WALAILAK PROCEDIA
Walailak Procedia 2019; 2019(1): IC4IR.176
http://wjst.wu.ac.th/index.php/wuresearch

Hybrid pre-processing models for heart disease prediction based on socioeconomic status and major risk factors: Thai heart study

Chalinee PARTANAPAT*, Chuleerat JARUSKULCHAI and Chanankorn JANDAENG

Division of Management of Information Technology, School of Informatics, Walailak University, Nikon Si Thammarat 80161, Thailand

(*Corresponding author's e-mail: chalinee.pa@mail.wu.ac.th)

Abstract

Heart disease is the leading cause of death in all over the world over past ten years. The ability to identify the risk factors related to an effective diagnosis is very important for improving accuracy on heart disease prediction. Major Risk Factors such as ECG, Angiography, an imaging modality for blood vessels, hypertension, diabetes, are currently the most accurate method for diagnosis. However, physical diagnosis based on only biological risk factors, which sometimes are reported to wrong diagnosis and treatment, which prompted this study to investigate alternate solutions. This paper is to enhance the prediction accuracy of the presence of heart disease based on SES (Socioeconomic Status) related to biological risk factors with reduced number of attributes. We examined whether every single SES measures like income or education addressed this bias and derived an approach of relevance to traditional risk factors by employing discretization and hybrid feature selection methods. Originally, thirteen biological risk factors are involved for predicting heart disease. In our work, we reduce number of biological risk factors by hybrid feature selection methods and discretization on some of continuous and numerical risk factors, and add five more SES factors towards prediction. Seven feature selection methods with four hybrid ones and discretization of equal depth are applied for reducing number of attributes to achieve more effective accuracy on prediction. Four classifiers are employed to predict the diagnosis. The observations exhibit that after adjusting discretization on numerical risk factors with Relief Attribute Eval algorithm combined with Bayes, our proposed method gives the highest accuracy with 94.01% classified by SVM. Thirteen biological attributes are reduced to six attributes and SES as income are involved. This experiment concludes that low income can cause high risk for heart disease. Discretization on continuous and numerical risk factors can improve performance on prediction accuracy. Performance of classification accuracy based on unsupervised discretization is compared, SVM is proved to be the best one for this study. The novelty of hybrid combination models and discretization methods are proved to enhance on classifying heart disease problems. Equal Depth Discretization with feature selection by Relief Attribute Evaluation and Bayes gives the better accuracy, when compared with no discretization and without feature selection.

Keywords: Pre-processing model, socioeconomic factor, Thai heart disease, data mining

Introduction

Many researchers have been studying about heart disease prediction based on only major risk factors using several data mining techniques to enhance performance of diagnosis. In many developing countries including Thailand, socioeconomic status also have impacts on heart disease mortality. Heart disease is the major cause of death and expected to become one of the major health problems and increasing in low-middle income countries including Thailand. The total of 58,681 death of Thai Patients are due to heart disease or 7 patients per hour in average (Nittaya, 2015). Normally, the deaths raised from heart disease in Thailand occur due to work overload, mental stress, and many other problems. Low Socioeconomic Status has been found a profound influence on heart disease of Thai patient (Vathesatogkit et al. 2012). This correlation motivates the greater prevalence of higher levels of blood pressure, obesity, and smoking, among people with lower levels of education and/or occupation. Other factors such as unhealthy life-styles, occupational or environmental hazards, could play a determinant role on the cause of coronary heart disease (Egeland et al. 2003). This study describes demographic and socioeconomic characteristics and their association to the heart disease and explores the high predictive risks of heart disease.

According to the heart disease database of Thai patient, most of diagnosis depends on physician's experiences. Effective treatment needs to be executed accurately and efficiently. Data mining can be a tool that has a set of techniques and algorithms for discovering the patterns and relationships of knowledge hidden in the Thai heart disease database. Feature Selection is a preprocessing technique used to remove irrelevant attributes and to identify the significant attributes, which play a dominant role in the task of classification. This leads to the dimensionality reduction. By applying different approaches, features can be reduced. The reduced feature set improves the accuracy of the classification task in comparison of applying the classification task on the original data set.

In this study, we use feature selection algorithms seeking for a subset of relevant features to use in model construction in order to simplify and reduce over-fitting of the models. The wrapper and filter methods are the two main categories that have been applied. Here, we proposed 4 classifiers working with 6 feature selection methods from filter group and 2 from Wrapper Group for classification. The redundant and insignificant attributes will be removed by using proposed and hybrid attribute

selection models. In order to prove the algorithms of attribute selection method, the attribute selection method that gives the higher accuracy after removing attributes for classification are combined. Then the reduced data are fed into sequence of classifiers classified to obtain better accuracy. However, some dataset characteristics can't be solved efficiently with general feature selection and normal classification algorithms (Gupta et al. 2011), such problem is solved using evolutionary algorithms. Additionally, many feature selection algorithms (Liu & Setiono, 1996) are shown to work effectively on discrete data or even more strictly, on binary data (and/or binary class value). The problem to attack is; Given data sets with numeric attributes (the range of each numeric attribute could be very wide), find an algorithm that can automatically discretize the numeric attributes as well as remove irrelevant/redundant ones. In order to deal with numeric attributes, a common practice for those algorithms is to discretize the data before conducting feature selection. Finally, we apply discretization before feature selection on some continuous or numerical features which is predefined by physician, and compare the accuracy results between filter group of feature selection of heart disease dataset with and without discretization.

Heart disease

Heart Disease is a disease in which a waxy substance called plaque) builds up inside the coronary arteries. These arteries supply oxygen-rich blood to your heart muscle. Existing research on heart disease has established that it is not a single condition, but refers to any condition in which the heart and blood vessels are injured and do not function properly, resulting in serious and fatal health problems (Chilnick, 2008; Health, 2010; King, 2004; Silverstein et al. 2006). According to survey of WHO, 17 million total global deaths are due to heart attacks and strokes. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. Heart disease was the major cause of casualties in the United States, England, Canada and Wales as in 2007. Heart disease kills one person every 34 seconds in the United States (http://en.wikipedia.org/wiki/Heart_disease). The World Health Organization has estimated that 17.7 million deaths occur worldwide, every year due to the cardiovascular diseases. Half the deaths in the United States and other developed countries occur due to cardiovascular diseases. It is also the chief reason of deaths in numerous developing countries. On the whole, it is regarded as the primary reason behind deaths in adults (http://chineseschool.netfirms.com/heart-disease-causes.html). In the year 2015, total death of 58,681 Thai Patients are due to Heart Disease (7 patients per hour in avg.). However, symptoms of each person are different. In Thailand, the common risk factors are hypertension, accounted for 83.2%, LDL accounted for 59.5%, diabetes accounted for 50.7%, smoking accounted for 32.1%. In this study, we apply various proposed preprocessing models to prove the major risk factors caused of heart disease for Thai population is the same as defined from physician. In majority of the cases, there is no early symptom and the disease is identifiable only in the advanced stage. Some common symptoms of heart disease are Chilnick (2008), Health (2010), Crawford (2002): chest pain (Angina pectoris); strong compressing or flaming sensation in the chest, neck or shoulders; discomforts in chest area; sweating, light-headedness, dizziness, shortness of breath; pain spanning from the chest to arm and neck, and that amplifying with exertion; cough; palpitations; fluid retention. The causes of heart disease are unclear, but age, gender, family history, and ethnic background are all considered to be the major causes in different researchers (Chilnick, 2008; Health, 2010; King, 2004; Silverstein et al. 2006). Other factors like eating habits, fatty foods, lack of exercise, high cholesterol, hypertension, pollution, life style factors, obesity, high blood pressure, stress, diabetes and lack of awareness have also been claimed to increase the chance of developing heart disease (Shantakumar & Kumaraswamy, 2009). Heart research, further, has found that the majority of the disease occurrence is noticed in people between the ages of 50-60 (Chilnick, 2008; Health, 2010; Silverstein et al. 2006), and in female more than male. For Thai heart study, a number of primary risk factors used in the experiment for prediction analysis are age, sex, hypertension, bad cholesterol, good cholesterol, blood sugar, exercise, cigarette smoking, alcohol consumption, diabetes, obesity, family history, and Triglyceride. These factors are used to analyze the Thai heart disease in this study. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. In this paper, analysis of various data mining techniques with some of feature selection algorithms and hybrid preprocessing models were used and helpful for medical analysts or practitioners for accurate and efficient heart disease diagnosis.

