AN UNSUPERVISED MACHINE LEARNING APPROACH FOR HEART DISEASE PREDICTION

LIM YU JIUN

UNIVERSITI TUNKU ABDUL RAHMAN

AN UNSUPERVISED MACHINE LEARNING APPROACH FOR HEART DISEASE PREDICTION

LIM YU JIUN

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Science (Honours) Applied Mathematics with Computing

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

September 2024

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature : 

Name : Lim Yu Jiun

ID No. : 2106963

Date : 06/09/2024

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"AN UNSUPERVISED MACHINE LEARNING APPROACH FOR HEART DISEASE PREDICTION"** was prepared by **LIM YU JIUN** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Science (Honours) Applied Mathematics with Computing at Universiti Tunku Abdul Rahman.

Approved by,

Signature	:	*loh*
Supervisor	:	Loh Wing Son
Date	:	21/10/2024

Signature	:	
Co-Supervisor	:	
Date	:	

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my research supervisor, Mr Loh Wing Son, who had contributed to the successful completion of this project. His invaluable advice, guidance and continuous encouragement have been a source of inspiration throughout the development of the research. I deeply appreciate his enormous patience and willingness to help me navigate through challenges, ensuring that I stayed focused and motivated. His dedication to supervising my project and fostering my academic growth will always be remembered with sincere appreciation.

ABSTRACT

Heart disease, also known as cardiovascular disease, persists as a primary cause of mortality on a global scale, necessitating effective prediction methods. This study introduces a novel modelling approach utilising the Self-Organising Map (SOM), an unsupervised machine learning approach, for heart disease prediction with the incorporation of Particle Swarm Optimisation (PSO), a metaheuristic optimisation algorithm. The SOM model was employed to analyse and cluster patient data based on intrinsic patterns without requiring predefined labels, allowing for the identification of individuals with heart disease. The results demonstrated the SOM's capability in distinguishing between healthy and diseased individuals, offering a robust approach for early detection of heart disease. By integrating PSO to fine-tune SOM's hyperparameters, the SOM model achieved superior predictive performance, with an accuracy of 94.44%, precision of 100%, recall of 92.86%, F1-score of 96.30%, and a quantisation error of 0.024. Furthermore, this study explored the impact of SOM's visualisation techniques, such as heatmaps and the Unified Distance Matrix (U-Matrix), on the comprehension of cardiovascular conditions. The U-Matrix of the optimised SOM model provided insightful visualisation that revealed two distinct clusters, effectively illustrating the health status of individuals with similar health conditions concerning heart disease. These visualisations afford a more profound understanding of the hidden relationships within the heart disease data, enhancing the model's interpretability and facilitating a better management of heart disease. The findings suggest the potential of integrating SOM into clinical workflows, offering a potent tool for healthcare professionals in the fight against heart disease.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	Machine Learning
PCA	Principal Component Analysis
t-SNE	t-distributed Stochastic Neighbour Embedding
SOM	Self-Organising Map
PSO	Particle Swarm Optimisation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
NM-PCA	Normalised Mutual Information induced Principal Component Analysis
MLP	Multilayer Perceptron
ABC	Artificial Bee Colony
ANN	Artificial Neural Network
BMU	best matching unit
ACC	accuracy
SN	sensitivity
SP	specificity
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
QE	Quantisation Error
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree
NB	Naive Bayes
LR	Logistic Regression
KNN	K-Nearest Neighbours
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic

CHAPTER 1

INTRODUCTION

1.1 General Introduction

The heart, a vital organ, unceasingly works to infuse every part of the human body with vibrant red blood. Yet, the heart is susceptible to malfunctioning and damage, despite being one of the strong and muscular internal organs. Heart disease, also known as cardiovascular disease, has been among the most fatal illnesses across the globe, affecting all walks of life. Nearly 17.9 million lives have been taken every year due to heart disease, as reported by The World Health Organization (WHO, 2021). Heart disease encompasses various conditions, such as coronary artery disease, arrhythmias and heart failure. These conditions produce a considerable burden on healthcare systems and impact patients' quality of life.

Researchers and clinicians continually seek innovative approaches to enhance disease diagnosis and optimise analyses of patients' conditions. Thriving in the era of science and technology, Machine Learning (ML) has become a promising technique in enhancing medical diagnostics, with its ability to handle and analyse large medical datasets. ML is an algorithm in which data is learnt by computer machines without human intervention. They are widely implemented for complex data analysis and prediction (Sheeba, et al., 2022). ML primarily comprises two basic types: Supervised Learning and Unsupervised Learning. For the past few years, supervised learning such as Logistic Regression (LR), Support Vector Machine (SVM) has been popularly implemented for analysis and prediction in almost every field. The intersection of ML and healthcare has even brought remarkable impacts to the world (Lin and Hsieh, 2015). Nevertheless, unsupervised learning models are barely studied to verify its effectiveness in analysing and predicting illness.

