the value of the i_{th} attribute in x, and n is the number of attributes.

$$(x) = \Pi(c_i)P(x|c_i)k_i = 1 \quad (8)$$

where k is the number of classes and c_i is the x_{th} class. Thus, Equation (2.3) can be calculated by normalizing the numerators of the right-hand-side of the equation.

6. Neural network (NN)

Neural Network (NN) is computer systems developed to mimic the operations of the human brain by mathematically modeling its neuro-physiological structure. NN have been shown to be effective at approximating complex nonlinear functions [42]. For classification tasks, these functions represent the shape of the partition between classes. In NN, computational units called neurons replace the nerve cells and the strengths of the interconnections are represented by weights, in which the learned information is stored. This unique arrangement can acquire some of the neurological processing ability of the biological brain such as learning and drawing conclusions from experience. NN combine the flexibility of the boundary shape found in K-nearest neighbor with the efficiency and low storage requirements of discriminant functions. Like the k-nn, NN are data driven; there are no assumed model characteristics or distributions, as is the case with discriminant analysis [43]. Single hidden layer feed forward network is the most widely used model form for modeling, forecasting, and classification [44]. The model is characterized by a network of three layers of simple processing units connected by acyclic link. The relationship between the output (y) and the inputs $(x_1, x_2, \ldots, x_p)$ has the following mathematical representation:

$$y_t = w_0 + \sum_{j=1}^{q} w_j \cdot g\left(w_{0.j} + \sum_{i=1}^{p} w_{i.j} \cdot x_{t.i}\right) + \varepsilon_t, \quad (9)$$

where $w_{i.j}(i = 0,1,2,\ldots,p, j = 1,2,\ldots,q)$ and $w_j(j = 0,1,2,\ldots,q)$ are model parameters often called connection weights; p is the number of input nodes; and q is the number of hidden nodes. Data enters the network through the input layer, moves through hidden layer, and exits through the output layer. Each hidden layer and output layer node collects data from the nodes above it (either the input layer or hidden layer) and applies an activation function. Activation functions can take several forms. The type of activation function is indicated by the situation of the neuron within the network. In the majority of cases input layer neurons do not have an activation function, as their role is to transfer the inputs to the hidden layer. The logistic and hyperbolic functions are often used as hidden layer and output transfer function for classification problems that are shown in Eqs. (10) and (11), respectively. Other transfer functions can also be used such as linear and quadratic, each with a variety of modeling applications.

$$Sig(x) = \frac{1}{1 + \exp(-x)} \quad (10)$$

$$Tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)} \quad (11)$$

The simple network given is surprisingly powerful in that it is able to approximate the arbitrary function as the number of hidden nodes when q is sufficiently large.

7. Rule induction (RI)

Rule induction is one of the most important techniques of machine learning. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools of data mining at the same time. Usually rules are expressions of the form

if (attribute - 1, value - 1) and (attribute - 2, value - 2) and . and (attribute - n, value - n) then (decision, value). (12)

Some rule induction systems induce more complex rules, in which values of attributes may be expressed by negation of some values or by a value subset of the attribute domain. Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labeled as attributes and a decision. We will restrict our attention to rule induction which belongs to supervised learning: all cases are pre-classified by an expert.

8. Support vector machine (SVM)

Support vector machines (SVM) is a new pattern recognition tool theoretically founded on Vapnik's statistical learning theory [45]. Support vector machines, originally designed for binary classification, employs supervised learning to find the optimal separating hyper plane between the two groups of data. Having found such a plane, SVM can then predict the classification of an unlabelled example by asking on which side of the separating plane the example lies. SVM acts as a linear classifier in a high dimensional feature space originated by a projection of the original input space, the resulting classifier is in general non-linear in the input space and it achieves good generalization performances by maximizing the margin between the two classes. In the following this research give a short outline of construction of support vector machine. Consider a set of training examples as follows:

$$\{(x_i y_i)\ x_i\ \in\ R^n, y_i \in \{+1, -1\};\ \ i = 1,2, \dots, m, \quad (13)$$

where the x_i are real n-dimensional pattern vectors and the y_i are dichotomous labels. SVM maps the pattern vectors $x \in R^n$ into a possibly higher dimensional feature space $(z = \emptyset(x))$ and construct an optimal hyperplane $w.z + b = 0$ in feature space to separate examples from the two classes. For SVM with L1 soft-margin formulation, this is done by solving the primal optimization problem as follows:

$$Min\frac{1}{2}||w|| + C\sum_{i=1}^{m}\epsilon_i \quad s.t.\, y_i(w.z_i + b) \geq 1 - \varepsilon_i x\ \varepsilon_i \geq 0, i = 1,2, \dots, m, \quad (14)$$

