



The Relation Between Respiratory & Acute Coronary Syndrome Using Data Mining Techniques

العلاقة بين الجهاز التنفسي والتهاب الشريان التاجي الحادة باستخدام تقنيات التقيب عن البيانات

by

HANI ABDULLAH YOSEF ABU DOULEH

**A dissertation submitted in fulfilment
of the requirements for the degree of
MSc INFORMATION TECHNOLOGY MANAGEMENT
at
The British University in Dubai**

**Dr. Sherief Abdallah
April 2018**

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Hani Abdullah Yosef Abu Douleh

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

In healthcare world, data science is one of the most important sciences that helps in predicting diseases, and despite the availability of medical data from laboratory tests, most medical institutions in middle east region still do not benefit from these data in diseases analysis and prediction. The purpose of this study is to diagnose acute coronary syndrome using the widely available respiratory diagnosing tools and laboratory test results using data mining classification techniques (Decision tree, Gradient boosted tree, Neural Network, and Naïve Bias). In this study I've split one dataset of patients who have attended to emergency departments in Abu Dhabi hospitals to two datasets (Respiratory and Cardiac), then applied the data mining algorithms on each dataset and one time on the original dataset. This study found that respiratory features such as sO₂(Oxygen saturation), O₂Hb (Oxyhemoglobin), pH and HHb (Deoxyhemoglobin) values can predict if the patient is an acute coronary syndrome or there is a possibility be affected by this disease.

ملخص

في عالم الرعاية الصحية ، يعتبر علم البيانات من أهم العلوم التي تساعد على التنبؤ بالأمراض ، وعلى الرغم

من توافر المعطيات الطبية من الاختبارات المعملية ، فإن معظم المؤسسات الطبية في منطقة الشرق الأوسط

لا تزال لا تستفيد من هذه البيانات في تحليل الأمراض والتنبؤ بها. إن الغرض من هذه الدراسة هو العثور

على سبب ألم الصدر وتشخيص متلازمة الشريان التاجي الحاد بشكل خاص عن طريق استخدام اجهزة

التشخيص التنفسى المتاحة على نطاق واسع بالإضافة إلى نتائج اختبار الخوارزميات (شجرة القرار، الشجره

المعززه، الشبكه العصبيه و الإنحياز) مع تقنيات تصنيف التنبؤ عن البيانات. قسمت الدراسه مجموعه

بيانات واحدة من المرضى الذين حضروا إلى أقسام الطوارئ في مستشفيات أبو ظبي إلى مجموعه بيانات

(الجهاز التنفسى والقلب) ، ثم طبقت خوارزميات استخراج البيانات مرة واحدة لكل مجموعه بيانات ومرة

واحدة على مجموعه البيانات الأصلية. وجدت هذه الدراسه ان خصائص الفحوصات التنفسيه مثل تشعب

الأكسجين، نسبة ديوکسي هيمو غلوبين، حمضية الدم ونسبة الأوکسی هيمو غلوبين يمكن ان تتنبأ اذا كان

المريض مصابا بمتلازمة الشريان التاجي او ان هناك امكانيه قريبه للاصابة بهذا المرض.

Dictionary

Word	Decryption
Radiometer	Radiometer Medical Aps the Blood gas and Cardiac analyzer manufacturer company since 1935, based in Denmark.
ACS	Acute Coronary syndrome
RACS	Respiratory Acute Coronary Syndrome
HF	Heart Failure

Contents

1	Chapter 1: Introduction	1
1.1	Dissertation document Structure	1
1.2	Overview.....	2
1.3	Problem Definition.....	2
1.4	Motivation.....	2
1.5	Research Goals.....	3
1.6	Research Characteristics	3
1.7	Research Questions.....	3
2	Chapter 2: Literature Review	4
2.1	Introduction.....	4
2.1.1	Data Mining	4
2.1.2	Data Mining and Diseases prediction	5
2.1.3	Heart failure early prediction using data mining.....	6
2.1.4	Respiratory and Blood gases analysis	7
2.1.5	Acute coronary syndrome and Troponin T analysis	8
2.2.0	Literature Conclusion.....	8
3	Chapter 3: Research Design and Methodology	9
3.1	Research Design.....	9
3.1.1	Research Methodologies	10
3.1.2	Data Sets and Collections	10
3.1.3	Data preparation challenges	10
3.1.4	Class Labels	12
3.1.5	Duplicate Records.....	14
3.1.6	Attributes Elimination and Missing Values	14
3.1.7	Final Datasets.....	14
3.1.8	Respiratory Dataset.....	14
3.1.9	Cardiac Dataset	15
3.2	Classification Algorithms	15
3.3	Decision Tree	19
3.3.1	Prediction Results for Decision Tree	20
3.3.2	Decision Tree Performance Analysis.....	20

3.3.2.1	Respiratory Dataset (DT)	20
3.3.2.2	Cardiac Dataset (DT)	21
3.3.2.3	The Combined Datasets (DT)	22
3.3.3	Decision Tree Features Importance Analysis	22
3.3.3.1	Respiratory dataset (DT)	22
3.3.3.2	Cardiac Dataset (DT)	24
3.3.3.3	The Combined Datasets (DT)	24
3.4	Gradient Boosted Tree	26
3.4.1	Gradient Boosted Tree Prediction Results	26
3.4.2	Gradient Boosted Tree Performance analysis.....	26
3.4.2.1	Respiratory Dataset (GBT)	27
3.4.2.2	Cardiac Dataset (GBT).....	27
3.4.2.3	The Combined Datasets (GBT).....	28
3.4.3	Gradient Boosted Tree Features Importance Analysis.....	29
3.4.3.1	Respiratory Dataset (GBT)	29
3.4.3.2	Cardiac Dataset (GBT).....	31
3.4.3.3	The Combined Datasets (GBT).....	33
3.5	Neural Network (NN)	34
3.5.1	Neural Network Prediction Results.....	34
3.5.2	Neural Network Performance analysis	35
3.5.2.1	Respiratory Dataset (NN).....	35
3.5.2.2	Cardiac Dataset (NN).....	36
3.5.2.3	The Combined Datasets (NN).....	36
3.5.3	Neural Network Features Importance Analysis	37
3.5.4	Respiratory Dataset (NN).....	38
3.5.5	Cardiac Dataset (NN).....	40
3.5.6	The Combined Datasets (NN).....	41
3.6	Naïve Bias (NB).....	41
3.6.2	Naïve Bias Prediction Results.....	41
3.6.2.1	Naïve Bias Performance analysis.....	42
3.6.2.2	Respiratory Dataset (NB).....	42
3.6.2.3	Cardiac Dataset (NB)	42
3.6.2.4	The Combined Datasets (NB)	43

3.6.3	Naïve Bias Features Importance Analysis	44
3.6.3.1	Respiratory Dataset (NB).....	44
3.6.3.2	Cardiac Dataset (NB).....	45
3.7	Discussion	47
4	Chapter 4: Conclusion.....	52
	References.....	54

List of Figures

Figure #	Description
Figure 1	Cardiac and Respiratory
Figure 2	First raw data from SQL Server before any modification
Figure 3	Data after duplicate deletion
Figure 4	Data after conversion rows to columns
Figure 5	Data after cleaning from unwanted feature, removing 0s values and converted date to integer
Figure 6	Rapid Miner Main process with Discretize and Replace Missing value operator
Figure 7	Rapid Miner Enabling and Disabling operators
Figure 8	Rapid Miner Process Cross validation
Figure 9	Rapid Miner Bayesian Boosting operator
Figure 10	Rapid Miner Bagging operator
Figure 11	Decision Tree output- Respiratory dataset
Figure 12	Decision Tree output
Figure 13	Decision Tree output – The Combined Datasets
Figure 14	R1: Respiratory Class R1 for gradient boosted tree - HHb
Figure 15	R1: Respiratory class for gradient boosted tree Oxygen saturation– sO2
Figure 16	R2: Respiratory class for gradient boosted tree pH
Figure 17	Cardiac Dataset for gradient boosted tree – NT-proBNP
Figure 18	Cardiac Dataset for gradient boosted tree – NT-proBNP, PCT and Age
Figure 19	The Combined Datasets R1 Class for gradient boosted tree – sO2
Figure 20	The Combined Datasets HPACS class for gradient boosted tree – NT-proBNP
Figure 21	Sigmoid function
Figure 22	Sigmoid Graph
Figure 23	Neural Network Attribute weight – Respiratory Dataset
Figure 24	Neural Net attributes importance using Node weight – Cardiac Dataset
Figure 25	O2Hb importance in Naïve Bias – Respiratory Dataset
Figure 26	sO2 importance in Naïve Bias – Respiratory Dataset
Figure 27	Nt-ProBNP importance in Naïve Bias – Cardiac Dataset
Figure 28	PCT importance in Naïve Bias – Cardiac Dataset

List of Tables

Table #	Decryption
Table 1	Related studies predicted heart failure
Table 2	Final Attributes for Respiratory Dataset
Table 3	Final Attributes for Cardiac Dataset
Table 4	Respiratory Dataset True Classes
Table 5	Cardiac Dataset True Classes
Table 6	Combined Dataset True Classes
Table 7	Decision Tree Accuracy – All Datasets
Table 8	Decision Tree confusion matrix- Respiratory Dataset
Table 9	Decision Tree confusion matrix – Cardiac Dataset
Table 10	Decision Tree confusion matrix – The Combined Datasets
Table 11	Gradient Boosted Tree Accuracy –All Datasets
Table 12	Gradient boosted tree confusion matrix - Respiratory Dataset
Table 13	Gradient boosted tree confusion matrix - Cardiac Dataset
Table 14	Gradient boosted tree confusion matrix – The Combined Datasets
Table 15	Neural Network Result Accuracy –All Datasets
Table 16	Neural Network confusion matrix - Respiratory Dataset
Table 17	Neural Network confusion matrix - Cardiac Dataset
Table 18	Neural Network confusion matrix – The Combined Datasets
Table 19	Neural Net Sigmoid values – Respiratory Dataset
Table 20	Neural Net Sigmoid values – Cardiac Dataset
Table 21	Neural Net Sigmoid values – The Combined Datasets
Table 22	Naïve Bias Accuracy results – All Datasets
Table 23	Naïve Bias confusion matrix - Respiratory Dataset
Table 24	Naïve Bias confusion matrix - Cardiac Dataset
Table 25	Naïve Bias confusion matrix – The Combined Datasets
Table 26	All Methods Accuracy Summary- All Datasets
Table 27	Algorithms Performance Summary – Respiratory Dataset
Table 28	Algorithms Performance Summary – Cardiac Dataset
Table 29	Algorithms Performance Summary – the Combined Datasets

1 Chapter 1: Introduction

In this part, a general structure of the dissertation is presented. An overview of the topic, the problem definition is featured. Research motivation is characterized. The point of this research has been obviously expressed. The research questions are portrayed quickly. The used techniques are illustrated.

1.1 Dissertation document Structure

In this part, the reader will be able to understand the whole document structure. This document is divided into four chapters to ensure the ease of reading and better understanding:

Chapter 1: Includes the document structure, overview of the topic, problem definition, motivation and the research goals. Moreover, some questions which this dissertation answers.

Chapter 2: the literature review, in this chapter, will organize the topic by the headline and will not mix-up the related work, this was proffered to ensure content and structure easy reading and understanding.

Chapter 3: The research engine, it contains the methodology, data collection, preprocessing steps and displays the raw results with initial comments, and the study output, it contains the research analysis, discussions and limitations.

Chapter 4: the conclusion of the research followed by references.

1.2 Overview

Heart failure, respiratory and acute coronary syndrome are diseases cause chest pain. this dissertation will discuss the three diseases and will use data mining and exploration to link those diseases together to predict the right reason for chest pain. Data mining is a non-trivial extraction of understood, already obscure and potentially valuable data around information (Sudhakar & Manimekalai,2014), Heart attack diseases remain the primary driver of death around the world, the identification at a previous level will keep the diseases medical specialists produce data with an abundance of hidden data present, and it's not correctly being used correctly for forecasts (Masethe,2014). Respiratory Distress Syndrome (RDS) is a serious lung disease with a high death rate (Schmidt et al, 2013). Acute coronary syndrome (ACS) caused by increased cardiac troponin in the blood which is monitored essentially after each open-heart surgery, showing perioperative myocardial cell injury (Lehrke Et al 2004).

Feeling chest pain is a massive problem as chest pain have many reasons, and up to date chest pain diagnoses are hard as it is related to the most critical functions in the human body, Lung and Heart and both parts are very sensitive and lead to human death if ignored or were not taken seriously. Most of the emergency patients who attend with chest pain will stay at the hospital at least for six hours as the required diagnoses and laboratory testing routines take time, and patient needs to be monitored during the stay at the hospital to avoid stork or blood coagulation. This study came to merge both respiratory and specific type of heart failure called acute coronary syndrome to be a novel study which discusses and analyze the relationship between both diseases and predict the correct reason of chest pain.

1.3 Problem Definition

The problem lies in the need to speed up the diagnosis of chest pain in emergency departments and ambulances as well reduce doctors and nurses stress working in this area to provide more time for patient care. moreover, finding additional options for diagnosing cardiac diseases with lower cost as the cardiac diagnose equipment and techniques are not available in all emergency departments due to its high cost.

1.4 Motivation

There are many studies that have predicted heart failure diseases using data mining and exploration techniques, but a few studies have predicted respiratory diseases which eventually may lead to problems in the heart for direct contact with the functions of the body, which in turn affect the heart and everything related to it directly such as the efficiency of blood vessels. This study will integrate heart and respiratory problems into one database and use them to predict heart failure or respiratory failure which causes the patient chest pain.

1.5 Research Goals

The purpose of this study is to help staff in the emergency departments and trauma centers to discover the cause of chest pain using respiratory analysis results quickly and take the necessary action as soon as possible to avoid patients any clots of cardiac or venous, which often lead to the death of the patient if the chest pain was not taken seriously.

1.6 Research Characteristics

This research is novel due to several reasons, first the difficulty of obtaining the medical data due to its sensitivity, and this has been achieved because of the nature of my work for the manufacturer of the medical devices directly responsible for heart disease and respiratory systems. Second, this research is the first to integrate respiratory laboratory tests and heart diseases work on the creation of medical links between the functions of the two systems, for example the effect of the proportion of oxygen in the blood on the secretion of the enzyme Troponin T, which warns if it's percent increases that problem may occur in the heart at any future moment. Thirdly, this research depends on patients testing results without any patient's habits or characteristics definition, dataset doesn't know whether the patient was smoker, drinking alcohol, has blood pressure, chest pain type, medication, etc.., the only known facts about patients were the age and sex, which means the dependency of this research was only pure blood testing results, and actually this what happens in emergency rooms, hard to know patient medical history.

1.7 Research Questions

- Does combining data from respiratory and cardiac results in higher precisions?
- Can respiratory diagnosis equipment's be a step one in diagnosing cardiac (Chest Pain)?
- Can existing classification techniques achieve reasonable precession when predicting chest pain using cardiac and respiratory results?
- Can respiratory diagnosis equipment's be a step one of diagnosing Hypertrophy of the heart muscle?

2 Chapter 2: Literature Review

This research has combined two totally different topics, Cardiac and Respiratory, and even in my meetings with doctors and specialists, no one give a direct answer about how I can combine respiratory and cardiac to predict the chest pain, and this was understood, as there is no direct relation between them, only experience and family history can answer this question. It was not easy to find studies related to both topics, that's why I've spilt the literature review into sections, each section lists each topic individually.

2.1 Introduction

In this part, the related work of previous studies in the same field will be listed and will be categorized under different headlines and at the end, it will be simplified to help this dissertation to reach its goals. I have divided the literature review into four parts ends with a conclusion and learned lessons as below:

1. Data Mining
2. Data Mining and Diseases prediction
3. Heart Failure early prediction
4. Respiratory and blood gases analysis
5. Acute coronary syndrome and Troponin T analysis

2.1.1 Data Mining

Data Mining Definitions by previous researchers:

1. Data Mining and exploration is the methods of removing concealed learning from the huge data. Learning should not be self-evident and must be new and have the capacity to use it. Moreover, it can be "the nontrivial extraction of beforehand obscure, certain and conceivably valuable data from information (Han et al, 2011).
2. It is "the exploration of extricating helpful data from huge databases". It is one of the undertakings during the time spent information revelation from the database (Han et al, 2011).
3. (Masethe,2014) also defined Data mining as information revelation strategy to analyze data and typify it into valuable data.
4. Data Mining is a non-trivial extraction of understood, already obscure and potential valuable data around information (Sudhakar & Manimekalai,2014).
5. Data mining has officially settled as a novel field for investigating shrouded designs in the colossal datasets. Medicinal science is another field where a lot of information is produced utilizing distinctive clinical reports and other patient side effects (Taneja, 2013).

6. Data mining uses two systems: regulated and unsupervised learning. In administered taking in, a making ready set is created to learn display parameters though in unsupervised adapting no preparation set is utilized (e.g., k-means clustering is unsupervised) (SA, 2013).

7. Data Mining is a standout amongst the most imperative and rousing zone of research with the goal of finding significant data from colossal informational indexes. In display time, Data Mining is getting to be prevalent in human services field on the grounds that there is a need of an effective investigative system for recognizing obscure and significant data in wellbeing data (Tomar & Agarwal, 2013).

2.1.2 Data Mining and Diseases prediction

There is a lot of studies used data mining in diseases prediction, and especially the classification techniques, healthcare industry collects colossal sums of healthcare information and that require to be mined to find covered up data for successful choice making. Find of covered up designs and connections regularly go unexploited (Chitra & Seenivasagam, 2013). data mining procedures are valuable for foreseeing the different maladies in the restorative field. Malady forecast plays a vital part in information mining. There are distinctive sorts of illnesses foreseeing in information mining specifically heart illnesses, lung cancer and breast cancer (Banu & Gomathy, 2014).

The healthcare industry collects colossal information of social insurance which, lamentably, are not mined to find concealed data for forcing basic leadership. The revelation of hidden examples and connections frequently goes unexploited. Propelled information mining strategies can help cure this circumstance. The therapeutic analysis is viewed as an imperative yet confused assignment that should be executed precisely and proficiently. The computerization of this framework would be greatly invaluable. Unfortunately, all specialists don't have aptitude in each subspecialty and in addition, there is a lack of asset people at specific spots. In this manner, a programmed therapeutic analysis framework would likely be exceedingly useful by combining every one of them (Chapman et al ,2000).

List of studies predicted heart failure, techniques and number of features/attributes used:

Author/s	Techniques	Attributes
(Ordonez et al,2001)	Association rules	25
(Rani, 2011)	Classification: Neural Network	13
(Nahar et al,2013)	Predictive Apriori/ Tertius	14
(Sundar et al, 2012)	genetic algorithm / CANFIS	14
(Jabbar et al,2011)	Clustering/ Association rule mining / Sequence number,	14
(Ishtake & Sanap, 2011)	Decision Tree/ Neural Network/ Naive Bayes	15
(Hsieh et al, 2012)	(EVAR) / Machine learning / Markov blanket	
(Atkov et al,2012)	artificial neural network/ genetic polymorphisms	
(Pattekari & Parveen, 2012)	Naive bias	15

(Patil & Kumaraswamy, 2009)	MAFIA / Clustering/ K-Means	13
-----------------------------	-----------------------------	----

Table 1: Related studies predicted heart failure

(Kaur & Singh, 2014) listed the and defined association and classification Techniques which used in data mining:

Association: One of the top-known dtata mining methods. In Association, a pattern is found based on a relationship of the thing on other things in the same exchange. For instance, the association method is used in heart diseases forecast as it tells us the relationship of diverse properties utilized for analysis and sorts out the understanding with all the risk figure which are required for the expectation of disease.

Classification: An old data mining procedure based on machine learning. Essentially, classification is utilized to classify each item in a set of data into one of a predefined set of classes or bunches. Classification strategy makes use of scientific methods such as neural network, linear programming and decision trees.