Aside from outcome prediction, interpretable ML models should be implemented for a comprehensive understanding of the disease (Jiang, et al., 2023). Unsupervised learning is well known for its data-driven approach without requiring labelled data, unlike supervised learning. What makes unsupervised learning algorithms significant is its capability to discover hidden

patterns and trends in the data, on top of dimensionality reduction. Principal Component Analysis (PCA), t-distributed Stochastic Neighbour Embedding (t-SNE) are great examples of unsupervised learning algorithms which aid in reducing dimensionality at the stage of data preprocessing.

Moreover, clustering is among the tasks that could be performed by unsupervised learning models. Consequently, more accurate information for clinical assessment and management is available when clustering and real-time analysis are connected (Jiang, et al., 2023). Self-Organising Map (SOM) presents as one kind of neural network that can be developed to analyse data and predict outcomes of diseases. Besides, it is capable of discovering hidden patterns and can be used further to determine whether diseases are present or not (Rankovic, et al., 2023). In this context, this project seeks to adopt and evaluate SOM, as one of the unsupervised ML approaches to predict heart disease, contributing to early heart disease detection and prevention.

1.2 Importance of the Study

One of the primary significances of conducting this study is to accurately predict heart disease with an unsupervised ML model. By having a reliable computational model, it could contribute to early detection of heart disease, providing the correctly predicted output coupled with the comprehensive understanding of the disease along with its risk factors. Individuals suffering from heart disease face physical limitations, pain, anxiety and other unfavourable symptoms. Research that improves prevention and early detection of the disease directly enhances patients' well-being. Hence, undoubtedly early detection of heart disease indicates possible treatment in advance and recovery.

Besides, the modern era has allowed ML to be capable of analysing and predicting heart disease with patients' conditions, thereby raising the public's alert on heart disease. Future research on ML applications ought to improve prediction accuracy by utilising real-world data and external validations. By implementing and optimising ML techniques, accurate heart disease prediction becomes achievable with the existing medical data that accumulates as time goes. These techniques can offer the advantage of automated disease detection, making the process less complicated, costly, and more reproducible. Furthermore, such an approach enables screening of a large

number of patients using readily available clinical data from hospitals. Hence, to make it the best possible model in heart disease prediction, a metaheuristic technique can be incorporated into the developed unsupervised ML model to enhance the model prediction performance.

1.3 Problem Statement

Despite medical advancement, heart disease has been a significant challenge in the healthcare field, necessitating continuous research and innovative approaches. Some patients exhibit symptoms like chest pain, shortness of breath, or fatigue, while others seek routine check-ups. The healthcare professionals face several critical issues such as assessing every patient's risk of developing heart disease. Traditional risk factors such as age, sex, blood pressure, cholesterol levels provide only minimal information as conventional models have limited ability to model the complex relationship within several variables. Henceforth, a comprehensive approach is necessary to consider both obvious and hidden risk factors in disease detection.

Heart disease data is complex. It includes clinical measurements, lifestyle and environmental factors. Traditional models struggle to handle the multidimensional data effectively, let alone predicting the disease. It makes the identification of subtle patterns that precede clinical symptoms challenging with manual mankind analysis. These patterns may not be obvious and evident in structured data alone. Hence, manual analysis can be time-consuming, especially when dealing with numerous patients. To illustrate, trained professionals usually conduct specialised procedures such as echocardiograms and electrocardiograms, which are time-consuming, costly and require continuous effort. In addition, electrocardiograms may occasionally be unable to confirm the presence of heart disease in patients, leading to potential diagnostic challenges and uncertainties in their medical evaluation. Furthermore, it is unknown if there are distinct patient subgroups among the heart disease patients. Perhaps some patients share common symptoms, genetic factors or lifestyle habits. Understanding these subgroups can lead to tailored interventions and personalised treatment.

1.4 Aim and Objectives

The project aims to explore and develop an unsupervised learning model capable of predicting heart disease, thereby increasing the efficiency of disease diagnosis. The objectives of this project are as follows:

1. To develop an unsupervised machine learning model for heart disease prediction.
2. To incorporate a metaheuristic optimisation algorithm to enhance the performance of the developed unsupervised machine learning model.
3. To evaluate the classification performance of the unsupervised machine learning model based on performance metrics.

