National College of Ireland

Diagnosis of Cardiovascular Diseases using Hybrid Feature Selection and Classification Algorithms

MSc Research Project
Data Analytics

Sandip Mondal
x16133111

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Sandip Mondal
Student ID:	x16133111
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dr. Paul Stynes
Submission Due Date:	11/12/2017
Project Title:	**Diagnosis of Cardiovascular Diseases using Hybrid Feature Selection and Classification Algorithms**
Word Count:	6875

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	9th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Diagnosis of Cardiovascular Diseases using Hybrid Feature Selection and Classification Algorithms

Sandip Mondal
x16133111
MSc Research Project in Data Analytics

9th December 2017

Abstract

Current diagnostic systems in order to identify cardiovascular diseases (CVDs) such as Echocardiography (ECG) require highly skilled physicians to evaluate complex combinations of clinical and pathological data. Inaccurate decision decision making is the challenge in the process and thus can't be permitted in healthcare industry. Data mining methodologies can be applied to large medical datasets to extract insights that aid healthcare professionals in the diagnosis of cardiovascular diseases. In CVDs data mining, classification categorize a patient as having CVDs or free from it based on their similarities to previous examples of other patients. The classification accuracy rate is highly influenced by feature selection technique which eliminates features or attributes with practically no or little information from the dataset. Thus, feature selection and classification algorithms are considered as a concern of global "combinatorial optimization". The aim of this research is to investigate the optimal hybrid model of feature selection and classification algorithms in the diagnosis of cardiovascular diseases based on three performance metrics namely accuracy, sensitivity and specificity. It followed the Cross Industry Standard Process for Data Mining (CRISP-DM). The effect of hybrid feature selection and classification algorithms is examined on heart disease dataset acquired from University of California, Irvine - Machine Learning Repository (UCI-ML). The feature selection algorithm used is Particle Swarm Optimization (PSO). The classification algorithms used are Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree. The hybrid feature selection and classification algorithms are evaluated based on accuracy, sensitivity and specificity with the objective of achieving superior predictive performance. Results demonstrated that hybrid combination of PSO with SVM (PSO_SVM) achieves superior predictive performance over other models. The research will thus empower physicians to diagnose cardiovascular diseases and initiate timely treatment without the intervention of a trained cardiologist.

Keywords: *Cardiovascular Diseases (CVDs), Data Mining, Feature Selection, Particle Swarm Optimization (PSO), Classification Algorithm, Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest, C5.0 Decision Tree*

1 Introduction

Cardiovascular Diseases (CVDs) has been identified as the major cause of death globally in the past decades (*World Health Organization*; May, 2017; *European Public Health Alliance*; February, 2011; *ESCAP*; 2010; *Australian Bureau of Statistics*; 2014; *South Africa Statistics*; 2013). According to Paladugu (2010), correct and timely diagnosis is desirable for its cure. The widely recognized clinical methods for diagnosing CVDs such as Electrocardiogram (EKG) and Echocardiography (ECG) are complex bearing in mind the number of factors that the physician has to evaluate (*The National Institute of Health of U.S. Department of Health and Human Services*; 2017). Hence, diagnosing CVDs demands experienced and highly skilled physicians (Das et al.; 2009). According to Ming Hsu and Camerer (2005) and Jennifer S. Lerner and Kassam (2014), the action of making vital decisions by humans is optimal, however it is poor when the amount of data to be classified is vast. Poor and inaccurate decision making due to extensive stress and over workload can't be permitted in healthcare industry. Medical diagnosis is thus an essential yet complex task that should be executed accurately and effectively (Hideg and Kleef; 2013). So, the automation of this process is desirable. Motivated by the world-wide growing mortality of CVDs patients every year and the accessibility to huge amount of patients' data from which to fetch helpful knowledge, researchers are using data mining methods for assisting healthcare professionals in the diagnosis of CVDs (Princy and Thomas; 2016; Kalaiselvi; 2016; Masethe and Masethe; 2014).

Data mining techniques combines statistical analysis, database technologies and machine learning to extract undiscovered data and relationships from large datasets (Mai Shouman and Stocker; 2012) which are difficult to comprehend with conventional statistics (Jabbar; 2012; B.N.Lakshmi and G.H.Raghunandhan; 2011). In CVDs data mining, classification categorize a patient as having CVDs or free from it based on their similarities to previous examples of other patients. So, classification involves dividing up group of instances so that each is assigned to one of many mutually exhaustive and exclusive categories known as classes. The term "mutually exhaustive and exclusive" means that each object must be assigned to precisely one class, i.e. a patient can be either having or free from CVDs (Bramer; 2013; B.N.Lakshmi and G.H.Raghunandhan; 2011). The term classifier refers to the function, implemented by a classification algorithm that maps input data to a class and enable predictions based on historical data (Kalaiselvi; 2016). Few examples of classifiers are Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Trees. The classifier performance is highly influenced by feature selection algorithms, a procedure of discarding features or attributes with practically no or little information. Features such as blood sugar may not be a risk factor and is of little information in diagnosing CVDs. In the event of large number of features, the dataset size will be large and not clean, therefore the classifier performance would be adversely affected (Parthiban and R.Subramanian; 2007; Elbedwehy; 2012). Few examples of feature selection algorithms are Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Artificial Bee Colony (ABC). Combination of feature selection and classification algorithms are considered as a concern of global "combinatorial optimization" thereby enhancing accuracy, sensitivity and specificity (Liu and Yu; 2005). Accuracy is the measure of the proportion of instances that are correctly classified. Sensitivity denotes the proportion of positive instances that are correctly classified as positive i.e. what amount correct instances are predicted from the all the instances and specificity is a measure of the the proportion of negative instances are that correctly classified as negative (Shwartz and David; 2014).

The aim of this research is to investigate the optimal hybrid model of feature selection and classification algorithms in the diagnosis of cardiovascular diseases based on three performance metrics namely accuracy, sensitivity and specificity. The innovation of this research is in the combination of feature selection and classification algorithms that has not been experimented in diagnosing CVDs. However, this research is limited to the use of heart disease dataset acquired from the Machine Learning Repository of University of California, Irvine (UCI-ML) given that it is available to be used by the machine learning community for the empirical analysis of machine learning algorithms (*Machine Learning Repository of University of California, Irvine*; 2017). Thus, the goal of this research is two-fold:

i To examine if feature selection algorithms could be utilized effectively to avoid the curse of dimensionality in the diagnosis of cardiovascular diseases.

ii To examine if hybrid feature selection and classification algorithms can be utilized to improve the accuracy, sensitivity and specificity in diagnosing cardiovascular diseases.

This paper is organized as follows. Section 2 presents a literature review on diagnosis of cardiovascular diseases and hybrid implementations of data mining algorithms wherein we compare previous works and specify the foreseen contributions. This section also documents a comparative study of data mining tools. Section 3 describes the research methodology. The implementation is illustrated in section 4. Section 5 outlines the evaluation and discussions on the results. Finally, we conclude this research in section 6.

2 Literature Review

The following literature review presents a critical analysis (Webster and Watson; 2002) on data mining algorithms and its use in the diagnosis of cardiovascular diseases in Section 2.1, on data mining tools in Section 2.2 and finally, concludes with a summary of the insights gained which forms the base of our research in Section 2.3.

2.1 Diagnosis of Cardiovascular Diseases Using Data Mining Techniques

This section begins with an exploration on cardiovascular diseases (CVDs). Next, a review on data mining and feature selection algorithms is presented followed by classification algorithms.

Cardiovascular Diseases (CVDs) are a class of ailments that involve the heart. It occurs when indiscretion exists in the stream of blood or bruise heart muscles as a consequence of insufficient supply of oxygen (A Sudha and Jaishankar; 2012). According to Heller (2008), the most common types of CVDs are Congenital Heart Disease, Congestive Heart Failure and Coronary Heart Disease. Symptoms like chest pain, heart palpitations, fatigue, lethargy or daytime sleepiness are likely an indication of CVDs (Fogoros; 2009). The

clinical techniques for diagnosing CVDs such as Electrocardiogram (EKG) and Echocardiography (ECG) are complex bearing in mind the number of factors that the physician has to evaluate (*The National Institute of Health of U.S. Department of Health and Human Services*; 2017).

As opposed to complex clinical investigations, an automated data mining process aid medical practitioners make near-perfect diagnosis easily (Mai Shouman and Stocker; 2012; Princy and Thomas; 2016). Data mining is the investigation of huge datasets to extract concealed and hidden patterns, connections and information which are hard to understand with conventional statistics (Jabbar; 2012; Huan Liu and Zhao; 2010; Sivagowry and Durairaj; 2014). Data mining in healthcare is therefore a field of high significance for forecasting and a more profound understanding of medical data (Ruben; 2009). In data mining, feature selection algorithms is a procedure of discarding features or attributes with practically no or little information. Features such as blood sugar may not be a risk factor and is of little information in diagnosing CVDs. In the event of large number of features, the dataset size will be large and not clean which will adversely affect the classifier performance would be adversely affected (Parthiban and R.Subramanian; 2007; Elbedwehy; 2012). Few examples of feature selection algorithms are Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and Artificial Bee Colony (ABC).