2.1.3 Heart failure early prediction using data mining

Most of the previous work used direct characteristics of heart failure reasons, for example, Smoker, Chest torment write, pulse, stoutness, practices and so on which truly help to recognize the patient case as a matter of fact just, for example, hypertension is outstanding influences venous and heart and can without much of a stretch reason stork or enormous heart failure. The heart is a critical organ or portion of our body. Life is itself subordinate to the proficient working of the heart. In case the operation of the heart is not legitimate, it will influence the other body parts of a human such as a brain, kidney etc. It is nothing more than a pump, which pumps blood through the body. On the off chance that circulation of blood in the body is wasteful the organs like brain endure and, in the event, that the heart stops working through and through, passing happens inside minutes. Life is totally subordinate to the productive working of the heart (Sudhakar & Manimekalai, 2014). Cardiovascular diseases are one of the highest death reasons modern world (Delen and Olson, 2008).

According to (WHO) the world health organization, more than twelve million demises happen around the world because of heart problems. It is likewise one of the most illnesses in India which causes greatest losses (Taneja, 2013).

Several researchers are using statistical and data mining tools to help healthcare professionals in the diagnosis of heart disease (Xing et al, 2007).

Cardiac problems and diseases, also increasing, include a major portion of not able to be communicated to other diseases. In 2010, of all noticed worldwide deaths, (Kotaska et al, 2010). According to World Health Organization and Research for International Tobacco Control, 1 million are predicted to die due to cardiac diseases. In fact, CVDs was the only biggest reasone of death in the world accounting for more than a third death. (Taneja, 2013) implemented a prediction model and used data from cardiac hospital for 7,080 patients and he utilized three strategies, J48 tree, Naïve Bias and Neural network, this examination utilizes two methods for try, initial one was to utilize all the 15 accessible characteristics in the dataset, and the second uses 8 chose qualities, in every procedure test the exactness was higher when they utilized the selected attributes, not all attribute which empowers analyst to expand their precision to over

90%, and removed out the unwanted attribute. (Chaurasia, 2017) used data mining techniques to evaluate and predict heart failure using Cleveland dataset, in this study researcher used three approaches, CART, ID3 and decision tree the good thing of this research was ability of investigating the distinction between techniques, by detailed analysis of every method result, which motivate other researchers to keep ongoing of enhancing the current studies regardless of whether that review achieves high precision. (Banu & Gomathy, 2014) implemented data mining techniques model to predict heart failure in this study researcher used three techniques, K-mean based MAFIA, K-mean based MAFIA with ID3 and K-mean based MAFIA with ID3 and C4.5 in this research it was recognizable that no distinction between strategies in the study contrasts between strategies leads to diverse result and this was an issue which (10) have to maintain a strategic distance in consideration, utilizing different strategies is required to discover more design and valuable analysis.

(Xing et al, 2007) listed some reasons which indicate the increment danger of Heart disease

- Parents and family experience of heart disease
- Smoking
- Cholesterol
- wrong diet
- out of range blood pressure
- out of range blood cholesterol
- Obesity rate
- Physical laziness
- Hyper tension

2.1.4 Respiratory and Blood gases analysis

(Schmidt et al, 2013) defined respiratory as Acute respiratory distress syndrome (ARDS) is a serious lung malady with a high death rate. Patients getting its most serious structures, with extreme hypoxemia, have the most exceedingly terrible anticipation, and death rate can surpass 60 % In these cases, extracorporeal pulmonary assistance (oxygenation and CO₂ elimination from the blood), also experienced as extracorporeal membrane oxygenation (ECMO)

(COPD: Chronic obstructive pulmonary disease) is the fourth leading cause of death worldwide (Cerretelli, 1976). The calculated annual costs of COPD are Twenty-Four billion, and 70% are connected to exacerbations requiring hospitalization (West et al, 1983). Adding that to their large number of acute morbidity, death rate, and cost, exacerbations are also linked with main slip in long-term quality of life and lung function (West et al, 1983).

(Grocott et al, 2009) developed a research method discover out the distinction of po₂, pCO₂ and sO₂ concentration in the blood sometime recently, amid and after climbing everts mount, the study was conducted on 10 climbers and result appeared the sO₂ stay steady but the other parameters were variation, which showed that the circumstance, area, and blood pressure increment and decrease the oxygenation concentration in the blood which influencing other body capacities such as lung, brain, and heart as anticipated to be demonstrated in this research.

The pH value in the umbilical artery is the best examination to evaluate the presence and intensity of the fetal acidosis, as they reflect the corrosive base status of the fetal tissue. The value in the umbilical vein reflects the blood that returns to the baby because of the exchange of CO₂ and O₂ through the placenta (Kotaska et al, 2010).

2.1.5 Acute coronary syndrome and Troponin T analysis

(Sudhakar & Manimekalai,2014) defined Coronary HF illness: It is also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.

(Lehrke Et al 2004) said: Increased cardiac troponins in blood are watched after essentially each open-heart surgery, showing perioperative myocardial cell injury. researchers looked for to decide the ideal time point for blood testing and the particular cutoff value of cardiac troponin T (cTnT) for risk assessment in patients experiencing cardiac surgery. blood tests for estimation of cTnT were taken before heart surgery, 4 and 8 h after aortic cross clamping, and each 24 h amid the primary postoperative week or until release. cTnT was estimated quantitatively by a one-advance compound immunoassay in light of electrochemiluminescence innovation (Elecsys 2010; Roche). The lower identification point of confinement of this measure is 0.01 _g/L with a prescribed analytic edge of 0.03 _g/L for unconstrained AMI. The intraassay CVs (between-day imprecision informational index of no less than 11 runs) were 20% at 0.015 _g/L, 10% at 0.03 _g/L, and 5% at 0.08 _g/L. (Del-Carlo et al, 2004) developed a study to determine if serial determinations of cardiovascular troponin T (cTnT) in decompensated heart failure (HF) are predictive of clinical events (death, need for readmission for new episode of HF decompensation, or both) during 1 year of follow-up. 62 patients with decompensated HF were enrolled in this associate. The main estimation of cTnT (cTnT1) was from a blood test drawn inside 4 days of healing facility confirmation; the second measurement(cTnT2) was on blood acquired 7 days after the fact. (Prasad et al, 2006) clinical occasions (16 death, 10 readmissions, 23 consolidated readmissions, and death) happened during the follow-up. independent predictors of clinical events were: cTnT1 _ .020 ng/mL(P _ .050), cTnT2 _ .020 ng/mL (P _ .050), and serum sodium _ 135 mEq/L (P _ .050). In view of levels of cTnT1 and cTnT2 _ .020 ng/mL () or _0.020 ng/mL(–), patients were partitioned into 2 gatherings: aggregate 1 (cTnT1_-, cTnT2_- or cTnT1_-, cTnT2_), assemble 2 (cTnT1_-, cTnT2_- or cTnT1_-, cTnT2_). Group 2 patients had higher rates of death (45.0% versus 71.4%, P .050), hospital readmission (35.0% versus 61.9%, P _ .050), and clinical occasions (55.0% versus 90.5%, P _ .010) than group patients. The Conclusion, persistently increased cTnT levels (_0.020 ng/mL) are prescient of higher rates of death and healing facility readmission for decompensated HF.

2.2.0 Literature Conclusion

There were no studies found talking about the combination of respiratory and cardiac, each topic has its own characteristics and specialties. This research will combine data mining, respiratory and cardiac to predict chest pain.

3 Chapter 3: Research Design and Methodology

In this chapter research design, methodologies, data collection, and processing are demonstrated with the data mining tool used.

3.1 Research Design

One dataset of 105,257 records were used in this research, the dataset was a combination of two blood testing analyzers, the first analyzer measures blood gases, and the second one measures cardiac, this dissertation dataset was divided into two datasets, first called Respiratory dataset to evaluate respiratory, and the second called cardiac dataset to predict cardiac, and at the end the whole dataset will be used to predict respiratory and cardiac at the same time.

This research was designed by combining those datasets of blood machine testing results, research designed to employ data mining to evaluate and investigate the patient's data who visited emergency and accident departments in Abu Dhabi since June 2017 till Jan 2018. This study combines the results of patients with chest pain and respiratory problems. Total of 105,257 test records collected, only the patients who examine both respiratory and Cardiac were taken as seen in figure 1, and the rest were ignored. Eventually, 1,791 records were selected for the final data preprocessing and cleaning.

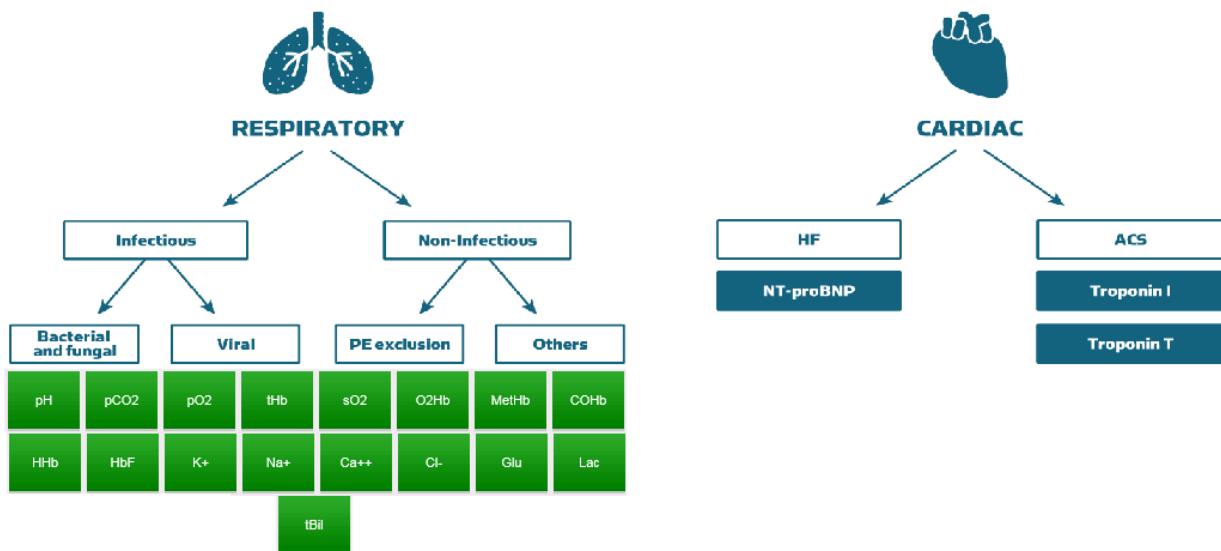


Figure 1: Cardiac and Respiratory

3.1.1 Research Methodologies

For respiratory dataset, I used Data Mining classification techniques to classify the patient in the dataset into three groups, Normal, Respiratory type1(R1) and Respiratory type2 (R2).

For Cardiac dataset, I used Data Mining classification techniques to classify the patient in the dataset into three groups, Normal, High possibility of Acute Coronary Syndrome (HPACS), and already has Acute Coronary Syndrome (ACS).

For the original dataset, Normal, Respiratory (R1 or R2), Respiratory + ACS (R-ACS) and Acute Coronary Syndrome (ACS).

using Rapid Miner version 7.4, the datasets was processed to predict the class label. data mining techniques will be applied to each dataset separately then will discuss the results.

3.1.2 Data Sets and Collections

The main dataset used includes blood gas and Cardiac results of UAE based patients from June 2017 till Jan 2018. The used data format was unstructured and required different steps of normalization to have it ready to be used with Rapid Miner.

By virtue of my work with a medical company that performs clinical and laboratory tests in the fields of heart and respiratory system, I obtained a database of the company's assists replacement medical devices from Abu Dhabi hospitals (Shiekh Khalifa Medical City and Mafraq Hospital) and connected that devices to interface engine program called AQURE POC to extract the results of the patients to SQL database instead of extracting results of each device individually, extracting data from devices was step one of data preparation.

3.1.3 Data preparation challenges

The obtained data was huge but dirty, 105,257 records for a large number of repeat patients and most of the results were deficient and does have many missing values, I used SQL SERVER instead of Rapid Miner to filter out the results, delete the recurrence and take the results of the patients who examined the blood gases and Cardiac testing only and ignore the rest, each patient in the database has 28 records as the database format was test per row as each patient performs 22 test per panel which means we have to convert the table of 28 rows to one row to hold that data and to facilitate the work of Rapid Miner. At the end total of 306 records was obtained, all patient has valid age and sex which is required in the research analysis.

The steps of data peroration were as below:

1. Extracted the required fields from AQURE POC system to one SQL server view holds all results with patient demographics, device location, and test performed.

2. Created SQL Server table called Allress and imported data from the view from AQUIRE POC system to my dissertation database, total 105,257 records. As shown in Figure 2.
3. Created another SQL table called AllRessV001 hold all results in AllRess but deleted all repeated patient tests. total 45,713 record. As shown in Figure 3.
4. Created another SQL table called AllRessV002 but converted rows to columns (pivoting) and data was transferred from AllRessV001 to AllRessV002 using vb.net desktop program made by me, I preferred this step to keep my data in SQL format a SQL provides better search and filtration experience. Total record at this point was 1763 record, each record holds all patient features as shown in Figure 4.
5. Another table called AllRessV003 was created after data fill-up from the previous step, but in this table, I imported only the correct patients who have tested blood gases and Cardiac at same period to ensure better accuracy for the study. This ends the table with 306 records as shown in figure 5.
6. It was required to convert date from date format to age and this was done in table AllRessV004 as shown in figure 5.
7. Last was removing the unnecessary feature from the dataset, and with the discussion with Dr. Fayha (Consultant at Dubai Hospital) she advised to keep only 24 parameters out of 28 parameters and this will be covered in detail features selection.

	TimeAnalyzed	DisplayName	TestName	ValueString	HospitalName	Sex	DateOfBirth	PID
1	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	pCO2	4.30889755185233	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
2	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	tBil	30.0298709869385	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
3	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	pCO2(T)	4.30889755185233	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
4	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	pCO2	5.68579220628238	MAFRAQ HOSPITAL	2	04/29/1988	PID_470
5	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	tBil	16.2009410858154	MAFRAQ HOSPITAL	2	04/29/1988	PID_470
6	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	pCO2(T)	5.68579220628238	MAFRAQ HOSPITAL	2	04/29/1988	PID_470

Query executed successfully. | HDT (10.50 SP2) | sa (62) | master | 00:00:01 | 105257 rows

Figure 2: First raw data from SQL Server before any modification

Results Messages

	TimeAnalyzed	DisplayName	TestName	ValueString	HospitalName	Sex	DateOfBirth	PID
1	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	pCO2	4.30889755185233	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
2	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	tBil	30.0298709869385	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
3	2017-06-05 16:19:35.0000000 +04:00	MQ-SICU	pCO2(T)	4.30889755185233	MAFRAQ HOSPITAL	1	08/28/1958	PID_101
4	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	pCO2	5.68579220628238	MAFRAQ HOSPITAL	2	04/29/1988	PID_470
5	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	tBil	16.2009410858154	MAFRAQ HOSPITAL	2	04/29/1988	PID_470
6	2017-06-05 16:53:07.0000000 +04:00	MQ-ED1	pCO2(T)	5.68579220628238	MAFRAQ HOSPITAL	2	04/29/1988	PID_470
7	2017-06-06 06:55:42.0000000 +04:00	MQ-AQT90 ER_	TnT	0.01	MAFRAQ HOSPITAL	1	07/04/1984	PID_315

Query executed successfully. | HDT (10.50 SP2) | sa (66) | master | 00:00:00 | 45713 rows

Figure 3: Data after duplicate deletion

Results Messages

	Hospitalname	PID	Sex	DOB	pO2(T)	Na+	tHb	MetHb	tBil
1	MAFRAQ HOSPITAL	PID_10	1	04/14/1946	7.6974893654993	131.952838132177	11.3536410759193	0.3	-7.68266677856445
2	SKMC	PID_1000	1	08/30/1995	0	0	0	0	0
3	SKMC	PID_1003	1	07/07/1950	0	0	0	0	0
4	AHHY Hayer	PID_1005	1	02/08/1962	0	0	0	0	0
5	AHHY Hayer	PID_1006	1	03/01/2015	0	0	0	0	0
6	MAFRAQ HOSPITAL	PID_1007	1	12/14/1964	0	0	0	0	0

Query executed successfully. | HDT (10.50 SP2) | sa (72) | master | 00:00:00 | 1791 rows

Figure 4: Data after conversion rows to columns

Results Messages

	Sex	Age	pO2(T)	Na+	tHb	MetHb	tBil	HCO3-	COHb	pO2
1	1	71	7.6974893654993	131.952838132177	11.3536410759193	0.3	-7.68266677856445	25.3858779774964	0.6	57.73594268
2	1	59	2.37416234061617	152.642460753906	3.42947765496515	1.2	100.402694702148	24.715397908432	0.9	17.80768953
3	1	51	3.63515996842249	140.67748191581	6.21801896935899	0.5	6.57755041122437	27.7547184805879	0.9	27.26595356
4	2	65	25.1786061779041	143.298057666696	9.79317701081396	0.5	7.38613510131836	17.9490927496696	1.2	188.8551570
5	1	30	8.0396626357339	138.868493399364	8.84013124900055	0.5	46.4031982421875	28.6719230540398	-0.3	60.30245435
6	2	30	9.00604557932277	140.51235601408	8.47150096165953	0.8	13.6126899719238	24.681943460354	0.4	67.55092559
7	1	26	10.982544219195	132.127269087462	9.25228452798883	0.8	-2.3164699776062	22.0547215229581	0.7	82.38224152

Query executed successfully. | HDT (10.50 SP2) | sa (73) | master | 00:00:00 | 306 rows

Figure 5: Data after cleaning from unwanted feature and removing zero values and converted date to integer

Most of the researchers used Rapid Miner to reprocess data, convert data types, delete redundancy and more data cleansing function, but in this study, SQL Server played an important rule to prepare and make data ready, for me SQL was faster and more reliable tool saved time and efforts, that's why we will see fewer operators inside Rapid Miner, only main prediction operators were used.

3.1.4 Class Labels

Defining the class label model was based on pathologists from Dubai Hospital and Rashid Hospital both in Dubai-UAE, in Dubai Hospital I worked with Dr. Fayha Ahmad (Head of pathology and clinical chemistry laboratories) to identify the respiratory class label while in Rashid Hospital I worked with Dr. Salwa Nouri head of chemistry laboratories to identify the

Acute coronary syndrome. Dataset attributes were presented to both doctors and discussed all details.

For Respiratory, Dr. Fayha Ahmad listed the testing outcomes for pO₂ and pCO₂ as the key risk factors for diagnosing patients with respiratory issues type1 or respiratory issues type 2.

On that, the respiratory class label generation as below:

pO₂ > 60 mmHg → Normal

pO₂ < 60 mmHg AND pCO₂ <= 50mmHg → R1

pO₂ < 60 mmHg AND pCO₂ > 50 mmHg → R2

Respiratory Class labels:

1. Normal
2. Respiratory Type1: R1
3. Respiratory Type2: R2

For Cardiac dataset Acute coronary syndrome (ACS), Doctor Salwa explained the Troponin T characteristics which affect patients, as per Dr. Salwa Troponin T for normal patient should be always less or equal 0.01 ng/mL, in case of chest pain Troponin must be measured 3 times in 3 hours, if the result of Troponin T raised than 0.01 ng/mL and stay less than 1.0 u\dl this means that the patient will be exposed to coronary artery disease any time, but if the result exceeds 1.0 this means patient is already exposed the problem and immediately need to be treated.

