An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach

Thesis submitted to the Department of Computer Science and Engineering in the partial fulfillment of the requirement for the Degree of M.Sc. in CSE (Evening) program

Submitted by

Rajib Kumar Halder

Student ID: M180305745

Batch: 7th Session: Summer-2018

Supervised by

Professor Dr. Mohammed Nasir Uddin

Chairman, Department of Computer Science and Engineering

Department of Computer Science and Engineering

Jagannath University

Dhaka - 1100, Bangladesh

February, 2021

An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach

Submitted by

Rajib Kumar Halder

Student ID: M180305745

Supervised by

Professor Dr. Mohammed Nasir Uddin

Submitted to the Department of Computer Science and Engineering of Jagannath University in the partial fulfillment of the requirement for the Degree of M.Sc. in CSE

Thesis Evaluation Committee:

1. Supervisor …….....
   Professor Dr. Mohammed Nasir Uddin
   Chairman, Department of Computer Science and Engineering

2. Member …….....

3. Member …….....

External Member …….....

Thesis Approval

Student Name: Rajib Kumar Halder

Student ID: M180305745

Thesis Title: An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach

We the undersigned, recommend that the thesis completed by the student listed above, in the partial fulfillment of the requirement for the Degree of M.Sc. in CSE, be accepted by the Department of Computer Science and Engineering, Jagannath University for deposit.

Supervisor Approval

……

Professor Dr. Mohammed Nasir Uddin

Departmental Approval

……

Professor Dr. Mohammed Nasir Uddin

Chairman, Department of Computer Science and Engineering

Jagannath University

Dhaka -1100, Bangladesh

Dedicated to my beloved parents

Abstract

Cardiovascular disease, which describes a range of conditions, is known as heart disease. Predicting cardiovascular disease to ensure the treatment area is one of the most important challenges. Machine learning methods help to manipulate a large amount of data and without human intervention, automatically identify the hidden patterns. In the preliminary stage of cardiovascular disease, it helps physicians to make the right decision. The main objective of this research work is to develop an intelligent agent to predict cardiovascular disease to inquire about responsible steps before any unwanted event occurs. We proposed an ensemble method-based Multilayer Dynamic System (MLDS) that can be able to increase its existing knowledge in every layer. Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator, Lasso, and Extra Trees classifier have been used for feature selection. Finally, for classification Random Forest (RF), Naïve Bayes (NB), Gradient Boosting (GB) classifiers have been used to construct the ensemble method. When the above classifiers failed to classify correctly in any layer, we applied K Nearest Neighbor algorithm to find the neighborhood point of the test data. The effectiveness of the proposed model has been validated by using a real dataset (70,000 records) collected from Kaggle and obtained the accuracy of 88.84%, 89.44%, 91.56%, 92.72%, and 94.16% based on the different proportions between training and testing data 50:50, 60:40, 70:30, 80:20, and 87.5:12.5 respectively. Our proposed model provided a maximum AUC=0.94 which means it has a 94% chance to classify positive class and negative class when the proportion between training and the testing dataset was 87.5:12.5. As well as three different datasets Cleveland (303 instances), Hungarian (294 instances), and Framingham (4240 instances) have been applied to this model and obtained the accuracy of 98.88%, 99.53%, 99.98%, 98.36%, 96.66%, 97.77% and 94% based on different partitions of datasets. Besides, the proposed model has been compared with five existing prediction models and the results showed that the proposed model can effectively predict cardiovascular disease.

Keywords: Machine Learning, Cardiovascular Disease, Features Selection, Ensemble Model, Classification

Acknowledgment

In this very special moment, first and foremost I would like to express my heartiest gratitude to the almighty God for allowing me to accomplish this M.Sc. study successfully. I am grateful to my respectable supervisor Professor Dr. Mohammed Nasir Uddin, who has given me constructive comments, suggestions, inspiration and guidance during this research work.

I would like to thank our honorable examiners of the examination committee for their valuable comments and suggestions that helped to improve manuscript.

I am thankful to my family members for their unconditional support and encouragement. I wish to say thank to all the staffs of Department of Computer Science and Engineering of Jagannath University.

Rajib Kumar Halder

February, 2021

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview:

According to the World Health Organization (WHO), 17.9 million deaths due to cardiovascular disease are reported every year and one-third of these deaths occur prematurely in people under the age of 70 years [1]. Around 23.6 million deaths will be cardiovascular disease-related deaths by the year 2030, which is very much alarming. The report from WHO suggests that, in 2018, around 17.3 million people died from heart-related diseases among which stroke case death was nearly 6.2 million and the coronary heart-related disease caused an estimated 7.3 million deaths [2]. Symptoms include tightness in the chest, pressure in the chest, discomfort in the chest (angina) and pain in the chest, shortness of breath, numbness, weakness or coldness in the legs or arms if the blood vessels narrow in those parts of the body, pain in the neck, throat, upper abdomen or back. Some significant reasons for cardiovascular disease, such as age, smoking, sugar, obesity, depression, hypertension, high blood pressure, cholesterol, poor diet, physical inactivity are present [3]. Cardiovascular disease is also caused by coronary artery damage, damage to the whole or part of the heart, or poor supply of the organ with nutrients and oxygen. Some types of cardiovascular diseases are inherited, such as hypertrophic cardiomyopathy, dilated cardiomyopathy, right ventricular arrhythmogenic cardiomyopathy.

It is necessary to watch cardiovascular symptoms and concerns with the doctor. Early-stage detection of cardiovascular disease can reduce the death rate. But it's not so easy because medical information has redundancy, multi-attribution, incompleteness, and a close relationship with time. Analyzing and take the proper decision from the massive volumes of data effectively becomes a major problem. Machine learning technique helps in creating prediction models that can process and analyze large amounts of complex medical data and predict the absence or presence of cardiovascular disease in the body with accurate results. In Machine Learning a computer program is assigned to perform some tasks and a machine has learned from its experience and machine takes the decision and does predictions based on data [4]. When unseen data is provided machine learning techniques enable a machine to make proper decisions based on a build-in analytical model. As well as machine learning techniques take less time for the prediction with more accuracy. An intelligent prediction system for cardiovascular disease would facilitate cardiologists in taking quicker decisions so that more patients can receive treatments within a shorter period of time, resulting in saving millions of lives.

1.2 Problem Statement:

Most of the data of hospital in Bangladesh is not only confidential but also, it's not well organized. The data that was tried to collect for research purposes, was denied to provide by the authority. The data that is allowed to collect was mostly the patient's name, age and sex which are not enough for prediction. Cardiovascular diseases depend on a lot of variables like age, weight, systolic blood pressure, diastolic blood pressure, amount of cholesterol, blood sugar in body, drug addiction, physical activity / inactivity, chest pain etc. The collection of these data was against the hospital rules. Another problem is that little amount of data records and a single dataset are not enough to test the efficiency of a model. The main problem in building a prediction model is to select an appropriate preprocessing mechanism to improve classification accuracy because sometimes the collected data contains missing values and many irrelevant features that are not involved in decision making. Identifying the missing values, removing the missing values, or filling the missing values with median is an important task before classification. Irrelevant features that are not involved in decision making increase searching times, decrease prediction accuracy. So, selecting the most effective features based on features selection techniques is a key task. Several types of research have been done to build a cardiovascular disease prediction system in the last period. Most of the systems are one-layer filtering system. These systems are not able to increase their knowledge from their existing resources and participate in further prediction process after completing a classification process to improve the accuracy of their prediction. So, it’s a major issue to build a multilayer dynamic system to predict cardiovascular disease.

1.3 Objectives:

The objectives of this research work are:

- To develop an intelligent agent to predict cardiovascular disease to inquire about responsible steps before any unwanted event occurs.
- To identify the key patterns or features from the datasets by using effective feature selection techniques to highlight key features to improve the classification accuracy.
- To select the good proportion between training and testing data for which a model performs well.
- To continue the classification process by introducing multilayer techniques and increasing systems knowledge from its resource to improve the proposed model performance.

1.4 Contribution:

To fulfill the objectives and to present an efficient cardiovascular disease prediction system the contribution of this research work is given below in details:

- Our proposed model works as a multilayer dynamic system (MLDS). Layer to layer prediction occurs in this system. In MLDS we have introduced how to continue the classification process from one layer to another layer and how a system can be able to develop its existing knowledge from one layer to another layer from its immediate previous

layer after completing the validation process (details are given in Chapter 5: Table 5.3, 5.4, 5.5, 5.6, 5.7).

- Besides increasing existing knowledge, we have used three classification algorithms to introduce ensemble method in every layer to improve predictive performance and find the optimal result.
- Not only for better classification purposes but also for feature selection we have also used the ensemble technique in our proposed MLDS. We have used multiple features selection techniques to select the effective features. We selected those features that are common within a fixed range among the five algorithms (details are given in Chapter 4: Table 4.2).
- A small amount of data was used in previous existing studies such as Cleveland (303 instances), Hungarian (294 instances), Framingham (4238 instances) and the authors compared their proposed model's performance to other studies. But our proposed model able to show good performance by using real and more data than other existing studies. In our research work, we have used Kaggle heart disease dataset that contains 70,000 instances as well as we have applied other existing research work datasets on our proposed model and compared our proposed MLDS performance to their model. The compression boxes are given in Chapter 5: Table 5.8, 5.9.
- To select the best proportion between the training and testing dataset we have divided the whole dataset (70000 instances) into five partitions (50:50, 60:40, 70:30, 80:30, 87.5:12.5) and applied our proposed MLDS to every proportion and observed the classification performance (details are given in Chapter 5).

1.5 Organization of the thesis:

In this section, which chapter consist which type of information are discussed:

In chapter 1, the overview of the whole proposed method, existing problems, the objectives of this research work and to fulfill mentioned objectives, the contributions of this research work are described.

Chapter 2, begins with information about types of cardiovascular disease (CVD). This chapter also consists information of different machine learning techniques such as classification algorithms, feature selection algorithms, ensemble learning model and dataset details.

In chapter 3, existing works which are used to predict cardiovascular disease are mentioned and by understanding these related works, limitations of these works are identified and written in details.

In chapter 4, The whole proposed method is described in detail. The whole research work is divided into five parts.

In chapter 5, performance evaluations metrics such as confusion matrix, detection accuracy, true positive rate, false positive rate, true negative rate, false negative rate, roc curve, area under curve results, layer to layer prediction progress in table format are generated. These results are then

compared with other model's performance evaluations metrics results. These results are presented in table format and for better visualization graphical presentation is also given.

In chapter 6, as a conclusion the final outcome of this research work is written and why this research work is better and future work are discussed.

Chapter 2

Theoretical Background

2.1 Cardiovascular Disease (CVD):

Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels. There are several types of cardiovascular diseases such as coronary heart disease, stroke, congenital heart disease, aortic aneurysm and dissection, deep venous thrombosis (DVT) and pulmonary embolism, peripheral arterial disease, etc. [5].

A. Coronary heart disease: Disease of the blood vessels which supply the muscle of the heart. Coronary heart disease kills more than 7 million people a year.

Risk factors: High blood pressure, high blood cholesterol, physical inactivity, use of tobacco, unhealthy diet, diabetes, advancing age, hereditary (genetic disposition are major risk factors. Other contributing factors include poverty, inadequate learning, poor mental health (depression), inflammation and blood clotting issues.

B. Stroke: Blood supply disruption to the brain causes strokes. Either blockage (ischaemic stroke) or rupture of the blood vessel (haemorrhagic stroke) can lead to this. Almost 6 million peoples are killed by strokes each year.

Risk factors: High blood pressure, atrial fibrillation, high blood cholesterol, use of tobacco, unhealthy diet, physical inactivity, diabetes, and age progression.

C. Congenital heart disease: Genetic factors or adverse exposures during gestation may be responsible for malformations of the heart structures present at birth. Holes in the heart, irregular valves, and abnormal chambers of the heart are examples.

Risk factors: Maternal alcohol use, expectant mothers use of medicines (for example, thalidomide, warfarin), maternal infections such as rubella, poor maternal nutrition (low folate intake), close parental blood relationships (consanguinity).

D. Aortic aneurysm and dissection: Dilation and rupture of the aorta.

Risk factors: Advancing age, long-term elevated blood pressure, Marfan syndrome, congenital heart disease, syphilis, and other contagious and inflammatory disorders.

E. Deep venous thrombosis (DVT) and pulmonary embolism: Blood clots in the veins of the leg that can move to the heart and lungs and dislodge.

Risk factors: Surgery, obesity, cancer, prior DVT episode, new birth, use of oral contraceptives and hormone replacement therapy, lengthy stretches of immobility, such as when travelling, elevated homocysteine blood levels.

F. Peripheral arterial disease: Disease of the arteries that supply the legs and arms. Risk factors with regard to coronary heart disease.

G. Other cardiovascular diseases: Cardiac tumors, vascular tumors of the brain, heart muscle disorders (cardiomyopathy), diseases of the heart valve, heart lining disorders.

Inflammation, drugs, high blood pressure, unhealthy diet, trauma, toxins and alcohol are other factors that can harm the heart and blood vessel system.