Paul et al. (2016) studied Genetic Algorithm (GA) feature selection mechanism in diagnosing CVDs. GA imitates the mechanism of natural selection and simulate the survival of the fittest among individuals over consecutive generations for solving a problem. In GAs, there exist a pool or a population of possible solutions for the given problem. These solutions then go through recombination and mutation (simillar to natural genetics), creating new children, and the procedure is repeated over various generations. Every individual (or candidate solution) is alloted a fitness value (based on how well it answer the given problem) and the fitter individuals are given a higher chance to mate and yield more "fitter" individuals or solutions. The study demonstrated the efficiency of GA in selecting the optimal attributes in a dataset. Another feature selection algorithm namely Particle Swarm Optimization (PSO) was examined by (Alex C Alberto and Nadal; 2012). The author defined PSO, as a mechanism based on the principle similar to bird flocking in an uniform regular fashion in search of food. Initially, they are not known where the food is and therefore move in a random fashion in search of food in an area. Each bird is attracted to some degree to the best location it has found so far, and also to the best location any member of the population has found. After some time the population can coalesce around one location or can continue to move. PSO uses the same scenario to solve optimization problems, where birds are represented as particle and food is represented as the solution (I. Nahar and Chen; 2013). This is expected to move the swarm toward the best solutions for eliminating attributes with practically no or little information. This study also demonstrated the efficiency of PSO in selecting the optimal attributes in a dataset. Bharti and Singh (2015) compared GA and PSO in detecting CVDs. The paper demonstrated that PSO has the same effectiveness in finding the true global optimal solution as GA, however PSO has significantly better computational efficiency because it implements statistical analysis and formal hypothesis testing. A investigation by MN Ab Wahab and Atyabi (2015) and Benxian Yue and Liu (2007) affirmed PSO as more computationally effective than GA however, it would had been more interesting if their performances were evaluated with real-life experiments. These insights therefore motivated the use of PSO in this research as the feature selection mechanism.

Though feature selection methodologies are important to eliminate redundant and unimportant features, the accuracy of the subset generated is more important. Subsequently, it is essential to validate it by utilizing classification methodologies and learn the underlying procedures in Bio-Informatics (Parthiban and R.Subramanian; 2007; Elbedwehy; 2012; Liu and Yu; 2005; Wang; 2007). Bhatia (2013) executed a hybrid approach of GA and classification algorithms for detecting Heart Diseases and got 79.62% accuracy. The study affirmed that the hybrid approach using multiple data mining algorithms had better performance than used individually in healthcare applications. Similar studies by Mokeddem (2013) and Chitra and Seenivasagam (2013) too demonstrated that hybrid models are more accurate and is an approving outcome in the literature. This research therefore considers hybrid approach involving feature selection and classification algorithms as its base for diagnosing CVDs.

In data mining, classification algorithms assigns instances in a dataset to target categories or classes. The objective of classification is to accurately predict the target category for each instance in the dataset. For example, a classification model could be utilized if a patient is free or having CVDs (Bramer; 2013; B.N.Lakshmi and G.H.Raghunandhan; 2011). The term classifier refers to the function, implemented by a classification algorithm that maps input data to a category and enable predictions based on historical data (Kalaiselvi; 2016). Few examples of such algorithms are Artificial Neural Network (ANN), Support Vector Machines (SVM), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Trees.

Niwas et al. (2005) examined the performance of Artificial Neural Network (ANN) classifier in detecting CVDs from ECG signals. ANNs are designed according to the biological neural networks. It endeavors data mining algorithms to learn and solve problems in a way that a human brain would. The perceptron is the

unit of neural networks and comprises of one or more inputs, a processor (sum and activation function) and a single output (Gupta and Bajaj; 2014) as seen in Figure 1. Therefore, a perceptron follows a "feed-forward" model, meaning the inputs are sent to a neuron, processed and result in output (Acharya; 2017; Aditya; 1991). Results demonstrated an accuracy of 99.02%. The investigation however did not considered clinical information, for example, cholesterol, blood pressure and age. A similar experiment for diagnosing CVDs was conducted by Alty et al. (2003) using Support Vector Machines (SVM) by measuring pulse volume at the finger tip. SVMs classifies both linear and non-linear data by means of a mapping function. This mapping function divides instances into different classes by using a hyper-plane. For instance in Figure 2, hyper-plane x1-x2 divides instances into two classes represented by squares and circles. According to the paper, SVM classifier yields a classification accuracy of >85%. However, this investigation too did not considered clinical data and focused on physical examination. A comparative study of SVM and ANN on CVDs clinical dataset was conducted by Kumari and Godara (2011). Performance of these algorithms were evaluated based on accuracy, sensitivity and specificity. Results demonstrated SVM to be better classifier than ANN in diagnosing CVDs. However, it did not examine the performance of both ANN and SVM in combination with PSO. This research therefore considers to investigate the same.

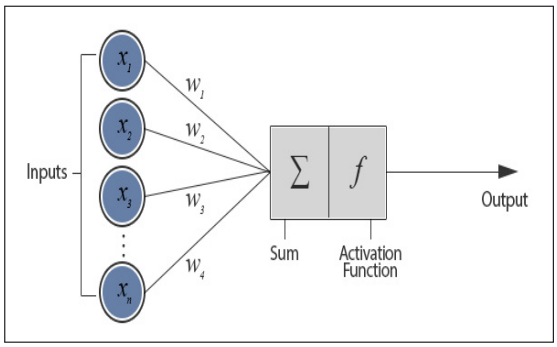

Figure 1: Artificial Neural Network Perceptron

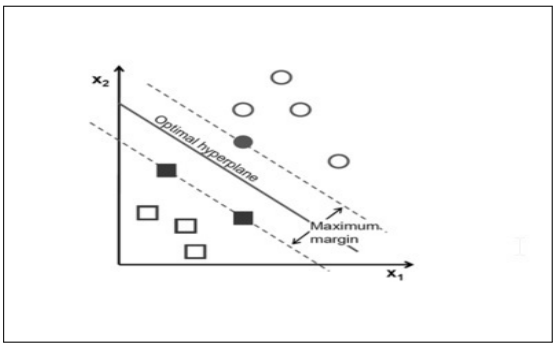

Figure 2: Support Vector Machines (SVM)

Mai Shouman and Stocker (2012) investigated K-Nearest-Neighbour (KNN) classification algorithm to aid healthcare professionals in the diagnosis of CVDs. According to the author, KNN is one of the successful data mining techniques used in classification problems. The paper investigated if KNN can enhance the accuracy in the diagnosis of CVD patients when applied on benchmark dataset. The results demonstrated that applying KNN could achieve higher accuracy than neural networks. However, the application of the algorithms on benchmark dataset rather than real-life examples was a constraint in the study. This research therefore, examines the performance of KNN in combination with PSO on real-life examples.

Elnaz Pashaei and Aydin (2015), studied the performance of C5.0 Decision Tree classifier in improving medical diagnosis reliability. C5.0 is an improved version of decision tree algorithm presented by Ross Quinlan (Kuhn and Johnson; 2013). A decision tree is a decision support algorithm that uses a tree-like graph or model of decisions and their possible consequences as seen in Figure 3. The experiment was implemented on micro-array cancer dataset from UCI-ML repository. Results demonstrated that C5.0 has relatively lower error rates compared to other decision tree classifiers such as C4.5 and CART. Consequently it is more accurate and considerably faster. This research plans to examine the performance of C5.0 decision tree classifier in combination with PSO on heart disease dataset as opposed to micro-array cancer dataset.

Shan Xu and Zhu (2017) studied Random Forest in predicting cardiovascular diseases. Random Forest classifier comprises of a collection of decision trees as seen in Figure 4. The difference between C5.0 and Random Forest lays in the fact that C5.0 is a single tree whereas a random forest is an ensemble of decision trees. An individual tree is constructed by applying an algorithm A on the training set S and an additional random vector, θ where θ is sampled independently and identically from some distribution (Shwartz and David; 2014). The paper illustrated how data pre-processing plays a critical role and is an obligatory step in data mining process. The presence of missing values could degrade the classifier performance and are either discarded or supplanted with an approximate value using data imputation methodologies. Results demonstrated an accuracy of 91% in detecting CVDs. In this paper feature selection was performed using Genetic Algorithm (GA). This research therefore considers to investigate the performance of Random Forest classifier in combination with PSO feature selection methodology.

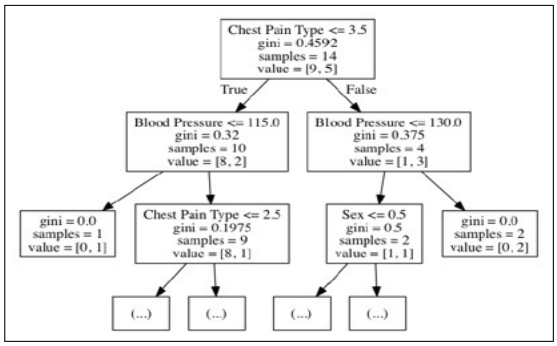

Figure 3: C5.0 Decision Tree

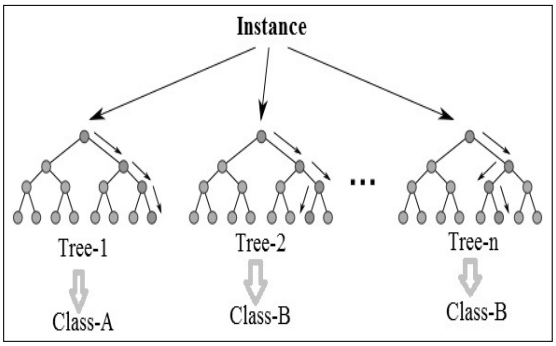

Figure 4: Random Forest

In addition, investigations by G. Subbalakshmi and Rao (2011) and K.Srinivas and A.Govrdhan (2010) demonstrated Naïve Bayes as an effective classifier for diagnosing CVDs as well. Naïve Bayes is based on Bayes' Theorem and is applied in automated medical diagnosis (Rish; 2001). Naïve Bayes is thus a probabilistic method for constructing classifiers. An advantage of Naïve Bayes is that it only requires a small amount of data for classification. Similar to the reviewed works on ANN and SVM, both the studies it did not examine the performance of Naïve Bayes in combination with PSO. This research therefore considers to investigate the same.

Majority of the published scholar and research papers based on CVDs diagnosis uses UCI-ML heart disease dataset for its investigation. It has 76 attributes in total. However, the papers referred to using a subset of 14 attributes. Few of such papers are mentioned in Table 1.