where C is a regularization parameter used to decide a tradeoff be- tween the training error and the margin, and ε_i $(i = 1,2, \dots, m)$ are slack variables. The above problem is computationally solved using the solution of its dual form:

$$Max_{\propto}\sum_{i=1}^{m}\propto_i - \frac{1}{2}\sum_{j=1}^{m}\propto_i\propto_j y_i y_j k(x_i, y_i) \quad s.t\sum_{i=1}^{m}\propto_i y_i = 0;\ 0 \leq \propto_i \leq C, i = 1,2, \dots, m, \quad (15)$$

where $k(x_i, y_i) = \emptyset(x_i).\emptyset(x_j)$ is the kernel function that implicitly define a mapping $\emptyset$. The resulting decision function is:

$$f(x) = sgn\left\{\sum_{i=1}^{m}\propto_i y_i k(x_i, x) + b\right\}. \quad (16)$$

All kernel functions have to fulfil Mercer theorem, however, the most commonly used kernel functions are polynomial kernel and radial basis function kernel, respectively.

$$k(x_i, x_j) = (a(x_i, x_j) + b)^d, \quad k(x_i, x_j) = exp(-g||x_i, x_j||^2), \quad (17)$$

SVM differ from discriminant analysis in two significant ways. First, the feature space of a classification problem is not assumed to be linearly separable. Rather, a nonlinear mapping function (also called a kernel function) is used to represent the data in higher dimensions where the boundary between classes is assumed to be linear [46]. Second, the boundary is represented by SVM instead of a single boundary. Support vectors run through the sample patterns which are the most difficult to classify, thus the sample patterns that are closest to the actual boundary [46]. Over fitting is prevented by specifying a maximum margin that separates the hyper plane from the classes. Samples which violate this margin are penalized which is a parameter often referred to as C [47], [48].

C. Non-communicable disease dataset

1. Wisconsin biopsy breast cancer (WBBC)

Biopsy Data on Breast Cancer Patients [56]. This breast cancer database was collected from Dr. William H. Wolberg, University of Wisconsin Hospitals, and Madison from. He observed biopsies of breast tumors for 699 patients up to 15 July 1992; nine attributes has been scored on a scale of 1 to 10, and the output is also known.

2. Wisconsin diagnostic breast cancer (WDBC)

Ten real-valued features are computed for each cell nucleus such as radius attribute, texture attribute, perimeter attribute, area attribute, compactness attribute, concavity attribute, concave points attribute, symmetry attribute, and fractal dimension attribute [57]. The mean, standard error, and "worst" or largest of these features were computed for each image, resulting in 31 features. Result is predicting diagnosis: B = benign or M = malignant and data sets are linearly separable using all 31 input features.

3. Colon cancer (CC)

The data collection from one of the first successful trials of adjuvant chemotherapy for colon cancer [58]. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-

FU is a moderately toxic (as these things go) chemotherapy agent. The output consists two records per person, one record for recurrence and one record for death.

4. Echocardiogram (ECG)

All the patients suffered heart attacks at some point in the past. The patients are still alive and some are not. The survival variables and still-alive variables, when taken together, show whether a patient survived following the heart attack [59]. The past researchers addressed problem to predict from the other variables whether or not the patient will survive at least one year. The difficult part is correctly predicting that the patient will not survive.

5. Heart disease (HD)

Instance-based prediction of heart-disease presence with the Cleveland database [60]. In particular way, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. The Cleveland database has concentrated on simply attempting to distinguish presence (values 1-4) from absence (value 0).

6. Indonesian diabetes patient (IDP)

The dataset collected from the government public hospital in Palembang, Indonesia [61]. The patients included only type-2 diabetes, whereas other types of diabetes type-2 were excluded. The final data obtained 435 cases, where 347 cases in class "TRUE" and 88 cases in class "FALSE". There are 11 clinical attributes: (1) Gender, (2) BMI, (3) Blood Pressure (BP), (4) Hyperlipidemia, (5) Fasting blood sugar (FBS), (6) Instant blood sugar, (7) Family history, (8) Diabetes Gest history, (9) Habitual Smoker, (10) Plasma insulin, (11) Age.

7. Lung cancer (LC)

The dataset contain of survival patients with advanced lung cancer from the North Central Cancer Treatment Group [62]. Performance scores rate how well the patient can perform usual daily activities. The attributes such as institution code, time, age, sex, ph.ecog, ph.karno, pat.karno, meal.cal, weight loss. The output is censoring status 1=censored and 2=dead.

8. Pima Indian dataset (PID)

In Pima datasets consist of female patients at least 21 years old of Pima Indian heritage [63]. The attributes consist of 1) Number of times pregnant; 2) Plasma glucose concentration 2 hours in an oral glucose tolerance test; 3) Diastolic blood pressure (mm H_g); 4) Triceps skin fold thickness (mm); 5) 2-Hour serum insulin (mu U/ml); 6) BMI(weight in kg/(height in m)^^{2}); 7) Diabetes pedigree function; 8) Age (years); 9) Class variable (0 or 1).

