2021-10

HEART DISEASE DIAGNOSIS BY COMBINING SELF-ORGANIZING NEURAL NETWORK AND SUPPORT VECTOR MACHINE

YOSEPH, DIRIRSA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF COMPUTING

HEART DISEASE DIAGNOSIS BY COMBINING SELF-ORGANIZING NEURAL NETWORK AND SUPPORT VECTOR MACHINE

YOSEPH DIRIRSA

Addis Ababa, Ethiopia

October 2021

HEART DISEASE DIAGNOSIS BY COMBINING SELF-ORGANIZING NEURAL NETWORK AND SUPPORT VECTOR MACHINE

YOSEPH DIRIRSA

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar University, Institute of Technology, BDU in partial fulfilment of the requirements for the degree of Master of Science in the Computer Science in Computing Faculty.

Advisor: Dr. Elefelious Getachew

Addis Ababa, Ethiopia

October 2021

DECLARATION

I, the undersigned, declare that this thesis comprises of my own work. In compliance with internationally accepted practices, I have duly acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the university and can evoke penal action from the sources which have not been properly cited or acknowledged.

Name of the student: Yoseph Dirirsa

Signature: _____

Date of submission: October 2021

Place: Addis Ababa

This thesis has been submitted for examination with my approval as a university advisor

Advisor Name: Elefelious Getachew (PhD)

Advisor's Signature: _____

Bahir Dar University

Bahir Dar Institute of Technology

School of Research and Graduate Studies

Faculty of Computing

THESIS APPROVAL SHEET

Student:

_____	_____	_____
Name	Signature	Date

The following graduate faculty members certify that this student has successfully presented the necessary written final thesis and oral presentation for partial fulfilment of the thesis requirements for the Degree of Master of Science in computer science

Approved by:

Advisor:

_____	_____	_____
Name	Signature	Date

External Examiner:

_____	_____	_____
Name	Signature	Date

Internal Examiner:

_____	_____	_____
Name	Signature	Date

Chair Holder:

_____	_____	_____
Name	Signature	Date

Faculty Dean:

_____	_____	_____
Name	Signature	Date

ACKNOWLEDGEMENTS

In the first place, I would like to thank my almighty God, as nothing would be possible without his divine support. I am honestly grateful to my advisor, Dr. Elefelious Getachew for his supervision, advice and guidance while I was working on this thesis. Thank you so much for showing me to an interesting research direction and for the valuable comments you gave me.

Finally, I would like to extend gratitude to my family, friends and relatives for their moral support and encouragement during this research work. Thank you all.

Yoseph Dirirsa

ABSTRACT

Heart Disease or Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels including coronary heart disease, cerebrovascular disease, rheumatic heart disease and others. It is the number one cause of death globally, taking an estimated 17.9 million lives each year, and one third of these deaths occur in people under 70 years of age, which is working groups. The prevalence and the expense of treating heart disease had projected to increase in many countries by 2030.

In the last decades, many researches had conducted on heart disease to minimize deaths caused due to these diseases. However, heart disease remains as the number one cause of death globally. Therefore, it is very essential to conduct a study that initiates alternative means to diagnose heart diseases. Accordingly, this study aims to bring new insight by combining Self Organizing Neural Network (SONN) and Support Vector machine (SVM) algorithms on different kernel functions to classify heart diseases.

In this research, we used 303 clinical datasets collected from Cleveland Clinic Foundation. These datasets had pre-processed to make it useful for the experimentation and we conducted all the experimentations by using Python on PyCharm IDE. First, we experimented and evaluated the dataset by using SVM alone on different kernel functions. Then we experimented by combining SONN and SVM algorithms. To combine these two algorithms, first SONN used to get the cluster representation of each input variables, then these clusters fed to the SVM as a feature to classify Heart disease. Evaluations on the combined model showed that SONN has increased the accuracy and specificity of SVM on Polynomial Kernel function to diagnose Heart disease.

Keywords: Heart Disease, Support Vector Machine, Self-Organizing Neural Network, Kohonen Self-Organizing Map, Machine Learning

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

- ACC- American Cardiac College
- AHA- American Heart Association
- AI- Artificial Intelligence
- ASCVD- Atherosclerotic Cardiovascular Disease
- BFE- Backward Feature Elimination
- CVDs- Cardiovascular diseases
- KSOM- Kohonen Self Organizing Map
- ML- Machine Learning
- RBF- Radial Basis Function
- SONN- Self Organizing Neural Network
- SVM – Support Vector Machine
- WHO- World Health Organization

LIST OF FIGURES

LIST OF TABLES

Chapter one

1. Introduction

1.1.Background

A heart disease or cardiovascular disease (CVD) is the death of a segment of the heart muscles caused by a loss of blood supply to it. The blood cut off when an artery supplying the heart muscle had blocked by a blood clot (Justin & Tim, 2017). Most often, the blockage may cause by build-up of fat, cholesterol and other substances, which form a plaque in the arteries that feed the heart (coronary arteries). The plaque could break away and forms a clot. As plaque continues to accumulate, the patient's coronary arteries detrimentally narrow over time and reduce blood flow to the heart, thus increase the risk of heart failure, heart attack or stroke (Miao et al., 2016).

Nowadays, heart disease is ubiquitous cause of morbidity and mortality in most countries. According to WHO (2019) CVD are the number one cause of death globally, taking an estimated 17.9 million lives each year, it represents about 31% of the global deaths. In addition, the people who are living in low and middle-income countries are the most affected, three quarters of the world's deaths from CVDs occurring in low- and middle-income countries.

The diagnosis of heart disease includes medical and family history, checking blood pressure, cholesterol level, and lifestyle. In addition the patient may be referred for further tests to help confirm CHD, including electrocardiogram (ECG), exercise stress tests, X-rays, Echocardiogram, blood tests, coronary angiography, radionuclide tests, MRI scans, CT scans (UKNHS, 2020). Heart disease treated with a combination of changing lifestyle, medications and, in some cases, surgery (NHS, 2020). The cost of treating heart disease is very costly (Liu et al., 2002; Chang et.al, 2012).

To reduce this impact of CVD many researches had undertaken, some by using Machine Learning. Machine Learning is a subset of Artificial Intelligence that enables a system to learn from data rather than through explicit programming (Judith & Daniel, 2018) it became as one of the mainstays of information technology. With increasing amounts of data being available in many fields, there is good chance that data analysis will become a necessary ingredient for technological progress (Alex & Vishwanathan, 2008).

ML uses a variety of algorithms, the algorithms will iteratively learn from data to improve, describe and predict outcomes. As the algorithms take training data, it is possible to produce more precise model. A machine-learning model is a file or an output generated when you train your ML algorithm with data. After training, when you provide a model with an input then you will be given an output (Judith & Daniel, 2018).

The adoption of data driven machine-learning methods can be found throughout science, and technology, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modelling, marketing, and policing (Jordan & Mitchell, 2015).

These days, ML plays a key role in many health sectors, including the development of new medical procedures, the handling of patient data and the treatment of chronic diseases (Mike, 2019). Healthcare providers began using tools and technologies that are built from ML models that use anomaly detection algorithms to predict heart attacks, strokes, sepsis and other serious complications. These tools use data from patients' medical records, daily evaluations, and measurements of vital signs, such as heart rate, cholesterol and blood pressure, to alert staff of imminent patient risks so they can immediately take preventive actions (Dale, 2019).

During the last decades, many researchers have conducted studies on heart disease diagnosis, by using machine learning algorithms that includes Support vector Machine, Naive Bayes Classifier, Nearest Neighbour, Logistic Regression Linear Regression, Decision tree, Random forest, and Artificial Neural Network, and ensemble learning (by combining these algorithms), but heart disease remains as a top agenda on public health issue. Hence, this study aims to provide a new insight by combine Self-Organizing Neural

Network (SONN) with Support Vector Machine (SVM) on different kernel functions to diagnose heart disease.

1.2. Statement of Problem

Nowadays, cardiovascular disease (CVD) is a ubiquitous cause of morbidity and mortality in most countries. According to World Health Organization cardiovascular diseases are the number one cause of death globally, taking an estimated 17.9 million lives every year, which represents 31% of all deaths (WHO, 2019).

The poorest people in low and middle-income countries are the most affected. At least three quarters of the world's deaths from CVDs occur in low and middle income countries (WHO, 2017). On top of this, the disease affects the working age groups, which leads to a large economic impact on developing countries.

Besides that, treating heart attack is very costly. For example, in 1990 roughly 25% of South African healthcare expenditures were devoted to the treatment of CVD, in Korea, national spending on CHD was $2.52 billion in 2005 and in UK coronary heart disease cost £1.73 billion in 1999 (Thomas et.al, 2005; Chang et.al, 2012; Liu et al., 2002). Thus, people in low- and middle-income countries who suffer from CVDs have less access to health care services, since the treatment of the disease is costly.

Most researches indicate that earlier detection and primary prevention of CVD reduces the risk of heart attack like myocardial infarction (MI) and heart failure (Tamam & Yasmine, 2019). A research indicates that up to 45% of patients who are admitted to a hospital with heart failure die within 1 year of admission and the majority die within 5 years of admission (Ponikowski et al., 2014).

The prevalence of all CVD had projected to increase in many countries by 2030, for example in US CVD had projected to increase by 40.5% and in China by more than 50% (Heidenreich et al., 2011; Moran et al., 2011). Similarly, the expense of treating CVD had projected to increase by triple in United State between 2010 and 2030, from $272.5 billion to $818.1 billion (Heidenreich et al., 2010).

During the last decades, many researchers have conducted studies on heart disease classification and prognosis using different machine learning algorithms. Still heart disease remains as the number one cause of death and projected to continue likewise. SVMs classifiers prove to be quite popular and successful. However, any of the researchers did not conduct heart disease prediction by combining SONN and SVM algorithms using different kernel functions. Hence, it is very essential to conduct a study that can bring new insight and means to diagnose heart disease. Accordingly, this study aims to combine SONN and SVM algorithms on different kernel functions to diagnose heart disease.

1.3. Research Questions

The study will strive to answer the following research questions:-

i. Which SVM kernel function has better accuracy to diagnose heart disease?
ii. What are risk factors for heart disease prediction?
iii. Which SVM kernel function utilizes SONN's output to perform better heart disease classification?

1.4. Objective

The general and specific objectives of this study are:

1.4.1 General Objective

The general objective of this research is to combine SONN and SVM algorithms on different kernel functions to classify heart disease.

1.4.2 Specific objective

In order to achieve the general objectives, the following specific objectives had carried out

i. Process the dataset to be suitable for the experimentation

ii. Examine the performance of SVM on different kernel function

iii. Design and develop a model to combine SONN and SVM

iv. Examine ways that can increase the performance of the combined model

1.5. Scope and limitation of the Study

This study deals with Heart Disease prediction and includes all age and sex groups. Due to time and budget constraints, the researcher has limited the scope of the research on combining two algorithms only. Besides, the researcher has not collected the datasets and the study has limitation in this regard, which may hinder the investigation of another attributes that can determine heart disease.

1.6. Significance of the Study

People in low- and middle-income countries who suffer from CVDs have less access to health care services, since the treatment of the disease is costly. Thus, this study will help to minimize the burden of physicians in making decisions during diagnosing patients. Furthermore, the output of the study will also give valuable information to researchers who want to conduct research on predicting heart disease. In addition, the result of this study can be additional input for policy makers to devised alternative means to diagnose Heart Disease and investigation.

1.7. Organization of the Thesis

The thesis has organized into five chapters. The first chapter is introduction that consists background of the study, statement of the problem, research questions, objective of the study, scope and limitation of the study, and significance of the study.

