



T.C.

ISTANBUL ALTINBAS UNIVERSITESI  
GRADUATE SCHOOL OF SCINCES  
ENGINEERING

Master of Information Technology

**HEART DISEASE DIAGNOSTIC USING DATA  
MINING TECHNIQUES**

Asaad Qasim Shareef

Thesis supervisor

Asst.Prof. Dr. Sefer KURNAZ

Istanbul (2018)

# **HEART DISEASE DIAGNOSTIC USING DATA MINING TECHNIQUES**

by

**Asaad Qasim Shareef**

Information Technology

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree

Master

ALTINBAŞ UNIVERSITY

2018

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

---

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Asst. Prof. Dr.Zeynep ALTAN

Software Engineering, Istanbul  
Beykent University

---

Asst. Prof. Dr. Sefer KURNAZ

Software Engineering, Istanbul  
Altinbaş University

---

Prof. Dr. Osman N. UCAN

Electrical Engineering ,  
Istanbul Altinbaş University

---

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

---

Assoc. Prof. Dr.Oğuz ATA

Head of Department

---

Assoc. Prof. Dr,Oğuz BAYAT

Director

Approval Date of Graduate School of  
Science and Engineering: \_\_\_\_/\_\_\_\_/\_\_\_\_

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Asaad Qasim Shareef

## DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time. It is also dedicated to my brothers and sisters they supported me in this study .it is also dedicated to my wife which support me throughout my studied and my new daughter Joanna who inspired me to reach this moment



## ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Sefer Kurnaz of the Faculty computer of Science at Altinbaş. The door to Dr. Kurnaz office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

Finally, I must express my very profound gratitude to my parent's and to my wife and my brothers and sisters and the parents of my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Asaad Qasim Shareef

## **ABSTRACT**

### **HEART DISEASE DIAGNOSTIC USING DATA MINING TECHNIQUES**

Shareef , Asaad Qasim Shareef,

M.Sc., Information Technologies, Istanbul Altınbaş University,

Supervisor: Asst.Prof. Dr. Sefer KURNAZ

Date: May 2018

Pages: 49

Lately, large masses of data have been generated due to the ongoing approaches in biotechnology and fitness sciences areas. It combines clinical information and genetic data which included in Electronic Health Records (EHRs). On the other side, it is required to recognize symptoms, which can wrongly convince the human health in addition to placing economic burdens on their shoulders, in an early stage to avoid many difficulties. Lately, several data mining procedures have played a vital role in developing automated operations that can identify syndromes efficiently and correctly. In this thesis, we satisfy some of the research disciplines that have employed either the data mining procedures for identifying symptoms. Additionally, a set of well-known data mining methods including Decision Trees (j48), Naïve Bayes, Multilayer Perceptron (MLP), and Random Forest (RF) has been assessed in performing the classification task using a publicly available heart diseases dataset.

**KEYWORDS:** Data Mining, Heart Disease, Classification, Healthcare, Syndrome detection..

## ÖZET

### VERİ MADENCİLİĞİ TEKNİKLERİ KULLANARAK KALP HASTALIĞI TANISI

Shareef , Asaad Qasim Shareef,

M.Sc., Bilgi Teknolojisi, Altınbaş Üniversitesi,

Danışman : Yrd. Dr. Sefer KURNAZ

Tarih: Mayıs 2018

Sayfalar : 49

Son zamanlarda, biyoteknoloji ve fitness bilimleri alanlarındaki devam eden yaklaşımlar nedeniyle büyük veri yığınları oluşturulmuştur. Elektronik Sağlık Kayıtlarında (EHR'ler) bulunan klinik bilgileri ve genetik verileri birleştirir. Öte yandan, birçok zorluğu önlemek için erken aşamada ekonomik yükleri omuzlarına yerleştirmenin yanı sıra insan sağlığını da inandırabilecek semptomları tanımak gerekmektedir. Son zamanlarda, birçok veri madenciliği prosedürü, sendromları etkili ve doğru bir şekilde tanımlayabilen otomatik operasyonların geliştirilmesinde hayati bir rol oynamıştır. Bu tezde, belirtileri tanımlamak için veri madenciliği prosedürlerini uygulayan bazı araştırma disiplinlerini sağlarız. Buna ek olarak, Karar Ağaçları (j48), Naïve Bayes, Çok Katmanlı Perceptron (ÇKP) ve Rastgele Orman (RO) gibi bir dizi iyi bilinen veri madenciliği metodu, kamuya açık kalp hastalıkları veri setini kullanarak sınıflandırma görevini gerçekleştirirken değerlendirilmiştir.

ANAHTAR KELİMELER: Veri Madenciliği, Kalp Hastalığı, Sınıflandırma, Sağlık, Sendrom tespiti



## TABLE OF CONTENTS

### Pages

<b>LIST OF TABLES .....</b>	<b>xi</b>
<b>LIST OF FIGURES .....</b>	<b>xii</b>
<b>1. INTRODUCTION .....</b>	<b>1</b>
1.1 HEART DISEASE .....	1
1.2 DATA MINING .....	2
1.3 CONTRIBUTION .....	3
1.4 RISK FACTOR IN HEART DISEASE .....	4
1.4.1 Risk Factors in details.....	4
1.5 THESIS ORGANIZATION .....	5
<b>2. RELATED WORKS .....</b>	<b>6</b>
2.1 CLASSIFICATION TECHNIQUES.....	6
2.2 CLUSTERING .....	7
2.3 ASSOCIATION .....	7
2.4 REGRESSION .....	8
2.5 RELATED WORKS BASED ON DATA MINING.....	8
2.6 SUMMARY .....	12
<b>3. COMPARATIVE STUDY.....</b>	<b>14</b>
3.1 DATASET.....	14
3.1.1 Attributes of dataset.....	14
3.2 DATA MINING TECHNIQUES.....	15
3.2.1 Decision tree Algorithm. ....	16
3.2.2 J48 Algorithm.....	16
3.2.3 Naïve Bayes Algorithm.....	16

3.3	WEKA .....	16
3.3.1	Input Data in WEKA. ....	17
3.3.2	J48 Result. ....	21
3.3.3	Naïve Bayes Result.....	23
3.3.4	Multi-preceptron Result.....	25
3.3.5	Support Vector Machine Result. ....	27
3.4	SUMMARY .....	29
<b>4.</b>	<b>PROPOSED FRAMEWORK.....</b>	<b>31</b>
4.1	FRAMEWORK DESCRIPTION .....	32
4.1.1	Pre-processing Phase. ....	32
4.1.2	Hybrid Classification Phase.....	32
4.2	Preprocessing Implementation using PHP code.....	33
4.3	Decision Tree Implementation using PHP code.....	35
4.4	Random Forest Implementation using PHP code .....	37
<b>5.</b>	<b>EXPERIMENTAL RESULTS .....</b>	<b>38</b>
5.1	DEVICE AND TOOL CAPABILITES.....	38
5.2	IMPLEMENTATION SCREENSHOTS .....	38
5.3	EXPERMINTAL RESULT AND DESCUSSION .....	44
<b>6.</b>	<b>CONCLUSION.....</b>	<b>47</b>
<b>7.</b>	<b>REFERENCE .....</b>	<b>48</b>

## LIST OF TABLES

	<u>Pages</u>
Table 2.1: A summary of some related work based on data mining techniques .....	12
Table 3.1: J48 Result from WEKA program .....	22
Table 3.2: Native Bayes Result from WEKA program .....	24
Table 3.3: Multi-layer perceptron Result from WEKA program .....	26
Table 3.4: SVM Result from WEKA program .....	28
Table 5.1: Tools and device used to preform proposed framework .....	38
Table 5.2: Confusion Matrix.....	44
Table 5.3: Compartive Reuslt between Proposed hybrid algorithm and other algorithms.....	45

## LIST OF FIGURES

	<u>Pages</u>
Figure 2.1: The different data mining techniques used in healthcare management.....	6
Figure 3.1: Number of instances of heart disease Dataset using WEKA program.....	19
Figure 3.2: Statistical values of Age attribute using WEKA program .....	19
Figure 3.3: Statistical values of Sex attribute using WEKA program .....	19
Figure 3.4: Statistical values of Chest Type attribute using WEKA program.....	20
Figure 3.5: J-48 Result using WEKA program.....	21
Figure 3.6: Naïve Bayes Result using WEKA program .....	23
Figure 3.7: Neural Network using WEKA program.....	25
Figure 3.8: SVM Result using WEKA program.....	27
Figure 3.9: Accuracy and Error Diagram of Comparative Study between Datamining techniques used in WEKA Program .....	29
Figure 3.10: Time Diagram of Comparative Study between Datamining techniques used in W Program	30
Figure 4.1: Proposed Framework .....	31
Figure 4.2: PHP pre-processing code.....	33
Figure 4.3: PHP decision tree code .....	35
Figure 4.4: PHP Random forest code.....	37
Figure 5.1: Data Set from UCI location .....	39
Figure 5.2: Uploading Data Set to the server .....	40
Figure 5.3: Pre-processing for data set in the server.....	41
Figure 5.4: Decision Tree form .....	42
Figure 5.5: Accuracy and Error Diagrams .....	46

# 1. INTRODUCTION

A variety of applications use data-mining procedures on a large scale. In the healthcare industry, for example, data mining plays an essential role in predicting or diagnosing diseases with reasonable accuracy. One critical application is to diagnose the heart diseases or cardiovascular as these diseases are recognized as the leading condition of mortality globally in our modern society [1].

## 1.1 HEART DISEASE

Accordingly the World Health Organization (WHO), Extra than seventeen million individual died from cardiovascular illnesses in 2013, and around three million of this mortality happened before the age of sixty [2]. However, 90% of that mortality was estimated to be preventable if patients have correctly been diagnosed early and they improved their habits such as healthy eating, exercise, and a like [3]. In traditional healthcare environments, diagnosis of a disease depends on doctor's decision for identifying it as the most likely cause depending on a person's symptoms. However, this leads to unwanted mistakes that are resulting in more medical costs and affecting the property of service afforded to patients. Instead, expert systems [4] could be applied to emulate the decision-making ability of a human expert for answering not only simple issues like "What is the normal age of inmates who have heart illness?", "Recognize the female inmates who are eligible, and who have been treated for heart illness?", but also complex ones like "Given inmate reports, foresee the possibility of inmates who diagnosed a heart illness?".