Troponin T (TNT) <= 0.01 ng/mL → Normal

Troponin T (TNT) > 0.01 ng/mL and Troponin T (TNT) < 1.0 → High possibility for ACS

Troponin T (TNT) >= 1.0 ng/mL → Already ACS

Cardiac Class Label:

1. Normal
2. High Possibility - Acute Coronary Syndrome: HPACS
3. Already has Acute coronary syndrome: ACS

For the combined dataset, total of four classes where generated,

1. Normal: When pO₂ > 60 mmHg and Troponin T (TNT) <= 0.01 ng/mL
2. Respiratory (R): When ((pO₂ < 60 mmHg AND pCO₂ <= 50mmHg) OR (pO₂ < 60 mmHg AND pCO₂ > 50 mmHg)) AND Troponin T (TNT) <= 0.01.
3. R-ACS: When ((pO₂ < 60 mmHg AND pCO₂ <= 50mmHg) OR (pO₂ < 60 mmHg AND pCO₂ > 50 mmHg)) AND Troponin T (TNT) >= 1.00.

4. ACS: When Respiratory formula = Normal and Troponin T (TNT) \geq 1.00.

3.1.5 Duplicate Records

As mentioned in challenges, many patients were repeated, due to nature of blood gas and troponin as it needs monitoring, I have selected the first test made for each patient as this study trying to simulate the critical patients in emergency rooms.

3.1.6 Attributes Elimination and Missing Values

Attribute selection was applied based on domain knowledge of this dissertation consultants, we kept the main blood gases test and remove the calculated and non-effect tests such as (Anion_p, HbF, Hct, SBC, and cBaseECF) and (Hospital Name and Department Name)

Temperature attribute was removed as well, as all available patient has the same temperature, so it will not change the Prediction Results, Patient ID acting as an identity which will won't help in prediction then I removed it. There were two attributes with zero value for most of the patients, such as beta-hCG and p50(act) the deletion for that attributes were better than keeping them to ensure higher and more efficient prediction Results.

3.1.7 Final Datasets

In this part a summary tables for each dataset used in this dissertation Respiratory and Cardiac dataset.

3.1.8 Respiratory Dataset

Attribute	Description	Normal Value
Age	Age in years	[0-115]
Sex	Gender of patient	1 = male 2 = female
Na+	Sodium - mmol/l	[118.53-162.65]
HCO3-	Bicarbonate - mmol/l	[2.94-45.80]
tHb	Total Hemoglobin (%)	[2.16-15.57]
MetHb	Methemoglobin (%)	[-0.4-4]
tBil	Total bilirubin μ mol/L	[3.420-23.94]
COHb	Carboxyhemoglobin (%)	[-0.5-11.8]
pO2	Oxygen partial pressure - mmHg	[75 - 100]
pH	Acidity or alkalinity - pH unit	[7.004-7.746]
Lac	Lactate - mmol/l	[0.351-20.128]
tO2	Oxygen content - mmol/l	[1.442-14.031]
sO2	Oxygen saturation (%)	[12.1-100]
Glu	Glucose - mmol/l	[1.129-40.006]
Ca++	Calcium ion - mmol/l	[0.522-1.932]

Cl-	Chloride - mmol/l	[90.120-133.205]
K+	Potassium - mmol/l	[1.939-8.170]
pCO2	Carbon dioxide partial pressure - mmHg	[38-42]
HHb	Deoxyhemoglobin (%)	[0-87.3]
cBase(B)	Standard Base - mmol/l	[-25.984-14.708]
O2Hb	Oxyhemoglobin (%)	[11-97.9]
pO2(T)	Oxygen partial pressure T in mmHg	[75 - 100]
pCO2(T)	Carbon dioxide partial pressure T in mmHg	[38-42]
pH(T)	Acidity or alkalinity T in pH unit	[7.004-7.746]
(variable to be predicted)	Class of Respiratory types	Normal R1: Respiratory Type1 R2: Respiratory Type1

Table 2: Final Attributes for Respiratory Dataset

3.1.9 Cardiac Dataset

Attribute	Description	Normal Value
Age	Age in years	[0-115]
Sex	Gender of patient	1 = male 2 = female
PCT	Procalcitonin - ug/l	[0.10 - 85]
NTproBNP	BNP - pg/mL	[10-263]
Ddimer	ng/mL	[-0.4-4]
Troponin T	Troponin T - ng/mL	[<=0.01]
(variable to be predicted)	Class of Acute Coronary Syndrome	Normal HPACS ACS

Table 3: Final Attributes for Cardiac Dataset

3.2 Classification Algorithms

I applied four classification algorithms on each dataset and changed the datasets characteristics for some of them to meet the algorithms processing requirements, and kept changing the replace missing values operator criteria, as it is medical data I tried to replace the missing values by minimum, maximum and average. Tried discretizing age, used bagging, boosting operators and kept changing the algorithm properties to hit the maximum accuracy and prediction performance. I only mentioned the best trial results and ignored the lower results to reduce text.

The used algorithms as below:

1. Decision Tree
2. Gradient Boosted Trees
3. Neural Network
4. Naïve Bias

To ensure the ease use of Rapid Miner, I put all required operators and started to switch them on-off, this saved a lot of time, as during the dissertation preparation I had a difficult time adding and removing operators to the rapid miner and keep versions of the processes and sometimes overwrite some processes and lose the work. Below figures show Rapid miner interfaces. For every test, I used different criteria to ensure the highest performance and accuracy, for instance, I used decision Tree with discretizing age to random age ranges, I used replace missing values as well and tested each operator three times, tried with replacing with minimum, maximum and average, I tried to hide the noisy attributes which affected the accuracy in some tests, as well I used with bagging and boosting for every algorithm and without, also used tried to use the options available in each operator seeking the best accuracy and performance as will be listed down in each algorithm analysis.

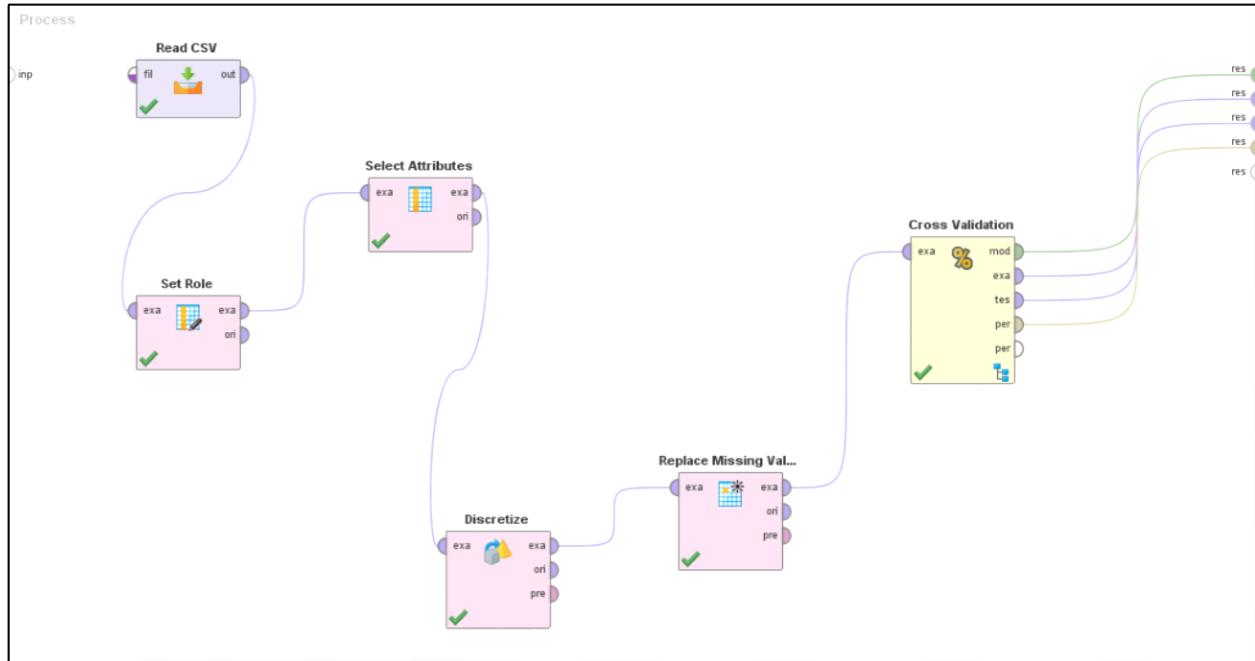


Figure 6: Rapid Miner Main process with Discretize and Replace Missing value operator

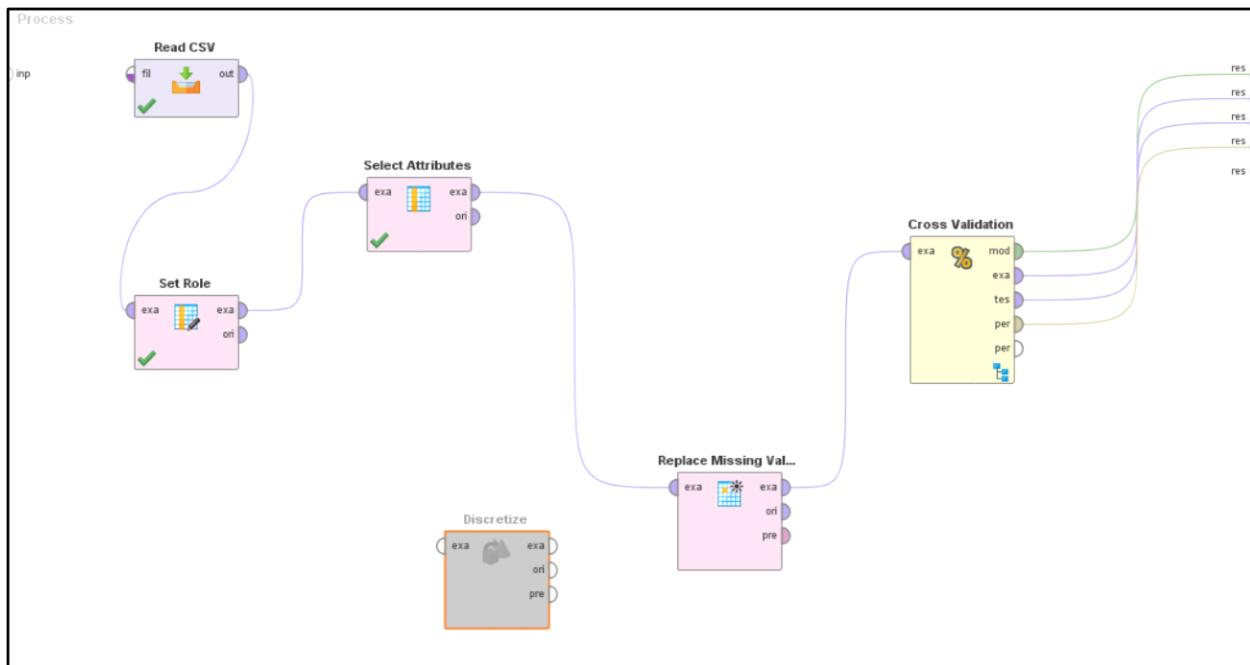


Figure 7: Rapid Miner Enabling and Disabling operators

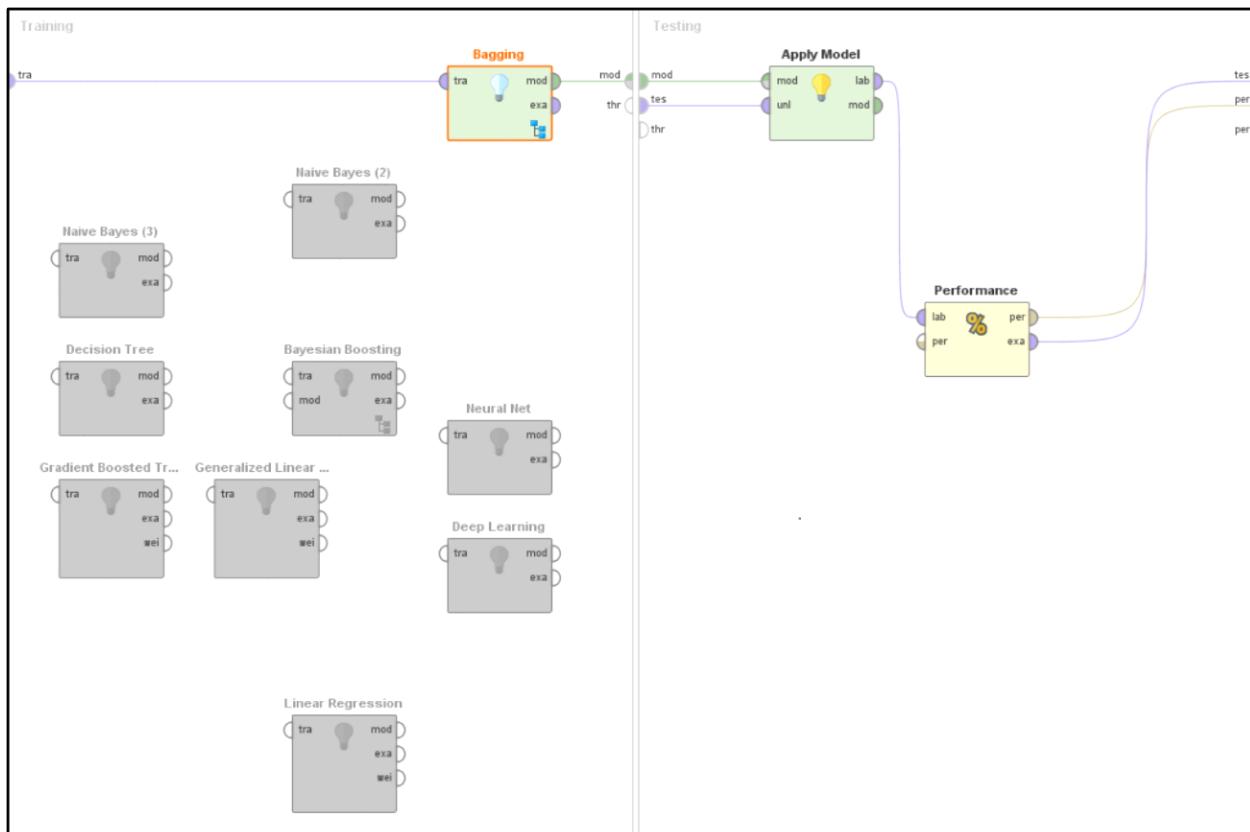


Figure 8: Rapid Miner Process Cross validation

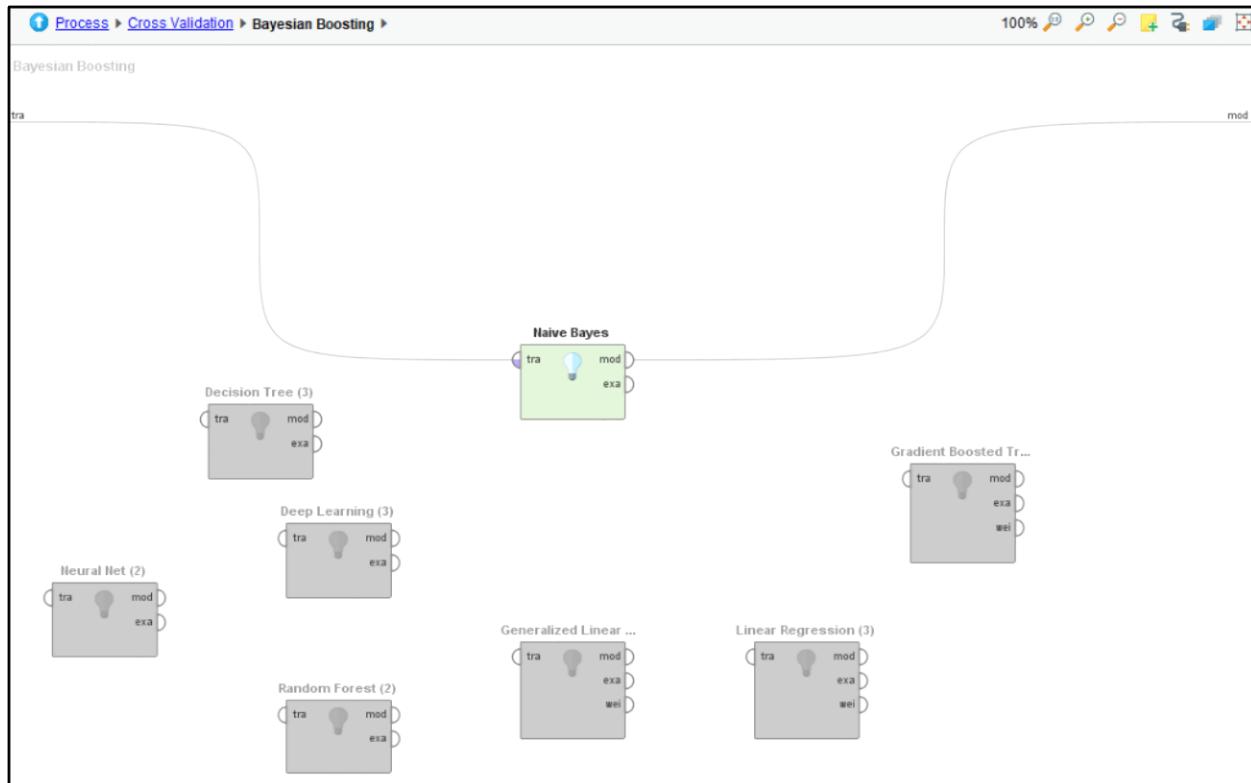


Figure 9: Rapid Miner Bayesian Boosting operator

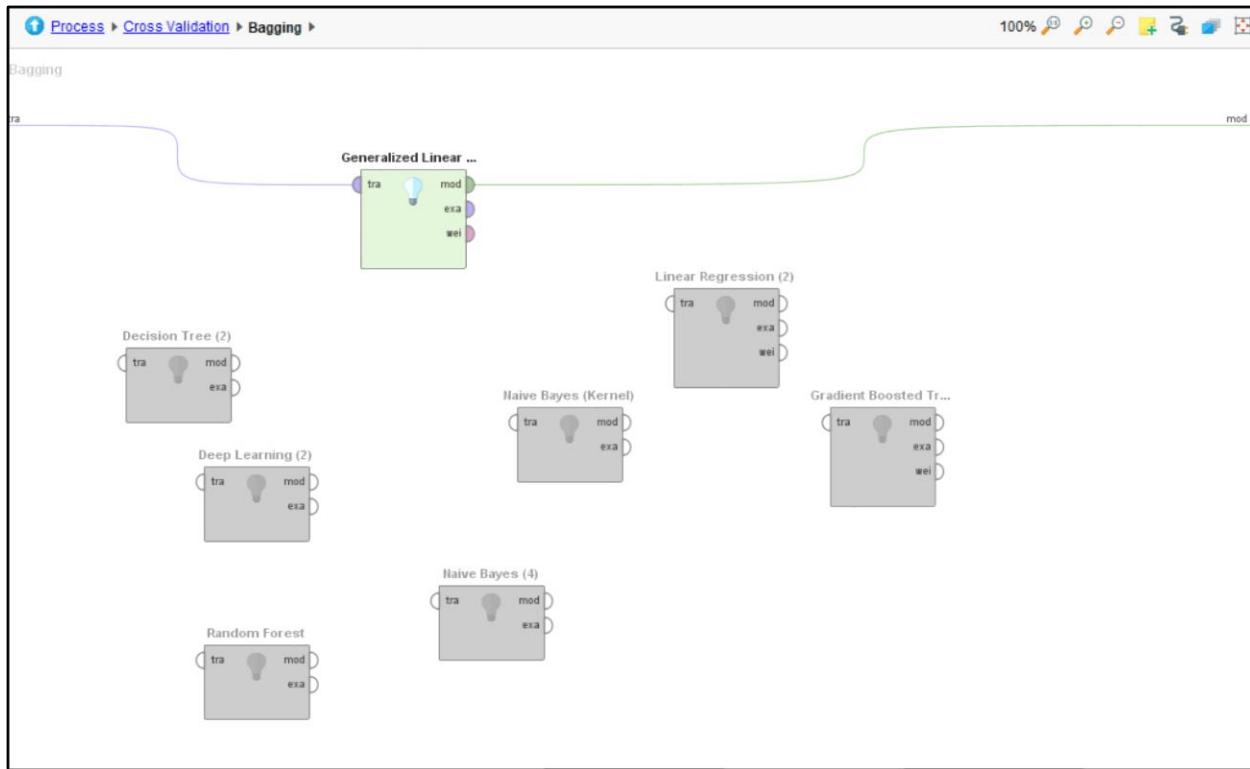


Figure 10: Rapid Miner Bagging operator

Before analysis, below are the actual class labels count for each dataset of 306 patients:

Respiratory Classes	R1	R2	Normal
Count	186	32	88

Table4: Respiratory Dataset True Classes

Cardiac Classes	HPACS	ACS	Normal
Count	71	7	228

Table5: Cardiac Dataset True Classes

Respiratory & Cardiac Classes	R	ACS	R-ACS	Normal
Count	159	19	59	69

Table6: Combined Dataset True Classes

3.3 Decision Tree

Decision tree classification gives a fast and viable system for classifying datasets (Aitkenhead, 2008). it chooses which attribute has more importance to decide, Furthermore, it might be acknowledged those root of the tree, dependent upon those root value, the other tree component will be ordered and sorted in such discernable approach. Decision trees are quick and simple to use, it gives a straightforward structure of the classifier.