2.2 Machine Learning (ML):

Machine Learning (ML) is a kind of Artificial Intelligence (AI) technique which allows the system to obtain knowledge with no explicit programming. The main intention of ML technique is to enable the computers to learn with no human assistance. Machine learning methods have generally gained significantly high precision in classification-based problems [6]. The various method has been used for knowledge abstraction by using well-known techniques like features selection, classification, clustering [7],[8]. All Machine learning techniques are divided into three sections: one is supervised learning, the second is unsupervised learning and the third is reinforcement learning.

- In Supervised learning a function that designing an input to an output based on input-output pairs. Supervised ML techniques utilize from what it is has gained knowledge from the previous and present data with the help of labels to forecast events [9]. Train data are labeled in supervised machine learning. Example: SVM, Naive Bayes, Random Forest, Linear Regression, Logistic Regression classification etc.
- In the unsupervised machine learning technique, there is not any labeled data as supervised learning. Data are grouped based on their similarity or measure Euclidian distance with others by assigning centroid points such as clustering techniques [7],[10]. The core concept of clustering is to group several data or objects into a group or several groups where each group contains data that has similarities or very close to other data [11]. Examples: Simple K-means, Hierarchical, Expectation maximization clustering etc.
- Semi-supervised learning consists of a small amount of labeled data with a large amount of unlabeled data during training. It's a part of unsupervised learning and supervised learning. It is preferable in case where the acquired labeled data need skillful and appropriate resources to train or learn from it [9].
- Reinforcement ML techniques interact with the environment by actions and locates errors or rewards. Trial and error search and delayed rewards are some of the common features of reinforcement method. It enables the systems and software programs to identify the ideal behavior in a specific context to increase the performance [9].

2.3 Feature Selection:

In Data Mining and Machine Learning, feature selection plays a significant role. This lowers data dimensionality. In the collection of features, two methods are used, one filter method and another wrapper method [12]. In the filter system, features are chosen in various statistical tests based on their scores, which calculate the significance of features by their correlation with the dependent variable or the outcome variable. By evaluating the usefulness of a subset of features with the dependent variable, the wrapper approach seeks a subset of features [13]. Features selection is necessary for the machine learning process because sometimes irrelevant features affect the classification performance of the machine learning classifier. Feature selection improves the classification accuracy and reduces the model execution time [8].

2.3.1 Correlation Attribute Evaluator:

Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average [14].

2.3.2 Gain Ratio Attribute Evaluator:

Evaluates the worth of an attribute by measuring the gain ratio with respect to the class [15].
GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

2.3.3 Information Gain Attribute Evaluator:

Evaluates the worth of an attribute by measuring the information gain with respect to the class [16].
InfoGain(Class,Attribute)= H(Class) - H(Class | Attribute)

2.3.4 Lasso:

Least absolute shrinkage and selection operator eliminate the zero features from the feature's subset. By updating the absolute value of features coefficient Lasso select the features. The features having high values of coefficients will be included in selected feature subsets. LASSO performs excellently with low coefficients feature values [8].

2.3.5 Extra Trees classifier:

It produces a large number of unpruned decision trees from the training dataset and in the case of regression, forecasts are made by integrating the prediction of decision trees or by the use of majority voting in the case of classification. Each tree is provided with a random sample of K number of features from which each decision tree must select the best feature.

2.4 Ensemble Method:

The Ensemble Model is built in parallel with a series of individual models whose outputs are merged with a decision fusion technique to provide a single solution to a given problem [17]. The models that are used to construct an ensemble model are called base mode. The main advantage of the ensemble model is that the modeling of final decision-making judgment based on the various classifiers. The model builds in two steps: all the base learners are used in a parallel style where the generation from a learner has an impact on the other learners and the next decisions or results of all base learners are combined in two different way namely, majority voting and weighted averaging [18]. Voting is divided into two categories: one is hard voting and another is soft voting. In hard voting, each classifier votes individually, and the majority of these votes is accepted. In soft voting, each classifier defines the probability for a particular target class on each data point. By averaging these probabilities, the target label with the greatest average provides the vote [19].

2.5 Classification Technique:

Classification algorithms are used to predict the target class where predefined labels are assigned to instances by properties. Supervised learning technique is used for classification: Training set $\rightarrow$ Classification algorithms $\rightarrow$Unseen data$\rightarrow$ Prediction result.

2.5.1 Random Forest:

Random Forest is an example of an ensemble method involving the collection of decision trees. In the Random Forest algorithm, samples are drawn randomly and decision trees are built for the random sample and the process is repeated [20]. It avoids the missing values and outliers by following steps: data analysis and data pre-processing and corrects the overfitting to their training dataset [21]. This ensemble classifier incorporates several decision trees to get the best result. Decision Trees mainly apply bootstrap aggregating or bagging [22]. For example, a given data, X= $\{X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n\}$ which repeats the bagging from b=1 to B. The unseen samples x' is made by averaging the predictions $\sum_{b=1}^{B} fb(\mathrm{x}')$ from every individual trees on x' :

$$j = \frac{1}{B}\sum_{b=1}^{B} fb(X') \qquad (2.1)$$

The uncertainty of prediction on tree is made through its standard deviation:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(fb(X') - \hat{f})^2}{B-1}} \qquad (2.2)$$

2.5.2 Naïve Bayes:

Naïve Bayes classifier or the Bayesian theorem is another classification technique that is utilized for predicting a target class. It depends on probabilities in its calculations [23],[24]. Based on Bayesian theory, each quantity has a statistical distribution by which a test sample can be categorized [12]. Naive Bayes does not consider the correlation between attributes [25]. For example, every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussain function with prior probability $P(X_f)$ = priority ∈ (0:1)

$$P(X_{f1,}\, X_{f1,} \dots\dots\dots, X_{fn}\,|c) = \prod_{i=1}^{n} P(X_{fi}|c)$$

$$P(X_{fi}|c_i) = \frac{P(c_i \mid X_f)\, P(X_f)}{P(C_i)} \qquad (2.3)$$

c ∈ {begin, malignant}

At last the testing data is categorized based on the probability of association:

$$c_{nb} = \arg\ \max\ P(c_k) \prod_{i=1}^{n} P(X_{fi}|c_k), \quad \text{for k =1,2}$$

2.5.3 Gradient Boosting:

For regression and classification issues, the gradient boosting machine learning method is used. In the form of an ensemble of decision trees that are constructed in a stage-wise process, it may generate a prediction model [26]. In gradient boosting, decision trees are generally used. The main advantage of gradient boosting is that the residual last time is reduced in each calculation and the residual gradient direction can be reduced to create a new model in order to decrease the residual [27]. In boosting, every new tree is a fit on a modified version of the original dataset.

Algorithm: Gradient Boosting

1. $F_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_{i,}, \rho)$
2. For m=1 to M do:
3. $\widehat{y_i} = -\left[\frac{\partial L\left(y_{i,} F(x_i)\right)}{\partial F_{x_i}}\right]_{F(x)=F_{m-1}(x)}, i = 1, \dots\dots, N$
4. $\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} [\widehat{y_i} - \beta h(x_i; \alpha_m)]^2$
5. $\rho_m = \arg min \sum_{i=1}^{N} L(y_{i,} F_{m-1}(x_i) + \rho h(x_i; \alpha))$
6. $F_m(x) = F_{m-1}(x) + \rho h(x, \alpha_m)$
7. end For
8. end

2.5.4 K Nearest Neighbor (KNN):

KNN is a supervised machine learning algorithm [8]. It is used for both classification and regression. Features similarity is used to predict the values of new data points. The Euclidean or Manhattan or Hamming methods are used to calculate the distance between test data and each row of training data. Usually, to find the similarity, it works with distance [28].

Algorithm: K Nearest Neighbor
1. Input: x, S, d
2. Output: class of x
3. For $(x', l')\epsilon S$ do
4. Compute the distance d(x', x)
5. end For
6. Sort the |S| distances by increasing order
7. Count the number of occurrences of each class I_j among the k nearest neighbors
8. Assign to x the most frequent class

2.6 Dataset Description:

a. Cardiovascular Disease Dataset:

This dataset has been collected from Kaggle. Kaggle is an online community of data scientists and machine learning practitioners. This community provides different types of survey data on different topics. This dataset was uploaded by Svetlana Ulianova is a data scientist at Ryerson University, Toronto, Ontario, Canada. It has total 70000 instances. It contains 12 features where 11 features are independent and one is dependent.

Data description

There are 3 types of input features:
Objective: factual information.
Examination: results of medical examination.
Subjective: information given by the patient.

Features:

1. Age | Objective Feature | int (days)
2. Height | Objective Feature | int (cm)
3. Weight | Objective Feature | float (kg)
4. Gender | Objective Feature | categorical code
5. Systolic blood pressure | Examination Feature | int
6. Diastolic blood pressure | Examination Feature | int
7. Cholesterol | Examination Feature | 1: normal, 2: above normal, 3: well above normal
8. Glucose | Examination Feature | 1: normal, 2: above normal, 3: well above normal

9. Smoking | Subjective Feature | binary
10. Alcohol intake | Subjective Feature | binary
11. Physical activity | Subjective Feature | binary
12. Presence or absence of cardiovascular disease | Target Variable | binary

b. Cleveland and Hungarian Heart Disease Dataset:

These two datasets have been collected from UCI Machine Learning Repository. UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. These two datasets contain 76 features, but all published experiments refer to using a subset of 14 of them.

Creators:

1. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
2. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

Features:

1. Age | Integer
2. Sex | Integer | 1=male,0=female
3. Cp | Chest pain type | Discrete | 1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptomatic
4. Trestbps | Resting blood pressure| Integer
5. Chol | Serum cholestoral in mg/dl | Integer
6. Fbs | Fasting blood sugar > 120 mg/dl | Integer | 1=true, 0=false
7. Restecg | Resting electrocardiographic results | Integer | 0=normal, 1=having ST-T wave abnormality, 2=showing probable or left ventricular hypertrophy by Estes criteria
8. Talach | Maximum heart rate achieved| Integer
9. Exang | Exercise induced angina | Integer | 1= yes, 0=no
10. Oldpeak | ST depression induced by exercise relative to rest | Real
11. Slope | The slope of the peak exercise ST segment | Integer
12. Number of major vessels | Number of major vessels (0-3) colored by flourosopy that range between 0 and 3 | Integer
13. Thal | Thal | Integer
14. Num | The predicted attribute | Integer | 0=no, 1= yes

c. Framingham Heart Disease Dataset:

This dataset has been collected from Kaggle. It has total 4240 instances. It contains 16 features where 15 features are independent and one is dependent.

Features:

1. Gender | binary (0=male, 1=female)
2. Age | int (years)
3. Education | int
4. Current Smoker | binary (0=no, 1=yes)
5. Cigs per day | int
6. BPMeds | binary
7. Prevalent Stroke | binary (0=no, 1=yes)
8. Prevalent Hypertension | binary (0=no, 1=yes)
9. Diabetes | binary (0=no, 1 =yes)
10. Cholesterol | int
11. Systolic blood pressure | float
12. Diastolic blood pressure | float
13. BMI | float
14. Heart Rate | int
15. Glucose | int
16. Target | binary (0=no, 1=yes)

Chapter 3

Related Work

3.1. An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection:

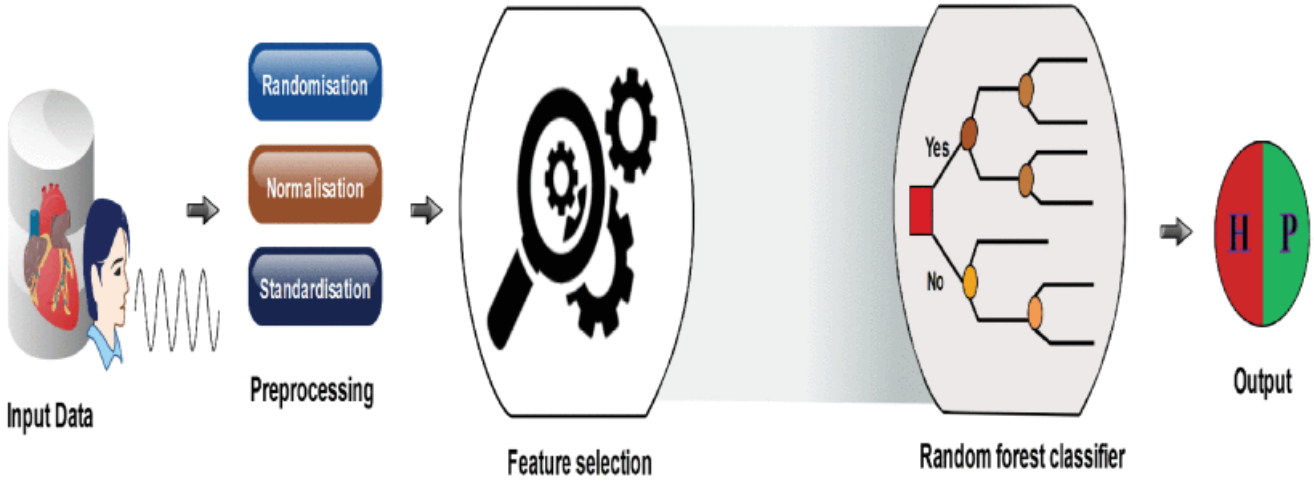

Figure 3.1: Architecture of the model of Ashir Javeed et al. [29].