The second chapter covers reviews of related literatures that helps to understand basic concepts related to Heart Disease, approaches to heart disease prediction and machine learning techniques, and related works. The third chapter covers the methodology and the techniques followed in this research. It discusses data preparation for the experiment and

models that are implemented .along with the performance evaluation on the combined model.

The fourth chapter covers experimental results and discusses. In this chapter, different experiments had conducted using the selected algorithms and their corresponding interpretations together with the performance of the developed models and results. Finally, the fifth chapter provides conclusions and recommendations based on the findings.

Chapter Two

2. Literature Review

2.1. Heart Disease

2.1.1. Overview of Heart Disease

The term “Cardiovascular diseases”, “cardiac disease” or “heart disease” are a group of malfunctions of the heart and blood vessels. It includes coronary heart disease, cerebrovascular disease, Congenital Heart Disease, rheumatic heart disease, Heart Valve Disease and other conditions. The most common type of heart disease is coronary artery disease, which affects the normal blood flow to the heart which can cause a heart attack (WHO, 2007; CDC, 2020). From all CVD deaths almost 80% of it is due to heart attacks and strokes (WHO, 2007). A heart attack is the death of a segment of heart muscles, which is caused by a loss of blood supply. The blood cut off when an artery supplying the heart muscle had blocked by a blood clot (Justin & Tim, 2017).

According to WHO (2019), CVDs are the number one cause of death globally, taking an estimated 17.9 million lives each year. Of an estimated 58 million deaths globally from all causes in 2005, CVD accounted for 31% which is almost 18 million deaths. This is equal to the death caused due to infectious diseases, nutritional deficiencies, and maternal and perinatal conditions combined. It is important to recognize that almost half of these deaths (46%) occur on individuals whose age is under 70 years, which is the most productive period of life.

Furthermore, studies show that heart disease mortality differs in geographic disparities (Michael et al, 2016). This is due to the risk factors disparity occurring in different location. There are a number of underlying determinants of CVDs or "the causes of the causes", these includes urbanization, population ageing, poverty, air pollution stress and hereditary factors (WHO, 2017).

Different researches show that the rates of cardiovascular disease are falling in wealthier countries, during the second half of the 20th century deaths from diseases such as heart

attacks and strokes fell by 50%. In most developed countries like the US, UK and Western Europe deaths related to CVD's fell by 80%, but this is not a global phenomenon. The findings, of World Health Organization and World Heart Federation, showed that in several low and middle-income countries the epidemic is still developing (Theconversation, 2014). The poorest people in low- and middle-income countries are the most affected, at least three quarters of the world's deaths from CVDs occur in low- and middle-income countries (WHO, 2017).

With the aging population, the prevalence of all CVD had projected to increase. It is projected that by 2030, 40.5% of the US population would have some form of CVD (Heidenreich et al., 2011). Similarly, in China cardiovascular disease in adult's ages 35 to 84 years projected to increase by 50% between 2010 and 2030 based on population aging and growth alone. (Moran et al., 2011)

High blood cholesterol, High blood pressure, and smoking are indicated as a key risk factors for heart disease (Virani et al., 2020). Several other medical conditions and lifestyle choices can also put people at a higher risk for heart disease, including Age, Obesity, Diabetes, Metabolic syndrome, Family history of heart attack, Lack of physical activity, Stress, Illicit drug use, A history of preeclampsia, and An autoimmune condition (Thomas, 2005).

2.1.2. Diagnosis of Heart Disease

Coronary heart disease (CHD) had usually diagnosed after a risk assessment and some further tests. If a Physician thinks you may be at risk of CHD, they may do a risk assessment for cardiovascular disease, heart attack or stroke (UKNHS, 2020).

The diagnosis include medical and family history, checking blood pressure, and a blood test to assess cholesterol level, lifestyle (how much exercise you do and whether you smoke). All these factors considered as part of the diagnosis. In addition the patient may be referred for further tests to help confirm CHD, including electrocardiogram (ECG), exercise stress tests, X-rays, Echocardiogram, blood tests, coronary angiography, radionuclide tests, MRI scans, CT scans (UKNHS, 2020). The commonest cause in

coronary heart disease is past heart attack (myocardial infarction) which is responsible for around half of all new cases of heart failure (Petersen et al., 2002).

Physicians also use ACC/AHA (the American College of Cardiology (ACC) and the American Heart Association (AHA)) to estimate the 10-year risk of developing a first atherosclerotic cardiovascular disease (ASCVD) event, which was defined as nonfatal MI, coronary heart disease (CHD) death, or fatal or nonfatal stroke, in individuals who were initially free from ASCVD (Tamam & Yasmine, 2019).

ACC/AHA Risk Calculator uses ASCVD Risk Algorithm, this algorithm uses the 10 risk factors to asses risk: age, gender, race, total cholesterol (mg/dL), high-density lipoprotein cholesterol (mg/dL), systolic blood pressure (mmHg), Diastolic blood pressure (mmHg), treatment for high blood pressure, diabetes, and smoking (American College of Cardiology, 2020).

2.1.3. Treatment of Heart Disease

Heart disease can be treated in several ways. The aims of treatment are to reduce symptoms and delay progression of the disease, reduce hospitalisation, and extend and improve the quality of life (Petersen et al., 2002). Cardiac rehabilitation helps to prevent another heart problems and it can also help patients to form a healthy living habits. Anyone who has had a heart problem previously, such as a heart attack, heart failure, stroke or heart surgery, can benefit from cardiac rehabilitation. (CDC, 2021)

Heart disease can effectively treat with a combination of changing lifestyle, medications and, in some cases, surgery. Cardiac rehabilitation is an important for anyone who is recovering from any heart related diseases like heart failure, heart attack, stroke, or any type which require surgery. According to NHS (2020) cardiac rehab is a supervised program that includes the following:-

- Physical activity
- Education on healthy living habits, which includes healthy eating, taking medicine as prescribed, physical activity and ways to help you quit smoking

- Helping patients to find ways to minimize stress and improve mental health

In United Kingdom (UK) coronary heart disease cost £2.42 billion in informal care out of which 24.1% of production losses were attributable to mortality and 75.9% due to morbidity. The total annual cost of all coronary heart disease was, the highest of all diseases in which comparable analyses had done (Liu et al., 2002).

Similarly, in South Korea, estimated national spending on CHD in 2005 was 2.52 billion USD. Most of the cost was attributable to medical, followed by productivity loss due to morbidity and premature death, transportation, and informal caregiver costs. Annual per-capita cost of treating MI, excluding premature death cost, was $3.2 million, which is about 2 times higher than the cost for angina, which is 1.6 million USD (Chang et al., 2012).

In addition, in United States the total direct medical costs of CVD projected to triple, between the year 2010 and 2030, from $272.5 billion to $818.1 billion (Heidenreich et al., 2010).

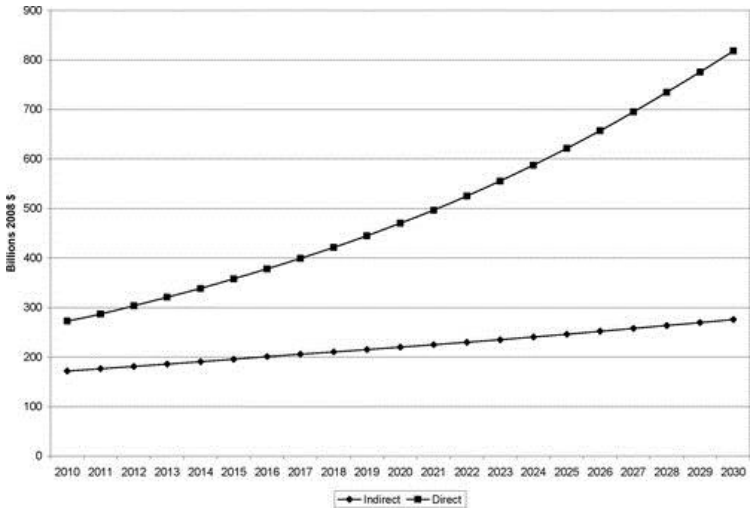

Figure 1: Projected direct and indirect costs of all CVD, 2010 to 2030 (Heidenreich et al., 2010)

2.2. Machine learning in Health care

Machine learning is one of today's most rapidly growing technical fields, lying at the intersection of computer science, artificial intelligence, data science and statistics. It addresses the question of how to build computers that can improve automatically through experience without human intervention (Jordan & Mitchell, 2015).

Current progress in technologies and ML has been driven both by the development of new learning algorithms and theory and by the on-going explosion in the availability of online data and low-cost computation. The adoption of ML can be found throughout technology, science, commerce and others, leading us to more evidence-based decision-making across many fields, including health care, education, manufacturing, finance, policing, and marketing (Jordan & Mitchell, 2015).

The increased availability of electronic health data processing had created major opportunity in healthcare for both discovery and practical applications to improve healthcare. It can result in advances on disease diagnosis, treatment, drug discovery, remote healthcare monitoring, discover patterns from medical data sources and reduction in healthcare costs and it can can predict heart attacks, strokes, sepsis and other serious complications (Saleem & Chishti, 2020; Shailaja et al., 2018; Dale, 2019).

These tools use data from patients' medical history, daily evaluations, and measurements of vital signs in real-time, such as heart rate and blood pressure, to alert medical workers of imminent patient risks so they can take preventive actions that can save lives (Dale, 2019).

According to Statnikov (2011), we can also use machine-learning algorithms in health care to:

- Build classification models that assign patients/samples into two or more classes. it can be used for outcome prediction, diagnosis, and other classification tasks
- Use regression models to predict survival, length of stay in the hospital, laboratory test values, etc.

- Identify novel or outlier patients/samples. Such models used to discover deviations in sample handling protocol when doing quality control of assays, etc.

- Group samples/patients into a number of clusters based on their similarity of features. These methods helps to discovery disease sub-types and for other tasks.

2.3. Machine Learning Algorithms

ML uses a variety of algorithms that learn from data to describe data, and predict outcomes. As the algorithm takes training data, it is possible to produce a model that can precisely predict. ML is now essential for creating analytics models (Judith & Daniel, 2018). An algorithm is a step of procedure taken to accomplish a specific task, it is the idea behind any reasonable computer program; it must solve a general and well-specified problems (Skiena, 2020). Algorithms allow computers to learn automatically without any interventions or assistance from humans, computers can adjust actions accordingly (Expert System, 2017).

Machine learning algorithms often categorized:

- **Supervised machine learning algorithms:** In machine learning, classification is a supervised learning approach (Mandy, 2017). It can learn from past data that are labelled to predict future events. (Expert System, 2017). Some examples of supervised learning could be speech recognition, handwriting recognition, biometric identification, document classification, heart disease prediction, etc. (Mandy, 2017).

- **Unsupervised machine learning algorithms:** is when the data we used to train is neither classified nor labelled. Unsupervised learning identifies pattern in the data to describe a hidden structure. This kind of learning does not figure out the right output, but it explores the data to draw inferences from datasets and can automatically identify structure in data (Expert System, 2017).

- **Semi-supervised machine learning algorithms** this kind of algorithms fall between supervised and unsupervised learning, since they use both labelled and

unlabelled data for training, usually it uses dataset that has small amount of labelled and a large amount of unlabelled data. We can use this kind of algorithms when the acquired labelled data requires skilled and relevant resources in order to learn from it. Otherwise, acquiring unlabelled data generally does not require additional resources (Expert System, 2017).

2.3.1. Self-Organizing Neural Network (SONN)

The Self-Organizing Neural Network (SONN) or Self Organizing Map (SOM) is a special case of neural network algorithm without any hidden layer and it uses a competitive learning technique to train itself in an unsupervised manner (Pohl et al., 2012).