Socioeconomic risk factors

All new cases of Heart Disease in Thailand cannot be predicted using only primary or major risk factors like smoking, hypertension, cholesterol, diabetes, or family history. Therefore, studies aiming at the discovery of socioeconomic risk factors have been proposed in this research. Socioeconomic risk factors have been found to be the most important indicators associated with causing heart disease for Thai population. The term "socioeconomic status" covers a wide range of aspects, sometimes referred to as "social class," but it includes aspects of education, income, occupation, living conditions, income inequality, and many other socioeconomic aspects of life (Kaplan & Keil, 1993). An association between lower socioeconomic status (SES) and poorer health has been observed for several years. This research summarizes the evidence for an important association between socioeconomic status and heart diseases, both considering independently and conducting experiment with major or primary risk factors together.

It has been apparently appeared that heart disease patients has increased in industrialized countries, but variable patterns and trends are found within individual countries (Marmot, 1989; Lehman, 1976; Helsing & Comstock, 1977). Early studies indicated a positive association between CHD and SES, with the highest SES groups having the most disease (Cassel et al., 1971; Salonen, 1982. However, an inverse association has been observed in those countries, with similar patterns in New Zealand, Australia, and the Scandinavian countries (Jones et al. 1988; Leren et al. 1988; Koskenvuo et al. 1981; Simons et al. 1986). The lowest SES groups now have the highest CHD rates. Correspondingly, lower SES groups also have the least

favorable health characteristics, including obesity, cigarette smoking, hypertension, and lack of physical activity. However, the relative affluence of lower SES groups in many industrialized countries also leads to increased risk not balanced by positive health habits (Helsing & Comstock, 1977; Heller et al. 1984; Burr & Sweetnam, 1984). The relationship between SES and blood pressure or body mass index is less consistent than that of smoking. In the USA, the Stanford Five-City Project showed that age-adjusted mean systolic and diastolic blood pressure, body mass index, and prevalence of hypertension were inversely associated with level of education in both sexes (Winkleby et al. 1992; Winkleby et al. 1990). In England, men in lower socioeconomic groups had higher age-adjusted mean systolic blood pressure but not diastolic blood pressure (Smith et al. 1990). Different socioeconomic indicators may describe different aspects of socioeconomic position, e.g. education indicates skills requisite for acquiring positive social, psychological, and economic resources; occupation measures prestige, responsibility, physical activity, and work exposure; income reflects spending power, diet, and medical care (Winkleby et al. 1992; Helmert et al. 1990; Luoto et al. 1994). Although it is apparent that dynamic changes have occurred in CHD and SES patterns during this century, there is little information on trends between SES risk factors and heart disease. Thus, individuals at high risk for heart disease must be identified accurately and prevention programs must be designed to meet the needs of those specific groups.

Many studies have confirmed that the primary risk factors for heart disease are cigarette smoking, hypertension, serum cholesterol level, sedentary lifestyle, and diabetes. However, this study shows that these risk factors influence about 80% of overt cases of coronary heart disease. Numerous investigators have studied other possible risk factors, such as obesity, hostility, stress, noise, and coping styles. Based on the evidence of present lifestyle and eating behavior, we believe SES should be added to the list of potential risk factors for coronary disease. For example, Tian et al. reported that blood pressure was inversely associated with level of education in an urban population (Tian et al. 1996); and middle-aged male workers with lower educational attainment or heavier labor intensity had increased levels of cardiovascular risk factors in a study sample from seven steel and metal plants (Siegrist et al. 1990). Men with higher SES tended to have lower mean blood pressure levels, smoked fewer cigarettes per day, and had lower relative risks of being obese and of being current smokers (Zhijie et al. 2000). For decades, analyses of critical statistic methods have demonstrated both positive and inverse relation between SES and heart disease. This research was conducted on Thai Population with 5 socioeconomic risk factors which are income, education, occupation, jobs, and living condition. The Thai heart study was analyzed to determine whether the observed SES risk factors mentioned above are associated with heart disease.

The conceptualization and measurement of SES risk factors have been reviewed by several authors. The debate on the conceptualization of SES is concerned with sociological theory. According to medical criteria and sociological theory of background of Thai population, the most commonly used measures, indexes, and ecological measures of social class are reviewed below.

Education

Education is the most widely used measure of SES in epidemiologic studies. There are a variety of reasons for this risk factor. The level of education is usually fixed after young adulthood, it is likely that poor education relatively influences poor health among adults. The risk associated with low educational attainment is roughly comparable to many traditional CHD risk factors such as high cholesterol levels (Fiscella & Franks, 2004). However, education should be correlatively considered with social, behavioral, and psychological factors subject to age, because of the various background and lifestyle of each person. In this study, the education level of individuals was collected on an ordinal scale. For the purposes of this analysis, four categories were defined: high school graduate, vocational graduate, bachelor degree and higher than bachelor degree.

Income

Measures of income are obviously an important indicator of SES. Income provides ability to afford and access to goods and services, including quality education and medical care, which may protect against disease. In addition, lower income reflects the influence of poorer health. However, income was reported in only 15% of the articles published in the American Journal of Epidemiology that included measures of social class (Liberatos et al. 1998). The level of income is complicated, because it can be included with other sources of income than wages such as noncash benefits like food stamps or Medicare. We classified income factor into 4 categories; less than 10,000, 10,000 to 20,000, 20,000 to 50,000, and more than 50,000.