AUTHOR	TECHNIQUE	NO. OF FEATURES	ACCURACY
Kumari and Godara (2011)	Support Vector Machines (SVM)	14	84.12%
Abdullah and Rajalaxmi. (2012)	Decision Tree	14	63.33%
J.Soni (2011)	Association Rules	14	81.51%
K.Srinivas and A.Govrdhan (2010)	K-NN	14	45.67%
	Decision Tree	14	52%
	Naive Bayes	14	84.14%
Rajkumar and Reena (2010)	Decision Tree	14	64.40%
	Naive Bayes	14	52.33%

Table 1: Data Mining Algorithms Experimented on 14 attributes of UCI-ML Heart Disease Dataset

Data mining evaluation methodologies are business applications dependent, wherein one techniques may stand great over another in light of the application assessed. Consequently, the performance metrics considered in this research was adjusted to cardiovascular diseases diagnosis. According to the contributions from literature (Aravinthan and Vanitha; 2016; K.Sudhakar and Manimekalai; 2014; Parthiban and R.Subramanian; 2007; Elbedwehy; 2012; Rohilla and Gulia; 2015; Yuanning Liu and Wang; 2011), the performance of classifiers were mainly evaluated based on accuracy, sensitivity and specificity. Therefore this research uses the same for evaluating each of the models actualized.

2.2 Data Mining Tools

Several papers (Marjia Sultana and ShorifUddin; 2016; Mohammed Abdul Khaled and Dash; 2013; D. P. Shukla and Sen; 2014) has reviewed algorithms and methods used for data mining in general but comparatively not much work has been done on the data mining tools for practitioners (Wimmer and Powell; 2015). X. Chen and Williams (2007) compared 12 open source tools against several aspects such as "general characteristics, data source accessibility, data mining functionality, and usability". Auza (2010) reviewed open source data mining tools such as Rapid Miner, Weka, Orange and R.

Rapid Miner, formerly Yale, has transformed into a licensed software product rather than open source; nevertheless, Rapid Miner Community Edition continues to be free and open source. Rapid Miner can perform "process control, connect to a repository, import and export data, data transformation, modeling and evaluation". While the open source version has many features, certain functionalities are disabled. One such instance is data sources. CSV and MS Excel are only supported by the open source version and has

no access to databases (*The Rapid Miner Platform*; 2017). WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is an open source software program that was developed on the Java platform (Wimmer and Powell; 2015). It supports a "collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions" (Ian H. Witten and Hall; 2011). R is a free and open source package for data mining, statistical analysis and graphing. R is traditionally a Command Line Interface (CLI); however, there are numerous freely available open source tools that can be integrated into R. One such tool, is R-Studio which provides a Graphical User Interface (GUI) for R. R can be utilized for a wide range of statistical and analytics tasks including but not limited to "clustering, regression, time series analysis, text mining, and statistical modeling" (*The R Project for Statistical Computing*; 2017). Table 2 compares Rapid Miner, WEKA and R in terms of the data mining methodologies supported (Wimmer and Powell; 2015):

ALGORITHM	**R**	**RAPID MINER**	**WEKA**
K-Nearest Neighbour	✓	✓	✓
Association Rule Mining	✓	✓	✓
Neural Networks	✓	✓	✓
Decision Tree	✓	✓	✓
Support Vector Machines	✓	✓	✓
Feature Selection	✓	Few	✓
Time Series Analysis	✓	Few	✓
Big Data Processing	✓	Few	✓

Table 2: Comparison of Data Mining Tools

As seen from Table 2 above, WEKA and R outperforms Rapid Miner. According to PEHLİVANLI (2011), though R performs better than WEKA in terms of interactivity, mechanism for supporting data structures, graphics for visualization and analysis and mechanism to handle missing values, R has poor feature selection implementation support compared to WEKA. Therefore, this research uses R as it's analysis tool for conducting experiments except for feature selection which has been implemented in WEKA.

2.3 Summary

After reviewing the literature, the first task is to study and verify the different hypotheses associated with diagnosing CVDs. There are so many articles in the review that do not recommend any single classifier for getting better results. Despite the fact that there are some certain circumstances where a couple of the classifiers give comparatively better results, these cannot be a benchmark in all scenarios such as an algorithm used alone or in a hybrid mode. As an abstract representation,the main contributions of the review reflected in this research are as follows:

- Data mining algorithms in hybrid mode outperforms algorithms when used individually in diagnosing cardiovascular diseases. Therefore, classifiers such as Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree are hypothesized to perform better when combined with Particle Swarm Optimization (PSO) feature selection technique.
- UCI-ML heart disease dataset has 76 features in total. However, majority of the published scholar and research papers refers to using a subset of 14 attributes.
- Accuracy, sensitivity and specificity are the most common performance metrics when evaluating models for diagnosing cardiovascular diseases.

3 Methodology

This research examines the empirical relationship between a set of features such as age, serum cholesterol and blood sugar with the probability of being diagnosed with cardiovascular diseases (CVDs). Researcher Yin (2013) defines "a case study as an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used". The methodology begins by defining the scope and purpose of this research as discussed in subsection 3.1. Data extraction, its exploratory analysis and descriptive statistics is portrayed in subsection 3.2. Subsection 3.3 explains the pre-experimental setup and subsection

3.4 briefly describes the experimentation conducted. Finally, the methodology for interpreting the results is discussed in subsection 3.5.

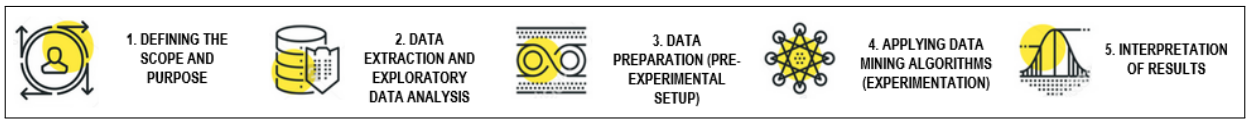

Figure 5: Methodology of the Research

3.1 Defining Scope and Purpose

This step focuses on comprehending the objectives and requirements from a business perspective and converting this knowledge into a data mining problem definition. A preliminary plan for achieving the defined objectives is also outlined (Awang and Palaniappan; 2008). Motivated by the world-wide growing mortality of CVDs patients and complexity in its clinical diagnosis, the business objectives of this research is to automate the diagnosis of CVDs by means of data mining techniques. The preliminary plan is to use an available heart disease dataset and compare different hybrid feature selection and classification algorithms based on different performance metrics for finding an optimal model.

3.2 Data Extraction and Exploratory Data Analysis

In the second step, Cleveland heart disease dataset is extracted from UCI-ML repository (*Machine Learning Repository of University of California, Irvine*; 2017) to comprehend the effect of hybrid feature selection and classification algorithms.

Out of the 76 attributes in the actual dataset only 14 attributes was used for this study as per the insights gained from literature review and is described in table 3.

Table 3: Attribute Description of UCI-ML Heart Disease Dataset (*Attribute Description: UCI-ML*; 2017)

SERIAL #	ATTRIBUTE	DESCRIPTION
1	Age	Age in Years
2	Sex	*1=Male, 0=Female*
3	Cp	Chest pain type (*1=Typical Angina, 2=Atypical Angina, 3=Non-Anginal Pain, 4=Asymptomatic*)
4	Trestbps	Resting blood sugar(in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (*1=true, 0=false*)
7	Restecg	Resting Electrocardiographic results (*0=Normal, 1=ST-T wave abnormality, 2=Left Ventricularhypertrophy*)
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The Slope of the peak exercise ST segment (*1=Upsloping, 2=Flat, 3=Downsloping*)
12	Ca	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	The heart status (*3=Normal, 6=Fixed defect, 7=Reversible defect*)
14	Num	Class Label (*0=No Disease, 1=Disease*)

Next, exploratory data analysis (EDA) is performed on the extracted dataset to understand the data better such as class imbalance, features included, correlation between features, missing values and outliers (Martinez; 2010). The concern of class imbalance is examined to assure that there exists near about equal number of 'Disease' and 'No-Disease' patients in the extracted dataset for optimal experimentation. EDA is performed by means of mosaic plots, scatter matrix and correlation matrix.

3.3 Data Preparation (Pre-Experimental Setup)

In the third step, pre-experimental setup is laid down as per the insights gained from step 3.2. Missing data is supplanted with an approximate value using mode data imputation methodology for categorical attributes and mean for numerical attributes (Sivagowry and Durairaj; 2014). Further, normalization is

used to re-scale the numerical data values. It scales the data values between 0 and 1 thus enabling the same range of values for each of the inputs (Patro and Sahu; 2015). Thereafter, the dataset is separated into larger portion for training and a smaller portion of testing the hybrid model in 70% - 30% ratio respectively (Marjia Sultana and ShorifUddin; 2016).

3.4 Applying Data Mining Algorithms (Experimentation)

In the fourth step, the experimentation is carried out. In this research, 6 experiments are conducted. Each of the experiments conform to 6 different hybrid models of feature selection and classification algorithm. The feature selection algorithm used is Particle Swarm Optimization (PSO). The classification algorithms used are Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree.

- **Experiment #1 (PSO_SVM):** The first experiment is designed to comprehend the performance of hybrid Support Vector Machines (SVM) and Particle Swarm Optimization (PSO) in diagnosing CVDs. Therefore, the classifier used in experiment 1 is SVM.
- **Experiment #2 (PSO_ANN):** The second experiment is designed to comprehend the performance of hybrid Artificial Neural Networks (ANN) and Particle Swarm Optimization (PSO) in diagnosing CVDs. Therefore, the classifier used in experiment 2 is ANN.
- **Experiment #3 (PSO_NB):** The third experiment is designed to comprehend the performance of hybrid Naïve Bayes and Particle Swarm Optimization (PSO) in diagnosing CVDs. Therefore, the classifier used in experiment 3 is Naïve Bayes.
- **Experiment #4 (PSO_KNN):** The fourth experiment is designed to comprehend the performance of hybrid K-Nearest Neighbour (KNN) and Particle Swarm Optimization (PSO) in diagnosing CVDs. Hence, the classifier used in experiment 4 is KNN.
- **Experiment #5 (PSO_RF):** The fifth experiment is designed to comprehend the performance of hybrid Random Forest and Particle Swarm Optimization (PSO) in diagnosing CVDs. Therefore, the classifier used in experiment 5 is Random Forest.
- **Experiment #6 (PSO_C5.0):** Finally, the sixth experiment is designed to comprehend the performance of hybrid C5.0 Decision Tree and Particle Swarm Optimization (PSO) in diagnosing CVDs. Hence, the classifier used in experiment 6 is C5.0 Decision Tree.