3.3.1 Prediction Results for Decision Tree

Tried decision tree in four creation types, gain ratio, information gain, Gini index, and accuracy, the results differ for each creation type, not forgetting that I tried with each creation type to change the replace missing value by (minimum, maximum and average) the highest accuracy was when using Accuracy creation type with replacing missing value by maximum. The accuracy of the decision tree model for each dataset listed below in Table7.

Decision Tree Creation Type: Accuracy			
Respiratory Dataset Accuracy			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	88.22%	88.55%	88.22%
Cardiac Dataset			
Replace missing values by	Not Used		
Results	96.73%		
The Combined Datasets			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	91.52%	91.85%	91.52%

Table 7: Decision Tree Accuracy – All Datasets

3.3.2 Decision Tree Performance Analysis

In this part, deep analysis of the confusion matrix of decision tree highest accuracy trial will be listed, this will illustrate the class label accuracy and class precision for each dataset separately.

3.3.2.1 Respiratory Dataset (DT)

Table 8 lists decision tree confusion matrix for respiratory dataset.

	true R1: 186	True R2: 32	true Normal: 88	class precision
pred. R1	172	14	7	89.12%
pred. R2	2	18	0	90.00%
pred. Normal	12	0	81	87.10%
class recall	92.47%	56.25%	92.05%	

Table 8: Decision Tree confusion matrix- Respiratory Dataset

For the class R1: decision tree succeeded to predict 172 out 186 which means excellent prediction performance, with a low error rate as 14 patients were predicted as respiratory type 2 (R2), this indicated the ability of the decision tree to predict R1 class.

R2 Class: was not that good in predicting the R2 class as 14 out of 32 patients were wrongly predicted, the tree was able to predict 18 out of 32 patients correctly 56.25% and this is poor performance.

For NORMAL class: 81 patients out of 88 were predicted correctly with 92.05% class prediction accuracy and this is excellent comparing to R2 class.

Decision tree prediction summary:

1. High prediction performance for class R1
2. Low prediction performance for class R2
3. High prediction performance for class NORMAL
4. Overall very good performance as there are 7 out 88 False-Positive results. (Sick but predicted as Normal)

3.3.2.2 Cardiac Dataset (DT)

Confusion matrix analysis for cardiac dataset listed in table 9,

	true Normal/ 288	true HPACS/ 71	true ACS/ 7	class precision
pred. Normal	228	3	0	98.70%
pred. HPACS	0	67	6	91.78%
pred. ACS	0	1	1	50.00%
class recall	100.00%	94.37%	14.29%	

Table 9: Decision Tree confusion matrix – Cardiac Dataset

Class NORMAL: Starting with the good news, class NORMAL shows 100% prediction accuracy, decision tree didn't predict any patient as ACS, 228 patients out of 228 were predicted correctly as Normal.

Class HPACS: Very good performance prediction for class HPACS (High possibility of Acute Coronary Syndrome) as 67 patients out of 71 patients was predicted correctly with the percentage of 94.34%.

Class ACS: The problem which appeared was predicting only one true ACS out of 7 patients and the rest 6 patients were predicted as HPACS, the prediction accuracy for this class was the lowest at 14.29% and this is due to the sensitivity of Troponin T which decided ACS class.

Decision tree prediction summary for the cardiac dataset:

1. High prediction **error** rate for ACS – 14.29%
2. Excellent prediction for class NORMAL – 100%
3. Very good prediction for class HPACS – 94.37%
4. ACS class prediction was difficult due to the narrow result value in the dataset. The predictor was able to decide if the patient has a high possibility of ACS but was not able to say it is true ACS.

3.3.2.3 The Combined Datasets (DT)

Confusion matrix analysis for the combined datasets listed in table 10.

	true R/159	true R-ACS/59	true Normal/69	true ACS/19	class precision
pred. R	146	3	3	0	96.05%
pred. R-ACS	4	54	0	2	90.00%
pred. Normal	9	0	65	1	86.67%
pred. ACS	0	2	1	16	84.21%
class recall	91.82%	91.53%	94.20%	84.21%	

Table 10: Decision Tree confusion matrix – The combined Dataset

Class R: decision tree succeeded to predict 146 out 159 which is a good rate of 91.82%, only 13 patients were predicted wrongly.

Class R-ACS: prediction was excellent, 54 out of 59 patients were correct and almost the rest were predicted as R class only, but patients are truly having ACS and Respiratory problems.

Class NORMAL: were excellent, with success rate of 92.75%, as 65 out of 69 patients were predicted correctly but the problem was in predicting one patient as ACS while the patient was Normal, and this is the false negative prediction.

Class ACS: shows very good prediction performance, it predicted 16 patients out of the 19.

Decision tree summary:

1. High prediction performance rate for ACS the class performance was 84.21% only.
2. High prediction performance rate for R-ACS class with the performance of 91.53% only.
3. Overall very good prediction performance for ACS and R-ACS.

3.3.3 Decision Tree Features Importance Analysis

In this part, the most important attributes which affected the analysis will be listed and discussed for each dataset used.

3.3.3.1 Respiratory dataset (DT)

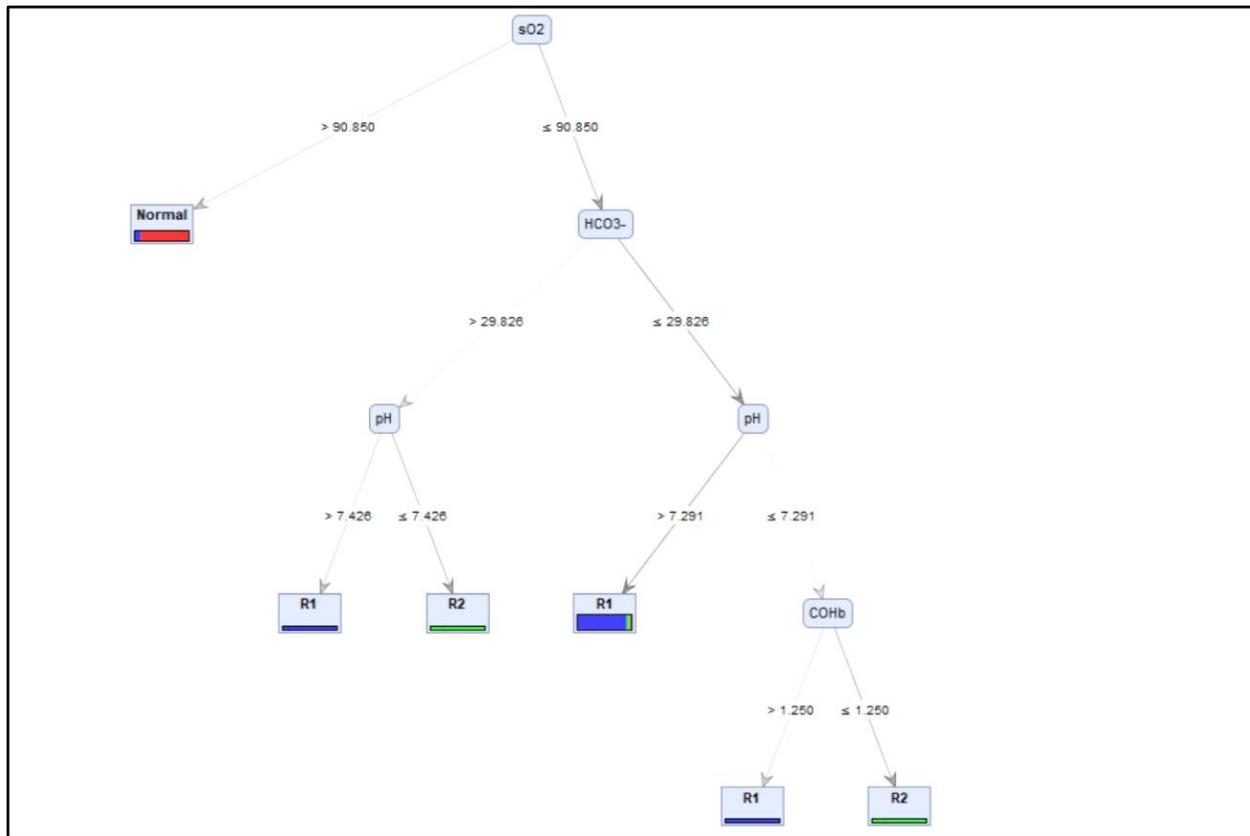


Figure 11: Decision Tree output- Respiratory dataset

The most effective features as per decision tree were sO2 played an important role to predict the class label as the below criteria:

```

sO2 > 90.850: Normal {R1=12, R2=0, Normal=85}
sO2 ≤ 90.850
|   HCO3- > 29.826
|   |   pH > 7.426: R1 {R1=7, R2=0, Normal=0}
|   |   pH ≤ 7.426: R2 {R1=0, R2=13, Normal=0}
|   HCO3- ≤ 29.826
|   |   pH > 7.291: R1 {R1=164, R2=11, Normal=3}
|   |   pH ≤ 7.291
|   |   |   COHb > 1.250: R1 {R1=3, R2=0, Normal=0}
|   |   |   COHb ≤ 1.250: R2 {R1=0, R2=8, Normal=0}
  
```

[sO2], [HCO3-], [pH] and [COHb] are the important features with relation to respiratory as below:

If sO2 was >90.850 then patient is Normal else check [HCO3-]. If [HCO3-] > 29.826 then check [pH] as if [pH] > 7.426 then patient has Respiratory type1 else then patient has respiratory type 2. But if [HCO3-] ≤ 29.826 then check [pH] as if [pH] > 7.291 patient has a respiratory type 1 else if [pH] ≤ 7.291 then [COHb] need to be checked, as if [COHb] > 1.250 then patient has respiratory type 1 else then patient will be respiratory type2.

3.3.3.2 Cardiac Dataset (DT)

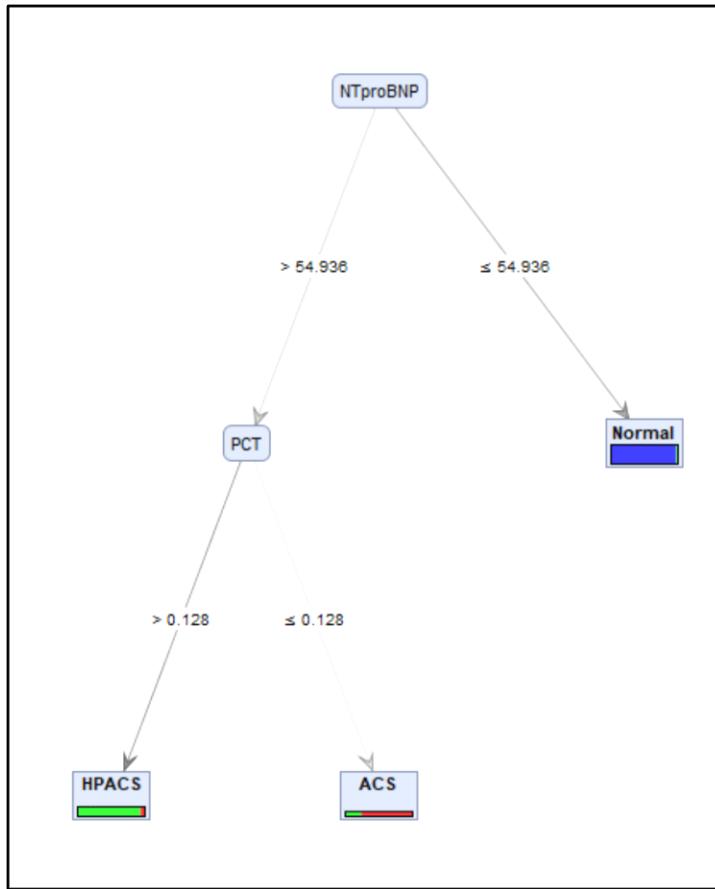


Figure 12: Decision Tree output

The most effective features as per decision tree were [NTproBNP] and [PCT] played the main role to predict the cardiac class label as the below criteria:

```
NTproBNP > 54.936
| PCT > 0.128: HPACS {Normal=0, HPACS=67, ACS=4}
| PCT ≤ 0.128: ACS {Normal=0, HPACS=1, ACS=3}
NTproBNP ≤ 54.936: Normal {Normal=228, HPACS=3, ACS=0}
```

[NTproBNP] relation to cardiac (Acute coronary syndrome):

If [NTproBNP] was > 54.936 then check [PCT], If [PCT] > 0.128 then patient chest pain hi possibility to be caused by coronary syndrome, else if [PCT] ≤ 0.128 then patient probably has already ACS (Acute coronary syndrome) and needs an immediate action. But if [NTproBNP] ≤ 54.936 then patient for sure has no ACS and very less chance of future ACS problems.

3.3.3.3 The Combined Datasets (DT)

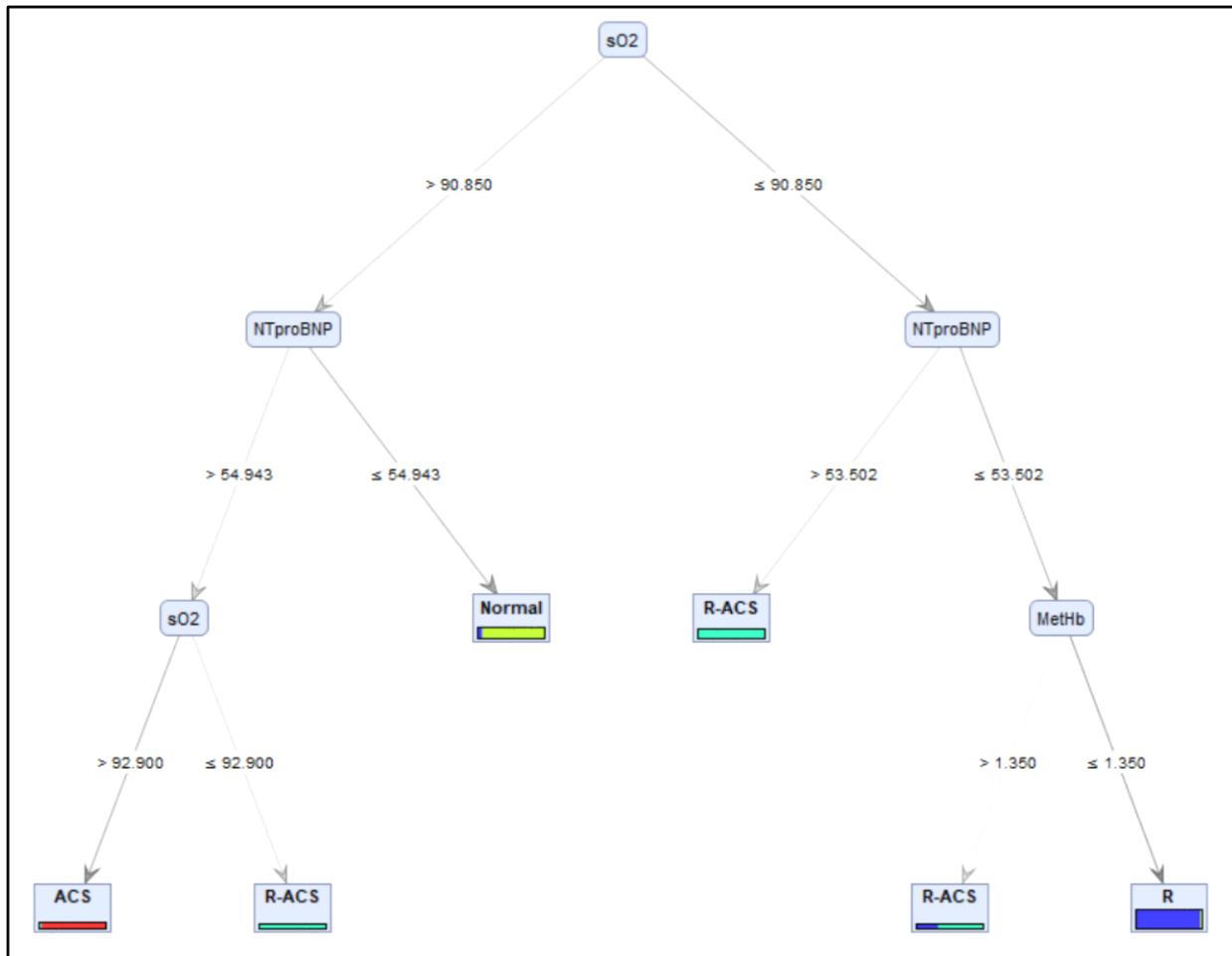


Figure 13: Decision Tree output – The Combined Datasets

The most effective features as per decision tree were [sO2], [NTproBNP] and [MetHb] played an important role to predict the class label as the below criteria:

- sO2 > 90.850
 - | NTproBNP > 54.943
 - | | sO2 > 92.900: ACS {R=0, R-ACS=1, Normal=0, ACS=19}
 - | | sO2 ≤ 92.900: R-ACS {R=0, R-ACS=5, Normal=0, ACS=0}
 - | NTproBNP ≤ 54.943: Normal {R=6, R-ACS=0, Normal=66, ACS=0}
- sO2 ≤ 90.850
 - | NTproBNP > 53.502: R-ACS {R=1, R-ACS=51, Normal=0, ACS=0}
 - | NTproBNP ≤ 53.502
 - | | MetHb > 1.350: R-ACS {R=1, R-ACS=2, Normal=0, ACS=0}
 - | | MetHb ≤ 1.350: R {R=151, R-ACS=0, Normal=3, ACS=0}

Feature relations to each class label:

If sO₂ was >90.850 then check [NTproBNP], If [NTproBNP] > 54.943 then check [sO₂] again if [sO₂] >92.900 then patient chest pain caused by ACS, else if [NTproBNP] <54.934 then most probably patient is (NORMAL).

But if [sO₂] <= 90.850 then see if [NTproBNP] > 53.502 if yes then most probably patient has both a respiratory and coronary problems (R-ACS). But if [NTproBNP] <= 53.502 then check [MetHb], is if [MetHb] <= 1.350 then patient for sure has only respiratory problems (R).

3.4 Gradient Boosted Tree

(Lawrence et al, 2004) defined gradient boosted trees as the processing of standard classification tree analysis that endeavors to limit these confinements by utilizing arrangement blunders to iteratively refine the trees utilizing an irregular example of the preparation information and consolidating the different trees iteratively created to characterize the information.