Occupation

Occupation is an important status characteristic in modern societies. Numerous studies have indicated that people definitely rank occupations in terms of prestige and status. Health outcomes for specific occupations have been examined in numerous studies. The categorizations of SES are based on issues of status and roles, power, prestige, lifestyle, job characteristics, income and education, traditions, beliefs, and values (Susser et al. 1985). Classification of occupation represents something quite different from job characteristics but associated with each other. The US Bureau of the Census has categorized occupations since 1897 (Lynch et al. 1996). This scale, based on putative social rank, divides occupations into the following 12 ordered categories: professional; technical and kindred workers; managers and administrators (other than farm); sales workers; clerical and kindred workers; craftsmen and kindred workers; operatives (except transport); transport equipment operatives; laborers (except farm); farmers and farm managers; service workers (excluding household); and private household workers (Lynch et al. 1996). However, numerous difficulties are associated with the terms of occupational measures. The breadth of occupational groupings could reflect the impact of predicting heart disease. For example, the chief executive officer of a large multinational corporation and a proprietor of a small family business would be the same rank in some systems, or a

skilled manual worker might have an income that considerably exceeds that of a university professor. In fact, Karasek et al. (Kuller) have argued that classifications based on characteristics such as decision latitude, time pressure, intellectual discretion, and other job-related characteristics provide a better way of grouping occupations with respect to SES. In this study, we classified occupation into 4 groups subject to Thai culture and history information as Manager, intermediate, self-employed, and technical.

Jobs

Employment Status is also possible to characterize people in terms of current employment status. The important consideration is the association between employment status and health (Kaplan & Keil, 1993). However, it is critical to distinguish between those who are able to work but cannot find employment and those who are unable to work for health reasons. We simply categorize level of job into 2 groups; employed and unemployed.

Living condition

Measures of living condition indicated conditions under which people live. For example, ownership of a house, automobile has been used in some research (Fox & Goldblatt, 1982; Kaplan & Salonen, 1990). Such measures will be highly correlated with income and education, and also lifestyle differences. In this study, living condition is classified into 2 groups: rented and owner. A large number of studies have used measures of living condition as in which subjects live, like geographic units with respect with to income, education, occupation, crowding, condition of housing, value of homes or rental prices. (Hongmei et al. 2006; Mokeddem et al. 2013; Chilnick, 2008; Health, 2010; King, 2004; Silverstein et al. 2006). It has been cleared that the effect of affluence on lifestyle is an increase on consciousness in high socioeconomic status groups, resulting in improved health behaviors and lower risk.

Related works

The heart disease prediction for Thai Heart Disease Patient can be divided into two paradigms in the context as followings;

The first part is the socioeconomic risk factors. To predict heart disease independently using socioeconomic status factors. Luepker et al. (2015) applied Linear Regression and Logistic Regression methods using SAS Statistical Software. The result showed that Age and sex are associated with education and income levels are significant for Heart Disease Cause. Lynch et al. (1996) analyzed the relation between Socioeconomic Status and Risk of All-Cause Cardiovascular Mortality. Twenty three risk factors were conducted in the experiment. Association between socioeconomic factors was assessed by Cox proportional hazard models using SAS Statistical Software. The result showed that each of the 23 risk factors used was significantly associated, especially when adjusting age with income. Zhijie, et al. (2000) conducted the survey to investigate the association between socioeconomic status and cardiovascular heart disease major risk factors in urban China. A sample of 4000 people aged 15-69 years, stratified by sex and 10-year age groups, was drawn randomly from urban areas of the city. Four socioeconomic indicators (education, occupation, income, and marital status), blood pressure, body mass index, and cigarette smoking were determined in the survey. Their findings showed that education levels seemed to be the most significant factor of the four socioeconomic indicators associated with cardiovascular risk factors.

In the second part, the data mining techniques are applied to classify the patients whether they have the chance of occurrence of heart disease. Aha and Kibler used instance based algorithms to predict heart disease. They used C4.5 algorithm and achieved an accuracy of 74.8% (Rupali & Patil, 2014). Elma et al. (2013) presented a new classifier which combines K Nearest Neighbor and Naïve Bayes Classifier to improve accuracy for predicting heart disease dataset and achieved an accuracy of 82.96%. Bhatia et al. (2008) developed SVM Based Decision Support System for Heart Disease Classification with inter-Coded Genetic Algorithm to Select Critical Features. To achieve this, the experiment was conducted on UCI dataset via optimization of Kernel. Shouman et al. (2011) proposed diagnosing heart disease patients using decision tree J48 and Bagging Algorithms and introduced the feature Selection methods as Information Gain, GINI, and Gain Ratio which can identify ranking of most significant features. The highest accuracy is 84.1% measured by Equal Frequency. Saravanakumar & Rinesh (2014) proposed Effective Heart Disease Prediction. To get the effective result, they used Maximal Frequent Itemset Algorithm (MAFIA) for feature selection classified based on Non Linear Integrals.

Several studies have focused on finding out efficient data mining methods for heart disease diagnosis based on just major risk factors and statistical methods for socioeconomic risk factors. Our approach is an attempt to predict efficiently diagnosis with reduced number of factors (attributes) that contributes more accuracy and better performance toward prediction based on both major risk factors and socioeconomic risk factors by using hybrid preprocessing models with various classifiers. The aim of this study is to analyze effects between each socioeconomic risk factors and primary risk factors on Thai Heart Study and to propose the approaches and algorithms for predicting Thai Heart Study most accurately. The remainder of this paper is organized in the following manner. Section methods explain the dataset used, the algorithms for feature selection and illustrates classification process and outcomes. Section results and discussion is dedicated to the evaluation of the mining results and discussion. Finally, we draw an experiment's summary in section conclusion.

Methods

Dataset

Thai Heart Disease Dataset from Rama Hospital of 500 records and 20,000 records with 18 attributes were used in this experiment comparatively. The 13 major risk factors (attributes), which have the corresponding predefined values by medical field are listed in **Table 1**.

Table 1 Description of attributes for major risk factors.

No.	Attributes	Descriptions	Values
1	Gender	Male or Female	0 = Female 1 = Male
2	Age	Age in years	Continuous
3	HDL	Good cholesterol in mg/dl	0 = >= 40 mg/dl 1 = < 40 mg/dl
4	LDL	Bad cholesterol in mg/dl	0 = <= 130mg/dl 1 = >130 mg/dl
5	fbs	Fasting Blood Sugar	0 = <= 120mg/dl 1 = >120 mg/dl
6	hp	Hypertension	0 = No 1 = Yes
7	Exercise	Physical Activity	0 = No Exercise 1 = Less Than 3d/wk 2 = Greater or Equal than 3d/wk
8	Smoke	Smoking Habit	0 = Non Smoking 1 = Smoking
9	Alcoholic	Alcoholic Habit	0 = Non Drinker 1 = Drinker
10	Diabetes	Diabetes Mellitus	0 = No 1 = Yes
11	Obesity	Measured By BMI	0 = No 1 = Yes
12	Hereditary	Family Member Diagnosed with HD	0 = No 1 = Yes
13	Tg	Triglyceride	0 = 0 – 200 mg/dl 1 = >200 mg/dl

The other 5 socioeconomic risk factors (attributes) are assessed into the analysis to get more optimal and effective prediction result in evaluating whether the relationship is independent from or associated with major risk factors to heart disease. The socioeconomic risk factors are listed in **Table 2**.

Table 2 Description of attributes for socioeconomic risk factors.

No.	Attributes	Descriptions	Values
	Income	Income levels	0 = "<=10,000" 1 = ">10,000 - 20,000" 2 = ">20,000 - 50,000" 3 = ">=50,000"
2	Edu	Education Levels	0 = Secondary 1 = Vocational 2 = Bachelor 3 = Higher than Bachelor
3	Occupation	Occupation Levels	0 = Manager 1 = Intermediate 2 = Self-Employed 3 = Technical
4	Jobs	Employment status	0 = Unemployed 1 = Employed
5	liv. cond.	Living condition	0 = Rented 1 = Owner

The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with no heart disease and value "0" for patients with heart disease.