"Find the most significant risk factor that results in heart disease?" Of course, using such systems could decrease pharmaceutical errors, and reduce practice exception, but surprisingly it can diagnose results. Discovering knowledge use procedures of data mining in vast volumes of data through detecting patterns and summarizing data into a format that can be understood.

## 1.2 DATA MINING

Data-mining could be a track associated with many ideas to supply important data that's shaped to apply in varied applications. Nevertheless, data is obtainable that embody a secret consciousness, data-mining procedure exploration for the balance, proportions, outlier events during this information and instant as mysterious consciousness. Then this consciousness is often utilized in varied forms. Predictive is singular in of the mission that uses the deep expertise, Elicited and fashioned to declare unknown amount supported this consciousness tomorrow. Forecast task is collect as 1 of 2 kinds; one will either plan to foretell some unprocurable data amounts or until trends or forecast the section marker to a few data also is joined on classification. Formerly the classification pattern is constructed supported a coaching set, a category description of associate degree target may be foretold, supported the characteristic values of an article and therefore that property values from this categories. This forecast is applied to estimate that missing integral amounts or increment/ decrement biases inside time/related information. This principal plan moves to do an outsized variety of last amount values to think about likely coming prices. During investigation mission generality, it'll be required to declaration binary logic state, i.e., presence or absence of health problem within the person in keeping with his medical report. Then, classification is often handled as analysis whereas satisfying a prospect of the survival from a condition

In fact, there are three principal methods of data-mining this can do utilized to a classification previously random data in predefined classes or to categorize new data into pre-existing categories. This could be done by examining the data that has previously been classified, learning the rules of classification and applying those rules to new data. Or to identify relationships between them and develop a pattern of these relationships. These patterns could be then used as a reference to predict future behavior [6]. Surveys such as [3, 4, 5] have been discussed the impact and power of data mining techniques in predicting systems.

### **1.3 CONTRIBUTION**

Inside our thesis, we have focused on data mining classification methods these are capable of forecasting a certain consequence based on a specified input. In particular, we have utilized four classifiers and create comparative study to analyze a medical dataset that recorded previously to diagnosis heart disease. The main contribution of this thesis, apply hybrid techniques (Decision tree and Random forest) of highest accuracy from comparative study.

We suggest a replacement structure for the application that works on real-life datasets to create effective data-mining standards taking into thought all the appropriate discourse circumstances. Its value considering here that the method isn't conation on however the discourse circumstances area unit collected however on however these circumstances area unit seized to accumulate consistent and proper arrangements. The purpose of the structure is that the use of context circumstances to realize higher forecast and precision of the information mining technique. The structure is found towards medical datasets, and therefore the cases wont see mistreatment the mining, particularly classification rules to develop call support systems associated to the medical domain.

A number of trials have been constructed to compare the accuracy of the implemented classifiers on a different size full training dataset with 14 attributes. Results showed that Decision tree outperforms other classifiers with an accuracy rate of 99.0%, which provided a more effective and comprehensive forecasting mechanism that can be integrated with the medical information system to assist in the analysis of the heart diseases in the earlier stages.

## **1.4 RISK FACTORS IN HEART DISEASE DATASET**

The properties (clinical circumstances) can be explained within the standing of a correctly decided situation. A situation illustrates a typical position once a doctor requires to spot the possibility of an inmate having the most blood vessels < fifty percent or > fifty percent narrowing as a period of illness danger. The data-mining purpose may be a natural alternative, for creating a forecast basis by mining the current information repository including the significant quantity of data. A standard data mining utilization would need a giant collection of data parameters to question the prediction model to lead to the expected preference. Nevertheless, by rigorously choosing the context circumstances the user is created to present solely a touch assemblage of input [18] whereas system assumes the remainder supported discourse factors.

### **1.4.1 Risk Factors in details**

Let us study the characteristics of the dataset in details:

1. (age) Age in years
2. (sex)Sex
  - Value 1: Male
  - Value 0: Female
3. (chest pain) chest pain type
  - Value 1: Typical angina,
  - Value 2: Atypical angina
  - Value 3: Non-anginal pain
  - Value 4: Asymptomatic
4. (trestbps) resting blood pressure
5. (chol) Serum cholestoral in mg/dl
6. (fbs) (Fasting blood sugar >120 mg/dl)
  - Value 1: True and Value 0: False
7. (restecg)resting electrocardiographic results
  - Value 0: Normal
  - Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV)
  - Value 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria



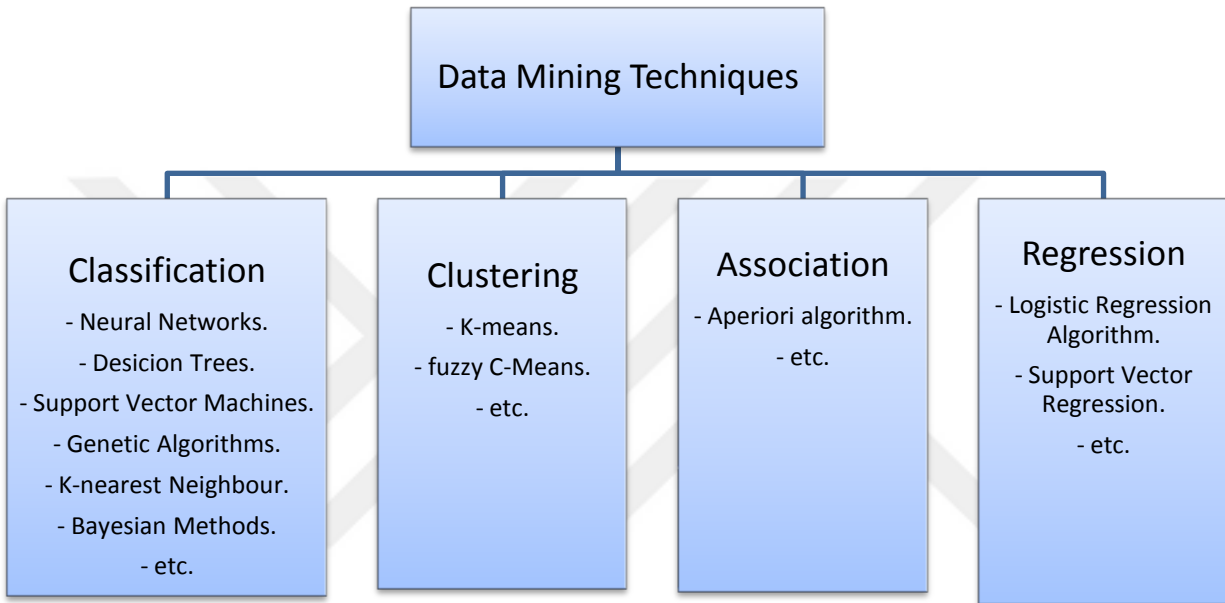
8. (thalach) Maximum heart rate achieved
9. (exang) Exercise induced angina (1 = yes; 0 = no)
10. (oldpeak) ST depression induced by exercise relative to rest
11. (slope) The slope of the peak exercise ST segment
  - Value 1: upsloping,
  - Value 2: flat,
  - Value 3: downsloping
12. (ca) Number of major vessels (0-3) colored by flourosopy
13. (thal) the heart status
  - Value 3: Normal,
  - Value 6: Fixed defect,
  - Value 7: Reversable defect
14. (num)Diagnosis of heart disease (angiographic disease status)
  - Value 0: <50% Diameter narrowing
  - Value 1: >50% Diameter narrowing

## **1.5 THESIS STRUCTURE**

The remainder the thesis is prearranged as follows. The reviewing of some related works to the proposed approach is presented in chapter II. Comparative study introduced in chapter III. Chapter IV discusses the research methodology. The results and discussions are obtainable into Part V. The final conclusions including later works are offered in chapter VI.

## 2. RELATED WORKS

Mostly, there several data mining techniques that are adopted in health care equivalent to Classification, Clustering, Association, and regression as shown in figure 2.1. A quick description concerning every one of them is provided next.



**Figure 2.1:** The different data mining techniques used in healthcare management.

### 2.1 CLASSIFICATION TECHNIQUES

Classification breaks data units into distinct groups. The categorization procedure foretells the aim category for every data points. To Illustrate, An inmate is grouped like “great risk” or “below risk” inmate with their illness model victimization data organization approach. It's a supervised coaching procedure having known category divisions. Binary and Multi-level area unit the two arrangements of classification, During a binary arrangement, Solely two potential conditions adore, “High” or “Low” risk patient is also thought of whereas the multiclass strategy has quite two purposes to Illustrate, “large,” “moderate” and “fading” risk inmate [4].

Classification contains two footsteps. The initial round is designing structure, which is applied to explain the practice dataset. The next round is a designing method wherever the assembled form that applied to classification. This efficiency about this categorization is measured according to the rate of experience units or experience dataset those are accurately classified [5]. There is a big set of an outsized cluster of techniques that are employed in care supervision to complete this coordination processing which includes: J-48, SVM, K-nearest neighbor, neural networks, Bayesian methods, etc...

## **2.2 CLUSTERING**

Clustering is the initial assignment into data-mining including a standard method concerning arithmetical data interpretation applied in various areas, use pattern identification, image processing, data retrieval, and bioinformatics [6]. Clustering is AN unsupervised training methodology that's distinct from categorization. Grouping is in contrast to distribution because it has no predefined levels. In collecting big data are divided into the structure of little separate subgroups or clusters. Clustering divided the info points supported the correlation live [4]. Clustering procedures discover getting the info such that an object in the equivalent cluster is a lot of regarding alternative (one another) than other teams. Several clustering procedures used in healthcare administration such as K-means, Fuzzy C-Means, etc.

## **2.3 ASSOCIATION**

Association Rule Mining developed abundant delayed from than machine learning also is directed facing larger management of this investigation field of databases. Although association rule mining was first presented as a storage box analysis agent, that should since enhanced example from that most important means for performing unsupervised exploratory data study aloft a comprehensive field from investigation including investment fields, Including the biology also bioinformatics [2]. Usually, Association is one among the essential approaches of data mining that's wont to conclude the acquainted patterns, new relationships between groups of data things within the data repository [4]. [5]. A priori algorithm is one of the association algorithms that are generally adopted in care management.