3.4.1 Gradient Boosted Tree Prediction Results

Gradient boosted tree was tested, using different trees count, the replace missing value method also used by (minimum, maximum and average) the highest accuracy was when using 20 trees with replacing missing value by Average. The accuracy of the gradient boosted tree algorithm for each dataset listed below in Table11. For cardiac dataset, as there are no missing values, so replace missing values operator was not used, the highest accuracy trial was 96.73% using 20 trees. But for the combined dataset, used replace missing value operator with average and the highest trial value was 93.16%.

Gradient Boosted Tree Accuracy: 20 Trees			
Respiratory Dataset Accuracy			
Replace missing values by	<i>Minimum</i>	<i>Maximum</i>	<i>Average</i>
Results	88.24%	87.23%	88.87%
Cardiac Dataset			
Replace missing values by	Not Used		
Results	96.73%		
The Combined Datasets			
Replace missing values by	<i>Minimum</i>	<i>Maximum</i>	<i>Average</i>
Results	92.83%	92.51%	93.16%

Table 11: Gradient Boosted Tree Accuracy –All Datasets

3.4.2 Gradient Boosted Tree Performance analysis

In this part, deep analysis of the confusion matrix of the gradient boosted tree highest accuracy trial will be listed, this will illustrate the class label accuracy and class precision for each dataset separately.

3.4.2.1 Respiratory Dataset (GBT)

Gradient boosted tree confusion matrix analysis listed in table x,

	true R1/186	true R2/32	true Normal/88	class precision
pred. R1	175	14	9	88.38%
pred. R2	0	18	0	100.00%
pred. Normal	11	0	79	87.78%
class recall	94.09%	56.25%	89.77%	

Table 12: Gradient boosted tree confusion matrix – Respiratory Dataset

For the class R1: boosted tree succeeded to predict 175 out 186 which means excellent prediction performance, with a low error rate as 11 patients were predicted as (NORMAL), this indicated the ability of the boosted tree to predict R1 class.

R2 Class: 14 wrong predictions occur out of 32 as R1, while it was R2 the class was performance is lower than R1 class as it reached 56.25%.

For NORMAL class: 79 patients out of 88 were predicted correctly with 89.77% class prediction accuracy and this is excellent comparing to R2 class.

Decision tree prediction summary:

1. High prediction performance for class R1
2. Low prediction performance for class R2
3. High prediction performance for class NORMAL
4. Total of 11 patients were R1 and predicted as NORMAL.

3.4.2.2 Cardiac Dataset (GBT)

Gradient boosted tree confusion matrix analysis listed in table x,

	true Normal/ <u>228</u>	true HPACS/ <u>71</u>	true ACS/ <u>7</u>	class precision
pred. Normal	228	3	0	98.70%
pred. HPACS	0	68	7	90.67%
pred. ACS	0	0	0	0.00%
class recall	100.00%	95.77%	0.00%	

Table 13: Gradient boosted tree confusion matrix – Cardiac dataset

Class NORMAL: Super Excellent performance for the class NORMAL, it shows 100% prediction accuracy, gradient boosted tree didn't predict any patient as ACS, 228 patients out of 228 were predicted correctly as Normal.

Class HPACS: Very good performance prediction for class HPACS (High possibility of Acute Coronary Syndrome) as 68 patients out of 71 patients was predicted correctly with a percentage of 95.77%.

Class ACS: 0 percent correct prediction for this class and as mentioned in decision tree this is due to the sensitivity of Troponin T which formulate the ACS class.

Gradient boosted tree prediction summary for the cardiac dataset:

1. High prediction **error** rate for ACS – 100%
2. Excellent prediction for class NORMAL – 100%
3. Very good prediction for class HPACS – 95.77%
4. ACS class prediction was difficult due to the narrow result value in the dataset. The predictor was able to decide if the patient has a high possibility of ACS but was not able to say it is true ACS.

3.4.2.3 The Combined Datasets (GBT)

Gradient boosted tree confusion matrix analysis listed in table x,

	true R/159	true R-ACS/59	true Normal/69	true ACS/19	class precision
pred. R	150	3	4	0	95.54%
pred. R-ACS	1	55	0	2	94.83%
pred. Normal	8	0	64	1	87.67%
pred. ACS	0	1	1	16	88.89%
class recall	94.34%	93.22%	92.75%	84.21%	

Table 14: Gradient boosted tree confusion matrix- The combined dataset

Class R: Gradient boosted tree succeeded to predict 150 out 159 which is a good rate of 94.34%, only 9 patients were predicted wrongly.

Class R-ACS: prediction was excellent, 55 out of 59 patients were correct and almost the rest were predicted as R class, and one patient was predicted as ACS class.

Class NORMAL: were excellent, with success rate of 91.30%, as 64 out of 69 patients were predicted correctly but the problem was in predicting one patient as ACS while the patient was Normal, and this is the false negative prediction.

Class ACS: shows very good prediction performance, it predicted 16 patients out of 19 as ACS, and two patients were predicted as R-ACS and one was predicted as normal.

Gradient boosted tree performance summary:

1. Low prediction error rate for ACS the class performance was 84.21% success.
2. Low prediction error rate for R-ACS class with performance of 93.22% success.
3. Overall great prediction performance, ACS and R-ACS.

3.4.3 Gradient Boosted Tree Features Importance Analysis

In this part, the most important attributes which affected the analysis will be listed and discussed for each dataset used.

3.4.3.1 Respiratory Dataset (GBT)

In gradient boosted trees, I selected to generate 20 trees pre-class so total 60 trees were generated, it was noticeable that all H1 (ACS) class related to K+ and pH, K+ appeared as a root for 10 trees, pH appeared as a root for 5 trees and the rest was distributed between Ca++ and MetHb, while in R1 (Respiratory) class O2Hb and sO2 appeared, O2Hb appeared 19 times as a tree root while sO2 appeared one time only.

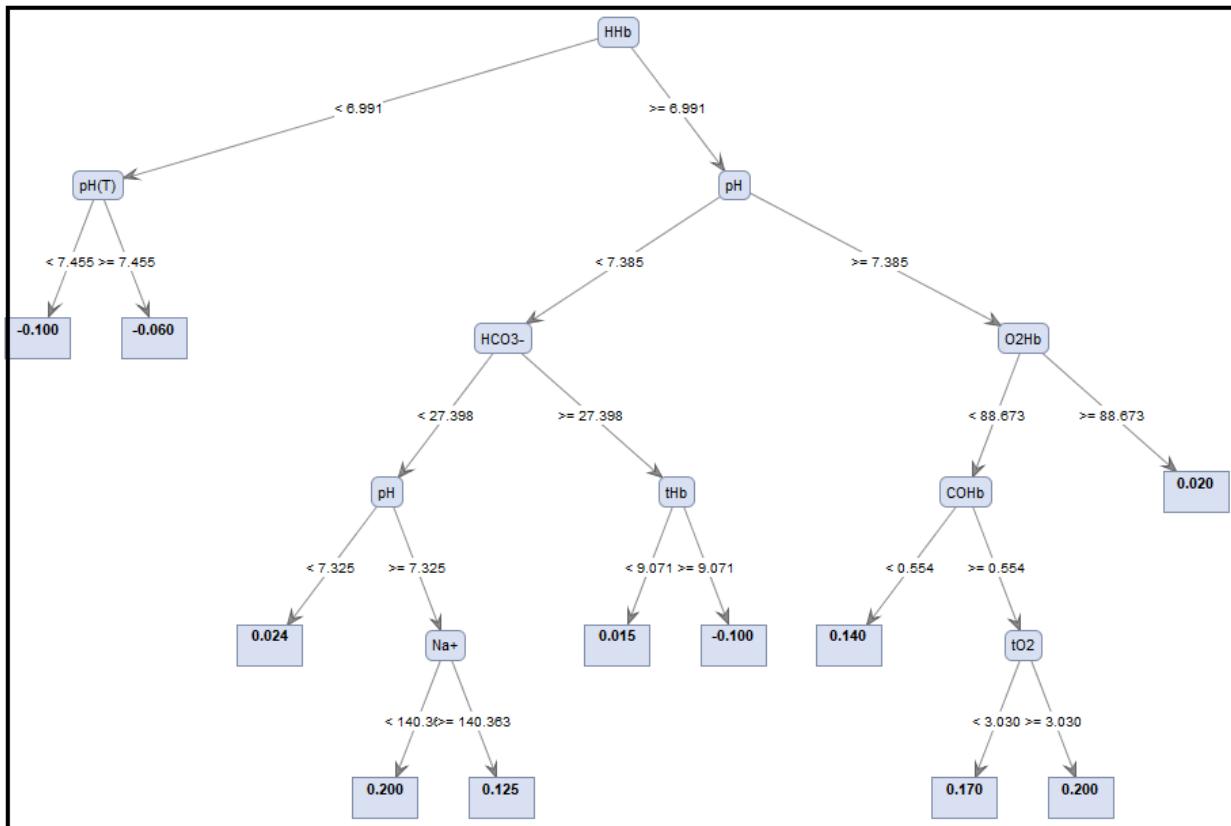


Figure 14: R1: Respiratory Class R1 for gradient boosted tree - HHb

HHb was the root for R1 Class tree, the dataset features importance was as below from left to right and root has the highest importance:

Root channel 1: HHb → pH → O2Hb → COHb → tO2
 Root channel 2: HHb → pH → HCO3- → tHb → Na+

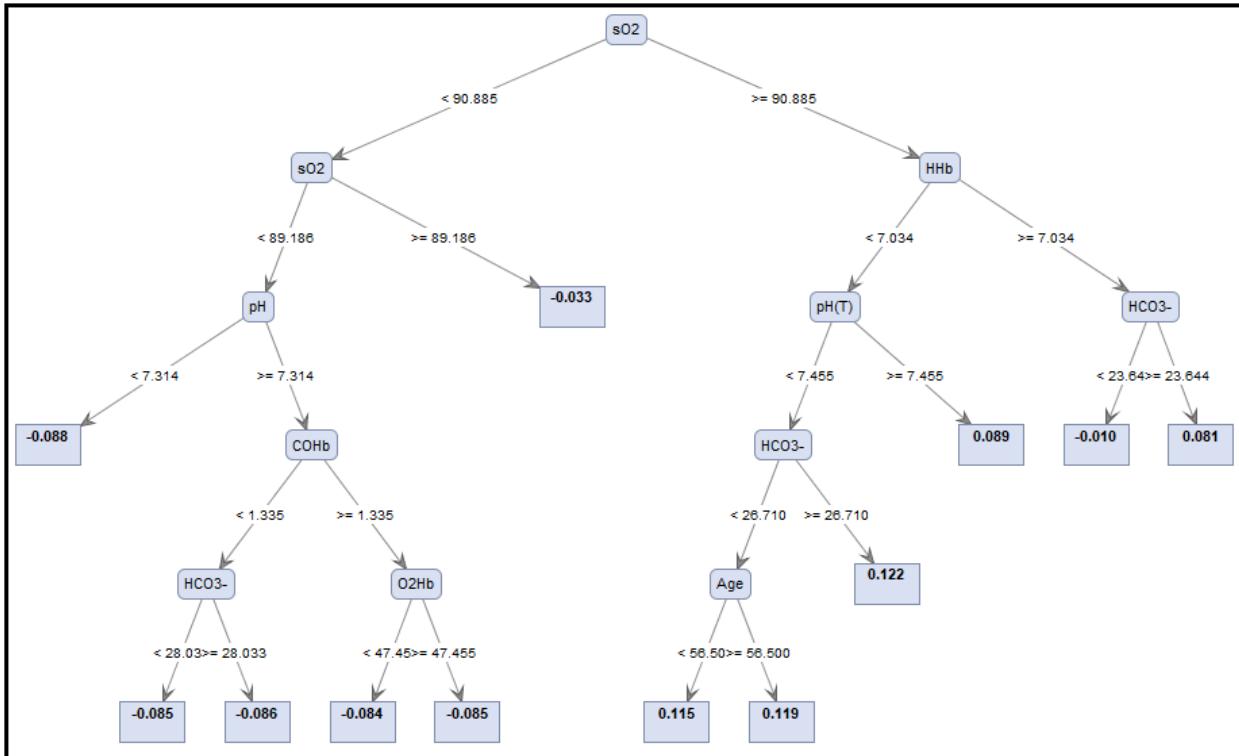


Figure 15: R1: Respiratory class for gradient boosted tree Oxygen saturation– sO2

sO2 as well, was found as a root in many trees of the R1 class, and here is the tree features importance from left to right.

Root channel 1: sO2 → HHb → pH(T) and HCO3- → Age

Root channel 2: sO2 → pH → CHOb- → O2Hb

While for class R2 pH and pH(T) were the roots as shown in figure 16.

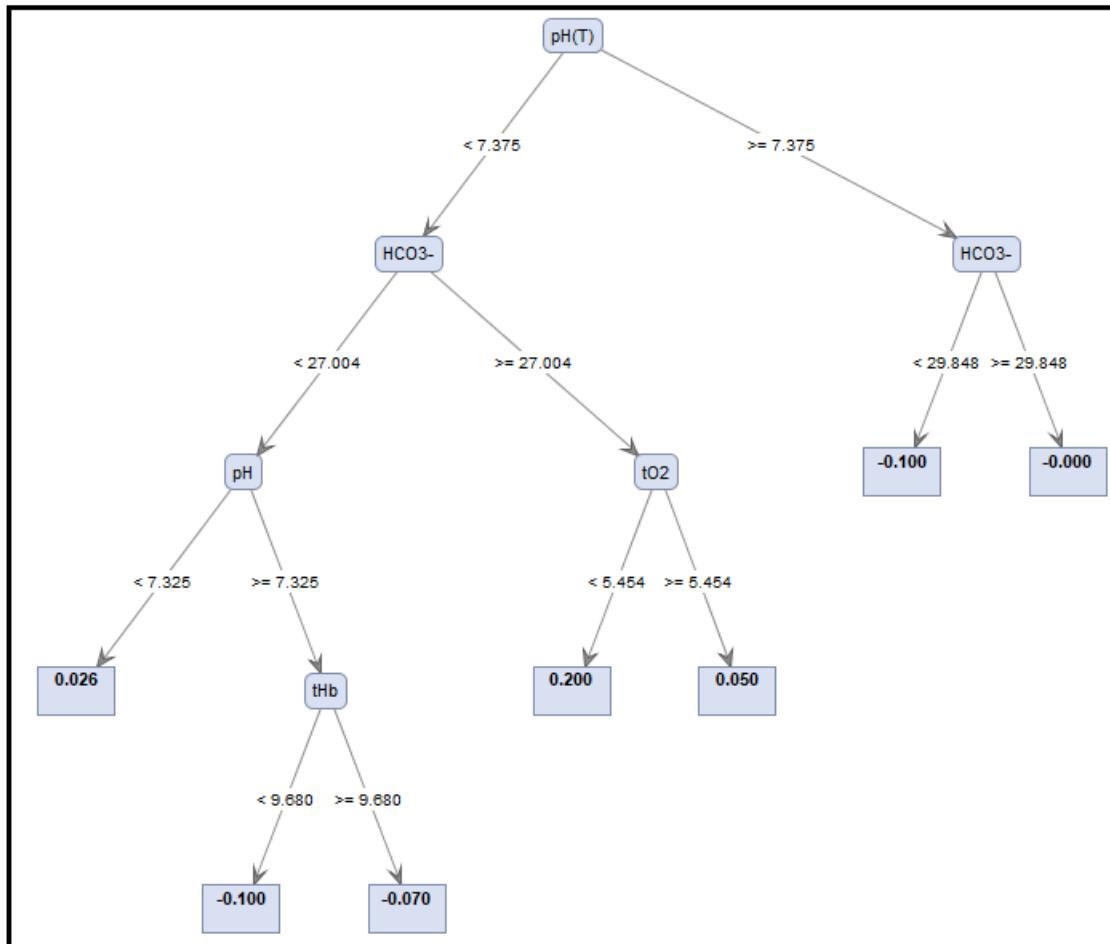


Figure 16: R2: Respiratory class for gradient boosted tree pH

pH & pH(T) were the root in most of the R2 Class trees, the dataset features importance was as below from left to right and root has the highest importance:

Root channel 1: pH(T) → pH → HCO3- → tO2 → tHb

3.4.3.2 Cardiac Dataset (GBT)

Out of 20 tables for all class labels, NT-proPNB appeared 18 times, PCT appeared two times as a root of the classes tree as shown in figure 17.

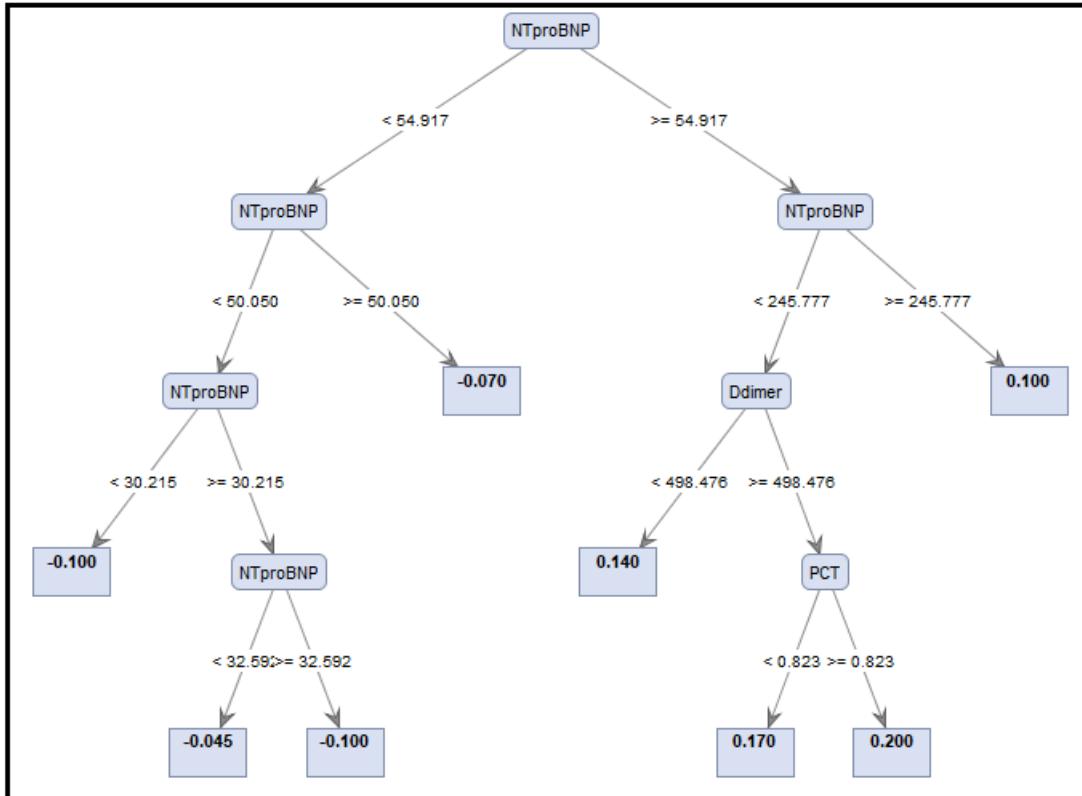


Figure 17: Cardiac Dataset for gradient boosted tree – NT-proBNP

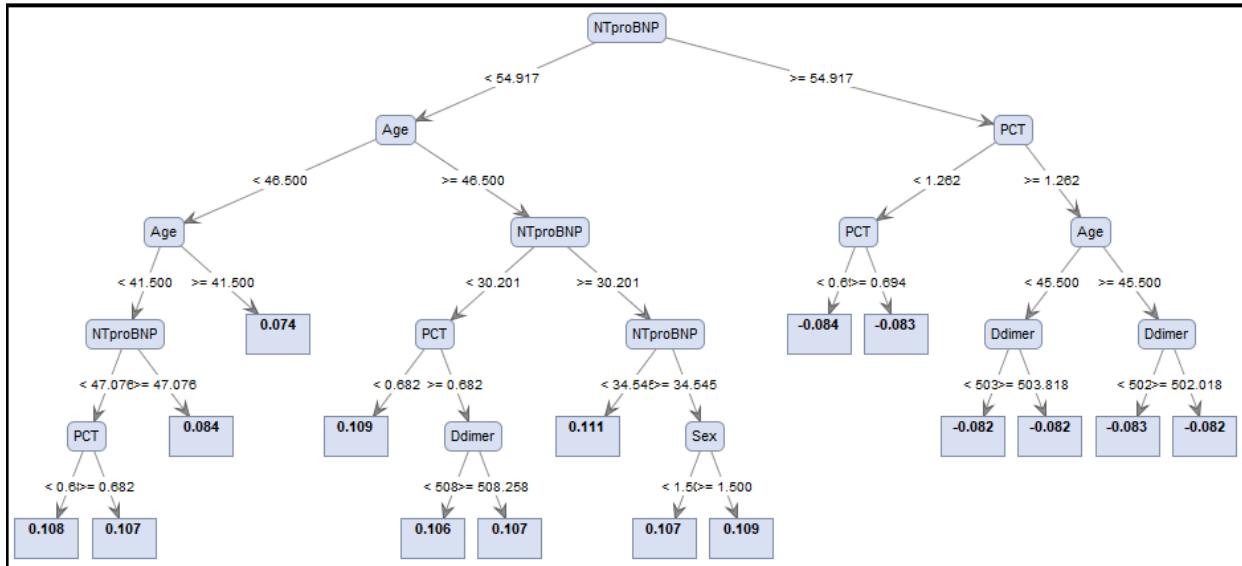


Figure 18: Cardiac Dataset for gradient boosted tree – NT-proBNP, PCT and Age

NT-ProBNP was the root for most of the classes it is very strong and important feature to predict cardiac, and this is the reason of the high accuracy for the algorithm. the dataset features importance was as below from left to right and root has the highest importance:

Root channel 1: NT-ProBNP → Age → PCT → D-Dimer

3.4.3.3 The Combined Datasets (GBT)

For the combined dataset, I was worried to see other than the early found respiratory important features + the Cardiac features as the roots of the tree as this indicates that there is something wrong in the analysis, sO₂, NT-ProBNP, and HHb only were the roots, and they were the most important features as shown in figure 19 and 20.