Feature selection

Feature Selection is the process of detecting and eliminating irrelevant, weakly relevant or redundant attributes or dimensions in a given data set. Feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. Removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models with greater accuracy.

1) *Correlation Based Feature Selection (CFS):* The CFS algorithm evaluates subsets of features on the basis of good feature subsets contain features highly correlated with the classification.

2) *Bayes Theorem:* describes the probability of an event based on conditions that might be related to the event. This is a conditional probability, the probability that one proposition is true provided that another proposition is true. It has been used to try to clarify the relationship between theory and evidence, supposing one is interested in whether a person has heart disease, and knows the person's age. If heart disease is related to age, then, using Bayes' theorem, information about the person's age can be used to more accurately assess the probability that they have heart disease.

3) *CFS and Bayes Theorem*: We proposed a new feature selection method by combining CFS with Bayes Theorem. Each attribute in CFS algorithm is compared pair wise to find the similarity and the attributes are compared to class attribute to find the amount of contribution it provides to the class value, based on these, the attributes are removed. The selected attributes from the CFS algorithm is fed into Bayes theorem for further reduction. Bayes theorem calculates the conditional probability for each attribute and the attribute which has highest conditional probability is selected.

4) *Wrapper Subset Evaluation:* The method considers the selection of a set of features as a search problem, where different combinations are prepared, evaluated, and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy to detect the possible interactions between variables.

5) *Chi-squared attribute evaluation:* This algorithm is used to test of association or independence of features. It will tell us whether there is a large difference between collected numbers and expected numbers. If the difference is large, it indicates that there may be something causing a significant change. A significantly large difference will be defined as no interaction between variables.

6) *Info Gain attribute evaluation:* The expected reduction in Entropy is caused by partitioning the examples according to a given attribute. Entropy characterizes the impurity of a collection of examples.

7) *Random Tree:* selects features based on random with replacement method and group every subset in a separate subspace. It is relatively robust to outliers and noise. It runs efficiently on large databases and also increases accuracy and decrease training time.

8) *SVM Atttribute Evaluation:* This algorithm evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM.

Classification used for prediction

Classification is one technique of data mining, which is a supervised learning method to extract models describing important data classes or to predict future trends. Our work intends to use four classifiers: Decision Tree J48, Naïve Bayes, MLP and SVM to diagnose the presence of heart disease in patients.

- Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data.
- Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Our work uses J48 decision tree for classification.
- Multi Layer Perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. It consists of multiple layers of nodes in a directed graph, and each layer is fully connected to the next one. MLP utilizes a supervised learning technique called back propagation for training the network. MLP can distinguish data that are not linearly separable.
- Support Vector Machine (SVM) is a supervised ML method, used for classification. It is widely used to produce a predicting model. For each given test input, SVM predicts which of two possible classes forms the input, making it a non-probabilistic binary linear classifier (Karaolis, 2010).

Discretization

Some columns may contain so many values that the algorithm cannot easily identify interesting patterns in the data from which to create a model. *Discretization* is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values. Many data mining systems work best with qualitative data, where the data values are discrete descriptive terms such as young and old. However, lots of data are quantitative, for example, with age being represented by a numeric value rather a small number of descriptors. One way to apply existing qualitative systems to such quantitative data is to transform the data. In data mining view, discretization is data pre-processing method that transforms quantitative data to qualitative data. Commonly attribute values in quantitative data are in medical data dataset. Discretization divides the value range of the quantitative attribute into a finite number of intervals. There are many learning algorithms are available to handle qualitative data. Even some algorithms can directly deal with quantitative data, but the learning often is not that much efficient and effective. So there is need to convert the quantitative data to qualitative data. In this study, we apply discretization to some of continuous and quantitative values to achieve the new range of values from each of risk factors. Dataset reduction requires categorical data. Consequently, data discretization is the first step. We used an approximate equal interval binning method to bin the data variables into a small number of categories.

Experiments

Experiments were conducted with WEKA 3.7 tool. Thai heart disease dataset consisted of 500 random records and 20,000 full dataset with 18 attributes were applied for prediction. First of all, we applied 500 dataset with all risk factors together and find out the accuracy. Secondly, the feature selection methods, which are wrapper and filter group including the proposed hybrid ones are applied to achieve the important attributes. Thirdly, the reduced numbers of attributes from feature selection will be fed to various classifiers. Next step, discretization is applied to some of continuous and categorical risk factors with both 500 and 20,000 dataset and then compare the accuracy. Last step, the results of different data mining algorithms using performance criteria are compared and analyzed.

Results

The experiments are produced by the proposed wrappers. We first calculated average accuracies of the wrapper algorithms coupled with BN classifier. In addition, each wrapper strategy produces feature subsets and we made experiments to measure effectiveness of these features used for heart disease diagnosis. The results of the experiments are evaluated with accuracy. Additionally, we involved the performance of classifiers without Feature Selection in **Table 3**. The list of features or risk factors for heart disease dataset generated by the wrapper algorithms is shown in **Table 4**.

Table 3 The performance evaluation of wrapper based feature selection algorithms.

Wrapper Algorithms	Accuracy of different Classifier methods			
	BN	SVM	MLP	J48
GA Wrapper	89.50	88.82	85.60	84.55
BFS Wrapper	87.50	84.53	83.50	84.18
SFFS Wrapper	88.18	87.17	81.89	83.50
Without FS	86.50	84.50	80.88	80.12

After applying feature selection generated by wrapper, **Table 3** shows the accuracy acquired using10-fold cross validation. GA wrapper with BN produces the most efficient feature model with the diagnosis accuracy of 89.50%. Instead, it can be examined that feature selection algorithms produces a strength features compared to full dataset (without FS). In addition, these classification performance results show that the feature model engendered with GA wrapper approach is powerful. The accuracy of different classifier methods with wrapper feature selection is shown in **Table 3**.

Table 4 The Selected Features by different wrapper Techniques.

No	Feature	Selected Features by FS approaches					
		GA wrapped BN	GA wrapped SVM	GA wrapped MLP	GA wrapped J48	BFS wrapped BN	SFFS wrapped BN
1	Gender	P	P	P	P	P	P
2	Age	P	P	P	P	P	P
3	Good Cholesterol	P	P		P		
4	Bad Cholesterol	P		P		P	P
5	Fasting Blood Sugar		P	P		P	
6	Hypertension		P	P	P		
7	Exercise	P		P			
8	Smoke	P	P		P	P	P
9	Alcoholic	P	P	P			
10	Diabetes	P	P	P	P		
11	Obesity		P				
12	Hereditary	P	P				
13	Triglyceride	P	P		P		
14	Income	P	P	P			
15	Education		P	P	P	P	P
16	Occupation						
17	Jobs		P			P	P
18	Living Condition						

After getting the result from wrapper feature selection, GA wrapper with BN classifier is proved to be the best accuracy of 89.50% However, the same dataset are reduced using the other attribute selection method, which are filter feature selections. The attribute selector methods are automated, where Gain Ratio Attribute Evaluation, One Attribute Eval, Chi-Squared Attribute Evaluation, and Relief Attribute Evaluation are applied. After applying these attribute selection methods, the number of reduced attributes by each attribute selection methods are shown in **Table 5**.