NCDs datasets have been picked up from internet repositories, primarily from the UCI Machine Learning Repository. This research used 8 secondary datasets, that consist of diabetes, heart, and cancer datasets (Table-2).

Table-2. NCDs dataset for classification task.

Researcher	Abbr.	Instance	Attr.	Class
[49]-[51]	WBBC	699	12	2
[51], [52]	WDBC	569	31	2
[53]	CC	1858	17	2
[49], [51], [52]	ECG	132	12	2
[22], [50]	HD	303	13	4
[54], [23]	IDP	435	11	2
[20], [51]	LC	228	12	2
[19], [52], [54], [55]	PID	768	8	2

D. Classification algorithm validation

This research uses a stratified 10-fold cross-validation method for learning data and testing data. It splits the training data into 10 equal parts and then performs the learning process 10 times. As shown in Table 3, this research selected another part of dataset for testing and used the remaining nine parts for learning. Then, it calculated the average values and the deviation values from the ten different testing results. Thus, this research employ the stratified 10-fold cross validation, because this method has known as a standard and state-of-the-art validation method in practical terms [64].

Table-3. Stratified 10 fold cross validation method.

n-validation	Dataset's Partition									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

E. Classification algorithm evaluation

This section explains the evaluation using accuracy and AUC. This research applies Area under Curve (AUC) as an accuracy indicator in our experiments

to evaluate the performance of the classification algorithm. AUC is area under ROC curve. The performance accuracy of the classification algorithm performed accuracy by confusion matrix (Table 4) and AUC (Eq. 24).

Table-4. Confusion matrix.

Parameter	Formula	
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	(18)
Sensitivity (TP Rate)	$\frac{TP}{TP+FN}$	(19)
Specificity (FP Rate)	$\frac{FP}{FP+TN}$	(20)
Positive Predictive Value (PPV)	$\frac{TP}{TP+FP}$	(21)
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$	(22)

Huang [65] recommended AUC and accuracy in evaluating learning algorithm, and AUC should be preferred over accuracy. AUC is equivalent to the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example. Hand and Till [66] present the following simple approach to calculating AUC of a classifier for binary classification

$$Å = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \quad (23)$$

where n_0 and n_1 are the numbers of positive and negative examples, respectively, and $S_0 = \sum r_i$, where r_i is the rank of the ith positive example in the ranked list.

In some research, Lessmann *et al.* [67] and Li et al. [68] stated the use of the AUC to improve cross study comparability. The AUC has advantage to improve convergence across empirical experiments significantly, because it separates predictive performance from operating conditions, and represents a general measure of predictive. A rough guide for classifying the accuracy of a diagnostic test using AUC is the traditional system, presented by Belle [27]. This research added the symbols for easier reading and understanding of AUC (Table-5).

Table-5. AUC evaluation.

AUC	Classification	Symbol
0.90 - 1.00	excellent	
0.80 - 0.90	good	
0.70 - 0.80	fair	
0.60 - 0.70	poor	
< 0.60	failure	

F. Classification algorithm benchmark

In comparison test, there are three families of statistical tests that can be used for benchmarking two or more classifiers over multiple datasets:

Parametric tests (the paired t-test and ANOVA), non-parametric tests (the Wilcoxon and the Friedman test) The non-parametric test that assumes no commensurability of the results (sign test).

Demsar suggests the Friedman test for multiple benchmark classifiers, which relies on less restrictive assumptions [69]. Based on this recommendation, the Friedman test is applied to compare the AUCs in different classifiers. The Friedman test is calculated on the average ranked (R) performances of the classification algorithms on each dataset.

Let r_j^i be the rank of the j-th of C algorithms on the i-th of D datasets. The Friedman test has aim to compare the average ranks of algorithm $R_j = \frac{1}{D}\sum_{i-1}^{D} r_j^i$. Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be fair. The statistic of Friedman is calculated as follows, and distributed according to x_F^2 with $C-1$ degrees of freedom, when variable D and C are big enough.

$$x_F^2 = \frac{12D}{C(C+1)}\left[\sum_j^D {R_j}^2 - \frac{C(C+1)^2}{4}\right] \quad (24)$$

If the null-hypothesis is rejected, it can be proceeded with a post-hoc test. When all classifiers are compared to each other, the Nemenyi test should be applied. Two classifiers have significantly different performance if the corresponding average ranks differ by at least the critical difference, shown by

$$CD = q_a \sqrt{\frac{C(C+1)}{D}} \quad (25)$$

where critical values q_a are based on the studentized range statistic.