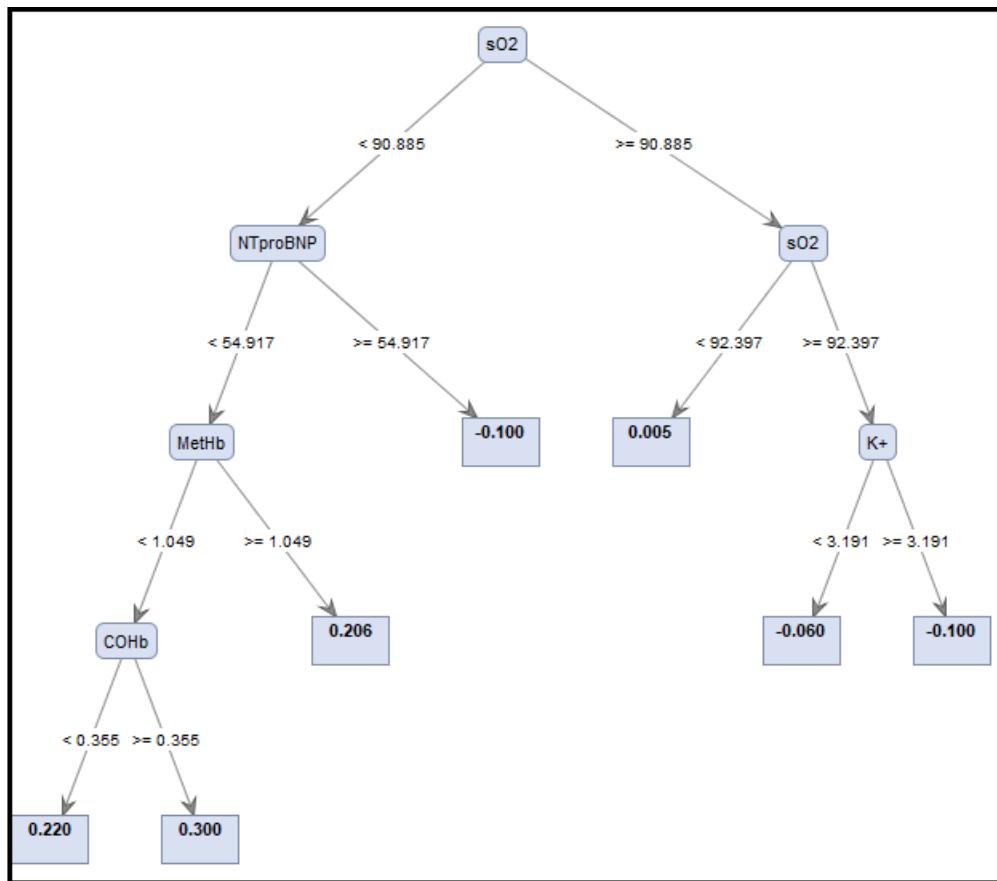


Figure 19: The Combined Datasets R1 Class for gradient boosted tree – sO₂

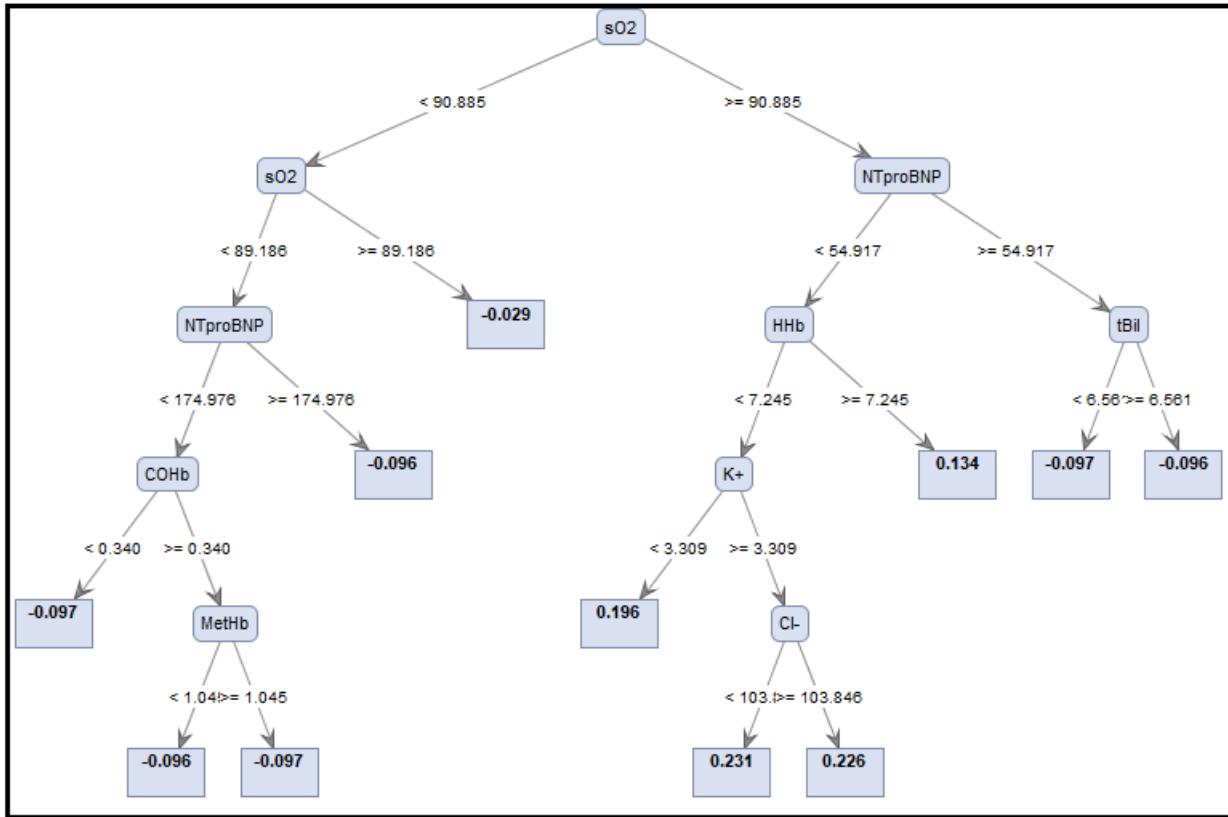


Figure 20: The Combined Datasets HPACS class for gradient boosted tree – NT-proBNP

3.5 Neural Network (NN)

In this part, I have used Neural Network to evaluate the dataset accuracy and performance.

An artificial neural network (ANN), is a scientific show or computational demonstrate based on the biological neural network. In other words, it is an emulation of the biological neural system

It maps a group of data input onto a group of appropriate output data. It consists of three layers input layer, hidden layer & output layer. A connection between layers and weights are assigned to each connection. The main function of neurons of the input layer is to split input x_i into neurons in the hidden layer. The neuron of hidden layer adds input signal x_i with weights w_{ji} of respective connections from input layer (Kaur & Singh, 2014).

3.5.1 Neural Network Prediction Results

For Neural Network algorithm, tried many options, such as replacing missing values, boosting, bagging and changed the training cycle and the number of hidden layers, the best trial was when used 700 training cycle with replacing missing value method by maximum to reach 87.25%. but

the process time was the problem as it costs up to 1.41 minutes for respiratory and the combined datasets. For cardiac dataset using the same customization the accuracy reached to 96.06%, and for the combined dataset, I got 87.88% using replace missing value by maximum.

Neural Network Accuracy: 700 Training Cycle via Bagging			
Respiratory Dataset Accuracy			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	83.65%	87.25%	85.30%
Cardiac Dataset (Without bagging)			
Replace missing values by	Not Used		
Results	96.06%		
The Combined Datasets			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	87.26%	87.88%	87.55%

Table 15: Neural Network Result Accuracy- All Datasets

3.5.2 Neural Network Performance analysis

In this part, analysis of the confusion matrix of the Neural Network highest accuracy trial will be listed, this will illustrate the class label accuracy and class precision for each dataset separately.

3.5.2.1 Respiratory Dataset (NN)

Neural Network confusion matrix analysis listed in table 16,

	true R1/ 186	true R2/ 32	true Normal/ 88	class precision
pred. R1	176	18	11	85.85%
pred. R2	1	14	0	93.33%
pred. Normal	9	0	77	89.53%
class recall	94.62%	43.75%	87.50%	

Table 16: Neural Network confusion matrix – Respiratory Dataset

For the class R1: Neural Network succeeded to predict 176 out 186 which means excellent prediction performance, with a low error rate as 11 patients were predicted as (NORMAL).

R2 Class: bad performance in predicting the R2 class as 18 out of 32 patients were wrongly predicted, and this is worse than decision tree and gradient boosted tree, low performance reached to 43.75%.

For NORMAL class: 77 patients out of 88 were predicted correctly with 87.50% class prediction accuracy and this also less than the previous algorithms so far.

Neural Network prediction summary:

1. High prediction performance for class R1
2. Low prediction performance for class R2
3. Good prediction performance for class NORMAL

3.5.2.2 Cardiac Dataset (NN)

Neural Network confusion matrix analysis listed in table x,

	true Normal: 228	true HPACS: 71	true ACS: 7	class precision
pred. Normal	227	4	0	99.27%
pred. HPACS	1	67	7	89.33%
pred. ACS	0	0	0	0.00%
class recall	99.56%	94.37%	0.00%	

Table 17: Neural Network confusion matrix - Cardiac Dataset

Class NORMAL: The Neural network was excellent but less than trees for this class, as it achieved 99.56% accuracy, Neural Network predicted only one patient wrongly out 228 NORMAL patients.

Class HPACS: Very good performance prediction for class HPACS (High possibility of Acute Coronary Syndrome) as 67 patients out of 71 patients was predicted correctly with a percentage of 94.37%.

Class ACS: Superbad performance in predicting ACS, 0% was correct, and the reason as mentioned before is related the narrow difference between the test results values.

Neural Network prediction summary for the cardiac dataset:

1. Not able to predict ACS class with 0% success.
2. Very good prediction rate for class NORMAL – 99.56%
3. Very good prediction for class HPACS – 94.37%
4. ACS class prediction was difficult, and this will be discussed at the dissertation recommendation.

3.5.2.3 The Combined Datasets (NN)

Neural Network confusion matrix analysis listed in table 18,

	true R/ 159	true R-ACS/ 59	true Normal/ 69	true ACS/ 19	class precision
pred. R	147	8	9	0	89.63%
pred. R-ACS	3	48	1	2	88.89%
pred. Normal	9	1	59	2	83.10%
pred. ACS	0	2	0	15	88.24%
class recall	92.45%	81.36%	85.51%	78.95%	

Table 18: Neural Network confusion matrix – The Combined Datasets

Class R: decision tree succeeded to predict 147 out 159 which is good rate of 92.45%, only 12 patients were predicted wrongly, 9 as Normal and 3 as R-ACS.

Class R-ACS: prediction was good, 48 out of 59 patients were correct and almost the rest were predicted as R class, 1 as Normal and 2 as ACS.

Class NORMAL: were very good, with success rate of 85.51%, as 59 out of 69 patients were predicted correctly but the problem was in predicting one patients as R-ACS while patient was Normal.

Class ACS: 15 out of 19 patients were predicted correctly and this needs attention as this is related to heart, 4 mistakes in this department out of 19 is high number.

Neural Network prediction summary for the combined dataset:

1. Very good prediction performance rate for R class, it is true that Neural network predicted 9 as Normal while they were Respiratory, but this is not big problem as no patients were predicted as ACS.
2. Two patients were predicted as ACS while they are having a high possibility to be affected and this is not problem.
3. The problem when Neural network predict ACS as NORMAL, the algorithm predicted 2 patients out of 59 but even though this is accepted number but still need enhancement.

3.5.3 Neural Network Features Importance Analysis

Figure 21 shows the sigmoid function which calculates the weight of the hidden nodes in the Neural Network, in Table 6 we have three classes and in each class in next to each node there is the sigmoid (weight value) which indicated the attribute importance in predicting the class the positive value is our concern and I will ignore the negative values.

$$z = \sum_{i=1}^m w_i x_i + bias$$

Sigmoid Function is: $\sigma(z) = \frac{1}{1+e^{-z}}$

Figure 21: Sigmoid function

(Towards Data Science ,2017) The only important thing about sigmoid function graph in Figure 22 is first, its curve, and second, its derivative. Here are some more details:

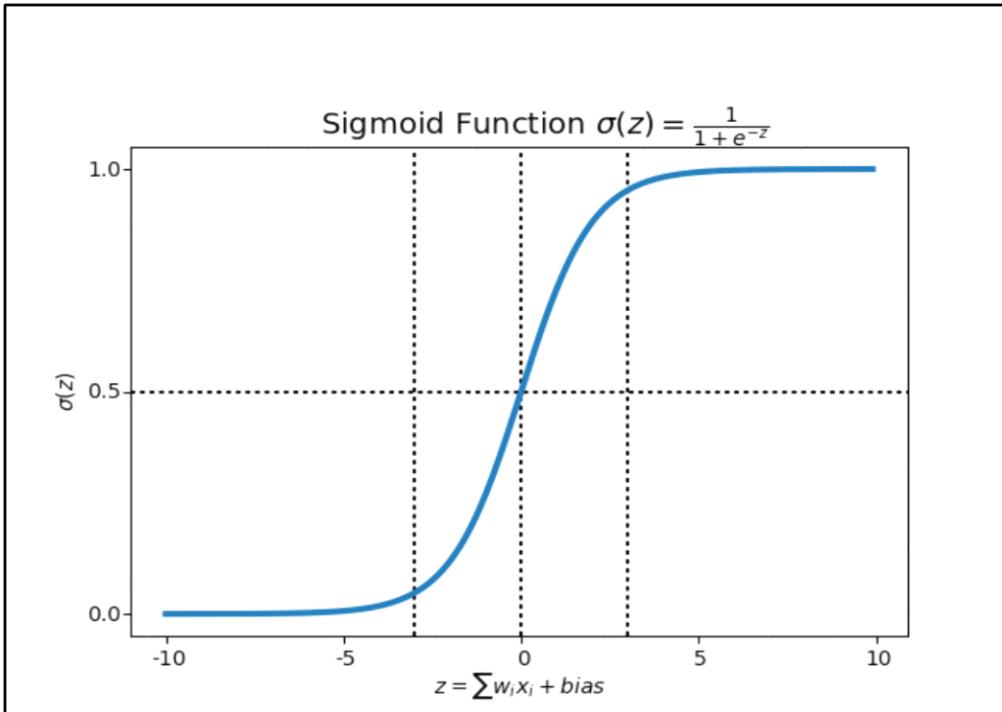


Figure 22: Sigmoid Graph

The sigmoid function produces comparable outcomes to step work in that the output is near 0 and 1. The curve crosses 0.5 at $z=0$, which we can set up rules for the activation function, for example, If the sigmoid neuron's output is bigger than or equivalent to 0.5, it outputs 1; if the output is littler than 0.5, it outputs 0. The sigmoid function does not have a snap on its curve. It is smooth, and it has an extremely pleasant and straightforward subordinate of $\sigma(z) * (1-\sigma(z))$, which is differentiable wherever on the curve. The analytics inference of the subordinate can be found on Stack Overflow here if you need to see it. In any case, you don't need to know how to infer it. On the off chance that z is exceptionally negative, at that point the output is roughly 0; if z is extremely positive, the output is around 1; however, around $z=0$ where z is neither too extensive or too little (in the middle of the two-external vertical dabbed network lines in figure 23), we have generally more deviation as z changes.

3.5.4 Respiratory Dataset (NN)

The analysis of the Neural Network from the generated network, the node with thick connections has higher importance than one with normal thickness, figure 15 shows that clearly.

Bagging has created 10 networks, the analysis of the Neural Network from the generated networks tells that the node with thick connections has higher importance than one with normal thickness, sample Network shown in figure 23.