1.5 Scope and Limitation of the Study

This study involves the development of an unsupervised learning model based on a heart disease dataset consisting of relevant features such as age, blood pressure and cholesterol levels. Proper data preprocessing like dimensionality reduction or feature engineering, and SOM parameter optimisation are essential procedures. In addition, swarm-based optimisation algorithms such as Particle Swarm Optimisation (PSO) are to optimise the model for better prediction performance. Evaluation metrics such as accuracy, recall, and precision are used to evaluate the model performance.

Despite the promising potential of ML, several limitations must be acknowledged. Besides the variables studied, other health conditions, lifestyle changes, or external events should be concerned as they may influence heart disease predictions, which is not available in this study. Besides, while SOMs are effective at representing the distribution of structured input vectors, they can sometimes create misleading representations in areas outside the input space (Astudillo and Oommen, 2014). The initial placement of weight vectors is crucial, as neurons in low-density regions might not be chosen as Best Matching Units (BMUs) or updated adequately. As training progresses and the neighbourhood radius shrinks, some neurons may end up in areas with no data points, being ignored for the remaining training process (Astudillo and Oommen, 2014).

CHAPTER 2

LITERATURE REVIEW

Numerous studies have showcased the impactful application of ML models in the healthcare field. There are myriads of means and techniques to predict heart disease. This section reviews the common methods used for predicting heart disease, as well as the existing research to highlight the unsupervised ML models coupled with optimisation in the healthcare industry.

2.1 Conventional Approach

The usual process of predicting heart disease is by undergoing multiple clinical tests, obtaining medical images of patients and analysing them. Specialised procedures like echocardiograms and electrocardiograms are typically conducted by trained specialists, requiring significant time, resources, and effort due to their complexity and cost (Verma, et al., 2016). As technology keeps evolving, most of the researchers used deep learning instead of conventional models to predict heart disease because of their excellent performance with medical images like echocardiograms. For instance, Arnaout, et al. (2021) performed an ensemble of neural networks to distinguish between the congenital heart disease and the normal heart using a dataset of over 1300 echocardiograms. Correspondingly, researchers have also found that recurrent neural networks work well with both medical images and videos, including ultrasound standard planes. This approach allows them to observe the foetus' detailed cardiac anatomy, which is crucial for the diagnosis of congenital heart disease (Chen, 2017). These ultrasound planes are capable of clearly visualising core anatomical for disease diagnosis.

Nevertheless, deep learning techniques, particularly in the medical field, are considered as black boxes with little to no interpretability, despite their high accuracy and sensitivity (Kaur and Ahmad, 2024). Hence, ML algorithms have been widely used by researchers to make predictions which have high accuracy and provide better and easier understanding for clinicians. For instance, Mohapatra, et al. (2022) used an approach of stacking classifiers model in base level and meta level to predict heart disease where a set of ML models

were used. The authors reported 92% accuracy with sensitivity of 92.6% and specificity of 91%. Other than supervised ML models, unsupervised ML models also have been applied by researchers to predict various diseases. To illustrate, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a popular unsupervised clustering algorithm was used by Kaur and Ahmad (2024) to perform congenital heart disease prediction. The clusters obtained were treated as attributes to the dataset which were then classified using ML models including SVM and Random Forest (RF). Such a cluster-based approach allowed them to achieve a high Area Under the ROC Curve (AUC) of 0.91.

2.2 Dimensionality Reduction

A remarkable medical application of unsupervised ML algorithms worth noting is the implementation of t-SNE for dimensionality reduction and feature selection in various disease analyses. In the research by Jiang, et al. (2023), it is known that sepsis, a disease triggered by the body's overwhelming response to an infection, has a complex multidimensional collection of variables. Hence, t-SNE and PCA were applied to identify subgroups in the septic death cohort. The high-dimensional dataset was reduced to a two-dimensional space, allowing Jiang, et al. (2023) to visualise the grouping effect. The t-SNE and PCA were able to reveal that two distinct phenotypes in the cohort were significantly separable in the plane space, using patients with septic death as an example.

Besides, in the study by Sheeba, et al. (2022), Normalised Mutual Information induced Principal Component Analysis (NM-PCA) was proposed to reduce feature dimension before modelling for heart disease prediction, decreasing the lengthy computing time. The proposed model was found performed the best with an accuracy of 91.7% when NM-PCA was applied.