STEP #	MODEL PSEUDO CODE
	Start Feature Selection
1	Generate initial particles and setup the PSO parameters; Set iteration = 0 and execute the subset selection process from step 2–4.
2	Set iteration counter = iteration counter + 1.
3	Evaluate the fitness of each particle; Manipulate the particles' movement on to the next position.
4	If the stopping condition is satisfied (PSO parameter defined in step 1), go to step 2, else, go to next step.
	End Feature Selection
	Start Classification
5	Train the classifier (SVM, ANN, Naïve Bayes, KNN, Random Forest and C5.0 Decision Tree) by the features selected by PSO.
6	Calculate accuracy, sensitivity and specificity of all the models.
7	Interpret results and compare models for superior predictive performance over each other.
	End Classification

Table 4: Pseudo Code of the Cardiovascular Diseases Diagnostic System

3.5 Interpretation of Results

In the fifth step, the impact of all the hybrid and individual models are evaluated based on accuracy, sensitivity and specificity. Results are interpreted and models are compared for superior predictive performance

over each other. Finally, in the sixth step conclusions are derived based on the results obtained. It was concluded as to which model actually has better performance.

The methodology therefore aids in constructing the cardiovascular diagnostic systems' architectural blue-print for carrying out research and development of the systematic process as illustrated in the Design section.

4 Implementation

In this section, the implementation of this research has been illustrated. This research is a quantitative case study and follows Cross Industry Standard Process for Data Mining (CRISP-DM). The motivation driving utilizing CRISP-DM as the research process is its stages which could be duly structured, organized and characterized, enabling a task to be effectively comprehended or updated as opposed to other processes such as SEMMA and KDD (Daniel; 2005). Data extraction, its exploratory analysis and descriptive statistics has been described in subsection 4.1, subsection 4.2 explains the pre-experimental setup and finally, subsection 4.3 briefly describes the experimentation conducted. Figure 6, portrays the implementation of the hybrid feature selection and classification algorithms pictorially.

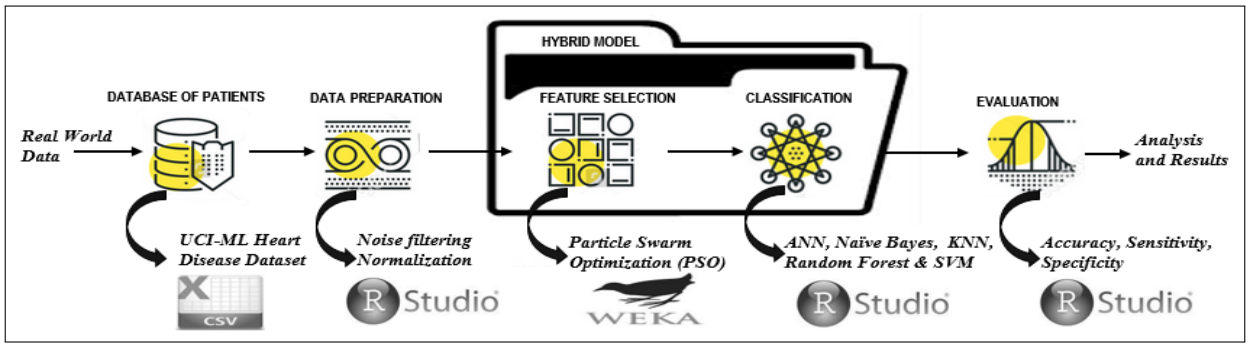

Figure 6: Implementation of the Cardiovascular Diseases Diagnostic System

4.1 Data Extraction and Exploratory Data Analysis

The effect of hybrid feature selection and classification algorithms has been experimented on heart disease dataset acquired from University of California, Irvine - Machine Learning Repository (UCI-ML). Data from Cleveland dataset concerning heart disease diagnosis was extracted using R-Studio as seen in Figure 7. This research has followed the rules laid out in the UCI - Code of Ethics document (*Code Of Ethics of University of California, Irvine*; 2017).

```
# Read Dataset : UCI-ML Heart Disease Dataset
cardiovascular.disease.data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data",header=FALSE,sep=",",
na.strings = '?')
# Assign Column Names
names(cardiovascular.disease.data) <- c( "age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal", "num")
```

Figure 7: Extraction of UCI-ML Heart Disease Dataset via R-Studio

A random sample of the extracted dataset has been illustrated in Figure 8.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	1

Figure 8: Sample Data from UCI-ML Heart Disease Dataset

Post extraction, exploratory data analysis (EDA) and descriptive statistics was carried out to understand the data better such as class imbalance, attributes included, correlation between attributes, missing values and outliers. Two cases were investigated for comprehending class imbalance: #1: How many have

cardiovascular diseases ('Disease vs 'No-Disease')? (see Figure 13) #2: What is the distribution of cardiovascular diseases by gender (Males vs Females)? (see Figure 14). Both the Figures depicts a near equal number of instances in each of the categories thereby nullifying class imbalance in the dataset. An analysis also revealed the presence of 6 missing values in the dataset.

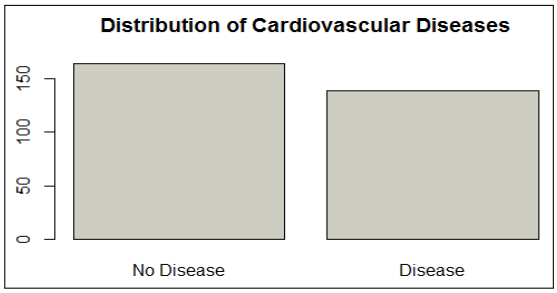

Figure 9: Distribution of CVDs

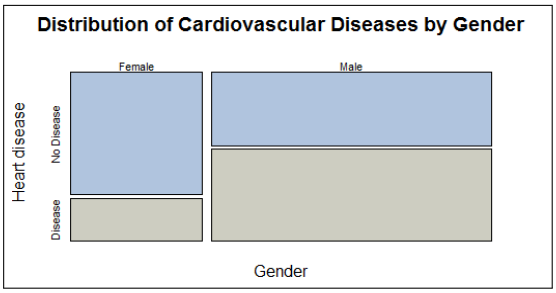

Figure 10: Distribution of CVDs by Gender

An investigation on correlation between the numerical attributes namely age, trestbps (resting blood sugar in mm Hg on admission to hospital), chol (serum cholesterol in mg/dl) and thalach (maximum heart rate) (*Attribute Description: UCI-ML*; 2017) has been executed. As seen in Figure 11, a positive correlation exists between age, trestbps and chol i.e. as one increases the other increases as well and vice-versa whereas, thalach exhibited a negative correlation i.e. as age increases maximum heart rate decreases and vice-versa.

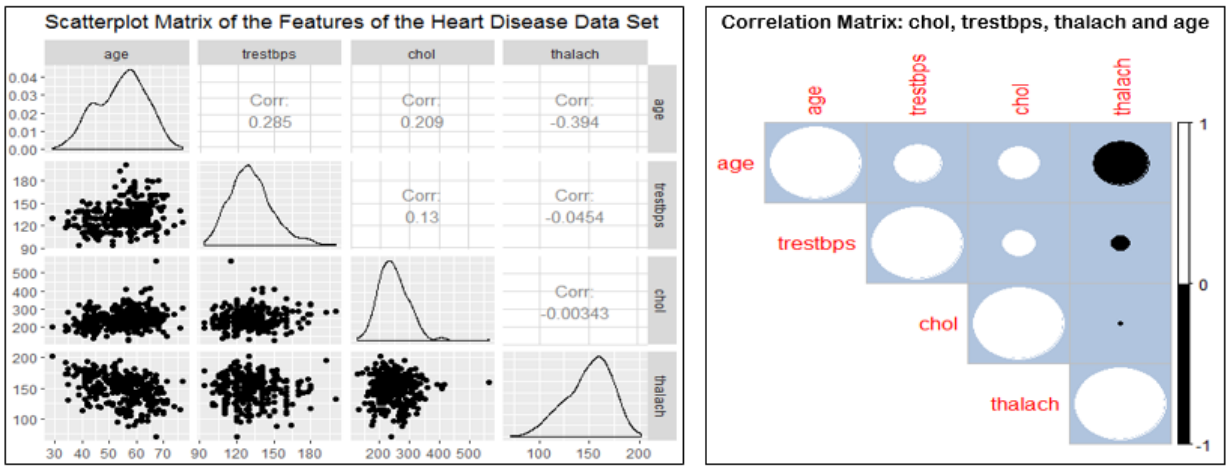

Figure 11: A Study of Correlation in UCI-ML Heart Disease Dataset

An interpretation of descriptive statistics (see Figure 12), conveyed the mean, median and mode of the attributes being different. Therefore, the distribution was considered to be asymmetrical wherein data tends towards the higher or lower range of the dataset.

```
> summary.data.frame(cvd)
      age             trestbps          chol            thalach
 Min.   :29.00   Min.   : 1.00   Min.   :126.0   Min.   : 71.0
 1st Qu.:48.00   1st Qu.:15.00   1st Qu.:211.0   1st Qu.:133.5
 Median :56.00   Median :23.00   Median :241.0   Median :153.0
 Mean   :54.44   Mean   :22.99   Mean   :246.7   Mean   :149.6
 3rd Qu.:61.00   3rd Qu.:29.00   3rd Qu.:275.0   3rd Qu.:166.0
 Max.   :77.00   Max.   :50.00   Max.   :564.0   Max.   :202.0
```

Figure 12: Descriptive Statistics of UCI-ML Heart Disease Dataset

Further exploration revealed outliers in the dataset. Patients diagnosed with CVDs primarily has these outliers, meaning their cholesterol and blood pressure values were above the normal scale. In real clinical diagnosis it would not had been considered as outliers or irrelevant data. Therefore this research sensibly considers to keep these outliers as they are in the dataset.