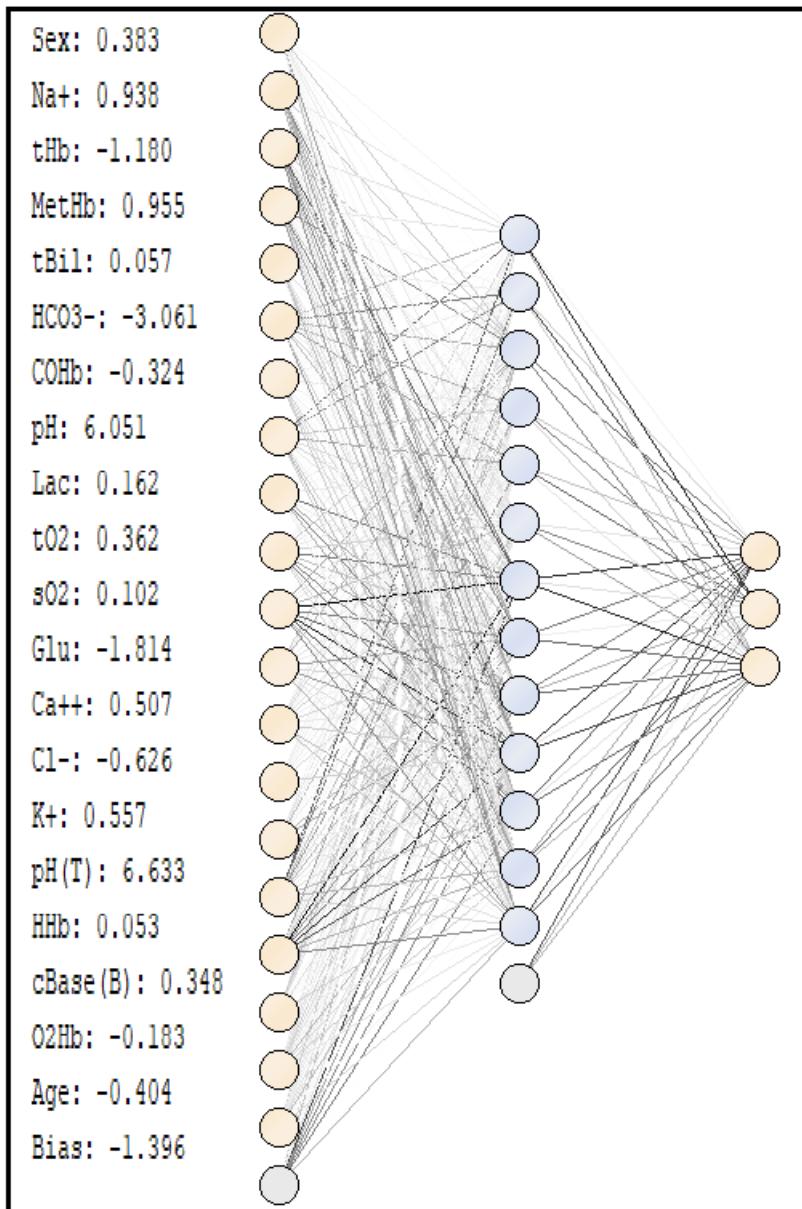


Figure 23: Neural Network Attribute weight – Respiratory Dataset

(SAS institute Inc ,2000) The nearest values to 0 nodes indicate the heights importance nodes, so the attribute importance order as listed in table xx which ignores negative and values above 1, for better understanding, when looking at the report in Table xx generated by Rapid miner we need to notice the Sigmoid values for each node, sigmoid function defined as shown in figure 22. as in rapid miner the hidden layer size value is set to -1, then layer size would be calculated from the number of attributes of the input example set, the hidden layers equal (number of attributes + number of classes) / 2 + 1 (Rapid Miner ,2018), in our case we have $(20 + 3) / 2 + 1 = 13$ nodes in the hidden layer + the bias node.

Attribute	Sigmoid
HHb	0.053
tBil	0.057
sO2	0.102
Lac	0.162
cBase(B)	0.348
tO2	0.362
Sex	0.383
Ca++	0.507
K+	0.557
Na+	0.938
MetHb	0.955

Table 19: Neural Net attributes importance using Node weight – Respiratory Dataset

3.5.5 Cardiac Dataset (NN)

In cardiac dataset I used the Neural Network without bagging as shown in figure 24.

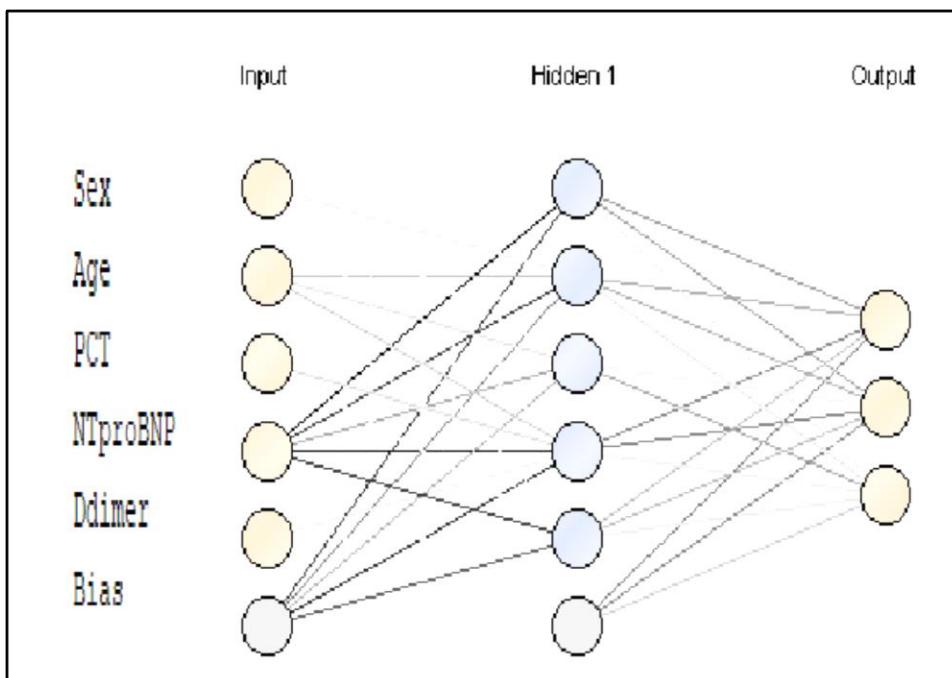


Figure 24: Neural Net attributes importance using Node weight – Cardiac Dataset

Important features listed in table 20

Attribute	Sigmoid
NT-ProBNP	0.808
D-Dimer	5.100
Age	5.879

Table 20: Neural Net attributes importance using Node weight – Cardiac Dataset

3.5.6 The Combined Datasets (NN)

Same for the combined dataset, feature importance as shown in table 21,

Attribute	Sigmoid
Age	0.053
NT-ProBNP	0.057
D-Dimer	0.102
sO2	0.162
O2Hb	0.348
pH	0.362

Table 21: Neural Net Sigmoid values – The Combined Dataset

3.6 Naïve Bias (NB)

Naïve Bayes is the premise for numerous machine-learning and data mining strategies (Tang & Maclennan, 2005). The calculation is used to make models with prescient capabilities. It gives better approaches of investigating and understanding data. It learns from the “evidence” by calculating the relationship between the target (i.e., subordinate) and other (i.e., autonomous) factors. (Palaniappan, & Awang, 2008)

3.6.2 Naïve Bias Prediction Results

In this part, Naïve Bias operator used, and tried it with bagging, boosting and with discretizing age, but the highest accuracy was when using the operator directly inside the cross-validation. Used replace missing values 3 times as (Min, Max, and average), the respiratory dataset accuracy as listed in table 22 achieved the lowest compared with the previous algorithms, and the surprise was in the cardiac dataset, as a noticeably decreased in the accuracy reached up to 20% of the previous methods, it scored only 77.38%. but the combined dataset showed better accuracy using replace missing value operator by minimum, 81.34% was the result.

Naïve Bias Accuracy			
Respiratory Dataset Accuracy			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	81.34%	79.73%	80.35%
Cardiac Dataset			
Replace missing values by	Not Used		
Results	77.38%		
The Combined Datasets			
Replace missing values by	<u>Minimum</u>	<u>Maximum</u>	<u>Average</u>
Results	83.66%	82.99%	83.38%

Table 22: Naïve Bias Accuracy results – All Datasets

3.6.2.1 Naïve Bias Performance analysis

In this part, deep analysis of the confusion matrix of the Naïve bias highest accuracy trial will be listed, this will illustrate the class label accuracy and class precision for each dataset separately.

3.6.2.2 Respiratory Dataset (NB)

Naïve Bias confusion matrix analysis listed in table x,

	true R1/186	true R2/32	true Normal/88	class precision
pred. R1	161	22	6	85.19%
pred. R2	10	10	4	41.67%
pred. Normal	15	0	78	83.87%
class recall	86.56%	31.25%	88.64%	

Table 23: Naïve Bias confusion matrix - Respiratory Dataset

For the class R1: Naïve Bias was able to predict 161 out 186 patients correctly, with a low error rate as 10 patients were predicted as respiratory type 2 (R2), and 15 as NORMAL.

R2 Class: poor prediction performance as 22 out of 32 patients were wrongly predicted as R1 while they were R2.

For NORMAL class: 78 patients out of 88 were predicted correctly with 88.64% class prediction accuracy and this is being the best class for respiratory dataset prediction.

Naïve Bias prediction summary:

1. Good prediction performance for class R1
2. Low prediction performance for class R2
3. Very good performance for class NORMAL
4. Overall fair performance for Naïve Bias.

3.6.2.3 Cardiac Dataset (NB)

Naïve Bias confusion matrix analysis for Cardiac dataset listed in table 24,

	true Normal/228	true HPACS/71	true ACS/7	class precision
pred. Normal	210	14	0	93.75%
pred. HPACS	2	22	1	88.00%
pred. ACS	16	35	6	10.53%
class recall	92.11%	30.99%	85.71%	

Table 24: Naïve Bias confusion matrix – Cardiac Dataset

Class NORMAL: class NORMAL shows 92.11% prediction accuracy, Naïve Bias was not able to beat the previous algorithms for this class, it predicted 210 patients correctly out of 228.

Class HPACS: very poor prediction for class HPACS (High possibility of Acute Coronary Syndrome) as only 22 patients out of 71 patients was predicted correctly with a percentage of 30.99%.

Class ACS: Here is the great news, in Naïve bias the ACS class finally was predicted correctly with very high success rate reached to 85.71%, it is very high compared to the previous algorithms, Naïve bias technique was able to detect the correct class label for ACS but unfortunately it was not that good in predicting the other classes as the previous algorithms.

Naïve Bias prediction summary for the cardiac dataset:

1. fair prediction rate for ACS – 85.71%
2. Good prediction for class NORMAL – 92.11%
3. Poor prediction performance for class HPACS – 30.99%
4. Only ACS class prediction in Naïve Bias was better than the other algorithms.

3.6.2.4 The Combined Datasets (NB)

Naïve Bias confusion matrix analysis listed in table 25,

	true R/ 159	true R-ACS/ 59	true Normal/ 69	true ACS/ 19	class precision
pred. R	137	7	4	0	92.57%
pred. R-ACS	10	47	3	4	73.44%
pred. Normal	12	3	58	1	78.38%
pred. ACS	0	2	4	14	70.00%
class recall	86.16%	79.66%	84.06%	73.68%	

Table 25: Naïve Bias confusion matrix – The Combined Datasets

Class R: Naïve Bias succeeded to predict 137 out 159 which is less than all the previous algorithms with success rate of 86.61%, 24 patients were predicted wrongly but the good thing none of the patients were predicted as ACS.

Class R-ACS: prediction was excellent, 47 out of 59 patients were correct and the 7 as R, 3 as Normal and two patients were predicted as ACS. And this can be accepted as R-ACS and ACS almost same.

Class NORMAL: were not excellent, as four patients were predicted as ACS while they are NORMAL, but still, this is not a big issue, the problem if the scenario happened in the opposite direction if the patients were ACS and predicted as Normal and this is the false – negative.

Class ACS: one case only false negative appeared in ACS class, and the rest four patients were predicted as R-ACS, which does not lead to a big medical problem.

Naïve Bias performance summary:

1. Fair prediction rate for R the class performance was 86.16%.
2. Good prediction rate for R-ACS class with the performance of 79.66%.
3. Overall good prediction performance for ACS and R-ACS.

3.6.3 Naïve Bias Features Importance Analysis

In this part, the most important attributes which affected the analysis will be listed and discussed for each dataset used.

3.6.3.1 Respiratory Dataset (NB)

In Naïve bias, O2Hb attribute tends to be between 85.0 to 99.0 when class is normal, else of that values distributes between Respiratory and Heart Failure ACS class as shown in figure 24

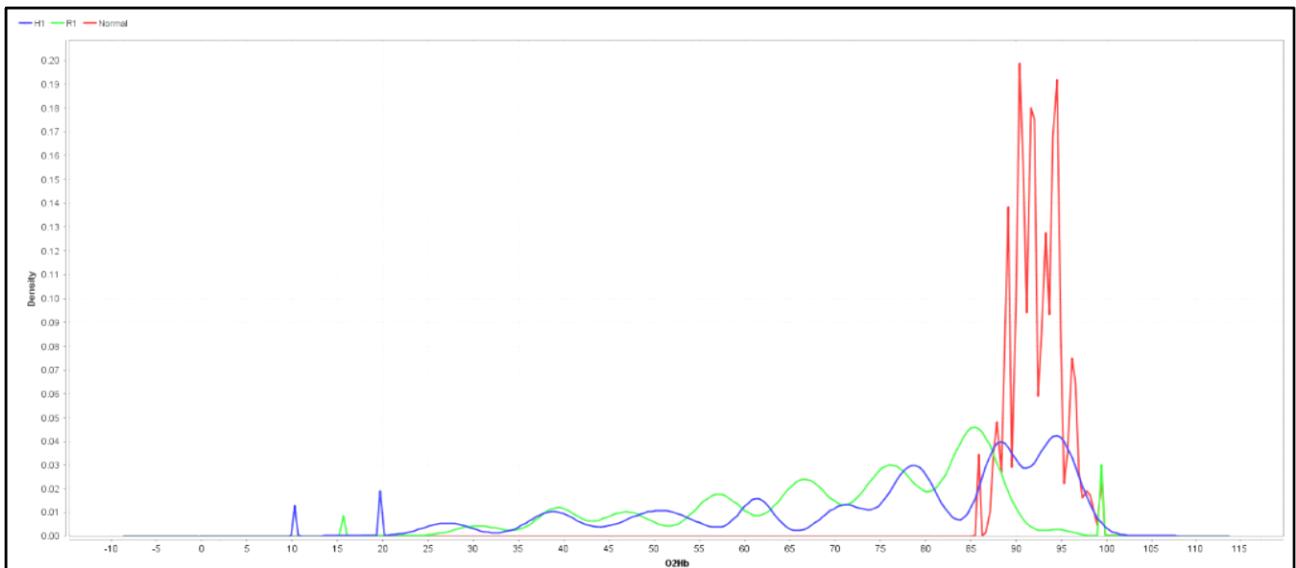


Figure 25: O2Hb importance in Naïve Bias – Respiratory Dataset

In Naïve bias, sO2 attribute tends to be between 85.0 to 105.0 when class is normal, else values distributes between R1 and R2 classes as shown in figure 25.

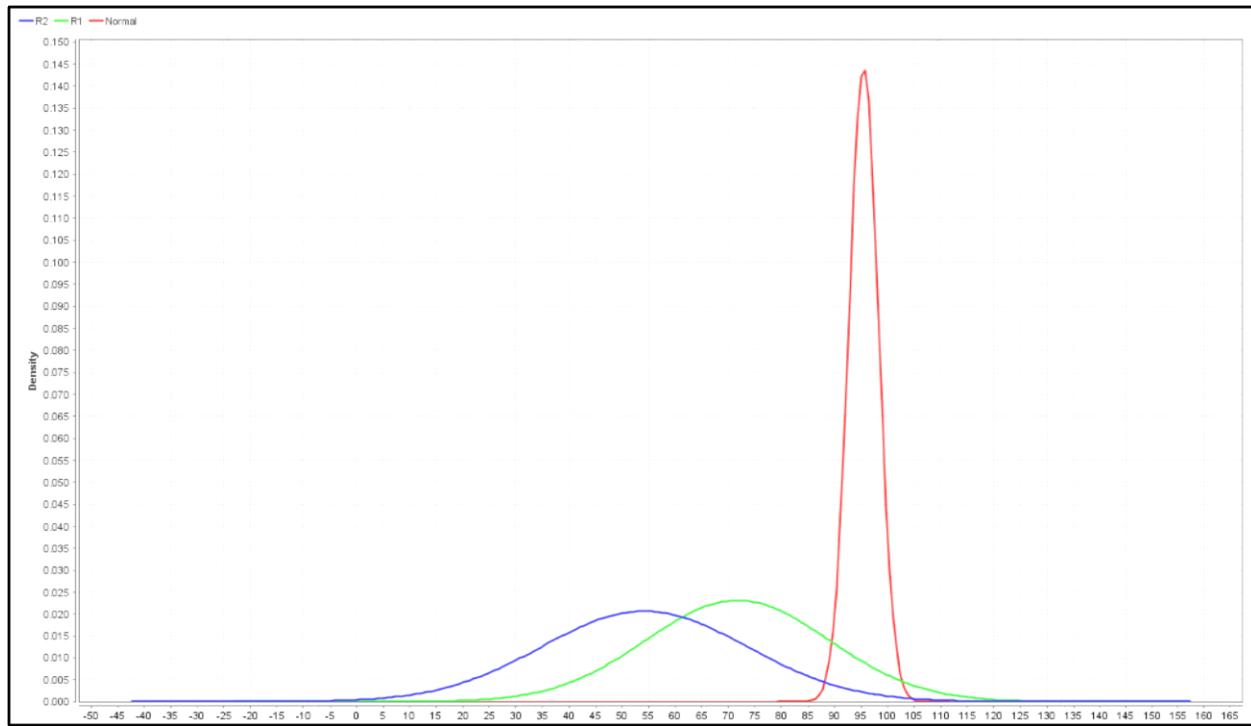


Figure 26: sO2 importance in Naïve Bias – Respiratory Dataset

3.6.3.2 Cardiac Dataset (NB)

In Naïve bias, NT-ProBNP attribute tends to be between 0 to 54 when class is NORMAL, and if more than 54 to 325 then patient class is HPACS, and if from 100 to 450 then patient is ACS class as shown in figure 26.

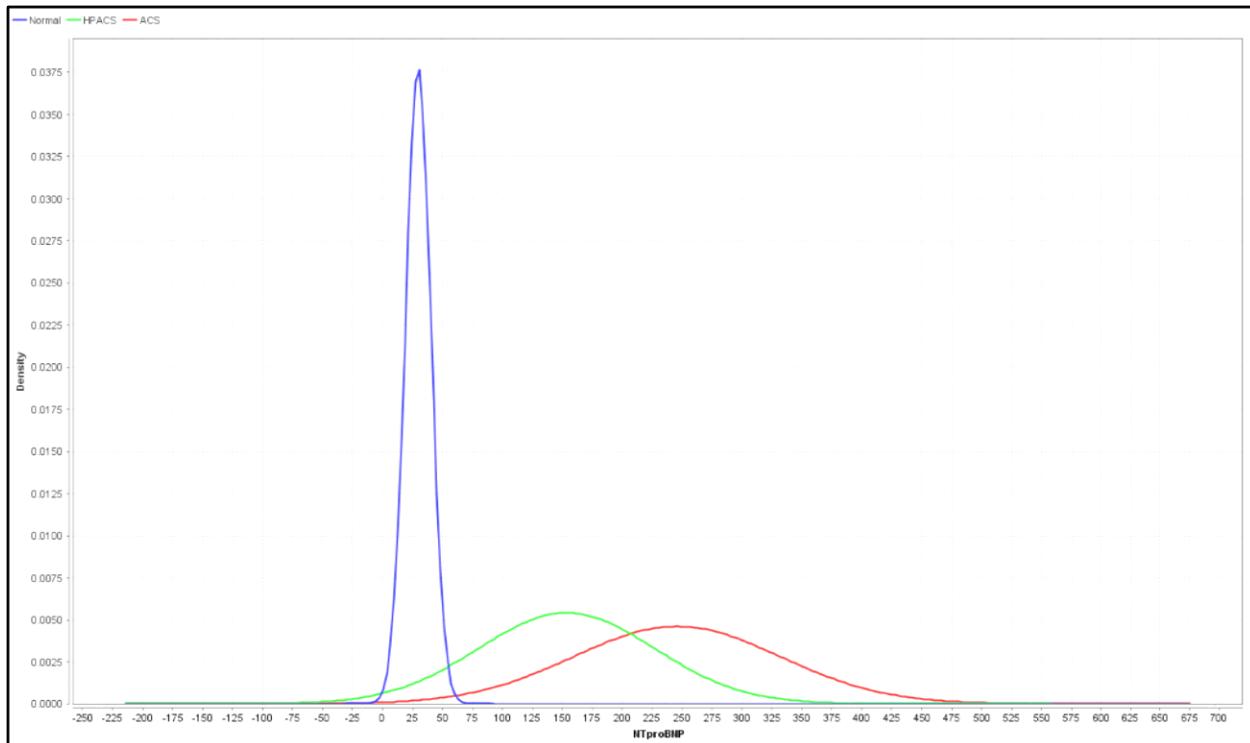


Figure 27: Nt-ProBNP importance in Naïve Bias – Cardiac Dataset

PCT does not make difference as the variance between NORMAL and ACS is very less as shown in figure 27.