## **2.4 REGRESSION**

Regression is applied to get out duties that demonstrate the association between various variables. An analytical form is created by applying the training dataset. In analytical modeling, two sets from variables called dependent variable, also an independent variable which regularly represented using 'Y' and 'X.' Regression is widely used in the pharmaceutical area for foretelling the diseases[4]. Many regression methods used in healthcare control.

## **2.5 RELATED WORK BASED ON DATA MINING**

In [7], they proposed and developed a system to verify and observance coronary artery illness. They need to use Cleveland heart dataset that it was received from UCI. They need to be enforcement their applied to 303 cases with thirteen features choose from seventy-six features obtainable within the dataset. Two experiences a look at (Holdout experience and Cross-Validation experience) are performed mistreatment 3 algorithms Known as, Bayes Net, Support vector machine (SVM), and Functional Trees (FT) for live a system's detection ability apply WEKA tool. Throughout the primary experience, SVM produced 88.3% accuracy whereas, throughout the second experience, Bayes Net and SVM both produced 83.8% efficiency and FT produced 81.5% accuracy. Then, cross-validation experience is enforcement once more mistreatment seven best feature that is Select using the best initial choice algorithm. When the feature reduction step the accuracy is increased to be 84.5%, 85.1%, and 84.5% for Bayes Net, FT, and SVM, severally.

In [8], heart diseases are diagnosed apply Naïve Bayes algorithm. The utilized dataset is received by one amongst this leading diabetic study academy in Chennai. They used information collection includes five hundred cases. In their experiments, they used the WEKA tool with seventieth of proportion split so as to perform the classification method. The results have shown that Naive Bayes has 86.419% accuracy.

In [9], they aimed to examine the completion of logistic artificial neural networks (ANNs), regression and decision tree guides to foretelling diabetes or prediabetes apply ordinary danger circumstances. Research about two societies in Guangzhou, China, seven hundred and thirty-five cases proved they own diabetes. Or Prediabetes and seven hundred and fifty-two right. Directions did. Recovered. A regular survey continued commanded to receive knowledge about demographic features, people diabetes past, anthropometric dimensions including lifestyle danger agents. Later they advanced 3 predictive patterns apply twelve facts variables also 1 output variable of this inquiry knowledge; we judged this 3 patterns concerning that precision, specificity, and sensitivity. The logistic regression pattern performed a classification. efficiency regarding 76.13% by a sensitivity regarding 79.59% and a specificity of 72.74%. The Artificial Neural Network pattern gave a classification efficiency regarding 73.23% by a sensitivity regarding 82.18% and a specificity regarding 64.49%, also the decision tree (C5.0) gave a classification accuracy regarding 77.87% by a sensitivity regarding 80.68% also the specificity regarding 75.13%. The decision tree pattern (C5.0) owned the highest classification efficiency, supported use of each logistic regression pattern also each ANN produced the ominous intelligence.

In [10], the new method applies a comparative of Genetic Algorithm (GA) and decision trees for the diagnosis of diabetes was given. A comparative of the C4.5 decision tree and GA models were applied to improve the accuracy, rate, and the diagnosis of diabetes. In the decision tree, feature identification and choice in every node states mean that the feature is extra effective than others in data choice. Accordingly, the user can get the final judgment more accurate and quicker. In this research, Pima data of decision tree including 768 people by 8 features were assessed. The suggested method has produced 89.7% identification accuracy.

In [11], the primary purpose of their research is to analyze this achievement from methods those are used to foretell diabetes by applying data mining methods. They have compared machine learning classifiers (K-Nearest Neighbors, J48 Decision Tree, and Random Forest, S V M) to distinguish inpatients by diabetes mellitus. Those methods own experimented among data units taken from the UCI machine learning data container. Achievements about those algorithms become contained inside both the states, i.e., a dataset with noisy data (before pre-processing) a

dataset set out noisy data (after pre-processing) and compare regarding Efficiency, Specificity, and Sensitivity.

In the state that is before pre-processing, the decisions have shown that the decision tree J48 classifier produces a greater precision of 73.82 % than the 3 classifiers. In another state that is later pre-processing the dataset, both KNN (k=1) and Random Forest execution important higher than those extra 3 classifiers and they provide 100% efficiency.

In [12], they attempted to predict the liver infection applying Naïve Bayes and SVM Classification algorithms. They have applied the ILPD dataset of UCI that involves 560 status and 10 attributes. The review from the couple algorithms is included in terms of both accuracy and performance time. The MATLAB is applied in the implementation. The test events have given that SVM has the accuracy of 79.66%. While Naive Bayes has the accuracy of 61.28%.

In [13], they proposed a technique for classifying liver cases applying a dataset received of UCI. They have applied decision tree, MLP, SVM, RF and Bayesian Network classifiers by the WEKA tool. Following feature choice, the test events have given that the accuracies are 69.1252%, 70.669%, 70.8405%, 70.8405%, and 71.8696% for J48, SVM, MLP, Bayesian Network, and Random Forest, respectively.

In [14] they have suggested a technique for classifying breast cancer into benign or malignant. They have applied MLP (multilayer perceptron) and the RBF (radial basis function). The applied dataset is the Wisconsin Diagnostic and Prognostic Breast Cancer datasets as feature choice (of UCI) which includes 699 cases each by 11 features. The tests have been performed at 683 states just because the neglected states have lost data. The test results have confirmed that MLP has 88% accuracy while RBF has 97% accuracy.

In [15], their chief aim is to compare the production of another data mining techniques in breast cancer prediction. The applied dataset is the Wisconsin Diagnostic and Prognostic Breast Cancer datasets for feature choice (of UCI) which includes 699 cases every by 11 features. In order to use their tests, they have used the Rapid Miner tool. The used algorithms are discriminant analysis, multilayer perceptron MLP, Logistic Regression, Decision tree, NB, Supper vector

machine, and K-NN. This test events have given that the accuracies of the several algorithms are 94.15%, 96.2%, 94.8%, 96%, 96.5%, 96.2%, and 95.5% for discriminant analysis, multilayer perceptron MLP, SVM, Naïve Bayes, Decision tree, Logistic Regression, and K-NN, respectively.

In [16], their chief aim is to compare the production from both Naive Bayes and back propagation classifiers in classifying hepatitis disease (A, B, C, and D, also E). The applied dataset is received of UCI and includes 155 states among 20 features several. Truly, just 50 features are applied to their job. The test events have given that the accuracies are 97% and 98% for the Naive Bayes classifier and Back propagation classifier, respectively.

In [17], the researchers have applied C4.5, ID3, and the CART algorithms to diagnose some hepatitis infection. The accepted dataset is received of UCI. They have adopted the WEKA tool in their trials. The trial events have conferred that the accuracies are 83.2%, 64.8%, and 71.4% for the CART, ID3, and C4.5, respectively.

## 2.6 SUMMARY

**Table 2.1:** A summary of some related work based on data mining techniques

Ref. Num	Year	Objective	Dataset/Source	Data Mining Techniques	Accuracy
[11]	2015	Coronary artery disease detection	Cleveland heart dataset / UCI	Bayes Net SVM Functional Tree	84.5% 85.1% 84.5%
[12]	2015	Heart disease diagnosing	-/ research institute in Chennai	Naive Bayes	86.419%
[13]	2013	predicting diabetes or prediabetes	-/two communities in Guangzhou, China	Logistic Regression, ANN Decision Tree (C5.0)	76.13% 73.23% 77.87%
[14]	2016	Diagnosis of diabetes	Pima dataset/UCI	A combination of GA and C4.5 decision tree models	89.7%
[15]	2015	predicting diabetes	--/UCI	J48 Decision Tree K-Nearest Neighbors Random Forest	86.46% 100% 77.73%
[16]	2015	Predicting the liver disease	ILPD dataset/ UCI	SVM Naïve Bayes	79.66% 61.28%
[17]	2014	Classifying liver patients	--/UCI	J48 MLP SVM Random Forest Bayesian Network	70.669% 70.840% 70.840% 71.869% 69.125%
[18]	2012	Classifying breast cancer into benign or malignant	Wisconsin Diagnostic dataset /UCI	MLP RBF	88% 97%
[19]	2014	Breast cancer prediction	Wisconsin Diagnostic dataset /UCI	SVM Naïve Bayes K-NN	86.5% 86.2% 85.5%



[20]	2011	Classifying hepatitis disease into A, B, C, D, E	--/UCI	Naive Bayes Back propagation	97% 98%
[21]	2011	diagnose the hepatitis disease	--/UCI	CART ID3 C4.5	83.2% 64.8% 71.4%

As shown in Table 2.1. We see the related work based on data mining techniques and see the year of the research, Objective, Dataset/Source, Data Mining Techniques type and the Accuracy for every Technique.



### **3. COMPARATIVE STUDY**

Heart illness is the principal purpose of mortality globally. Heart illness is also the disease with the most significant disease in current global measures by The (WHO). World Health Organization and recognizes no decrease in mortality induced by heart illness tomorrow. New research foretells heart illness to be the first circumstances of dying in 2030.

#### **3.1 DATASET**

The dataset includes four data containers concerning heart illness examination. All properties are numeric-valued. These data were gathered of these locations:

1. Clinic Foundation in Cleveland (used)
2. University Hospital in Switzerland
3. Institute of Cardiology in Budapest
4. Medical Center

##### **3.1.1 Properties of Dataset**

The 14 attributes included in the Cleveland data set are:

1. Age
2. Gender
3. Chest pain type
4. Blood pressure when resting
5. Cholesterol
6. Blood sugar when fasting
7. Electrocardiographic results when resting
8. Maximum heart rate
9. Exercise induced angina
10. ST depression induced by exercise relative to rest

11. The slope of the peak exercise for the ST depression segment
12. Number of major vessels colored by flouroscopy
13. Heart status (Normal, defect)
14. Predicted status (Yes/No heart disease)

### **3.2 DATA MINING TECHNIQUES**

Data-mining is the identified title for all agents that can be utilized when seeking for associations and biases in massive volumes of data, mainly used on data recording no such patterns when assessed by the private eye. The confidence of data-mining is to be capable of deducing information and making an understanding of massive volumes of data.