Furthermore, to extract the best characteristics from the new subset without losing important information, Kapila and Saleti (2023) employed dimensionality reduction techniques such as t-SNE after feature selection, with the aim of detecting breast cancer. As t-SNE managed to produce a subset from which the best characteristics were extracted, these features were then applied to the suggested ensemble model for breast cancer detection. Hence, it is evident that t-SNE is effective in dimensionality reduction, contributing to a decent model performance in disease predictions.

2.3 Unsupervised Learning Approach

Unsupervised ML models have adeptly discerned patterns across a wide array of data types, encompassing of rainfall data and sedimentation data for hydrological field (Loh, et al., 2021; 2024), text document data for text mining (Yang and Lee, 2012), imaging data for medical fields which includes the diagnosis of tumours and abnormal cell growths. On top of that, in the medical field, they have also been applied to categorise diseases based on patients' symptoms and medical histories, thereby identifying the patient subgroups requiring personalized treatment (Rankovic, et al., 2023).

For instance, a study on different ML models including the unsupervised SOM was carried out by Rankovic, et al. (2023) to predict 17 different chronic diseases such as coronary heart disease, asthma and hypertension. A high cholesterol is observed as the most prevalent disease in Serbia by analysing the heatmap represented by SOM. Aside from considering clinical test attributes, the coronary heart disease diagnosis using electrocardiogram signal is achievable and it was studied by Rath, et al. (2022) with the development of SOM. An accuracy of 93.5% was achieved with predictions made by SOM alone. On top of that, the combination of SOM and autoencoder, an unsupervised neural network, that was further proposed exhibited the best performance with an accuracy of 98.4%.

Following the studies mentioned above, it is evident that SOM is effective in predicting outcomes. A hybrid approach based on SOM and other models yields even better performance. In fact, there are a number of research done in similar ways. For instance, in order to predict the risk of fatal heart disease incidence in Type 2 Diabetes Mellitus, a common form of diabetes, SOM was proposed by Zarkogianni, et al. (2018) to be combined with Hybrid Wavelet Neural Network which yielded a decent performance with AUC of 71.48%. As such issues associated with unbalanced nature of data and nonlinearities could be resolved. Another case study by Osman and Alzahrani (2019) on epilepsy disease, a brain disease, offered an alternative way for identifying unknown patterns in the epilepsy dataset using SOM. This study presented an automatic epilepsy diagnostic method based on SOM using radial basis function neural networks, achieving an overall detection accuracy of

97.47%. It was the best prediction performance in comparison to other methods such as wavelet chaotic neural network and radial basis function neural network. Henceforth, SOM demonstrates its effectiveness in predictions, while also being observed to be highly compatible with other models in hybrid approaches.

2.4 Optimisation Algorithms

Optimisation is a great helper in enhancing effectiveness and performance of ML algorithms. Many researchers have incorporated optimisation on classifiers to improve prediction accuracy after removing irrelevant features. For instance, after incorporating Genetic Algorithm (GA) to obtain optimal subset of features, Soni, et al. (2011) reported that the accuracy of Decision Tree (DT) and Naive Bayes (NB) further increased in heart disease prediction, at 99.2% and 96.5% respectively. Amin, et al. (2013) employed Artificial Neural Network (ANN) and GA for heart disease prediction. Using GA to optimise the connection weights of the neural network for prediction produced a result showing an accuracy of 89%.

Some researchers have explored the fusion of two optimisation algorithms in disease prediction. For example, aiming to predict heart disease accurately, Sheeba, et al. (2022) introduced a hybrid optimisation approach where the Moth-Flame Optimisation algorithm was combined with the Deer Hunting Optimisation algorithm to optimise its weight function. After the implementation, there was a significant increase in the accuracy of the optimised Recurrent Neural Network at 91.4%. Besides, Subanya and Rajalaxmi, Lin and Hsieh made significant contributions by hybridising their models with swarm intelligence-based Artificial Bee Colony (ABC) algorithms in 2014 and 2015, respectively. Subanya and Rajalaxmi (2014) applied ABC algorithm to select the best features and found that ABC-SVM performed better than feature selection with reverse ranking on Cleveland heart disease. It is shown when ABC-SVM achieved an accuracy of 86.76% while the method of reverse ranking yielded an accuracy of 85%. Besides, a hybrid evolutionary algorithm using endocrine-based PSO and ABC coupled with SVM was proposed by Lin and Hsieh (2015) for extracting the optimal feature subsets. The superiority of such hybrid algorithm was proven when the performance of SVM with feature

selection achieved higher accuracy then that without feature selection. The endocrine-based PSO achieved the highest accuracy, reaching 91.35%.

