CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Cardiovascular and Chronic Respiratory Diseases Prediction System

by

Wajid Shah

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the
Faculty of Computing
Department of Computer Science

2018

I dedicate my dissertation work to my family, teachers, and friends. The special feeling of gratitude to my loving parents and wife for their love, endless support, and encouragement.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY

ISLAMABAD

CERTIFICATE OF APPROVAL

Cardiovascular and Chronic Respiratory Diseases Prediction System

by

Wajid Shah

MCS143027

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr Syed Nasir Mehmood Shah	AQKICSIT, Kahuta
(b)	Internal Examiner	Dr. Muhammad Arshad Islam	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Aleem	CUST, Islamabad

_____

Dr. Muhammad Aleem

Thesis Supervisor

February, 2018

_____   _____

Dr. Nayyer Masood   Dr. Muhammad Abdul Qadir

Head   Dean

Dept. of Computer Science   Faculty of Computing

February, 2018   February, 2018

Author's Declaration

I, **Wajid Shah** hereby state that my MS thesis titled "**Cardiovascular and Chronic Respiratory Diseases Prediction System**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

Wajid Shah

Registration No: MCS143027

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "*Cardiovascular and Chronic Respiratory Diseases Prediction System*" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

Wajid Shah

Registration No: MCS143027

Acknowledgements

All praise is to ALLAH (S.W.T). The creator and Sustainer of all seen and unseen worlds. First and foremost, I would like to express my gratitude to ALLAH Almighty for his countless blessings upon me to complete this work. Secondly, I would like to express my sincerest thanks to my supervisor **Dr. Muhammad Aleem** for his valuable guidance and assistance. I am sincerely grateful for his support, encouragement, and technical advice in the research area. He has taught me, both consciously and unconsciously, many techniques and strategies during my research. I pray for his wellbeing and success in the future.

I am highly indebted to my parents and my family, for their support, assistance and encouragement throughout the completion my degree. They form the most important part of my life. After ALLAH (S.W.T) they are the sole source of my being in this world. No words can ever be sufficient for the gratitude I have for my parent and for my family. A special thanks to my brothers for their support and encouragement during my studies. I would also like to extend my gratitude to **Mr. Shafiq**, a Ph.D. Scholar at CUST and **Mr. Yasir Noman Khalid**, a Ph.D. Scholar at CUST and a generous friend for his valuable support and guidance.

Finally, I would like to thank my beloved wife. She was always there cheering me up and stood by me through the good times and bad.

I pray to ALLAH (S.W.T) that may He bestow me with true success in all fields in both worlds and shower His blessings upon me for the betterment of all Muslims and the whole Mankind.

Ameen

Wajid Shah

Abstract

Cardiovascular diseases and chronic respiratory diseases are a global threat and approximately cause 19 million deaths worldwide annually. Due to such high deaths, there is needed to tackle the reasons behind these diseases. The research identified several reasons and among them, the most important reason is a lack of ability to predict symptoms of these diseases. In this study, we proposed a system that predicts the vital symptoms of these diseases, thus help to diagnose the diseases and the patient can start treatment on time.

We used the University of Queensland vital signs dataset, containing patient monitoring data and vital signs recorded during 32 surgical situations at Royal Adelaide Hospital. We tackle this problem as a classification task. The proposed system is a part of a PMC-Tele-Health system where vital signs are collected using sensors and based on this data of vital signs, the systems alarm the caregivers and medical experts about the patient potential forthcoming medical situation.

The classifiers used are Support Vector Machine (SVM), Nave Bayes, and Decision Tree. The vital signs collected using sensors are treated as classification features whereas classes are the medical situation of the patient. We used 10 vital signs for cardiovascular prediction and chronic respiratory disease prediction. We deployed a forecasting model using a technique like linear regression and polynomial regression degree 2, 3, and 4. The patient vital signs dataset is divided into two parts to predict the patient situation and to analyze the effect of data size on prediction accuracy. Therefore, 50,000 milliseconds samples with 10 millisecond interval are used to train there regression model (i.e., linear regression, polynomial regression of degree 2, 3, and 4). In addition, regression model (i.e., linear regression, polynomial regression of degree 2, 3, and 4) are trained on 100,000 milliseconds samples with a 10-millisecond interval. This division of sample dataset is actually the 10 minutes (50,000 samples with 10 milliseconds interval) observation of patient and 150 minutes (100,000 samples with 10 milliseconds interval) observation of patient. To predict the next critical situation of patient we analyze that first engagement in this situation will be available in 60 seconds which is caregiver or nurse so we forecast the next 60 seconds data. When providing the medical expert in this situation

will take approximately 3 minutes so, we also forecast the next 03 minutes data. When trains the regression techniques on 50,000 milliseconds samples with 10 millisecond interval to forecast the next 60 seconds of vital signs. Moreover, 100,000 milliseconds samples with 10 millisecond interval to forecast the next 03 minutes of vital signs values. The Linear regression performs well on 50,000 milliseconds samples with 10 millisecond interval of data to forecast the next 60 seconds of vital signs values and polynomial perform well 100,000 samples data to forecast the next 3 minutes vital signs values. To predict the patient critical situation on the forecasted values gain by regression techniques, we apply the three classification techniques Support Vector Machine (SVM), Nave Bayes, and Decision Tree for prediction. When hybrid regression model (combination of two or more regression models) was used with 50,000 samples with 10 millisecond interval to forecast the next 60 seconds, the decision tree obtained the high prediction accuracy by scoring F-measure of 0.90. Similarly, when hybrid regression model was used with 100,000 samples with 10 millisecond interval to forecast the next three minutes, the Nave Bayes obtained the high prediction accuracy by scoring F-measure of 0.99. During this research, I have found that when dataset has 50.000 samples, the decision tree should be used for predicting the heart failure and chronic respiratory disease. If the dataset has 100,000 samples then Nave Bayes should be used for predicting the heart failure and chronic respiratory disease. To conclude that vital signs used in this research have good predictive power and have shown significant accuracy.

Contents

List of Figures

List of Tables

Abbreviations

HR   Heart Rate:

                    Heart rate derived from the ECG sensor.

NBP   Non-invasive Blood Pressure:

                    Blood oxygen saturation derived from pulse oximetry.

NBP(Sys)   Systolic Blood Pressure:

                    Measured with a non-invasive blood pressure cuff.

NBP(Dia)   Diastolic Blood Pressure:

                    Measured with a non-invasive blood pressure cuff.

NBP(Mean)   Mean Blood Pressure:

                    Measured with a non-invasive blood pressure cuff.

SpO2   Oxygen Saturation Percentage:

                    A pulse oximeter uses two frequencies of light

                    (red and infrared) to determine the percentage

                    of hemoglobin in the blood that is saturated with oxygen

ET CO2   End Tidal Carbon Dioxide:

                    Measured using side stream capnography.

IM CO2   Inspired Minimum Carbon Dioxide:

                    Measured using side stream capnography.

ET O2   End-Tidal Oxygen:

                    Concentration measured from the airway.

IN O2   Inspired Oxygen:

                    Concentration measured from the airway.

Chapter 1

Introduction

This chapter gives an introduction and background knowledge of Health Monitoring System, followed by the problem statement. The research question is defined and explained after problem statement; objectives of this research thesis are presented after research question. Finally, the structure of this thesis reports is presented in the last section.

1.1   Introduction

Cardiovascular diseases [1] and chronic respiratory [2] diseases are a global threat, medical experts and researchers are utilizing their best efforts to control the death rate caused by these diseases. The intensity of the problem is highlight by a survey conducted in 2016 by *World Health Organization* (WHO) [3]. In this survey, it has been observed that approximately 19 million people died due to this two diseases in 2016. Every year, approximately 17.5 million people die due to the *Cardiovascular Disease* (CVD) [1] while approximately 1.59 million people die due to the chronic respiratory diseases[2].

In this survey study, authors highlighted several reasons responsible for such high death rates. The study showed that people do not have access to better medical facilities and poor medical systems are one of the main factors. One of the major

shortcomings of such health system is the lack of estimation of important information related to major symptoms of these diseases. On time information related to major symptoms of these diseases play a vital role in reducing the number of deaths globally. The major symptoms include high blood pressure, high blood cholesterol, high blood glucose, and other conditions related to smoking raise the risk of heart and chronic respiratory diseases.

Another important factor behind such high deaths rates due to the *Cardiovascular Disease* (CVD) [1] and *Chronic Respiratory Diseases* (CRD) [2] is the lack of prediction related to the disease vital signs. To provide a prediction of vital signs, we need a healthcare system capable of providing urgent medical information in terms of processed vital signs data and health-alerts (prediction of the patients health). These health-alerts will help the medical specialist and caregivers to pay urgent attention to current and futuristic health situation of the patient. Therefore, we propose a prediction system for caregivers and medical experts. For experimentation, our study uses medical vital signs dataset from the University of Queensland [4]. The dataset contains an extensive variety of patient monitoring data and vital signs that were recorded during 32 surgical situations where patients experienced anesthesia at the Royal Adelaide Hospital. The dataset contains 13 minutes to 5 hours data of the patients. In this research, we employ 10 vital signs related to cardiovascular [1] and chronic respiratory disease[2] predictions. These vital signs include Heart Rate (HR) which is the speed of the heartbeat measured using the number of contractions of the heart per minute(i.e., *heart beats per minutes* (bpm)) and derived from the ECG sensor. *Non-invasive Blood Pressure* (NBP) is the pressure of circulating blood on the walls of blood vessels. *Systolic Blood Pressure* NBP(Sys) pressure is the peak pressure in the arteries, which occurs near the beginning of the cardiac cycle when the ventricles are contracting and measured with a non-invasive blood pressure cuff[5]. *Diastolic Blood Pressure* NBP (Dia) pressure is minimum pressure in the arteries, which occurs near the end of the cardiac cycle when the ventricles are filled with blood and measured with a non-invasive blood pressure cuff. *Mean Blood Pressure NBP* (Mean) measured with a non-invasive blood pressure cuff. *Oxygen Saturation Percentage* (SpO2) stands for

peripheral capillary oxygen saturation, an estimate of the amount of oxygen in the blood and measured with a pulse oximeter. *End-tidal Carbon Dioxide* (ETCO2) the amount of *Carbon Dioxide* (CO2) in exhaled air from the body and measured using side stream Capnography and Capnography is the monitoring of the concentration or partial pressure of *Carbon Dioxide* (CO2) in the respiratory gases. *Inspired Minimum Carbon Dioxide* (IMCO2) measured using side stream Capnography. *End-tidal Oxygen* (ETO2) concentration is the partial pressure of oxygen at the end of exhalation and measured from the airway and *Inspired oxygen* (INO2) concentration measured from the airway.We experimented with several machine learning based forecasting models like linear regression [6] and 2, 3, 4 degrees of polynomial regressions [7]. To train the regression models (for forecasting of the vital signs), we used 50,000 to 100,000 samples of vital signs with a 10 milliseconds interval. These sampling partitions are based on 10 minutes (50,000 samples with 10 milliseconds intervals) of the patient in observation. Second samples of records were based on 150 minutes of observed vital sign data (i.e., containing 100,000 samples with 10 milliseconds interval). For a patient with a critical condition, generally, first 60 seconds are crucial and require urgent help from caregivers or medical experts. Therefore, we focus on the prediction of the vital signs for the next 60 seconds so that in-time medical attention could be provided to the patient. Moreover, considering the availability of a medical expert for a hospitalized patient a 03 minute prediction of the vital signs has been performed too. For forecasting 60 seconds of vital signs, we performed training of the machine learning predictor using 50000 samples. For forecasting next 03 minutes of vital signs, the model was trained using 100000 samples. We observed that the linear regression[6] produces a high prediction accuracy (for next 60 seconds) with a training set of 50,000 samples. Similarly, for a training set of 100000 samples, the linear regression is able to predict vital signs accurately up to next 03 minutes.

This study also investigates the patient situation wither the patient is in a critical situation or not. Our proposed e-health system helps the caregivers to know the critical situation of the patient and the causes of that critical situation (i.e., the vital signs responsible for the criticality). Therefore, the proposed system helps

the caregiver or medical expert to treat the patient accordingly. To represent the patients critical situations, several classes (of the patient current situation) have been proposed that represent a certain situation of a patient (shown in Table 12 and Table 13). To predict the vital signs, several classification techniques were used such as *Support Vector Machine* (SVM) [8], *Nave Bayes*[9], and *Decision Tree*[10]. The forecasted values vital signs by regression techniques are used as input features and assigned classes are then used as output features. First, these classification models(i.e., *Support Vector Machine* (SVM) [8], *Nave Bayes*[9], and *Decision Tree*[10] are trained on assigned class data set(as seen in Table 3.9 and Table 3.10). The forecasting model predicts vital sign values and then these values are provided to train the classification model which predicts a class label. The class label gives us the information about abnormal vital signs which cause *Cardiovascular* [1] and *Chronic Respiratory Diseases*[2]. This early detection helps in the interpretation of abnormal vital signs.

1.2 Purpose

The goal of this study is to develop cardiovascular and chronic respiratory disease prediction system for caregivers and medical experts. The system will be able to provide information about aggravating vital signs related to cardiovascular and chronic respiratory diseases. This early detection will help the medical advisors and caregiver to take necessary early actions.

1.3 Background

Accurate prediction and timely health-alerts help save precious lives. The proposed system will help caregivers and medical experts to react and make a decision on time and prevent delays in the critical medical situations. This system

will forecast the vital signs related values for the potential dangerous condition of a patient. The proposed system predicts the class (condition of the patient) based on the forecasted vital signs. The patients condition is categorized into 38 different classes representing the vital sign related situation. At a certain time, a vital sign value can be one of the three types i.e., normal, high and low. We have defined these 38 classes by consulting the domain experts (i.e., medical experts).Major objectives of the proposed system are to help caregivers and medical experts to take in-time and necessary actions for a potentially worsening situation of a patient. The proposed system will help to save precious lives of patients under observation related to cardiovascular [1]] and chronic respiratory [2] diseases. There is a number of medical systems [11–20], which propose health monitoring and urgent health alerts (prediction of the patients health) based solutions. The number of systems have been developed [11–20] to meet the demands of emerging telehealth services, However, a majority of systems do not provide forecasting of the vital signs using patients historical data to predict the chronic respiratory diseases (i.e., hypoxemic and hypercapnic[2]) and cardiovascular disease (i.e. heart failure [1]) collectively.

1.4   Problem Statement

To the best of our knowledge, there exist no such systems that provide a health monitoring and prediction (forecasted values) for chronic respiratory diseases (i.e. hypoxemic and hypercapnic[2]) and cardiovascular disease (i.e. heart failure[1]). Critical analysis of the literature has led us to the following research gaps:

Research Question 1: For cardiovascular disease, which regression model (i.e., Linear Regression, Polynomial Degree 2,3,4) will be able to forecast the vital signs(i.e., heart rate, blood oxygen, systolic blood pressure, diastolic blood pressure and mean blood pressure) with high accuracy and minimum F-measure error?

Research Question 2: By using the forested vital signs values which classification model (i.e., SVM, Decision Tree, and Naive Bayes) will help to predict the

critical situation of a cardiovascular patient with high accuracy?

Research Question 3: Which regression model (i.e., Linear Regression, Polynomial Degree 2,3,4) will forecast the vital signs for chronic respiratory diseases (such as, systolic blood pressure, carbon dioxide, inspired minimum carbon dioxide, oxygen concentration and inspired oxygen concentration) high accuracy and minimum F-measure?

Research Question 4: Which classification model (i.e., SVM, Decision Tree, and Naive Bayes) helps to predict chronic respiratory diseases by using the forested vital signs values and provide the high accuracy for classification?

Research Question 5: Does hybrid approach (combining two or more regression techniques) will produce high accuracy?

1.5 Scope

In this research, we propose a vital signs and patients health situation based prediction system for chronic respiratory diseases[2] (i.e. hypoxemic, and hypercapnic) and cardiovascular disease (i.e. heart failure[1]). The proposed prediction system is equally useful for both the medical specialists and caregivers. For building a high-accuracy prediction system, we first forecast the vital signs on the basis of historical data of patient by using regression techniques (i.e. Linear Regression[6] and Polynomial regression degree 2,3, and 4) [7]. Forecasted the vital signs values (using the regression techniques) are then analyzed for the performance metrics (i.e., mean square error and mean square percentage error). After the performance analysis, we select most optimal regression model for forecasting the vital sign values. We then employ three different prediction algorithms such as SVM[8], Decision Tree[10], and Nave Bayes[9] to the medical data (vital-signs data related to the mentioned heart failure, hypoxemic and hypercapnic diseases). Next, we analyze the accuracy of the predicted data using different performance metric (i.e., Precision, Recall, and F-measures) for the employed data-mining techniques. Afterwards, we will select the most optimal model (i.e., having higher accuracy or

F-Measures) and use the selected technique in the proposed PMC-telehealth system (shown in Figure.3.1) for predicting the cardiovascular and chronic respiratory diseases. The proposed system tends to be a dynamic tool, incorporating further data for real-time adoption.

1.6 Significance of the Solution

The prediction system will be an integral part of the proposed PMC-tele-health system (shown in Figure.3.1) where the medical data (vital signs) will be collected using different sensors. To forecast the patients situation, the system will be able to analyze and predict the upcoming health situation of a patient. The proposed PMC-tele-health system will be able to alarm the caregivers and medical experts about patients potential forthcoming worsening medical condition.

1.7 Dissertation Organization

In chapter 2 we present comprehensive literature review related to telehealth system based on predicting the diseases conditions or vital signs. Chapter 3 explains the proposed methodology adopted to answer the 04 research questions, identified in this study. Chapter 3 also presents the details related to the dataset and the evaluation metrics to evaluate the results. The experiment and results are discussed in Chapter 4. Moreover, Chapter 4 highlights the importance of the proposed approach. Chapter 5 concludes this dissertation and presents potential future research directions.

Chapter 2

Literature Review

2.1 Introduction

This chapter presents comprehensive literature review based on the state of the art techniques relevant to the proposed research topic. Different data mining and machine learning approaches used by various researchers have been focused in this chapter.

Numbers of systems have been developed to meet the demands of emerging tele-health services. In[21], authors performed a survey of data mining techniques on healthcare data. In this study the combinations of different data-mining techniques such as classification (i.e., *Decision Tree*[10], *Support Vector Machine*[8], *Bayesian*[9], *K-Nearest Neighbors* (KNN) [22]), clustering, association, and regression are explored for the effectiveness of such combinations for disease prediction. Authors in [21] concluded that there is no single data-mining technique, which provides consistent results for all types of diseases. Mostly, the performance of data mining techniques depends on the type of dataset. Therefore, authors concluded that hybrid or integrated data mining technique such as fusion of different classifiers, clustering, and association-based approaches generally produce a more accurate prediction. Authors observed that combination of two techniques likes *Particle Swarm Optimization* SVM (PSO-SVM), *Fuzzy* KNN, *Association Rules*

PSO SVM (AR-PSO-SVM) resulted in achieving more accurate prediction.

Chronic disease has progressively become a major death cause in Taiwan in 2012. In [23], authors determine the impact of important physiological indicators and clinical test values for various chronic illnesses. The authors [23], investigate five chronic diseases i.e., *hypertension, diabetes, cardiovascular disease, a disease of the liver, and renal disease.* Authors used three data mining techniques such as *K-Nearest Neighbor* [22], *linear discriminant analysis*, and *sequential forward selection.* Firstly, linear discriminant analysis with sequential forward selection is used to extract the risk factors. These risk factors are then used by K-Nearest Neighbor model for early-warning. The proposed technique produces up to 80% correct early warning alarms. In contrast with our proposed methodology, authors [23] do not provide prediction related to cardiovascular and chronic respiratory.

In [16], authors unearth some problems related to telehealth systems. These issues are very crucial in the medical and healthcare domains. Authors proposed a framework for a real-time and continua-based care *Guideline Recommendation System* (Cagurs), which utilizes mobile device platforms to provide caregivers (of chronic patients) with real-time care guidelines. Cagurs system enables vital-signs data to be transmitted automatically between different devices using the continua standard [24]. Moreover, the proposed system adopts the episode mining approach to monitor and predict anomalous conditions (i.e., hypertension and hypoxemia) of patients. In addition to that, the system offers related recommendations and care-guidelines to caregivers. These system recommendations and guidelines help the caregivers to provide in time preventative measures. In contrast to our work, forecasting using historical data was not exercised in [16].

Medical experts for predicting the diseases and related symptoms use data mining algorithms. In [25], authors compared different data mining techniques in the healthcare domain. Algorithms such as *Nave Bayes* [9], *J48*[10], *K-Nearest Neighbors* [22], and C4.5[26] were used for classification in order to diagnose diseases like heart disease, cancer, AIDS, brain cancer, diabetes, dengue, and hepatitis C. Comparison study revealed a high accuracy by decision table (up to 97.77%) for cancer disease prediction. Moreover, up to 83.6% accuracy was reported using

C4.5 for diabetes mellitus. The J48 algorithm produced 81.8% accuracy for prediction of HIV/AIDS. Up to 75.97% accuracy for prediction of kidney disease was attained using decision making. For heart disease, an accuracy of up to 60% is achieved using Nave Bayes.

In another similar work by [14], authors proposed an approach to predict the coronary heart disease using data mining techniques. Authors in [14], predict the heart disease in a patient using data mining techniques such as *Iterative Dichotomiser 3* (ID3) [27], *Classification And Regression Tree* (CART) [28] and *Decision Tree* (DT) [10]. The CART produced 83.49% accurate prediction for heart diseases. Moreover, up to 82.50%, the accurate prediction was reported using DT for heart disease. The ID3 algorithm produced up to 72.93% accurate prediction for heart disease.

Data mining technique for disease prediction is also investigated by [15]. Authors in [15]focus on using different data mining techniques(such as *Nave Bayes*[9], *K-Nearest Neighbors* (K-NN) [22], *Decision Tree* [10], C4.5[26], *Artificial Neural Network*(ANN) [29] and *Fuzzy Decision Tree* [30]) to predict heart, breast cancer, diabetes diseases. The authors used *Nave Bayes*[9], *K-Nearest Neighbors* (K-NN) [22], and *Decision Tree* [10] machine learning algorithms for predicting the heart disease. Authors in [15] found that *Nave Bayes* [9] algorithm produced high accuracy up to 74% for predicting the heart disease. For breast cancer prediction, authors in [15] used machine learning algorithms such as C4.5[26], *Artificial Neural Network*(ANN) [29] and fuzzy decision trees [30]. The fuzzy decision tree algorithm, predicted the breast cancer with high accuracy as compared to other data mining techniques such as C4.5 and ANN. Moreover, the authors used homogeneity based algorithm with genetic algorithm for diabetes prediction. Homogeneity with genetic algorithm achieved high prediction accuracy as compared to other data mining techniques (such as *Nave Bayes*[9], *K-Nearest Neighbors* (K-NN) [22], *Decision Tree*[10], C4.5[26], *Artificial Neural Network*(ANN) [29] and *Fuzzy Decision Tree* [30]).