SONNs are different from other artificial neural networks in that they use a neighbourhood function to preserve the topological properties of the input space and they had used to create a cluster representation of multi-dimensional inputs, which simplifies complexity and reveals meaningful relationships (George, 2010).

The SONN algorithm has a special property of creating spatially organized of various features of each input signals and their abstractions better than other Artificial Neural Networks (Kohonen, 1990). It transfers complex high-dimensional inputs into simpler forms in lower dimensions (usually two dimensions) by observing relationships between data (Widiyaningtyas et al., 2019).

It is consisted of an array of neurons or cells, which arranged on two dimensional or rectangular (or hexagonal) sheets (Xuegong & Yanda, 1993). The SONN models are associated with the neurons or nodes of usually two-dimensional grid. The SONN algorithm constructs the mapping such that, more similar input vectors will be clustered to nodes that are closer in the grid, whereas less similar models will be located farther away in the grid (Kohonen, 2013). The basic structure of self-organizing map network shown in Figure 2

The cells with single index i. The input vector $X(t) = [x_1(t), x_2(t) \dots x_n(t)]^T \in R^n$, is connected in parallel to all the cells, but via different weight vectors $m_i(t)=[M_{i1}(t),$

$M_{i2}(t)\ldots M_{in}(t)\ ]^T \in R^n$, which are adapted according to the input data set during the self-organizing learning procedure at a given time t (Xuegong & Yanda,1993).

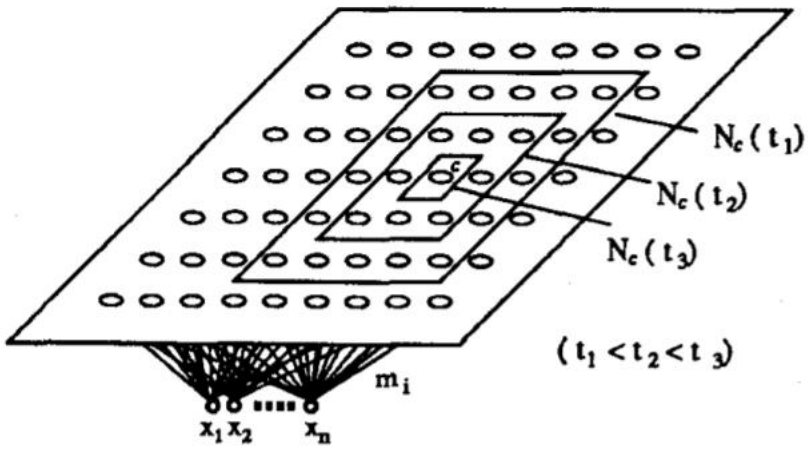

Figure 2: The structure of SOM network (Xuegong & Yanda, 1993).

In the learning procedure, each m_i first initialized with some small random values. Then the data to be analysed are presented as input vectors repeatedly in their original order or some random order. Each time an input X (t) is presented, we find among all the cells the best-matching cell c defined as the one that

$$\|x - m_c\| = \min \|x - mi\|$$

Where the min ||.|| is a norm of the Euclidian distance or some other distance measurement. Around this cell, we define a neighbourhood $N_c(t)$ as the range of lateral interaction, which is also illustrated in Figure 2. The basic update on weight-learning or weight-adapting process carried out by the following equation:

$$M_i(t+1) = \begin{cases} M_i(t) + \propto (t)[x(t) - M_i(t)], & i \in N_c(t)] \\ M_i(t), & i \notin N_c(t) \end{cases}$$

Where $0 < \propto (t) < 1$a scalar factor that controls the learning rate and it is decreases with as time goes by

The result of this lateral interaction is that after adequate iterations, the network will tend to be "spatially organized" according to the structure of the input data set that was feed to it. The cells will become tuned to specific input vectors or groups of them, that is, each cell will respond only to some specific patterns in the input pattern set (if it responds to any of them). And the locations of the cells responding to different inputs tend to be ordered in accordance with the topological relations among the patterns in the input set. In this way, the topological relationship in the original space is optimally preserved on the map. This property is why the algorithm named as self-organizing map and is what makes the network quite powerful in certain applications (Xuegong & Yanda, 1993).

SONN is widely applied for clustering problems, clustering means partitioning a data into a set of clusters. (Kohonen, 2013; Vesanto & Alhoniemi, 2000) Clustering is a form of unsupervised learning that can be used to discover useful patterns in data (Talavera, 1999).

The SONN has been used extensively as a visualization tool in exploratory data analysis. It is widely used in many application domains, such as economy, industry, management, sociology, geography, text mining, and management of very large document collections etc. New, promising applications exist in bioinformatics (Marie et al., 2018; Kohonen, 2013). In addition to these, one may mention a few specific applications, e.g., profiling of the behaviour of criminals, categorization of galaxies, categorization of real estates, etc. (Kohonen, 2013).

2.3.2. Support Vector Machine (SVM)

In machine learning, support vector machines (SVM) also known as support vector networks are a type of supervised learning models with associated learning algorithms that analyse data, it is used for classification and regression analysis (Alexandre, 2017), but it were initially developed for classification and have been extended for regression (Yu & Kim, 2012).

SVMs can be utilized in applications like Handwriting recognition, Intrusion detection, Cancer Diagnosis and Prognosis, Face detection, Email classification, and Gene classification (Saumya, 2020). It is an effective classification method with significant

advantages such as the absence of local minima, an adequate generalization to new objects, and a representation that depends on few parameters (Maldonado & Weber, 2009).

The objective of SVM is to find a hyper plane in a multi-dimensional space (the number of features) that distinctly classifies the data points. Many possible hyperplanes could be drown (Rohith, 2018)

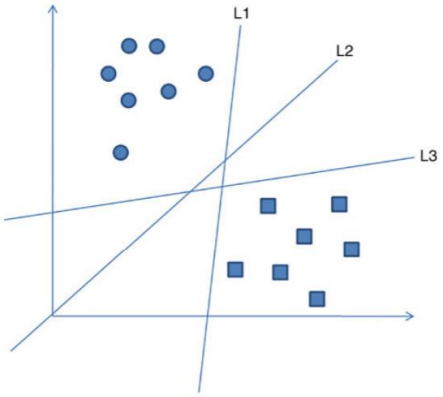

Figure 3: Linear classifiers (hyper plane) in two-dimensional spaces (Yu & Kim, 2012).

The hyper plane must have a wider margin that has a maximum distance between support vectors of both classes. If the hyper plane's margin is maximized then future data points could classified with more confidence (Rohith, 2018).

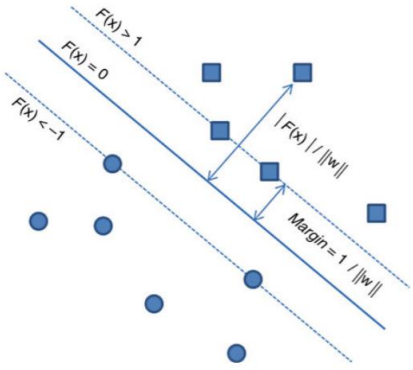

Figure 4: SVM classification functions: the hyper plane maximizing the margin in a two-dimensional space (Yu & Kim, 2012).

SVM algorithms use a group of mathematical functions that had known as kernels (Saumya, 2020). A kernel function is a method used in SVM for helping to solve binary classification problems. They provide shortcuts to avoid complex calculations. The advantage of using kernel is that we can go from lower to higher dimensions and perform smooth calculations with the help of it (TechVidvan, 2021).

The main limitation of SVM lies in the choice of the kernel function, and the classification accuracy rate of SVM is highly influenced by the Kernel function we used (Patle & Chouhan, 2013).

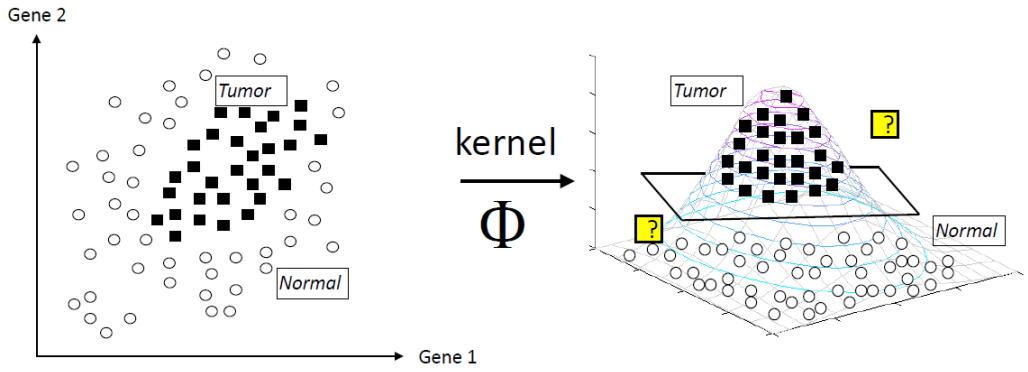

Figure 5: Kernel trick transforms not linearly separable data into higher feature space (Statnikov, 2011)

As shown in the figure 5, kernel trick helps to transform Data that is not linearly separable in the input space into a Data that is linearly separable in the feature space.

The kernel function takes data as input and manipulates it to transform it into the desired form. Different SVM uses differing kinds of kernel functions. These functions are of different kinds, the most widely used are linear, radial basis function (RBF), polynomial, and sigmoid (Saumya, 2020). They are different in case of making the hyper plane decision boundary between the classes

Linear kernel: This kernel function is used to classify linearly separable data (Goel & Srivastav, 2016). It is regarded as the most straightforward kernel function. The inner

product of $K(X_i, X_j) = (X_i \bullet X_j)$ plus an optional constant c represents linear kernel (Al-Mejibli et al., 2018).

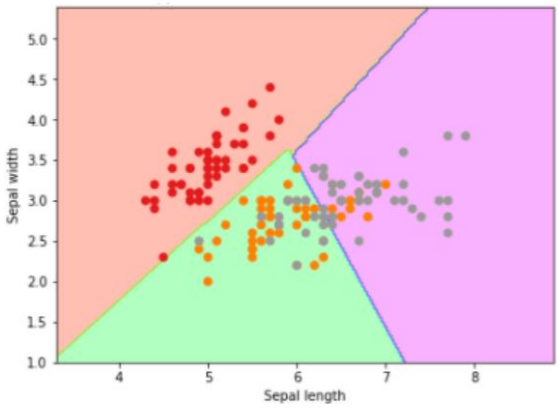

Figure 6: Support vector Machine with linear kernel (Saumya, 2020)

Linear kernel is one of the most common kernels that prove to be the best function when there are large datasets with lots of features. The linear kernel consumes less time for computation and it is mostly preferred for text-classification problems since such kind of classification problems can be separated linearly (Saumya, 2020).

Polynomial kernel: Sometimes the data may not be linear separable due to noise in data or bad feature representation. The solution for this problem is to map data into different space in such a way that in the new feature space these samples would be linearly separable. Polynomial kernel is one of the kernel functions used for non-linear separable and it's applied by using mathematical function $K(X_i, X_j) = (\gamma X_i \bullet X_j + C)^{d}$ (Al-Mejibli et al., 2018).