The statistically meaningful correlations between data objects that are obtained with data-mining, frequently referenced as a classifier and will be from now on in this statement, can be utilized to new data and the possibility of every result can be obtained.

The classifier is constructed using a collection of practice data. Training data is data with a previously identified result for the recommended research. After the classifier has been determined, it is essential that it is utilized to information that was not a piece of the training data. The assessment could differently result in wrongly great efficiency for the classifier. An instance of practice data could be a database including inmates and their pharmaceutical studies. The inmates all have the identical information reported and not just random data. The inmates also have a yes/no determination for a particular illness, e.g., the heart infection.

### **3.2.1 Decision Tree Algorithm**

Decision Trees does a procedure so generally applied in data-mining. The purpose signifies to generate the collection of dictates which can foretell the particular issue variable depended on a set from input data. A Decision Tree depends on peaks and edges. These sides express a way or a decision-influencing on that subsequent vertices.

### **3.2.2 J48 Technique**

J48 is a public Java as a programming language development concerning the C4.5 method, applied while a data mining software WEKA. Each C4.5 technique is an expansion on this ID3 algorithm also is utilized to start a Decision Trees that can be applied as categorization.

### **3.2.3 Naive Bayesian technique**

The Bayesian assignment is similarly applied in data mining like Decision Trees and can forecast the possibility of the class group. The Bayesian assignment technique depends on. Bayes Theorem also is several usually applied in machine.learning. There are various distinct variants from the Bayesian distribution where Naive-Bayes is the most popular. It has been and still is a prevalent method to apply when performing.spam.filters and different varieties about writing categorization.

## **3.3 WEKA SOFTWARE**

Waikato Environment for Knowledge Analysis. (WEKA) is a public-implementation reference software programmed into Java received at the University of Waikato in New Zealand. That is a pre-programmed of data mining methods without owning to perform each technique principles of scratch. WEKA has provided RF, J48, perceptrons and Naive Bayesian.

### 3.3.1 Data Input in WEKA program

WEKA claims that an input repository does both a. csv records or. arff records. We obtained the original Heart illness data records (numeric values) from the Heart illness original source and converted it into a arff-records.

```
% property 'gender'
% value: 0      means: female
% value: 1      means: male

% property 'cp'
% value: 1      means: typ angina
% value: 4      means: asympt
% value: 3      means: non anginal
% value: 2      means: atyp angina

% property 'fbs'
% value: 1      means: t
% value: 0      means: f

% property 'restecg'
% value: 0      means: normal
% value: 1      means: abnormality
% value: 2      means: hyper

% property 'exang'
% value: 0      means: no
% value: 1      means: yes
```

% property 'slope'

% value: 1 means: up

% value: 2 means: flat

% value: 3 means: down

% Relabeled values into attribute 'thal'

% value: 6 means: fixed defect

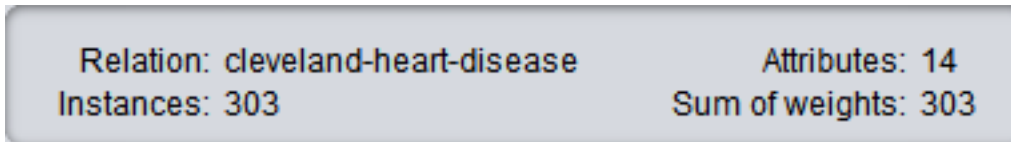
% value: 3 means: normal

% value: 7 means: reversable defect

% Relabeled values into attribute 'num'

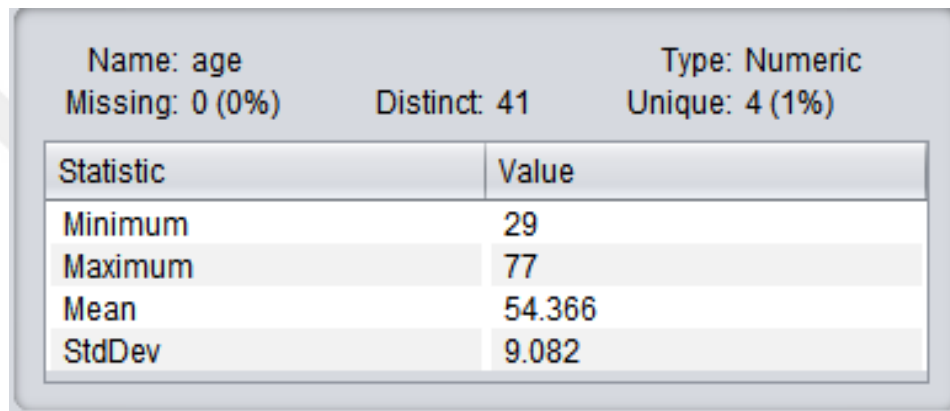
% value: '0' means: '<50'

% value: '1' means: '>50 '



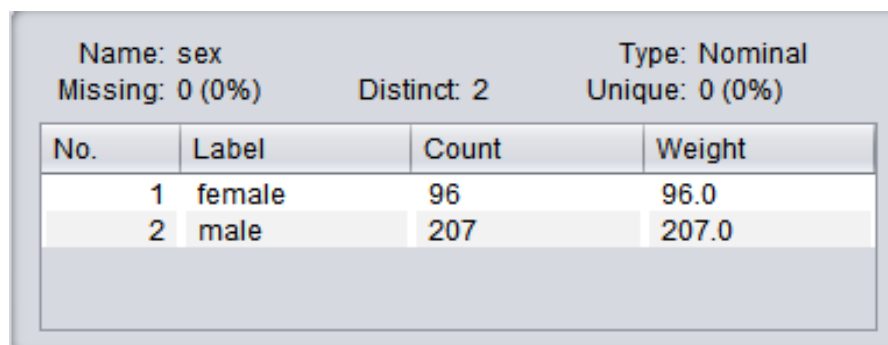
**Figure 3.1:** Number of instances of heart disease Dataset using WEKA program

As shown in figure 3.1, the dataset contains 303 records (patients) with 14 properties



**Figure 3.2:** Statistical values of Age attribute using WEKA program

As shown in figure 3.2. explain the Statistical values, minimum age 29 years old to maximum age 77 years old as presented, the mean (Average of data), standard deviation, Distinct (How many the data repeated(41)), Unique (the data without repeated(4)) and missing data(0)



**Figure 3.3:** Statistical values of Sex attribute using WEKA program

As shown in figure 3.3. This dataset contains 96 female patients and 207 male patients Distinct (How many the data repeated (2)), Unique (the data without repeated (0)) and missing data (0)

Name: cp		Type: Nominal	
Missing: 0 (0%)		Distinct: 4	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	typ_angina	23	23.0
2	asympt	143	143.0
3	non_anginal	87	87.0
4	atyp_angina	50	50.0

**Figure 3.4:** Statistical values of Chest Type attribute using WEKA program

As shown in figure 3.4, the dataset contains 303 records (patients) with chest pain in different types (23 patients have typical angina type, 143 patients have asympt, 87 patients have non anginal (non patient) and 50 patients have a typical angina). Distinct (How many the data repeated (4)), Unique (the data without repeated (0)) and missing data (0)



### 3.3.2 J48 result from WEKA program

Time taken to build model 0.13 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	235	77.5578 %
Incorrectly Classified Instances	68	22.4422 %
Kappa statistic	0.5443	
Mean absolute error	0.1044	
Root mean squared error	0.2725	
Relative absolute error	52.0476 %	
Root relative squared error	86.5075 %	
Total Number of Instances	303	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
	0.830	0.290	0.774	0.830	0.801	0.546	0.809	0.767
	0.710	0.170	0.778	0.710	0.742	0.546	0.809	0.779
Weighted Avg.	0.776	0.235	0.776	0.776	0.774	0.546	0.809	0.772

Figure 3.5: J-48 Result using WEKA program

**Table 3.1:** J48 Result from WEKA program

<b>Result</b>	<b>Values</b>
Correctly Classified Instances	235
In Correct Classified Instances	68
TP Rate	0.776
FP Rate	0.235
Precision	0.776
Recall	0.776
ROC Area	0.809
Time	0.13 Second

In this experiment, we conduct experiment using WEKA software and applied J-48 technique on our dataset and conduct result as shown in figure 3.5. We have 235 correctly classified patients out of 303 records with 77% precision, 77% recall and accuracy 80.9% in 0.13 second as discussed in table 3.1.

### 3.3.3 Naive Bayes result from WEKA program

```
Time taken to build model 1.07

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 253      83.4983 %
Incorrectly Classified Instances 50      16.5017 %
Kappa statistic 0.6661
Mean absolute error 0.0738
Root mean squared error 0.2299
Relative absolute error 36.8026 %
Root relative squared error 72.9665 %
Total Number of Instances 303

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area
          0.867   0.203   0.836     0.867   0.851     0.667  0.904
          0.797   0.133   0.833     0.797   0.815     0.667  0.904
Weighted Avg.  0.835   0.171   0.835     0.835   0.835     0.667  0.904
```

Figure 3.6: Naive Bayes Result using WEKA program

**Table 3.2:** Native Bayes Result from WEKA program

<b>Result</b>	<b>Values</b>
Correctly Classified Instances	253
In Correct Classified Instances	50
TP Rate	0.835
FP Rate	0.171
Precision	0.835
Recall	0.835
ROC Area	0.904
Time	1.07 second

In this experiment, we conduct experiment using WEKA software and applied Naive Bayes technique on our dataset and conduct result as shown in figure 3.6. We have 253 correctly classified patients out of 303 records with 83% precision, 83% recall and accuracy 90.4% in 1.07 second as discussed in table 3.2.