G. Experimental infrastructure

The experiment equipped with infrastructure consists RapidMiner and SPSS. Rapidminer toolkit as an open-source system consisting of a number of data mining algorithms to automatically analyze a large data collection and extract useful knowledge [70]. SPSS Statistic also known as PASW (Predictive Analytics SoftWare) is specifically made for analyzing statistical data and thus it offers a great range of methods, graphs and charts. The hardware used CPU: HP Z420 Workstation, Processor: Intel® Xeon® CPU E5-1603 @ 2.80 GHz, RAM: 8, 00 GB, and OS: Windows 7 Professional 64-bit Service Pack 1.

H. Implementation

The data type has been set an integer, Boolean, polynomial for three or more inputs, binomial for two outputs. Then, data cleaning was applied on the datasets selected. Regarding the missing data analysis, missing data has been removed from the NCDs datasets. Other than missing data analysis, NCDs datasets were also cleaned to remove noisy data. Some of classification algorithms weren't unable to handle missing data such as LDA, LR and NN. Unnecessary space characters or other spelling mistakes were also cleaned in the datasets, the implementation procedure. Then, the data pre-processing steps have been completed using all 8 NCDs datasets (WBBC, WDBC, CC, ECG, HD, IDP, LC and PID) have been used to run the 5 classification algorithms (DT, k-NN, LDA, LR, NB, NN, RI, and SVM). For all algorithms, splitting the data into learning and testing splits have been selected as the validation method. 10-fold cross validation has been implemented on the same datasets for the selected algorithms. Parameters have been adjusted for optimal performance, shown in Table-6.

Table-6. Parameter setting.

Classifier	Item	Value
DT	**K**	**2 class**
	Max run	10
	Max optimization	100
	Measure type	
	Divergence	
k-nn	k	10
LDA		
LR	Feature selection	M5 prime
	Min tolerance	0.05
NB	Parameter	Laplace correction
NN	Training cycles	500
	Learning rate	0.3
	Momentum	0.2
RI	Criterion	Information gain
	Sample ratio	0.9
	Pureness	0.9
	Minimal prune benefit	0.25
SVM	Type	C-SVC
	Kernel	Linier
	C	0.0
	Cache	80
	Epsilon	0.5

EXPERIMENTAL RESULTS

In this section, the performance results of classification algorithms will be discussed accordingly. The accuracy of classification algorithm is actually evaluated, when pointing at the performance results of the classifier. The accuracy is calculated by determining the percentage of instances correctly classified [10]. The accuracy values of the multiple dataset implementations according to each classifier can be seen in Table-7.

Table-7. Classification accuracy using 8 datasets.

	WBBC	WDBC	CC	ECG	HD	IDP	LC	PID
DT	93.67	93.99	49.89	96.07	49.22	94.72	72.35	70.46
k-nn	88.58	95.85	53.99	94.46	52.45	94.49	73.24	74.89
LDA	91.92	94.42	57.11	92.86	58.48	90.43	79.82	77.21
LR	76.69	96.00	57.05	93.39	59.11	93.39	81.33	76.69
NB	93.49	95.99	53.39	93.39	52.86	97.25	78.12	75.79
NN	96.66	82.25	51.51	93.39	57.09	61.77	83.38	74.74
RI	91.57	94.28	39.03	96.25	51.82	93.33	67.98	74.10
SVM	97.54	95.34	54.46	95.89	54.8	79.31	72.32	76.71

Table-8 reports the AUCs of all classification algorithms. In last column of Figure-2, the mean rank R_j of each classifier over all datasets, which constitutes the basis of the Friedman test.

Table 8: AUC of classification algorithm.

	WBBC	WDBC	CC	ECG	HD	IDP	LC	PID
DT	0.919	0.938	0.500	0.500	0.001	0.924	0.500	0.595
k-nn	0.940	0.987	0.520	0.960	0.067	0.977	0.664	0.794
LDA	0.500	0.500	0.500	0.500	0.289	0.500	0.500	0.500
LR	0.824	0.988	0.625	1.000	0.294	1.000	0.500	0.824
NB	0.987	0.977	0.516	1.000	0.280	0.976	0.809	0.805
NN	0.985	0.934	0.563	0.993	0.281	0.500	0.896	0.797
RI	0.906	0.954	0.453	0.500	0.245	0.883	0.641	0.744
SVM	0.992	0.958	0.597	0.980	0.071	0.500	0.500	0.828

The best classification model on each dataset is highlighted with boldfaced print. The highest Friedman score (R) is LDA, followed by DT, RI, and SVM. In statistical significance testing, the P-value is the probability of achieving a test statistic at least as extreme as the one that was actually observed, hence assuming that the null hypothesis is true. Oftenly, the research is used "rejects the null hypothesis" when the P- value is less than the predetermined significance level (α), showing the observed result would be highly unlikely under the null hypothesis.