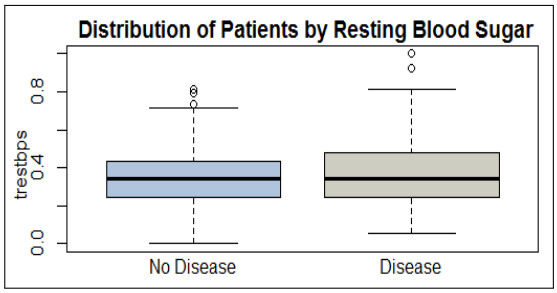

Figure 13: Distribution of Patients by Blood Sugar Amount in Blood Sample

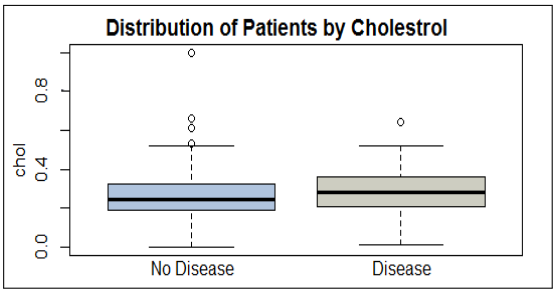

Figure 14: Distribution of Patients by Cholesterol Amount in Blood Sample

These insights gained from the exploratory data analysis and descriptive statistics facilitated in data preparation or preprocessing.

4.2 Data Preparation (Pre-Experimental Setup)

The Cleveland dataset used for this research has 303 rows, 5 numerical and 9 categorical attributes. The presence of missing values can degrade the classifier performance and henceforth the instances with missing data is supplanted with an approximate value using data imputation methodologies (Sivagowry and Durairaj; 2014). The dataset has 6 missing values and asymmetrical distribution as per the knowledge gained from section 4.1. Since the missing values are from categorical attributes it was supplanted with mode value of the respective columns. Normalization is used to answer asymmetrical distribution. It scales the data values between 0 and 1 enabling the same range of values for each of the inputs. Data type conversions is executed to empower the model comprehend the character of the data for each attribute. For example, age can be any number but sex can be either male or female. In this way, age must be of numeric data type whereas, sex must be of type factor. As seen in Figure 15, R-Studio has been used for necessary data type conversions.

```
# Data Type Conversion
cardiovascular.disease.data$age = as.numeric(cardiovascular.disease.data$age)
cardiovascular.disease.data$sex=as.factor(cardiovascular.disease.data$sex)
cardiovascular.disease.data$cp=as.factor(cardiovascular.disease.data$cp)
cardiovascular.disease.data$trestbps=as.factor(cardiovascular.disease.data$trestbps)
cardiovascular.disease.data$chol=as.numeric(cardiovascular.disease.data$chol)
cardiovascular.disease.data$fbs=as.factor(cardiovascular.disease.data$fbs)
cardiovascular.disease.data$restecg=as.factor(cardiovascular.disease.data$restecg)
cardiovascular.disease.data$thalach=as.numeric(cardiovascular.disease.data$thalach)
cardiovascular.disease.data$exang=as.factor(cardiovascular.disease.data$exang)
```

Figure 15: Data Type Conversion

Following Marjia Sultana and ShorifUddin (2016), the dataset is divided into a larger part training data (70%) and a smaller portion testing data (30%). This separation also retained balanced number of 'Disease' or 'No-Disease' categories in both training and testing datasets for optimal examination. 'Caret' library in R-Studio has been used for this separation as seen in Figure 16.

```
# Splitting Training and Testing Data
library(caret)
set.seed(10)
inTrainRows <- createDataPartition(cardiovascular.disease.data$num,p=0.7,list=FALSE)
trainingCardiovascularData <- cardiovascular.disease.data[inTrainRows,]
testingCardiovascularData <-  cardiovascular.disease.data[-inTrainRows,]
```

Figure 16: Division of Cleveland Data set into Training and Testing Data

4.3 Applying Data Mining Algorithms (Experimentation)

Each of the six classifiers has been combined independently with a feature selection algorithm forming a hybrid model for diagnosing CVDs. Therefore, the unified perspective of the hybrid model includes two stages namely Feature Selection (Phase 1) and Model Classifier Learning (Phase 2) as seen in Figure 17.

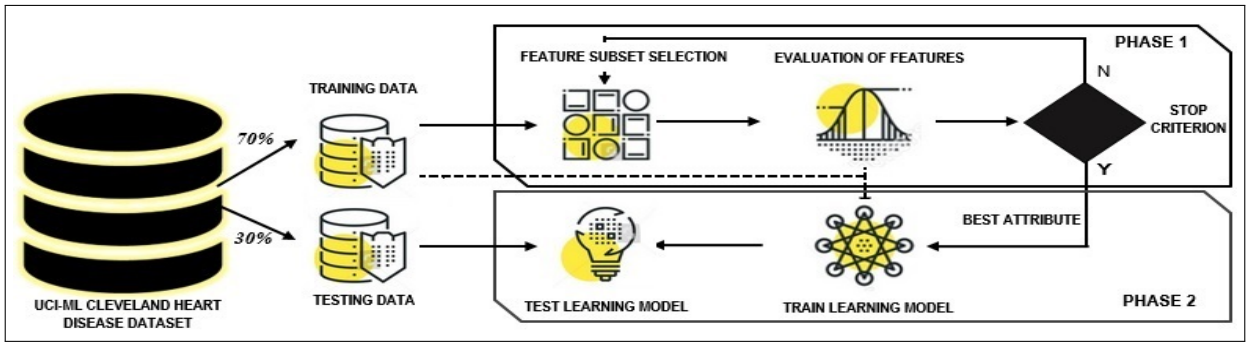

Figure 17: Unified View of the Cardiovascular Diseases Diagnostic Model

Feature selection is further sub-partitioned into three stages and the candidate set obtained thereafter was utilized for filtering the data for training the model:

i Generation of candidate set containing subset of original variables utilizing a research strategy,

ii Evaluation of the candidate set and estimation of utility of variables in the candidate set and,

iii Determination of the predictive ability of the present set of selected variables.

For each hybrid model actualized, prepared data goes to the classifiers via the feature selection methodology. Particle Swarm Optimization (PSO) feature selection methodology is executed using WEKA tool. PSO algorithm condenses the number of attributes in the dataset to 7 from 14 as seen in Figure 18. The selected attributes are namely sex, cp, thalach, exang, oldpeak, ca and thal.

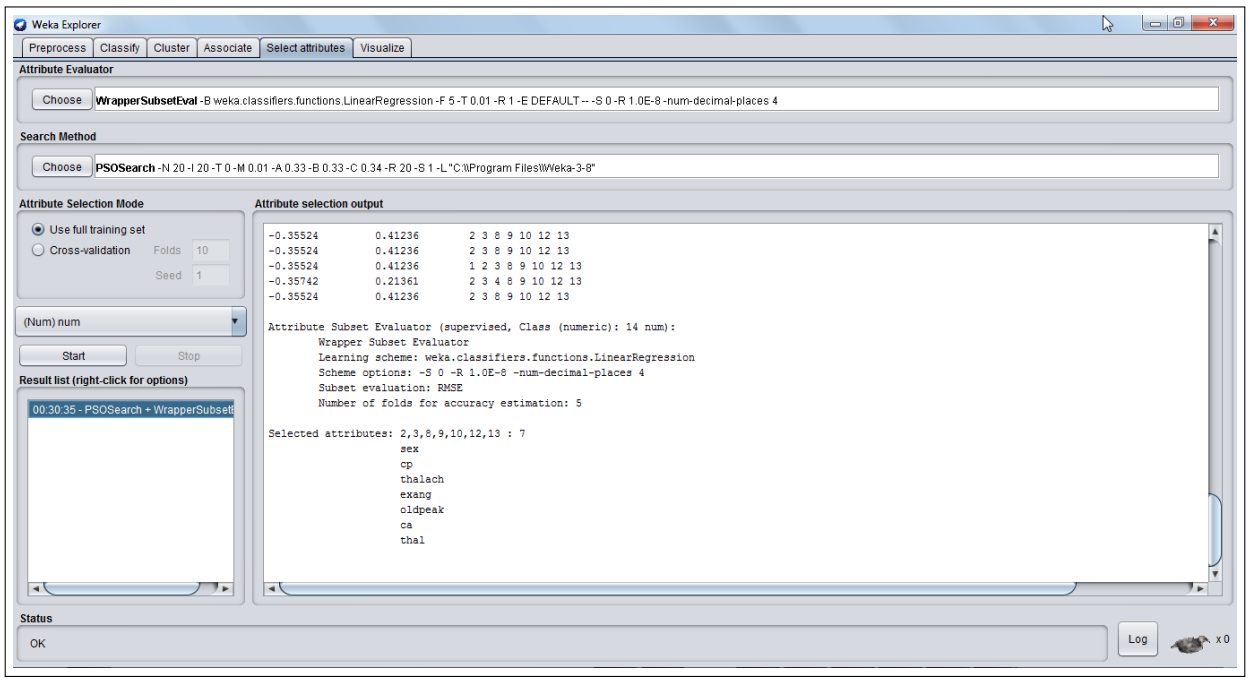

Figure 18: Execution of Particle Swarm Optimization (PSO) in WEKA

After reducing the dataset using PSO, the 7 optimal attributes is utilized to train the classifiers using R-Studio. For instance in experiment #2 PSO_ANN (Particle Swarm Optimization combined with Artificial Neural Network), Figure 19 depicts the use of sex, cp, thalach, exang, oldpeak, ca and thal attributes for training the ANN classifier out of the 14 attributes available in the actual dataset. Ten neurons has been used to train the ANN model as seen in Figure 20.

```
# Experiment 2 (PSO_ANN)
library(nnet) #Its drawback is that it only allows for a single hidden layer of neurons.
library(devtools)
library(caret)
neurons<-10 # declare number of nuerons
set.seed(100)
ann_hy<-nnet(num ~ sex+cp+thalach+exang+oldpeak+ca+thal, trainingCardiovascularData, size=neurons)
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044412703516c34f1a4684a5/nnet_plot_update.r')
plot.nnet(ann_hy)
ann_test_pred_hy<-predict(ann_hy, testingCardiovascularData, type = "class")
cm <-confusionMatrix(ann_test_pred_hy, testingCardiovascularData$num)
draw_confusion_matrix(cm)
```