Machine learning based approaches like SVM[8], *Neural networks*[22], *Random forest*[31], *Bagging and Boosting* [32], ANN[29], KNN[22] and *Naive Bayes*[9] were

used by authors in [33]. Authors present a comprehensive review to predict *Parkinson* disease by using machine learning based approaches. Machine learning based approaches are used to classify between healthy and *Parkinson* infected persons. The authors observed that SVM produced high accuracy up to 99.49% for predicting the *Parkinson* disease. The ANN technique produced up to 96.77% prediction accuracy. Moreover, KNN technique predicted the disease with accuracy up to 96.07%. For Parkinson disease, accuracy up to 95% was achieved by using decision tree classification technique. The *Naive Bayes* technique predicted the disease with 74.31% accuracy.

In a similar work [34], authors studied the ability of machine learning based approaches such as *Nave Bayes*[9] and *Decision Tree* (J48) [10] to accurately predict the survivability of patients diagnosed with lung cancer. Survivability is defined as someone who lives beyond a 5year post disease period. The authors found that Naive base approach predicted the lung cancer with up to 92.37% accuracy and the decision tree (i.e., J48) produced up to 94.43% accuracy. Observation showed that both algorithms produce accurate results (up to 90%).

In [17], authors used different data mining techniques on dengue dataset. Authors in[17], used different data mining algorithms such as *Nave Bayes*[9], *Decision Tree* (J48) [10], *Sequential Minimal Optimization* (SMO) [35], *RANDOM Tree*[31], and REP tree to predict the dengue disease in a patient. Observations show that *Nave Bayes*[9] predict the dengue disease by producing 100% accuracy and J48[10] predict the dengue disease with producing 99.70% accuracy. The random tree has produced 91.03% prediction accuracy for dengue disease. Moreover, REP algorithm predicted the dengue disease up to 75.60% accuracy.

In [20], authors recommended the seamless integration of *Wireless Body Area Network* (WBAN) and *Mobile Cloud Computing* (MCC). This integration provides tremendous opportunities for pervasive healthcare systems. The proposed system collects patients vital signs and movements using miniaturized wearable or implantable sensors. The proposed system can be used in home, hospital, and outdoor environment. Patient vital signs data is collected from these locations and forwarded to medical servers (where different data mining techniques were

implemented) for analyses and prediction. Because of MCC elasticity, scalability, and pay-as-you-go pricing model, it can potentially provide huge cost savings, flexible high throughput, and ease of use of WBAN services.

The authors in[13], discussed that in developing countries chronic respiratory diseases represent a challenge to the public health because of the frequency, severity, projected trends, and economic impact of the chronic respiratory diseases. International agencies (i.e., Centers for Disease Control and Prevention[36], WHO[3] etc.) could assist by defining essential drugs and equipment, and encourage the use of generics. WHO recommended the essentials list of drugs, which should be provided at low cost to developing countries. WHO also recommends the awareness program in developing countries to control the tobacco usage which is the major cause of the chronic respiratory diseases.

Intelligent Heart Disease Prediction Systems (IHDPS) has been proposed by authors in [11] for predicting heart-related diseases. Authors in [11], used data mining techniques (such as *Decision Tree*[10], *Nave Bayes*[9], and *Neural Network* (NN)) on cleveland heart disease dataset. IHDPS discovered and extract hidden knowledge (patterns and relationships) associated with heart disease from the historical heart disease dataset. The proposed system helps the medical expert to make the intelligent medical decision to save the heart patient life. In [11], authors used 15 medical vital signs (i.e., *age, sex, chest pain, resting blood pressure, scrum cholesterol, fasting blood sugar, thalach, old peak, slope, CA, thal, slope, exang , electrocardiographic*) from cleveland heart disease dataset to diagnose the heart patient. Nave Bayes algorithm produced the highest prediction accuracy up to 95% on 432 samples of vital signs. Moreover, the decision tree prediction accuracy was up to 94.93% on 106 samples of vital signs. Neural Network algorithm produced up to 93.54% prediction accuracy on 298 samples of vital signs.

In another similar work in[19], authors used different data mining techniques to predict the heart disease. Authors in [19], compared the performance and accuracy of the different algorithm of *Decision Tree* [10] classification techniques such as J48, *Logistic Model Tree* (LMT), and *Random Forest* (RF). In [19], authors used WEKA tool for experiments, performance and accuracy comparison on cleveland

dataset of heart disease. Authors used different heart disease attributes (such as *age, sex, cp, trestbps, chol, exang continuous maximum heart rate achieved, thalach, od peak ST, slope, Ca, thal, class*). The different classification algorithms (i.e., J48, LMT, and RF) of decision tree produced different prediction accuracy on same data shown in Table 2.1.Authors in [19], concluded that J48 technique turned out to be the best classifier for heart disease prediction with up to 56.76% accuracy. The LMT algorithm has prediction accuracy up to 55.77% for heart disease.

TABLE 2.1: Performance of classification techniques.

	J48	Logistic Model Tree Algorithm	Random Forest Algorithm
Prediction Accuracy	56.76%	55.77%	¡55%
Training Error	0.1423221	0.1656716	0
Test Error	0.1666667	0.237931	0.2

In [37], authors proposed the prediction system to predict the parkinson disease in a patient by using the symptoms like olfactory loss feature and sleep disorder. Authors in [37], used data mining techniques such as *Support Vector Machine* (SVM) and *Classification Tree* (CT) [28] for predicting the parkinson disease. Authors used olfactory loss feature from 40 items of *University of Pennsylvania Smell Identification Test* (UPSIT) and *Sleep Behavior Disorder Feature* (SBDF) from *Rapid Eye Movement Sleep Behavior Disorder Screening Questionnaire* (RBDSQ), obtained from the *Parkinson's Progression Marker's Initiative* (PPMI) database. Observation shows that *Support Vector Machine* (SVM) produced up to 85.48% prediction accuracy. Moreover, the classification tree technique produced up to 81.72% prediction accuracy.

In public health scenario, the prediction of heart diseases is the major challenge. In [38], the authors focused on predicting the heart disease by using the data mining techniques such as *Nave Bayes, Decision Tree* (DT), and KNN. Authors in[38],

achieved 52.33% prediction accuracy of heart disease using *Nave Bayes* algorithm. For heart disease, the *Decision Tree* produced 52% prediction accuracy. Moreover, the KNN achieved 45.67% prediction accuracy for heart disease.

In a similar work in [39], authors compared the predictive accuracy between two data mining technique such as nave Bayes and decision tree on different medicals dataset (such as heart dataset, lung cancer dataset and dermatology dataset). Authors in [39], observed that same data mining techniques produced different prediction accuracy on the different medical dataset. The Nave Bayes and decision tree techniques produced up to 80% prediction accuracy on heart disease dataset. For lung cancer disease, decision tree techniques produced up to 90.59% prediction accuracy. Moreover, for lung cancer, Nave Bayes produced up to 82.31% prediction accuracy. For Dermatology disease, Nave Bayes produced up to 82.32% prediction accuracy and decision tree achieved up to 79.61% prediction accuracy.

TABLE 2.2: Comparison of different techniques from the literature review

Year	Data description	Approaches	Accuracy	Ref.
2010	Heart disease dataset	**Naive Bayes**	**52.33%**	[38]
		Decision tree	52%	
		KNN	45.67%	
2012	Lung Cancer (SEER dataset)	**Decision Tree(C4.5)**	**94.43%**	[34]
		Nave Bayes	92.37%	
2013	Cleveland Heart Disease database	**Nave Bayes**	**95%**	[11]
		Decision Tree	94.93%	
		Neural Network	93.54%	
2013	Heart disease dataset (Cleveland Clinic Foundation)	**CART**	**83.49%**	[14]
		DT	82.50%	
		ID3	72.93%	
2013	Heart disease	**Nave Bayes**	**74%**	[15]
	Heart disease	KNN	¡ 74%	
	Heart disease	DT	¡74%	
2013	Cancer	**Decision Tree**	**97.77%**	[25]
	Blood Bank sector	J48	89.9%	
	Diabetes Mellitus	C4.5	82.6%	
	HIV/AIDS	C4.5	81.8%	
	Tuberculosis	KNN	78%	
	Hepatitis C	Decision Rule	73.20%	
	Heart Disease	**Naive Bayes**	**60%**	
2013	Lung cancer dataset	**Decision trees**	**90.59%**	[39]
		Nave Bayes	82.31%	
	Heart dataset	Decision trees	80.71%	
		Nave Bayes	**80.75%**	
	Dermatology	Decision trees	79.61%	
		Nave Bayes	**82.32%**	
2014	Parkinson's disease dataset	**SVM**	**85.48%**	[37]
		Classification Tree	81.72%	
2015	Parkinson's disease dataset	**SVM**	**99.49%**	[33]
		ANN	96.77%	
		KNN	96.07%	
		Decision Tree	95%	
		Random Forest	90.26%	
		Nave Bayes	74.31%	
2015	Dengue Dataset	**Nave Bayes**	**100%**	[17]
		Decision Tree(J48)	99.70%	
		Random Tree	91.03%	
2015	Cleveland database of UCI repository	Decision Tree (J48)	56.76%	[19]
		Random Forest	56%	
		Logistic Model Tree	55.77%	

TABLE 2.3: Comparison of Literature review techniques with my technique

Author		[21]	[23]	[16]	[14]	[25]	[15]	[33]	[34]	[17]	[20]	[11]	[19]
UQM Dataset		✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
CVD		✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓
CRD		✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Regression Technique	LR	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	PND2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	PND3	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	PND4	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Classification Technique	DT	✓	✗	✗	✓	✓	✓	✗	✓	✓	✗	✓	✓
	SVM	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	NB	✓	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✗
Heart Failure Disease	HR	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
	SpO2	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
	NBP(Sys)	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗
	NBP(Dia)	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✓
	NBP(Mean)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Hypoxemic and Hypercapnic Disease	SpO2	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
	imCO2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	etCO2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	etO2	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
	inO2	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗

*(LR: Linear regression model, PND2: Polynomial degree 2, PND3: Polynomial degree 3, PND4: Polynomial degree 4, UQM: University of queen marry, CVD: cardiovascular disease, CRD: Chronic respiratory disease)

Chapter 3

Methodology

3.1 Introduction

In the previous chapter, the critical review has been represented which showed that the majority of studies are based on prediction for the medical system. However, to the best of our knowledge, there exist no such systems that provide a health monitoring and prediction for chronic respiratory diseases (i.e., hypoxemic and hypercapnic[2]) and cardiovascular disease (i.e., heart failure[1]). In this chapter, we discuss the proposed methodology. The first section discusses the PMC-Tele-health system. Then, we discuss proposed methodology, dataset detail, and vital signs related data acquisition. After that data normalization process is discussed. Data mining techniques such as regression model and classification algorithms are also discussed. The remainder of the chapter represents details related to the proposed methodology.

3.2 PMC-Tele-Health System

The proposed system is the main part of the *Patient Medical Condition Tele-health* (PMC Tele-health) system (as shown in Figure.3.1). Figure.3.1 shows the data collection layer. The vital signs related data is collected using different biomedical

sensors. The biomedical sensors transmit the vital signs data to the adjacent routing equipment via wired or wireless transmission. The network routers transfer the aggregated data over the Internet to the *Lifesaver system*. The lifesaver system represents the computing and storage facility to provide services (such as a real-time monitor of patient status, disease status prediction, alert generator, and user interfaces). Different users (such as hospitals, clinics, caregivers, and patients, etc) may acquire multiple lifesaving services using a variety of interfaces such as personal computers, and mobile phones. One of the key components of the lifesaver system is *Critical Condition Vital Signs* (CCVS) prediction system. The proposed prediction system will be based on an optimal (accurate) data mining technique to provide high prediction accuracy. Users layer shown in Figure 1 highlights the several potential users of the system (such as patients, caregivers, clinics, and hospitals). The users of the system will get different forms of data such as *vital signs values, visual charts*, and *health-alerts*. Our system uses the patients biomedical data (vital-signs) from the PMC-Tele-Health system for recommending the best prediction technique for cardiovascular and chronic respiratory diseases. We use UQ [4] vital signs dataset of 32 patients for experimentation purpose. The UQ [4] medical dataset has been collected using medical sensors and represents real hospital traces of 32 different patients. The medical traces contain data of several important vital signs (such as *Heart Rate* (HR), SpO2, NBP (Systolic), NBP (Diastolic), and NBP (Mean)) for cardiovascular disease (i.e., heart failure). Moreover, for chronic respiratory disease (i.e., hypoxemic and hypercapnic), there are several vital signs available in the dataset (such as *SpO2, End-Tidal CO2* (etCO2), *Inspired Minimum CO2* (imCO2), *End-Tidal Oxygen* (etO2), and *Inspired Oxygen* (inO2)).
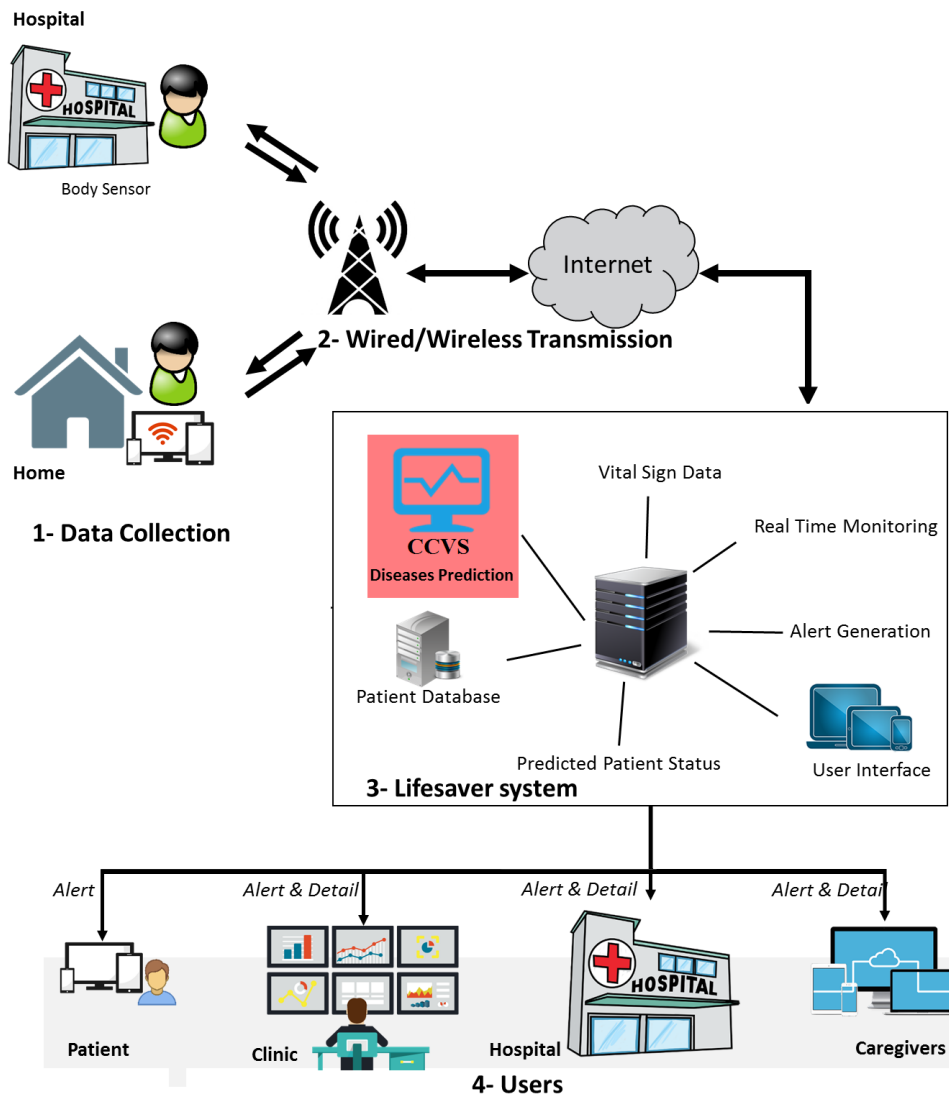
FIGURE 3.1: A Generic framework of PMC-Tele-health system.

3.3 Proposed Methodology

In order to evaluate the research questions, we adopt a unique methodology (as shown in Figure 3.2). For diseases prediction system based experiments, we acquire vital sign data from the University of Queensland[4]. The dataset contains different vital signs for 32 different patients based case studies (as explained in Section 3.4. The dataset is pre-processed to check any anomalies (i.e., Low HR, Low SpO2, High HR etc.). The proposed model aims to forecast vital signs for a future time period of sixty seconds and up to three minutes. The cardiovascular

and chronic respiratory disease related vital signs predication systems are implemented using separate regression and classification models (i.e., for predicting vital signs for next 60 seconds to 03 minutes) as shown in Figure 3.2). Vital signs have different ranges or levels: low, normal, and high to represent the patient condition. Table 9,10,11, shows the three levels (i.e., low, normal, and high) of several vital signs related to cardiovascular and chronic repertory disease. For classification, we use a combination of different ranges (or levels) of vital signs to assign a unique label or patient status class with the consultation and recommendation of the medical specialist. The assignment of different classes for cardiovascular and chronic respiratory disease is explained in Section 3.6. The forecasting model first forecasts the vital sign for a specific time period (i.e., 60 seconds and 3 minutes intervals). Moreover, these forecasted vital signs are passed to the classification model (machine learning based model) which classifies the patients situation into one of the 38 classes mentioned in Table 12,13. The class label provides the overall situation of the patient and helps to know the cause (the important vital sign) of the patient situation for the cardiovascular and chronic respiratory diseases. This early detection helps in the interpretation of abnormal vital signs (i.e., Low HR, High HR, Low SpO2 etc.).

For training and validation of the machine learning model, we split the dataset into two parts. To train the regression models (for forecasting of the vital signs), we used 50,000 to 100,000 samples of vital signs with a 10 milliseconds interval. These sampling partitions are based on 10 minutes (50,000 samples with 10 milliseconds intervals) of the patient for observation. Second samples of records are based on 150 minutes of observed vital sign data (i.e. containing 100,000 samples with 10 milliseconds interval). For a patient with a critical condition, generally, first, 60 seconds are crucial and require urgent help from caregivers or medical experts. Therefore, we focus on the prediction of the vital signs for the next 60 seconds, so that, in-time medical attention could be provided to the patient. Moreover, considering the availability of a medical expert for a hospitalized patient, the prediction for 03 minutes of the vital signs is also performed. For forecasting 60

seconds of vital signs, we first performed training of the machine learning predictor using 50,000 samples. Then for forecasting next 03 minute of vital signs, the model is trained using 100,000 samples. For evaluation purpose, few standard metrics will use for evaluating the performance of each algorithm. These standard measures include *F-Measure, Recall, Precision, Mean Square Error* (MSE), and *Mean Absolute Percentage Error* (MAPE), which are further elaborated in Section 3.11. After applying these metrics, results are comprehensively explained in Chapter 4.
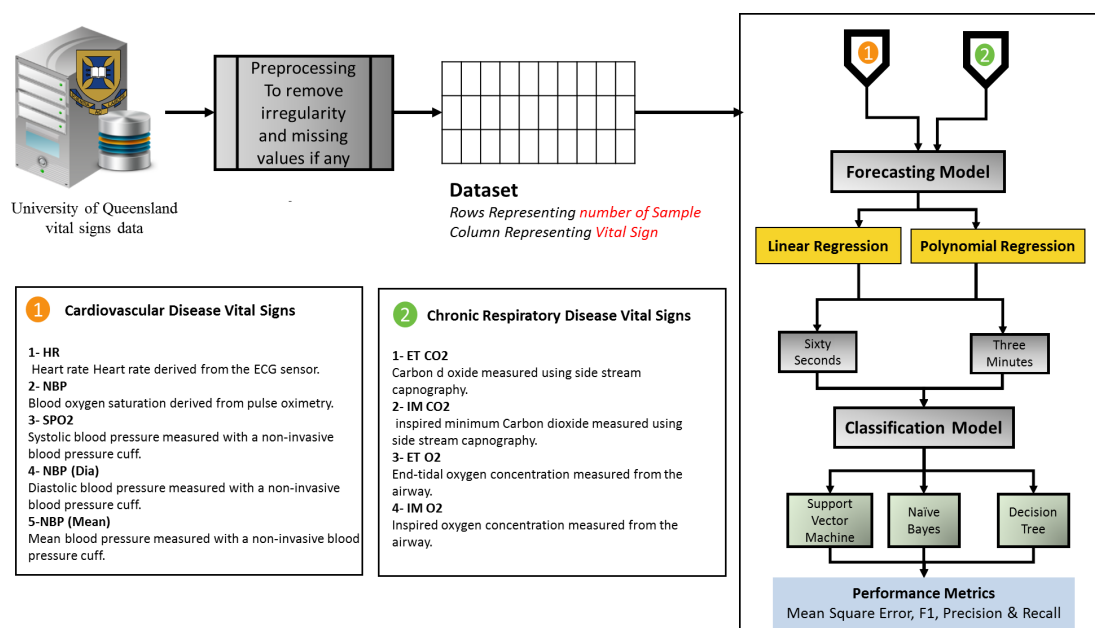
FIGURE 3.2: Proposed methodology for predicting the CVD and CRD.