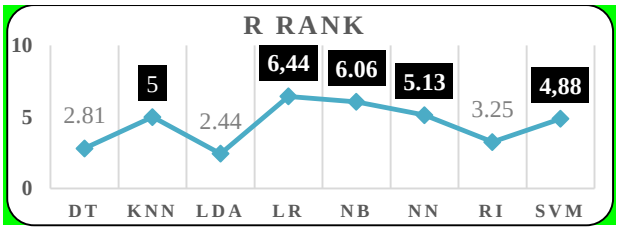

Figure-2. R Rank of classification algorithm.

In this research, it set the statistical significance level (α) to be 0.05. It means that there is a statistically significant difference, if P-value < 0.05. From the result of experiment, P-value is 0.0001, this is lower than the significance level α=0.05, hence one should reject the null hypothesis, and there is a significant difference,

statistically. For detecting particular classifiers differ significantly, it can be used a Nemenyi post hoc test. Nemenyi post hoc has ability to calculates all pairwise benchmarks between different classifiers and find which performance differences of models exceed the critical difference. The results of the pairwise benchmarks of classification algorithms are shown in Table-9.

Table-9. Pairwise of Nemenyi post hoc test.

	DT	k-nn	LDA	LR	NB	NN	RI	SVM
DT	0	-2.449	1.452	-2.214	-3.041	-1.336	-1.478	-0.768
k-nn	-2.449	0	2.889	-0.425	-1.921	-0.065	1.164	0.929
LDA	1.452	2.889	0	-3.736	-4.301	-3.504	-2.599	-2.14
LR	-2.214	-0.425	-3.736	0	-0.783	0.15	1.329	1.111
NB	-3.041	-1.921	-4.301	-0.783	0	0.803	2.311	1.673
NN	-1.336	-0.065	-3.504	0.15	0.803	0	0.89	1.188
RI	-1.478	1.164	-2.599	1.329	2.311	0.89	0	-0.138
SVM	-0.768	0.929	-2.14	1.111	1.673	1.188	-0.138	0

P-value results of Nemenyi post hoc test are shown in Table-10. P-value < 0.05 results are highlighted with boldfaced and grayscale print, furthermore there is a statistically significant difference between two classification algorithms, in a column and a row.

Table-10. P-value of Nemenyi post hoc test.

	DT	k-nn	LDA	LR	NB	NN	RI	SVM
DT	1	0.044	0.19	0.062	0.019	0.223	0.183	0.467
k-nn	0.044	1	0.023	0.684	0.096	0.95	0.282	0.384
LDA	0.19	0.023	1	0.007	0.004	0.01	0.035	0.007
LR	0.062	0.684	0.007	1	0.459	0.885	0.226	0.303
NB	0.019	0.096	0.004	0.459	1	0.448	0.054	0.138
NN	0.223	0.95	0.01	0.885	0.894	1	0.403	0.274
RI	0.183	0.282	0.035	0.226	0.054	0.403	1	0.894
SVM	0.467	0.384	0.007	0.303	0.138	0.274	0.894	1

As shown in Table-10, LDA outperforms other models in most NCDs datasets. In terms of R value (Figure-3) and AUC mean (M) (Figure-4), NB also has the highest value, followed by LR, NN and k-nn. Based on P-value results (Table-7), actually there is no significant difference between LR, NN, SVM, k-NN and RI. This result confirmed Upadhyaya [23] result that NN and LR seem to be the techniques used in models, they are performing well in NCDs prediction, relatively. SVM actually has optimal generalization for small sample data like ECG with AUC 0.98. Meanwhile, in this experiment SVM perform not too well, as it will require feature selection for improving the performance of classifier.

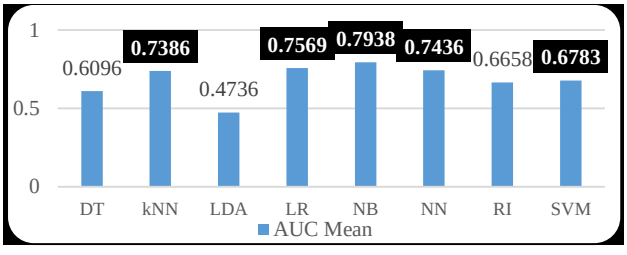

Figure-3. AUC mean (M) of 8 classification algorithms.

The top of AUC Mean held by NB, LR, NN, SVM, k-nn. From P-value analysis, there is a significant difference between k-nn, NB compare to DT algorithm. This is may be due to the irrelevant attribute and noisy class of NCDs datasets. LDA and k-nn models also poor

performing and to be failure in the most NCDs datasets. Significant difference table resulted by Nemenyi post hoc test shown in Table-11.