Figure 19: Execution of PSO_ANN in R-Studio

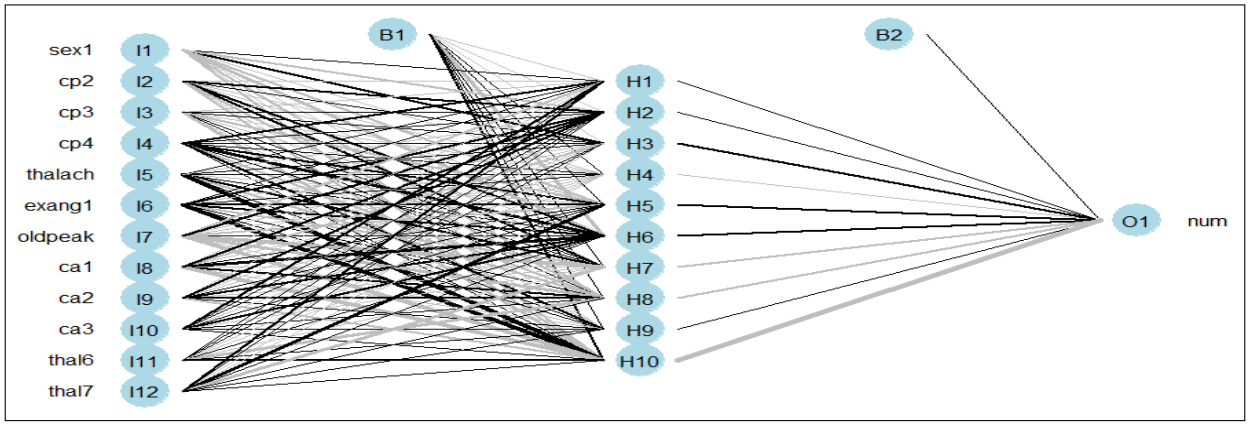

Figure 20: Generation of Perceptron in PSO_ANN

Similarly, the remaining experiments namely PSO_SVM, PSO_NB, PSO_KNN, PSO_RF and PSO_C5.0 are conducted and was evaluated based on accuracy, sensitivity and specificity.

5 Evaluation

We comprehend that data mining evaluation methodologies are business applications dependent and hence, the performance metrics considered in this research is adjusted to cardiovascular diseases diagnosis. According to the insights gained from literature review, the impact of all the 6 hybrid models, Particle Swarm Optimization (PSO) combined with Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree is evaluated based on accuracy, sensitivity and specificity. These performance metrics are calculated by using the values in the confusion matrix. Confusion matrix is a mechanism to demonstrate how the classifier is confused while predicting. Figure 21, illustrates the confusion matrix of a binary classification case such as 0/1, true/false, dead/alive and, disease/no-disease.

Predicted Class	True/Actual Outcome: Patient is free from CVDs	
	No Disease (Patient is free from CVDs)	Disease (Patient has CVDs)
No Disease (Patient is free from CVDs)	True Positive (TP)	False Positive (FP) (Patients has been diagnosed free from CVDs)
Disease (Patient has CVDs)	False Negative (FN) (Patients has been diagnosed with CVDs)	True Negative (TN)

Figure 21: Confusion Matrix for Binary Classification

$$Accuracy = \frac{(TP + FN)}{(TP + FP + FN + TN)} * 100$$

$$Sensitivity = \frac{(TP)}{(TP + FN)} * 100$$

$$Specificity = \frac{(TN)}{(TN + FP)} * 100$$

Table 5: Performance Metrics Formula

5.1 Experiment #1 (PSO_SVM)

The first experiment is evaluated to comprehend the performance of hybrid Support Vector Machines (SVM) and Particle Swarm Optimization (PSO). The model correctly classifies 74 (accuracy = 83.1%) instances

while 15 (16.9%) of the instances are classified incorrectly as seen in Figure 22. The model correctly identifies 39 patients out of 48 patients who are free from CVDs and 9 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 81.2%. Besides, the model correctly identifies 35 patients out of 41 patients to have CVDs and the 6 are identified incorrectly to be free from the disease. This result gives the model a specificity of 85.4%.

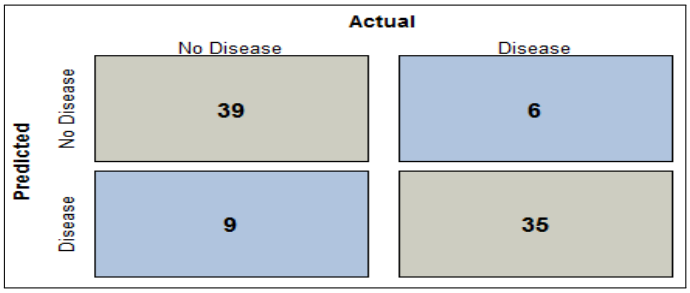

Figure 22: Confusion Matrix of PSO_SVM

PERFORMANCE METRIC	VALUE
Accuracy	83.1%
Sensitivity	81.2%
Specificity	85.4%

Table 6: Performance Evaluation of PSO_SVM

5.2 Experiment #2 (PSO_ANN)

The second experiment is evaluated to comprehend the performance of hybrid Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO). The model correctly classifies 67 (accuracy = 75.3%) instances while 22 (24.7%) of the instances are classified incorrectly as seen in Figure 23. The model correctly identifies 38 patients out of 48 patients who are free from CVDs and 10 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 79.2%. Besides, the model correctly identifies 29 patients out of 41 patients to have CVDs and the 12 are identified incorrectly to be free from the disease. This result gives the model a specificity of 70.7%.

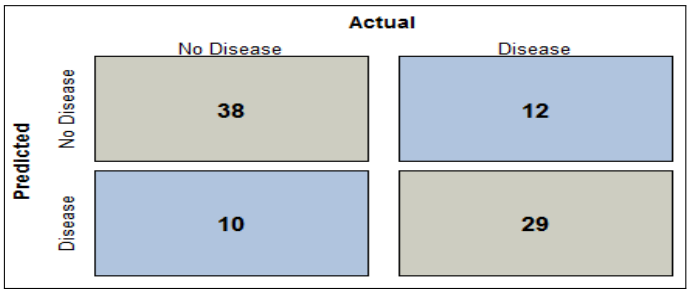

Figure 23: Confusion Matrix of PSO_ANN

PERFORMANCE METRIC	VALUE
Accuracy	75.3%
Sensitivity	79.2%
Specificity	70.7%

Table 7: Performance Evaluation of PSO_ANN

5.3 Experiment #3 (PSO_NB)

The third experiment is evaluated to comprehend the performance of hybrid Naïve Bayes and Particle Swarm Optimization (PSO). The model correctly classifies 71 (accuracy = 79.8%) instances while 18 (20.2%) of the instances are classified incorrectly as seen in Figure 24. The model correctly identifies 40 patients out of 48 patients who are free from CVDs and 8 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 83.3%. Besides, the model correctly identifies 31 patients out of 41 patients to have CVDs and the 10 are identified incorrectly to be free from the disease. This result gives the model a specificity of 75.6%.

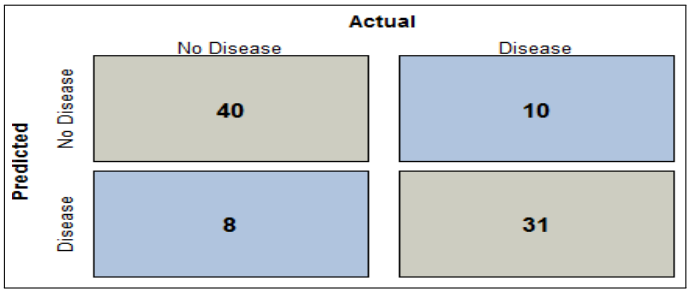

Figure 24: Confusion Matrix of PSO_NB

PERFORMANCE METRIC	VALUE
Accuracy	79.8%
Sensitivity	83.3%
Specificity	75.6%

Table 8: Performance Evaluation of PSO_NB

5.4 Experiment #4 (PSO_KNN)

The fourth experiment is evaluated to comprehend the performance of hybrid K-Nearest Neighbour (KNN) and Particle Swarm Optimization (PSO). The model correctly classifies 66 (accuracy = 74.2%) instances while 23 (25.8%) of the instances are classified incorrectly as seen in Figure 25. The model correctly identifies 37 patients out of 48 patients who are free from CVDs and 11 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 77.1%. Besides, the model correctly identifies 29 patients out of 41 patients to have CVDs and the 12 are identified incorrectly to be free from the disease. This result gives the model a specificity of 70.7%.

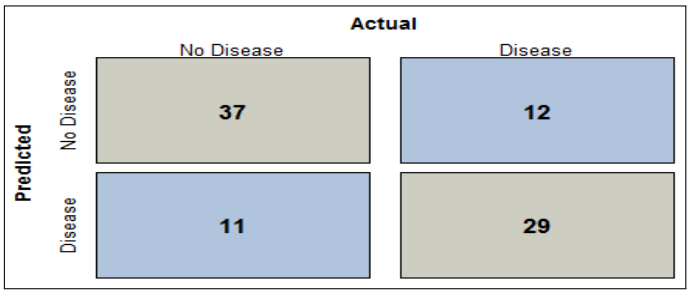

Figure 25: Confusion Matrix of PSO_KNN

PERFORMANCE METRIC	VALUE
Accuracy	74.2%
Sensitivity	77.1%
Specificity	70.7%

Table 9: Performance Evaluation of PSO_KNN

5.5 Experiment #5 (PSO_RF)

The fifth experiment is evaluated to comprehend the performance of hybrid Random Forest and Particle Swarm Optimization (PSO). The model correctly classifies 71 (accuracy = 79.8%) instances while 18 (20.2%) of the instances are classified incorrectly as seen in Figure 26. The model correctly identifies 38 patients out of 48 patients who are free from CVDs and 10 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 79.2%. Besides, the model correctly identifies 33 patients out of 41 patients to have CVDs and the 8 are identified incorrectly to be free from the disease. This result gives the model a specificity of 80.5%.