Ashir Javeed et al. [29], has proposed a smart learning method for better diagnosis of heart disease based on random search algorithm and structured random forest model. This proposed model uses the Random Search Algorithm (RSA) to select features and refine the random forest classifier for the prediction of heart failure. Data was collected from the machine learning UCI repository. The total number of instances in the dataset is 303 among them 297 instances have complete attributes information while six instances have missing details. First, they used Conventional Random Forest model for heart disease detection and obtained an accuracy 90% but while they implemented an RSA-based optimized random forest model then obtained an accuracy 93.33%. The proposed model produced 3.3% higher accuracy than the conventional random forest model while using only 7 features.

Limitations: Independent features has a great impact on target value. During the features selection in this research work no intelligence was used to select the features subset because RSA is fully random location generating process. Suppose a dataset of N=100 features. In RSA based features selection technique need to generate a total N-1=99 subset of features because we do not know for which subset of features model will give good accuracy. It's a time-consuming process.

3.2. Classification models for heart disease prediction using feature selection and PCA:

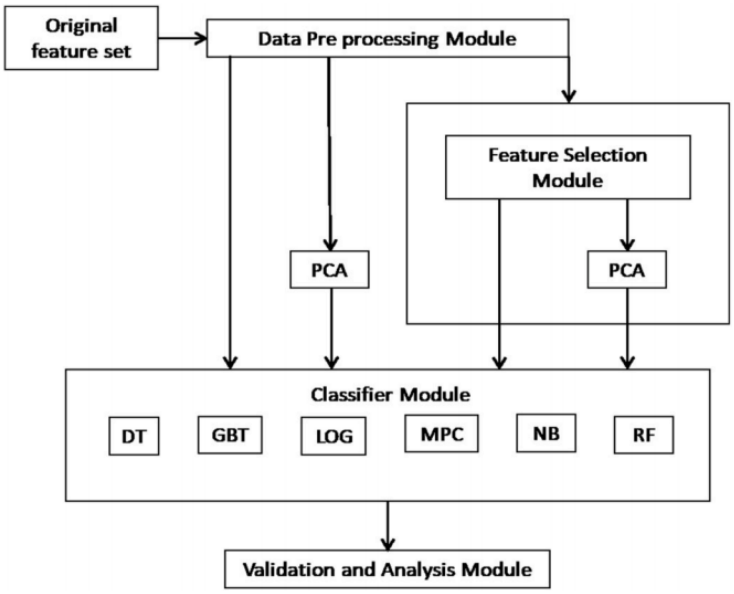

Figure 3.2: Architecture of proposed model of Anna Karen Garate-Escamila et al. [30].

Anna Karen Garate-Escamila et al. [30], has proposed a classification model for heart disease prediction using feature selection and PCA. They used six classifiers, principle component analysis (PCA), and chi-squared feature selection techniques. They used Cleveland (283 instances), Hungarian (294 instances), and Cleveland-Hungarian (577 instances) dataset for this experiment. They performed four types of experiments: i) classified the raw data with all six classifiers, ii) applied the chi-square feature selection technique to obtain and validate the features with the classifiers, iii) reduced the datasets obtained by chi-square and then applied PCA and iv) the final experiment was the direct use of PCA from raw data. Chi-square and principal component analysis with random forest (CHI-PCA) obtained the highest performance, with an accuracy of 98.7%, 99.0%, and 99.4% for Cleveland, Hungary, and Cleveland-Hungarian (CH) data sets, respectively.

Limitations: It is difficult to identify how many principal components to keep- in practice because after implementing PCA on the dataset, original features of a dataset will turn into principal components that's are the linear combination of original features and not as readable, interpretable as original features. So, it has a great possibility to miss some information compared to the original list of features during the selection of principle component.

3.3. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms:

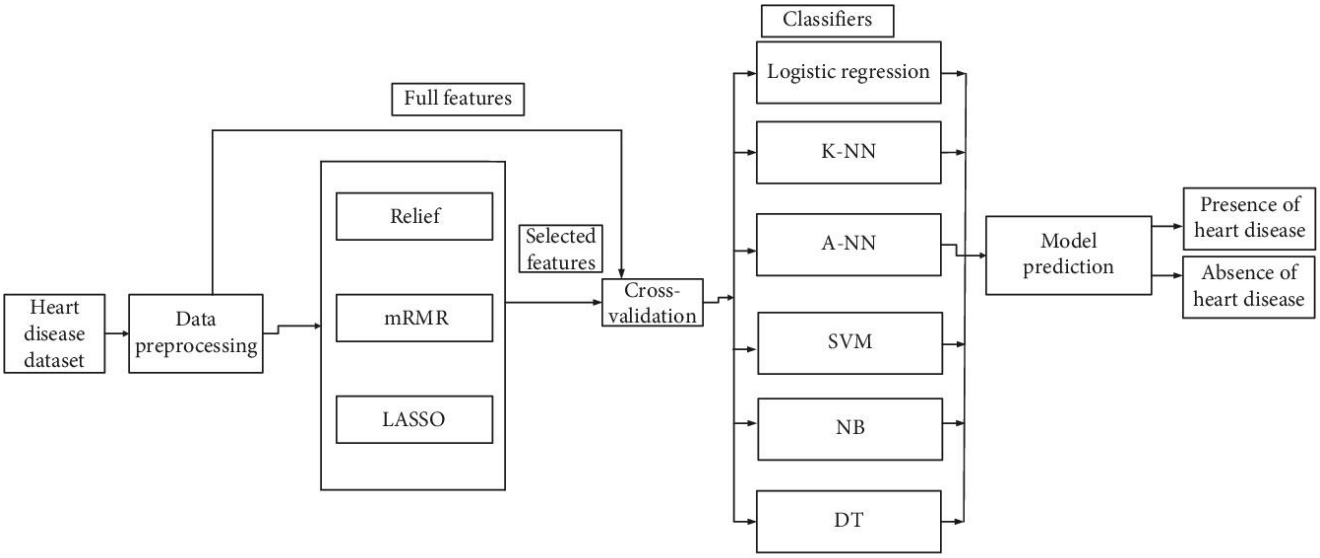

Figure 3.3: Architecture of the model of Amin Ul Haq et al. [8].

Amin Ul Haq et al. [8], has Proposed a hybrid intelligent system framework for the prediction of heart disease. The total number of instances in the dataset is 303 among them 297 instances have complete attributes information while six instances have missing details. The researcher used three feature selection algorithms (Relief, mRMR, and LASSO), the K-fold cross-validation method, and seven classifiers (LR, K-NN, ANN, SVM (kernel RBF and kernel linear), NB, DT, and RF). They recorded the accuracy of the different classifiers based on extracted features by using different feature selection algorithms.

Limitations: The disadvantage of this method is that the training algorithm has to be rerun from scratch k times.so, It takes too much times to complete one computation task in case of train the model. This system is not suitable for large dataset.

3.4. Identification of Cardiovascular Diseases Using Machine Learning:

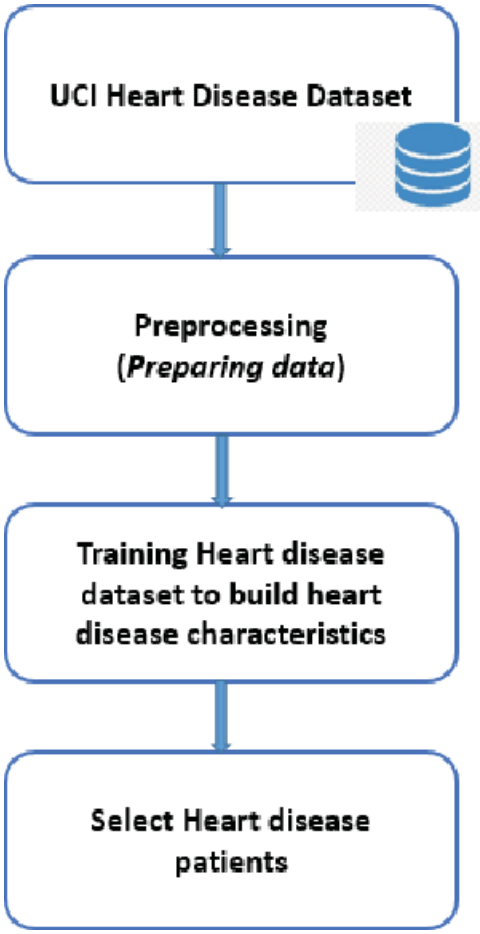

Figure 3.4: Architecture of the model of Nabaouia Louridi et al. [28].

Nabaouia Louridi et al. [28], Proposed a machine learning model to identify Cardiovascular Disease (CVD) where they used 303 records with 13 attributes. They used Support Vector Machine (SVM), K-NN, Bayes Naif (BN) for classification and found the highest accuracy 86.8% by SVM-linear kernel are used for classification.

Limitations: No algorithm for feature selection was introduced in this research work. Researcher only handled the missing value in preprocessing unit but features selection plays an important role to get good accuracy and it reduces the execution time. It helps to select effective features that have great impact on target value.

3.5. An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk:

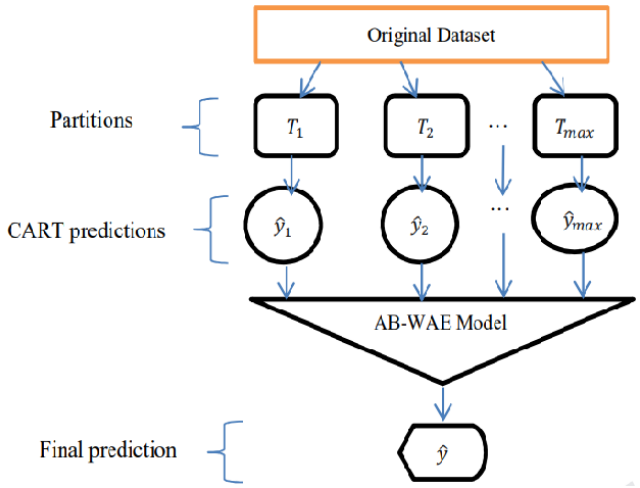

Figure 3.5: Architecture of the model of Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang [31].

Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang [31], has Proposed an improved ensemble learning approach for the prediction of heart disease risk. The proposed strategy requires the use of a mean-based splitting technique to divide the sample into smaller subsets arbitrarily. Different partitions are modeled by using the classification and regression tree (CART) algorithm. An accuracy dependent weighted aging classifier ensemble (WAE) is used to generate a homogeneous ensemble from different CART models. In this research work, two heart disease datasets were used, the Cleveland dataset (303 instances) obtained from the and the Framingham dataset (4238 instances).

Limitations: No optimization algorithm was deployed to select effective attributes in this system. This system cannot handle noisy data and have a chance to create noisy decision tree.

All the above existing models are one-layer filtering system. These models are not capable to increase their existing knowledge from its resource. In our research, we have found after completing a classification process in a layer the correctly classified data can have a great impact on the remaining incorrectly classified data. If correctly classified data is used as a training set for the next classification process after completing the validation process in a layer it has a probability to extract a new hidden pattern from the remaining test data. Correctly classified data of a layer can be used as new knowledge for the system. So, we have proposed a multilayer dynamic prediction system (MLDS) where after completing a classification process in any layer the proposed MLDS is ready to attend further prediction process of cardiovascular disease for the rest of the incorrect classified data with additional knowledge to get a more accurate result.