Table-11. Significant differences of Nemenyi post hoc test.

	DT	k-nn	LDA	LR	NB	NN	RI	SVM
DT	N	Y	N	N	Y	N	N	N
k-nn	Y	N	Y	N	N	N	N	N
LDA	N	Y	N	Y	Y	Y	Y	Y
LR	N	N	Y	N	N	N	N	N
NB	Y	N	Y	N	N	N	N	N
NN	N	N	Y	N	N	N	N	N
RI	N	N	Y	N	N	N	N	N
SVM	N	N	Y	N	N	N	N	N

CONCLUSIONS

Data mining has many prediction techniques, the most of used technique for classification. The benchmark of classification algorithm is required to show the optimal performance in every algorithm. This framework is proposed for comparing the performance of classification algorithms for NCDs prediction. It is comprised of 8 NCDs datasets, 8 classification algorithms, 10 fold cross validation method, and AUC as an indicator of accuracy.

The significance of AUC between models will use the Friedman test and Nemenyi test. The optimal classifier for NCDs prediction showed by AUC Mean, such as NB (0.7938); LR (0.7569); NN (0.7436); k-nn (0.7386); SVM (0.6783), and there is no significant different both of them. DT, RI and LDA has poor result of AUC Mean. The NCDs datasets have noisy data and irrelevant attribute. The outcome proved that k-nn, NB, SVM and NN robust with noisy dataset, meanwhile irrelevant attribute problem can be handled with pre-processing technique for improving accuracy rate.

ACKNOWLEDGEMENT

This research was supported by a grant from LPDP Minister of Finance of Indonesia No. Kep56/LPDP/2014.

REFERENCES

[1] WHO, "Global status report on noncommunicable diseases," 2010.

[2] M. K. A. Ghani, R. K. Bali, R. N. G. Naguib, I. M. Marshall, and N. S. Wickramasinghe, "Critical analysis of the usage of patient demographic and clinical records during doctor-patient consultations: a Malaysian perspective," Int. J. Healthc. Technol. Manag., vol. 11, no. 1/2, p. 113, 2010.

[3] D. H. Sutanto, N. S. Herman, and M. K. A. Ghani, "Trend of Case Based Reasoning in Diagnosing Chronic Disease: A Review," Adv. Sci. Lett., vol. 20, no. 10, pp. 1740-1744, 2014.

[4] M. K. A. Ghani, R. K. Bali, R. N. G. Naguib, I. M. Marshall, and N. S. Wickramasinghe, "Electronic health records approaches and challenges: a comparison between Malaysia and four East Asian countries," Int. J. Electron. Healthc. vol. 4, no. 1, p. 78, 2008.

[5] WHO, World Health statistics 2014. 2014.

[6] J. C. Montoya, C. L. Rebulanan, N. A. C. Parungao, and B. Ramirez, "A look at the ASEAN-NDI : building a regional health R & D innovation network," pp. 1–10, 2014.

[7] Novo Nordisk, "(2013). The Blueprint for Change Programme No 5. Where economics and health meet: changing diabetes in Indonesia," no. March, p. 28.

[8] F. S. Khan, F. Maqbool, S. Razzaq, K. Irfan, and T. Zia, "The Role of Medical Expert Systems in Pakistan," World Acad. Sci. Eng. Technol., vol. 2, no. 1, pp. 280–282, 2008.

[9] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, vol. 40, no. 6. 2001.

[10] J. T. L. Wang, M. J. Zaki, H. T. T. Toivonen, and D. Shasha, Data Mining in Bioinformatics. 2005.

[11] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Feature selection and

classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Syst. Appl., vol. 38, no. 5, pp. 5947–5957, 2011.

[12] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," Expert Syst. Appl., vol. 37, no. 12, pp. 8102–8108, Dec. 2010.

[13] D. C. Li, C. W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets.," Comput. Biol. Med., vol. 40, no. 5, pp. 509–18, May 2010.

[14] N. H. Barakat, A. P. Bradley, S. Member, and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," IEEE Trans. Inf. Technol. Biomed. vol. 14, no. 4, pp. 1114-1120, 2010.

[15] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," Expert Syst. Appl. vol. 38, no. 7, pp. 8311-8315, Jul. 2011.

[16] M. Fallahnezhad, M. H. Moradi, and S. Zaferanlouei, "A Hybrid Higher Order Neural Classifier for handling classification problems," Expert Syst. Appl., vol. 38, no. 1, pp. 386-393, Jan. 2011.

[17] F. Gagliardi, "Instance-based classifiers applied to medical databases: diagnosis and knowledge extraction," Artif. Intell. Med., vol. 52, no. 3, pp. 123-39, Jul. 2011.