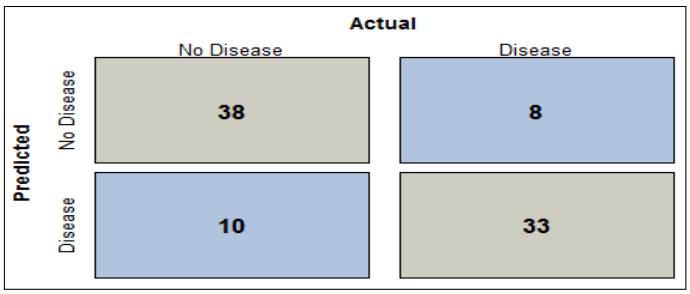

Figure 26: Confusion Matrix of PSO_RF

PERFORMANCE METRIC	VALUE
Accuracy	79.8%
Sensitivity	79.2%
Specificity	80.5%

Table 10: Performance Evaluation of PSO_RF

5.6 Experiment #6 (PSO_C5.0)

The last experiment is evaluated to comprehend the performance of hybrid C5.0 Decision Tree and Particle Swarm Optimization (PSO). The model correctly classifies 68 (accuracy = 76.4%) instances while 21 (23.6%) of the instances are classified incorrectly as seen in Figure 27. The model correctly identifies 37 patients out of 48 patients who are free from CVDs and 11 are identified incorrectly to have the disease while they actually didn't have. This result gives the model a sensitivity of 77.1%. Besides, the model correctly identifies 31 patients out of 41 patients to have CVDs and the 10 are identified incorrectly to be free from the disease. This result gives the model a specificity of 75.6%.

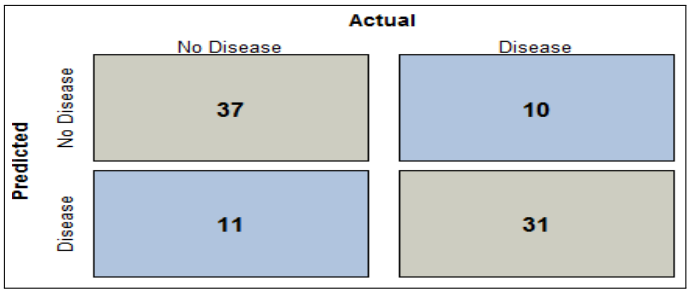

Figure 27: Confusion Matrix of PSO_C5.0

PERFORMANCE METRIC	VALUE
Accuracy	76.4%
Sensitivity	77.1%
Specificity	75.6%

Table 11: Performance Evaluation of PSO_C5.0

5.7 Discussion

A detailed discussion of the findings from the 6 experiments has been presented in this section. Results has been interpreted and models were compared for superior predictive performance over each other. The results of all the experiments in this research has been combined in Table 12 and Figure 28.

EXPERIMENT # (MODEL)	ACCURACY	SENSITIVITY	SPECIFICITY
Experiment #1 (PSO_SVM)	83.1%	81.2%	85.4%
Experiment #2 (PSO_ANN)	75.3%	79.2%	70.7%
Experiment #3 (PSO_NB)	79.8%	83.3%	75.6%
Experiment #4 (PSO_KNN)	74.2%	77.1%	70.7%
Experiment #5 (PSO_RF)	79.8%	79.2%	80.5%
Experiment #6 (PSO_C5.0)	76.4%	77.1%	75.6%

Table 12: Performance Comparison of all the Models

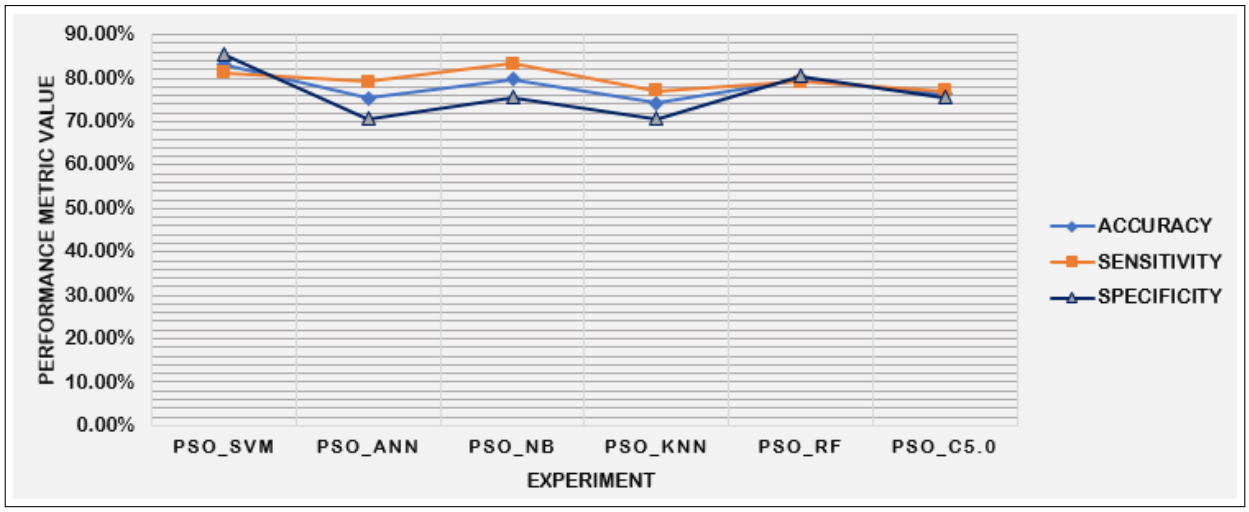

Figure 28: Performance Comparison of all the Models Graphically

As seen from Table 12 and Figure 28, PSO_SVM with a value of 83.1% outperforms other models in terms of accuracy. It is followed by both PSO_NB and PSO_RF with an accuracy of 79.8%. PSO_KNN with an accuracy of 74.2% is the worst performer in terms of accuracy. In terms of sensitivity, PSO_NB outperforms other models with a value of 83.3%. It is followed by PSO_SVM with a value of 81.2%. Both PSO_KNN and PSO_C5.0 with a value of 77.1% are the worst performer in terms of sensitivity. PSO_SVM with a value of 85.4% again leads the list in terms of specificity. It is followed by PSO_RF with a value of 80.5%. Both PSO_ANN and PSO_KNN are the poorest performers based on specificity with a value of 70.7%. It is very clear that PSO_SVM outperforms other models in two performance measures out of three. Therefore this research concludes a hybrid combination of Particle Swarm Optimization (PSO) and Support Vector Machines (SVM) as having the superior predictive performance over other models using Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree classification algorithms.

6 Conclusion and Future Work

In this research, the aim is to test which hybrid model of feature selection and classification algorithms classifies cardiovascular diseases (CVDs) efficiently for improved and reliable diagnosis. The feature selection algorithm used is Particle Swarm Optimization (PSO). The classifier used are Support Vector Machines (SVM), Artificial Neural Network (ANN), Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest and C5.0 Decision Tree. Data is collected from University of California, Irvine - Machine Learning Repository (UCI-ML). The hybrid models are evaluated based on accuracy, sensitivity and specificity with the objective of achieving superior predictive performance. The most effective model for diagnosing CVDs is a hybrid combination of Particle Swarm Optimization (PSO) and Support Vector Machines (SVM). This research illustrated that data mining techniques can be used efficiently to diagnose cardiovascular diseases. The end result of this research can be utilized as an auxiliary tool by physicians to help them to make more consistent diagnosis of CVDs.

The work can be further enhanced and expanded by training the models with dataset of other disease such as pulmonary diseases and cancer. This would enable diagnosis of other major diseases besides CVDs. This work can also be extended by considering patient's physical attributes besides clinical attributes for diagnosing CVDs or other diseases.

Acknowledgment

I would like to first thank my supervisor Dr. Paul Stynes, Vice-Dean of Academic Programmes and Research at School of Computing, National College of Ireland. Dr. Stynes, always provided his assistance whenever I ran into a trouble spot or needed clarifications about my research or writing. He invariably allowed this research to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank my friend, Dr. Somshubra Paul for providing his clinical insights aligned to the data set used in this research. Without his passionate participation and inputs, the experiment could not have been successfully conducted.

I would also like to acknowledge my wife, Mrs Arpita Chakraborty and friend, Dr. Indrakshi Dey for their committed and patient help with the proof reading and insightful inputs. I am gratefully indebted to them for their very valuable comments on this thesis.

References

A Sudha, P. G. and Jaishankar, N. (2012). Utilization of data mining approaches for prediction of life threatening disease survivability, *IJAC* .

Abdullah, A. and Rajalaxmi., R. (2012). A data mining model for predicting the coronary heart disease using random forest classifier, *International Conference in Recent Trends in Computational Methods, Communication and Controls* .

Acharya, A. (2017). Comparative study of machine learning algorithms for heart disease prediction, *Helsinki Metropolia University of Applied Sciences* .

Aditya, A. (1991). Learning algorithm for neural network, *Dissertation (Ph.D.), California Institute of Technology* .

Alex C Alberto, G. A. L. and Nadal, J. (2012). Detection of heart disease using binary particle swarm optimization, *Proceedings of the Federated Conference on Computer Science and Information Systems* pp. 177–182.

Alty, S. R., Millasseau, S. C., Chowienczyc, P. J. and Jakobsson, A. (2003). Cardiovascular disease prediction using support vector machines, *2003 46th Midwest Symposium on Circuits and Systems*, Vol. 1, pp. 376–379 Vol. 1.

Aravinthan, K. and Vanitha, D. M. (2016). A comparative study on prediction of heart disease using cluster and rank based approach, *International Journal of Advanced Research in Computer and Communication Engineering* .