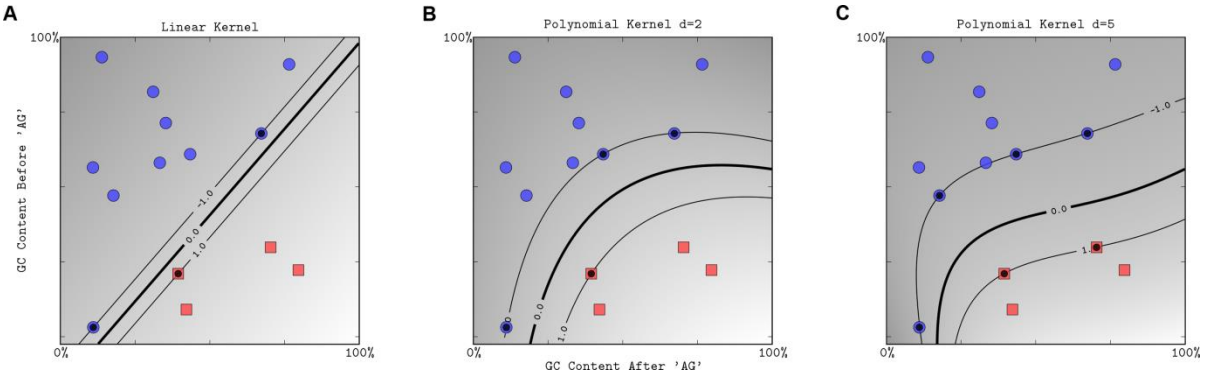

Figure 7: The effect of the degree of a polynomial kernel (Ben-Hur et al., 2008)

As Shown in figure 7, the polynomial kernel of degree 1 means the data is linearly separable using a straight line and polynomial kernels of higher than 1 degree allow a more flexible decision.

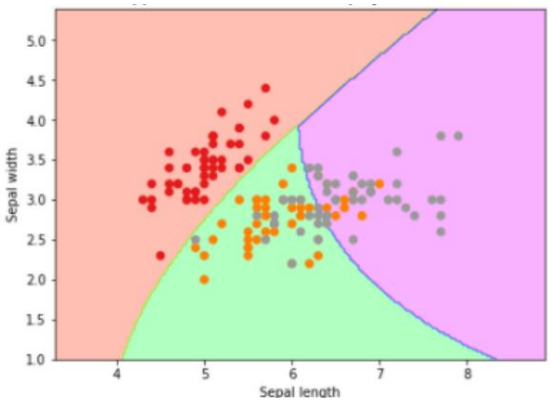

Figure 8: Support vector Machine with polynomial kernel of degree three (Saumya, 2020)

Polynomial Kernel is a more generalized representation of the linear kernel and it gives good results for problems where all the training data is normalized (Saumya, 2020). It is widely used in natural language processing (NLP).

Radial Base Function (RBF) kernel: The Radial basis function (RBF) kernel, also known Gaussian kernel, is represented in equation $K(X_i , X_j) = \exp(-\gamma \|X_i - X_j\|^2)$ Where γ is a parameter that used to maximize the kernel (Al-Mejibli et al., 2018).

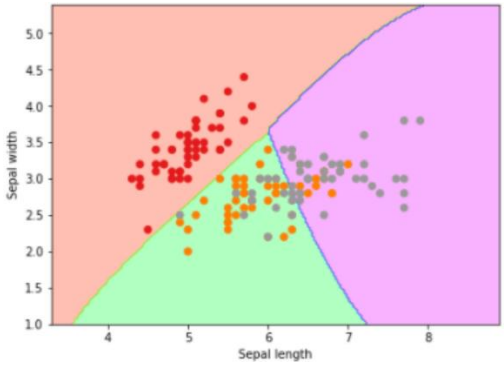

Figure 9: Support vector Machine with RBF kernel (Saumya, 2020)

RBF kernel is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data (Saumya, 2020).

Sigmoid Kernel: The sigmoid kernel equation is represented as $K(X_i , X_j)= \tanh (\gamma X_i \bullet X_j + C)$ Where γ is a scaling parameter and C represents the shift parameter that used to control the threshold of mapping of the input data (Al-Mejibli, 2018).

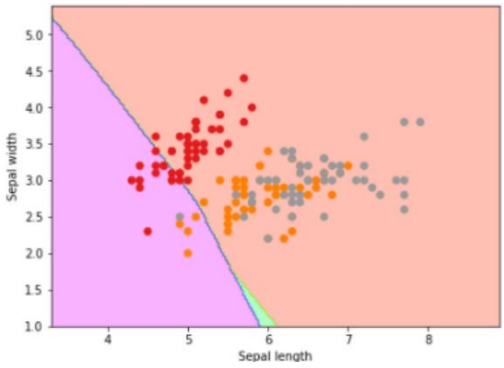

Figure 10: Support vector Machine with Sigmoid kernel (Saumya, 2020)

Sigmoid is mostly preferred for neural networks, it is similar to a two-layer perceptron model of the neural network, which works as an activation function for neurons (Saumya, 2020).

2.3.3. Other algorithms for Heart Disease Classification

- **Naive Bayes Classifier (Generative Learning Model)**: It is a family of simple "probabilistic classifiers" based on Bayes' Theorem with an assumption of independence among predictors. It assumes that no pair of features is dependent on each other or the presence of a particular feature in a class is unrelated to the presence of any other feature even if these features depend on each other or other features. Naive Bayes model is particularly useful for very large data sets and is easy to build. It is known to outperform even highly sophisticated classification algorithms (Mandy, 2017).

- **K-Nearest Neighbour:** is a classification algorithm that takes a bunch of labelled points and uses them to learn how to label other. The "k" represents the number of neighbours it checks to label a new point. First it looks at the labelled points closest to that new point (its nearest neighbours), and has the neighbours vote for

it, so that whichever label the most of the neighbours have is the label for the new point (Mandy, 2017).

- **Logistic Regression (Predictive Learning Model):** It is a statistical model for analysing a data set when one or more independent variables that determine an outcome. It is supervised classification algorithm. The output is measured with a variable that has two possible outcomes only (dichotomous variable). The goal of logistic regression is to estimate the probabilities of events to describe the relationship between the dichotomous characteristic of interest and a set of independent variables. It is better than other binary classification like nearest neighbour because it explains the outcome quantitatively (Mandy, 2017).

- **Linear Regression**: is a type of Supervised Learning is the most basic type of regression. Simple linear regression allows us to understand the relationships between two continuous variables by using a simple equation known as the regression equation which is an average value of Y. Linear regression is often used for the resolution of classification problems. It can be used with as an hybrid with other machine learning algorithms like decision tree to classify heart disease (Srinivas et al., 2020).

- **Decision Trees:** Decision tree builds classification or regression tree in the form of a tree-like model of decisions. It breaks down a data set into smaller subsets while the associated decision tree is incrementally developed, it will finally give us tree structure with decision nodes and leaf nodes. A decision node has two or more branches represent an outcome of the test; the leaf node represents a class label or decision. The topmost decision node in a tree, which corresponds to the note that starts the graph, called root node. Decision trees can solve both categorical and numerical data (Mandy, 2017).

- **Random Forest:** is an ensemble model that used for classification and regression problems, which operate by constructing a multitude of decision trees at training time. Random decision forests use many decision trees that correct decision trees' habit of over fitting to their training set (Mandy, 2017).

- **Neural Network:** Artificial Neural Network is a variety of deep learning technology, neural network consists of neurons, arranged in layers, which receives an input and delivers an output. Each neuron takes an input, and then it passes the output on to the next layer by applying a function to it. Each node weights the importance of input data from each of its predecessors, and it is these weightings, which are tuned in the training phase to adapt a neural network to a particular problem (Mandy, 2017).

2.4. Related works on Heart Disease prediction

Mahmoodi, (2017) used a data of 270 people with 13 features, as a Machine Learning classifier. The researcher used a fuzzy system and support vector by using MATLAB software and simulated by a system of core i5 and windows7, as the operating system, to diagnose patients with heart disease. The results show that the accuracy and sensitivity of these two indicators were 85% and 85.8% respectively.

Alty et al., (2003), conducted a study aimed to develop a method for rapidly assessing a patient's arterial stiffness and risk of developing cardiovascular disease (CVD) without resorting to laborious blood tests. They used simple measurement of a patient's volume pulse measured at the fingertip (digital volume pulse) using an infrared light absorption detector placed on the index finger is sufficient to predict their CVD risk. Support vector machine (SVM) classifier has been found to make 85% prediction accuracy.

Pooya et al. (2010) used support vector machines (SVMs) by modifying the kernel width by using trial-and-error approach to predict the existence of heart disease. They used existing medical and demographic data of the patients as and input for classifying the heart condition of the subject as normal or abnormal. They identified the significance of using the categorical feature-based classification method. Their result shows that SVM can classify existence of heart disease with 84% accuracy.

Tsehay, (2019) in this research the researcher used Kaggle heart disease dataset and for heart disease classification purpose they used support vector machine-learning algorithm.

Using this algorithm the researcher found out 73.41% accuracy on heart disease classification.

Ketut et al., (2016) have used KNN algorithm with parameter weighting method. Out of 13 parameters, they only use eight parameters. The result shows that the accuracy KNN algorithm has improved when they used only 8 parameters, comparing to 13 parameters. They also indicate that KNN algorithms showed better performance than Naive Bayes and Decision Tree.

Aniruddha et al., (2019) used a data which is obtained from the National Health and Nutritional Examination Survey (NHANES). To predicting the existence of Coronary Heart Disease (CHD) they used a two-layer CNN and for feature selection they used least absolute shrinkage and selection operator (LASSO). Using these they found out that CNN can identify the presence and absence of CHD with 77% and 81.8% respectively, and the average accuracy was 79.5%.

Widiyaningtyas et al. (2019) they used Self-Organizing Map (SOM) which is unsupervised algorithm for the classification of Coronary Heart Disease. The results showed that the most optimal level of accuracy in the data comparison of testing and training is 20-80 ratios. Using this they obtained 62.5% accuracy, 60.33% precision, 63.33% recall and the error rate was 37.5%.

To the knowledge of the researcher, the only other research that tries to combine the SONN and SVM is a research conducted by Nilashi et al., this research uses on Principal Component Analysis (PCA) for feature selection and as an algorithm they used Self-Organizing Map, Fuzzy Support Vector Machine (Fuzzy SVM). And to handle missing values they used two imputation techniques. The researchers used incremental PCA and Fuzzy SVM so that incremental learning of the data can reduce the computation time of classification. They used Cleveland and Statlog datasets, for analysis. Their result showed that the use of incremental Fuzzy SVM has improved the accuracy of heart disease classification.

Table 1: Summary of Related Works

Title	Algorithm	Finding	Gap
Designing a heart disease prediction system using support vector machine (2017)	fuzzy system and support vector machine	85% accuracy and 85.8 Sensitivity	Didn't use kernel function and dimensionality reduction
Cardiovascular disease prediction using support vector machines (2003)	They used Simple measurement of a patient's volume pulse measured at the finger-tip (digital volume pulse) using an infrared light. Then to classify they used support vector machine	accuracy 85%	None
A Support Vector Machine Approach for Predicting Heart Conditions (2010)	support vector machine, by modifying the kernel width by using trial-and-error	accuracy 84%	They used Gaussian kernel only
A Support Vector Machine Based Heart Disease Prediction, (2019)	support vector machine	accuracy 73.41%	They didn't use any kernel function
Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters (2016)	K-Nearest Neighbor with parameter weighting method	accuracy 81.85%	They dropped attributes without any tool or experimentation
An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction (2019)	two-layer Convolutional Neural Network and Least Absolute Shrinkage and Selection Operator (LASSO) regression	accuracy 79.5%	None
Self-Organizing Map (SOM) For Diagnosis Coronary Heart Disease (2019)	Self-Organizing Map	accuracy 62.5%	None
Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates, (2020)	Principal Component Analysis (PCA), Self-Organizing Map, Fuzzy Support Vector Machine	86.61 % accuracy, 81.31 specificity and 86.93 sensitivity	They used RBF kernel function only

From the literature, it is evident that different machine learning approaches had been used to investigating heart disease prediction. SVMs classifiers prove to be quite popular and successful. The use of Convolutional Neural Network and K-Nearest Neighbour are also appearing to be a popular choice for achieving good performance. However, any of the researchers did not conduct heart disease prediction by combining SONN and SVM algorithms using different kernel functions. Hence, this research aims at diagnosing heart disease by combining SONN and SVM algorithms on different kernel functions.