3.4 Data Collection

The first step is to collect the vital sign data. Our study acquires medical dataset from *University of Queensland*[4]. The dataset contains an extensive variety of patient monitoring data and vital signs. The vital signs data were recorded during 32 surgical situations mentioned in Table 3.2 of anesthesia patients at the *Royal Adelaide Hospital.* The dataset contains 13 minutes to 5 hours data of different patients. The detail about dataset attribute is mentioned in Table

3.1. For the prediction of hypoxemic, hypercapnic, and heart failure, we extract most relevant vital signs (recommended by a medical expert) from the UQ dataset mentioned in Table 3.3. There are several vital signs available in the dataset for cardiovascular disease (i.e., heart failure) such as heart rate, SpO2, NBP (Systolic), NBP (Diastolic). The vital signs of chronic respiratory disease (i.e., hypoxemic and hypercapnic[2]) such as SpO2,end-tidal CO_2, inspired minimum CO_2, end-tidal oxygen, and inspired oxygen

TABLE 3.1: Complete list of vital signs available in UQ dataset

Index	Vital Signs	Index	Vital Signs	Index	Vital Signs	Index	Vital Signs
1	Time	18	[ART (Dia)]	35 35	[Tidal Volume]	52	ECG
2	Relative Time(MS)	19	[ART (Mean)]	36	[Minute Volume]	53	Pleth
3	Clock	20	etDES	37	RR	54	CO2
4	HR	21	inDES	38	[Set Tidal Volume]	55	ART
5	[ST-II]	22	etISO	39	[Set RR]	56	EEG
6	Pulse	23	inISO	40	[Set I:E Ratio]	57	AWP
7	SpO2	24	etSEV	41	[Set PEEP]	58	AWF
8	Perf	25	inSEV	42	[Set PAWmax]	59	AWV
9	etCO2	26	etN2O	43	[Set PAWmin]	60	[AWP-Spiro]
10	imCO2	27	inN2O	44	[Mechanical Ventilation]	61	[AWF-Spiro]
11	awRR	28	MAC	45	[Tidal Volume (Spiro)]	62	[AWV-Spiro]
12	[NBP (Sys)]	29	etO2	46	[Tidal Volume In (Spiro)]	63	[Num Patient Alarms]
13	[NBP (Dia)]	30	inO2	47	[Minute Volume Exp (Spiro)]	64	[Num Technical Alarms]
14	[NBP (Mean)]	31	Temp	48	[Minute Volume In (Spiro)]	65	Alarm
15	[NBP	32	BIS	49	[Lung		

TABLE 3.2: 32 Patients records in millisecond with 10-millisecond interval

Index	Patient File Name	Number of Samples (with 10 ms Interval)
1	case01	683214
2	case02	69222
3	case03	1262888
4	case04	776842
5	case05	340181
6	case06	461345
7	case07	293887
8	case08	265701
9	case09	579384
10	case10	145821
11	case11	331006
12	case12	1422298
13	case13	463055
14	case14	333546
15	case15	62465
16	case16	976898
17	case17	102707
18	case18	61850
19	case19	120627
20	case20	591259
21	case21	245664
22	case22	759641
23	case23	142834
24	case24	266348
25	case25	502863
26	case26	798226
27	case27	929720
28	case28	385673
29	case29	600409
30	case30	248121
31	case31	1189072
32	case32	376933

3.5 Preprocessing and Normalization

The data normalization deals with removal of redundant (duplicate) and partial data. UQ dataset contains different surgical situations for a single patient into many files. Therefore, we combine the different files of a patient into a single file. After that, we remove irrelevant vital signs from these files (can be seen

Table 3.3: 10 most relevant vital signs for hypoxemic, hypercapnic, and heart failure diseases

Index	Vital Sign
1	HR
2	SpO2
3	NBP(Sys)
4	NBP(Dia)
5	NBP(Mean)
6	SpO2
7	imCO2
8	etCO2
9	etO2
10	inO2

in Table 3.4).We make two sets of the dataset, one is for cardiovascular disease with relevant vital signs and the other one is for the chronic respiratory disease. The dataset contain 19,829,316 samples for all 32 patients. After preprocessing, we apply normalization on the relevant vital signs. We remove all the redundant (duplicate) data of the 32 patients. After that, we remove the missing data (i.e., vital sign containing null values or missing values) from 32 patients files (as shown in Table 3.5). After preprocessing up to 10 percent of data is removed. The same process is repeated for both diseases.

Table 3.4: Sample raw data of a patient

Time	Relative Time(ms)	Clock	HR	ST-II	Pulse	SpO2	Perf	etCO2
00:00:00_000	0	20:15	100	-	-	-	-	30
00:00:00_010	10	20:16	99	77	-	-	-	30
00:00:00_020	20	20:17	100	33	33	55	-	30
00:00:00_030	30	20:17	-	22	-	-	-	30
00:00:00_040	40	20:17	88	-	23	-	-	30
00:00:00_050	50	20:17	99	-	-	60	-	30
00:00:00_060	60	20:18	-	44	-	-	-	30
00:00:00_070	70	20:18	50	-	44	-	-	30
00:00:00_080	80	20:18	60	-	-	60	-	30
00:00:00_090	90	20:20	76	44	-	55	-	30
00:00:00_100	100	20:20	56	33	-	-	-	33
00:00:00_110	110	20:25	88	43	-	-	-	33

TABLE 3.5: Sample preprocessed and normalized data of a patient of cardio-vascular diseases

RelativeTime(MS)	HR	SpO2	NBP(Sys)	NBP(Dia)	NBP(Mean)
0	55	97	113	60	72
10	55	100	113	60	72
20	56	100	113	60	72
30	59	100	113	60	72
40	60	100	113	61	72
50	61	100	113	68	72
60	49	100	113	60	72
70	55	100	113	60	72
80	56	100	113	60	72
90	64	100	113	60	72
100	55	100	113	68	72
110	57	100	113	60	72
120	58	100	113	60	72

3.6 Vital signs and classes

In order to label the data with actual classes representing a patients situation, the vital sign global ranges [40] are used with the consultation of a medical specialist. The three patient situation classes (or labels) are represented as high, normal and low along with the specific vital sign. The normal, low and high ranges of the vital signs for both diseases are shown in Table 3.6,3.7,3.8.

TABLE 3.6: Vital Sign Value Range when it is considered as Normal

Vital sign	Normal Ranges
spo2	BETWEEN 94 AND 100
etCO2	BETWEEN 30 AND 43
imCO2	BETWEEN 1 AND 20
etO2	BETWEEN 60 AND 90
inO2	BETWEEN 60 AND 90
HR	BETWEEN 60 AND 100
NBP(Sys)	BETWEEN 100 AND 130
NBP(Dia)	BETWEEN 60 AND 90

We combine these ranges with a single vital sign and to the combination of the vital signs for defining output classes. If all the concerned vital signs have low values/ranges(mentioned in Table 3.7) then the corresponding output class will show Low (Low means the critical condition of the patient and need urgent attention)

TABLE 3.7: Vital Sign Value Range when it is considered as low

Vital sign	Low Ranges
spo2	Less than 94
etCO2	Less than 30
imCO2	Less than 1
etO2	Less than 60
inO2	Less than 60
HR	Less than 60
NBP(Sys)	Less than 100
NBP(Dia)	Less than 60

TABLE 3.8: Vital Sign Value Range when it is considered as High

Vital sign	High Ranges
spo2	Greater than 100
etCO2	Greater than 43
imCO2	Greater than 20
etO2	Greater than 90
inO2	Greater than 90
HR	Greater than 100
NBP(Sys)	Greater than 130
NBP(Dia)	Greater than 90

which is the patient situation. If more than one vital sign have a low range (mentioned in Table 3.7) and all others are in normal range(mentioned in Table 3.6), then output class will name the low vital signs with Low keyword(i.e., Low HR) and ignored the normal vital signs. Same is the case for high range (high range means a critical situation of the patient and need urgent attention mentioned in Table 3.8) vital signs. We also named the combination of high and low vital signs with *Low* and *High* keywords(i.e., Low HR, High spO2). Table 3.9 presents the details related to cardiovascular disease while Table 3.10 presents disease classes for chronic respiratory diseases. These patients classes have been defined after the consultation of the medical experts.

3.7   Chronic Respiratory Disease

Respiratory failure is a syndrome in which the respiratory system fails in one or both of its gas exchange functions i.e. oxygenation and carbon dioxide elimination.

TABLE 3.9: Cardiovascular Diseases Classification model of 38 output class

Class ID	Class Label	Class ID	Class Label	Class ID	Class Label
1	Low	14	Low HR, spO2, NBP(Sys)	27	High spO2 NBP(dia)
2	Normal	15	Low HR spO2, NBP(dia)	28	High NBP(Sys) NBP(dia)
3	High	16	Low HR NBP(Sys) ,NBP(dia)	29	High HR spO2, NBP(Sys)
4	Low HR	17	Low spO2 NBP(Sys), NBP(dia)	30	High HR spO2, NBP(dia)
5	Low spO2	18	Low HR spO2, NBP(Sys) ,NBP(dia)	31	High HR, NBP(Sys), NBP(dia)
6	Low NBP(Sys)	19	High HR	32	High spO2 NBP(Sys) ,NBP(dia)
7	Low NBP(Dia)	20	High spO2	33	High HR spO2, NBP(Sys) , NBP(dia)
8	Low HR, spO2	21	High NBP(Sys)	34	High HR Low NBP(sys) ,NBP(dia)
9	Low HR, NBP(Sys)	22	High NBP(Dia)	35	High HR Low NBP(dia)
10	Low HR, NBP(dia)	23	High HR spO2	36	Low HR High NBP(sys) ,NBP(dia)
11	Low spO2, NBP(Sys)	24	High HR, NBP(Sys)	37	Low HR High NBP(sys)
12	Low spO2, NBP(dia)	25	High HR, NBP(dia)	38	Low SPO2 High NBP(sys)
13	Low NBP (Sys), NBP(dia)	26 26	High spO2, High spO2, NBP(Sys)		

In practice, it may be classified as either hypoxemic or hypercapnic.

TABLE 3.10: Chronic Respiratory Diseases Classification model of 38 output class

Class ID	Class Label	Class ID	Class Label	Class ID	Class Label
1	Low	14	Low etCO2 ,spO2	27 27	High etCO2 ,etO2
2	Normal	15	Low etCO2 ,imCO2	28	High spO2 ,imCO2
3	High	16	Low etCO2 , etO2	29	High spO2 ,etO2
4	Low spO2	17	Low spO2, imCO2	30	High imCO2 ,etO2
5	Low etCO2	18	Low spO2 ,etO2	31	High etCO2 ,spO2,imCO2
6	Low imCO2	19	Low imCO2 ,etO2	32	High etCO2 ,spO2,etO2
7	Low etO2	20	Low etCO2 ,spO2,imCO2	33	High etCO2 ,imCO2,etO2
8	Low inO2	21	Low etCO2 ,spO2,etO2	34	High spO2 ,imCO2,etO2
9	High etCO2	22	Low etCO2 ,imCO2,etO2	35	High etCO2 ,spO2,imCO2 ,etO2
10	High spO2	23	Low spO2 ,imCO2,etO2	36	High etCO2 Low imCO2 ,etO2
11	High imCO2	24	Low etCO2 ,spO2,imCO2 ,etO2	37	High etCO2 ,Low etO2
12	High etO2	25	High etCO2 ,spO2	38	Low etCO2 High imCO2 ,etO2
13	High inO2	26	High etCO2, imCO2		

3.7.1   Hypoxemic respiratory failure (type I)

Hypoxemic respiratory failure is characterized by a Pulmonary Arterial Oxygen Tension (PaO2) lower than 60 mm Hg with a normal or low Pulmonary Arterial Carbon Dioxide Tension (PaCO2). This is the most common form of respiratory failure, and it can be associated with virtually all acute diseases of the lung, which generally involve fluid filling or collapse of alveolar units[41]. Some examples of

type I respiratory failures are cardiogenic or noncardiogenic pulmonary edema, pneumonia, and pulmonary hemorrhage.

3.7.2 Hypercapnic respiratory failure (type II)

Hypercapnic respiratory failure is characterized by an *Arterial Carbon Dioxide Tension* (PaCO2) higher than 50 mm Hg. Hypoxemia is common in patients with hypercapnic respiratory failure who are breathing in room air. The pH depends on the level of bicarbonate, which, in turn, is dependent on the duration of hypercapnia. Common etiologies include drug overdose, neuromuscular disease, chest wall abnormalities, and severe airway disorders (e.g., asthma and *Chronic Obstructive Pulmonary Disease* [COPD]).

3.8 Cardiovascular Diseases

Heart failure is a chronic, progressive condition in which the heart muscle is unable to pump enough blood through to meet the body's needs for blood and oxygen. There are two types of left-sided heart failure i.e., systolic failure and Diastolic failure

3.8.1 Systolic failure

In this type of failure, the left ventricle loses its ability to contract normally. The heart cannot pump with enough force to push enough blood into circulation.

3.8.2 Diastolic failure

Diastolic failure (also called diastolic dysfunction). In this type of failure, the left ventricle loses its ability to relax normally (because the muscle has become stiff).

The heart cannot properly fill with blood during the resting period between each beat.

3.9 Regression model

Regression analysis explores the relationship between a quantitative response variable and explanatory variables. Regression model has two main objectives. Firstly, identify the statically significant relationship between these two variables. Secondly, forecast the new observations on response variable based on explanatory variables. In short, these variable are of two types i.e. dependent variable and independent variable. The dependent variable is the one whose value is required to be forecasted (i.e. vital signs values in our case) whereas the independent variable is used to explain dependent variable as input. Followings are some regression models which are used in this research.

3.9.1 Linear regression

Linear regression[6] is fitting a line to a data set of observations and then using this line to predict unobserved values. It is represented by a linear equation:

$$y = \beta_0 + \beta_1 x \tag{3.1}$$

where:

$y(dependent variable) =$ is function of x,
$\beta_0 \qquad\qquad\qquad = $ is the intercept or constant and
$\beta_1 i \qquad\qquad\qquad = $ is the slope or coefficient

3.9.2 Polynomial regression

Using higher order polynomials to fit our data is Polynomial Regression [7] since our data may not be appropriate to fit into a straight line sometimes. The degree of input variable should be equal to 2 then it is called polynomial regression of degree 2. The degree of input variable should be equal to 3 then it is called polynomial regression of degree 3. Moreover, the degree of input variable should be equal to 4 then it is called polynomial regression of degree 4. In linear regression, we use the linear formula (3.1), however, we can use higher degree (i.e., degree 2, 3, 4) polynomials to fit our data too:

For degree 2:

$$y = \beta_0 x^2 + \beta_1 x + \beta_2 \tag{3.2}$$

For degree 3:

$$y = \beta_0 x^3 + \beta_1 x^2 + \beta_2 x + \beta_3 \tag{3.3}$$

For degree 4:

$$y = \beta_0 x^4 + \beta_1 x^3 + \beta_2 x^2 + \beta_3 x + \beta_4 \tag{3.4}$$

3.10 Classifiers Studies

Various methodologies have been used by different researchers starting from linear regression to neural networks. As per literature review (already discussed in chapter 2), many authors compared different data mining techniques (already discussed in chapter 2) on the medical domain to predict different diseases (in chapter 2). Therefore, after a survey of different authors publication, we have found that SVM, Decision Tree, and Nave Bayes have high prediction accuracy in the medical domain on different diseases. Therefore, we choose Support Vector Machine[8], Decision Tree[10], and Nave Bayes[9] for our experimental study.

3.10.1   Support Vector Machine

Support Vector Machines (SVM) [8] are the most well-known and discussed machine learning algorithms. It remains in mainstream around the time they were created in the 1990s and keep on being the go-to technique for a high-performing algorithm with a little tuning. SVM is a discriminative classifier, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new example. On the basis of this training, the algorithm is able to predict unknown values.

3.10.2   Nave Bayes

In machine learning, Nave Bayes [9] classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes[9] has been studied extensively since the 1950s. Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling.

3.10.3   Decision Tree

Decision tree[10] learning uses a decision tree (as a predictive model) to go from observations about an item to conclusions about the item's target value. The goal is to create a model that predicts the value of a target variable based on several input variables. A decision tree is one of the predictive modeling approaches used in statistics, data mining, and machine learning.

3.11   Evaluation

In order to evaluate the proposed methodology, some standard metrics are applied for evaluating the performance of each algorithm. These metrics are *Precision, Recall, F-Measure, Mean Square Error* and *Mean Absolute Percentage Error*, which

is further elaborated below. To understand precision and recall terms, following figure (Figure3.3) presents the Confusion Matrix:

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

FIGURE 3.3: Confusion Matrix.

Consider an example of laptop database. We have 80 laptops in the database and a search was conducted to find the specific laptops and 60 records were retrieved. In retrieved 60 records, only 45 records were relevant Now, we can define the confusion matrix which is shown in Figure.3.3

True positive = the number of relevant records retrieved is 45 records

False positive= the number of irrelevant records retrieved is 60-45=15 records

False negative= the number of relevant records not retrieved is 80-45=35 records

where Actual shows the realistic record of laptops in the database and Predicted presents the performed search results on the database. Equation (3.5) and Equation (3.6) defines the Precision and Recall.

1. **Precision:**

   The proportion of correct positive classifications(true positives) from cases that are predicted as positive is called precision [42].

   $$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{3.5}$$

2. **Recall:**

   Proportion of correct positive classifications(true positives) from cases that are actually positive is called recall [42].

   $$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3.6}$$

   Now, we can calculate these metrics for search applied on laptop database Precision = TP/(TP+FP) = 45/(45+15)=0.75 Recall = TP/(TP+FN) = 45/(45+35)=0.56

3. **F-Measure:** F-measure is the harmonic mean of Precision and Recall [42].

   $$F - Measure = 2 \times (\frac{Precision \times Recall}{Precision + Recall}) \tag{3.7}$$

   Example: Using the above results of Precision and Recall, in order to find F-measure value, below steps will be followed Precision.Recall=0.75*0.56=0.42 Precision+Recall=0.75+0.56=1.31 F-measure= 2* (0.42/ 1.31)= 2*0.32 = 0.64

4. **Mean Square Error:** Mean Square Error (MSE) is used to evaluate the regression model. MSE measures the expected squared distance between what your predictor predicts for a specific value and what the true value is. Minimum mean square error score represents that the regression model produced high prediction accuracy. Below is the formula used to calculate the mean square error.

   $$MeanSqaureError = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2 \tag{3.8}$$

   where:

   $Y' =$ is a vector of n predictions and

   $Y \ =$ is the vector of observed values of the variable being predicted

5. **Mean Absolute Percentage Error:** Means Absolute Percentage Error (MAPE) is used to evaluate the regression model. MAPE is a measure of

prediction accuracy of a forecasting method and expresses accuracy as a percentage. The formula of mean absolute percentage error is mentioned below.

$$MeanAbsolutePercentageError = \frac{100}{n} \sum_{t=1}^{n} |\frac{A_t - F_t}{A_t} \qquad (3.9)$$

Where is the actual value and is the forecast value? The difference between At and Ft is divided by the Actual value At again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. multiplying by 100 makes it a percentage error.

Chapter 4

Results

4.1 Introduction

This chapter presents the comprehensive results of the proposed methodology, followed in Chapter 3. Initially, different phases of the dataset collection are elaborated. Then results of different regression models (i.e., *linear regression, polynomial degree 2, polynomial degree 3, polynomial degree 4*) and classification models (i.e. *SVM, nave bayes and decision tree*), implemented on considered cases are explained. For each disease, five different patient cases (randomly selected) are examined. Each case considers 60 seconds and 3 minutes forecasting values. The significance of the obtained results is also discussed in detail.

4.2 Cardiovascular Disease Experimentation

For training and prediction of vital signs related to cardiovascular disease, we use UQ dataset[4].The dataset is treated with standard pre-processing and normalization procedures (already discussed in section 3.5). For the prediction of heart failure disease, we extracted most relevant vital signs (recommended by medical experts) from the UQ dataset. There are several vital signs available in the dataset

for cardiovascular disease (discussed in section 3.6). However, we have used vital signs presented in Table 4.1 for predicting heart failure

TABLE 4.1: Input Vital Sign for Cardiovascular Diseases

Feature No	Vital sign
1	spo2
2	HR
3	NBP Sys
4	NBP Dia
5	NBP Mean

4.2.1 Sixty seconds forecasted value

In this section, regression models (i.e., *linear regression[6], polynomial regression (degree 2, degree 3 and degree 4)[7]*) are trained on the 50,000 samples with 10 milliseconds interval for each vital sign(i.e., HR or NBP(Sys) etc.) of heart failure. The trained models (i.e., linear regression[6], polynomial regression (degree 2, 3 and 4)[7]) of each vital sign are then used to forecast the next 60 seconds of the same vital sign. These forecasted values of vital signs are used as an input for classification techniques (like SVM[8], Nave Bayes[9] and Decision Tree[10]) with output class) already explained in section 3.6. As it is supervised learning method, therefore, we have evaluated results through data mining techniques (i.e., regression techniques and classification techniques used in this research) from the results of standard evaluation measures (i.e., best regression model have minimum MSE and MAPE, classification model have precision, recall and f-measure score equal to 1 means that prediction accuracy is high etc.). The same process is repeated for 100,000 samples with 10 milliseconds interval to forecast the next 3 minutes. Results of a performance metric of data mining techniques on a different range of dataset (i.e., 50,000 samples to forecast the 60 seconds and 100,000 samples to forecast the 3 minutes) are shown and discussed in below sections with tables and charts. Hybrid approach (combining two or more regression model) results are also discussed in detail in following sections.

1. **Classification using Linear Regression**

   In this section, the results of linear regression and all the three classification techniques are shown in Table 4.2 and Figure 4.1. In Table 4.2, the first column shows the three types of classification algorithms (i.e., SVM, *nave bayes, and decision tree*) with heart failure vital signs. The second column shows the results of linear regression evaluation methods (i.e., MSE, and MAPE). The low MSE or MAPE value against each vital sign is highlighted. The highlighted value means that linear regression forecast the vital signs with minimum error rate and high accuracy. Linear regression values are repeated for every classification technique to analyze the classification performance on same forecasted values (i.e. vital sign forecasted values obtained by linear regression). The third column presents adopted classification techniques (i.e., SVM, *nave bayes and decision tree*) with precision, recall, and f-measure. The third column exhibits those classification techniques that have predicted high accuracy. The high value (i.e., up to 1) of precision, recall, and f-measure means that the classification technique has predicted the heart failure with high accuracy. We have focused on MAPE results and F-measure results to evaluate the data mining techniques. Results illustrated in Table 4.2 are also presented in the form of a chart. These results are based on linear regression model which is trained on 50,000 samples with a 10-millisecond interval to forecast the next 60 seconds data. The MAPE result shows that SpO2 has produced minimum error rate up to 0.01 therefore, SpO2 has high accuracy. Moreover, HR has produced the second minimum error rate up to 12.5 scores. NBP(Mean) is third in a row according to the MAPE score up to 13.76. NBP(sys) and NBP(Dia) last in the queue with MAPE score are up to 14. The performance metrics column in Table 4.2 has three different types of evaluation techniques (i.e., precision, recall and f-measure) for evaluating the classification techniques. According to the result in Table 4.2, decision tree has achieved high prediction accuracy score up to 0.14. Moreover, nave Bayes prediction accuracy is up to 0.10. SVM prediction accuracy on forecasted values is not satisfactory. Overall linear

regression forecast the SpO2 with MAPE up to 0.01, which is a satisfactory result. Decision tree predicts the heart failure disease with f-measure up to 0.14 prediction accuracy which is also satisfactory.

TABLE 4.2: Classification of 60 Seconds forecasted value by Linear Regression of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.0001	0.01	0.07	0.18	0.10
	HR	56.57	12.5			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	74.34	13.76			
Support Vector Machine	spo2	0.0001	0.01	0	0	0
	HR	56.57	12.5			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	74.34	13.76			
Decision Tree	spo2	**0.0001**	**0.01**	**0.18**	**0.12**	**0.14**
	HR	56.57	12.5			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	74.34	13.76			
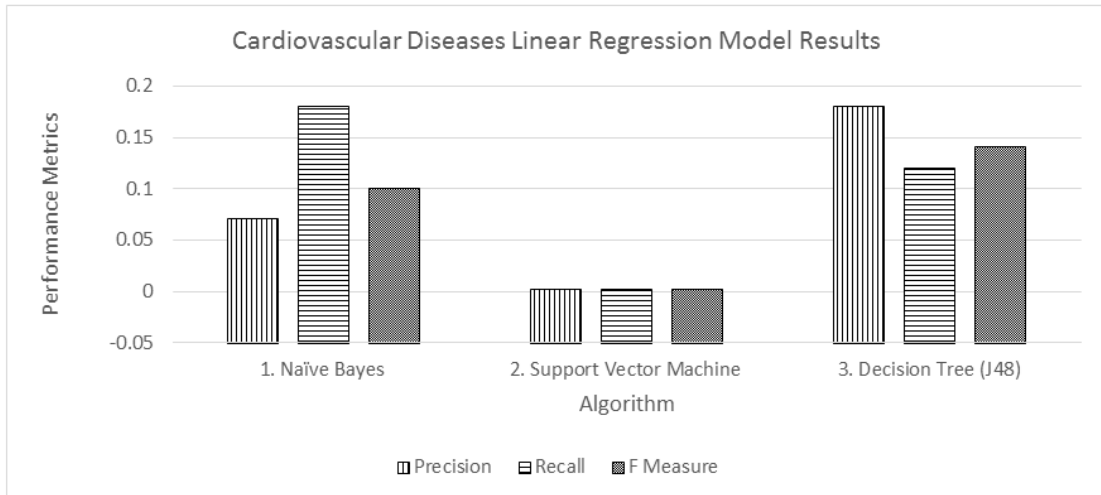
FIGURE 4.1: Classification of 60 Seconds forecasted value by Linear Regression of CVD.