### 3.3.4 Multilayer Preceptron (Neural Network) result

---

Time taken to build model: 1.04 seconds

=== Stratified cross-validation ===  
=== Summary ===

Correctly Classified Instances	245	80.8581 %
Incorrectly Classified Instances	58	19.1419 %
Kappa statistic	0.6141	
Mean absolute error	0.0772	
Root mean squared error	0.2544	
Relative absolute error	38.477 %	
Root relative squared error	80.7429 %	
Total Number of Instances	303	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
	0.824	0.210	0.824	0.824	0.824	0.614	0.894	0.903
	0.790	0.176	0.790	0.790	0.790	0.614	0.889	0.878
Weighted Avg.	0.809	0.194	0.809	0.809	0.809	0.614	0.891	0.892

Figure 3.7: Neural Network Result using WEKA program

**Table 3.3:** Multi-layer perceptron Result from WEKA program

<b>Result</b>	<b>Values</b>
Correctly Classified Instances	245
In Correct Classified Instances	58
TP Rate	0.809
FP Rate	0.194
Precision	0.809
Recall	0.809
ROC Area	0.891
Time	1.04 second

In this experiment, we conduct experiment using WEKA software and applied Neural Network (perceptron) technique on our dataset and conduct result as shown in figure 3.7. We have 245 correctly classified patients out of 303 records with 80% precision, 80% recall and accuracy 89% in 1.04 second as discussed in table 3.3.

### 3.3.5 SVM (Support Vector Machine) Result

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	255	84.1584 %
Incorrectly Classified Instances	48	15.8416 %
Kappa statistic	0.678	
Mean absolute error	0.1805	
Root mean squared error	0.2873	
Relative absolute error	89.9636 %	
Root relative squared error	91.1884 %	
Total Number of Instances	303	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	FRC Area
	0.897	0.225	0.827	0.897	0.860	0.681	0.836	0.798
	0.775	0.103	0.863	0.775	0.817	0.681	0.836	0.771
Weighted Avg.	0.842	0.169	0.843	0.842	0.841	0.681	0.836	0.786

Figure 3.8: SVM Result using WEKA program

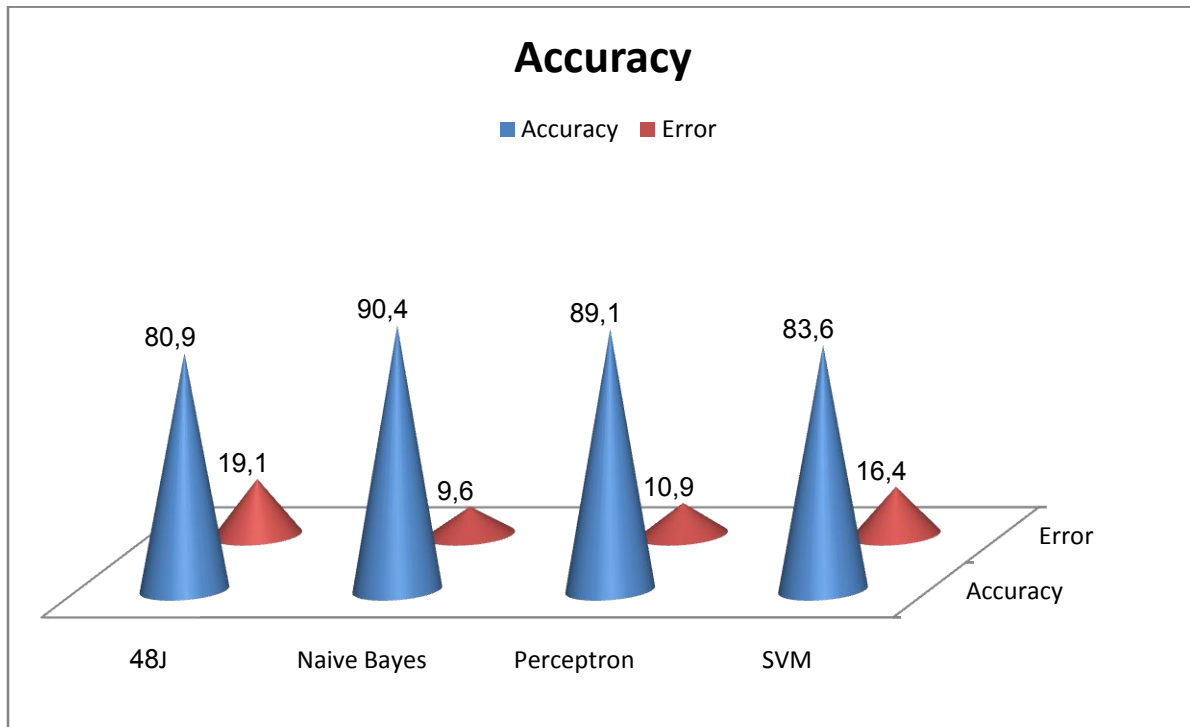
**Table 3.4:** SVM Result from WEKA program

<b>Result</b>	<b>Values</b>
Correctly Classified Instances	255
In Correct Classified Instances	48
TP Rate	0.840
FP Rate	0.169
Precision	0.843
Recall	0.842
ROC Area	0.836
Time	0.15 second

In this experiment, we conduct experiment using WEKA software and applied SVM technique on our dataset and conduct result as shown in figure 3.8. We have 255 correctly classified patients out of 303 records with 84% precision, 84% recall and accuracy 83% in 0.15 second as discussed in table 3.4.

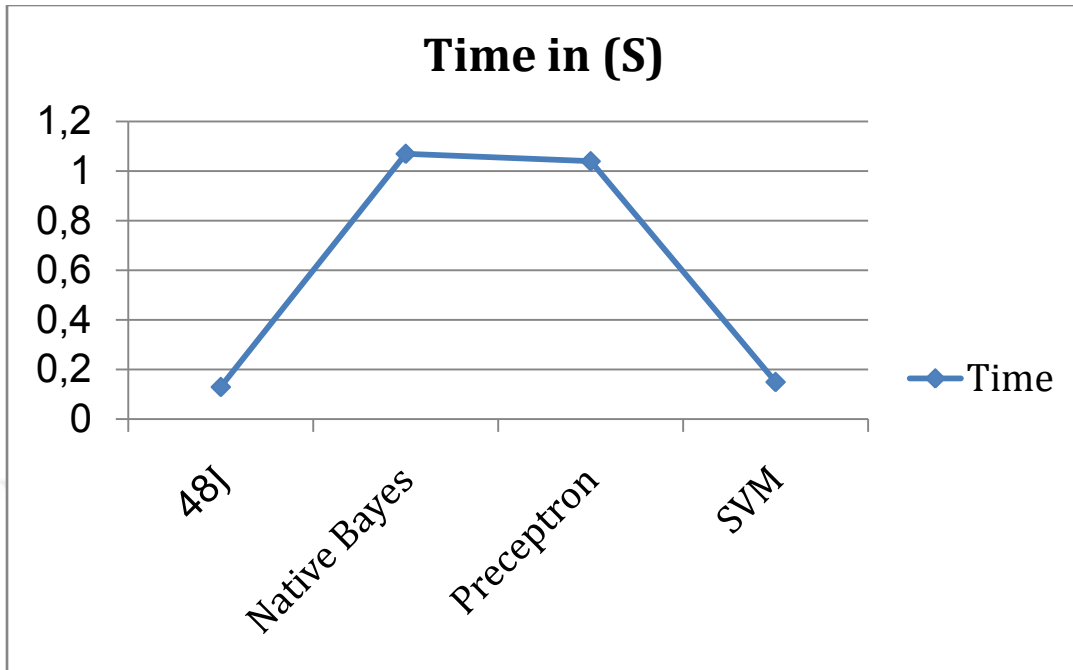


### 3.4 SUMMARY



**Figure 3.9:** Accuracy and Error Diagram of Comparative Study between Data Mining techniques used in WEKA Program

As conclusion from these experiments as shown in figure 3.9 and figure 3.10, Naive Bayes has a highest performance but it is the slower one. However, J-48 is the fastest one but the lowest performance. So, we need to balance between performance and time consumption.



**Figure 3.10:** Time Diagram of Comparative Study between Data mining techniques used in WEKA Program

## 4. PROPOSED FRAMEWORK

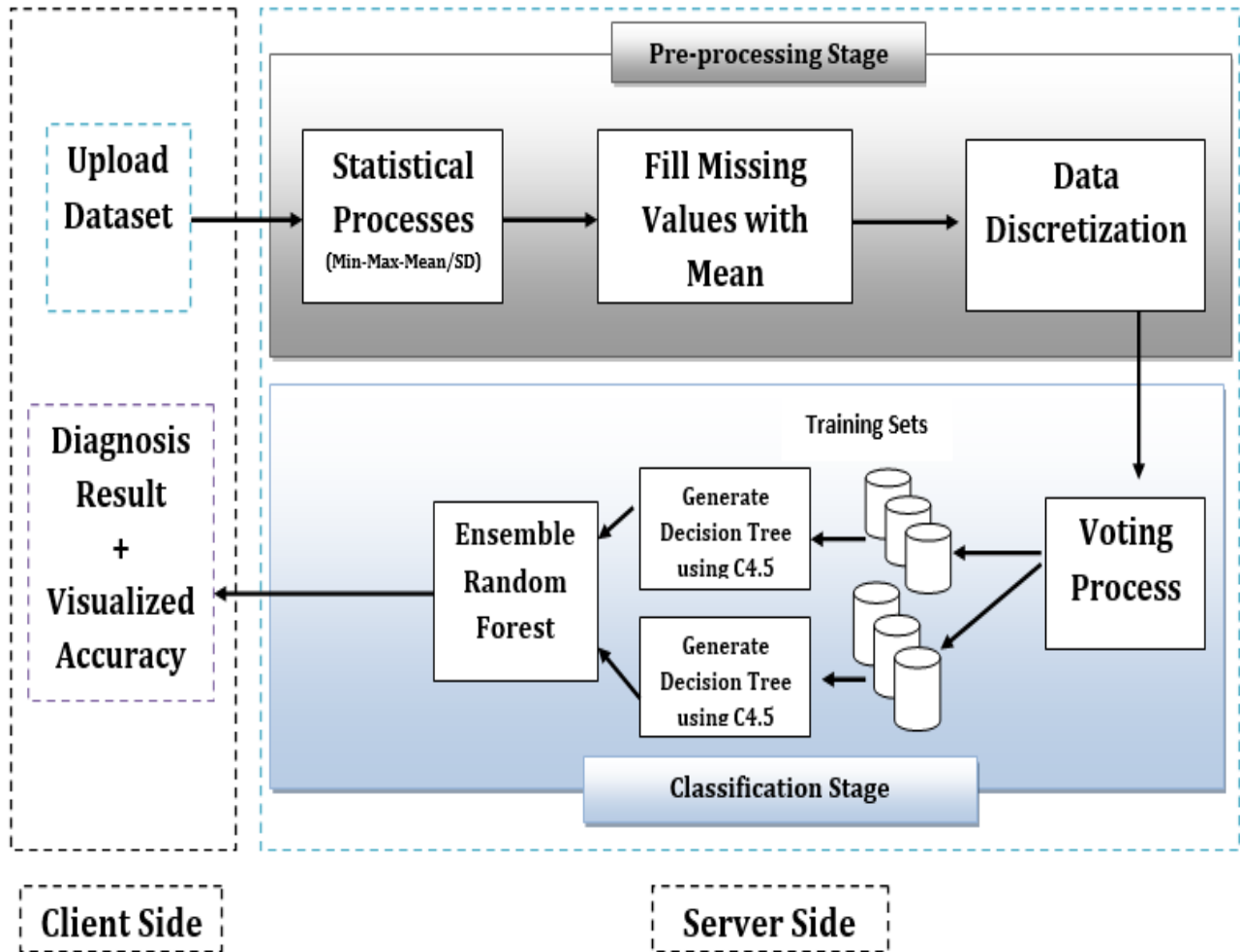


Figure 4.1: Proposed Framework

## 4.1 FRAMEWORK DESCRIPTION

We have designed a simple proposed framework for evaluating the hybrid algorithms as depicted in Fig. 4.1.