In fact, a wide range of such uses of metaheuristic algorithms can be found in illness prediction in the medical sector (Nssibi, et al., 2023; Malakar, et al., 2023). Considering the effectiveness of metaheuristic algorithms, Malakar, et al. (2023) have favoured to employ PSO for feature selection when applying them to medical datasets that are associated with chronic illnesses. According to the authors, this approach was selected over the others primarily because it can be used to address a wide range of challenging optimisation. They were able to determine the best-case scores for Heart disease, Breast Cancer, Chronic Kidney disease, and Diabetes to be 96.72%, 99.82%, 100.00%, and 84.41%, respectively after a comprehensive series of tests. Furthermore, with the purpose of predicting heart disease, Verma, et al. (2016) developed a hybrid model using two ML classifiers, K-Nearest Neighbours (KNN) and Multilayer Perceptron (MLP) with PSO as well and achieved an accuracy of 90.28%. Moreover, in the study by Raja and Pandian (2020), the popular PSO optimisation technique was proven effective when it was combined to improve performance while maintaining the distinctive characteristics of Fuzzy Clustering Means (FCM). Premature convergence, a limitation of FCM, was prevented by incorporating PSO and resulted in a higher accuracy of 95.42%. This indicates that the predicted model, PSO-FCM, is successful in predicting the onset of diabetes earlier. In short, the integration of PSO with other optimisation techniques for hybrid approaches yields favourable results.

2.5 Summary

All in all, unsupervised ML models including t-SNE and SOM have gained popularity in the medical field for higher efficiency in disease diagnosis. The methods discussed above demonstrate how various algorithms enhance evaluation metrics such as accuracy and precision. However, it is noticeable that there remains scope for further improvement in the developed models. Therefore, it is undeniable that optimisation like PSO is essential to make it the best possible ML model in analysing and predicting disease outcomes.

To provide a concise overview of the unsupervised ML models discussed above, a summary of analysis, specifically SWOT (strengths,

weaknesses, opportunities, and threats) analysis, is carried out as shown in tables below. As stated in Table 2.1, in terms of dimensionality reduction, it is clear that t-SNE excels at visualising high-dimensional data in lower dimensions, preserving local structures. However, it can be computationally expensive for large datasets. In contrary, PCA is limited to capture linear relationships while maintaining orthogonality of components.

Apart from that, unlike supervised learning models, SOMs are effective for clustering and dimensionality reduction, as described in Table 2.2. They improve the understanding of relationships between variables but require careful handling of parameters. Furthermore, a few optimisation algorithms, which are PSO, GA and ABC, are analysed as shown in Table 2.3. It is clear that PSO is efficient and globally searches for optimal solutions. It is computationally less complex compared to GA, besides having relatively faster convergence than ABC. However, it may converge to local optima and requires parameter tuning.

Table 2.1: SWOT Analysis of t-SNE and PCA.

Algorithm	Strength	Weakness	Opportunity	Threat
t-SNE	Excellent at visualising high-dimensional data in lower dimensions	Sensitive to hyperparameters	Enhanced feature selection and pattern discovery	Computationally expensive for large datasets
PCA	Maintains orthogonality of components	May not capture nonlinear relationships	Integration with other algorithms for better results	Sensitivity to outliers and noise

Table 2.2: SWOT Analysis of SOM.

Algorithm	Strength	Weakness	Opportunity	Threat
SOM	Effective for clustering and dimensionality reduction	Sensitivity to initialization parameters	Better understanding of data topology	Potential for model overfitting
	Nonlinear dimensionality reduction technique	Interpretability can be challenging	Application in various fields such as image processing, natural language processing	

Table 2.3: SWOT Analysis of Optimisation Algorithms.

Algorithm	Strength	Weakness	Opportunity	Threat
PSO	Global search capability due to swarm dynamics	Convergence to local optima if not properly tuned	Application in various optimisation problems	Sensitive to parameters
GA	Suitable for optimization with constraints	Computational complexity	Development of parallel and distributed versions	Difficulty in handling high-dimensional problems
ABC	Less sensitive to parameter settings	Slower convergence compared to PSO and GA	Enhancement through hybridization with other algorithms	Vulnerable to noise and outliers

In a nutshell, the advantages of SOM such as excellent visualisation through heatmaps and effective clustering are evident. Pairing it with PSO can significantly boost the accuracy of SOM for better performance in heart disease prediction.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction

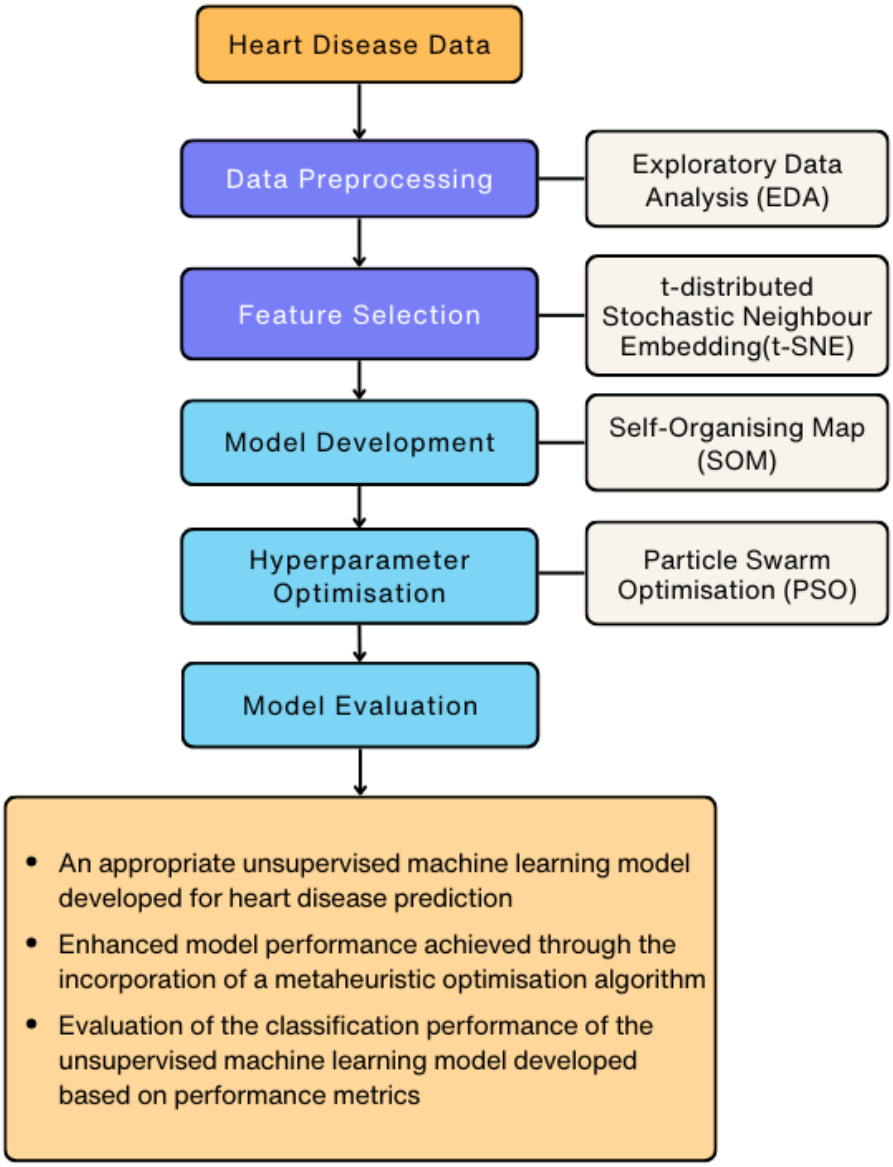

Figure 3.1: Flow Chart of the Proposed Project.

The figure above outlines the overall workflow of the project. The workflow commences with data preprocessing, where raw data is cleaned, scaled, and

prepared for analysis. Subsequently, low-dimensional visualisation using t-SNE provides a visual representation of the high-dimensional data. This aids in feature selection, identifying potentially relevant attributes for predicting heart disease while considering correlation between features.

An unsupervised SOM is then developed as the primary predictive model, utilising the R programming environment for model development and optimisation. The SOM model is implemented using the *kohonen* package, which provides a comprehensive set of functions for constructing and training SOMs. For optimising the hyperparameters of SOM, PSO is incorporated with the *pso* package, a metaheuristic algorithm known for its efficacy in fine-tuning model parameters, enhancing the model's predictive accuracy.

Finally, the trained SOM is evaluated using various performance metrics, such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's effectiveness in predicting the presence of heart disease. Additionally, visualisations like heatmaps can be plotted to identify patterns within the data and visualise the SOM's structure.

3.2 Dataset Description

The dataset used for this project is the Heart Disease Dataset, available on the UCI Machine Learning Repository. There are 303 instances in total. In addition, this dataset contains 13 features. Each feature is described in further detail in Table 3.1 below.

Table 3.1: Feature Description of the Heart Disease Dataset (UCI Machine Learning Repository, 1988).