Findings: *Linear regression model is best to forecast the next 60 seconds of SpO2 vital signs of heart failure and has achieved a minimum error rate of MAPE up to 0.01. Decision tree has predicted the heart failure with F-measure score 0.14, which is satisfactory*

2. **Classification using Polynomial Degree Two**

When implementing the same process on polynomial degree two[7] forecasted values, we achieved an improved result. The results of polynomial degree two regression and all the three classification techniques are shown in Table 4.3 and Figure 4.2. If we observe the MAPE closely, polynomial degree two regressions model forecast the SpO2 with minimum error score of MAPE up to 0.16, which is a satisfactory result. Linear regression forecast the HR with second minimum MAPE score up to 3.91. NBP (sys), NBP (Dia) and NBP (Mean) are third in a row and having minimum MAPE score up to 20. According to the result in Table 4.3, SVM has achieved high prediction accuracy score of f-measure up to 0.83. Moreover, decision tree prediction accuracy of f-measure is up to 0.76. Nave Bayes prediction accuracy on forecasted values is not satisfactory as it has achieved f-measure of 0.08. Overall polynomial degree two forecast the SpO2 with MAPE up to 0.16 which is a satisfactory result. SVM has predicted the heart failure disease with f-measure up to

0.83 prediction accuracy which is also quite satisfactory. Decision tree has also played a significant role to predict the heart failure as it has achieved f-measure up to 0.76.

TABLE 4.3: Classification of 60 Seconds forecasted value by Polynomial Degree Two of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.02	0.16	0.05	0.18	0.08
	HR	7.22	3.91			
	NBP(Sys)	465.28	20.6			
	NBP(Dia)	129.13	22.7			
	NBP(Mean)	170.3	20.77			
Support Vector Machine	spo2	0.02	0.16	**0.91**	**0.81**	**0.83**
	HR	7.22	3.91			
	NBP(Sys)	465.28	20.6			
	NBP(Dia)	129.13	22.7			
	NBP(Mean)	170.3	20.77			
Decision Tree	spo2	0.02	0.16	0.82	0.71	0.76
	HR	7.22	3.91			
	NBP(Sys)	465.28	20.6			
	NBP(Dia)	129.13	22.7			
	NBP(Mean)	170.3	20.77			
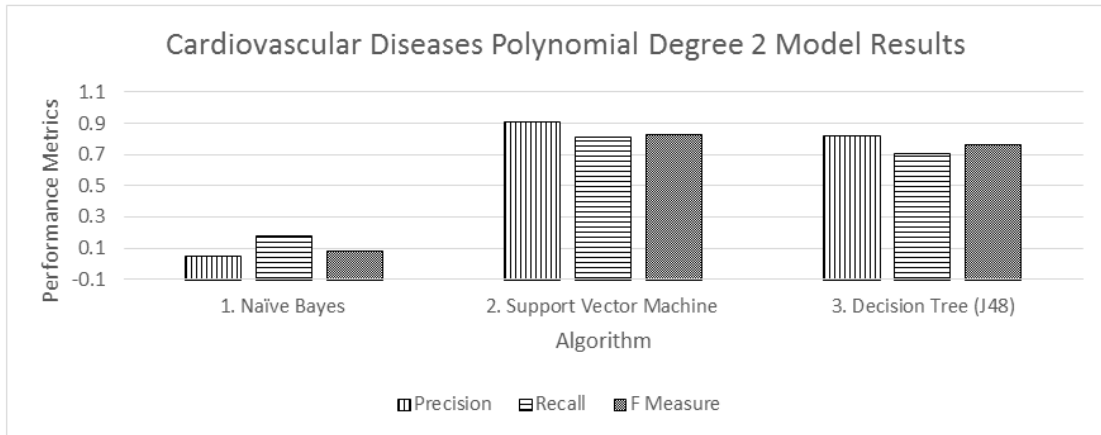
FIGURE 4.2: Classification of 60 Seconds forecasted value by Polynomial Degree Two of CVD

Findings: *Polynomial degree 2 regression model does not play any significant role in forecasting the vital signs of heart failure as compared to linear regression. Linear regression forecasted vital signs of heart failure with minimum MAPE is up to 0.01 and maximum MAPE is up to 14, while polynomial degree 2 regression forecasts the heart failure vital signs with minimum MAPE is up to 0.16 and maximum MAPE is up to 22. Prediction accuracy on polynomial degree 2 forecasted values of vital signs are quite satisfactory as compared to linear regression Polynomial degree 2 achieved maximum prediction accuracy up to 0.83 and minimum prediction accuracy is up to 0.08 while linear regression maximum prediction accuracy score is up to 0.14 and minimum prediction accuracy is up to 0. SVM and decision tree has predicted the heart failure with high accuracy.*

3. **Classification using Polynomial Degree Three** We have also implemented higher degree polynomial model and have evaluated its performance. We applied polynomial degree 3 to evaluate our model further. The results of polynomial degree 3 regression and all the three classification techniques (i.e., SVM, *nave bayes, and decision tree*) are shown in Table 4.4 and Figure 4.3. If we observe the MAPE closely, the error results do not improve with higher degree polynomial regression. The minimum score of MAPE is up to 0.08 and a maximum score of MAPE is up to 45. Moreover, it can be observed that the performance metrics results have not been improved

as well. The minimum prediction accuracy produced by classification techniques on the basis of polynomial degree 3 vital signs forecasted values are up to 0.001 and maximum prediction accuracy is up to 0.15. Polynomial degree 3 forecast the SpO2 with minimum MAPE up to 0.08. Forecasting the HR, NBP (Sys), NBP (Dia) and NBP (mean) by using the polynomial degree 3 is not quite satisfactory. Nave Bayes has predicted the heart failure with F-measure score of 0.15, which is quite high in this experiments and SVM has predicted well by achieving 0.10 F-measure. However, the prediction made by decision tree is not much satisfactory.

TABLE 4.4: Classification of 60 Seconds forecasted value using Polynomial Degree three of CVD

| Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.008	0.08	**0.13**	**0.18**	**0.15**
	HR	694	41			
	NBP(Sys)	1336	30			
	NBP(Dia)	690	45.7			
	NBP(Mean)	756	38			
Support Vector Machine	spo2	0.008	0.08	0.07	0.18	0.10
	HR	694	41			
	NBP(Sys)	1336	30			
	NBP(Dia)	690	45.7			
	NBP(Mean)	756	38			
Decision Tree	spo2	0.008	0.08	0.001	0.001	0.001
	HR	694	3.91			
	NBP(Sys)	1336	30			
	NBP(Dia)	690	45.7			
	NBP(Mean)	756	38			
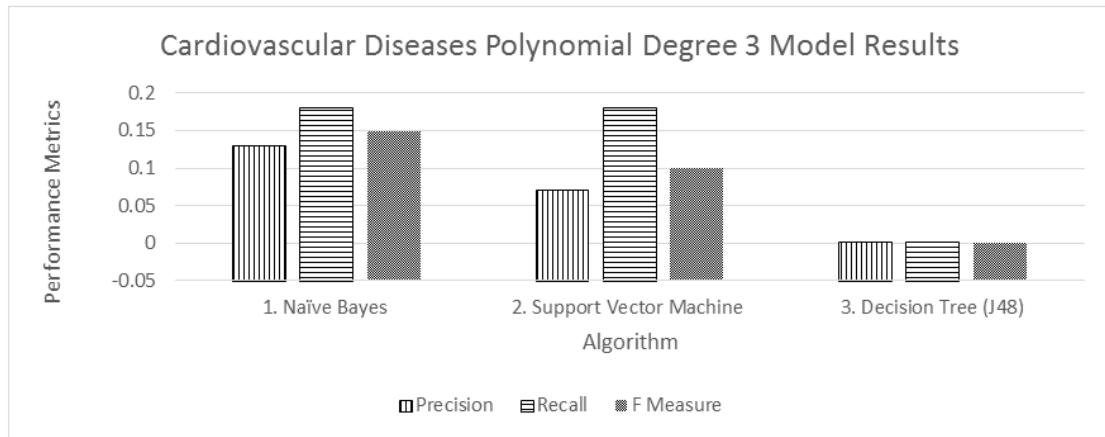
FIGURE 4.3: Classification of 60 Seconds forecasted value by Polynomial Degree three of CVD

Findings: *Findings: Polynomial degree 3 regression model does not play any significant role in forecasting the vital signs of heart failure as compared to polynomial degree 2 and linear regression. Linear regression model forecasted vital signs of heart failure with MAPE score is up to 0.01 to 14. Moreover, polynomial degree 2 regression model forecast the heart failure vital signs with MAPE score is up to 0.16 to 22, while polynomial degree 3 regression model has forecasted the heart failure vital signs with MAPE score is up to 0.08 to 45. Prediction accuracy on polynomial degree 3 forecasted values of vital signs is not much satisfactory as compared to polynomial degree 2 but better than linear regression. Polynomial degree 2 has achieved maximum prediction accuracy up to 0.83 and Polynomial degree 3 has achieved maximum prediction accuracy up to 0.15 while linear regression maximum prediction accuracy score is up to 0.14. In linear regression, decision tree has performed well and in polynomial degree 2 SVM has performed quite satisfactory and Nave Bayes has performed well with polynomial degree 3.*

4. **Classification using Polynomial Degree Four** In order to evaluate results further and to identify the best combination of forecasting measures, we have also evaluated polynomial regression with degree 4. The results of polynomial degree 4 regression and all the three classification techniques are shown in Figure 4.4 and Table 4.5. If we observe the MAPE closely, it can be seen that the error results have not improved with higher degree polynomial

regression. The minimum score of MAPE is up to 1.25 and a maximum score of MAPE is up to 103. We have also observed the performance metrics results havent been improved. The prediction accuracy produced by classification techniques (i.e., SVM, *nave bayes and decision tree*) on the basis of polynomial degree 4 vital signs forecasted values are up to 0.001. Polynomial degree 4 forecast the SpO2 with minimum MAPE up to 1.25. Forecasting the HR, NBP (Sys), NBP (Dia) and NBP (mean) by using the polynomial degree 4 is not quite satisfactory. Nave Bayes, SVM, and decision tree have predicted the heart failure with the f-measure score up to 0.001, which is not much satisfactory.

TABLE 4.5: Classification of 60 Seconds forecasted value by Polynomial Degree Four of CVD

| Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	**2.37**	**1.25**	0.001	0.001	0.001
	HR	2074	59			
	NBP(Sys)	7817	72			
	NBP(Dia)	3595	103			
	NBP(Mean)	4071	87			
Support Vector Machine	spo2	2.37	1.25	0.001	0.001	0.001
	HR	2074	59			
	NBP(Sys)	7817	72			
	NBP(Dia)	3595	103			
	NBP(Mean)	4071	87			
Decision Tree	spo2	2.37	1.25	0.001	0.001	0.001
	HR	2074	59			
	NBP(Sys)	7817	72			
	NBP(Dia)	3595	103			
	NBP(Mean)	4071	87			
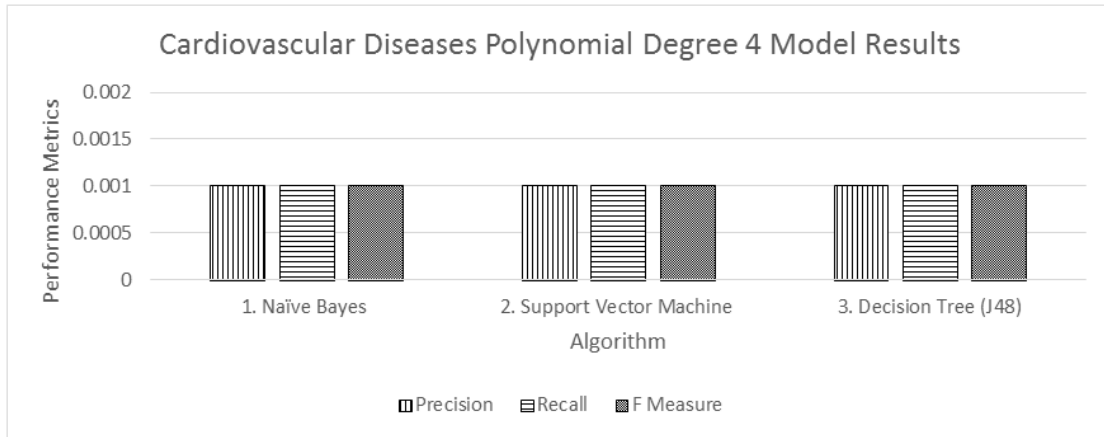
FIGURE 4.4: Classification of 60 Seconds forecasted value by Polynomial Degree Four of CVD

Findings: *Findings: When Polynomial regression degree 4 is used, F-measure score against all classifiers has dropped while MAPE score has increased*

5. **Classification using Hybrid Regression Model** In order to find the best combination of regression techniques, we chose the lowest error regression model from linear regression, polynomial degree 2, 3 and 4 as seen in Table 4.6 and Figure 4.5. Table 4.6 shows that linear regression and polynomial degree 2 forecast the vital signs with minimum error rate (i.e., MAPE score) out of four techniques (i.e., linear regression, polynomial degree 2, 3 and 4). Linear regression forecast with minimum MAPE on four vital signs (i.e., SpO2, NBP (Sys), NBP (Dia), and NBP (Mean)) and polynomial degree 2 forecasts with minimum MAPE on one vital sign (i.e., HR) and remaining polynomial degree 3 and 4 are forecasting the heart failure vital signs with a maximum MAPE score which is not much satisfactory. We used the hybrid approach (combining two or more regression techniques) forecasted values as shown in Table 4.6. These forecasted values of hybrid approach are used as an input to classification techniques (i.e., SVM, nave Bayes and decision tree).

Implementation of Hybrid Regression Model The results of hybrid approach and all the three classification techniques (i.e., SVM, nave Bayes, and decision tree) are shown in Table 4.7 and Figure4.6. We have observed

TABLE 4.6: Hybrid Regression Model for 60 seconds forecasted value for CVD

Best Model	Vital Sign	MAPE
Linear regression	spo2	0.01
Polynomial degree 2	HR	3.91
Linear regression	NBP(Sys)	14.57
Linear regression	NBP(Dia)	14.55
Linear regression	NBP(Mean)	13.76
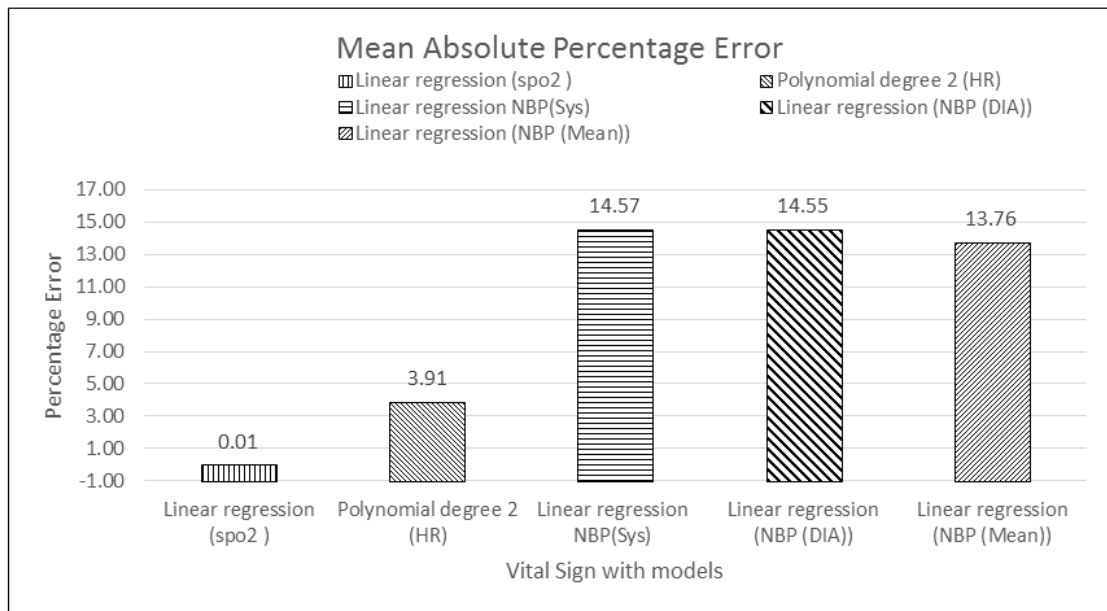
FIGURE 4.5: Hybrid Regression Model for 60 seconds forecasted value for CVD

the performance metrics results and found the improved results as compared to individual regression techniques. Decision Tree has achieved a high prediction accuracy score of F-measure up to 0.90. Moreover, SVM prediction accuracy of f-measure is up to 0.85. Nave Bayes prediction accuracy on forecasted values are not satisfactory and achieved f-measure up to 0.06.

TABLE 4.7: Cardiovascular Diseases Optimal Results With all classifier

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.0001	0.01	0.03	0.18	0.06
	HR	7.22	3.91			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	75.34	13.7			
Support Vector Machine	spo2	0.0001	0.01	0.91	0.84	0.85
	HR	7.22	3.91			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	75.34	13.7			
Decision Tree	spo2	0.0001	0.01	**0.93**	**0.89**	**0.90**
	HR	7.22	3.91			
	NBP(Sys)	233.28	14.57			
	NBP(Dia)	52.54	14.55			
	NBP(Mean)	75.34	13.7			
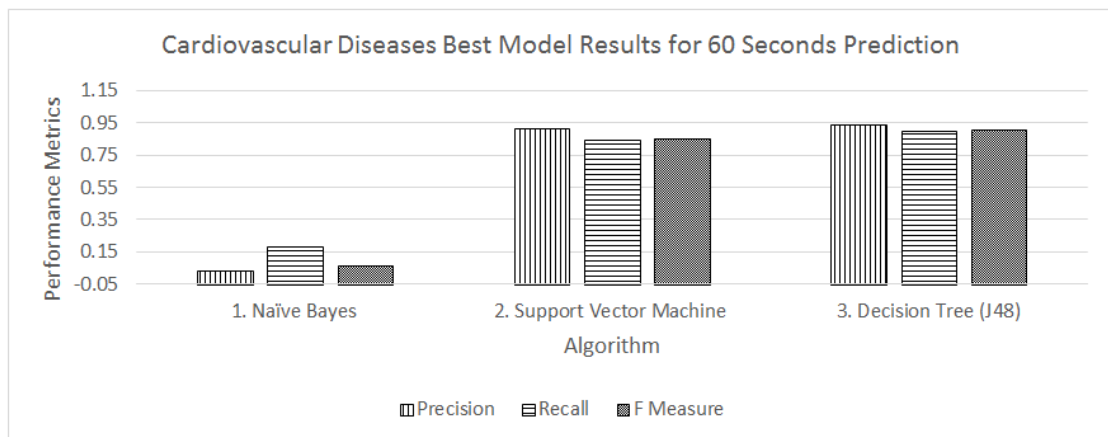
FIGURE 4.6: Cardiovascular Diseases Optimal Results With all classifier

Findings: *With top regression models (hybrid approach), the decision tree has obtained the best result by scoring F-measure of 0.90. While using the individual regression techniques (i.e., linear regression, polynomial degree 2, 3, and 4), polynomial degree 2 has achieved the prediction accuracy score up to 0.83. Moreover, polynomial degree 3 has achieved the prediction accuracy score up to 0.15 and linear regression prediction accuracy score was up to 0.14. Hence, using the hybrid approach, the highest prediction accuracy can be achieved.*

4.2.2   Three minutes forecasted value

Same process (presented in the previous section 4.2.1) is also performed with 3 minutes forecasted value

1. **Classification using Linear Regression** The results of linear regression and classification techniques are shown in Table 4.8 and Figure 4.7. Linear regression forecast the heart failure vital signs for next 3 minutes. Linear regression forecast the SpO2 with minimum MAPE score is up to 0.39. Moreover, NBP(Sys), NBP(Dia), and NBP(Mean) have minimum MAPE score up to 6 and HR have MAPE score up to 9.1. Nave Bayes[9] and Support Vector Machine[8] achieved heart failure prediction accuracy score up to 0.98 F-measure and decision tree achieved up to 0.99 F-measure scores. Therefore, prediction accuracy of heart failure has satisfactory result by using linear regression forecasted values.

TABLE 4.8: Classification of 3 minutes forecasted value by Linear Regression of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	**MSE**	**MAPE**	**Precision**	**Recall**	**F measure**
Nave Bayes	spo2	0.18	0.39	0.98	0.98	0.98
	HR	26.53	9.1			
	NBP(Sys)	39.71	6.97			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	12.7	6.4			
Support Vector Machine	spo2	0.18	0.39	0.98	0.98	0.98
	HR	26.53	9.1			
	NBP(Sys)	39.71	6.97			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	12.7	6.4			
Decision Tree	spo2	0.18	0.39	**0.99**	**0.99**	**0.99**
	HR	26.53	9.1			
	NBP(Sys)	39.71	6.97			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	12.7	6.4			

Findings: *When Linear Regression model is used with three minutes data, all classifier (i.e., SVM, nave bayes and decision tree) predict the heart failure with high accuracy. Decision tree classifier achieved the highest prediction accuracy up to 0.99 f-measure and SVM, nave bayes also predict the heart failure with 0.98 score. MAPE score of linear regression are not too high.*

2. **Classification using Polynomial Degree Two** When implementing the same process on polynomial degree two[7] forecasted values, we does not get improved results. The results of polynomial degree two regression and all the three classification techniques are shown in Table 4.9 and Figure 4.8.
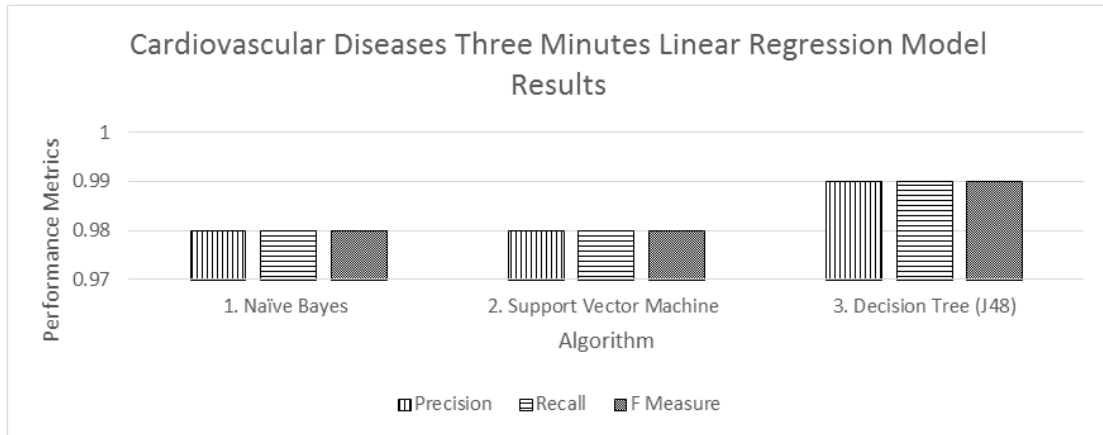
FIGURE 4.7: Classification of 3 minutes forecasted value by Linear Regression of CVD

If we observe the MAPE closely, polynomial degree two regressions model forecast the SpO2 with minimum error score of MAPE up to 0.44, which is satisfactory result. Polynomial degree two forecast the HR with second minimum MAPE score up to 1.86. NBP (sys), NBP (Dia) and NBP (Mean) are third in a row and having minimum MAPE score up to 8. According to the result in Table 4.9, decision tree achieved high prediction accuracy score of f-measure up to 0.83. Moreover, SVM prediction accuarcy of f-measure are upto 0.80. Nave bayes prediction accuracy on forecasted values are not satisfictory and achieved f-measure upto 0.001. Overall polynomial degree two forecast the SpO2 with MAPE up to 0.44 which is satisfactory result. Decision tree predicts the heart failure disease with f-measure up to 0.83 prediction accuracy which is also quite satisfactory.SVM is also good to predict the heart failure because SVM achieved f-measure up to 0.80.