### 4.1.1 Pre-processing Stage

- A. Statistical Processing
  - Compute Min, Max, Mean and Standard Deviation
- B. Filling Missing value
  - Any missing values would be replaced with Mean value, or archived value
- C. Data Discretization
  - We used an approximate equal interval binning method to bin the data variables into a small number of categories.
  - Entropy is an information theoretical equation that calculate of the ‘uncertainty’ involved in a training set. It assesses applicant cut arguments over an entropy-based process to choose limits for discretization.

### 4.1.2 Hybrid Classification processing Stage

- A. Sampling and voting
  - Numerous classifiers elective includes separating the training data in reduced equal subsections from data and construct a Decision Tree structure for each subsection of data. Election is based on variety or majority voting. Every discrete classifier donates a solo vote.
- B. Decision Tree
  - In the decision tree method, we need to pick the excruciating feature that reduces the value from entropy and exploiting the Information Gain. To recognize an excruciating feature from the Decision Tree, should compute the Information Gain to every feature also then choose a feature that exploits an Information Gain.

$$E = \sum_{i=1}^k -P_i \log_2 P_i \quad 4.1)$$

Where

- k is that value from classes of this objective feature
- Pi is a value of incidences from class i separated via the whole value from occurrences

### C. Ensemble Random Forest

- A collaborative process that associates the forecasts from composed numerous machine learning methods together to make more correctness of forecasts than any distinct model.
- Bootstrap Aggregation is a universal process that can be applied to decrease the adjustment for those methods that have high adjustment.
- Random Forests are an enhancement over bagged decision trees.

## 4.2 PREPROCESSING IMPLEMENTATION USING PHP CODE

```
class parsing{
public function array_parsing(){
    $handle = fopen("dataset/cleveland.dat", "r");
    $i=0;
    if ($handle) {
        while (($line = fgets($handle)) !== false) {
            list($age,$sex,$cp,$trestbps,$chol,$fbs,$restecg,$thalach,$sexang,$oldpeak,$slope,$sca,$thal,$num) = explode(",",$line);

            $newLine='$data['.$i.']= array(';

            $newLine .= "'age' =>".$age.", ";

            if( intval($sex)==0)
                $newLine .= "'sex' =>'female', ";

            else if(intval($sex)==1)
                $newLine .= "'sex' =>'male', ";

            if( intval($cp)==1)
                $newLine .= "'cp' =>'typ_angina', ";

            else if(intval($cp)==2)
                $newLine .= "'cp' =>'atyp_angina', ";

            else if(intval($cp)==3)
                $newLine .= "'cp' =>'non_anginal', ";

            else if(intval($cp)==4)
                $newLine .= "'cp' =>'asympt', ";

            $newLine .= "'trestbps'=>".$trestbps, ";
            $newLine .= "'chol'=>".$chol, ";
```

**Figure 4.2:** PHP Pre-processing code

## 4.3 DECISION TREE IMPLEMENTATION USING PHP CODE

### 4.3.1 Node Classes of Decision Tree

```
class DT_Data{
    var $number;
    var $match;
    var $unmatch;

    var $split_key;
    var $split_value;
}

class DT_Node{
    var $dtdata;
    var $left;
    var $right;
    var $terminal;
}
```

### 4.3.2 Main Classes of Decision Tree

```
function Decision_Tree($data)
{
    $this->data      = $data;
}
public function classify($base_key,$base_value,$false_value){

    $this->base_key    = $base_key;
    $this->base_value  = $base_value;
    $this->>false_value = $false_value;

    $this->bv_data = Decision_Tree::make_binary_variable_data($this->data,$base_key);

    $tree = Decision_Tree::make_decision_tree($this->bv_data,
    $this->base_key,
    $this->base_value,
    $this->base_key,
    $this->base_value);

    $this->tree = $tree;
    return $tree;
}
```

```

private function make_decision_tree($data,$base,$base_value,$split_key,$split_value){

    $delta_I_array = array();
    $dtnode = new DT_Node();
    $dtdata = new DT_Data();

    $dtdata = Decision_Tree::set DtData($data,$base,$base_value);
    $dtdata->split_key    = $split_key;
    $dtdata->split_value = $split_value;
    $dtnode->dtdata      = $dtdata;
    $dtnode->terminal    = false;

    $keys = array_keys($data[0]);
    foreach ($keys as $k => $key) {
        if ($key == $base) {
            continue;
        }
        $delta_I_array[$key] = CART::calc_delta_I($data,$base,$key);
    }

    $flg =0;
    foreach ($delta_I_array as $key => $value) {
        if($value != 0.0){$flg=1;}
    }
    if($flg==0){
        $dtnode->terminal= true;
        return $dtnode;
    }

    $split_key = array_keys($delta_I_array,max($delta_I_array));
}

```

**Figure 4.3:** PHP Decision tree code



## 4.4 RANDOM FOREST IMPLEMENTATION USING PHP CODE

```
function sample_data($data, $n = 0)
{
    $count = count($data);

    if ($n <= 0) {
        $n = $count;
    }

    $result = array();
    for ($i = 0; $i < $n; ++$i) {
        $result[] = $data[mt_rand(0, $count - 1)];
    }

    return $result;
}

function random_forest_classify($trees, $data)
{
    $candidates = array_map(
        function ($t) use ($data) { return $t->species($data); },
        $trees);

    return vote($candidates);
}
```

Figure 4.4: PHP Random Forest code

## 5. EXPERIMENTAL RESULT

In the previous chapter, the proposed hybrid ensemble classification algorithm (Decision Tree + Random forest) was introduced. In this chapter, we need comparison between the proposed hybrid ensemble classification algorithm and existing algorithms are needed to confirm it is efficiency. Lastly, the experiment of the proposed framework is needed to confirm it is usability, reliability, and efficiency.

### 5.1 DEVICE AND TOOL CAPABILITIES

**Table 5.1:** Tools and device used to preform proposed framework

Metric	Values
CPU	Intel core i7
RAM	4G
Operating system	Windows 10
Programming Language	PHP v4
Server Platform	Apache server

As shown in Table 5.1. We see the experiment that was used like operating system, programming Language and the server Platform

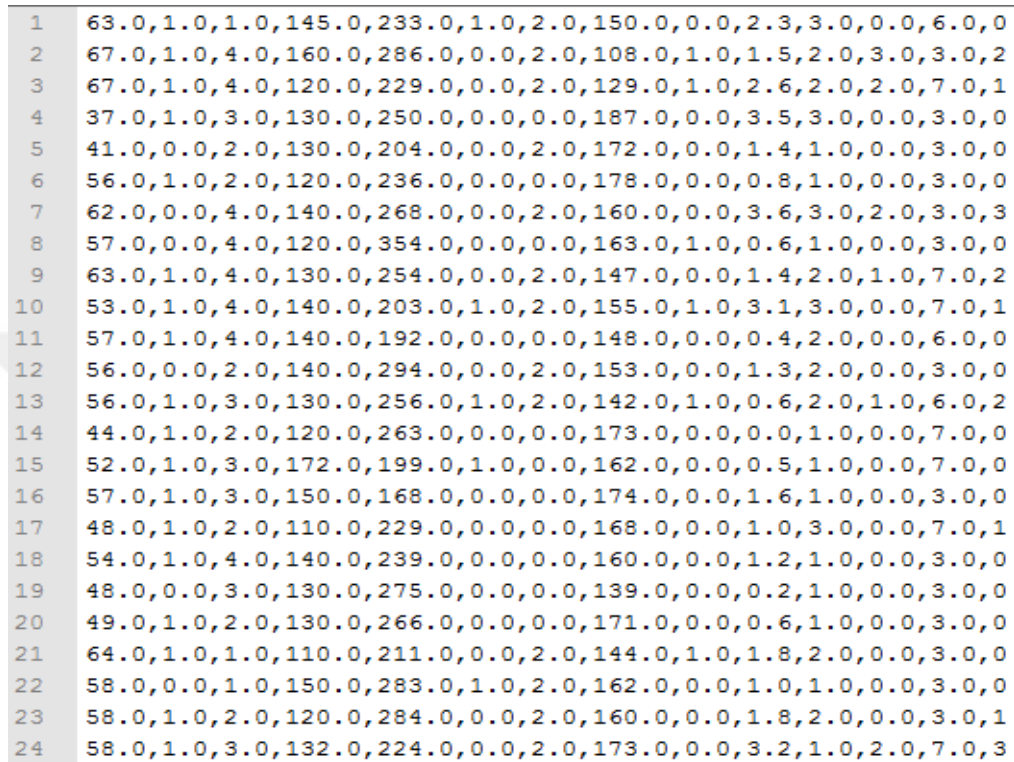
### 5.2 IMPLEMENTATION SCREEN SHOTS

The dataset includes four data containers concerning heart illness examination. All properties are numeric-valued. These data were gathered of these locations:

1. Clinic Foundation in Cleveland (used)
2. University Hospital in Switzerland
3. Institute of Cardiology in Budapest
4. Medical Center

We use the first dataset with 14 attributes and it has a few missing values and all records had a complete digaoosis values.