Table 5 Number of selected attributes by each attribute selection method.

Attributes Selection Methods	Number of Attributes Selected
Gain ratio attribute Evaluation	8(1, 2, 3, 4, 6, 8, 13, 14)
One attribute eval	8(1, 2, 4, 5, 8, 10, 13, 14)
Chi-squared attribute Evaluation	9(1, 2, 4, 5, 7, 8, 10, 13, 14)
Relief attribute evaluation	7(1, 2, 3, 6, 8, 10, 14)

As the result in the table shown above, the attributes are reduced as 7 applied by Relief Attribute Evaluation. Major risk factors, which are biological factors and some of behavioral factors like sex, age, LDL, Smoke, and Socioeconomic risk factor as education is selected in every attribute selection method. The study implies that sex, age, LDL, Smoke, and education are the significant risk factors for Thai heart disease patients. After reducing the number of attributes, the resulting data is given to the classification algorithms. The result shows improvement in accuracy for all of classifiers. **Table 6** represents comparison of accuracy of classifiers with reduced attributes.

Table 6 Accuracy of classifiers with reduced attributes (in%).

Attributes Selection Methods	Accuracy of				Average
	NB	J48	MLP	SVM	
Chi-squared attribute Evaluation	83.70	81.66	82.88	85.50	**83.44**
One attribute eval	87.00	84.70	85.18	87.50	**86.10**
Gain ratio attribute Evaluation	87.50	85.18	86.88	88.00	**86.89**
Relief attribute evaluation	88.89	85.97	87.50	89.50	**87.97**

As seen in the **Table 6**, Relief attribute evaluation selection method classified by SVM gives the best accuracy of 89.50%. The result is found that classification gives the better accuracy after applying feature selection methods. Age, sex, LDL, smoking, and low education are the significant risk factors for causing heart disease for Thai patients with the percentage of accuracy of 89.50%, classified by SVM classifier.

The four hybrid feature selector methods are applied in sequence, which are Chi-squared attribute evaluation with Bayes Theorem, One R with Bayes Theorem, Gain Ratio with Bayes Theorem, and Relief Attribute Evaluation with Bayes Theorem to determine the better accuracy on prediction. The numbers of reduced attributes selected by these four methods are shown in **Table 7**.

Table 7 Number of selected features by proposed hybrid selection methods: Chi-squared attribute evaluation with Bayes Theorem, One R with Bayes Theorem, Gain Ratio with Bayes Theorem, and Relief Attribute Evaluation with Bayes Theorem respectively.

Attributes Selection Methods	Number of Attributes Selected
Chi-squared attribute Evaluation + Bayes Theorem	9(1, 2, 3, 4, 5, 8, 10, 13, 14)
One R + Bayes Theorem	7(1, 2, 4, 8, 10, 13, 14)
Gain Ratio + Bayes Theorem	7(1, 2, 3, 4, 6, 8, 13)
Relief Attribute Evaluation + Bayes Theorem	6(1, 2, 3, 8, 10, 14)

As the hybrid feature selection methods shown above, the reduced numbers of attributes after applying Chi-squared attribute evaluation is 9, One attribute evaluation is 8, Gain Ratio attribute evaluation is 8, and Relief attribute evaluation is 7 respectively. These selected attributes are fed into Bayes Theorem, which produces the same number of 9 attributes if fed by Chi-squared attribute evaluation, reduces to 7 attributes if fed by One attribute evaluation and by Gain Ratio Attribute Evaluation, and reduces to 6 attributes when fed by Relief attribute evaluation. The result also indicates that Age, Sex, LDL, Smoke, and Education are the significant risk factors to cause heart disease. Nevertheless, after applying the hybrid feature selectors, this reduced dataset that is fed into four classification algorithms is proved in improvement of accuracy. The comparison of accuracy of classifiers with reduced attributes is represented in **Table 8**.

Table 8 Comparison of Classifiers Accuracy after feature selection between Bayes and Chi-squaredattributeEval, Bayes and One attributeEval, Bayes and Gain Ratio attribute Eval, and Bayes and Relief attribute Eval (in%).

Attributes Selection Methods	Accuracy of				Average
	NB	J48	MLP	SVM	
Chi-squared attr eval + Bayes	85.92	83.29	84.07	87.50	**85.20**
One attribute eval + Bayes	88.00	85.88	86.50	88.67	**87.26**
Gain ratio attr eval + Bayes	88.18	86.50	87.88	89.50	**88.02**
Relief attribute eval + Bayes	90.50	86.67	88.50	92.59	**89.57**

As a result shown in the **Table 8**, Relief Attribute Evaluation + Bayes Theorem attribute selection method classified by SVM gives the best accuracy of 92.59%. The result is shown that classification produces the better accuracy after applying hybrid feature selection methods. Age, sex, LDL, smoking, and low education are still the significant risk factors for causing heart disease for Thai patients. Diabetes is appeared to show the impact on prediction and can have the critical effects on Thai Heart Disease, associated with age, sex, cholesterol and smoking.

After getting the result from some of hybrid feature selection methods conducted, the new hybrid algorithm, Relief Attribute Evaluation with Bayes is applied and broadly accepted as high performance feature selectors, and chosen to propose as the feature selectors to enhance the performance of accuracy on Thai heart disease prediction and compare with the previous ones. In this phase, each feature selector produces reduced numbers of attributes selected to find the risk forecast of Thai heart patients. After applying the attribute selection methods, the number of reduced attributes by each attribute selection method, including CFS and Bayes Theorem are shown in **Table 9**.

Table 9 Attributes selected from 2 wrapper methods and 2 proposed hybrid feature selection methods.

Attributes Selection Methods	Attributes Selected
GA Wrapper BN	8 (1, 2, 3, 4, 7, 8, 10, 13)
Wrapper Subset Evaluation	7 (1, 2, 3, 4, 8, 10, 14)
CFS + Bayes Theorem	6 (1, 2, 4, 8, 10, 14)
Relief Attribute Evaluation + Bayes Theorem	6 (1, 2, 3, 8, 10, 14)

1 2 8 10
Sex, Age, Smoke, Diabetes

According to the result in the table shown above, the attributes are reduced as 6 applied by CFS + Bayes Theorem. Major risk factors, which are biological factors and some of behavioral factors like sex, age, LDL, Smoke, Diabetes and Socioeconomic risk factor as education is selected in every attribute selection method. The study implies that sex, age, LDL, Smoke, Diabetes and education are the significant risk factors for Thai heart disease patients. After reducing the number of attributes, the resulting data is given to the classification algorithms. The result shows improvement in accuracy for all of classifiers. **Table 10** represents comparison of accuracy of classifiers with reduced attributes.

Generally, results of our experiments perceived that the proposed algorithm has the highest classification accuracy in the literature. This study presents comparison among wrapper model, filter feature selection, and our proposed hybrid model. The result shows that Hybrid model of Relief Attribute Evaluation with Bayes Theorem gives the best accuracy of 92.59% for heart disease diagnosis in Thai patient. The comparison of classification accuracy between wrapper group and hybrid methods is shown in **Table 10**.