Feature	Description
age	Age of the patient (in years)
sex	Gender (Female: 0, Male: 1)
cp	Chest pain type (0: typical angina, 1: atypical angina, 2: non-angina pain, 3: asymptomatic)
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol (in mg/dl)
fbs	Fasting blood sugar > 120mg/dl (0: False, 1: True)
restecg	Resting electrocardiogram (0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (0: no, 1: yes)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of peak exercise ST segment (0: upsloping, 1: flat, 2: downsloping)
ca	Number of major vessels coloured by fluoroscopy (0 - 3)
thal	Thalassemia type (0: normal, 1: fixed defect, 2: reversible defect)
target	Heart disease (0: negative, 1: positive)

There are 5 numerical features, including age, trestbps, chol, thalach and oldpeak. The rest of the attributes are categorical with 3 being binary (sex, fbs, exang) and 5 having multiple categories (cp, restecg, slope, ca, thal). In addition, the study focuses on predicting the binary presence of heart disease as the target variable, classifying between individuals with heart disease and those without. This dataset is selected to be used to develop and test an unsupervised ML model for heart disease prediction.

3.3 Dimensionality Reduction

Despite the fact that conventional research aimed to predict heart disease effectively, the lack of dimensionality reduction and feature selection would significantly affect the accuracy (Barfungpa, et al., 2023). Therefore, an effective means using ML techniques to increase the prediction rate is essential for the diagnosis of heart disease.

Due to its powerful mapping capability, the t-SNE, which was created by Maaten and Hinton (2008), has gained popularity in the ML industry. As t-SNE tends to preserve both global and local structures, it is introduced in this study to reduce dimensionality of nonlinear data. It is useful for managing SOM because it promotes the easy comprehension of high-dimensional data and their transformation into a low-dimensional space (Barfungpa, et al., 2023). This approach computes pairwise similarity between the data points to convert the distances between them into Gaussian joint probabilities, preserving the similarity between the high-dimensional data and mapping it into a low-dimensional space.

Suppose that a set of n D-dimensional points $\mathrm{X} = \{x_1, x_2, \dots, x_n \mid x_i \in \mathcal{R}^D\}$ is to be mapped into a low-dimensional space, $Y = \{y_1, y_2, \dots, y_n \mid y_i \in \mathcal{R}^K\}$, where $K < D$. The t-SNE first computes the conditional probability of x_i choosing x_j as its neighbour, which is denoted by $p_{j|i}$ and is defined as

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \tag{3.1}$$

where

σ_i = vector variance of the Gaussian function centred on the data point x_i.

When $p_{i|i} = 0$, the joint probability p_{ij} in high-dimensional space is defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{3.2}$$

Similar to Equation (3.2), in low-dimensional space, when σ_i of conditional probability $q_{j|i}$ equals $\frac{1}{\sqrt{2}}$, the joint probability q_{ij} is defined as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|y_k - y_j\|^2)^{-1}} \tag{3.3}$$

Subsequently, t-SNE addresses the crowding problem by employing a heavy-tailed distribution to the embedded low-dimensional data points. Furthermore, Kullback–Leibler divergences between *Q* and *P* are calculated with the gradient descent below.

$$C = KL(P|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.4}$$

Then, the gradient of the Kullback–Leibler divergence is given by

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \tag{3.5}$$

From Equation (3.5), the solution is given by

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)}) \tag{3.6}$$

where

$Y^{(t)}$ = the solution obtained from the *i*th iteration,

η = learning rate,

$\alpha(t)$ = momentum of the *i*th iteration.

3.4 Model Development

In this study, SOM is proposed to be built to predict heart disease. The SOM is a special kind of ANN that Kohonen (1982) introduced. SOM is made up of a single layered grid of neurons in two dimensions, in contrast to the ANN

structure. Every grid node has a direct connection to the input vector, but there is no connectivity between the nodes. The grid is a representation of the map that organises itself based on the input data at each.