Chapter 4

Proposed Methodology

4.1 System Overview:

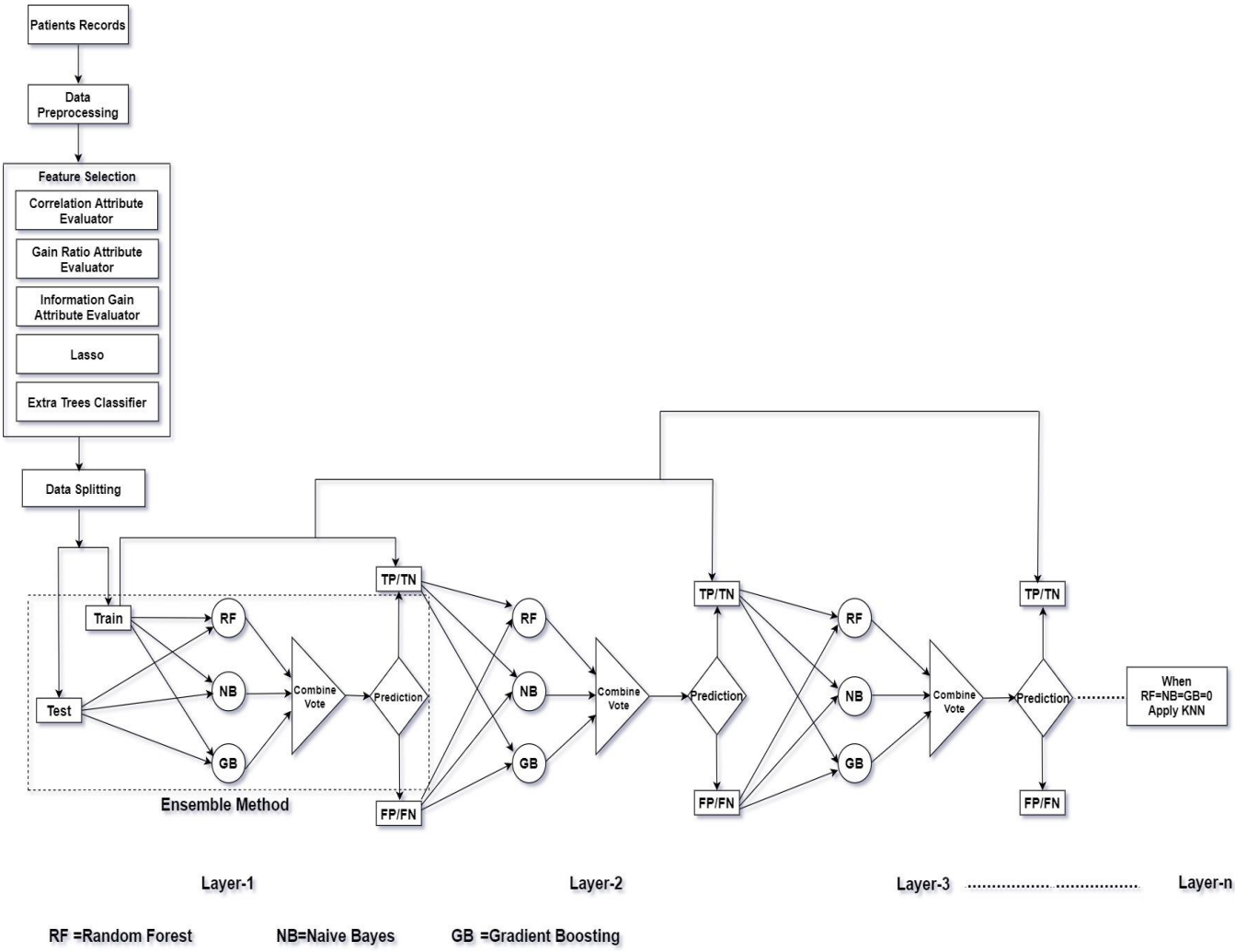

Figure 4.1: Proposed multilayer dynamic system (MLDS) to predict cardiovascular disease.

This section provides a detailed description about the proposed methodology of this model. Spyder (Scientific python development environment), WEKA tool, python language, a real dataset, and different machine learning algorithms have been used in this research work to predict cardiovascular disease. This model is a multilayer dynamic system so prediction occurs layer by

layer and every layer has more knowledge than the previous layer. The proposed model has been developed with the aim of classifying people with cardiovascular disease positive or negative. Correlation Attribute Evaluator algorithm, Gain Ratio Attribute Evaluator algorithm, Information Gain Attribute Evaluator, Lasso, Extra Trees Classifier have been used to select important features, and based on these selected features, the performance of the classifiers has been tested. Popular machine learning classifiers Random Forest (RF), Naïve Bayes (NB), Gradient Boosting (GB) have been used in this model. The methodology of the proposed system consists of five stages including (1) real data collection, (2) dataset preprocessing, (3) feature selection, (4) splitting dataset (5) applying classifiers to predict cardiovascular disease. All steps are illustrated in Figure 4.1.

4.1.1 Real Data Collection:

We have collected the dataset from Kaggle contains 70,000 records with 11 independent features and one dependent variable. These data were collected at the moment of medical examination and information given by the patient. Kaggle is an online community of data scientists and machine learning practitioners. This community provides different types of survey data on different topics. Details of all the features of this dataset, including some statistical calculations, are shown in Table 4.1

Table 4.1: Kaggle cardiovascular disease dataset attributes description with statistical calculation.

Serial Number	Variable Description	Min	Max	Mean	StdDev
1	age-int (days)	10798	23713	19468.866	2467.252
2	Height-int (cm)	55	250	164.359	8.21
3	Weight-float (kg)	10	200	74.206	14.396
4	gender-categorical code (f=female, m=male)				
5	ap_hi - int	-150	16020	128.817	154.011
6	ap_lo-int	-70	11000	96.63	188.473
7	Cholesterol-1: normal, 2: above normal, 3: well above normal	1	3	1.367	0.68
8	gluc 1: normal, 2: above normal, 3: well above normal	1	3	1.226	0.572
9	Smoke-bimary	0	1	0.008	0.283
10	Alco	0	1	0.054	0.226
11	active-binary	0	1	0.804	0.397
	Target- Presence or absence of cardiovascular disease-binary	0	1	0.5	0.5

4.1.2 Data Preprocessing

After collecting the dataset, gender column has been transferred from categorical word to numerical value i.e. gender=1 for female instead of "f" and gender=2 for male instead of "m". The patient's age has been converted from days to a year. Missing value handling is an important part of data analysis because attributes in a dataset give valuable information. If any value is missing it impacts on decision making. We have used isnull() function to detect the missing values. The following command is used:

```
In [5]: import pandas as pd
        DataFrame=pd.read_csv("E:/cardio_Data.CSV")
        DataFrame.isnull().sum()

Out[5]: age            0
        gender         0
        height         0
        weight         0
        ap_hi          0
        ap_lo          0
        cholesterol    0
        gluc           0
        smoke          0
        alco           0
        active         0
        target         0
        dtype: int64
```

From the above output we can see that there are no missing values in our dataset. From the above output we can see that there are no missing values in our dataset. There are some important functions to handle the missing value: dropna() function is used to drop the missing values and fillna() function is used to fill NA/NaN values using the specified method.

4.1.3 Feature Selection

It is necessary to extract effective features to improve accuracy and decrease the searching time of classification. For feature selection in our system, we have used five well-known features selection algorithms and these algorithms select important features. Rank wise features arrangement are given in Table 4.2

Table 4.2: Rank wise selected features using five feature selection algorithms.

Algorithm Name	Rank wise selected features	Algorithm Name	Rank wise selected features
Correlation Attribute Evaluator (using WEKA)	0.23816 1 age 0.22115 7 cholesterol 0.18166 4 weight 0.08931 8 gluc 0.06572 6 ap_lo 0.05448 5 ap_hi 0.03565 11 active 0.01549 9 smoke 0.01082 3 height 0.00811 2 gender 0.00733 10 alco	Lasso (using python libraries)	0.01455872 1 age 0.00550486 4 weight 0.00014236 6 ap_lo 0.00014108 5 ap_hi 0 7 cholestoral 0 8 gluc - 0 9 smoke - 0 10 alco - 0 11 active - 0 2 gender - 0.00104144 3 hight
Gain Ratio Attribute Evaluator (using WEKA)	0.072691 5 ap_hi 0.054584 6 ap_lo 0.034366 7 cholesterol 0.01393 1 age 0.008123 4 weight 0.008007 8 gluc 0.001284 11 active 0.000519 3 height 0.000402 9 smoke 0 10 alco 0 2 gender	Extra Trees classifier (using python libraries)	0.290111 1 age 0.177682 5 ap_hi 0.176635 4 weight 0.171074 3 height 0.109448 6 ap_lo 0.045414 7 cholesterol 0.011002 8 gluc 0.005456 2 gender 0.004683 11 active 0.004346 10 alco 0.004143 9 smoke
Information Gain Attribute Evaluator (using WEKA)	0.170065 5 ap_hi 0.106194 6 ap_lo 0.044944 1 age 0.036573 7 cholesterol 0.025608 4 weight 0.006092 8 gluc 0.000918 11 active 0.000332 3 height 0.000173 9 smoke 0 10 alco 0 2 gender		

From the above Table 4.2 we can see that age, cholesterol, weight, gluc, ap_lo, ap_hi are common features from serial 1 to 6 in four algorithms (Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain Attribute Evaluator, Lasso) out of five algorithms. These six features have great impact on target value. So, we have selected these six features (age, cholesterol, weight, gluc, ap_lo, ap_hi) from eleven independent features as an effective feature set.

Effect of age, cholesterol, weight, gluc, ap_lo, ap_hi on cardiovascular disease:

age: Adults aged 65 and older are more likely to suffer from cardiovascular disease than younger people. Aging can lead to changes in the heart and blood vessels that may increase the risk of cardiovascular disease in a person [32].

cholesterol: If there is so much cholesterol in our blood, it builds up in the walls of the arteries, activating a mechanism called atherosclerosis, a type of cardiac disease. The arteries are reduced and blood flow to the muscle of the heart is slowed or blocked [33].

gluc: High blood glucose from diabetes can damage blood vessels and the nerves that control our heart and blood vessels. The longer once have diabetes, the greater risk of developing heart disease [34].

ap_hi: By making blood vessels more rigid and damaging the inner lining, high blood pressure damages the blood vessels. The damaged lining increases the risk of deposition of fat, preventing the flow of blood. Because of blood vessel resistance, the heart must work harder to provide the body with oxygen-rich blood adequately [35].

weight: In several respects, obesity leads to heart failure. More body fat contributes to a higher blood flow, which allows it more difficult for the heart to pump all the excess fluid. This causes damaging changes in the structure and function of the heart over the years that can eventually lead to heart failure [36]. Hypertension and enlarged left ventricle (left ventricular hypertrophy) are also associated with excess weight, increasing the risk of heart failure [37].

ap_lo: Low blood pressure, which causes insufficient blood flow to the organs of the body, can lead to strokes, heart attacks and kidney failure [38].

4.1.4 Dataset Splitting

To select the good proportion between training and testing data we have divided our collected dataset into multiple proportions and applied our proposed MLDS to each of the proportion and tested which ratio gives us the best results (Shown in Chapter 5: Table 5.1, 5.2). We have used “train-test-split” method to divide the whole dataset into training and testing data. We have split the whole dataset into five partitions: 50:50 (35000 training and 35000 testing out of 70000), 60:40 (42000 training and 28000 testing out of 70000),70:30 (49000 training and 21000 testing out of 70000), 80:20 (56000 training and 14000 testing out of 70000) and 87.5:12.5 (61250 training and 8750 testing out of 70000). Different machine learning algorithms have been used for classification and extract the hidden pattern of the testing dataset with the help of the training set.

4.1.5 Applying Classifiers to Predict Cardiovascular Disease

Finally, three classifiers, Random Forest (RF), Naïve Bayes (NB), and Gradient Boosting (GB) have been applied to perform classification. Layer to layer prediction process occurs in this model. In each layer, three classifiers participate to classify the same testing dataset based on the same training dataset. The classifier which classifies more correctly than others is accepted to be reported its accuracy for the related layer. After completing classification in each layer, the correctly classified data (TP+TN) based on comparing with pre-defined target values in original Kaggle dataset is added to previous training data to enter into the next layer. Also, the incorrectly classified data (FP+FN) will participate in the next iteration as new testing data. This process will be continued until three classifiers provide zero accuracy (TP and TN=0). When all three classifiers are being failed to classify correctly, the proposed ensemble model tries to find the optimal number of nearest neighbors by utilizing K Nearest Neighbor (KNN) algorithm. This approach is called multilayer dynamic system (MLDS). The total accuracy for MLDS can be calculated with the formula presented below:

Number of train data in layer i: Number of train data in layer(i-1) + Number of TP, TN in layer (i-1)

Test data in layer i: Number of FP, FN in layer (i-1)

Where i = layer number; i =1, 2, 3, …., n

When Random Forest=0, Naïve Bayes=0, Gradient Boosting=0, KNN=0

$$\text{Total accuracy} = \frac{\sum_{i=1}^{n}(TP,TN)_i}{\text{total number of test data}} \qquad (4.1)$$

Chapter 5

Result Analysis

Performance matrix is used to measure the performance of a machine learning model. Matrix module provides necessary functions to compute performance evaluation metrics using skit-learn library. The evaluation of the model is performed with the confusion matrix. Four outcomes have been generated by the confusion matrix, namely TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) for different proportions of the dataset (50:50,60:40,70:30,80:20,87.5:12.5 respectively).

$$\begin{bmatrix}16813 & 680\\ 3224 & 14283\end{bmatrix}\begin{bmatrix}13429 & 510\\ 2446 & 11615\end{bmatrix}\begin{bmatrix}10106 & 369\\ 1402 & 9123\end{bmatrix}\begin{bmatrix}6707 & 260\\ 758 & 6275\end{bmatrix}\begin{bmatrix}4262 & 123\\ 388 & 3977\end{bmatrix}$$

The following measures are used for the calculation of the accuracy, precision, True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (5.1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5.2)$$

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \quad (5.3)$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \quad (5.4)$$

$$\text{True Negative Rate} = \frac{TN}{TN+FP} \quad (5.5)$$

$$\text{False Negative Rate} = 1\text{-TPR} \quad (5.6)$$

ROC curve shows the performance of a classification model at all classification thresholds. ROC is curve of probability. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. ROC curves are shown in Figure 5.1 for different proportions of the dataset (50:50,60:40,70:30,80:20,87.5:12.5 respectively).