Table 10 Comparison of the results between wrapper group and Hybrid methods.

Attributes Selection Methods	Accuracy classified by SVM
GA Wrapper BN	89.50
Wrapper Subset Evaluation	88.54
CFS + Bayes Theorem	90.94
Relief Attribute Evaluation + Bayes Theorem	92.59

As seen in the **Table 10**, Relief Attribute Evaluation+Bayes Theorem attribute selection method classified by SVM gives the best accuracy of 92.59%, CFS+Bayes Theorem attribute selection method classified by SVM produces the second best accuracy of 90.94%. Both of them are filter selection methods. It is found that classification gives the better accuracy after applying hybrid models of feature selection methods. Age, sex, LDL, smoking, Diabetes, and low education are the significant risk factors for causing heart disease for Thai patients with the percentage of accuracy of 92.59%, classified by SVM classifier. Compared with the first four feature selection methods in this experiment, Diabetes is not the critical factor as in the last four feature selection methods, for causing heart disease for Thai Patient. In addition, Relief Attribute Evaluation + Bayes Theorem attribute selection method which gives the highest accuracy and is performed better than CFS with Bayes and than gain ratio with bayes based on accuracy, shows that diabetes is the significant factor as it is selected and corresponding to presence of risk to Thai heart disease.

In the experiment conducted above, we set up with the dataset of 500 records as a pilot study. The full dataset in this study is performed to obtain if it has the better accuracy. According to the attribute selection method proposed above, we achieved 5 rank of best accuracy as shown in **Table 11**. Therefore, we applied full dataset of 20,000 with these five attribute selectors to see the difference of the number of attributes selected. **Table 11** shows the number of selected features by proposed wrapper group and hybrid selection methods differentiated by numbers of dataset.

Table 11 Number of selected features by proposed hybrid selection methods differentiated by Numbers of Dataset.

	N=500	N=20,000
Attributes Selection Methods	**Attributes Selected**	**Attributes Selected**
GA Wrapper BN	8 (1, 2, 3, 4, 7, 8, 10, 15)	10(1, 2, 3, 4, 7, 8, 10, 13, 14, 15)
Wrapper Subset Evaluation	7 (1, 2, 3, 4, 8, 10, 15)	9(1, 2, 3, 4, 7, 8, 10, 13, 15)
CFS + Random Tree	7(1, 2, 3, 5, 6, 10, 13)	9(1, 2, 3, 4, 5, 6, 8, 10, 13)
CFS + Bayes Theorem	6 (1, 2, 4, 8, 10, 14)	8(1, 2, 4, 6, 8, 10, 13, 14)
Relief Attribute Evaluation + Bayes Theorem	6 (1, 2, 3, 8, 10, 13)	8(1, 2, 4, 6, 8, 10, 13, 15)

According to the table shown above, full data set of 20,000 is proved to achieve more number of attributes selected in every attribute selectors, compared with 500 dataset. In addition, the hybrid feature selection methods in filter group shown above, Relief attribute evaluation with Bayes and CFS with Bayes produced the least numbers of attributes, that is 8. The more number of dataset, the more number of attributes selected. It can be inferred that the pilot study of 500 patients cannot be as a representative as a full dataset of Thai patients. Nevertheless, the accuracy of full dataset is slightly decreased, compared with pilot study. Therefore, this full dataset of Thai heart disease can be implied reliable. The result also indicates that Age, Sex, LDL, Smoke, Diabetes, Triglyceride and Education are the significant risk factors to cause heart disease. The comparison of the accuracy of each attributes selection methods classified by SVM differentiated by number of dataset is shown in **Table 12**.

Table 12 Comparison of the accuracy of each attributes selection methods classified by SVM.

	N=500	N=20,000
Method	**Classification Accuracy (%)**	**Classification Accuracy (%)**
GA Wrapper BN	89.50	88.82
Wrapper Subset Evaluation	88.54	87.18
CFS + Random Tree	90.55	87.89
CFS + Bayes Theorem	90.94	89.55
Relief Attribute Evaluation + Bayes Theorem	92.59	90.92

According to the performance measures of all feature selection models proposed above, some attributes are continuous values and have the predefined range of value from the medical theory, which is not well suited for lifestyle in the present time. For example, attribute such as age, in the clinical value, is continuous, which is not clear enough to be the input for heart disease prediction. In this section, we propose the method to enhance preprocessing models for prediction. We apply discretization on 6 risk factors that have continuous and predefined categorical values before feeding into the best feature selectors above, that is Relief Attribute Evaluation with Bayes. Discretization refers to the process of converting or the different continuous attributes, variables to discrete or nominal attributes using discretization. The new discretized value of each factor will be automated binning. However, many classification algorithms require that the values of the factors contain only discrete attributes, and some would work better on discretized or binarized data (Burr & Sweetnam, 1984). If those numeric data can be automatically transformed into discrete ones or new range of values, the classification algorithms in this study would be work more efficiently. Discretization designed to 6 attributes is shown in **Table 13**.

Table 13 Discretization on continuous and some of numeric factors of 6 risk factors.

Sr.No.	Attributes	Descriptions	Clinical Values	Discretized Values
1	Age	Age in years	Continuous	0 = 0 - 37
				1 = >=37 - 55
				2 = >= 55
2	LDL	Bad Cholesterol in mg/dl	0 = <= 130 mg / dl	0 = <= 100 mg / dl
			1 = > 130 mg / dl	1 = >= 100 - 130 mg / dl
				2 = >=130 - 160 mg / dl
				3 = >= 160 mg / dl
3	fbs	Fasting Blood Sugar	0 = <=120 mg/dl	0 = <=125 mg/dl
			1 = >120 mg/dl	1 = >125 mg/dl
4	hp	Hypertension	0 = No	0 = <=120 mm/hg
			1 = Yes	1 = 120 - 150 mm/hg
				2 = >=150 mm/hg
5	Obesity	Obesity (Measured By BMI)	0 = 0 - 25	0 = 19 - 22
			1 = > 25	1 = 23 - 25
				2 = 26 - 30
				3 = > 30
6	tg	Triglyceride	0 = 0 - 200 mg/dl	0 = 0 - 150 mg/dl
			1 = > 200 mg / dl	1 = > 150 - 199 mg / dl
				2 = >= 200 - 499 mg / dl

According to **Table 14**, the dataset both 500 and 20000 records, are performed comparatively, and classified using Relief Attribute Evaluation with Bayes with and without discretization, measured by accuracy. **Table 14** shows the comparison of the accuracy of different algorithms between with feature selection, without feature selection and Discretization, and with discretization and feature selection classified by 4 classifiers, namely Naïve Bayes, J48, MLP, and SVM, respectively.

Table 14 Comparison of the accuracy of different algorithms between with feature selection, without feature selection and Discretization, and with discretization and feature selection classified by 4 classifiers.