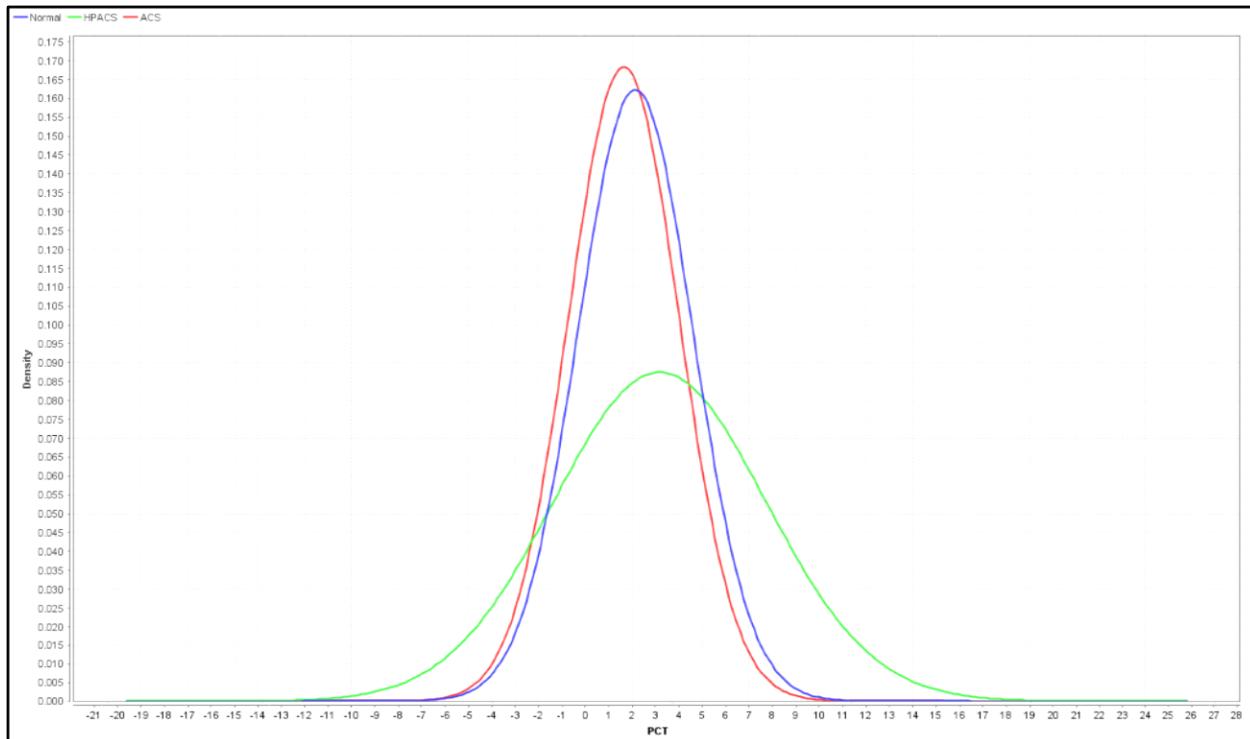


Figure 28: PCT importance in Naïve Bias – Cardiac Dataset

3.7 Discussion

The goal of this dissertation was to find the relation between respiratory and cardiac and find if respiratory features can be used to predict the cardiac features. the analysis of the performance and accuracy of each algorithm shows very interesting facts, and it is very true that accuracy is not the only way to measure how good was the test, in this dissertation I was able to confirm that Blood gases test results (Respiratory) can predict Cardiac ACS Acute coronary syndrome, this study proved that blood gas testing (Respiratory) can help in showing the respiratory types and limits and can predict if the patient is cardiac or not, this research tried to find the relation between respiratory and chest pain and it was able to prove that sO₂ and O₂Hb from Respiratory dataset with help of NT-ProBNP from Cardiac dataset can predict if the patient has ACS (Acute coronary syndrome) or there is a risk of ACS.

According to the analysis made based on the selected models, from accuracy prospective Gradient boosted tree had the highest accuracy, which the highest accuracy reached up to 88.87, for the respiratory dataset, 96.73% for the Cardiac dataset and 93.16 when I combined the datasets together. the execution performance was very low in the boosted trees and Neural network comparing to other models. Table 26 listed each dataset algorithms accuracy summery.

Respiratory Dataset				
Algorithm	Replace Missing	Time\Min	Customization	Accuracy
Decision Tree	Maximum	0.1	Creation Type: Accuracy	88.55%
G. Boosted Trees	Average	0.28	Use 20 trees	88.87%
Neural Network	Maximum	1.47	Via Bagging using 700 Training Cycle	87.25%
Naïve Bias	Maximum	0.2	Via boosting	81.34%
Cardiac Dataset				
Decision Tree	NA	0.1	Creation Type: Accuracy	96.37%
G. Boosted Trees	NA	0.11	Use 20 trees	96.73%
Neural Network	NA	0.49	Via Bagging using 700 Training Cycle	96.06%
Naïve Bias	NA	0.1	Via boosting	77.38%
The Combined Datasets				
Decision Tree	Maximum	0.1	Creation Type: Accuracy	91.85%
G. Boosted Trees	Maximum	0.15	Use 20 trees	93.16%
Neural Network	Maximum	1.59	Via Bagging using 700 Training Cycle	87.88%
Naïve Bias	Maximum	0.2	Via boosting	83.66%

Table 26: All Methods accuracy summary- All Datasets

Performance vector summary as listed for each dataset in Tables 27,28 and 29, shows good performance for the respiratory dataset, all algorithms used were able to predict class R1 and class Normal perfectly, but the problem was in class R2, with an average accuracy of 46.8%. therefor I can't confirm that R2 class was predicted correctly. For Cardiac dataset, algorithms did very well for NORMAL and HPACS classes but unfortunately all algorithms except Naïve Bias fails to predict ACS correctly, Naïve bias didn't perform well as decision tree for instance, but it was able to predict 6 patients out of 7 as ACS correctly, while the other algorithms predicted 1 out of 7 correctly and some of them got 0 correct prediction. Troponin T which indicates the ACS class is very sensitive, most of its values are between 0.01 and 0.09 while the other attributes in the cardiac dataset have a quite large range. But when I combined both respiratory and cardiac dataset, the performance went up, even the ACS class was predicted correctly by almost all the algorithms with success percent reached to 80.2% and this is great achievement and the goal of this dissertation, as this class was hard to be evaluated due to the narrow difference Troponin T values.

Respiratory Dataset:

	true R1: 186	true R2: 32	true Normal: 88	class precision
Decision Tree				
pred. R1	172	14	7	89.12%
pred. R2	2	18	0	90.00%
pred. Normal	12	0	81	87.10%
class recall	92.47%	56.25%	92.05%	
Gradient Boosted Tree				
pred. R1	175	14	9	88.38%

pred. R2	0	18	0	100.00%
pred. Normal	11	0	79	87.78%
class recall	94.09%	56.25%	89.77%	
Neural Network				
pred. R1	176	18	11	85.85%
pred. R2	1	14	0	93.33%
pred. Normal	9	0	77	89.53%
class recall	94.62%	43.75%	87.50%	
Naïve Bias				
pred. R1	161	22	6	85.19%
pred. R2	10	10	4	41.67%
pred. Normal	15	0	78	83.87%
class recall	86.56%	31.25%	88.64%	
	true R1/186	true R2/32	true Normal/88	class precision

Table 27: Algorithms Performance Summary – Respiratory Dataset

Cardiac Dataset:

	true Normal/ <u>288</u>	true HPACS/ <u>71</u>	true ACS/ <u>7</u>	class precision
Decision Tree				
pred. Normal	228	3	0	98.70%
pred. HPACS	0	67	6	91.78%
pred. ACS	0	1	1	50.00%
class recall	100.00%	94.37%	14.29%	
Gradient Boosted Tree				
pred. Normal	228	3	0	98.70%
pred. HPACS	0	68	7	90.67%
pred. ACS	0	0	0	0.00%
class recall	100.00%	95.77%	0.00%	
Neural Network				
pred. Normal	227	4	0	99.27%
pred. HPACS	1	67	7	89.33%
pred. ACS	0	0	0	0.00%
class recall	99.56%	94.37%	0.00%	
Naïve Bias				
pred. Normal	210	14	0	93.75%
pred. HPACS	2	22	1	88.00%
pred. ACS	16	35	6	10.53%
class recall	92.11%	30.99%	85.71%	
	true Normal/228	true HPACS/71	true ACS/7	class precision

Table 28: Algorithms Performance Summary – Cardiac Dataset

The Combined Dataset:

	true R/ 159	true R-ACS/ 59	true Normal/ 69	true ACS/ 19	class precision
Decision Tree					
pred. R	146	3	3	0	96.05%
pred. R-ACS	4	54	0	2	90.00%
pred. Normal	9	0	65	1	86.67%
pred. ACS	0	2	1	16	84.21%
class recall	91.82%	91.53%	94.20%	84.21%	
Gradient Boosted Tree					
pred. R	150	3	4	0	95.54%
pred. R-ACS	1	55	0	2	94.83%
pred. Normal	8	0	64	1	87.67%
pred. ACS	0	1	1	16	88.89%
class recall	94.34%	93.22%	92.75%	84.21%	
Neural Network					
pred. R	147	8	9	0	89.63%
pred. R-ACS	3	48	1	2	88.89%
pred. Normal	9	1	59	2	83.10%
pred. ACS	0	2	0	15	88.24%
class recall	92.45%	81.36%	85.51%	78.95%	
Naïve Bias					
pred. R	137	7	4	0	92.57%
pred. R-ACS	10	47	3	4	73.44%
pred. Normal	12	3	58	1	78.38%
pred. ACS	0	2	4	14	70.00%
class recall	86.16%	79.66%	84.06%	73.68%	

Table 29: Algorithms Performance Summary – All Datasets

All the used models show almost narrow results, but some of them were better when using via bagging, and some of them preferred to replace the missing items with Minimum and some of them made higher accuracy when replacing with Maximum. And here are the most noticeable differences.

1. Decision Tree: shows better performance when using accuracy in tree creation type and when replaced the missing values with maximum, there was 4% difference between the best and worst trial for decision tree. As gain ration -maximum resulted in 79.74 % while accuracy – Minimum achieved 88.87 %.

2. Gradient boosted tree: was better than decision tree, the testing was interesting, as the results were changing and responding when I modified the process, for instance, when used bagging the accuracy was lower in 4% but it increased when used the operator directly it reaches the best result.
3. Neural Network: it achieved the best accuracy at 96.06 %, the model achieved this number when I increased the training cycles from 500 to 700, I tried to increase the cycles more than 700 but I noticed that the accuracy started to decrease, it reached to the peak when I used 700 cycles via bagging. But the main issue in this technique was the execution time as the best result was achieved in 1.47 minutes.
4. Naïve Bias: was a strange algorithm and I was not able to find the reason of its differences than other algorithms, it does not provide any information for analysis after the test, but it was great in predicting ACS class while the rest algorithms failed to predict it.

In this dissertation, I found that some features (Attributes) affected the prediction, those attributes appeared in most techniques as a key attribute. NT-ProBNP, sO₂, O₂Hb, K+, MetHb, and tHb played an important role to predict the class labels.

4 Chapter 4: Conclusion

This research is novel as it was able combine two different topics not studied before and it identified the relation between cardiac and respiratory, many features in the respiratory dataset have an influence on the class label in the cardiac dataset and that was maybe known by doctors and specialist but not proved using data mining. The archived results of this research answered the research questions and shown interesting results to this project consultants (Dr. Fayha Ahmad and Dr. Salwa Nori).

First question, “Does combining data from respiratory and cardiac results in higher precisions?”

Answer: Yes, combining both cardiac and respiratory helped to increase the classification precession, the research found that sO₂ values have a direct relation to NT-proBNP, and NT-proBNP can predict the Troponin T values for specific age and gender. And here is the formula from Rapid Miner.

If sO₂ was >90.850 then check [NTproBNP], If [NTproBNP] > 54.943 then check [sO₂] again if [sO₂] >92.900 then patient chest pain caused by ACS, else if [NTproBNP] <54.934 then most probably patient is (NORMAL). But if [sO₂] <= 90.850 then see if [NTproBNP] > 53.502 if yes then most probably patient has both a respiratory and coronary problems (R-ACS). But if [NTproBNP] <= 53.502 then check [MetHb], is if [MetHb] <= 1.350 then patient for sure has only respiratory problems (R).

Second question, “Can respiratory diagnosis equipment’s (Blood Gases) be a step one in diagnosing cardiac (Chest Pain)?” Answer: Yes, because the respiratory (Blood gas) features (sO₂, MetHb, tO₂, and pH) have shown a direct influence on the cardiac features values (NT-ProBNP, PCT, and Troponin T)

Third question, “Can existing classification techniques achieve reasonable precession when predicting chest pain using cardiac and respiratory results?” Answer, Yes, it was clear from research results, that predicting respiratory type 2 from respiratory dataset had a low precision, and the accuracy of predicting Acute coronary syndrome from cardiac dataset shown very low performance, but when the two datasets united, the accuracy for respiratory and Cardiac increased.

Fourth question, “Can respiratory diagnosis equipment’s (Blood Gases) be a step one of diagnosing Hypertrophy of the heart muscle?” Answer: Yes, as K⁺ and sO₂ were able to predict NT-ProBNP values.

From prediction performance prospective, it is correct that accuracy is not the key of the prediction quality, confusion matrix showed that the prediction of two out three classes was excellent but for one class the results were bad with an average of less than 50% in each dataset. But after combining the respiratory and cardiac dataset the attributes united and did good work together and increased the performance vector from 50% to 80% for that classes. This means that this research recommends combining cardiac features to respiratory features to help diagnosing chest pain in critical departments such as emergency, critical units and operation rooms.

Finally, In the near future, I will be working on adding more attributes analyzed by Radiometer Medical Apps* to the dataset to predict the relationship between kidney diseases, respiratory and cardiac to find out the effect of kidney on the heart overall performance.

References

- Sudhakar, K. and Manimekalai, D.M., 2014. Study of heart disease prediction using data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(1).
- Taneja, A., 2013. Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, 6(4), pp.457-466.
- Chitra, R. and Seenivasagam, V., 2013. Review of heart disease prediction system using data mining and hybrid intelligent techniques. *ICTACT journal on soft computing*, 3(04), pp.605-09.
- Masethe, H.D. and Masethe, M.A., 2014, October. Prediction of heart disease using classification algorithms. In Proceedings of the world congress on Engineering and Computer Science (Vol. 2, pp. 22-24).
- Tomar, D. and Agarwal, S., 2013. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5), pp.241-266.
- Banu, M.N. and Gomathy, B., 2014, March. Disease forecasting system using data mining methods. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 130-133). IEEE.
- Schmidt, M., Tachon, G., Devilliers, C., Muller, G., Hekimian, G., Bréchot, N., Merceron, S., Luyt, C.E., Trouillet, J.L., Chastre, J. and Leprince, P., 2013. Blood oxygenation and decarboxylation determinants during venovenous ECMO for respiratory failure in adults. *Intensive care medicine*, 39(5), pp.838-846.
- Grocott, M.P., Martin, D.S., Levett, D.Z., McMorrow, R., Windsor, J. and Montgomery, H.E., 2009. Arterial blood gases and oxygen content in climbers on Mount Everest. *New England Journal of Medicine*, 360(2), pp.140-149.
- Kotaska, K., Urinovska, R., Klapkova, E., Prusa, R., Rob, L. and Binder, T., 2010. Re-evaluation of cord blood arterial and venous reference ranges for pH, pO₂, pCO₂, according to spontaneous or cesarean delivery. *Journal of clinical laboratory analysis*, 24(5), pp.300-304.
- Lehrke, S., Steen, H., Sievers, H.H., Peters, H., Opitz, A., Müller-Bardorff, M., Wiegand, U.K., Katus, H.A. and Giannitsis, E., 2004. Cardiac troponin T for prediction of short-and long-term morbidity and mortality after elective open heart surgery. *Clinical Chemistry*, 50(9), pp.1560-1567.
- Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.
- Delen, D. and Olson, D., 2008. Advanced data mining techniques. Springer. doi, 10, p.9780470172339.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

West, J.B., Lahiri, S.U.K.H.A.M.A.Y., Maret, K.H., Peters Jr, R.M. and Pizzo, C.J., 1983. Barometric pressures at extreme altitudes on Mt. Everest: physiological significance. *Journal of Applied Physiology*, 54(5), pp.1188-1194.

Xing, Y., Wang, J. and Zhao, Z., 2007, November. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *Convergence Information Technology*, 2007. International Conference on(pp. 868-872). IEEE.

Ordonez, C., Omiecinski, E., De Braal, L., Santana, C.A., Ezquerra, N., Taboada, J.A., Cooke, D., Krawczynska, E. and Garcia, E.V., 2001. Mining constrained association rules to predict heart disease. In *Data Mining*, 2001. ICDM 2001, Proceedings IEEE International Conference on (pp. 433-440). IEEE.

Rani, K.U., 2011. Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.

Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.P.P., 2013. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4), pp.1086-1093.

Sundar, N.A., Latha, P.P. and Chandra, M.R., 2012. Performance analysis of classification data mining techniques over heart disease database. *IJESAT] International Journal of engineering science & advanced technology ISSN*, pp.2250-3676.

Jabbar, M.A., Chandra, P. and Deekshatulu, B.L., 2011. Cluster based association rule mining for heart attack prediction. *Journal of Theoretical and Applied Information Technology*, 32(2), pp.196-201.

SA, S., 2013. Intelligent heart disease prediction system using data mining techniques. *International Journal of Healthcare & Biomedical Research*, 1, pp.94-101.

Hsieh, N.C., Hung, L.P., Shih, C.C., Keh, H.C. and Chan, C.H., 2012. Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. *Journal of medical systems*, 36(3), pp.1809-1820.

Atkov, O.Y., Gorokhova, S.G., Sboev, A.G., Generozov, E.V., Muraseyeva, E.V., Moroshkina, S.Y. and Cherniy, N.N., 2012. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of cardiology*, 59(2), pp.190-194.

Pattekari, S.A. and Parveen, A., 2012. Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp.290-294.

Patil, S.B. and Kumaraswamy, Y.S., 2009. Extraction of significant patterns from heart disease warehouses for heart attack prediction. IJCSNS, 9(2), pp.228-235.

Cerretelli, P., 1976. Limiting factors to oxygen transport on Mount Everest. Journal of applied physiology, 40(5), pp.658-667.

West, J.B., Boyer, S.J., Graber, D.J., Hackett, P.H., Maret, K.H., Milledge, J.S., Peters Jr, R.M., Pizzo, C.J., Samaja, M. and Sarnquist, F.H., 1983. Maximal exercise at extreme altitudes on Mount Everest. Journal of applied physiology, 55(3), pp.688-698.

Lawrence, R., Bunn, A., Powell, S. and Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote sensing of environment, 90(3), pp.331-336.

Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

Tang, Z. and Maclennan, J., 2005. Data mining with SQL Server 2005. John Wiley & Sons.

Kaur, G., Pandey, O.P., Singh, K., Homa, D., Scott, B. and Pickrell, G., 2014. A review of bioactive glasses: their structure, properties, fabrication and apatite formation. Journal of Biomedical Materials Research Part A, 102(1), pp.254-274.

Towards Data Science (2017) Available at: <https://towardsdatascience.com/multi-layer-neural-networks-with-sigmoid-function-deep-learning-for-rookies-2-bf464f09eb7f> (Accessed: 15 February 2018).

Rapid Miner (2018) Available at:
https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/neural_net.html (Accessed: 15 February 2018).

SAS Institute Inc (2000) Available at:

ftp://ftp.sas.com/pub/neural/importance.html#linmod_comp_wgt (Accessed: 22 March 2018)