Chapter Three

3. Methodology

This chapter presents the methodology that we have used to address the research problems and to achieve the research objective. It shows the source of the dataset, how we pre-processing the dataset, designing a model to combine the two algorithms, feature selection, tools we use to process the data, and how we conduct performance evaluation on each algorithm.

3.1. Research Type

The type of this research is experimental type, which aimed at developing Heart Disease diagnosis model by combining SONN and SVM algorithms.

3.2. Source of Data

Data sources for such researches can be obtained from Cleveland, Hungarian, Switzerland and Long Beach VA that have dataset as indicated in the table below

Table 2: Heart Disease Datasets

Data Source	Total Dataset	Positive	Negative	Number of missing value
Cleveland	303	139	164	6
Long Beach VA	200	149	51	698
Hungarian	188	106	188	782
Switzerland	123	115	8	273

As can be observed from the table Cleveland have large population and small number of missing values. Thus, the dataset for this research had obtained from Cleveland Clinic Foundation that contains 303 clinical data. This dataset have13 attributes, age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, stress test (ST) depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels coloured by fluoroscopy, and thalassemia.

3.3. Data Pre processing

In Machine Learning, Data pre-processing refers to the technique of cleaning and organizing the raw data to make it suitable for training and building Machine Learning models using different algorithms. A real-world data mostly contains noisy, missing values, and may be in an unusable format. Hence we cannot directly use them for machine learning models. Data pre-processing is required tasks for making it suitable for a machine-learning model, which also increases the accuracy and efficiency of a machine-learning model.

In our dataset when the data is screened, four and two missing values found in the slope and number of major vessels coloured by fluoroscopy respectively. The missing values had handled by using imputation with mean value. In addition, the diagnosed heart disease result contains values 0 for Negative result and 1 up to 4 for Positive result that indicate the severity of the disease. Our experiments on this database are concentrated on separating the cases with heart disease (values 1, 2, 3, 4) from no heart disease cases (value 0). Hence, we convert all severity of heart disease in to value 1.

3.4. Data Processing Tool

Python language was used on PyCharm IDE for the experimentation; we used python because it is one of the most preferred language for scientific computing, data science, and machine learning. It increases performance and productivity by enabling the use of low-level libraries and high-level APIs (Raschka et al., 2020).

3.5. Algorithm

In this research, two algorithms, namely Self-Organizing Neural Network (SONN) and Support Vector Machine (SVM) had combined to classify heart disease.

SVM classification algorithm has been widely applied in bioinformatics (Pavlidis, 2004) similarly SONN has shown promising applications in bioinformatics (Marie et al., 2018; Kohonen, 2013).

3.5.1. SONN ALGORITHM:

To select the number of nodes for the SONN we used Pohl et al. (2012) approach, in their research they used a 3x3 map resulting in 9 clusters. The following steps had used for SONN algorithm

Input: all independent data

Output: cluster of each input

Step 1: Read the Dataset.

Step 2: Split the dataset into dependent and independent dataset

Step 3: Read the Independent Dataset.

Step 4: Cluster each input

3.5.2. SVM algorithm

In our experiment we used a polynomial kernel of degree 3, since having higher degrees more than three tend to over fit. The following steps had used for SVM algorithm

Input: all data

Output: Heart Disease classification.

Step 1: Read the Dataset.

Step 2: Remove some feature by using Backward Feature Elimination

Step 3: Train SVM using different Kernel Functions

Step 4: Test SVM on each Kernel Functions

Step 5: Evaluate model performance

3.5.3. Combined Algorithms

The following steps had used for the combined algorithm

Input: all 13 features.

Output: Prediction of Heart Disease.

Step 1: Read the Dataset.

Step 2: Split the dataset into dependent and independent dataset

Step 3: Read the Independent Dataset.

Step 4: Cluster each input using SONN

Step 5: process the result to the format that could inserted in the input

Step 6: Insert the processed SONN result as a feature in the input variables

Step 7: Train SVM using different Kernel Functions by the updated dataset

Step 8: Test SVM on each Kernel Functions

Step 9: Evaluate model performance

3.5.4. Feature selection

Feature selection is of considerable importance in classification. A dataset that has large number of features impose a high computational cost and a high cost of data acquisition. On the other hand, a low-dimensional feature representation reduces the risk of over fitting (Maldonado & Weber, 2009). Removal of features that contain noise or having no effect

at all can increase the search speed and the accuracy rate. Feature selection is a part of supervised classification (Patle & Chouhan2013).

In this research we used Backward Feature Elimination (BFE). BFE has better computational performance than Forward Feature Selection (FFS) (Khaire & Dhanalakshmi, 2019).

In this technique, the following steps are used:

Step 1: at a given iteration the SVM algorithm had trained on all input features.

Step 2: Then we remove one input feature at a time and train the same model on n-1 input features n times.

Step 3: The input feature whose removal has produced the smallest increase in the error rate had removed, leaving us with n-1 input features.

Step 4: The classification then repeated using n-2 features, and so on

3.5.5. Combined Model

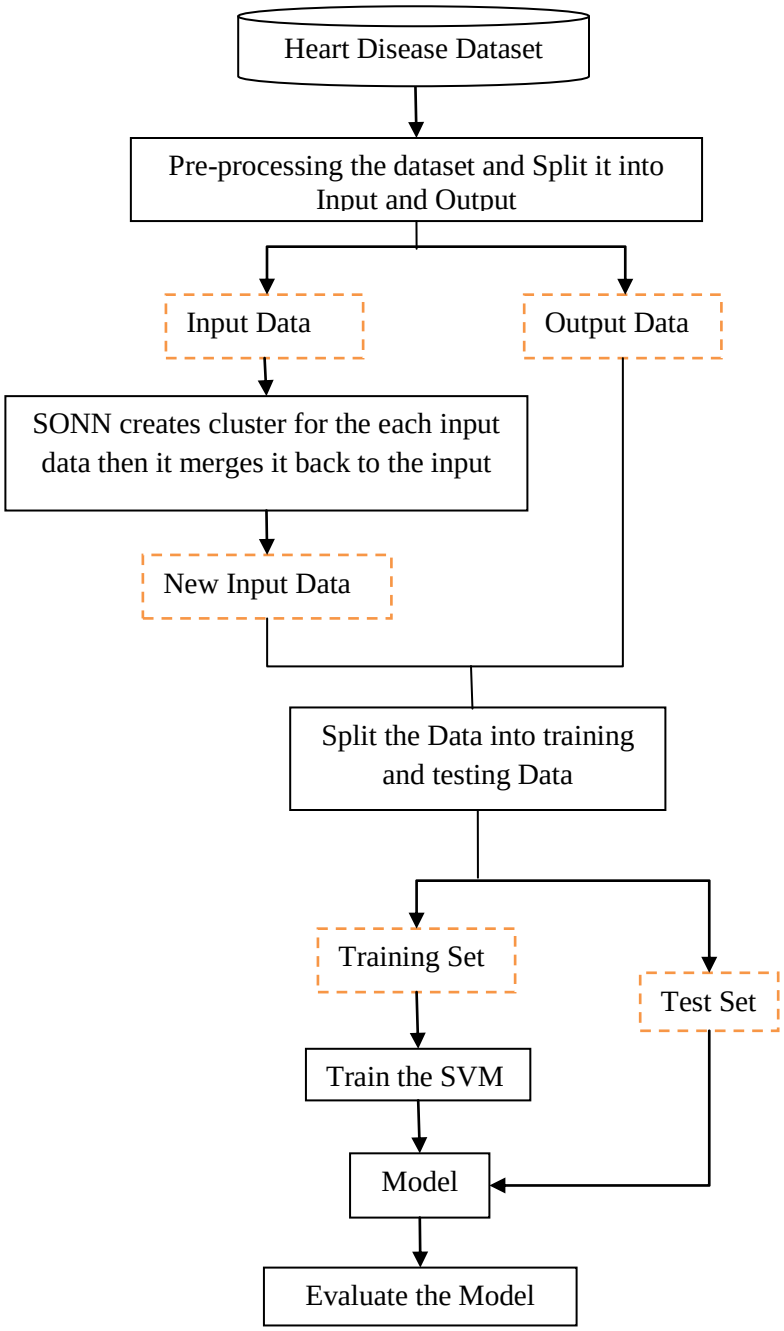

Figure 11: Combined model SONN and SVM

3.6. Evaluation

The predictive capability of the ML models had greatly affected by the training/testing ratios, 75/25 presenting the best performance of the models (Nguyen et al., 2021). Hence, in our experiment we used 75% for training and 25% for testing purposes.

All the algorithms have been evaluated in terms of performance by means of well-known quantitative metrics; accuracy, sensitivity and specificity.

These measurements had calculated using True Positive (TP), False positive (FP), True negative (TN) and False negative (FN). TP means person diagnosed with the disease and tested. The result showed that the person had the disease that goes with the idea of (Goel & Srivastav, 2016). FP subjects without the disease with the value of a parameter of interest above the cut-off (Simundic, 2009), TN subjects without the disease with the value of a parameter of interest below the cut-off (Simundic, 2009) and FN means person actually has the disease but the test is showing that the person doesn't have disease (Goel & Srivastav, 2016).

The measurements had calculated as follows

- **Accuracy**: it defines the actual results (Goel & Srivastav, 2016). It is defined as:
    - Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)
- **Sensitivity**: had expressed in terms percentage and defines the proportion of true positive subjects with the disease in a total group of subjects with the disease. Sensitivity actually, defined as the probability of getting a positive test result in subjects with the disease. Hence, it relates to the potential of a test to recognize subjects with the disease (Simundic, 2009). It is defined as:
    - Sensitivity = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)

- **Specificity**: is a measure of the accuracy of diagnostic test, complementary to sensitivity. It had defined as a proportion of individuals without the disease and with negative test result from the total of subjects without disease (Simundic, 2009). It is defined as:
    - Specificity = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

Chapter Four

4. Experimental result and Discussion

4.1. Results

4.1.1. Introduction

In this section, the results of the experiments have presented based on the model training simulations experimented. The objective of the research was to develop a model to predict Heart Disease by combining SONN and SVM.

The experiment had carried out using 303 data obtained from Cleveland Clinic Foundation. From the 303, 75% had used for training and 25% for testing purposes. Then the experiment had carried out to evaluate the performance of the presented model.