TABLE 4.9: Classification of 3 minutes forecasted value by Polynomial Degree Two of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	**0.23**	**0.44**	0.001	0.001	0.001
	HR	1.5	1.86			
	NBP(Sys)	120	11.836			
	NBP(Dia)	17.7	9.57			
	NBP(Mean)	22.6	8.5			
Support Vector Machine	spo2	**0.23**	**0.44**	1.0	0.67	0.80
	HR	1.5	1.86			
	NBP(Sys)	120	11.836			
	NBP(Dia)	17.7	9.57			
	NBP(Mean)	22.6	8.5			
Decision Tree	spo2	**0.23**	**0.44**	**1.0**	**0.71**	**0.83**
	HR	1.5	1.86			
	NBP(Sys)	120	11.836			
	NBP(Dia)	17.7	9.57			
	NBP(Mean)	22.6	8.5			

Findings: *Polynomial degree 2 regression model is not good in forecasting the vital signs of heart failure as compared to linear regression. Linear regression model forecasted vital signs of heart failure with MAPE score is up to 0.39 to 9.1, and polynomial degree 2 regression model forecast the heart failure vital signs with MAPE score is up to 0.44 to 11.8. Prediction accuracy on polynomial degree 2 forecasted values of vital signs are not satisfactory as compared to linear regression. Polynomial degree 2 achieved maximum prediction accuracy up to 0.83 while linear regression maximum prediction*
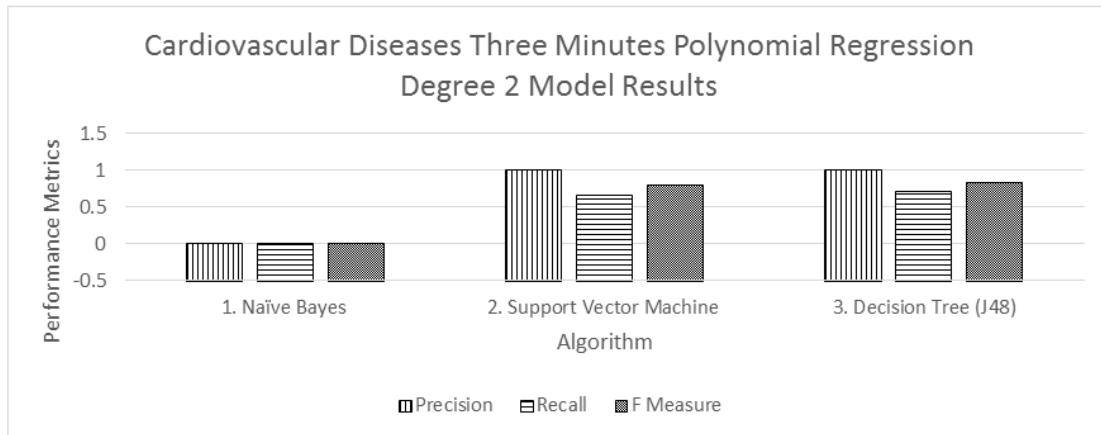
FIGURE 4.8: Classification of 3 minutes forecasted value by Polynomial Degree Two of CVD

accuracy score is up to 0.99. In linear regression decision tree quite satisfactory and in polynomial degree 2 decision tree performed well.

3. **Classification using Polynomial Degree Three** We also implemented higher degree polynomial model and evaluated its performance. We applied polynomial degree three to evaluate our model further. The results of polynomial degree three regression and all the three classification techniques (i.e., SVM, nave bayes, and decision tree) are shown in Table 4.10 and Figure 4.9. If we observe the MAPE closely, the error results do not improve with higher degree polynomial regression. The minimum score of MAPE is up to 0.79 for SpO2 vital sign and maximum score of MAPE is up to 26.8 for NBP(Dia). We also observe the performance metrics results that do not improve as well. The minimum prediction accuracy produced by classification techniques on the basis of polynomial degree 3 vital signs forecasted values are up to 0.05 and maximum prediction accuracy is up to 0.76. Polynomial degree 3 forecast the SpO2 with minimum MAPE up to 0.79. Forecasting the HR and NBP (Sys) by using the polynomial degree 3 is satisfactory with minimum MAPE up to 2. Moreover, NBP (Dia) and NBP (mean) by using the polynomial degree 3 are not quite satisfactory with minimum MAPE up to ¿18. Decision tree predict the heart failure with f-measure score is up to 0.76, which is high in this experiments and SVM are also predict good with up to

0.74 f-measure. Moreover, decision tree prediction is not satisfactory, which is 0.05 f-measure.

TABLE 4.10: Classification of 3 minutes forecasted value by Polynomial Degree three of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	**0.79**	**0.79**	1.0	0.03	0.05
	HR	28.4	1.86			
	NBP(Sys)	5.88	2.05			
	NBP(Dia)	163.41	26.80			
	NBP(Mean)	124.6	18.6			
Support Vector Machine	spo2	**0.79**	**0.79**	1.0	0.59	0.74
	HR	28.4	1.86			
	NBP(Sys)	5.88	2.05			
	NBP(Dia)	163.41	26.80			
	NBP(Mean)	124.6	18.6			
Decision Tree	spo2	**0.79**	**0.79**	**1.0**	**0.61**	**0.76**
	HR	28.4	1.86			
	NBP(Sys)	5.88	2.05			
	NBP(Dia)	163.41	26.80			
	NBP(Mean)	124.6	18.6			

Findings: *Polynomial degree 3 regression model is not good in forecasting the vital signs of heart failure as compared to polynomial degree 2 and linear regression. Linear regression model forecasted vital signs of heart failure with MAPE score is up to 0.39 to 9.1, and polynomial degree 2 regression model forecast the heart failure vital signs with MAPE score is up to 0.44 to 11.8, while polynomial degree 3 regression model forecast the heart failure vital signs with MAPE score is up to 0.79 to 26.8. Prediction accuracy*

FIGURE 4.9: Classification of 3 minutes forecasted value by Polynomial Degree three of CVD

on polynomial degree 3 forecasted values of vital signs are not satisfactory as compare to polynomial degree 2 and linear regression. Linear regression maximum prediction accuracy score is up to 0.99. Polynomial degree 2 achieved maximum prediction accuracy up to 0.83 while Polynomial degree 3 achieved maximum prediction accuracy up to 0.76. In linear regression , polynomial degree 2 and polynomial degree 3 decision tree perform good.

4. **Classification using Polynomial Degree Four** In order to evaluate further and to find the best combination of forecasting we also evaluated polynomial regression with degree four. The results of polynomial degree 4 regression and all the three classification techniques are shown in Table 4.11 and Figure 4.10. If we observe the MAPE closely, the error results do not improve with higher degree polynomial regression. The minimum score of MAPE is up to 0.99 and maximum score of MAPE is up to 13.8. We also observe the performance metrics results are improved. The prediction accuracy produced by classification techniques (i.e., SVM, nave bayes and decision tree) on the basis of polynomial degree 4 vital signs forecasted values are up to 0.98. Polynomial degree 4 forecast the NBP (Sys) with minimum MAPE up to 0.99 and SpO2 with up to 1.08. Forecasting the HR, NBP (Dia) and NBP (mean) by using the polynomial degree 4 are satisfactory. Nave bayes, SVM and decision tree predict the heart failure with f-measure score is up to 0.98, which is quite satisfactory.

TABLE 4.11: Classification of 3 minutes forecasted value by Polynomial Degree Four of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	1.48	1.083	0.98	0.98	0.98
	HR	75.5	13.8			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	14.63	8.51			
	NBP(Mean)	11.23	5.92			
Support Vector Machine	spo2	1.48	1.083	0.98	0.98	0.98
	HR	75.5	13.8			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	14.63	8.51			
	NBP(Mean)	11.23	5.92			
Decision Tree	spo2	1.48	1.083	**0.98**	**0.98**	**0.98**
	HR	75.5	13.8			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	14.63	8.51			
	NBP(Mean)	11.23	5.92			

Findings: *When polynomial degree four models are used with three minutes forecasted data, all classifier achieved 0.98 high f-measures. Linear regression prediction accuracy is 0.99, which is higher than polynomial degree four 0.98. Polynomial degree four achieved high prediction accuracy as compared to polynomial degree 2 with 0.83 and polynomial degree 3 with 0.76.*

5. **Classification using Hybrid Regression Model** In order to find the best combination, we choose the lowest error regression model from linear regression, polynomial degree 2, 3 and 4 as seen in Table 4.12 and Figure 4.11. Table 4.12 shows that linear regression, polynomial degree 2 and polynomial
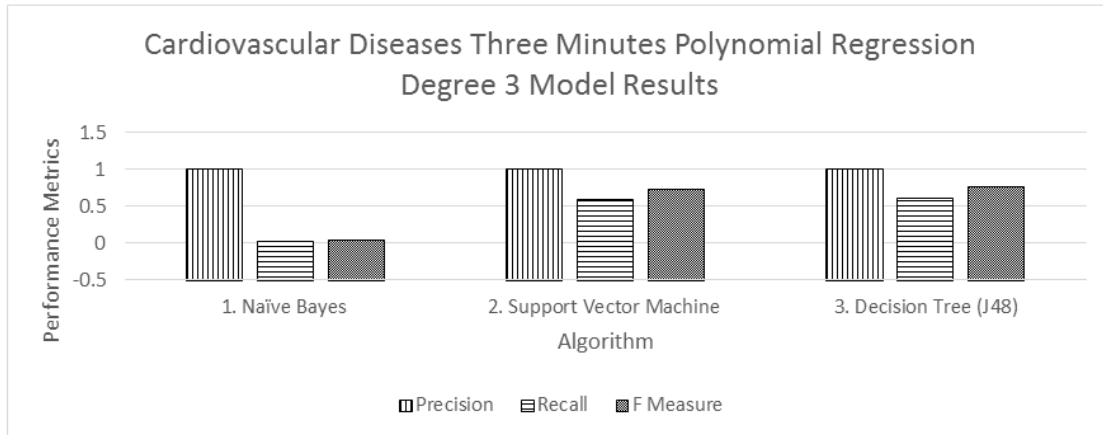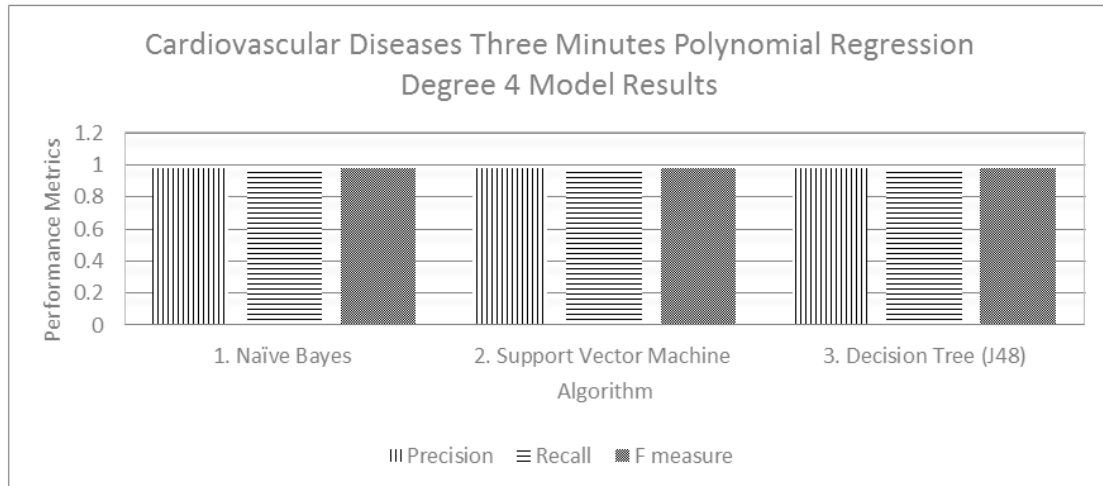
FIGURE 4.10: Classification of 3 minutes forecasted value by Polynomial Degree
Four of CVD

degree 4 forecast the vital signs with minimum error. Linear regression and
polynomial degree 4 forecast with minimum error rate(i.e., MAPE) on two
vital signs and polynomial degree 2 forecast with minimum error rate(i.e.,
MAPE) on the only HR and remaining polynomial degree 3 is not performed
well. We use these top rank models in classifier to get better classification
results.

TABLE 4.12: Hybrid Regression Model for 3 minutes forecasted value for CVD

Best Model	Vital Sign	MAPE
Linear Regression	spo2	0.39
Polynomial degree 2	HR	1.86
Polynomial degree 4	NBP(Sys)	0.99
Linear Regression	NBP(Dia)	5.77
Polynomial degree 4	NBP(Mean)	5.92

The results of all hybrid approach and classification techniques are shown
in Table 4.13 and Figure 4.12. Firstly we applied Nave Bayes[9] classifier
on the three minutes data. It performs well and achieved 0.96. Moreover,
when the forecasted vital sign is provided support vector machine, its scored
highest F measure of 0.99. Similarly, when best regression model is applied,
decision tree score the accuracy with the F-measure of 0.98.
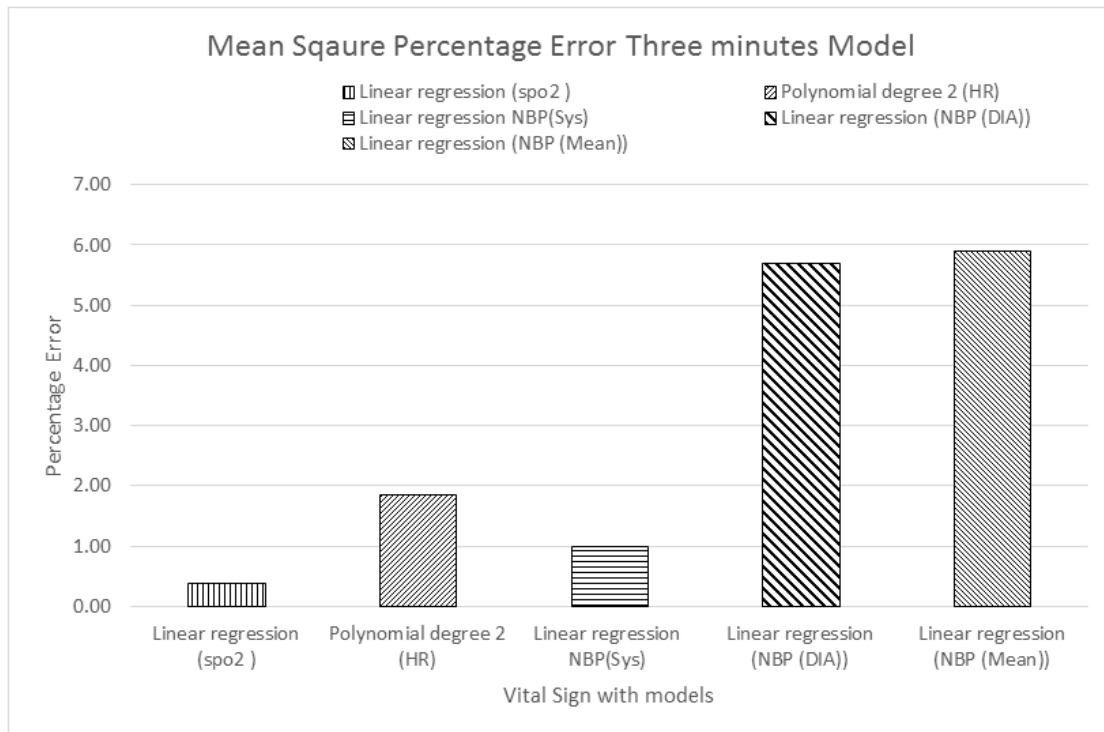
Figure 4.11: Best Model for 3 minutes forecasted value for CVD

Table 4.13: Classification of 3 minutes forecasted value by Best Model for CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.18	0.39	1	0.93	0.96
	HR	1.53	1.86			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	11.20	5.92			
Support Vector Machine	spo2	0.18	0.39	**0.99**	**0.99**	**0.99**
	HR	1.53	1.86			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	11.20	5.92			
Decision Tree	spo2	0.18	0.39	0.98	0.98	0.98
	HR	1.53	1.86			
	NBP(Sys)	1.009	0.99			
	NBP(Dia)	6.68	5.77			
	NBP(Mean)	11.20	5.92			
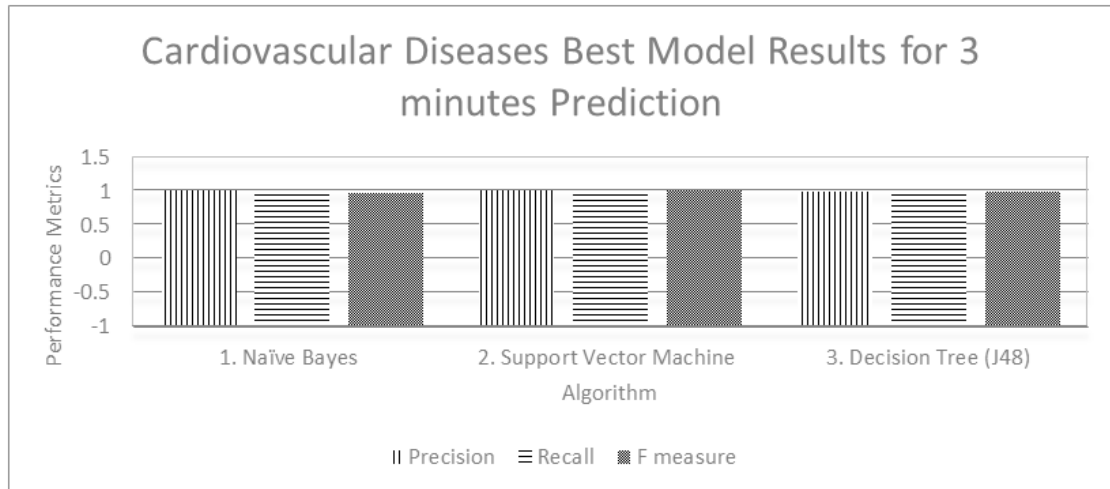
FIGURE 4.12: Classification of 3 minutes forecasted value by Best Model for CVD

Findings: *When top regression model was used, the Support Vector Machine obtained the best result by scoring F-measure of 0.99.*

4.2.3   Adaptation of Machine Error

Machine can be wrong or can have fault so we must consider this is our experiments. Error of machine can produce the wrong result and in medical field, error will result in the shape of patient death. Therefore, we must focus on this section and for this; we added new feature in the methodology by taking the mean of consecutive 100 samples of batches and create a new dataset. Dataset is than passed to the same process (presented in the previous section 4.2.1). We get upto 12 percent improved result(by using Linear Regression, Polynomial degree 2,Polynomial degree 3, and Polynomial degree 4) as compared to sixty seconds and three minutes forecasted values and f-measures result(by using SVM,Decision Tree and Naive Bayes) will also improved upto 10 percent. Following are the experiment and its result for mean of consecutive 100 samples of batches dataset.

1. **Classification using Linear Regression** The results of linear regression and classification techniques are shown in Table 4.14. Linear regression forecast the heart failure vital signs for mean of 100 consecutive sample of

batch. All the techniques(i.e. SVM,Decision Tree, Naive Bayes) produced same results with high accuracy upto 0.98 F-measure.

TABLE 4.14: Classification of forecasted value by Linear Regression of CVD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.99	2.35	0.97	0.98	0.98
	HR	34.86	6.34			
	NBP(Sys)	52.57	5.39			
	NBP(Dia)	19.80	5.58			
	NBP(Mean)	24.06	5.44			
Support Vector Machine	spo2	0.99	2.35	0.97	0.98	0.98
	HR	34.86	6.34			
	NBP(Sys)	52.57	5.39			
	NBP(Dia)	19.80	5.58			
	NBP(Mean)	24.06	5.44			
Decision Tree	spo2	0.99	2.35	**0.97**	**0.98**	**0.98**
	HR	34.86	6.34			
	NBP(Sys)	52.57	5.39			
	NBP(Dia)	19.80	5.58			
	NBP(Mean)	24.06	5.44			

Findings: *When using mean dataset Linear Regression model is used, all classifier (i.e., SVM, nave bayes and decision tree) predict the heart failure with high accuracy up to 0.98 F-measure. All classifier produced same result on for CVD disease*

1. **Classification using Polynomial Degree 2** The results of Polynomial Degree 2 and classification techniques are shown in Table 4.15. Polynomial Degree 2 forecast the heart failure vital signs for mean of 100 consecutive

sample of batch. SVM produced same results with high accuracy upto 0.96 F-measure.

TABLE 4.15: Classification of forecasted value by Polynomial Degree 2 of CVD

Classifier		Polynomial Degree 2		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	1.18	2.48	0.98	0.56	0.58
	HR	122.32	11.35			
	NBP(Sys)	1100.26	23.02			
	NBP(Dia)	470.95	26.50			
	NBP(Mean)	581.38	49.18			
Support Vector Machine	spo2	1.18	2.48	0.97	0.95	0.96
	HR	122.32	11.35			
	NBP(Sys)	1100.26	23.02			
	NBP(Dia)	470.95	26.50			
	NBP(Mean)	581.38	49.18			
Decision Tree	spo2	1.18	2.48	0.96	0.84	0.89
	HR	122.32	11.35			
	NBP(Sys)	1100.26	23.02			
	NBP(Dia)	470.95	26.50			
	NBP(Mean)	581.38	49.18			

Findings: *When using mean dataset Polynomial Degree 2 model is used, SVM classifier predict the heart failure with high accuracy up to 0.96 F-measure.*

1. **Classification using Polynomial Degree 3** The results of Polynomial Degree 3 and classification techniques are shown in Table 4.16. Polynomial Degree 3 forecast the heart failure vital signs for mean of 100 consecutive

sample of batch.All the techniques(i.e. SVM,Decision Tree, Naive Bayes) produced same results with high accuracy upto 0.98 F-measure.