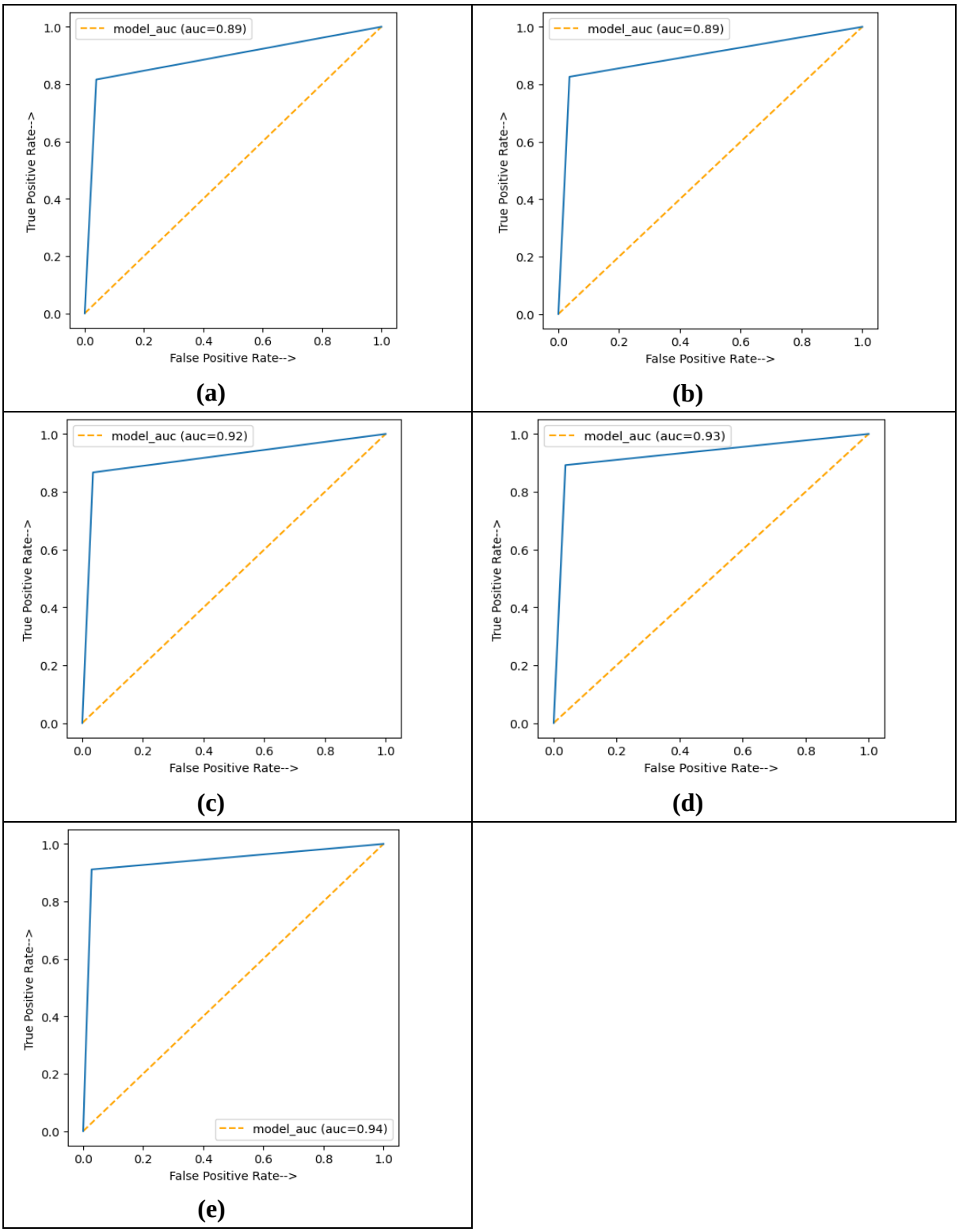

Figure 5.1: ROC curve and AUC curve: (a) 50:50, (b)60:40, (c)70:30, (d)80:20, (e) 87.5:12.5.

The blue dotted line is the ROC curve which plots the (x, y) = (FPR, TPR) points at all classification thresholds. A ROC curve to the top and left is a better model which means the proposed model is much better to classify testing datasets in their respective comparison to another proposed model. AUC represents the degree or measures of separability. AUC measures the whole two-dimensional area from (0,0) to (1,1) below the whole ROC curve (just like integral calculus). A model with higher AUC has a better chance to predict 0s as 0s and 1s as 1s. An excellent classifier has AUC near to the 1 which means it has a good measure of separability. From figure 5.1 (e) we can see that the AUC of this proposed MLDS is maximum when the dataset was divided into the ratio 87.5:12.5 and then the value of AUC is 0.94, it means the proposed MLDS has 94% chances to distinguish between cardiovascular disease absence or presence class correctly. The orange dotes line represents the AUC of this proposed model. This proposed MLDS is better than another model because it has AUC=0.94 is near to 1. The classification performance of the proposed MLDS with other machine learning algorithms has been compared. Same training datasets have been used to train XGBoost, Linear Regression, Support Vector Machine, K-Nearest-Neighbor, Decision Tree classification models.

The performance of our proposed MLDS has been compared with other existing classification algorithms as shown in Table 5.1 and its graphical representations are shown in Figure 5.2, 5.3.

Table 5.1: Performance evaluation metrics result of proposed (MLDS), XGB, LR, SVM, KNN, DT.

Model Name	**Parameter (%)**	**Train and Test Dataset Ratio**				
		50:50	**60:40**	**70:30**	**80:20**	**87.5:12.5**
Proposed MLDS	Accuracy	88.84	89.44	91.56	92.72	94.16
	Precision	95.45	95.79	96.11	96.02	97
	TPR	81.58	82.60	86.67	89.22	91.11
	FPR	3.88	3.65	3.52	3.73	2.80
	TNR	96.11	96.34	96.47	96.26	97.19
	FNR	18.41	17.39	13.32	10.77	8.88
XGB	Accuracy	73.66	73.62	73.70	73.52	73.87
	Precision	76.13	76.13	76.18	75.83	75.72
	TPR	68.97	69.16	69.14	69.40	70.10
	FPR	21.63	21.87	21.71	22.31	22.37
	TNR	78.36	78.12	78.28	77.68	77.62
	FNR	31.02	30.83	30.85	30.59	29.89
LR	Accuracy	70.56	70.70	70.54	70.7	71.10
	Precision	72.16	72.61	72.34	72.32	72.26
	TPR	67.00	66.88	66.74	67.51	68.29
	FPR	25.86	25.43	25.64	26.08	26.08
	TNR	74.13	74.56	74.35	73.91	73.91
	FNR	32.99	33.11	33.25	32.48	31.70
SVM	Accuracy	71.97	72.01	72.09	72.46	72.88
	Precision	77.33	77.53	77.40	77.47	77.55
	TPR	62.20	62.32	62.58	63.71	64.21
	FPR	18.24	18.21	18.35	18.70	18.49
	TNR	81.75	81.78	81.64	81.29	81.50
	FNR	37.79	37.67	37.41	36.28	35.78
KNN	Accuracy	68.92	68.87	69.15	69.22	69.23
	Precision	69.72	69.86	70.01	70.12	69.62
	TPR	66.94	66.88	67.25	67.51	67.99
	FPR	29.09	29.10	28.93	29.03	29.53
	TNR	70.90	70.89	71.06	70.96	70.46
	FNR	33.05	33.11	32.74	32.48	32.00
DT	Accuracy	62.84	63.53	63.14	63.14	63.37
	Precision	62.86	63.69	63.39	63.30	63.10
	TPR	62.83	63.67	62.64	63.34	63.98
	FPR	37.14	36.60	36.34	37.06	37.24
	TNR	62.85	63.39	63.65	62.93	62.75
	FNR	37.16	36.32	37.35	36.65	36.01

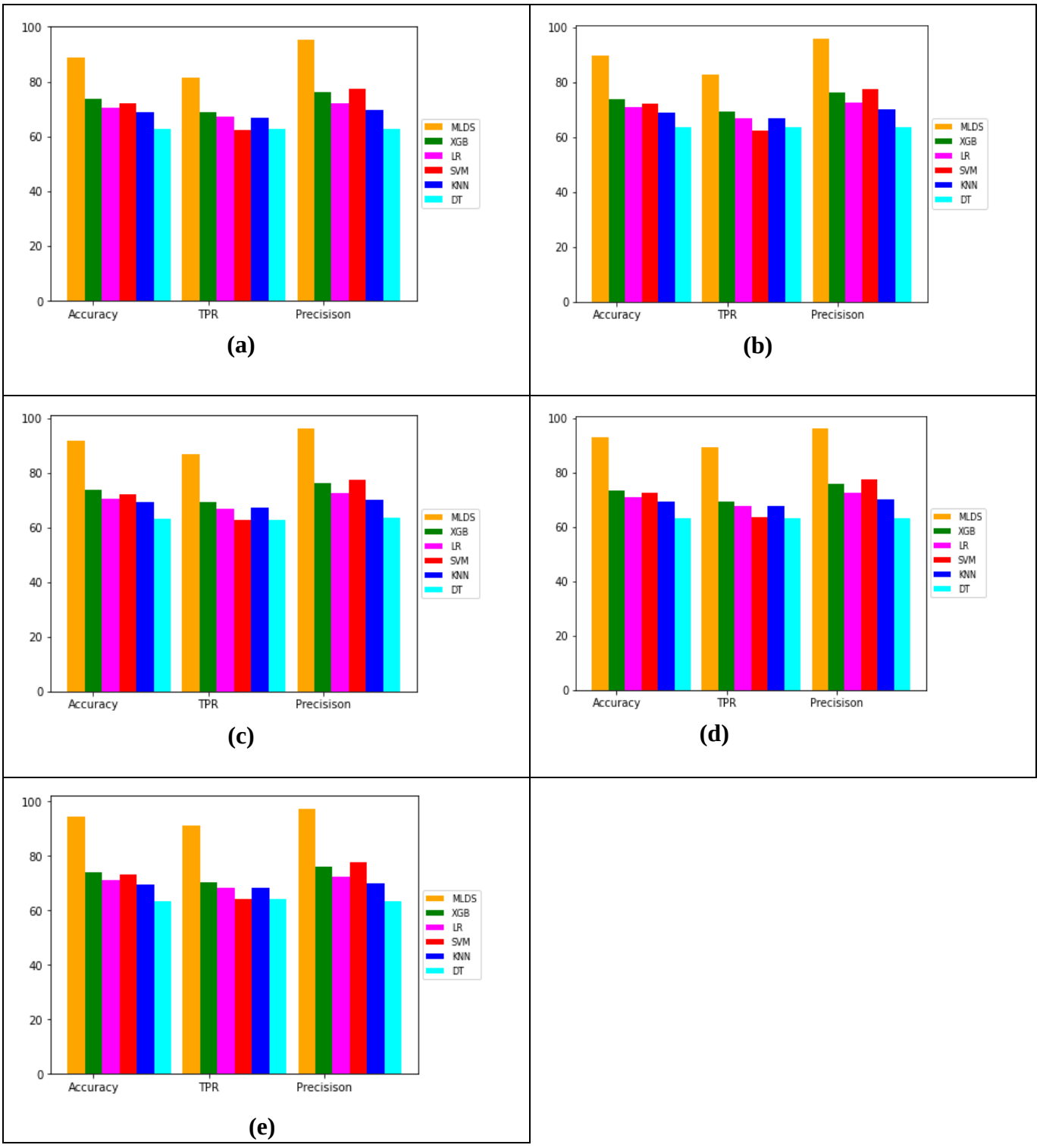

Figure 5.2: Result comparison (Accuracy, TPR, Precision) :(a) 50:50, (b)60:40, (c)70:30, (d)80:20, (e) 87.5:12.5.