The data contains 14 attributes, Age, Sex, chest pain type, Resting blood pressure, Serum cholesterol, Fasting blood sugar, Resting electrocardiographic results, Maximum heart rate achieved, Exercise induced angina, Stress Test (ST) depression induced by exercise relative to rest, The slope of the peak exercise ST segment, Number of major vessels coloured by fluoroscopy, Thalassemia and Diagnosis of heart disease. The following table shows the description of each attributes

Table 3: Description of the Heart Disease Dataset

Number	Attributes	Full name of attributes	Values
1	Age	Age in years	Continues number
2	Sex	Sex	1 = male; 0 = female
3	Cp	chest pain type	1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic
4	Trestbps	Resting blood pressure	in mm/Hg
5	Chol	Serum cholesterol	in mg/Dl
6	Fbs	Fasting blood sugar	Greater than 120 mg/dL 1 = true; 0 = false
7	Restecg	Resting electrocardiographic results	0, 1, 2

8	Thalach	Maximum heart rate achieved	Continues Number
9	Exang	Exercise induced angina	1 = yes; 0 = no
10	Oldpeak	ST depression induced by exercise relative to rest	Real number
11	Slope	The slope of the peak exercise ST segment	Between 1–3
12	Ca	Number of major vessels coloured by fluoroscopy	Between 0–3
13	Thal	Thalassemia	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Num	Diagnosis of heart disease	0 = absence; 1,2,3,4 = presence

The minimum and the maximum age group of the dataset goes from 29 up to 77 years old, and total number of positive with heart disease is 139 and negative is 164. Among these 97 were female and 206 were male, as shown in Figure 12

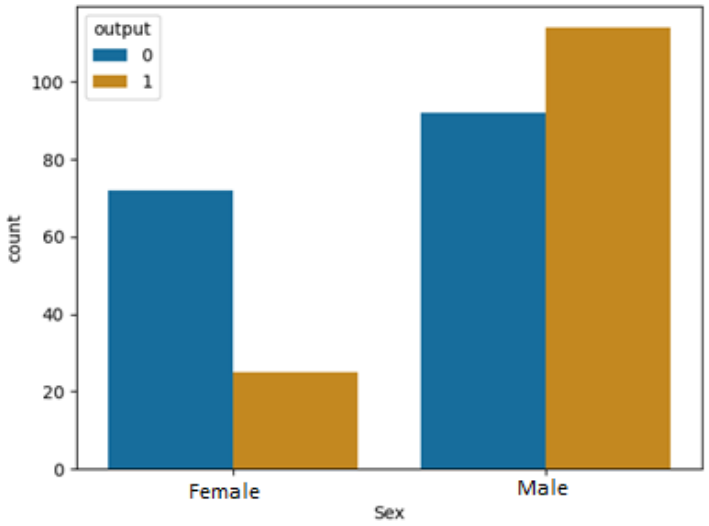

Figure 12: Male and female with positive and negative output

The figure shows that Males have heart disease than Females

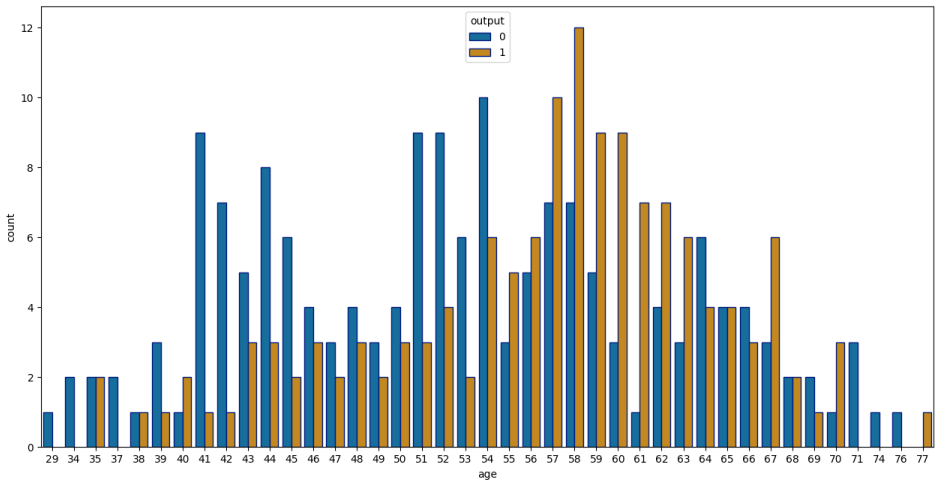

Figure 13: Comparing age with positive and negative output

As shown in figure 13 most of heart disease victims starts from age of 55 and became high at the age of 58 and above

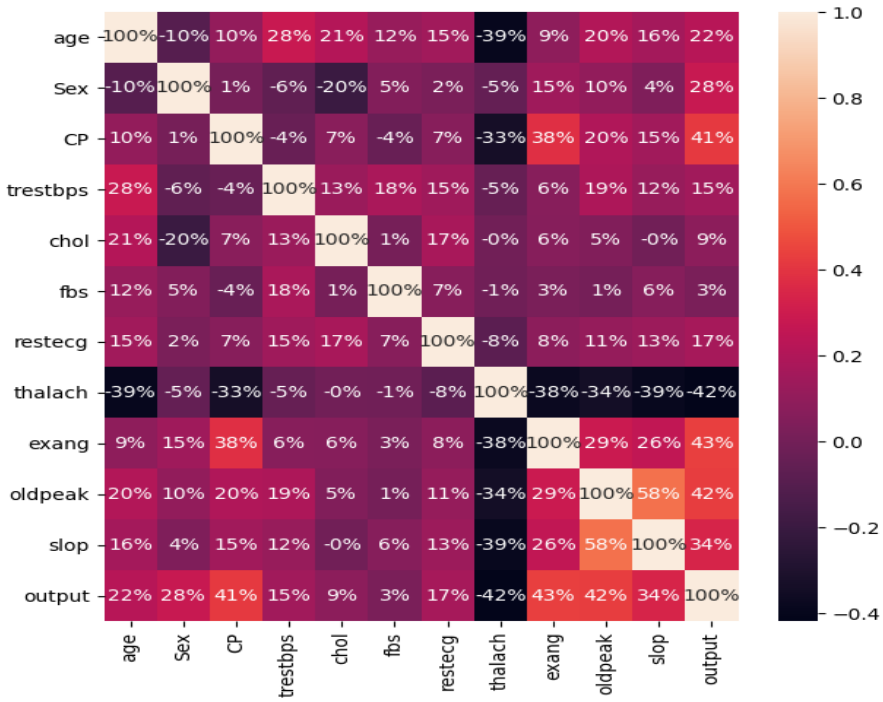

Figure 14: Correlation between attributes

The above Correlation figure for the independent and dependent variables shows a strong correlations among some independent variable (CP with exang, exang with oldpeak, old peak with slop, and slop with exang) and the dependent variable have high correlation with CP, exang, old peak and slop.

4.1.2. SVM Prediction.

After we pre-processed data to make it useful for the experiment. Then we ran the initial experiments using a different SVM Kernels Functions on the Cleveland data. Our experiments on this database are concentrated on separating the cases with heart disease (values 1, 2, 3, 4) from no heart disease cases (value 0).

We have used different performance metrics to measure the result of all the SVM Kernels. In this study, the performance measures used to evaluate each model include Accuracy, Sensitivity, and Specificity. Based on this the following results were found

Table 4: Classification efficiency of SVM on different Kernel functions without dropping any attribute

	Kernel Function			
	Linear	RBF	Sigmoid	Polynomial
Accuracy	0.8026	0.8158	**0.8553**	0.8421
Sensitivity	0.8	0.8286	0.8	**0.8571**
Specificity	0.8049	0.8048	**0.9024**	0.8293

Among the used performance criteria, Sigmoid Kernel has better accuracy and specificity on Classifying Heart disease. Nevertheless, sensitivity Polynomial has shown better prediction result.

We have used Backward Feature Elimination on each attribute and found that resting electrocardiographic results and sex does not improve the performance of the results, during diagnosis of heart disease. Thus, those features should neglected to get better results. Using Backward Feature Elimination, the following results found:

Table 5: Classification efficiency of SVM on different Kernels while dropping selected attribute

Dropped Attribute	Kernel Function			
	linear	RBF	sigmoid	Polynomial
Age	0.8026	0.8026	0.8289	0.8289
Sex	0.**8553**	0.8158	0.8421	0.**8553**
CP	0.7763	0.7894	0.8026	0.8289
trestbps	0.8026	0.8158	0.8158	0.8289
Chol	0.8421	0.8158	0.**8553**	0.8421
Fbs	0.8158	0.8026	0.8421	0.8289
restecg	0.8289	0.8289	0.8289	0.8158
Exang	0.**8553**	0.8289	0.8289	0.**8553**
oldpeak	0.8157	0.8158	0.8289	0.**8553**
Slop	0.8289	0.8289	0.8421	0.8289
Ca	0.8421	0.8158	0.8421	0.8158
Thal	0.7631	0.8421	0.7763	0.8421

As shown in the table, dropping CP (Chest Pain) and thal (Thalassemia) will decrease the prediction accuracy of SVM, which means that CP and thal are determinant factors to predict heart disease. Meanwhile we have seen that dropping resting electrocardiographic results and sex does improve the performance of the results.

However, we have noticed that dropping the combination of each attribute does not affect or improve the performance. For example, when we drop both sex and exang the accuracy has dropped to 81.57% on linear kernel function and it stays the same 85.53% on Polynomial kernel. When we drop old peak and exang on polynomial kernel the accuracy dropped to 82.89%, and when we drop old peak, exang and sex has dropped to 81.57%

We get highest prediction accuracy (85.53%) on six cases,

- First case is when we drop Sex (Maximum heart rate achieved) on linear Kernel function we get Sensitivity 82.86% and Specificity 87.82%

- Second, when we drop exang (Exercise induced angina) on linear Kernel function we get Sensitivity 85.71% and Specificity 85.37%

- The Third case is when we drop chol (Serum cholesterol level) on sigmoid Kernel function and get Sensitivity 80% and Specificity 90.24%

- The fourth case is when we drop Exang on Polynomial kernel function we get Sensitivity 88.57% and Specificity 82.93%

- The fifth case is when we drop oldpeak (ST depression induced by exercise relative to rest) on Polynomial Kernel we get Sensitivity 88.57% and Specificity 82.93%

- And the sixth, is when we drop Sex on Polynomial Kernel and get Sensitivity 77.14% and Specificity 92.68%

Since, these six approaches has shown better performance we will use each case for the combined model experimentation.

4.1.3. The combined model performance

For the combined model, we have first used SONN to learn cluster representation corresponding to input variables. Then the learned cluster representations fed to the SVM classifier as features for Heart disease classification. We have used the same performance metrics that we used in the SVM prediction to measure the result of the combined model. Based on this, the following results obtained

Table 6: Classification efficiency of combined model on different Kernels without dropping any attribute

Test number		Kernel Function			
		Linear	RBF	Sigmoid	Polynomial
1	Accuracy	0.8026	0.8158	0.8421	0.8289
	Sensitivity	0.8	0.8286	0.8	0.8
	Specificity	0.8049	0.8048	0.8781	0.8537
2	Accuracy	0.8289	0.8158	0.8421	**0.8553**
	Sensitivity	0.8286	0.8286	0.8	**0.8571**

	Specificity	0.8292	0.8048	0.8781	0.8537
3	Accuracy	0.7895	0.8026	**0.8553**	0.8158
	Sensitivity	0.8	0.8	0.8	0.8285
	Specificity	0.7805	0.8049	**0.9024**	0.8049
4	Accuracy	0.8026	0.8158	0.8421	0.8289
	Sensitivity	0.8	0.8	0.8	0.8285
	Specificity	0.8049	.8293	0.8781	0.8293
5	Accuracy	0.8158	0.8026	0.8421	0.8421
	Sensitivity	0.8	0.8	0.8	0.8286
	Specificity	0.8293	0.8049	0.8781	0.8537

Without dropping any attribute in the combined model, we get the highest prediction accuracy (85.53%) and Specificity (90.24%) on sigmoid Kernel function and highest Sensitivity (85.71%) on Polynomial Kernel functions.