TABLE 4.16: Classification of forecasted value by Polynomial Degree 3 of CVD

Classifier		Polynomial Degree 3		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.73	1.93	0.97	0.98	0.98
	HR	28.16	5.62			
	NBP(Sys)	87.66	4.89			
	NBP(Dia)	60.89	7.29			
	NBP(Mean)	66.39	6.47			
Support Vector Machine	spo2	0.73	1.93	0.97	0.98	0.98
	HR	28.16	5.62			
	NBP(Sys)	87.66	4.89			
	NBP(Dia)	60.89	7.29			
	NBP(Mean)	66.39	6.47			
Decision Tree	spo2	0.73	1.93	0.97	0.98	0.98
	HR	28.16	5.62			
	NBP(Sys)	87.66	4.89			
	NBP(Dia)	60.89	7.29			
	NBP(Mean)	66.39	6.47			

Findings:*When using mean dataset Polynomial Degree 3 model is used, all classifier (i.e., SVM, nave bayes and decision tree) predict the heart failure with high accuracy up to 0.98 F-measure. All classifier produced same result on for CVD disease*

1. **Classification using Polynomial Degree 4** The results of Polynomial Degree 4 and classification techniques are shown in Table 4.17. Polynomial Degree 4 forecast the heart failure vital signs for mean of 100 consecutive

sample of batch. SVM produced same results with high accuracy upto 0.97 F-measure..

TABLE 4.17: Classification of forecasted value by Polynomial Degree 4 of CVD

Classifier		Polynomial Degree 4		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	3.22	3.92	0.98	0.53	0.53
	HR	3.76	1.95			
	NBP(Sys)	6720.22	51.36			
	NBP(Dia)	2029.54	49.70			
	NBP(Mean)	3286.19	53.56			
Support Vector Machine	spo2	3.22	3.92	0.97	0.97	0.97
	HR	3.76	1.95			
	NBP(Sys)	6720.22	51.36			
	NBP(Dia)	2029.54	49.70			
	NBP(Mean)	3286.19	53.56			
Decision Tree	spo2	3.22	3.92	0.97	0.92	0.94
	HR	3.76	1.95			
	NBP(Sys)	6720.22	51.36			
	NBP(Dia)	2029.54	49.70			
	NBP(Mean)	3286.19	53.56			

Findings:*When using mean dataset Polynomial Degree 4 model is used, SVM classifier predict the heart failure with high accuracy up to 0.97 F-measure.*

4.3 Chronic Respiratory Disease Experimentation

In chronic respiratory disease vital sign, we used UQ dataset. The dataset is treated with standard preprocessing and normalization procedures (already discussed in section 3.5). For the prediction of heart failure disease, we extracted most relevant vital signs (recommended by medical experts) from the UQ dataset. There are several vital signs available in the dataset for chronic respiratory disease (discussed in section 3.6). However, we are using Table 4.18 vital signs for predicting hypoxemic and hypercapnic disease.

TABLE 4.18: Input vital sign chronic respiratory disease

Feature No	Vital sign
1	spo2
2	etCO2
3	imCO2
4	etO2
5	inO2

4.3.1 Sixty seconds forecasted value

In this section, regression models (i.e., linear regression[6], polynomial regression (degree 2, 3 and 4) [7]) are trained on the 50,000 samples with 10 milliseconds interval for each vital sign(i.e., HR or NBP(Sys) etc.) of hypoxemic and hypercapnic disease. Trained models (i.e., linear regression[6], polynomial regression (degree 2, 3 and 4) [7]) of each vital sign are then used to forecast the next 60 seconds of same vital sign. These forecasted values of vital signs are used as input for classification techniques (like SVM[8], Nave Bayes[9] and Decision Tree[10]) with output class(already explained in section 3.6). As it is supervised learning method, we evaluated data mining techniques(i.e., regression techniques and classification techniques used in this research) performance from already known

results (i.e., best regression model have minimum MSE and MAPE, classification model have precision, recall and f-measure score equal to 1 means that prediction accuracy is high etc.). Same process is repeated for 100,000 samples with 10 milliseconds interval to forecast the next 3 minutes. Results of performance metric of data mining techniques on different range of dataset (i.e., 50,000 samples to forecast the 60 seconds and 100,000 samples to forecast the 3 minutes) are discussed in below sections with tables and charts. Hybrid approach (combining two or more regression model) results are also discussed in detail in upcoming sections.

1. **Classification using Linear Regression** The results of linear regression and classification techniques are shown in Table 4.19 and Figure 4.13. Linear regression forecast the hypoxemic and hypercapnic disease vital signs for next 60 minutes. Linear regression forecast the SpO2 with minimum MAPE score is up to 0.75. Moreover, imCO2 achieved up to 4.81 MAPE score, etO2 and inO2 have minimum MAPE score up to 2.5, and etCO2 have MAPE score up to 45.64. Nave Bayes[9], SVM[8] and decision tree achieved hypoxemic and hypercapnic disease prediction accuracy score up to 0.001 F-measure. Prediction accuracy of hypoxemic and hypercapnic disease does not have satisfactory result by using linear regression forecasted values.

TABLE 4.19: Classification of 60 Seconds forecasted value by Linear Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.611	0.75	0.0001	0.0001	0.0001
	HR	359	45.64			
	NBP(Sys)	0.0034	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.59	2.5			
Support Vector Machine	spo2	0.611	0.75	0.0001	0.0001	0.0001
	HR	359	45.64			
	NBP(Sys)	0.0034	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.59	2.5			
Decision Tree	spo2	0.611	0.75	0.0001	0.0001	0.0001
	HR	359	45.64			
	NBP(Sys)	0.0034	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.59	2.5			

Findings: *When linear regression model is used, the all three technique obtained the same result by scoring F-measure of 0.0001, which is not satisfactory for hypoxemic and hypercapnic disease prediction.*

2. **Classification using Polynomial Degree Two** The result of linear regression is not quite satisfactory, so we implemented polynomial regression model to evaluate its performance. After implementing the polynomial degree two regression model the results of all classification techniques (i.e., SVM, nave bayes and decision tree) are shown in Table 4.20 and Figure 4.1. If we observe the MAPE closely, spo2 MAPE is increased from 0.75(achieved
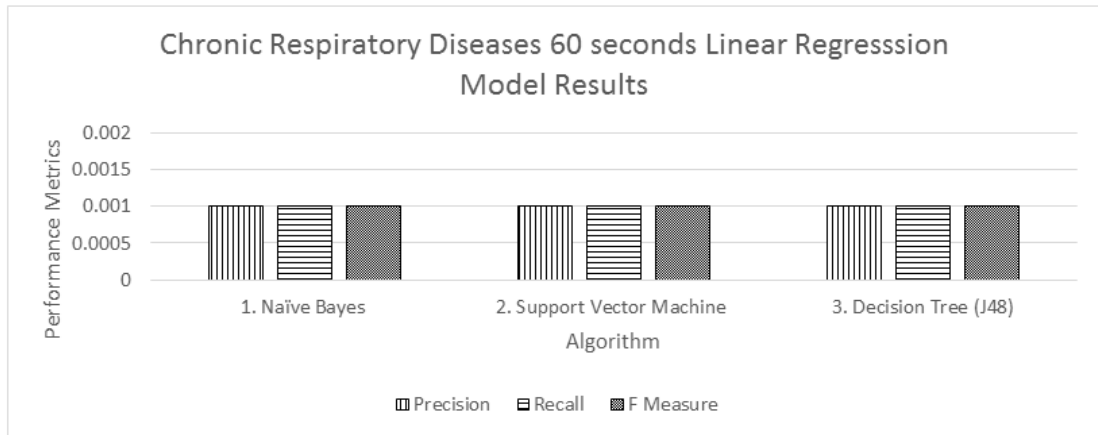
FIGURE 4.13: Classification of 60 Seconds forecasted value by Linear Regression of CRD

by linear regression) to 1.49(achieved by polynomial degree 2), which is not good. Other MAPE score of vital signs are also increased. The prediction accuracy of classification techniques is improved. Support vector machine classifier achieved f-measure score of up to 0.98. Similarly, following forms were able to achieve fair accuracy with the F-measure of 0.51 when using the Nave Bayes[9]. It is also revealed that using Support vector machine has the highest prediction accuracy of up to 0.98.

TABLE 4.20: Classification of 60 Seconds forecasted value by Polynomial Degree
2 Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	**MSE**	**MAPE**	**Precision**	**Recall**	**F measure**
Nave Bayes	spo2	2.73	1.49	1	0.34	0.51
	HR	233.53	31.04			
	NBP(Sys)	1.77	123.52			
	NBP(Dia)	105	11.5			
	NBP(Mean)	49	7.6			
Support Vector Machine	spo2	2.73	1.49	**0.98**	**0.98**	**0.98**
	HR	233.53	31.04			
	NBP(Sys)	1.77	123.52			
	NBP(Dia)	105	11.5			
	NBP(Mean)	49	7.6			
Decision Tree	spo2	2.73	1.49	0.0001	0.0001	0.0001
	HR	233.53	31.04			
	NBP(Sys)	1.77	123.52			
	NBP(Dia)	105	11.5			
	NBP(Mean)	49	7.6			

Findings: *When Polynomial regression degree 2 is used, the Support Vector Machine obtained the best result by scoring F-measure of 0.98 and the vital sign SPO2 obtain the mean square error of 1.49. Comparing with Linear Regression the results of F-measure improved from 0.0001 to 0.98 by using support vector machine.*

3. **Classification using Polynomial Degree Three** Polynomial degree two produces good performance metric to 0.98 but we need to test the higher degree polynomial for better results. We applied polynomial degree three to evaluate our model further. The results of polynomial degree three and
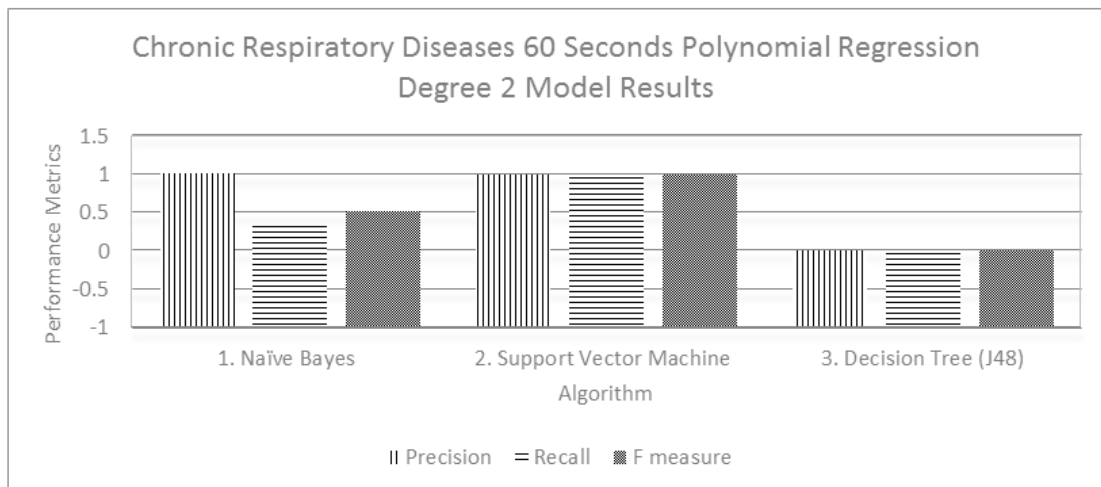
FIGURE 4.14: Classification of 60 Seconds forecasted value by Polynomial Degree 2 Regression of CRD

classification techniques (i.e., SVM, nave bayes and decision tree) are shown in Table 4.21 and Figure 4.15. Overall results are produces by polynomial degree three are not good as compare to polynomial degree two. But Nave Bayes[9] produces better in polynomial degree three.

Table 4.21: Classification of 60 Seconds forecasted value by PND-3 Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	5.87	2.01	**1**	**0.10**	**0.18**
	HR	49.1	14.4			
	NBP(Sys)	20.16	376			
	NBP(Dia)	1120	35			
	NBP(Mean)	15.7	3.6			
Support Vector Machine	spo2	5.87	2.01	0.001	0.001	0.001
	HR	49.1	14.4			
	NBP(Sys)	20.16	376			
	NBP(Dia)	1120	35			
	NBP(Mean)	15.7	3.6			
Decision Tree	spo2	5.87	2.01	0.001	0.001	0.001
	HR	49.1	14.4			
	NBP(Sys)	20.16	376			
	NBP(Dia)	1120	35			
	NBP(Mean)	15.7	3.6			

Findings: *When Polynomial regression degree 3 is used, the result are worst because polynomial degree 2 procedures better prediction results by scoring F-measure of 0.98. However, polynomial degree three obtained the prediction result by scoring F-measure of 0.18. therefore, comparing with polynomial degree 2 the results of F-measures drops from 0.98 to 0.18 but polynomial degree 3 have better prediction result from linear regression which is 0.001.*

4. **Classification using Polynomial Degree Four** In order to evaluate further and to find the best combination of forecasting we also evaluated polynomial regression with degree four. However, the results are worst as compare
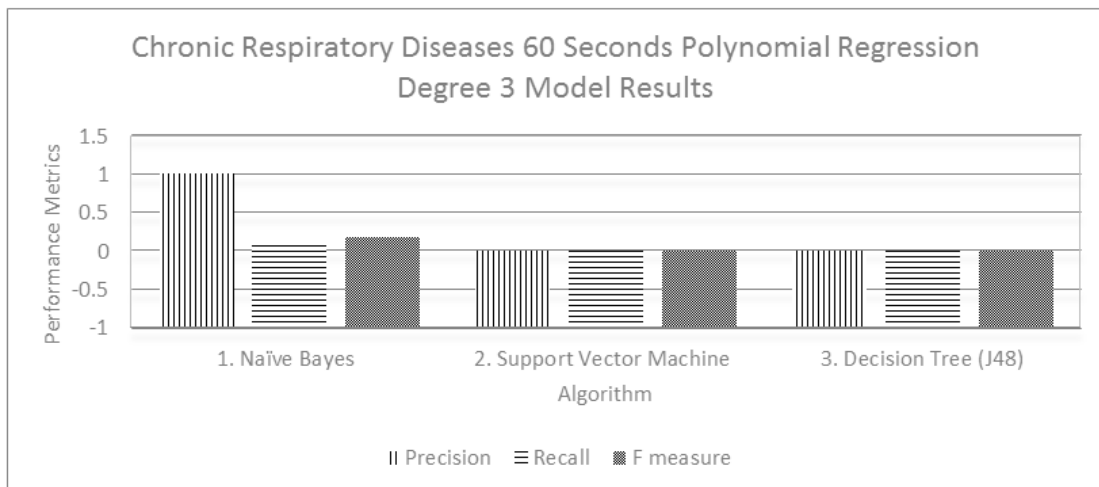
FIGURE 4.15: Classification of 60 Seconds forecasted value by PND-3 Regression of CRD

to polynomial degree two and degree three (shown in Table 4.22 and Figure 4.16). Polynomial degree four result drops score of F-Measure to 0.06 which means this technique doesnt produce the better result which we are expecting(shown in Table 4.22).

TABLE 4.22: Classification of 60 Seconds forecasted value by PND-4 Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	1.47	1.04	0.0001	0.0001	0.0001
	HR	507	42			
	NBP(Sys)	28	401			
	NBP(Dia)	488	19			
	NBP(Mean)	2175	43.1			
Support Vector Machine	spo2	1.47	1.04	0.0001	0.0001	0.0001
	HR	507	42			
	NBP(Sys)	28	401			
	NBP(Dia)	488	19			
	NBP(Mean)	2175	43.1			
Decision Tree	spo2	1.47	1.04	**1**	**0.03**	**0.06**
	HR	507	42			
	NBP(Sys)	28	401			
	NBP(Dia)	488	19			
	NBP(Mean)	2175	43.1			

Findings: *When Polynomial regression degree 4 is used, F-measure of all classifier drops which means that we cannot consider the polynomial degree 4 for predicting the hypoxemic and hypercapnic disease.*

5. **Classification using Hybrid Regression Model** In order to find the best combination, we choose the lowest error regression model from linear regression, polynomial degree 2,polynomial degree 3, and polynomial degree 4 as seen in Table 4.23 and Figure 4.17. Table 4.23 shows that linear regression and polynomial degree 3 performed well and have minimum MAPE score out of four techniques (i.e., linear regression, polynomial degree 2,
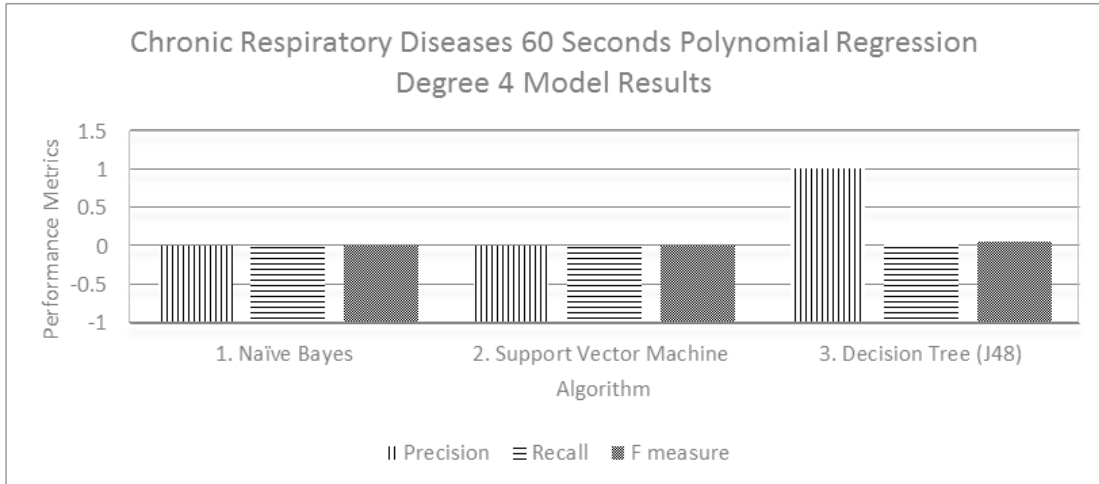
FIGURE 4.16: Classification of 60 Seconds forecasted value by PND-4 Regression of CRD

polynomial degree 3, and polynomial degree 4). Linear regression forecast with minimum error rate on four vital signs(i.e., SpO2,imCO2,etO2,inO2) and polynomial degree 3 forecast with minimum error rate on only one vital sign(i.e., etCO2) and remaining polynomial degrees are not forecasting well. We use these top rank models in classifier (i.e., SVM, nave bayes and decision tree) to get better prediction results.

TABLE 4.23: CRD Optimal regression Model for 60 seconds forecasted value

Best Model	Vital Sign	MAPE
Linear Regression	spo2	0.75
Polynomial degree 3	etCO2	14.4
Linear Regression	imCO2	4.81
Linear Regression	etO2	2.75
Linear Regression	inO2	2.5

Implementation of hybrid Regression Model The results of hybrid regression model and all classification techniques are shown in Table 4.24 and Figure 4.18. Firstly we applied Nave Bayes[9] classifier on the 60 seconds forecasted data of hybrid approach. Nave Bayes does perform well with the lowest MAPE score and has the highest prediction accuracy score with F-measure of up to 0.99. Moreover, when the forecasted vital signs (of hybrid regression model) are provided as input to support vector machine. SVM performed worst and having the prediction accuracy score of F-measure is
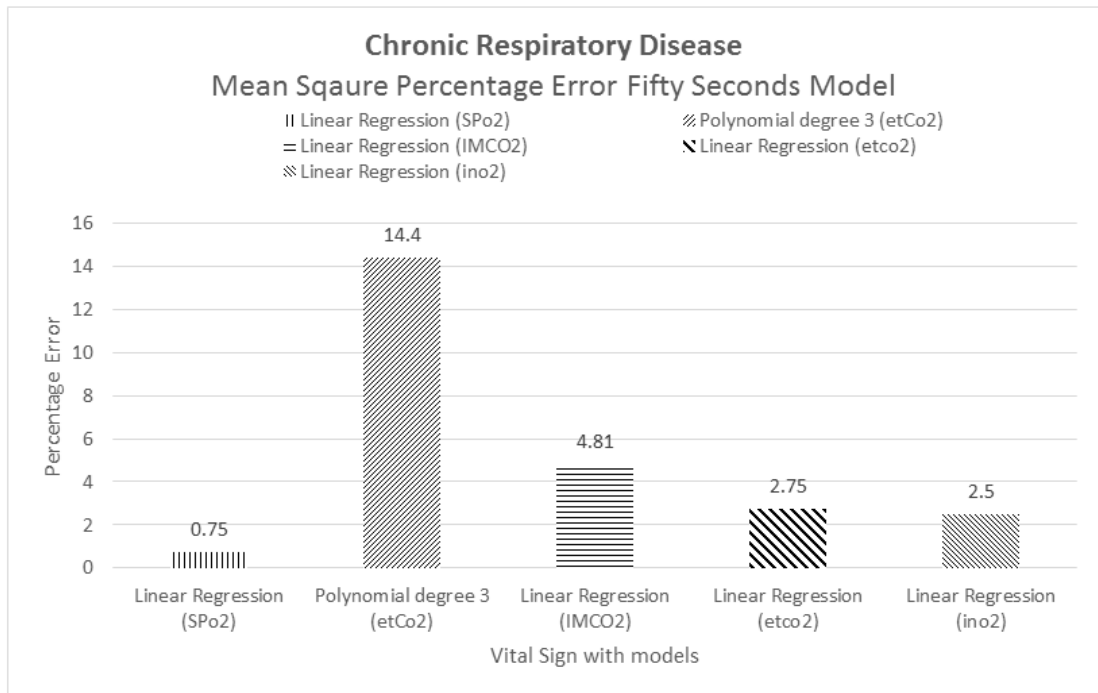
FIGURE 4.17: CRD Optimal Model for 60 seconds forecasted value

up to0.001. Similarly, when hybrid regression model is applied to decision tree, decision tree performed fair and has the prediction accuracy score with F-measure of up to 0.14.