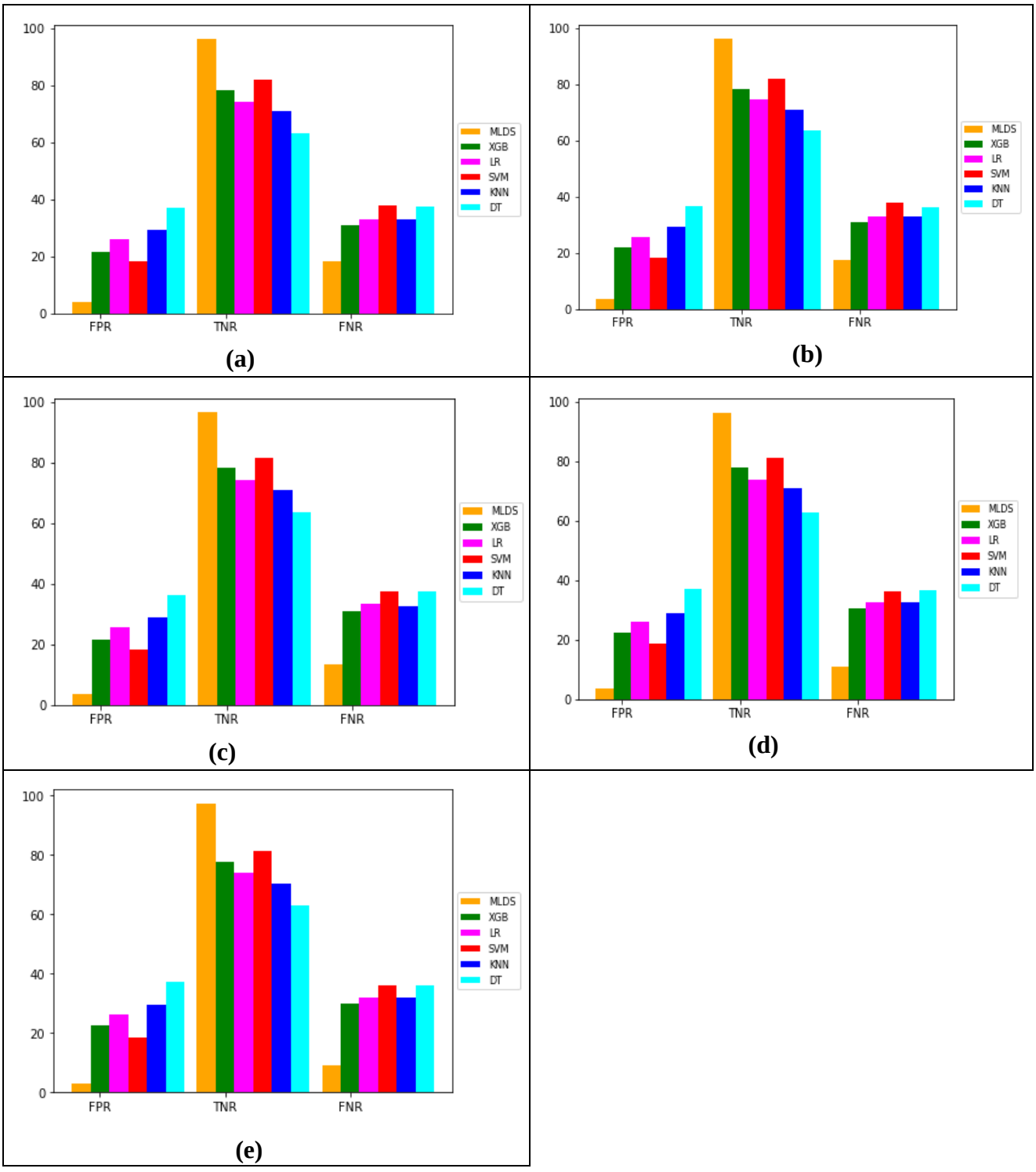

Figure 5.3: Result comparison (FPR, TNR, FNR) :(a) 50:50, (b)60:40, (c)70:30, (d)80:20, (e) 87.5:12.5.

The evaluation of our proposed MLDS has been performed with the confusion matrix and compared with the other classification algorithm's confusion matrix as shown in Table 5.2 and its graphical representations are shown in Figure 5.4, 5.5.

Table 5.2: Confusion matrix result of proposed (MLDS), XGB, LR, SVM, KNN, DT.

Model Name	**Split Ratio**	**Parameter**			
		TP	**FN**	**TN**	**FP**
Proposed MLDS	50:50	14283 out of 17507	3224	16813 out of 17493	680
	60:40	11615 out of 14061	2446	13429 out of 13939	510
	70:30	9123 out of 10525	1402	10106 out of 10475	369
	80:20	6275 out of 7033	758	6707 out of 6967	260
	87.5:12.5	3977 out of 4365	388	4262 out of 4385	125
XGB	50:50	12075 out of 17507	5432	13709 out of 17493	3784
	60:40	9725 out of 14061	4336	10890 out of 13939	3049
	70:30	7278 out of 10525	3247	8200 out of 10475	2275
	80:20	4881 out of 7033	2152	5412 out of 6967	1555
	87.5:12.5	3060 out of 4365	1305	3404 out of 4385	981
LR	50:50	11731 out of 17507	5776	12968 out of 17493	4525
	60:40	9404 out of 14061	4657	10393 out of 13939	3546
	70:30	7025 out of 10525	3500	7789 out of 10475	2686
	80:20	4748 out of 7033	2285	5150 out of 6967	1817
	87.5:12.5	2981 out of 4365	1384	3241 out of 4385	1144
SVM	50:50	10890 out of 17507	6617	14302 out of 17493	3191
	60:40	8763 out of 14061	5298	11400 out of 13939	2539
	70:30	6587 out of 10525	3938	8552 out of 10475	1923
	80:20	4481 out of 7033	2552	5664 out of 6967	1303
	87.5:12.5	2803 out of 4365	1562	3574 out of 4385	811
KNN	50:50	11720 out of 17507	5787	12403 out of 17493	5090
	60:40	9404 out of 14061	4657	9882 out of 13939	4057
	70:30	7079 out of 10525	3446	7444 out of 10475	3031
	80:20	4748 out of 7033	2285	4944 out of 6967	2023
	87.5:12.5	2968 out of 4365	1397	3090 out of 4385	1295
DT	50:50	11000 out of 17507	6507	10996 out of 17493	6497
	60:40	8954 out of 14061	5107	8836 out of 13939	5103
	70:30	6593 out of 10525	3932	6668 out of 10475	3807
	80:20	4455 out of 7033	2578	4385 out of 6967	2582
	87.5:12.5	2793 out of 4365	1572	2752 out of 4385	1633

TP = Total number of accurately identified people who are affected by cardiovascular disease.
TN= Total number of accurately identified people who are not affected by cardiovascular disease.
FP = Total number of incorrectly identified people who are not affected by cardiovascular disease.
FN= Total number of incorrectly identified people who are affected by cardiovascular disease.

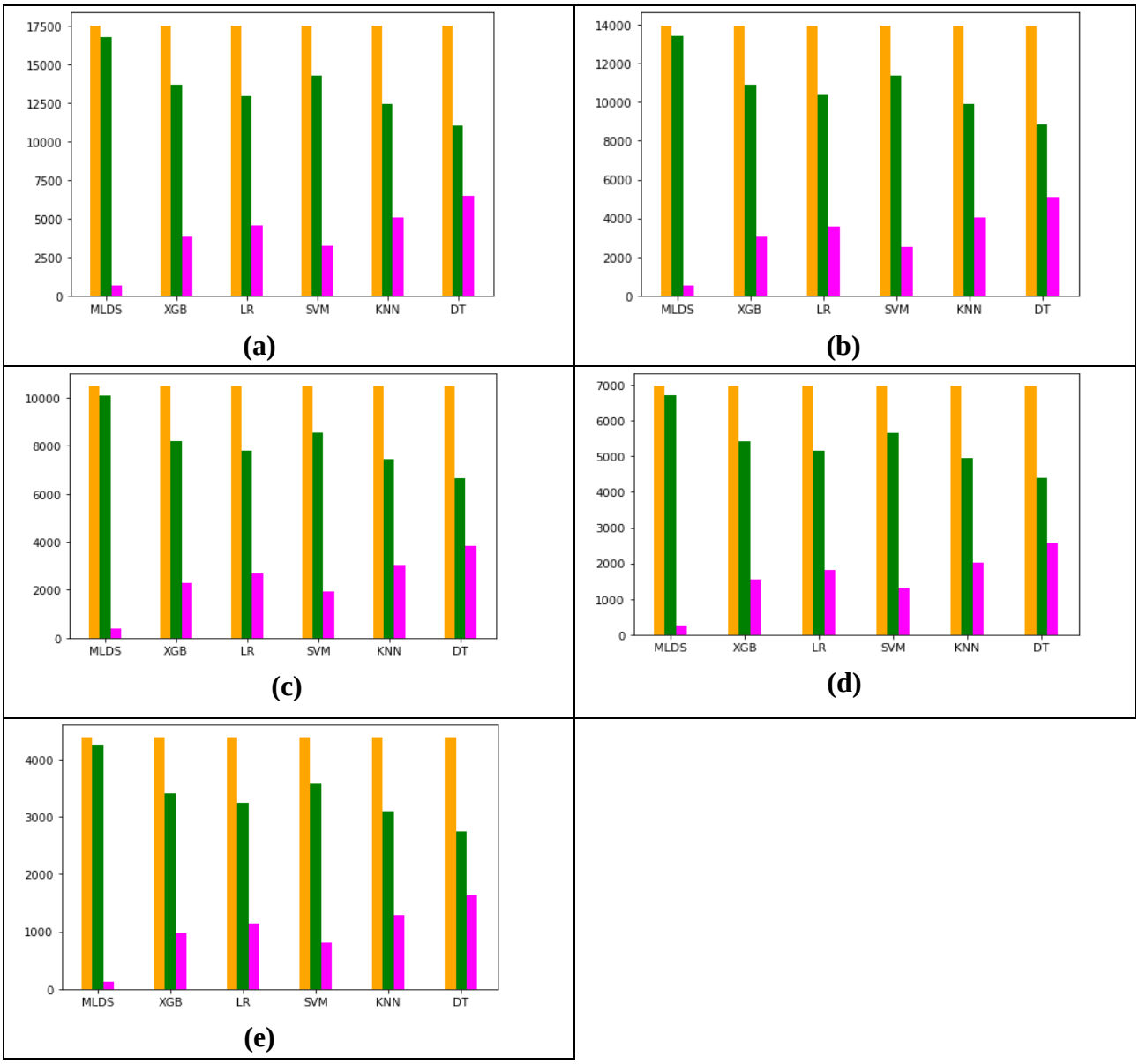

Figure 5.4: Comparison of TN, FP :(a) 50:50, (b)60:40, (c)70:30, (d)80:20, (e) 87.5:12.5.

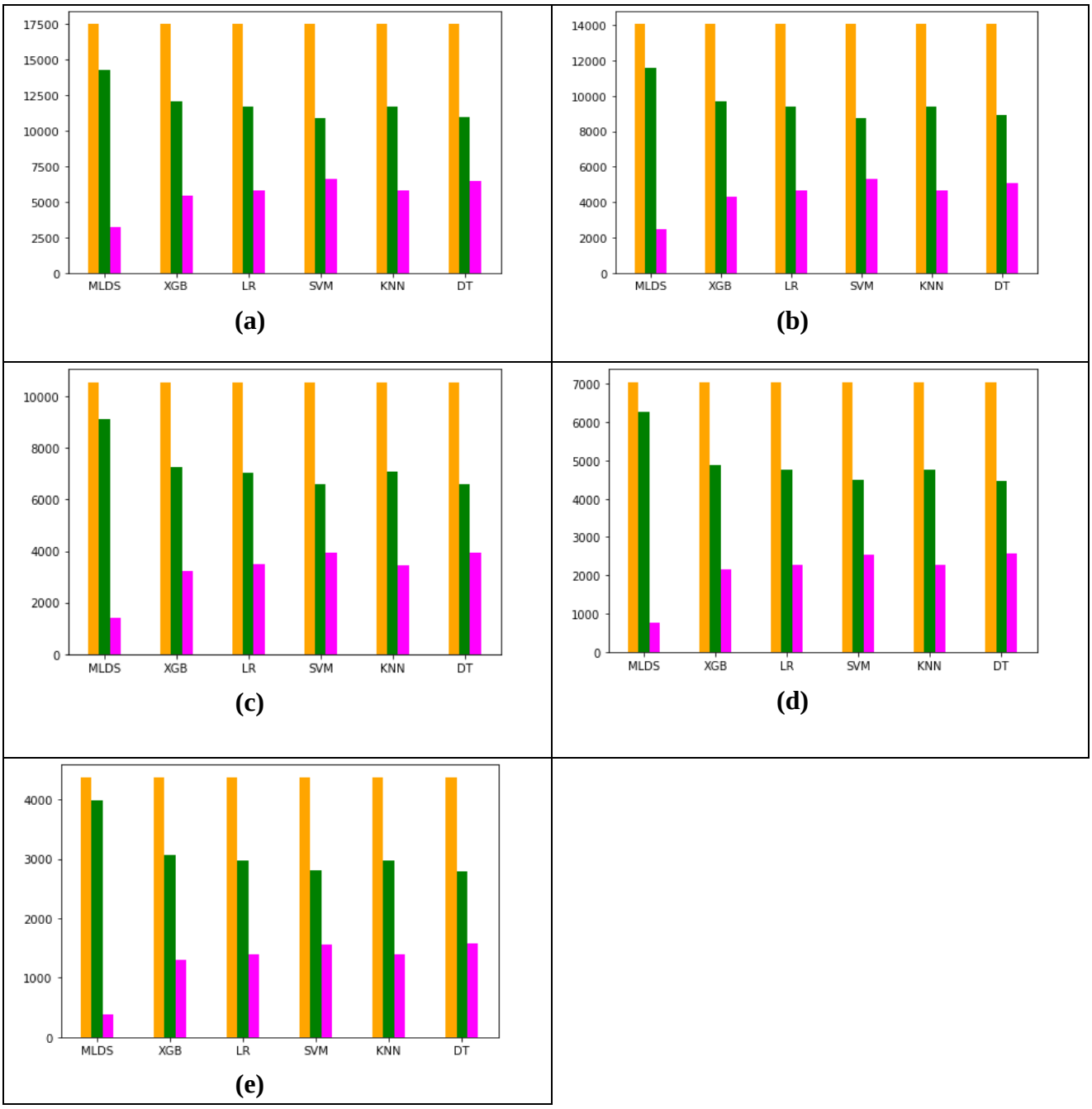

Figure 5.5: Comparison of TP, FN: (a) 50:50, (b) 60:40, (c)70:30, (d)80:20, (e) 87.5:12.5.