Attribute Description: UCI-ML (2017). Accessed: 2017-10-06.
URL: *http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names*

Australian Bureau of Statistics (2014). http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/3303.0~2015~Main%20Features~Australia%27s%20leading%20causes%20of%20death,%202015~3. Accessed: 2017-06-06.

Auza, J. (2010). 5 of the best and free open source data mining software. Accessed: 2017-10-17.
URL: *http://www.junauza.com/2010/11/free-data-mining-software.html*

Awang, R. and Palaniappan, S. (2008). Intelligent heart disease prediction system using data mining techniques, *IJCSNS* pp. 343–350.

Benxian Yue, Weihong Yao, A. A. and Liu, H. (2007). A new rough set reduct algorithm based on particle swarm optimization, *IWINAC, Springer Verlog* pp. 397–409.

Bharti, S. and Singh, D. N. (2015). Analytical study of heart disease prediction comparing with different algorithms, *International Conference on Computing, Communication and Automation* .

Bhatia, N. (2013). An analysis of heart disease prediction using different data mining techniques, *International Journal of Engineering Research & Technology* .

B.N.Lakshmi and G.H.Raghunandhan (2011). A conceptual overview of data mining, pp. 27–32.

Bramer, M. (2013). *Principles of Data Mining. 2nd edition*, Springer.

Chitra, R. and Seenivasagam, V. (2013). Review of heart disease prediction system using data mining and hybrid intelligent techniques, *ICTACT Journal On Soft Computing* .

Code Of Ethics of University of California, Irvine (2017). Accessed: 2017-10-06.
URL: *http://www.uci.ch/mm/Document/News/News/17/71/94/UCICodeofEthics_English.pdf*

D. P. Shukla, S. B. P. and Sen, A. K. (2014). A literature review in health informatics using data mining techniques, *International Journal of Software Hardware Research in Engineering* .

Daniel, L. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley Sons, Inc., Hoboken, New Jersey.

Das, R., Turkoglu, I. and Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles, *Expert Systems with Applications* **36**(4): 7675 – 7680.
URL: *http://www.sciencedirect.com/science/article/pii/S095741740800657X*

Elbedwehy, M. (2012). Detection of heart disease using binary particle swarm optimization, *Proceeding of the Federated Conference in Computer Science and Information System* .

Elnaz Pashaei, M. O. and Aydin, N. (2015). Improving medical diagnosis reliability using boosted c5.0 decision tree empowered by particle swarm optimization, *IEEE* .

ESCAP (2010). http://www.unescap.org/stat/data/syb2009/9.Health-riskscauses-of-death.asp. Accessed: 2017-06-06.

European Public Health Alliance (February, 2011). http://www.epha.org/a/2352. Accessed: 2017-06-06.

Fogoros, R. N. (2009). Key symptoms of heart disease. Accessed: 2017-10-06.
URL: *https://www.verywell.com/key-symptoms-of-heart-disease-1745915*

G. Subbalakshmi, K. R. and Rao, M. (2011). Decision support in heart disease prediction system using naive bayes, *Indian Journal of Computer Science and Engineering (IJCSE)* pp. 170–176.

Gupta, M. P. and Bajaj, P. (2014). Heart disease diagnosis based on data mining and neural network, *International Journal of Engineering Sciences Research Technology (IJESRT)* .

Heller, C. (2008). 5 common types of heart disease. Accessed: 2017-10-06.
URL: *http://ezinearticles.com/?5-Common-Types-of-Heart-Diseaseid=1073496*

Hideg, I. and Kleef, G. V. (2013). The consequences of faking anger in negotiations, *Journal of Experimental Social Psychology* .

Huan Liu, Hiroshi Motoda, R. S. and Zhao, Z. (2010). Feature selection: An everlasting frontier in data mining, *JMLR: The 4th Workshop on Feature Selection and Data Mining* .

I. Nahar, T. and Chen, Y. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, *Expert systems with applications* .

Ian H. Witten, E. F. and Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques, 3rd Edition*, Morgan Kaufmann, San Francisco.

Jabbar, M. (2012). Knowledge discovery from mining association rules for heart disease prediction, *JATIT* .

Jennifer S. Lerner, Ye Li, P. V. and Kassam, K. (2014). Emotions and decision making, *Psychological Review* .

J.Soni (2011). Intelligent and effective heart prediction system using weighted associative classifiers, *International Journal on Computer Science and Engineering* pp. 2385–2392.

Kalaiselvi, C. (2016). Diagnosing heart diseases using average k-nearest neighbour algorithm of data mining, *IEEE* .

K.Srinivas, B. R. and A.Govrdhan, D. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks, *International Journal on Computer Science and Engineering* pp. 250–255.

K.Sudhakar and Manimekalai, D. M. (2014). Study of heart disease prediction using data mining, *International Journal of Advanced Research in Computer Science and Software Engineering* .

Kuhn, M. and Johnson, K. (2013). Applied predictive modeling, *Springer* .

Kumari, M. and Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction, *International Journal Of Computer Science and Technology* .

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification, *IEEE* .

Machine Learning Repository of University of California, Irvine (2017). Accessed: 2017-10-06.
URL: *http://archive.ics.uci.edu/ml/about.html"*

Mai Shouman, T. T. and Stocker, R. (2012). Using data mining techniques in heart disease diagnosis and treatment, *IEEE* .

Marjia Sultana, A. H. and ShorifUddin, M. (2016). Analysis of data mining techniques for heart disease prediction, *IEEE* .

Martinez, W. L.; Martinez, A. R. . S. J. (2010). *Exploratory Data Analysis with MATLAB, second edition*, Chapman Hall/CRC.

Masethe, H. D. and Masethe, M. A. (2014). Prediction of heart disease using classification algorithms, *Proceeding of the World Congress on Engineering and Computer Science* .

Ming Hsu, Meghana Bhatt, R. A. D. T. and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making, *Science Magazine* .
URL: *www.sciencemag.org/cgi/content/full/310/5754/1680/DC1*

MN Ab Wahab, S. N.-M. and Atyabi, A. (2015). A comprehensive review of swarm optimization algorithms, *PLoS ONE* .

Mohammed Abdul Khaled, S. K. P. and Dash, G. (2013). A survey of data mining techniques on medical data for finding locally frequent diseases, *International Journal of Advanced Research in Computer Science and Software Engineering* pp. 1137–1145.

Mokeddem, S. (2013). Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm, *International Conference on Computational Science and Engineering* .

Niwas, S. I., Kumari, R. S. S. and Sadasivam, V. (2005). Artificial neural network based automatic cardiac abnormalities classification, *Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05)*, pp. 41–46.

Paladugu, S. (2010). Temporal mining framework for risk reduction and early detection of chronic diseases, *University of Missouri-Columbia* .

Parthiban, L. and R.Subramanian (2007). Intelligent heart disease prediction system using canfis and genetic algorithm, *International Journal of Biological and Life Sciences* .

Patro, S. G. K. and Sahu, K. K. (2015). A technical analysis of financial forecasting, *International Journal of Computer Sciences and Engineering* .

Paul, A. K., Shill, P. C., Rabin, M. R. I. and Akhand, M. A. H. (2016). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease, *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 145–150.

PEHLİVANLI, D. A. (2011). The comparison of data mining tools, *Department of Computer Engineering, İstanbul Kültür University* . Accessed: 2017-10-17.

Princy, T. and Thomas, J. (2016). Human heart disease prediction system using data mining techniques, *International Conference on Circuit, Power and Computing Technologies [ICCPCT]* .

Rajkumar, A. and Reena, M. G. (2010). Diagnosis of heart disease using datamining algorithm, *Global Journal of Computer Science and Technology* pp. 38–43.

Rish, I. (2001). An empirical study of the naive bayes classifier, *IJCAI Workshop on Empirical Methods in AI* .

Rohilla, J. and Gulia, P. (2015). Analysis of data mining techniques for diagnosing heart disease, *International Journal of Advanced Research in Computer Science and Software Engineering* .

Ruben, D. (2009). Data mining in healthcare: Current applications and issues.

Shan Xu, Zhen Zhang, D. W. J. H. X. D. and Zhu, T. (2017). Cardiovascular risk prediction method based on cfs subset evaluation and random forest classification framework, *IEEE 2nd International Conference on Big Data Analysis* .

Shwartz, S. and David, B. (2014). *Understanding Machine Learning*, Cambridge University Press, New York.

Sivagowry, S. and Durairaj, M. (2014). An intellectual technique for feature reduction on heart malady anticipation data, *International Journal of Advanced Research in Computer Science and Software Engineering* .

South Africa Statistics (2013). `http://www.statssa.gov.za/?page_id=737&id=3`. Accessed: 2017-06-06.

The National Institute of Health of U.S. Department of Health and Human Services (2017). Accessed: 2017-10-06.
URL: *https://www.nhlbi.nih.gov/health/health-topics/topics/hdw/diagnosis*

The Rapid Miner Platform (2017). Accessed: 2017-10-06.
URL: *https://rapidminer.com/products/*

The R Project for Statistical Computing (2017). Accessed: 2017-10-06.
URL: *https://www.r-project.org/*

Wang, X. (2007). Feature selection based on rough sets and particle swarm optimization, *Elsevier* .

Webster, J. and Watson, R. T. (2002). Analyzing the past to preparefor the future:writing aliterature review, *MIS* pp. 1137–1145.

Wimmer, H. and Powell, L. M. (2015). A comparison of open source tools for data science, *Proceedings of the Conference on Information Systems Applied Research Wilmington, North Carolina USA* .

World Health Organization (May, 2017). `http://www.who.int/mediacentre/factsheets/fs317/en/`. Accessed: 2017-06-06.

X. Chen, Y. Y. and Williams, G. (2007). A survey of open source data mining systems emerging technologies in knowledge discovery and data mining, *Springer* .

Yin, R. K. (2013). Case study research: Design and methods (applied social research methods), *5th edition. Los Angeles, Sage Publications;* .

Yuanning Liu, G. W. and Wang, S. (2011). An improved particle swarm optimization for feature selection, *Journal of Bionic Engineering* .