## 5.2.1 Dataset Screen Shots

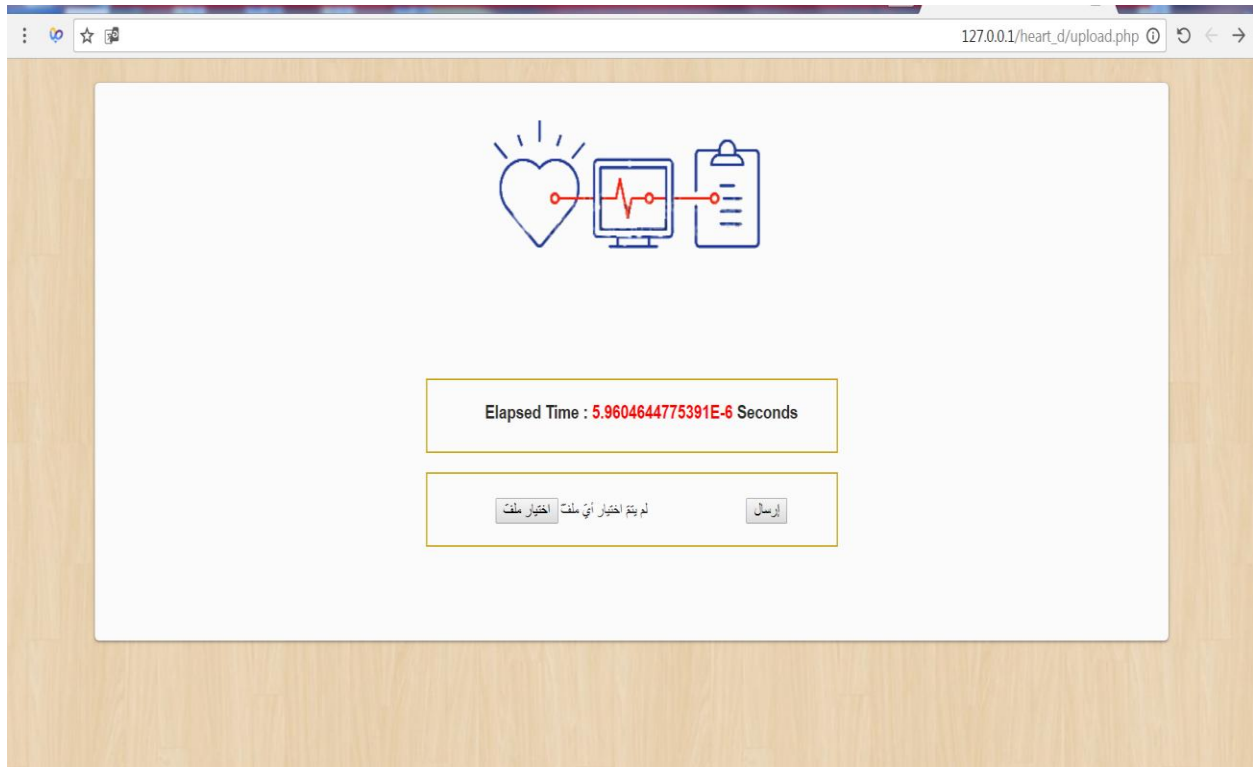


1	63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
2	67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
3	67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
4	37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
5	41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
6	56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
7	62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
8	57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
9	63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
10	53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1
11	57.0,1.0,4.0,140.0,192.0,0.0,0.0,148.0,0.0,0.4,2.0,0.0,6.0,0
12	56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0,0
13	56.0,1.0,3.0,130.0,256.0,1.0,2.0,142.0,1.0,0.6,2.0,1.0,6.0,2
14	44.0,1.0,2.0,120.0,263.0,0.0,0.0,173.0,0.0,0.0,1.0,0.0,7.0,0
15	52.0,1.0,3.0,172.0,199.0,1.0,0.0,162.0,0.0,0.5,1.0,0.0,7.0,0
16	57.0,1.0,3.0,150.0,168.0,0.0,0.0,174.0,0.0,1.6,1.0,0.0,3.0,0
17	48.0,1.0,2.0,110.0,229.0,0.0,0.0,168.0,0.0,1.0,3.0,0.0,7.0,1
18	54.0,1.0,4.0,140.0,239.0,0.0,0.0,160.0,0.0,1.2,1.0,0.0,3.0,0
19	48.0,0.0,3.0,130.0,275.0,0.0,0.0,139.0,0.0,0.2,1.0,0.0,3.0,0
20	49.0,1.0,2.0,130.0,266.0,0.0,0.0,171.0,0.0,0.6,1.0,0.0,3.0,0
21	64.0,1.0,1.0,110.0,211.0,0.0,2.0,144.0,1.0,1.8,2.0,0.0,3.0,0
22	58.0,0.0,1.0,150.0,283.0,1.0,2.0,162.0,0.0,1.0,1.0,0.0,3.0,0
23	58.0,1.0,2.0,120.0,284.0,0.0,2.0,160.0,0.0,1.8,2.0,0.0,3.0,1
24	58.0,1.0,3.0,132.0,224.0,0.0,2.0,173.0,0.0,3.2,1.0,2.0,7.0,3

**Figure 5.1:** Data Set from UCI location

As shown in figure 5.1, screen shot for the data that was taken from UCI location we see the data just numbers

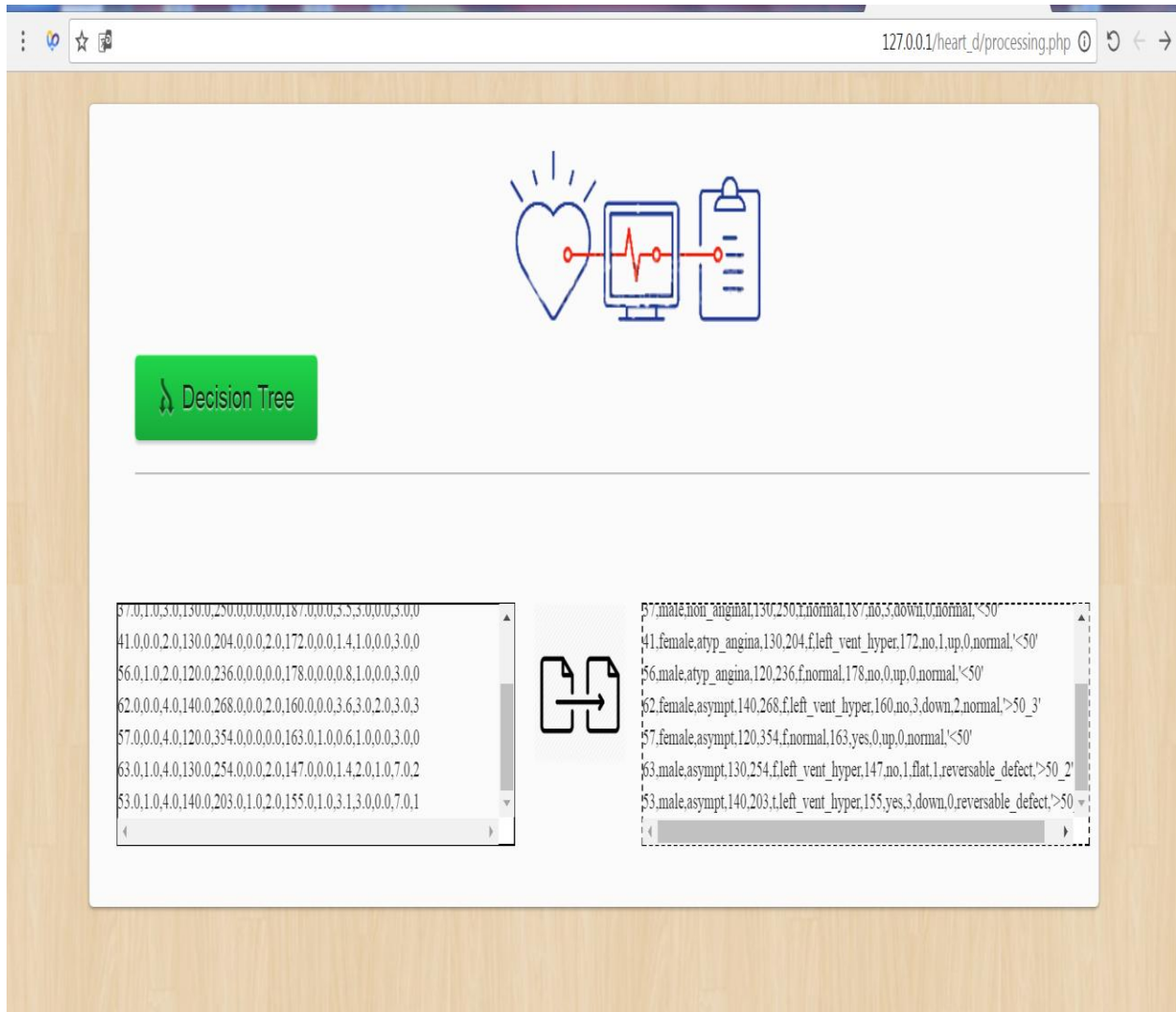
## 5.2.2 Uploading Screen Shots



**Figure 5.2:** Uploading Data Set to the server

As shown in figure 5.2, screen shot for the data when making uploading to the server

### 5.2.3 Pre-processing stage Screen Shots



**Figure 5.3:** Pre-processing for data set in the server

As shown in figure 5.3, screen shot for the data when making pre-processing filling missing value and making discrete for data in the server

## 5.2.4 Decision Tree stage Screen Shots

127.0.0.1/heart\_d/f48.php

SVM Random Forest

Relation Name	Num Attributes	Training Time	Testing Time	Tree Size	Leaves Number
heart-disease	14	0.39 seconds	0.01 seconds	227	127