TABLE 4.24: Classification of 60 Seconds forecasted value by Best Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.61	0.75	**0.99**	**0.99**	**0.99**
	HR	49.1	14.4			
	NBP(Sys)	0.003	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.5	2.5			
Support Vector Machine	spo2	0.61	0.75	0.001	0.001	0.001
	HR	49.1	14.4			
	NBP(Sys)	0.003	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.5	2.5			
Decision Tree	spo2	0.61	0.75	0.18	0.12	0.14
	HR	49.1	14.4			
	NBP(Sys)	0.003	4.81			
	NBP(Dia)	5.02	2.75			
	NBP(Mean)	5.5	2.5			

Findings: *When top regression model was used, the Nave Bayes obtained the best result by scoring F-measure of 0.99 which is higher in all experiments.*

4.3.2 Three minutes forecasted value

Same process (presented in the previous section 4.3.1) is also performed with 3 minutes forecasted value.
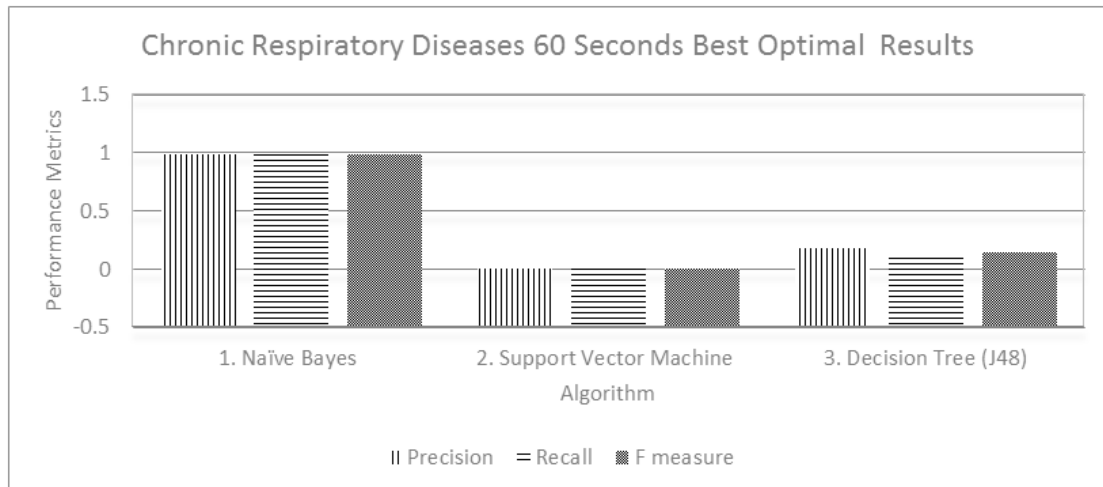
FIGURE 4.18: CRD Best Model Results for 60 Seconds Prediction

1. **Classification using Linear Regression** Same process is also done with 3 minutes forecasted value. The results of linear regression and classification techniques are shown in Table 4.25 and Figure 4.19. Linear regression forecast the SpO2 with minimum MAPE score up to 0.13 but not on etCO2 with minimum MAPE score up to 20. Other vital signs forecasted values are average and acceptable as shown in Table 35 . Nave Bayes[9] and Decision Tree[10] achieve prediction accuracy score up to 0.99 F-measure. SVM[8] achieved 0.0001 prediction accuracy scores in F-measure.

TABLE 4.25: Classification of 3 minutes forecasted value by Linear Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.04	0.13	**0.99**	**0.99**	**0.99**
	HR	88.75	20.04			
	NBP(Sys)	0.009	9.7			
	NBP(Dia)	16.7	5.27			
	NBP(Mean)	19.4	5.19			
Support Vector Machine	spo2	0.04	0.13	0.0001	0.0001	0.0001
	HR	88.75	20.04			
	NBP(Sys)	0.009	9.7			
	NBP(Dia)	16.7	5.27			
	NBP(Mean)	19.4	5.19			
Decision Tree	spo2	0.04	0.13	**0.99**	**0.99**	**0.99**
	HR	88.75	20.04			
	NBP(Sys)	0.009	9.7			
	NBP(Dia)	16.7	5.27			
	NBP(Mean)	19.4	5.19			

Findings: *When Linear Regression model is used with three minutes data, Nave Bayes and Decision tree classifier achieved high F-measure up to 0.99.*

2. **Classification using Polynomial Degree Two** The results of polynomial degree 2 and classification techniques are shown in Table 4.26 and Figure 4.20. Polynomial degree 2 produced minimum MAPE score for all the vital signs of hypoxemic and hypercapnic disease as shown in Table 4.26 . Support Vector Machine[8] achieves highest prediction accuracy f-measure score up to 0.74. Nave Bayes[9] and Decision Tree[10] achieve 0.22 and 0.11 F-measure respectively.
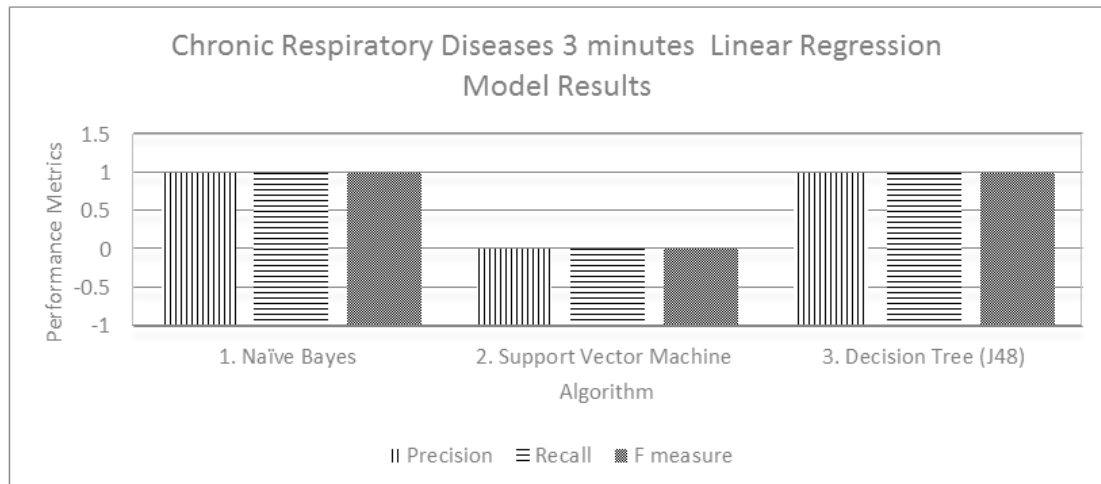
FIGURE 4.19: Classification of 3 minutes forecasted value by Linear Regression of CRD

TABLE 4.26: Classification of 3 minutes forecasted value by PND-2 of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.86	0.89	1	0.12	0.22
	HR	59	15.5			
	NBP(Sys)	0.014	11.4			
	NBP(Dia)	7.27	3.2			
	NBP(Mean)	23	5.4			
Support Vector Machine	spo2	0.86	0.89	**1**	**0.59**	**0.74**
	HR	59	15.5			
	NBP(Sys)	0.014	11.4			
	NBP(Dia)	7.27	3.2			
	NBP(Mean)	23	5.4			
Decision Tree	spo2	0.86	0.89	1	0.06	0.11
	HR	59	15.5			
	NBP(Sys)	0.014	11.4			
	NBP(Dia)	7.27	3.2			
	NBP(Mean)	23	5.4			
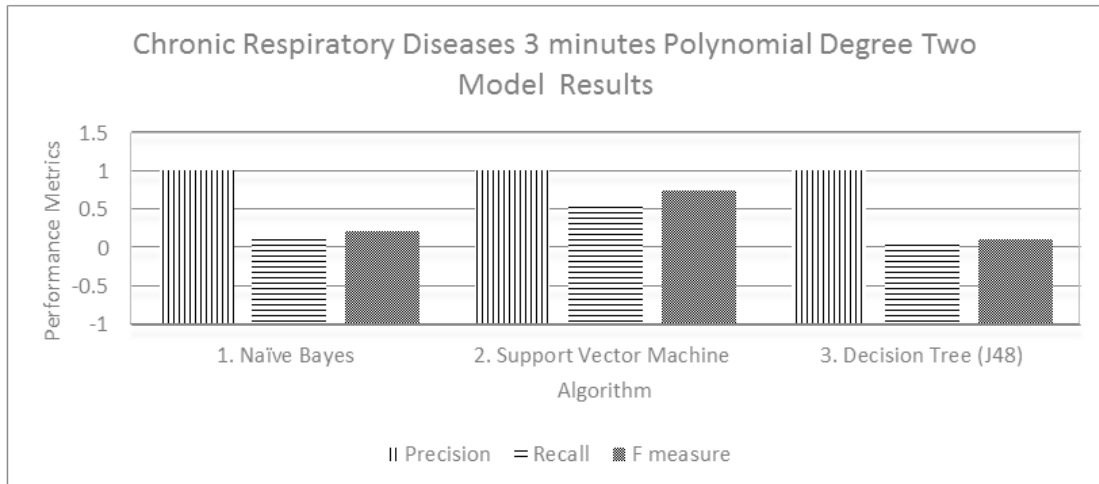
FIGURE 4.20: Classification of 3 minutes forecasted value by PND-2 of CRD

Findings: *When Polynomial regression with degree two is used with three minutes data, support vector machine achieved the F-measure of 0.74(shown in Table 4.26). But linear regression perform well as compare to polynomial degree two, because polynomial degree two drops the accuracy value from 0.99 to 0.74.*

3. **Classification using Polynomial Degree Three** The results of polynomial degree 3 and classification techniques are shown in Table 4.27 and Figure 4.21. Polynomial degree three forecast the hypoxemic and hypercapnic disease vital signs with minimum error rate as shown in Table 4.27.Nave Bayes[9] and Decision Tree[10] achieve highest prediction accuracy F-measure score up to 0.99. SVM[8] achieve 0.001 F-measure prediction accuracy respectively. Comparing with Linear Regression model and polynomial degree 2, polynomial degree 3 having same high F-measure accuracy as linear regression which is 0.99.

TABLE 4.27: Classification of 3 minutes forecasted value by PND-3 of CRD

| Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	1.79	1.19	**0.99**	**0.99**	**0.99**
	HR	182	26			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			
Support Vector Machine	spo2	1.79	1.19	0.001	0.001	0.001
	HR	182	26			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			
Decision Tree	spo2	1.79	1.19	**0.99**	**0.99**	**0.99**
	HR	182	26			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			
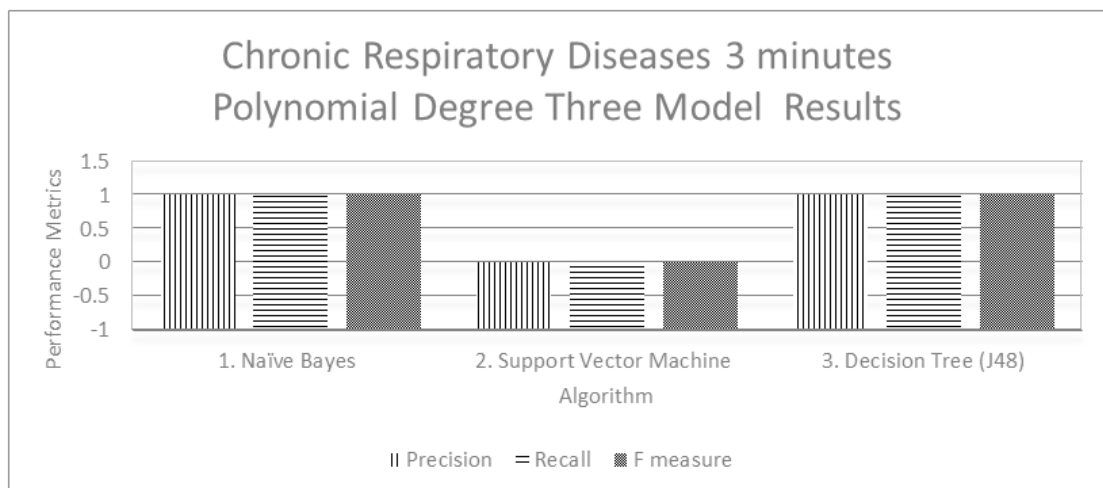
FIGURE 4.21: Classification of 3 minutes forecasted value by PND-3 of CRD

Findings: *When Polynomial degree three is used with three minutes data, Nave Bayes and Decision Tree have achieved high prediction accuracy F-measure score up to 0.99. Linear regression and polynomial degree 3 predict the same accuracy F-measure score up to 0.99.*

4. **Classification using Polynomial Degree Four** In order to evaluate further and to find the best combination of forecasting we also evaluated polynomial regression with degree four. The results of polynomial degree 4 regression and all the three classification techniques are shown in Table 4.28 and Figure 4.22. However, the result of polynomial degree 4 is worst as compare to linear regression, polynomial degree two, and degree three. Decision Tree[10] achieves highest F measure prediction accuracy score up to 0.27 with low MAPE is up to 0.087. Comparing with linear regression, polynomial degree 2 and 3, polynomial degree 4 drops all performance metric values but with low MAPE.

TABLE 4.28: Classification of 3 minutes forecasted value by PND-4 of CRD

| Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.02	0.087	0.0001	0.0001	0.0001
	HR	1846	79			
	NBP(Sys)	0.32	49			
	NBP(Dia)	78	9.96			
	NBP(Mean)	11.25	3.49			
Support Vector Machine	spo2	0.02	0.087	0.0001	0.0001	0.0001
	HR	1846	79			
	NBP(Sys)	0.32	49			
	NBP(Dia)	78	9.96			
	NBP(Mean)	11.25	3.49			
Decision Tree	spo2	0.02	0.087	**1**	**0.16**	**0.27**
	HR	1846	79			
	NBP(Sys)	0.32	49			
	NBP(Dia)	78	9.96			
	NBP(Mean)	11.25	3.49			
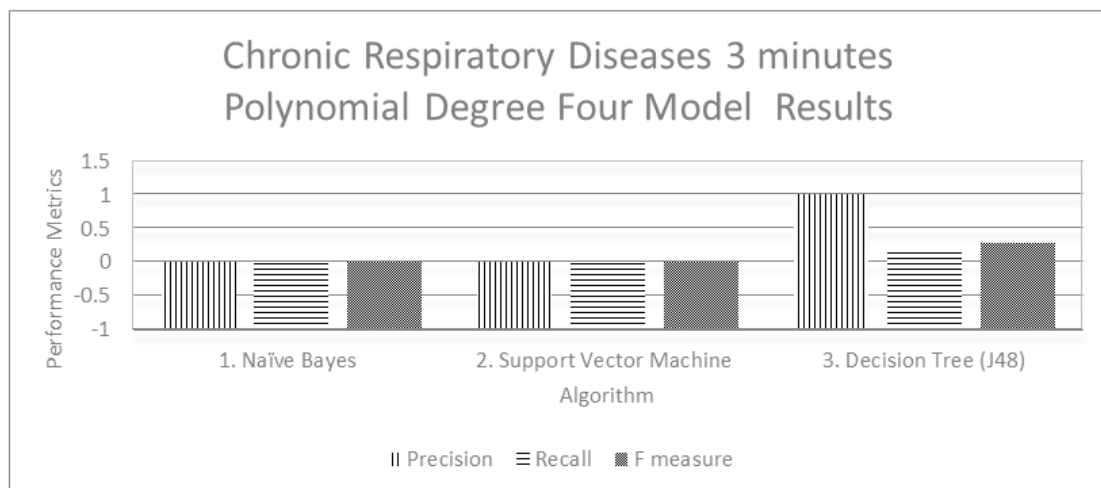
FIGURE 4.22: Classification of 3 minutes forecasted value by PND-4 of CRD

Findings: *When polynomial degree four models are used with three minutes data, all classifier achieved low prediction accuracy F-measure (shown in Table 4.28). Overall this technique drops all the performance matrix values and with low error (shown in Table 4.28) as compared to other 3 techniques (i.e., linear regression, polynomial degree 2, and 3).*

5. **Classification using Hybrid Regression Model** In order to find the best combination, we choose the lowest error regression model from linear regression, polynomial degree 2, polynomial degree 3 and polynomial degree 4 as seen in Table 4.29 and Figure 4.23. Table 4.29 shows that polynomial degree 3 and polynomial degree 4 forecast the hypoxemic and hypercapnic disease and have a minimum mean absolute percentage error out of four techniques (i.e., linear regression, polynomial degree 2, polynomial degree 3 and polynomial degree 4). Polynomial degree 3 forecasts well on four vital signs (i.e., etCO2, imCO2, etO2 and inO2) and polynomial degree 4 forecast well on the only SPO2 and remaining linear and polynomial degree 2 is not forecasting well. We use these top rank models in classifier to get better classification results

TABLE 4.29: CRD Best Optimal regression Model for 3 minutes forecasted value

Best Model	Vital Sign	MAPE
Polynomial degree 4	spo2	0.087
Polynomial degree 3	etCO2	15.5
Polynomial degree 3	imCO2	0.75
Polynomial degree 3	etO2	0.84
Polynomial degree 3	inO2	2.81

Implementation of Hybrid Regression Model The results of all classes are shown in Table 4.30 and Figure 4.24. Firstly we applied Nave Bayes[9] classifier on the three minutes data. Nave Bayes performs well by achieving prediction accuracy up to 0.99. Moreover, when the forecasted vital signs are provided as input to support vector machine, SVM scored lowest prediction accuracy up to 0.001. Similarly, when hybrid regression model is applied,
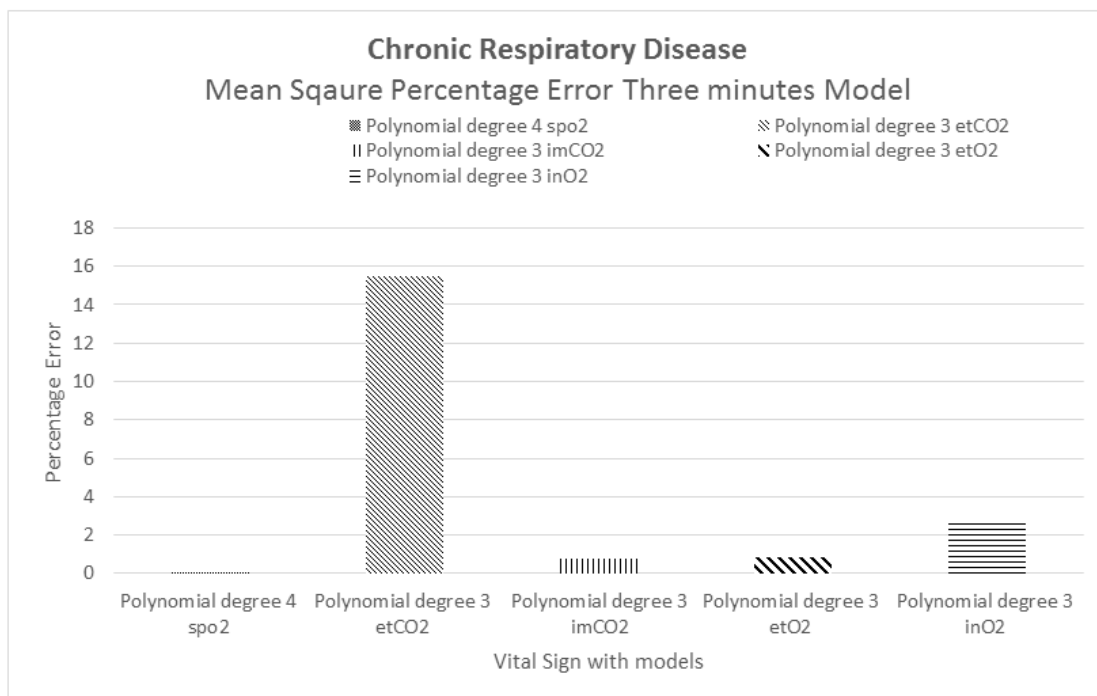
FIGURE 4.23: CRD Best Optimal regression Model for 3 minutes forecasted value

decision tree and nave bayes scored the highest prediction accuracy with the F- a measure of 0.99.

TABLE 4.30: Classification of 3 minutes forecasted value by Best Regression of CRD

Classifier		Linear Regression		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	0.02	0.087	**0.99**	**0.99**	**0.99**
	HR	59	15.5			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			
Support Vector Machine	spo2	0.02	0.087	0.0001	0.0001	0.0001
	HR	59	15.5			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			
Decision Tree	spo2	0.02	0.087	**0.99**	**0.99**	**0.99**
	HR	59	15.5			
	NBP(Sys)	7.6	0.75			
	NBP(Dia)	0.51	0.84			
	NBP(Mean)	7.26	2.81			

Findings: *When top regression model was used, the nave Bayes and decision tree obtained the best result by scoring F-measure of 0.99.*

4.3.3 Adaptation of Machine Error

Machine can be wrong or can have fault so we must consider this is our experiments. Error of machine can produce the wrong result and in medical field, error will be in the shape of patient death. Therefore, we must focus on this section and
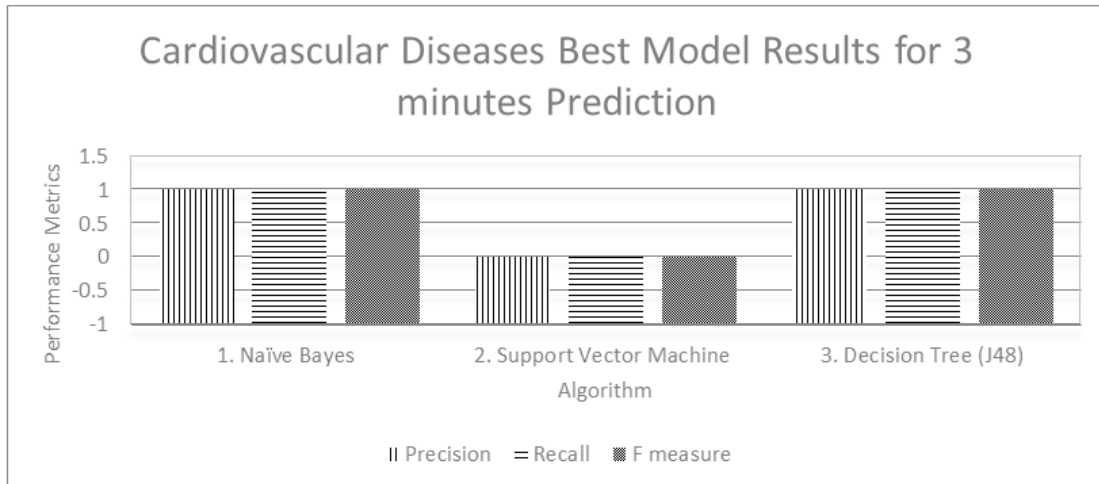
FIGURE 4.24: Classification of 3 minutes forecasted value by Best Regression of CRD

for this; we added new feature in the methodology by taking the mean of consecutive 100 samples of batches and create a new dataset. Dataset is than passed to the same process (presented in the previous section 4.2.1) We get upto 15 percent improved result(by using Linear Regression, Polynomial degree 2,Polynomial degree 3, and Polynomial degree 4) as compared to sixty seconds and three minutes forecasted values and f-measures result(by using SVM,Decision Tree and Naive Bayes) will also improved upto 10 percent. Following are the experiment and its result for mean of consecutive 100 samples of batches dataset.

1. **Classification using Linear Regression** The results of linear regression and classification techniques are shown in Table 4.31. Linear regression forecast the CRD vital signs for mean of 100 consecutive sample of batch. Error rate of linear regression are 15 percent less as compared to previous techniques. All the techniques(i.e. SVM,Decision Tree, Naive Bayes) produced same results with high accuracy upto 1 F-measure.