Table 5.3: Layer to layer classification result of MLDS (Train/Test Ratio: 50:50).

Layer No. i	Train Data, $x_i = x_{i-1} + (TP, TN)_{i-1}$	Test Data, $y_i = (FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP, TN)_i}{initial, y}$
1 (initial)	35000	35000	25682	9318	GB	73.37
2	60682	9318	3052	6266	NB	82
3	63734	6266	792	5474	RF	84.36
4	64526	5474	147	5327	RF	84.78
5	64673	5327	59	5268	NB	84.948
6	64732	5268	63	5205	RF	85.128
7	64795	5205	32	5173	RF	85.22
8	64827	5173	22	5151	RF	85.282
9	64849	5151	18	5133	RF	85.334
10	64867	5133	16	5117	RF	85.38
11	64883	5117	13	5104	RF	85.417
12	64896	5104	10	5094	GB	85.445
13	64906	5094	15	5079	RF	85.488
14	64912	5079	13	5066	RF	85.525
15	64934	5066	8	5058	NB	85.548
16	64942	5058	12	5046	RF	85.582
17	64954	5046	6	5040	GB	85.6
18	64960	5040	6	5034	RF	85.617
19	64966	5034	3	5031	RF	85.625
20	64961	5031	1	5030	RF	85.628
21	64970	5030	3	5027	GB	85.637
22	64973	5027	4	5023	RF	85.648
23	64977	5023	3	5020	RF	85.657
24	64980	5020	2	5018	RF	85.662
25	**64982**	**5018**	**0**	**5018**	**RF=GB= NB=0**	**85.662**
26	64982	5018	456	4562	KNN	86.965
27	65438	4562	291	4271	KNN	87.797
28	65729	4271	152	4119	KNN	88.231
29	65881	4119	116	4033	KNN	88.562
30 (final)	65997	4003	99	3904	KNN	88.84

Table 5.4: Layer to layer classification result of MLDS (Train/Test Ratio: 60:40).

Layer No. i	Train Data, x_i = x_{i-1} + $(TP, TN)_{i-1}$	Test Data, y_i= $(FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP, TN)_i}{initial, y}$
1 (initial)	42000	28000	20556	7444	GB	73.41
2	62556	7444	2444	5000	NB	82.14
3	65000	5000	786	4214	RF	84.95
4	65786	4214	109	4105	RF	85.339
5	65895	4105	46	4059	RF	85.503
6	65941	4059	35	4024	NB	85.628
7	65976	4024	21	4003	RF	85.703
8	65997	4003	26	3977	RF	85.796
9	66023	3977	16	3961	RF	85.853
10	66039	3961	9	3952	RF	85.885
11	66048	3952	10	3942	GB	85.921
12	66058	3942	14	3928	RF	85.971
13	66072	3928	11	3917	RF	86.01
14	66083	3917	7	3910	GB	86.035
15	66090	3910	5	3905	RF	86.053
16	66095	3905	4	3901	RF	86.067
17	**66099**	**3901**	**0**	**3901**	**RF=GB= NB=0**	**86.067**
18	66099	3901	357	3544	KNN	87.342
19	66456	3544	141	3403	KNN	87.846
20	66597	3403	210	3193	KNN	88.596
21	66807	3193	91	3102	KNN	88.921
22	66898	3102	81	3021	KNN	89.210
23 (final)	66979	3021	65	2956	KNN	89.442

Table 5.5: Layer to layer classification result of MLDS (Train/Test Ratio: 70:30).

Layer No. i	Train Data, x_i = x_{i-1} + $(TP,TN)_{i-1}$	Test Data, y_i= $(FP,FN)_{i-1}$	$(TP,TN)_i$	$(FP,FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N}\frac{(TP,TN)_i}{initial,y}$
1 (initial)	49000	21000	15428	5572	GB	73.46
2	64428	5572	1762	3810	NB	81.857
3	66190	3810	664	3146	RF	85.019
4	66854	3146	78	3068	RF	85.390
5	66932	3068	29	3039	RF	85.528
6	66961	3039	20	3019	RF	85.623
7	66981	3019	19	3000	RF	85.714
8	67000	3000	16	2984	RF	85.790
9	67016	2984	14	2970	NB	85.857
10	67030	2970	12	2958	RF	85.914
11	67042	2958	10	2948	GB	85.961
12	67052	2948	8	2940	RF	86
13	67060	2940	8	2932	RF	86.038
14	67068	2932	9	2923	RF	86.08
15	67077	2923	9	2914	GB	86.123
16	67086	2914	8	2906	RF	86.161
17	67094	2906	6	2900	RF	86.190
18	67100	2900	3	2897	RF	86.204
19	67103	2897	4	2893	RF	86.223
20	67107	2893	6	2887	RF	86.252
21	67113	2887	2	2885	RF	86.261
22	67115	2885	2	2883	GB	86.271
23	67117	2883	4	2879	RF	86.290
24	67121	2879	3	2876	RF	86.304
25	67124	2876	4	2872	GB	86.323
26	67128	2872	3	2869	RF	86.338
27	67131	2869	1	2868	RF	86.342
28	67132	2868	3	2865	RF	86.357
29	67135	2865	4	2861	RF	86.376
30	67139	2861	2	2859	RF	86.385
31	67141	2859	1	2858	GB	86.39
32	67142	2858	1	2857	RF	86.395
33	67143	2857	1	2856	RF	86.4
34	67144	2856	3	2853	RF	86.414
35	67147	2853	3	2850	GB	86.428
36	67150	2850	3	2847	RF	86.442

Layer No. i	Train Data, x_i = x_{i-1} + $(TP,TN)_{i-1}$	Test Data, y_i= $(FP,FN)_{i-1}$	$(TP,TN)_i$	$(FP,FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP,TN)_i}{initial,y}$
37	**67153**	**2847**	**0**	**2847**	**RF=GB= NB=0**	**86.442**
38	67153	2847	265	2582	KNN	87.704
39	67418	2582	156	2426	KNN	88.447
40	67574	2426	105	2321	KNN	88.947
41	67679	2321	70	2251	KNN	89.28
42	67749	2251	64	2187	KNN	89.585
43	67813	2187	27	2160	KNN	89.714
44 (final)	67840	2160	389	1771	KNN	91.566

Table 5.6: Layer to layer classification result of MLDS (Train/Test Ratio: 80:20).

Layer No. i	Train Data, x_i = x_{i-1} + $(TP,TN)_{i-1}$	Test Data, y_i= $(FP,FN)_{i-1}$	$(TP,TN)_i$	$(FP,FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP,TN)_i}{initial,y}$
1 (initial)	56000	14000	10290	3710	GB	73.5
2	66290	3710	1230	2480	NB	82.285
3	67520	2480	494	1986	RF	85.814
4	68014	1986	60	1926	RF	86.242
5	68074	1926	20	1906	RF	86.385
6	68094	1906	16	1890	RF	86.5
7	68110	1890	14	1876	NB	86.6
8	68124	1876	10	1866	RF	86.671
9	68134	1866	10	1856	RF	86.742
10	68144	1856	11	1845	RF	86.821
11	68155	1845	10	1835	RF	86.892
12	68165	1835	3	1832	GB	86.914
13	68168	1832	4	1828	RF	86.942
14	68172	1828	4	1824	RF	86.971
15	68176	1824	3	1821	GB	86.992
16	68179	1821	4	1817	RF	87.021
17	68183	1817	2	1815	GB	87.035
18	68185	1815	1	1814	GB	87.042
19	68186	1814	4	1810	RF	87.071

Layer No. i	Train Data, x_i = x_{i-1} + $(TP, TN)_{i-1}$	Test Data, y_i= $(FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP, TN)_i}{initial, y}$
20	68190	1810	2	1808	RF	87.085
21	68192	1808	1	1807	GB	87.092
22	68193	1807	1	1806	NB	87.1
23	68194	1806	1	1805	GB	87.107
24	68195	1805	1	1804	RF	87.114
25	68196	1804	1	1803	RF	87.121
26	18197	1803	2	1801	RF	87.135
27	68199	1801	1	1800	RF	87.142
28	68200	1800	1	1799	GB	87.15
29	68201	1799	4	1795	RF	87.178
30	68205	1795	1	1794	GB	87.185
31	68206	1794	2	1792	RF	87.2
32	68208	1792	1	1791	NB	87.207
33	68209	1791	1	1790	RF	87.214
34	**68210**	**1790**	**0**	**1790**	**RF=GB= NB=0**	**87.214**
35	68210	1790	161	1629	KNN	88.364
36	68371	1629	108	1521	KNN	89.135
37	68479	1521	69	1452	KNN	89.628
38	68548	1452	45	1407	KNN	89.95
39	68593	1407	38	1369	KNN	90.221
40	68631	1369	225	1144	KNN	91.828
41 (final)	68856	1144	126	1018	KNN	92.728

Table 5.7: Layer to layer classification result of MLDS (Train/Test Ratio: 87.5:12.5).

Layer No. i	Train Data, x_i = x_{i-1} + $(TP,TN)_{i-1}$	Test Data, y_i= $(FP,FN)_{i-1}$	$(TP,TN)_i$	$(FP,FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N}\frac{(TP,TN)_i}{initial,y}$
1 (initial)	61250	8750	6465	2285	GB	73.885
2	67715	2285	776	1509	NB	82.754
3	68491	1509	308	1201	RF	86.274
4	68799	1201	36	1165	RF	86.685
5	68835	1165	20	1145	RF	86.914
6	68855	1145	8	1137	GB	87
7	68863	1137	11	1126	RF	87.131
8	68874	1126	4	1122	GB	87.177
9	68878	1122	5	1117	RF	87.234
10	68883	1117	4	1113	NB	87.28
11	68887	1113	2	1111	RF	87.302
12	68889	1111	4	1107	RF	87.348
13	68893	1107	6	1101	RF	87.417
14	68899	1101	5	1096	RF	87.474
15	68904	1096	1	1095	RF	87.485
16	68905	1095	2	1093	RF	87.508
17	68907	1093	5	1088	RF	87.565
18	68912	1088	2	1086	RF	87.588
19	68914	1086	2	1084	RF	87.611
20	68916	1084	2	1082	GB	87.634
21	68918	1082	1	1081	GB	87.645
22	68919	1081	1	1080	RF	87.657
23	68920	1080	2	1078	RF	87.68
24	**68922**	**1078**	**0**	**1078**	**RF=GB= NB=0**	**87.68**
25	68922	1078	94	984	KNN	88.754
26	69016	984	64	920	KNN	89.485
27	69080	920	44	876	KNN	89.988
28	69124	876	20	856	KNN	90.217
29	69144	856	26	830	KNN	90.514
30	69170	830	21	809	KNN	90.754
31	69191	809	9	800	KNN	90.857
32	69200	800	7	793	KNN	90.937
33	69207	793	12	781	KNN	91.074
34	69219	781	4	777	KNN	91.12

Layer No. i	Train Data, x_i = x_{i-1} + $(TP, TN)_{i-1}$	Test Data, y_i= $(FP, FN)_{i-1}$	$(TP, TN)_i$	$(FP, FN)_i$	Classifier	Total Accuracy, $\sum_{i=1}^{N} \frac{(TP, TN)_i}{initial, y}$
35	69223	777	5	772	KNN	91.177
36	69228	772	3	769	KNN	91.211
37	69231	769	3	766	KNN	91.245
38	69234	766	3	763	KNN	91.28
39	69237	763	3	760	KNN	91.314
40	69240	760	1	759	KNN	91.325
42	69241	759	1	758	KNN	91.337
41	69242	758	1	757	KNN	91.348
42	69243	757	116	641	KNN	92.674
43	69359	641	56	585	KNN	93.314
44	69415	585	31	554	KNN	93.668
45	69446	554	29	525	KNN	94
46 (final)	69475	525	14	511	KNN	94.16

To check more efficiency of our proposed MLDS we have implemented other authors' models and applied our dataset to their model and compared the result with our proposed model as shown in Table 5.8.

Table 5.8. Comparison of MLDS with other authors model using same dataset.

Authors	**Method**	**Accuracy (%)**
Our study	MLDS	94
Ashir Javeed et al. [20]	RSA-RF	72.53
Anna Karen Garate-Escamila et al. [21]	PCA CH-PCA	61.36 66.98
Nabaouia Louridi et al. [19]	(SVM Kernel =Linear)	73.2
Amin Ul Haq et al. [8]	FS: Relief, CL: LR, C=100 FS:mRMR, CL: DT, C =100 FS: LASSO, CL: RF, C =100	71.95 71.73 68.36
Ibomoiye Domor Mienye et al. [22]	CART	74.0

Two different datasets Cleveland (303 instances), Hungarian (294 instances) and Framingham (4240 instances) have been used in this research work. All attributes of these three datasets is described in Chapter 2. These three datasets were commonly used in other authors studies. We

have applied these datasets to our proposed MLDS and tested the accuracy and compared to other authors system as shown in Table 5.9.