Classifiers	without Discretization and Feature Selection		with Feature Selection by Relief Attribute Eval + Bayes		with Discretization and Feature Selection by RAE + Bayes					
No. of Dataset	500	20,000	500	20,000	500	20,000	500	20,000	500	20,000
					4 Factors		5 Factors		6 Factors	
Naïve Bayes	85.70%	83.18%	90.50%	88.82%	92.67%	90.80%	93.50%	91.33%	94.78%	92.05%
J48	83.71%	81.67%	89.67%	87.55%	90.80%	88.82%	91.88%	89.02%	92.59%	90.67%
MLP	83.66%	80.67%	85.67%	83.59%	90.78%	87.59%	92.78%	90.55%	93.82%	91.89%
SVM	87.23%	85.55%	92.59%	90.92%	94.01%	91.92%	94.58%	92.58%	95.05%	93.10%

As a result in **Table 14**, the first column presents classification by 4 classifiers on heart dataset without feature selection and discretization. The second column shows classification by 4 classifiers with feature selection but not discretized. The last column presents discretization on 4, 5, and 6 risk factors respectively. Four factors that are discretized for the first round are age, fasting blood sugar, obesity, and triglyceride. Five factors discretized for the second round are the four factors performed in the first round and hypertension added. Six factors discretized for the third round are the five factors conducted in the second round and LDL included. From the results of all performance measures, we can say that the proposed system designed using discretization gives the higher accuracy than without discretization. Nevertheless, the accuracy of full heart dataset is a little bit less than that of pilot study, which is not much different. According to the result, discretization can improve the accuracy of all classifiers. Discretization on 6 risk factors with full dataset of 20,000 records of Thai heart disease patient classified by SVM gives the best accuracy as 93.10%, when compared with predefined value from medical theory of 85.55% as without feature selection and without discretization, and 90.92% as with feature selection but without discretization.

Conclusions

Performance of classification accuracy based on unsupervised discretization is compared with two different numbers of dataset. The more numbers of dataset, the more features selected are produced. It leads to lightly reduce the performance of prediction. The novelty of proposed approach is demonstrated. The combination of hybrid models and discretization methods are proved to enhance on classifying heart disease problems. Equal Depth Discretization with feature selection by Relief Attribute Evaluation and Bayes gives the better accuracy, when compared with no discretization and without feature selection. Support Vector Machine produces the best accuracy of 93.10%. As a way to validate the proposed system, we have tested with emphasis on heart disease on dataset taken from Rama hospital. Experimental results carried out on Heart disease dataset using the three approaches and it shows that discretization method improves the accuracy than traditional classifiers. The experiments confirm that our proposed method results show a significant performance in the form of classifier accuracy improvements. This prediction model helps the doctors in efficient heart disease diagnosis process with less information. Age, Sex, LDL, Triglyceride, Hypertension, Smoking, and Diabetes are strongly better indicators of being risky to be a Heart Disease for Thai Population when applied with full dataset and discretized, because they give the better accuracy. Age, LDL, Triglyceride, and Hypertension are strongly associated with heart disease after applying the discretization to adjust the new range of values.

References

Bhatia, S., Prakash, P., & Pillai, G.N. (2008). *SVM based decision support system for heart disease classification with inter-coded genetic algorithm to select critical features. In*: Proceedings of the World Congress on Engineering and Computer Science.

Burr, M.L., & Sweetnam, P.M. (1984). Family size and paternal unemployment in relation to myocardial infarction. *Journal of Epidemiol Community Health 34*, 93-95.

Cassel, J., Heyden, S., Bartel, A.G., Kaplan, B.H., Tyroler, H.A., Cornoni, J.C., & Hames, C.G. (1971). Incidence of coronary heart disease by ethnic group, social class, and sex. *Archives of Journal of Internal Medicine 128*, 901-906.

Chilnick, L.D. (2008). *Heart disease: An essential guide for the newly diagnosed*. Da Capo Press.

Crawford, M.H. (2002). *Current diagnosis & treatment in cardiology*. McGraw-Hill Professional.

Egeland, G.M., Tverdal, A., Selmer, R.M., & Meyer, H.E. (2003). Socioeconomic status and coronary heart disease risk factors and mortality: Married residents, three countries, Norway. *Norsk Epidemiologi 13*(1), 155-162.

Ferdousy, E.Z., Islam, M.M., & Matin, M.A. (2013). Combination of naïve bayes classifier and K-NN in the classification based Predictive models. *Computer and Information Science 6*(3), 48-56.

Fiscella, K., & Franks, P. (2004). Should years of schooling be used to guide treatment of coronary risk factors? *The Annals of Family Medicine Journal 2*(5), 469-473.

Fox, A.J., & Goldblatt, P.O. (1982). *Longitudinal study 1971-1975: England and Wales*. Office of Population Censuses and Surveys. London: Her Majesty's Stationery Office.

Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering 2*(2), 188-195.

Haan, M., Kaplan, G.A., & Camacho, T. (1987). Poverty and health: Prospective evidence from the Alameda County Study. *American Journal of Epidemiology 125*, 989-998.

Health, M. (2010). Heart disease. Retrieved from http://www.mamashealth.com/Heart_disease.asp

Heart disease. (2018). Retrieved from http://chineseschool.netfirms.com/heart-disease-causes.html

Heart disease. (2018). Retrieved from http://en.wikipedia.org/wiki/Heart_disease

Heller, R.F., Williams, H., & Sittampalam, Y. (1984). Social class and ischemic heart disease: use of the male: Female ratio to identify possible occupational hazards. *Journal of Epidemiology Community Health 38*, 198-202.

Helmert, U., Shea, S., Herman, B., & Greiser, E. (1990). Relationship of social class characteristics and risk factors for coronary heart disease in West Germany. *Journal of Public Health 104*, 399-416.

Helsing, K.J., & Comstock, G.W. (1977). What kinds of people do not use seat belts? *American Journal of Public Health 67*, 1043-1049.

Hongmei, Y., Yingtao, J., Jun, Z., Chenglin, P., & Qinghui, L. (2006). A multilayer Perceptron based medical decision support system for heart desease diagnosis. *Expert Systems with Applications 30*, 272-281.

Jabbar, M., Deekshatulu, B.L., Chandra, P., & Pillai, G.N. (2013). *Heart disease prediction using lazy associative classification. In*: International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing.

Jones, A., Davies, D.H., Dove, J.R., Collinson, M.A., & Brown, P. (1988). Identification and treatment of risk factors for coronary heart disease in general practice: A possible screening model. *British Medical Journal 296*, 1711-1714.

Kaplan, B.H, Cassel, J.C, Tyroler, H.A., Cornoni, J.C., Kleinbaum, D.G., & Hames, C.G. Occupational mobility and coronary heart disease. *Archives of Journal of Internal Medicine 128*, 938-942.

Kaplan, G.A., & Keil, J.E. (1993). Socioeconomic factors and cardiovascular disease: A review of the literature. *American Heart Association 88*, 1973-1998.

Kaplan, G.A., & Salonen, J.T. (1990). Socioeconomic conditions in childhood and ischaemic heart disease during middle age. *British Medical Journal 301*, 1121-1123.

Karaolis, M.A. (2010). Assessment of the risk factors of Coronary Heart Disease Events based on data mining with decision trees. *Information Technology in Biomedicine, IEEE Transactions on 14*(3), 559-566.

Khempila, A., & Boonjing, V. (2011). *Heart disease classification using neural network and feature selection. In*: Proceedings of the 21st International Conference on Systems Engineering.