Table 7: Classification efficiency of combined model by using the highest performance on previous SVM tests when we drop selected attributes

Test number		Linear Sex	Linear exang	Sigmoid chol	Poly exang	Poly sex	Poly oldpeak
1	Accuracy	0.8421	0.8421	0.8290	0.8289	0.8552	0.8157
	Sensitivity	0.8	0.8	0.8	0.8571	0.7714	0.8571
	Specificity	0.8781	0.8781	0.8537	0.8048	0.9268	0.780
2	Accuracy	0.8421	0.8684	0.8684	0.8552	0.8421	0.8421
	Sensitivity	0.8286	0.8286	0.8	**0.8857**	0.7714	**0.8857**
	Specificity	0.8781	0.9024	0.9268	0.8292	0.9024	0.8048
3	Accuracy	0.8553	0.8290	0.8553	0.8289	0.8684	0.8421
	Sensitivity	0.8286	0.8286	0.8	0.8286	0.7428	0.8571
	Specificity	0.8781	0.8293	0.9024	0.8293	0.9756	0.8293
4	Accuracy	0.8553	0.8421	0.8421	0.8552	**0.8815**	0.8552
	Sensitivity	0.8286	0.8	0.8	0.8571	0.7714	0.8571
	Specificity	0.8781	0.8781	0.8781	0.8537	**0.9756**	0.8536
5	Accuracy	0.8553	0.8553	0.8290	0.8421	0.8421	0.8552
	Sensitivity	0.8286	0.8	0.8	0.8286	0.7714	**0.8857**
	Specificity	0.8781	0.924	85.37	0.8537	0.9024	0.8292

In this experiment, we get highest prediction accuracy (88.15%) and specificity (97.56%) when we drop Sex on Polynomial Kernel function. In addition, we get the highest

Specificity (88.57%) on Polynomial Kernel functions when we drop Oldpeak (ST depression induced by exercise relative to rest) and exang (Exercise induced angina).

4.2. Discussion

Two types of machine learning techniques had been investigated in this research with variations of SVM kernel function; we have used Linear, RBF, Sigmoid and Polynomial to classify heart disease.

The overall accuracy of prediction model based on the combined algorithm on polynomial kernel is better on classifying heart disease than the SVM when we drop some features, and the combined model on Linear kernels were better on prediction when we use all features.

Table 8: Comparison of SVM and Combined Model without dropping attributes

	SVM				Combined Model			
	Linear	RBF	Sigmoid	Poly	Linear	RBF	Sigmoid	Poly
Accuracy	0.8026	0.8158	0.8553	0.8421	**0.8289**	0.8158	0.8553	**0.8553**
Sensitivity	0.8	0.8286	0.8	0.8571	**0.8286**	0.8286	0.8	0.8571
Specificity	0.8049	0.8048	0.9024	0.8293	**0.8292**	0.8048	0.9024	**0.8537**

As shown in table 8, for all the four SVM Kernel functions combined algorithms has achieved the same on RBF and Sigmoid kernels. While, in Linear and Polynomial Kernel Function, it shows that the combined model has improved the prediction.

Similarly as shown in table 9 below, When we drop selected attributes we have seen that Combined model have increase the Accuracy from 85.53% to 88.15% and the Specificity from 92.68% to 97.56% on Polynomial Kernel function of degree 3. Meanwhile, on the combined model the Sensitivity is found to be similar with SVM perdition.

Table 9: Comparison of SVM and Combined Model when we drop selected attributes

	SVM						Combined Model					
	Linea Sex	Linea exang	Sigmo chol	Poly Sex	Poly Exang	Poly oldpeak	Linear Sex	Linear exang	Sigmo chol	Poly Sex	Poly exang	Poly oldpeak
Accuracy	0.8553	0.8553	0.8553	0.8553	0.8553	0.8553	0.8421	0.8684	0.8684	**0.8815**	0.8552	0.8552
Sensitivity	0.8286	0.8571	0.8	0.7714	**0.8857**	**0.8857**	0.8286	0.8286	0.8	0.7714	**0.8857**	**0.8857**
Specificity	0.8782	0.8537	0.9024	**0.9268**	0.8293	0.8293	0.8781	0.9024	0.9268	**0.9756**	0.8292	0.8292

As shown in table 9, we have seen that when we drop some attributes the Combined Model on Polynomial Kernel function has improved Accuracy and Specificity but it has failed to improve the Sensitivity.

On our experimentation, we have seen that without dropping any attribute the linear kernel function performance has improved in all evaluation criteria when we add SONN cluster result as a feature in SVM. This finding goes with the idea that Linear Kernel functions most commonly used when there are a large number of features in the dataset (Huang & Lin, 2016).

The accuracy and specificity on the SVM has improved from 85.53 to 88.15 and 92.68 to 97.56 respectively on the Combined Model when we drop sex on the polynomial kernel function on degree 3. Similarly Ben-Hur et al. (2008) showed the use of a nonlinear kernel, like polynomial, leads to an improvement in classifier performance. They also found out that using polynomial kernel on degree 3 has better prediction than polynomial kernel on degree 7 and Gaussian kernel.

The sensitivities are higher on polynomial function than other kernel functions whether we drop feature or not, but we have seen an improvement on sensitivity when we drop an attribute. We have seen that when we drop exang (Exercise induced angina) or old peak on SVM and combined model, the sensitivity increased from 85.71% to 88.57%.

When we apply backward feature elimination, we have seen that CP (Chest Pain) and thal (Thalassemia) decreased the prediction accuracy; this is due to that Chest pain is highly

related with heart disease. According to Nilsson et al. (2003), Chest pain often relates to the possibility of manifestations of heart disease. Similarly Albus et al. (2017), also indicate that11% of patients with chest pain who present to a general practitioner and 25% of those who present to a cardiologist have chronic CHD. In addition to Chest pain Thalassemia (an inherited haemoglobin disorder resulting in chronic haemolytic anaemia) often causes pulmonary hypertension. Pulmonary hypertension represents the leading cause of heart failure (Aessopos et al., 2005).

Generally, the achieved accuracy on the combined model is obviously higher than the ones achieved by the SVM alone; this shows that SONN has provided a valuable feature for SVM to classify heart disease with better accuracy and specificity.

Chapter Five

5. Conclusion and Recommendation

The previous chapter discussed about the experimentation on heart disease diagnosis using SONN and SVM. In this chapter, conclusion on the overall work of the study and recommendation provided for further investigation on the remaining related issues.

5.1. Conclusion

Heart Disease are the number one cause of death globally, taking an estimated 17.9 million lives each year, representing 31% of all global deaths, out of which the poorest people in low and middle-income countries are the most affected. Studies indicate that three quarters of the world's deaths from heart related diseases occur in low- and middle-income countries. Furthermore, a high proportion of CVD occurs among adults of working age in developing countries that lead to a large impact on their economy, since treating heart attack is very costly. The prevalence of all CVD has projected to increase in many countries and the expense of treating CVD projected to increase in the coming decades. Hence, it vital to conduct research on heart disease that need global attention and further investigation, since it impose high impacts both on lives of the people and economy of the country. Accordingly, this study aims to create new insight by combining two algorithms namely SONN and SVM to classify heart disease.

For this study, we have used Cleveland Dataset that has 303 clinical data and it contains 13 features, from which 75% used for training and 25% for testing purposes. The experiments had conducted using Python language on PyCharm IDE. We used on each experiment Accuracy, Sensitivity and Specificity to measure the performance of the model.

With this performance metrics, we have seen that Unsupervised SONN with SVM on polynomial kernel have better accuracy on predicting heart disease. The accuracy that obtained in SVM improved from 85.53% to 88.15% in the combined model and, on

polynomial Kernel function of degree 3 when sex dropped. Similarly, the Specificity also improved from 92.68% to 97.56%.

Generally, we understood that combining Unsupervised SONN with SVM on polynomial kernel have better accuracy on classifying heart disease.

5.2. Recommendation

This study has provided an additional potential applicability of machine learning algorithm to establish an approach that helps to diagnose heart disease. With this, the main objective of this study had achieved because the results showed that combining these two algorithms improved the prediction.

Based on the findings of this study, the following recommendation forwarded to address related issues:

1. We have investigated a model that can diagnose heart disease using the available features. Nevertheless, the effect of other parameters need further study to improve efficiency of heart disease diagnosis.

2. Combining SONN with other machine learning algorithms such as K-Nearest Neighbour, Naïve Bayes, Random forest and other classification algorithms may help.

3. Further study need to develop using models that classify heart disease, and thereby test them whether the models can predict the disease (to estimate the risk of developing Heart disease before it occurs).

4. Since heart disease variables are very complex, it needs further investigation that requires expertise domain knowledge.

REFERENCES

1. Aessopos, A., Farmakis, D., Deftereos, S., Tsironi, M., Tassiopoulos, S., Moyssakis, I., & Karagiorga, M. (2005). *Thalassemia Heart Disease. Chest,* 127(5), 1523–1530. doi:10.1378/chest.127.5.1523.

2. Albus, C., Barkhausen, J., Fleck, E., Haasenritter, J., Lindner, O., & Silber, S. (2017). *The diagnosis of chronic coronary heart disease*. Deutsches Ärzteblatt International, 114(42), 712.

3. Alex, S. & Vishwanathan, S. (2008). *Introduction to Machine*. Cambridge, United Kingdom: Cambridge University Press.

4. Alexandre Kowalczyk (October 23, 2017) *Support Vector Machines Succinctly*. Syncfusion Inc.

5. Allan, B. (2015). *Accuracy of the ACC/AHA Cardiovascular Risk Calculator Is Challenged.* Journal Watch. Retrieved From https://www.jwatch.org/na37408/2015/03/25/accuracy-acc-aha-cardiovascular-risk-calculator-challenged

6. Al-Mejibli, I. S., Abd, D. H., Alwan, J. K., & Rabash, A. J. (2018, November). *Performance evaluation of kernels in support vector machine*. In 2018 1st Annual International Conference on Information and Sciences (AiCIS) (pp. 96-101). IEEE.

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

8. American College of Cardiology. (2020). *ASCVD Risk Estimator Plus Estimate Risk.* Retrieved From http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/

9. Dutta, A., Batabyal, T., Basu, M., & Acton, S. T. (2020). *An efficient convolutional neural network for coronary heart disease prediction*. Expert Systems with Applications, 159, 113408.

10. Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. PLoS computational biology, 4(10), e1000173.

11. CDC (2021). Content source: *National Centre for Chronic Disease Prevention and Health Promotion , Division for Heart Disease and Stroke Prevention*

12. Centres for Disease Control and Prevention. (2018). *Underlying Cause of Death, 1999–2018*. CDC WONDER Online Database. Atlanta, GA: Centres for Disease Control and Prevention

13. Chang, H. S., Kim, H. J., Nam, C. M., Lim, S. J., Jang, Y. H., Kim, S., & Kang, H. Y. (2012). *The socioeconomic burden of coronary heart disease in Korea*. Journal of Preventive Medicine and Public Health, 45(5), 291.

14. Dale, M. (2019). *How Machine Learning is Transforming Healthcare at Google and Beyond*. Retrieved From https://towardsdatascience.com/how-machine-learning-is-transforming-healthcare-at-google-and-beyond-d4f664b7e27c

15. George M. (2010). *Self-Organizing Maps*, National Technical University of Athens, Greece

16. Goel, A., & Srivastava, S. K. (2016, February). *Role of kernel parameters in performance evaluation of SVM*. In 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT) (pp. 166-169). IEEE.

17. Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., ... & Woo, Y. J. (2011). *Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association*. Circulation, 123(8), 933-944.

18. Huang, H.-Y., & Lin, C.-J. (2016). *Linear and Kernel Classification: When to Use Which?* Proceedings of the 2016 SIAM International Conference on Data Mining. doi:10.1137/1.9781611974348.25

19. Jordan, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects*. Science, 349(6245), 255-260.

20. Judith, H. & Daniel, K. (2018). *Machine Learning For Dummies®*, IBM Limited Edition: John Wiley & Sons, Inc.

21. Justin, C. & Tim, N. (2017). *How to spot and treat a heart attack.* Retrieved From https://www.medicalnewstoday.com/articles/151444.php

22. Ketut A., Muhammad S., & Dadang G. *Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters*: Dept. of Electrical Engineering, Universitas Indonesia, Indonesia. ISSN: 2180-1843 e-ISSN: 2289-8131 Vol. 8 No. 12 Journal of Telecommunication, Electronic and Computer Engineering January 2016

23. Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences. doi:10.1016/j.jksuci.2019.06.012

24. Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.

25. Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 37, 52–65. doi:10.1016/j.neunet.2012.09.018

26. Liu, J. L., Maniadakis, N., Gray, A., & Rayner, M. (2002). *The economic burden of coronary heart disease in the UK*. Heart, 88(6), 597-603.

27. Mahmoodi, M. S. (2017). *Designing a heart disease prediction system using support vector machine*. Journal of Health and Biomedical Informatics, 4(1), 1-10.

28. Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. Information Sciences, 179(13), 2208-2217.

29. Mandy, S. (2017). *Types of classification algorithms in Machine Learning.* Retrieved from: https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14

30. Marie Cottrell, Madalina Olteanu, Fabrice Rossi, Nathalie Villa-Vialaneix (2018). *Self-Organizing Maps, theory and applications*, 39 (1), pp.1-22. ffhal-01796059f

31. MayoClinic. (2019). *Heart attack*. Retrieved From https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106

32. Miao, K. H., Miao, J. H., & Miao, G. J. (2016). *Diagnosing coronary heart disease using ensemble machine learning*. International Journal of Advanced Computer Science and Applications, 7(10), 1-12.

33. Michele, C., Michael, K., Harrison, Q., Linda, S., Adam, V., and Sophia, G. (2016). *Changes in the Geographic Patterns of Heart Disease Mortality in the United States* Retrieved From https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836838/ US National Library of Medicine, National Institute of Health

34. Mike, T. (2019). *Ultra-Modern Medicine: Examples Of Machine Learning In Healthcare*. Retrieved From https://builtin.com/artificial-intelligence/machine-learning-healthcare

35. Moran, A., Gu, D., Zhao, D., Coxson, P., Wang, Y. C., Chen, C. S., & Goldman, L. (2010). *Future cardiovascular disease in China: Markov model and risk factor scenario projections from the coronary heart disease policy model–China*. Circulation: Cardiovascular Quality and Outcomes, 3(3), 243-252.

36. Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). *Influence of data splitting on performance of machine*

learning models in prediction of shear strength of soil. Mathematical Problems in Engineering.

37. NHS. (2020) *Coronary heart disease*. Retrieved From https://www.nhs.uk/conditions/coronary-heart-disease/

38. Nikhil, G. (2019). Why *is Python Used for Machine Learning?* Retrieved From https://hackernoon.com/why-python-used-for-machine-learning-u13f922ug

39. Nilashi, M., Ahmadi, H., Manaf, A. A., Rashid, T. A., Samad, S., Shahmoradi, L., & Akbari, E. (2020). *Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates*. International Journal of Fuzzy Systems, 22(4), 1376-1388.

40. Nilsson, S., Scheike, M., Engblom, D., Karlsson, L. G., Mölstad, S., Akerlind, I., & Nylander, E. (2003). *Chest pain and ischaemic heart disease in primary care.* British Journal of General Practice, 53(490), 378-382.

41. Patle, A., & Chouhan, D. S. (2013). SVM kernel functions for classification. 2013 International Conference on Advances in Technology and Engineering (ICATE). doi:10.1109/icadte.2013.6524743

42. Pavel, G., Jan, C., Eduard, H., Sylva, H. & Martin, S. (2018). *Global Correlates of Cardiovascular Risk: A Comparison of 158 Countries. US National Library of Medicine National Institutes of Health*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5946196/

43. Pavlidis, P., Wapinski, I., & Noble, W. S. (2004). *Support vector machine classification on the web*. Bioinformatics, 20(4), 586–587. doi:10.1093/bioinformatics/btg461

44. Petersen, S., Rayner, M., & Wolstenholme, J. (2002). *Coronary heart disease statistics: heart failure supplement 2002 edition*. University of Oxford.

45. Pohl, D., Bouchachia, A., & Hellwagner, H. (2012). Automatic sub-event detection in emergency management using social media. Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion. doi:10.1145/2187980.2188180

46. Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., ... & Filippatos, G. (2014). *Heart failure: preventing disease and death worldwide*. ESC heart failure, 1(1), 4-25.

47. Pooya Tabesh, Gino Lim, Suresh Khator. (2010) *A Support Vector Machine Approach for Predicting Heart Conditions* Proceedings of the 2010 Industrial Engineering Research Conference, A. Johnson and J. Miller, eds. University of Houston

48. Python. (2020). *The Python Tutorial*. Retrieved From https://docs.python.org/3/tutorial/index.html

49. Raschka, S., Patterson, J., & Nolet, C. (2020). *Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence*. Information, 11(4), 193.

50. Rohit S. (2021). 6 *Best Python IDEs for Data Science & Machine Learning* [2021]. JAN 4, 2021

51. Rohith, G. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

52. Saleem, T. J., & Chishti, M. A. (2020). *Exploring the applications of machine learning in healthcare*. International Journal of Sensors Wireless Communications and Control, 10(4), 458-472.

53. Saumya A., (2020) Seven Most Popular SVM Kernels, Retrieved from: https://dataaspirant.com/svm-kernels/

54. Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018, March). *Machine learning in healthcare: A review*. In 2018 Second international conference on electronics, communication and aerospace technology (ICECA) (pp. 910-914). IEEE.

55. Simundic, A. M. (2009). *Measures of diagnostic accuracy: basic definitions*. Ejifcc, 19(4), 203.

56. Skiena, S. S. (2020). *The algorithm design manual*. Springer International Publishing.

57. Statnikov, A. (2011). A gentle introduction to support vector machines in biomedicine: Theory and methods (Vol. 1). world scientific.

58. Talavera, L. (1999, January). *Feature selection as a pre-processing step for hierarchical clustering*. In ICML (Vol. 99, pp. 389-397).

59. Tamam, M. & Yasmine, A. *(2019). What is the American College of Cardiology (ACC) and the American Heart Association (AHA) risk score for atherosclerotic cardiovascular disease (ASCVD)?* Retrieved From https://www.medscape.com/answers/164214-100911/what-is-the-american-college-of-cardiology-acc-and-the-american-heart-association-aha-risk-score-for-atherosclerotic-cardiovascular-disease-ascvd

60. TechVidvan (2021) *SVM Kernel Functions – 'Coz your SVM knowledge is incomplete without it.* Retrieved from https://techvidvan.com/tutorials/svm-kernel-functions/

61. Theconversation. (2014). *Cardiovascular disease declines in rich countries but poor countries suffer more*. Retrieved from http://theconversation.com/cardiovascular-disease-declines-in-rich-countries-but-poor-countries-suffer-more-25277

62. Thomas, A. (2005). *Cardiovascular Disease in the Developing World and Its Cost-Effective Management*. Retrieved From https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.105.591792

63. Tsehay A. (2019) *A Support Vector Machine Based Heart Disease Prediction* Journal Of Software Engineering & Intelligent Systems Issn 2518-8739 31 December 2019, Volume 4, Issue 3

64. United Kingdom National Health Service. (2020). *Diagnosis-Coronary heart disease* retrieved from: (https://www.nhs.uk/conditions/coronary-heart-disease/diagnosis/)10 March 2020

65. Vesanto, J., & Alhoniemi, E. (2000). *Clustering of the self-organizing map*. IEEE Transactions on Neural Networks, 11(3), 586–600. doi:10.1109/72.846731

66. Virani S, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. (2020) *Heart disease and stroke statistics*—2020 update: a report from the American Heart Association external icon. *Circulation.* 2020;141(9):e139–e596.

67. Wang, S. C. (2003). *Artificial neural network. In Interdisciplinary computing in java programming* (pp. 81-100). Springer, Boston, MA.

68. WHO. (2017). *Cardiovascular diseases (CVDs)*. Retrieved From https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

69. WHO. (2019). *Cardiovascular Diseases*. Retrieved From https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

70. Widiyaningtyas, T., Zaeni, I. A. E., & Wahyuningrum, P. Y. (2019). *Self-Organizing Map (SOM) For Diagnosis Coronary Heart Disease*. 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). doi:10.1109/icitisee48480.2019.90

71. Xuegong Zhang, & Yanda Li. (1993). *Self-organizing map as a new method for clustering and data analysis*. Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). doi:10.1109/ijcnn.1993.714219.

72. Yu, H., & Kim, S. (2012). *SVM Tutorial — Classification, Regression and Ranking*. Handbook of Natural Computing, 479–506. doi:10.1007/978-3-540-92910-9_15

APPENDICES

Python Code

SONN Python Code

```
#import libraries
import pandas as pd

df=pd.read_csv("processed.cleveland.data.csv")
X=df.iloc[:,:-1].values #all of the rows and all of columns except the last column
Y=df.iloc[:,-1].values #all of the rows from the last column

from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X=sc.fit_transform(X)

from sklearn_som.som import SOM
HD_som = SOM(m=3, n=3, dim=13, lr=1, sigma=1, max_iter=3000, random_state=None)
HD_som.fit(X)
predictions = HD_som.predict(X)
print(predictions)
```

SVM Python Code

```
#import libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df=pd.read_csv("processed.cleveland.data.csv")
df=df.drop('exang',axis=1)
#we need to split x-independet/feature data with y-depenednt/target data
X=df.iloc[:,:-1].values #all of the rows and all of columns except the last column
Y=df.iloc[:,-1].values #all of the rows from the last column

from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X=sc.fit_transform(X)

#split the data again, into 75% training and 25% testing
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test=train_test_split(X,Y, test_size=0.25, random_state=1)

from sklearn import svm
clf=svm.SVC(kernel='linear', C=1) #linear rbf sigmoid
clf.fit(X_train,Y_train)

#to test modles accury on test dataset using confusion_matrix
```

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(Y_test,clf.predict(X_test))

TN=cm[0][0]
TP=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
#print the confusion Matrix
print(cm)

#print the models accuracy on the test data
print('Svm Test Accuracy= {}'.format((TP+TN)/(TP+TN+FN+FP)))
print('Sensitivity= {}'.format((TP)/(TP+FN)))
print('Specificity= {}'.format((TN)/(TN+FP)))
```

Combined algorithm Python Code

```
#import libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


df=pd.read_csv("processed.cleveland.data.csv")
df=df.drop('chol',axis=1)
#df=df.drop('Sex',axis=1)
#we need to split x-independet/feature data with y-depenednt/target data
X=df.iloc[:,:-1].values #all of the rows and all of columns except the last column
Y=df.iloc[:,-1].values #all of the rows from the last column

#running SOM on the data
from sklearn_som.som import SOM
#HD_som = SOM(m=10, n=10, dim c x=13, lr=0.1, sigma=1, max_iter=30000, random_state=None)
HD_som = SOM(dim=12)
HD_som.fit(X)


#concatinating result of SOM to X/The input

predictions = HD_som.predict(X)
conc=np.array(predictions)
conc=np.split(conc,conc.size)
X=np.concatenate((X,conc),axis=1)

from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X=sc.fit_transform(X)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test=train_test_split(X,Y, test_size=0.25, random_state=1)

from sklearn import svm
clf=svm.SVC(kernel='sigmoid',C=1) #linear rbf sigmoid
clf.fit(X_train,Y_train)

from sklearn.metrics import confusion_matrix
cm=confusion_matrix(Y_test,clf.predict(X_test))

TN=cm[0][0]
TP=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
#print the confusion Matrix
print(cm)

print('Svm Test Accuracy= {}'.format((TP+TN)/(TP+TN+FN+FP)))
print('Sensitivity= {}'.format((TP)/(TP+FN)))
print('Specificity= {}'.format((TN)/(TN+FP)))
```