Table 5.9: Comparison of MLDS with other authors model using their datasets.

Authors	**Dataset Name/ Number of instances/ Train-Test Ratio**	**Authors Model Accuracy**	**MLDS Accuracy**
Ashir Javeed et al. [20]	Cleveland (297 instances) (70:30)	93.33	98.88
Anna Karen Garate-Escamila et al. [21]	Cleveland (283 instances) (70:30)	98.7	98.88
Anna Karen Garate-Escamila et al. [21]	Hungarian (294instances) (70:30)	99	99.53
Anna Karen Garate-Escamila et al. [21]	CH (577 instances) (70:30)	99.4	99.98
Nabaouia Louridi et al. [19]	Cleveland (294 instances) (80:20)	86.8	98.36
Amin Ul Haq et al. [8]	Cleveland (297 instances) K Fold = 10	89	96.66
Ibomoiye Domor Mienyeet al. [22]	Cleveland (303 instances) (70:30)	93	97.77
Ibomoiye Domor Mienyeet al. [22]	Framingham (4238 instances) (70:30)	91	94

Chapter 6

Conclusion & Future Work

Cardiovascular disease is considered one of the leading causes of death around the world. Early diagnosis can help to prevent the progression of the disease. In this research work, we have proposed an ensemble method-based Multilayer Dynamic System (MLDS). The goal of this model is to predict whether a patient has heart disease or not. The layer to layer prediction process occurs in MLDS. In layer-to-layer prediction, this system can gain additional knowledge from the immediate previous layer to improve the detection accuracy. Irrelevant features reduce the performance of the diagnosis system and increase the computation time. So, the use of multiple feature selection algorithms to select the best features to improve the accuracy of the classification and reduce the execution time of the diagnostic method is another novel aspect of this research. Proposed MLDS has achieved a classification accuracy of 88.84%, 89.44%, 91.56%, 92.72%, and 94.16% based on the different proportions between training and testing 50:50, 60:40, 70:30, 80:20, and 87.5:12.5 respectively based on the Kaggle heart disease dataset. Correctly Classification probability of cardiovascular disease of the proposed system has been pointed out by the AUC curve. The AUC has shown that this system has a 94% chance to classify positive class and negative class when the proportion of training and the testing dataset was 87.5:12.5. The MLDS has gained better accuracy 98.88%, 99.53%, 99.98%, 98.36%, 96.66%, 97.77%, and 94% based on different partitions of Cleveland, Hungarian, Cleveland-Hungarian, and Framingham dataset. The suggested methodology has shown improved performance compared to other machine learning models. Besides, the proposed MLDS can be used to predict the risk of cardiovascular disease and help effectively provide clinical advice.

In future, this proposed method can be enhanced by the implementation of deep learning model to scan the data to search for features that correlate and combine them to enable faster learning without being explicitly told to do so. Future research may try to test more features to extract the hidden pattern and improve the detection accuracy. If a person suffers from cardiovascular disease, what kind of heart disease is it? We will add that module to our future research.

References

[1] Who.int. 2020. Cardiovascular Diseases (Heart Attack, Stroke). [online] Available at: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 [Accessed 6 January 2021]

[2] Who.int. 2021. *WHO | About Cardiovascular Diseases*. [online] Available at: <https://www.who.int/cardiovascular_diseases/about_cvd/en/> [Accessed 6 January 2021]

[3] C. Sowmiya, P. Sumitra .2017. Analytical study of heart disease diagnosis using classification techniques. *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS),* Srivilliputhur, India, 23-25 (March 2017).doi: 10.1109/ITCOSP.2017.8303115

[4] Susmita Ray .2019. A Quick Review of Machine Learning Algorithms. *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon),* Faridabad, India,14-16(February 2019), pp.35-39., doi: 10.1109/COMITCon.2019.8862451

[5] Cardiovascular Disease. [online]. Available at: <https://www.who.int/cardiovascular_diseases/en/cvd_atlas_01_types.pdf?ua=1/> [Accessed 6 January 2021]

[6] Md. Razu Ahmed, S M Hasan Mahmud1, Md Altab Hossin, Hosney Jahan, Sheak Rashed Haider Noori .2018. A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms. *IEEE 4th International Conference on Computer and Communications*, Chengdu, China, 7-10 (December 2018), pp.1951-1955.doi: 10.1109/CompComm.2018.8781022

[7] J. Dhanalakshmi, N. Ayyanathan .2019. An Implementation of Energy Demand Forecast using J48 and Simple K Means. *Fifth International Conference on Science Technology Engineering and Mathematics* ,Chennai, India, 14-15 (March 2019), pp.174-178. doi: 10.1109/ICONSTEM.2019.8918883

[8] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon, Shah Nazir ,Ruinan Sun. 2018. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, Vol. 2018, Article ID 3860146, 2 (Decmber 2018), pp.1-21

[9] R. Saravanan,Pothula Sujatha .2018. A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. *Second International Conference on Intelligent Computing and Control Systems*, Madurai, India,14-15 (June 2018), pp. 945-949. doi: 10.1109/ICCONS.2018.8663155

[10] Peshala T. Gamage, Md Khurshidul. Azad, Amirtaha Taebi, Richard H. Sandler, Hansen A. Mansy .2018. Clustering Seismocardiographic Events using Unsupervised Machine Learning. *IEEE Signal Processing in Medicine and Biology Symposium* , Philadelphia, PA, USA,01 (December 2018). doi: 10.1109/SPMB.2018.8615615

[11] Adhitya Nugraha, Mahendra Arista Harum Perdana, Heru Agus Santoso, Junta Zeniarja, Ardytha Luthfiarta, Ayu Pertiwi .2018. Determining The Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm. *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia,21-22 (September 2018), pp.225-228.doi: 10.1109/ISEMANTIC.2018.8549834.

[12] Hamidreza Ashrafi Esfahani, Morteza Ghazanfari .2017. Cadiovascular disease detection using a new ensemble classifier. *IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, Iran, 22 (December 2017), pp.1011-1014. doi: 10.1109/KBEI.2017.8324946

[13] Kazi Abu Taher , Billal Mohammed Yasin Jisan , Md. Mahbubur Rahman .2019. Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection, *International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, Dhaka, Bangladesh, pp.10-12 (January 2019), pp.643-646. doi: 10.1109/ICREST.2019.8644161

[14] Infochim.u-strasbg.fr. 2020. Correlationattributeeval. [online] Available at: <http://infochim.u-strasbg.fr/cgi-bin/weka-3-9-1/doc/weka/attributeSelection/CorrelationAttributeEval.html> [Accessed 29 December 2020]

[15] Infochim.u-strasbg.fr. 2020. Gainratioattributeeval. [online] Available at: <http://infochim.u-strasbg.fr/cgi-bin/weka-3-9-1/doc/weka/attributeSelection/GainRatioAttributeEval.html> [Accessed 29 December 2020]

[16] Weka.sourceforge.io. 2020. Infogainattributeeval (Weka-Dev 3.9.5 API). [online] Available at: <https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html> [Accessed 29 December 2020]

[17] Faliang Huang,Guoqing Xie,Ruliang Xiao .2009. Research on Ensemble Learning. *International Conference on Artificial Intelligence and Computational Intelligence*, Shanghai, China,7-8 (November 2009), pp.249-252. doi: 10.1109/AICI.2009.235

[18] Shikha Mehta, Priyanka Rana, Shivam Singh, Ankita Sharma, Parul Agarwal .2019. Ensemble Learning Approach for Enhanced Stock Prediction. *Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 8-10 (Aguest 2019). doi: 10.1109/IC3.2019.8844891

[19] Pau Suan Mung, Sabai Phyu .2020. Effective Analytics on Healthcare Big Data Using Ensemble Learning. *IEEE Conference on Computer Applications (ICCA)* , Yangon, Myanmar,27-28 (February 2020) . doi: 10.1109/ICCA49400.2020.9022853

[20] R. Gayathri Devi,P. Sumanjani .2015. Improved classification techniques by combining KNN and Random Forest with Naive Bayesian classifier. *IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, India, 20 (March 2015). doi: 10.1109/ICETECH.2015.7274997

[21] Kumar G Dinesh , K Arumugaraj, Kumar D Santhosh, V Mareeswari .2018. Prediction of Cardiovascular Disease Using Machine Learning Algorithms. *IEEE International Conference on Current Trends toward Converging Technologies*, Coimbatore, India, 1-3 (March 2018), pp. 1-7. doi: 10.1109/ICCTCT.2018.8550857

[22] Senthilkmar Mohan, Chandrasegar Thirumalal,Gautam Srivastava. 2019.Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, Vol. 7, 19 (June 2019) , pp. 81542 – 81554. doi: 10.1109/ACCESS.2019.2923707

[23] Mona Nasr, Essam Shaaban, Ahmed Samir .2019. A proposed Model for Predicting Employees Performance Using Data Mining Techniques: Egyptian Case Study. *International Journal of Computer Science and Information Security (IJCSIS)*,Vol. 17, 1 (January 2019), pp.31-40. ISSN:1947-5500

[24] Latha, C. and Jeeva, S., 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, p.100203

[25] Saiful Islam, Nusrat Jahan,Mst. Eshita Khatun. 2020. Cardiovascular Disease Forecast using Machine Learning Paradigms. *Fourth International Conference on Computing Methodologies and Communication*, Erode, India,11-13 (March 2020), pp.487-490. doi:10.1109/ICCMC48092.2020.ICCMC-00091

[26] Minh Pham ,Jing Lin,Yanjia Zhang . 2018. Diagnosing Voice Disorder with Machine Learning. *IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA,10-13 (December 2018), pp.5263-5266.,doi: 10.1109/BigData.2018.8622250

[27] Shengyu Lu,Beizhan Wang,Hongji Wang,Qingqi Hong .2018. A hybrid collaborative filtering algorithm based on KNN and gradient boosting. *13th International Conference on Computer Science & Education (ICCSE)*, Colombo, Sri Lanka, 8-11 (Aguest 2018), pp.432-436. doi: 10.1109/ICCSE.2018.8468751

[28] Nabaouia Louridi, Meryem Amar, Bouabid El Ouahidi .2019. IDENTIFICATION OF CARDIOVASCULAR DISEASES USING MACHINE LEARNING. *7th Mediterranean Congress of Telecommunications (CMT)*, Fès, Morocco, Morocco, 24-25 (October 2019). doi: 10.1109/CMT.2019.8931411

[29] Ashir Javeed, Shijie Zhou,Liao Yongjian,Iqbal Qasim,Adeeb Noor,Redhwan Nour .2019. An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. *IEEE Access* ,Volume 7 ,7 (November 2019), pp.180235 - 180243 .doi: 10.1109/ACCESS.2019.2952107.

[30] Gárate-Escamila, A., Hajjam El Hassani, A. and Andrès, E., 2020. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, 19, p.100330.

[31] Mienye, I., Sun, Y. and Wang, Z., 2020. An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, 20, p.100402.

[32] National Institute on Aging. 2020. Heart Health And Aging. [online] Available at: <https://www.nia.nih.gov/health/heart-health-and-aging#:~:text=Adults%20age%2065%20and%20older,risk%20of%20developing%20cardiovascular%20disease/>[Accessed 6 January 2021]

[33] Cholesterol And Heart Disease. 2020. [online] WebMD. Available at: https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk#1

[34] Information, H., Overview, D., Problems, P., Diabetes, &., Diabetes, a., Center, T. and Health, N., 2020. Diabetes, Heart Disease, And Stroke | NIDDK. [online] National Institute of Diabetes and Digestive and Kidney Diseases. Available at: <https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke#:~:text=Over%20time%2C%20high%20blood%20glucose,you%20will%20develop%20heart%20disease.&text=People%20with%20diabetes%20tend%20to,age%20than%20people%20without%20diabetes.> [Accessed 29 December 2020]

[35] Cardiosecur.com. 2020. High Blood Pressure – Causes And Connection To Heart Attacks | Cardiosecur. [online] Available at: <https://www.cardiosecur.com/magazine/specialist-articles-on-the-heart/high-blood-pressure-causes-and-connection-to-heart-attacks/>[Accessed 6 January 2021]

[36] US.News. 2020. [online] Available at: <https://health.usnews.com/health-care/for-better/articles/2016-10-19/how-obesity-can-affect-your-heart#:~:text=Obesity%20leads%20to%20heart%20failure,eventually%20lead%20to%20heart%20failure/>[Accessed 6 January 2021]

[37] Cleveland Clinic. 2020. Obesity & Heart Disease. [online] Available at: <https://my.clevelandclinic.org/health/articles/17308-obesity--heart-disease/> [Accessed 6 January 2021]

[38] MedicineNet. 2020. Low Blood Pressure Symptoms, Chart, Causes, And Treatments. [online] Available at: <https://www.medicinenet.com/low_blood_pressure/article.htm/>[Accessed 6 January 2021]