TABLE 4.31: Classification of forecasted value by Linear Regression of CRD

Classifier	Linear Regression			Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	6.41	3.71	1	1	1
	HR	4.36	4.10			
	NBP(Sys)	1.55	13.64			
	NBP(Dia)	36.40	11.88			
	NBP(Mean)	73.42	11.38			
Support Vector Machine	spo2	6.41	3.71	1	1	1
	HR	4.36	4.10			
	NBP(Sys)	1.55	13.64			
	NBP(Dia)	36.40	11.88			
	NBP(Mean)	73.42	11.38			
Decision Tree	spo2	6.41	3.71	1	1	1
	HR	4.36	4.10			
	NBP(Sys)	1.55	13.64			
	NBP(Dia)	36.40	11.88			
	NBP(Mean)	73.42	11.38			

Findings: *When using mean dataset Linear Regression model is used, all classifier (i.e., SVM, nave bayes and decision tree) predict the CRD with high accuracy up to 1 F-measure. All classifier produced same result for CRD disease*

1. **Classification using Polynomial Degree 2** The results of Polynomial Degree 2 and classification techniques are shown in Table 4.32. Polynomial Degree 2 forecast the CRD vital signs for mean of 100 consecutive sample of batch. All the techniques(i.e. SVM,Decision Tree, Naive Bayes) produced same results with high accuracy upto 1 F-measure.

TABLE 4.32: Classification of forecasted value by Polynomial Degree 2 of CRD

Classifier		Polynomial Degree 2		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	1.73	2.85	1	1	1
	HR	30.97	12.45			
	NBP(Sys)	3.22	30.18			
	NBP(Dia)	590.97	54.42			
	NBP(Mean)	266.37	58.05			
Support Vector Machine	spo2	1.73	2.85	1	1	1
	HR	30.97	12.45			
	NBP(Sys)	3.22	30.18			
	NBP(Dia)	590.97	54.42			
	NBP(Mean)	266.37	58.05			
Decision Tree	spo2	1.73	2.85	1	1	1
	HR	30.97	12.45			
	NBP(Sys)	3.22	30.18			
	NBP(Dia)	590.97	54.42			
	NBP(Mean)	266.37	58.05			

Findings: *When using mean dataset Polynomial Degree 2 model is used, all classifier (i.e., SVM, nave bayes and decision tree) predict the CRD with high accuracy up to 1 F-measure. All classifier produced same result for CRD disease*

1. **Classification using Polynomial Degree 3** The results of Polynomial Degree 3 and classification techniques are shown in Table 4.33. Polynomial Degree 3 forecast the CRD vital signs for mean of 100 consecutive sample of batch. All the techniques(i.e. SVM,Decision Tree, Naive Bayes) produced same results with high accuracy upto 1 F-measure.

TABLE 4.33: Classification of forecasted value by Polynomial Degree 3 of CRD

Classifier		Polynomial Degree 3		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	62.54	6.84	1	1	1
	HR	73.57	18.39			
	NBP(Sys)	1.16	13.45			
	NBP(Dia)	361.38	37.69			
	NBP(Mean)	269.18	35.00			
Support Vector Machine	spo2	62.54	6.84	1	1	1
	HR	73.57	18.39			
	NBP(Sys)	1.16	13.45			
	NBP(Dia)	361.38	37.69			
	NBP(Mean)	269.18	35.00			
Decision Tree	spo2	62.54	6.84	1	1	1
	HR	73.57	18.39			
	NBP(Sys)	1.16	13.45			
	NBP(Dia)	361.38	37.69			
	NBP(Mean)	269.18	35.00			

Findings: *When using mean dataset Polynomial Degree 2 model is used, all classifier (i.e., SVM, nave bayes and decision tree) predict the CRD with high accuracy up to 1 F-measure. All classifier produced same result for CRD disease*

1. **Classification using Polynomial Degree 4** The results of Polynomial Degree 4 and classification techniques are shown in Table 4.34. Polynomial Degree 4 forecast the CRD vital signs for mean of 100 consecutive sample of batch.Naive Bayes produced results with high accuracy upto 0.98 F-measure..

TABLE 4.34: Classification of forecasted value by Polynomial Degree 4 of CRD

Classifier		Polynomial Degree 4		Performance Metrics		
	Vital sign	MSE	MAPE	Precision	Recall	F measure
Nave Bayes	spo2	84.42	5.62	0.97	0.99	0.98
	HR	174.87	31.63			
	NBP(Sys)	7.41	45.44			
	NBP(Dia)	25.51	6.9			
	NBP(Mean)	3008.21	104.98			
Support Vector Machine	spo2	84.42	5.62	0.5	0.51	0.5
	HR	174.87	31.63			
	NBP(Sys)	7.41	45.44			
	NBP(Dia)	25.51	6.9			
	NBP(Mean)	3008.21	104.98			
Decision Tree	spo2	84.42	5.62	0.99	0.78	0.85
	HR	174.87	31.63			
	NBP(Sys)	7.41	45.44			
	NBP(Dia)	25.51	6.9			
	NBP(Mean)	3008.21	104.98			

Findings: *When using mean dataset Polynomial Degree 4 model is used, Naive Bayes classifier predict the CRD with high accuracy up to 0.98 F-measure.*

4.4 Results and discussion of Regression techniques

In order to find the best combination for both the diseases, we choose the best mean absolute percentage error from regression techniques i.e. Linear Regression,

polynomial degree 2, 3, and 4 as shown in Table 4.35,4.36. The below table contains all mean absolute percentage error which tells us, which regression model perform well and have the minimum error rate.

TABLE 4.35: Combine results of 4 regression techniques with heart failure and CRD

Vital Sign	Performance Metrics Heart Failure in 60 Seconds		Performance Metrics Heart Failure in 3 minutes	
	Reg. Model	*MAPE*	*Reg. Model*	*MAPE*
Heart Rate	Poly.deg.2	3.91	Poly.Deg.2	1.86
NBP(Sys)	Linear Reg.	14.57	Poly.Deg.4	0.99
NBP(Dia)	Linear Reg.	14.55	Linear Reg.	5.77
NBP (Mean)	Linear Reg.	13.76	Poly.Deg.4	5.92
SPO2	Linear Reg.	0.01	Linear Reg.	0.39

TABLE 4.36: Combine results of 4 regression techniques with heart failure and CRD

Vital Sign	Performance Metrics Chronic Respiratory in 60 Seconds		Performance Metrics Chronic v in 3 minutes	
	Reg.Model	*MAPE*	*Reg.Model*	*MAPE*
etCO2	Poly.Deg.3	14.4	Poly.Deg.3	15.5
imCO2	Linear Reg	4.81	Poly.Deg.3	0.75
etO2	Linear Reg	2.75	Poly.Deg.3	0.84
inO2	Linear Reg	2.5	Poly.Deg.3	2.81
SPO2	Linear Reg	0.75	Poly.Deg.4	0.087

When performing the regression on the vital signs of heart failure disease to forecast the next 60 seconds data, polynomial degree 2 forecast the HR with minimum MAPE rate and Linear regression forecast the NBP(Sys), NBP(dia), NBP(Mean) and SPO2 with minimum MAPE rate. To forecasting next 3 minutes data of

vital signs, polynomial degree 2 forecast the HR with minimum MAPE rate, polynomial degree 4 forecast the NBP (Sys) and NBP(Mean) with minimum MAPE rate and Linear regression on NBP(Dia) and SPO2. Similarly, when regression is performed on the vital sign of chronic respiratory disease to forecast the next 60 seconds data of vital signs, linear regression forecast the SPO2, inO2, etO2, imCO2 with minimum MAPE rate. Polynomial degree 3 forecast the etCO2 with minimum MAPE rate. To forecast 3 minutes data, polynomial degree performs well on etCO2, imCO2, etO2 and inO2. So overall linear regression performs well to forecast next 60 seconds data and polynomial performs well to forecast 3 minutes data. We also got minimum error rate on regression algorithm to forecasting the next 60 second and 3 minutes of CVD and CRD disease by using mean of consecutive 100 sample of batches. *Research question 1 and 3 are answered here that we can find the minimum error rate on regression algorithm to predict the heart failure and chronic respiratory disease.*

4.5 Results and discussion of classification

In order to find the best combination for both the diseases, we choose the highest f-measures from classification techniques i.e. Nave Bayes[9], Support vector machine[8] and a Decision tree[10] as seen in Table 4.37. The below Table 4.37 contains all the performance metric parameters which tell us, which classification techniques perform well and have the high prediction accuracy. It also tells us about the size of the dataset and its effect on prediction. We also got high prediction accuracy to predict the critical condition of patient of CVD and CRD disease by using mean of consecutive 100 sample of batches.

Research question 5 is answered here that we can find the high accuracy classification algorithm to predict the heart failure and chronic respiratory disease by using hybrid regression model (combining two or more regression techniques). The below Table 4.38 summarized the Table 4.37 and only show the high accuracy score of f-measure to understand the result of all research.

TABLE 4.37: Combine results of three classification techniques with heart failure and CRD

CT	RM	Performance Metrics Heart Failure in 60 Seconds			Performance Metrics Chronic Respiratory in 60 Seconds			Performance Metrics Heart Failure in 3 minutes			Performance Metrics Chronic Respiratory in 3 minutes		
		Pre	Rec	FM	Pre	Rec	FM	Pre	Rec	FM	Pre	Rec	FM
NB	LR	0.07	0.18	0.10	0.0001	0.0001	0.0001	0.98	0.98	0.98	0.99	0.99	0.99
	PN2	0.05	0.18	0.08	1	0.34	0.51	0.001	0.001	0.001	1	0.12	0.22
	PN3	0.13	0.18	0.15	1	0.10	0.18	1.0	0.03	0.05	0.99	0.99	0.99
	PN4	0.001	0.001	0.001	0.0001	0.0001	0.0001	0.98	0.98	0.98	0.0001	0.0001	0.0001
	H.A	0.03	0.18	0.06	0.99	0.99	0.99	1	0.93	0.96	0.99	0.99	0.99
SVM	LR	0.0	0.0	0.0	0.0001	0.0001	0.0001	0.98	0.98	0.98	0.0001	0.0001	0.0001
	PN2	0.91	0.81	0.83	0.98	0.98	0.98	1.0	0.67	0.80	1	0.59	0.74
	PN3	0.07	0.18	0.10	0.001	0.001	0.001	1.0	0.59	0.74	0.001	0.001	0.001
	PN4	0.001	0.001	0.001	0.0001	0.0001	0.0001	0.98	0.98	0.98	0.0001	0.0001	0.0001
	H.A	0.91	0.84	0.85	0.001	0.001	0.001	0.99	0.99	0.99	0.001	0.001	0.001
DT	LR	0.18	0.12	0.14	0.0001	0.0001	0.0001	0.99	0.99	0.99	0.99	0.99	0.99
	PN2	0.82	0.71	0.76	0.0001	0.0001	0.0001	1.0	0.71	0.83	1	0.06	0.11
	PN3	0.001	0.001	0.001	0.001	0.001	0.001	1.0	0.61	0.76	0.99	0.99	0.99
	PN4	0.001	0.001	0.001	1	0.03	0.06	0.98	0.98	0.98	1	0.16	0.27
	H.A	0.93	0.89	0.90	0.18	0.12	0.14	0.98	0.98	0.98	0.99	0.99	0.99

*Legend CT=Classification Techniques, RM=Regression Model, NB=Naïve Bayes, SVM=Support Vector Machine, DT=Decision Tree, LR=Linear Regression, PN2=Polynomial Degree 2, PN3=Polynomial Degree 3, PN4=Polynomial Degree 4. HA=Hybrid Approach, Pre=Precision, Rec=Recall, FM=F-Measures. ■ Not Acceptable, ■ Highly Acceptable, ■ Acceptable

When performing the classification of heart failure disease and hypoxemic and hypercapnic disease on 60 seconds forecasted data, the prediction of Decision tree[10] got the high prediction score up to 0.90. Similarly, when the classification is performed on 3 minutes forecasted data, the prediction of Nave Bayes[9] got the high score 0.99. It concluded that when forecasting the 60 seconds the decision tree should be used for prediction and when forecasting the 3 minutes then Nave Bayes[9] perform well. *Research question 2 and 4 are answered here that we can find the high accuracy classification algorithm to predict the heart failure and chronic respiratory disease.*

TABLE 4.38: summarized form of Table 42 for best classification technique

Classifier	Performance Metrics Heart Failure in 50 Seconds			Performance Metrics Chronic Respiratory in 50 Seconds			Performance Metrics Heart Failure in 3 minutes			Performance Metrics Chronic Respiratory in 3 minutes		
	Prec	Rec	FM	Prec	Rec	FM	Prec	Rec	FM	Prec	Rec	FM
Naïve Bayes	0.03	0.18	0.06	1	0.93	0.96	0.99	0.99	0.99	0.99	0.99	0.99
SVM	0.91	0.84	0.85	0.99	0.99	0.99	0.001	0.001	0.001	0.001	0.001	0.001
DT (J48)	0.93	0.89	0.90	0.98	0.98	0.98	0.18	0.12	0.14	0.99	0.99	0.99

Chapter 5

Conclusion and Future Work

According to the *World Health Organization* (WHO) [3], cardiovascular diseases[1] and chronic respiratory[2] diseases are deemed as the major cause of death globally. Approximately 19 million people have died due to cardiovascular and chronic respiratory diseases. Moreover, they have demonstrated that individuals do not prefer to take medication services from poor medical systems, which are the main causes of the high death rates. Cardiovascular diseases[1] and chronic respiratory[2] diseases are now turned as a worldwide threat, medical experts and scientists are consistently in process of finding effective ways to control the death rates caused by these infections. Another imperative factor behind the high death rate caused by the cardiovascular and chronic respiratory diseases is not having an efficient medical expert system to forecast the vital sign of these diseases. Timely identification of major symptoms of these diseases could play a fundamental role to decrease the number of deaths globally. The current state-of-the-art prediction support systems do not perform timely forecasting of cardiovascular (i.e., heart failure) and chronic respiratory (i.e., hypercapnic and hypoxemic) diseases. Moreover, they only perform classification and are not efficient enough to forecast the future vital signs of these diseases. To overcome these deficiencies, the CCVS system proposed in this thesis will facilitate the medical experts and caregivers to forecast the vital signs of diseases (i.e., Heart Failure, Hypoxemic, and Hypercapnic) and will generate the health alerts (i.e., Low heart rate, or high heart rate

etc. (Table 3.9, 3.10)). These alerts will help the medical experts or the caregivers to effectively examine the current medical condition of patient and treat them accordingly. In this study, we have developed a disease prediction system of cardiovascular (i.e. Heart failure) and chronic respiratory diseases (i.e., Hypercapnic and hypoxemic). The system will be able to provide information about abnormal (i.e., Low heart rate, or high heart rate etc. (Table 3.9, 3.10)) vital signs caused by cardiovascular and chronic respiratory disease. We have used ten vital signs for cardiovascular prediction and chronic respiratory disease prediction (already discussed section 3.8). We have deployed a forecasting model using a technique like linear regression[6] and polynomial regression degree 2,3, and 4. The forecasting model used to predict cardiovascular (i.e. Heart failure) and chronic respiratory diseases (i.e., Hypercapnic and hypoxemic) vital signs of next 60 seconds and three minutes. We also performed hybrid regression (combining two or more regression techniques) to find the high prediction accuracy.Hybrid regression model produced high accuracy as compared to linear regression and polynomial regression degree 2,3, and 4. With hybrid regression model(combining two or more regression techniques), the decision tree[10] has obtained the best prediction result by achieving F-measure score of 0.90 for heart failure disease and 0.98 for the chronic respiratory disease. Similarly, when hybrid regression model(combining two or more regression techniques) was applied on the next three minutes dataset, the Nave Bayes[9] classifier has obtained the best prediction result by achieving F-measure of 0.99 for both diseases. The overall outcomes of the study yield that the proposed system has significant capability to predict the vital signs of cardiovascular and chronic respiratory diseases with high accuracy up to 0.98 scores.

Future Work We set a number of goals for future research as only we tried vital sign prediction for cardiovascular diseases and chronic respiratory diseases.

1. We have chosen five renowned regression methods and each of them has shown a different level of mean square error. It would be an interesting idea to investigate the reason behind its occurrence.

2. In our research, ten different vital signs are included and adding more vital signs may have more predictive power.

3. Building a hybrid model for making such prediction remain another future goal for the research.

4. We can further extend our experiment on mean of consecutive 100 sample of batches for more predictive power.

Bibliography

[1] N. Y. S. D. of Health, "Types of cardiovascular disease," 2012, access date 10/10/2016. [Online]. Available: https://www.health.ny.gov/diseases/cardiovascular/heart_disease/types_of_cv.htm

[2] A. Rebuck, M. Kangalee, L. Pengelly, and E. Campbell, "Correlation of ventilatory responses to hypoxia and hypercapnia." *Journal of Applied Physiology*, vol. 35, no. 2, pp. 173–177, 1973.

[3] W. H. O. (WHO), "Cardiovascular diseases (cvds)." [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/

[4] D. Liu, M. Görges, and S. A. Jenkins, "University of queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research," *Anesthesia & Analgesia*, vol. 114, no. 3, pp. 584–589, 2012.

[5] T. G. Pickering, J. E. Hall, L. J. Appel, B. E. Falkner, J. Graves, M. N. Hill, D. W. Jones, T. Kurtz, S. G. Sheps, and E. J. Roccella, "Recommendations for blood pressure measurement in humans and experimental animals: part 1: blood pressure measurement in humans: a statement for professionals from the subcommittee of professional and public education of the american heart association council on high blood pressure research," *Circulation*, vol. 111, no. 5, pp. 697–716, 2005.

[6] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," *Radiology*, vol. 227, no. 3, pp. 617–628, 2003.

[7] S. Gupta, "A regression modeling technique on data mining," *International Journal of Computer Applications*, vol. 116, no. 9, 2015.

[8] S. WAN, "An overview on support vector machines," 2012.

[9] K. M. Leung, "Naive bayesian classifier," *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[10] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 6, 2013.

[11] S. SA, "Intelligent heart disease prediction system using data mining techniques," *Int J Healthcare Biomed Res*, vol. 1, pp. 94–101, 2013.

[12] A. Farmer, O. Gibson, P. Hayton, K. Bryden, C. Dudley, A. Neil, and L. Tarassenko, "A real-time, mobile phone-based telemedicine system to support young adults with type 1 diabetes," *Journal of Innovation in Health Informatics*, vol. 13, no. 3, pp. 171–177, 2005.

[13] N. Aït-Khaled, D. Enarson, and J. Bousquet, "Chronic respiratory diseases in developing countries: the burden and strategies for prevention and management," *Bulletin of the World Health Organization*, vol. 79, no. 10, pp. 971–979, 2001.

[14] V. Chaurasia, "Early prediction of heart diseases using data mining techniques," 2017.

[15] S. Vijiyarani and S. Sudha, "Disease prediction in data mining technique– a survey," *International Journal of Computer Applications & Information Technology*, vol. 2, no. 1, pp. 17–21, 2013.

[16] Y.-F. Lin, H.-H. Shie, Y.-C. Yang, and V. S. Tseng, "Design of a real-time and continua-based framework for care guideline recommendations," *International journal of environmental research and public health*, vol. 11, no. 4, pp. 4262–4279, 2014.

[17] K. A. Shakil, S. Anis, and M. Alam, "Dengue disease prediction using weka data mining tool," *arXiv preprint arXiv:1502.05167*, 2015.

[18] A. K. Tiwari, "Machine learning based approaches for prediction of parkinson disease," *Mach Learn Appl*, vol. 3, no. 2, pp. 33–39, 2016.

[19] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," *Heart Disease*, vol. 7, no. 1, 2015.

[20] J. Wan, C. Zou, S. Ullah, C.-F. Lai, M. Zhou, and X. Wang, "Cloud-enabled wireless body area networks for pervasive healthcare," *IEEE Network*, vol. 27, no. 5, pp. 56–61, 2013.

[21] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[22] P. Soucy and G. W. Mineau, "A simple knn algorithm for text categorization," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on.* IEEE, 2001, pp. 647–648.

[23] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8852–8858, 2012.

[24] C. H. Alliance, *IEEE STANDARDS ASSOCIATION AND CONTINUA HEALTH ALLIANCE JOIN FORCES TO DEVELOP END-TO-END, PLUG-AND-PLAY CONNECTIVITY FOR PERSONAL CONNECTED HEALTH*, Continua Health Alliance, 2013. [Online]. Available: http://standards.ieee.org/news/2013/ieeesa_continua.html

[25] M. Durairaj and V. Ranjani, "Data mining applications in healthcare sector: a study," *International journal of scientific & technology research*, vol. 2, no. 10, pp. 29–35, 2013.

[26] S. Ruggieri, "Efficient c4. 5 [classification algorithm]," *IEEE transactions on knowledge and data engineering*, vol. 14, no. 2, pp. 438–444, 2002.

[27] K. Archana, S Elangovan, "Survey of classification techniques in data mining," *International Journal of Computer Science and Mobile Applications*, vol. 2, pp. 65–71, 2014.

[28] D. G. Denison, B. K. Mallick, and A. F. Smith, "A bayesian cart algorithm," *Biometrika*, vol. 85, no. 2, pp. 363–377, 1998.

[29] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming.* Springer, 2003, pp. 81–100.

[30] J. Adamo, "Fuzzy decision trees," *Fuzzy sets and systems*, vol. 4, no. 3, pp. 207–219, 1980.

[31] G. Haitao, Z. Qingbao, and X. Shoujiang, "Rapid-exploring random tree algorithm for path planning of robot based on grid method [j]," *Journal of Nanjing Normal University (Engineering and Technology Edition)*, vol. 2, no. 14, 2007.

[32] J. R. Quinlan *et al.*, "Bagging, boosting, and c4. 5," in *AAAI/IAAI, Vol. 1*, 1996, pp. 725–730.

[33] S. Bind, A. K. Tiwari, A. K. Sahani, P. Koulibaly, F. Nobili, M. Pagani, O. Sabri, T. Borght, K. Laere, and K. Tatsch, "A survey of machine learning based approaches for parkinson disease prediction," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1648–1655, 2015.

[34] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the c4. 5 and a naive bayes classifier for the prediction of lung cancer survivability," *arXiv preprint arXiv:1206.1121*, 2012.

[35] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[36] C. f. D. C. a. Prevention., "Centers for disease control and prevention,." [Online]. Available: https://www.cdc.gov/

[37] R. Prashanth, S. D. Roy, P. Mandal, and S. Ghosh, "Parkinson's disease detection using olfactory loss and rem sleep disorder features," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.* IEEE, 2014, pp. 5764–5767.

[38] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm," *Global journal of computer science and technology*, vol. 10, no. 10, pp. 38–43, 2010.

[39] D. S. Jeyarani, G. Anushya, and A. Pethalakshmi, "A comparative study of decision tree and naive bayesian classifiers on medical datasets," *age*, vol. 30, pp. 31–40, 2013.

[40] wikipedia, "Global ranges of vital signs,," Tech. Rep., 2017. [Online]. Available: https://en.wikipedia.org/.

[41] Wikipedia, *Alveolar Units*, Wikipedia, 2017.

[42] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.