King, L. (2004). *Taking on heart disease*. Rodale

Koskenvuo, M., Kaprio, J., Romo, M., & Langinvainio, H. (1981). Incidence and prognosis of ischemic heart disease with respect to marital status and social class: A national record linkage study. *Journal of Epidemiology Community Health 35*, 192-196.

Kuller, L.H. Epidemiology of cardiovascular diseases: Current perspectives. *American Journal of Epidemiology 104*, 425-496.

Lehman, E.W. (1976). Social class and coronary heart disease: A sociological assessment of the medical literature. *Journal of Chronic Diseases 20*, 381-391.

Leren, P., Helgeland, A., Hjermann, I., & Holme, I. (1988). The Oslo study: CHD risk factors, socioeconomic influences, and intervention. *American Heart Journal 106*, 1200-1206.

Liberatos, P., Link, B.G., & Kelsey, J.L. (1988). The measurement of social class in epidemiology. *Epidemiologic Reviews 10*, 87-121.

Liu, H. & Setiono, R. (1996). *A probabilistic approach to feature selection: A filter solution. In*: Proceedings of the 13th International Conference on International Conference on Machine Learning.

Luepker, R.V., Rosamond, D., Murphy, R., Sprafka, J.M., Folsom, A.R., McGovern, P.G., & Blackburn, H. (2015). Socioeconomic status and coronary heart disease risk factor trends: The Minnesota heart survey. *Circulation 88*, 2172-2179.

Luoto, R., Pekkanen, J., Uutela, A., & Tuomilehto, J. (1994). Cardiovascular risks and socioeconomic status: Differences between men and women in Finland. *Journal of Epidemiology and Community Health 48*, 348-354.

Lynch, J.W., Kaplan, G.A., Cohen, R.D., Tuomilehto, J., & Salonen, J.T. (1996). Do cardiovascular risk factors explain the relation between socioeconomic status, risk of all-cause mortality, cardiovascular mortality, and acute myocardial infarction? *American Journal of Epidemiology 144*(10), 934-942.

Marmot, M. (1989). Socioeconomic determinants of CHD mortality. *International Journal of Epidemiology 18*, 2172-2179.

Mokeddem, S., Atmani, B., & Mokeddem, M. (2013). Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. *Computer Science and Information Technology 2013*, 41-51.

Nittaya, P. & Hataichanok, C. (2015). *World heart day 2015* (pp. 1-6). Ministry of Public Health.

Peter, T.J., & Somasundaram, K. (2012). *An empirical study on prediction of heart disease using classification data mining techniques. In*: Proceeding of IEEE Conference on Advances in Engineering, Science and Management.

Rajeswari, K., Vaithiyanathan, V., & Pede, S.V. (2013). Feature selection for classification in medical data mining. *International Journal of Emerging Trends & Technology in Computer Science 2*(2), 492-497.

Rupali, M.S., & Patil, R. (2014). Heart disease prediction system using naïve bayes and Jelmek-mercer Smoothing. *International Journal of Advanced Research in Computer and Communication Engineering 3*(5), 6787- 6792.

Salonen, J.T. (1982). Socioeconomic status and risk of cancer, cerebral stroke, and death due to coronary heart disease and any disease: A longitudinal study in eastern Finland. *Journal of Epidemiology Community Health 36*, 294-297.

Saravanakumar, S. & Rinesh, S. Effective heart disease prediction using frequent feature selection method. *International Journal of Innovative Research in Computer and Communication Engineering 2*(1), 2767-2774.

Sellapan, P., & Rafiah, A. (2008). Intelligent heart disease prediction system using data mining techniques. *International Journal of Computer Science and Network Security 8*(8), 343-350.

Shantakumar, B.P., & Kumaraswamy, Y.S. (2009). Extraction of significant patterns from heart disease warehouses for heart attack prediction. *International Journal of Computer Science and Network Security 9*(2), 228-235.

Shouman, M., Turner, T., & Stocker, R. (2011). *Using decision tree for diagnosing heart disease patients. In*: Proceedings of ninth Australian Data Mining Conference in Research and Practice in Information Technology.

Siegrist, J., Bernhardt, R., Feng, Z.C., & Schettler, G. (1990). Socioeconomic differences in cardiovascular risk factors in China. *International Journal of Epidemiology 19*, 905-991.

Silverstein, A., Silverstein, V.B. & Nunn, L.S. (2006). *Heart disease*. Twenty-First Century Books.

Simons, L.A., Simons, J., Magnus, P., & Bennett, S.A. (1986). Education level and coronary risk factors in Australians. *Medical Journal of Australia 145*, 446-450.

Sivagowry, S., Durairaj, M., & Persia, A. (2013). *An empirical study on applying data mining techniques for the analysis and prediction of heart disease. In*: International Conference on Information Communication and Embedded Systems.

Smith, G.D., Shipley, M.J., & Rose, G. (1990). Magnitude and causes of socioeconomic differentials in mortality: further evidence from the Whitehall Study. *Journal of Epidemiology and Community Health 44*, 265-270.

Stanford Five-City Project. *Preventive Medicine 21*, 592-601.

Susser, M., Watson, W., & Hopper, K. (1985). *Sociology in medicine*. 3rd eds. New York: Oxford University Press.

Theorell, T., Svensson, J., Knox, S., & Ahlborg, B. (1987). Blood pressure variations across areas in the greater Stockholm region: analysis of 74,000 18-year-old men. *Social Science & Medicine Journal 16*, 469-473.

Tian, H.G., Hu, G., Dong, Q.N., Yang, X.L., Nan, Y., Pietinen, P., & Nissinen, A. (1996). Dietary sodium and potassium, socioeconomic status, and blood pressure in a Chinese population. *Appetite 26*, 235-246.

Vanisree, K., & Singaraju, J. (2011). Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural network. *International Journal of Computer Applications 19*(6), 6-12.

Vathesatogkit, P., Sritara, P., Kimman, M., Hengprasit, B., E-Shyong, T., Wee, H.L., & Woodward, M. (2012). Associations of lifestyle factors, disease history and awareness with health-related quality of life in a Thai Population. *PloS One 7*(11), e49921.

Wing, S., Barnett, E., Casper, M., & Tyroler, H.A. (1992). Geographic and socioeconomic variation in the onset of decline of coronary heart disease mortality in white women. *American Journal of Public Health 82*, 204-209.

Winkleby, M.A., Fortmann, S.P., & Barrett, D.C. (1990). Social class disparities in risk factors for disease: Eight-year prevalence pattern by level of education. *Preventive Medicine 19*, 1-12.

Winkleby, M.A., Fortmann, S.P., & Rockhill, B. (1992.) Trends in cardiovascular disease risk factors by educational level: The

Winkleby, M.A., Jatulis, D.E., Frank, E., & Fortmann, S.P. (1992). Socioeconomic status and health: How education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health 82*, 816-820.

Zhijie,Y., Aulikki, N., Erkki, V., Guide, S., Zeyu, G., Gengwen, Z., Jaakko, T., & Huiguang, T. (2000). Associations between socioeconomic status and cardiovascular risk factors in an urban population in China. *Bulletin of the World Health Organization 78*(11), 1296-1305.