[18] H. C. Lin, C. T. Su, and P. C. Wang, "An application of artificial immune recognition system for prediction of diabetes following gestational diabetes.," J. Med. Syst., vol. 35, no. 3, pp. 283-9, Jun. 2011.

[19] F. Beloufa and M. a Chikh, "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm," Comput. Methods Programs Biomed., vol. 112, no. 1, pp. 92-103, Oct. 2013.

[20] N. A. Mat Isa and W. M. F. W. Mamat, "Clustered-Hybrid Multilayer Perceptron network for pattern recognition application," Appl. Soft Comput., vol. 11, no. 1, pp. 1457-1466, Jan. 2011.

[21] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification," Expert Syst. Appl., vol. 40, no. 13, pp. 5402–5412, Oct. 2013.

[22] H. Yoon, C. S. Park, J. S. Kim, and J.-G. Baek, "Algorithm learning based neural network integrating feature selection and classification," Expert Syst. Appl., vol. 40, no. 1, pp. 231–241, Jan. 2013.

[23] S. Upadhyaya, K. Farahmand, and T. Baker-Demaray, "Comparison of NN and LR classifiers in the context of screening native American elders with diabetes," Expert Syst. Appl., vol. 40, no. 15, pp. 5830–5838, Nov. 2013.

[24] B. Krawczyk and G. Schaefer, "A hybrid classifier committee for analysing asymmetry features in breast thermograms," Appl. Soft Comput. J., vol. 20, pp. 112-118, 2014.

[25] B. H. Cho, H. Yu, J. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, "Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels.," IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 2, pp. 247-56, Mar. 2008.

[26] H. R. Marateb, M. Mansourian, E. Faghihimani, M. Amini, and D. Farina, "A hybrid intelligent system for diagnosing microalbuminuria in type 2 diabetes patients without having to measure urinary albumin.," Comput. Biol. Med., vol. 45, pp. 34-42, Feb. 2014.

[27] V. Van Belle and P. Lisboa, "White box radial basis function classifiers with component selection for clinical prediction models," Artif. Intell. Med., vol. 60, no. 1, pp. 53–64, Jan. 2014.

[28] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," Expert Syst. Appl., vol. 41, no. 5, pp. 2239-2249, Apr. 2014.

[29] A. Sarwar and V. Sharma, "Comparative analysis of machine learning techniques in prognosis of type II diabetes," Ai Soc., vol. 29, no. 1, pp. 123–129, Apr. 2013.

[30] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, "Comparison between Neural Networks against Decision Mellitus," pp. 537–545, 2011.

[31] I. Kurt, M. Ture, and a. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," Expert Syst. Appl., vol. 34, no. 1, pp. 366-374, Jan. 2008.

[32] R. S. Wahono, N. S. Herman, and S. Ahmad, "A Comparison Framework of Classification Models for Software Defect Prediction," Adv. Sci. Lett., vol. 20, pp. 1945–1950, 2014.

[33] J. R. Quinlan, C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.

[34] E. Fix and J. L. H. Jr, "Discriminatory analysis-nonparametric discrimination: consistency properties." California Univ Berkeley, 1951.

[35] W. A. Chaovalitwongse, Y. Fan, S. Member, and R. C. Sachdeo, "On the Time Series K -Nearest Neighbor Classification of Abnormal Brain Activity," Seizure, vol. 37, no. 6, pp. 1005-1016, 2007.

[36] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," J. Risk Insur., vol. 69, no. 3, pp. 373-421, 2002.

[37] T. Yildiz, S. Yildirim, Y. Doç, and D. T. Altilar, "Spam filtering with parallelized KNN algorithm." Akademik Bilisim, pp. 627-632, 2008.

[38] G. G. Enas and S. C. Choi, "Choice of the smoothing parameter and efficiency of k-nearest neighbor classification," Comput. Math. With Appl. vol. 12, no. 2, pp. 235–244, 1986.

[39] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. ~ {Fisherfaces}: Recognition using class specific linear projection," Pami, vol. 19, no. 7, pp. 711-720, 1997.

[40] F. Tang and H. Tao, "Fast linear discriminant analysis using binary bases," Pattern Recognit. Lett., vol. 28, pp. 2209-2218, 2007.

[41] X. Yan, Linear Regression Analysis. 2009.

[42] G. P. Zhang, B. E. Patuwo, and M. Y. Hu, "A simulation study of artificial neural networks for nonlinear time-series forecasting," Comput. Oper. Res., vol. 28, pp. 381-396, 2001.

[43] V. L. Berardi and G. Q. Zhang, "The effect of misclassificaton costs on neural network classifiers," Decis. Sci. Institute, 1997 Annu. Meet. Proceedings, Vols 1-3, vol. 30, no. 3, pp. 364-366, 1997.