Correct Classified Data	InCorrect Classified Data	True Positive	False Positive	ROC
694 79.496%	179	0.795	0.124	0.912

exang = no  
| chol <= 0  
| | cp = typ\_angina

```
127.0.0.1/heart_d/f48.php

exang = no
| chol <= 0
| | cp = typ_angina
| | | trestbps <= 125: >50_1 (2.04/0.04)
| | | trestbps > 125: >50_2 (3.0)
| | cp = asympt
| | | thalach <= 123
| | | | slope = up
| | | | | age <= 55: >50_4 (3.0/1.0)
| | | | | age > 55: >50_1 (2.0/1.0)
| | | | slope = flat
| | | | | age <= 52
| | | | | | age <= 50: <50 (2.58/1.0)
| | | | | | age > 50: >50_2 (6.0/3.0)
| | | | | age > 52
| | | | | | restecg = left_vent_hyper: >50_1 (2.0/1.0)
| | | | | | restecg = normal: >50_1 (10.0/4.0)
| | | | | | restecg = st_t_wave_abnormality: >50_3 (7.0/1.0)
| | | | | slope = down: <50 (2.0/1.0)
| | | | thalach > 123
| | | | oldpeak <= 1
| | | | | fbs = t
| | | | | | trestbps <= 126: >50_1 (2.0/1.0)
| | | | | | trestbps > 126: >50_2 (2.0)
| | | | | fbs = f
| | | | | | trestbps <= 136
| | | | | | sex = female: >50_1 (2.0)
| | | | | sex = male
| | | | | | oldpeak <= 0
```

**Figure 5.4:** Decision Tree form

As shown in figure 5.4, screen shot for Decision Tree after finished the Pre-processing in the server

### 5.3 EXPERIMENTAL RESULT AND DISCUSSION

**Table 5.2:** Confusion Matrix

	Predicted patient with heart disease  (positive)	Predicted Healthy Persons  (negative)
Actual Patient with heart disease	True Predicted Patient as (TP)	False Predicted Person as (FN)
Actual Healthy Persons	False Predicted Patient as (FP)	True Predicted Person as (TN)

As shown in Table 5.2. We see matrix and how divided the data to True predicted, False Predicted, False predicted person and True predicted person to used in formula below.

For training and testing the data sets, we use ten-fold cross validation technique. This technique splits the dataset to 10 portions .9 portions are then applied to training and that tenth fragment is applied for testing. This is recurring, applying the alternative portion to the test section. Individually the data portion is utilized 1 for testing and 9 events for training. This is recurrent 10 events, including a novel portion doing the testing part. The average outcome is produced from the 10 runs. The accuracy of the applied procedures must be evaluated applying under titles about rightly classified instances, wrongly classified instances, FP rate, TP rate, recall, precision, ROC area, CPU Time, accuracy, error. We used several measures to evaluate the methods used on the heart diseases dataset as conferred below:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \quad (5.1)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \times 100\% \quad (5.2)$$

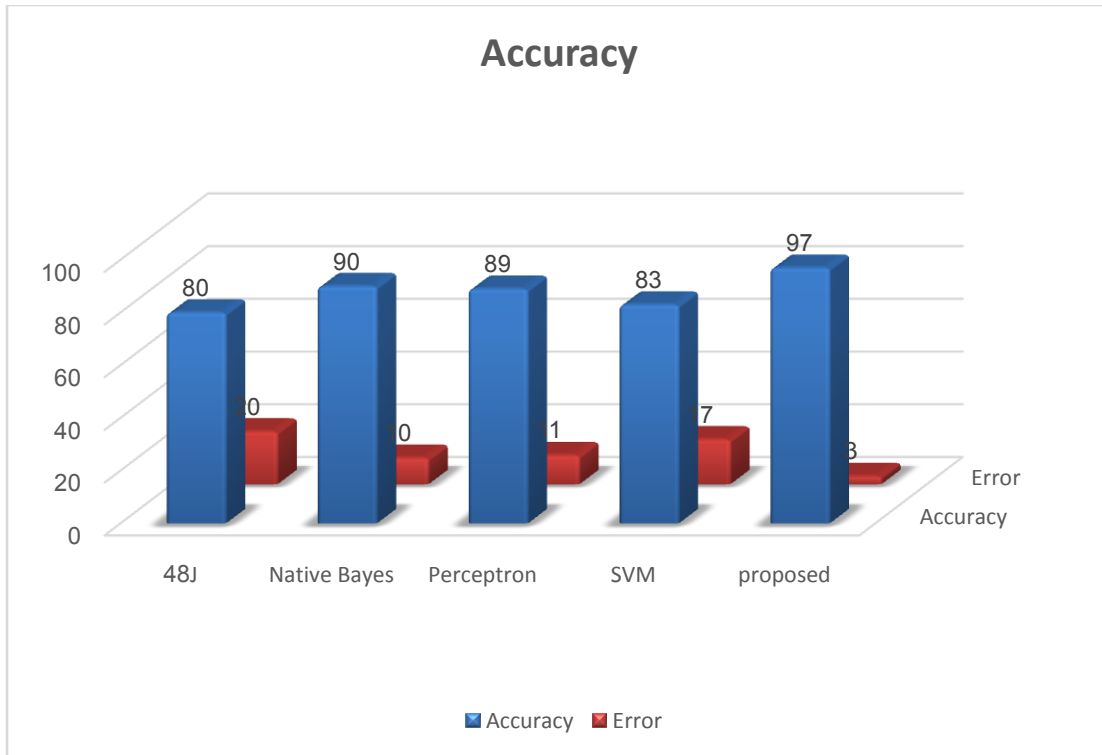
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (5.3)$$



Through analyzing results, we can conclude that a Random Forest classifier is giving that greatest accuracy, reflected by the multi-layer perceptron based classifier then, J48 classifier and lastly the Naïve Bayes classifier. Also, by comparing the results of both experiments we can see that the accuracy of the random forest classifier is enhanced on the large dataset in addition to J48 classifier, while the accuracy of the Naïve Bayes classifier and multi-layer perceptron based classifier is noticeably decreased with enlarging the dataset. Regarding the time, we can see that Naïve Bayes classifier become the best classifier in terms of CPU time instead of J48 classifier, and we can see the table 5.3 with the result, and Accuracy and Error Diagram of Comparative Study between proposed hybrid classification algorithm and other algorithms in figure 5.5

**Table 5.3:** Comparative Result between Proposed hybrid algorithm and other algorithms

Classifier	Sensitivity	Specificity	Accuracy
SVM	84%	84%	83%
J48	77%	77%	80%
Proposed Ensemble (iteration=50)	91%	89%	93%
Proposed Ensemble (iteration=100)	95%	92%	97%



**Figure 5.5:** Accuracy and Error Diagram of Comparative Study between proposed hybrid classification algorithm and other algorithms

## 6. CONCLUSION

Nowadays, data mining methods are playing an essential role in healthcare management. In this document, we perform an overview of some uses of data mining procedures in identifying, diagnosing, and foretelling various conditions and syndromes. Lastly a set of operation was led to assess this accuracy regarding a set of data mining procedures including Decision Trees (j48), Multilayer Perceptron, Naïve Bayes, and Random Forest into heart disease diagnosis. The experimental outcomes have shown that a Random Forest classifier is presenting the best achievement concerning accuracy by the large dataset.

## REFERENCES

- [1] Gupta, S., D. Kumar, and A. Sharma, Performance analysis of various data mining classification techniques on healthcare data. *International journal of computer science & Information Technology (IJCSIT)*, 2011. 3(4).
- [2] Jyoti Sony, Uma Ansari, Dinesh Sharma, Suita Sony "Predictive data mining for medical diagnosis: an overview of heart disease prediction" *International Journal of Computer Science and Engineering*, vol. 3, 2011
- [3] Ahmad, P., S. Qamar, and S.Q.A. Rizvi, Techniques of data mining in healthcare: A review. *International Journal of Computer Applications*, 2015. 120(15).
- [4] Tomar, D., and S. Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Biotechnology* 5, no. 5 (2013): 241-266.
- [5] Patel, S., and H. Patel. "Survey of data mining techniques used in the healthcare domain." *International Journal of Information* 6, no. 1/2 (2016).
- [6] Nagarajan, S., and R. M. Chandrasekaran. "Design and implementation of an expert clinical system for diagnosing diabetes using data mining techniques." *Indian Journal of Science and Technology* 8, no. 8 (2015): 771-776.
- [7] Otoom, A.F., E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour. Effective Diagnosis and Monitoring of Heart Disease. *International Journal of Software Engineering and Its Applications*. 9 (2015): 143-156.
- [8] Vembandasamy, K., Sasipriya, R. And Deepa, E. Heart Diseases Detection Using Naive Bayes Algorithm. *IJISSET International Journal of Innovative Science, Engineering & Technology*, 2(2015): 441-444.
- [9] Meng, X. H., Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors." *The Kaohsiung journal of medical sciences* 29, no. 2 (2013): 93-99.
- [10] Afshari, A. A., and S. M. Mirhosseini, "A New Approach in Diabetes Diagnosis by Hybrid of Genetic Algorithm and Decision Tree." *International Journal of Science* Volume-5, Issue-1, pp. 805-814, 2016.
- [11] Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." *Procedia Computer Science* 47 (2015): 45-51.
- [12] Vijayarani, S. And S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms. *International Journal of Science*." *Engineering and Technology Research (IJSETR)*, 4(2015): 816-820.

- [13] Gulia, A., R.Vohra, and P. Rani, "Liver Patient Classification Using Intelligent Techniques." (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2014): 5110-5115.
- [14] Raad, A., A. Kalakech, and M. Ayache. "Breast cancer classification using a neural network approach: MLP and RBF." networks 7, no. 8 (2012): 9.
- [15] Senturk, Z. K., and R. Kara. "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven different algorithms." Computer Science & Engineering 4, no. 1 (2014): 35.
- [16] Karlik, B., "Hepatitis Disease Diagnosis Using Back Propagation and the Naïve Bayes Classifiers." Journal of Science and Technology, 1 (2011): 49-62.
- [17] Sathyadevi, G., "Application of CART Algorithm in Hepatitis Disease Diagnosis." IEEE International Conference on Recent Trends in Information Technology (ICRTIT), MIT, Anna University, Chennai, 3-5 June 2011, 1283-1287.
- [18] Cleveland Clinic Foundation, "Heart Disease Data Set ", Available at: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>