[44] L. M. Silva, J. Marques de Sá, and L. a. Alexandre, "Data classification with multilayer perceptrons using a generalized error function," Neural Networks, vol. 21, no. 9, pp. 1302–1310, 2008.

[45] V. Vapnik, S. E. Golowich, and A. Smola, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing." pp. 281-287, 1998.

[46] R. Duda O., P. Hart E., and D. Stork G., Pattern Classification. 2000.

[47] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines.," Proc. Natl. Acad. Sci. U. S. A., vol. 97, no. 1, pp. 262-267, 2000.

[48] N. Cristiani and J. Shawe-Taylor, An Introduction to Support Vector Machines. 2000.

[49] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier," Expert Syst. Appl., vol. 38, no. 4, pp. 4600–4607, Apr. 2011.

[50] N. Yilmaz, O. Inan, and M. S. Uzer, "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases.," J. Med. Syst., vol. 38, no. 5, p. 48, May 2014.

[51] S. Belciug and F. Gorunescu, "Error-correction learning for artificial neural networks using the Bayesian paradigm. Application to automated medical diagnosis." J. Biomed. Inform. vol. 52, pp. 329–37, Dec. 2014.

[52] L. Y. Chuang, C. H. Yang, K. C. Wu and C. H. Yang, "A hybrid feature selection method for DNA microarray data," Comput. Biol. Med., vol. 41, no. 4, pp. 228-37, Apr. 2011.

[53] P. Ganesh Kumar, T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj, "Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm," Expert Syst. Appl., vol. 39, no. 2, pp. 1811-1821, Feb. 2012.

[54] J. Zhu, Q. Xie, and K. Zheng, "An improved early detefeaction method of type-2 diabetes mellitus using multiple classifier system," Inf. Sci. (Ny). vol. 292, pp. 1-14, Jan. 2015.

[55] M. A. Chikh, M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor.," J. Med. Syst., vol. 36, no. 5, pp. 2721-9, Oct. 2012.

[56] W. H. Wolberg, O. Mangasarian, and D. W. Aha, "Breast Cancer Wisconsin (Original) Data Set." UCI Machine Learning Repository, University of Wisconsin Hospitals Madison, Wisconsin, USA, 1992.

[57] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set." UCI Machine Learning Repository, University of Wisconsin Hospitals Madison, Wisconsin, USA, 1992.

[58] J. a. Laurie, C. G. Moertel, T. R. Fleming, H. S. Wieand, J. E. Leigh, J. Rubin, G. W. McCormack, J. B. Gerstner, J. E. Krook, J. Malliard, D. I. Twito, R. F. Morton, L. K. Tschetter, and J. F. Barlow, "Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and their combination of levamisole and fluorouracil," J. Clin. Oncol. vol. 7, no. 10, pp. 1447–1456, 1989.

[59] S. Salzberg and Evlin Kinney, "Echocardiogram Data Set." UCI Machine Learning Repository, the Reed Institute, Miami, 1988.

[60] D. W. Aha and D. Kibler, "Heart Disease Data Set." UCI Machine Learning Repository, Cleveland Clinic Foundation, 1988.

[61] B. A. Tama and F. S. Rodiyatul, "An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital," Telkomnika, vol. 9, no. 2, pp. 287-294, 2011.

[62] C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, and N. E. Klatt, "Prospective evaluation of prognostic variables from patient-completed questionnaires North Central Cancer Treatment Group," J. Clin. Oncol. vol. 12, no. 3, pp. 601-607, 1994.

[63] V. Sigillito, "Pima Indians Diabetes Database." UCI Machine Learning Repository, National Institute of Diabetes and Digestive and Kidney Diseases, 1990.

[64] ian H. Witten, E. Frank, and M. A. Hall, Data Mining : Practical Machine Learning Tools and Techniques, 3rd edition. 2006.

[65] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," IEEE Trans. Knowl. Data Eng., vol. 17, no. 3, pp. 299-310, 2005.

[66] D. J. Hand and R. J. Till, "A simple generalization of the area under the ROC curve to multiple class classification problems," Mach. Learn., vol. 45, pp. 171-186, 2001.

[67] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings," IEEE Trans. Softw. Eng., vol. 34, no. 4, pp. 485-496, 2008.

[68] D. C. Li, C. W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," Artif. Intell. Med., vol. 52, no. 1, pp. 45-52, May 2011.

[69] J. Demšar and J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," J. Mach. Learn. Res., vol. 7, pp. 1-30, 2006.

[70] D. Antonelli, E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, and N. Mahoto, "Analysis of diabetic patients through their examination history," Expert Syst. Appl., vol. 40, no. 11, pp. 4672-4678, Sep. 2013.