As stated by Rath, et al. (2022), there are four steps in the self-organisation process. Every connecting weight is initially set at random in the initialization process. The neuron that produces the least amount of value wins after each one computes its function. The winning neuron, often referred to as the best matching unit (BMU), locates its location within a close range of other neurons during the cooperation step. Lastly, by adjusting the connected weights, the individual discriminant values of each activated neuron are reduced in the adaptation stage. The stepwise procedure in SOM is as follows:

i. Initialisation:
   The weight vectors w_j are initialised randomly for each neuron.

ii. Sampling:
   A sample input vector x_i is chosen from a set of training input data.

iii. Cooperation:
   The subsequent step involves the identification of the BMU, w_j, with the closest weight vector. The similarity between the input vector x_i and weight vectors of each other neuron is determined by computing the distance between them, that is, Euclidean distance, as shown in Equation (3.7).

$$d_j(x) = \sqrt{\sum_{i=1}^{d} (x_i - w_{ji})^2} \tag{3.7}$$

where
d = number of features,
w_j = weight vector for the neuron j,
x = data example.

iv. Updating:

The weights of neurons are then updated with the weighting rule defined below.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha_i(t) \cdot \beta_{cj}(t)\left[x(t) - w_{ij}(t)\right] \quad (3.8)$$

$$\beta_j(t) = \exp\left(-\frac{w_j - w_j^{*2}}{2\sigma^2(t)}\right) \quad (3.9)$$

where

α = learning rate at time *t*,

j = index of BMU,

i = *i*th feature of the training example,

$\beta_j(t)$ = neighbourhood function,

$\sigma(t)$ = width of the kernel corresponding to neighbourhood radius.

Furthermore, in SOM, the topology, that is, the choice of grid shape—hexagonal or rectangular—significantly influences the map's ability to represent data. Hexagonal grids are normally favoured for their superior topological representation and visual appeal, as each neuron has six neighbours, creating a uniform neighbourhood structure that minimises distortions in the map's representation of high-dimensional data (Chaudhary, Bhatia and Ahlawat, 2014). Hence, this structure promotes excellent clustering and accurate data visualisation. In contrast, rectangular grids, while simpler and computationally efficient due to their alignment with Cartesian coordinates, can introduce distortions in the neighbourhood relationships due to the four-neighbour configuration (Chaudhary, Bhatia and Ahlawat, 2014). However, the decision between hexagonal and rectangular grids should be guided by the specific dataset and the characteristics of the map. Balancing accuracy with computational efficiency is crucial. A practical approach to determine the optimal grid type involves a trial-and-error methodology. By training SOMs with both grid types under consistent parameters and evaluating performance

using metrics like quantisation error, one can determine which grid type provides the best balance for the SOM model.

3.5 Hyperparameter Optimisation

The process of optimisation involves the identification of the optimal solution from a set of options available that either maximises or minimises an objective function for a particular problem (Nguyen, et al., 2023). The problems in which feasible solutions are restricted due to constraints can take place in real-world situations. As stated by Nguyen, et al. (2023), the following notation shows how the problems can be expressed:

$$\begin{gathered} \min_{x} f(x), \\ subject\ to \\ g_i(x) \leq 0, i = 1,2, \dots, m \\ h_j(x) \leq 0, j = 1,2, \dots, p \\ x \in X \end{gathered} \tag{3.10}$$

where

$g_i(x)$ and $h_j(x)$ = constraint functions,

X = set of possible values of x.

There are a few hyperparameters in SOM, including the size of the SOM grid, η - learning rate, σ – the bandwidth of the neighbourhood function shown in Equation (3.9) as well as the number of iterations. Their values remain constant throughout the training phase, and they can only be chosen in advance (Nguyen, et al., 2023). Aside from the hyperparameters, the model parameter, weight, can be optimised as well. They can be updated throughout the training phase.

In this study, the hyperparameters selected for optimisation include the grid size of the SOM (specifically dimensions *x* and *y*), the number of iterations, the learning rate, and the neighbourhood radius. The grid size determines the resolution of the SOM. Larger grids can capture more complex patterns but require more computation time. Besides, the number of iterations indicates the

number of training iterations or epochs the SOM undergoes, affecting how well the SOM learns from the data (Wehrens and Buydens, 2007). More iterations typically improve the model's ability to capture complex patterns but also lengthen computation time. Furthermore, the learning rate is a parameter that controls the extent to which the weights of the SOM nodes are adjusted during training (Wehrens and Buydens, 2007). It is typically defined within a range of values specified by a learning rate schedule. In addition, the radius determines the size of the neighbourhood around a winning node that is updated during training. It typically starts large and decreases over time.

The rationale for optimising multiple hyperparameters, rather than focusing on just one, is due to the inherent sensitivity of SOM to these hyperparameters. Addressing this weakness ensures more robust model performance. Hence, to optimise these hyperparameters, PSO is introduced in this study.

The PSO is a prevalent metaheuristic technique that can be applied to resolve optimisation issues (Nguyen, et al., 2023). It is metaheuristic as it explores potential solutions within a vast search space without making significant assumptions about the problem being optimised. This flexibility allows it to effectively tackle complex optimisation tasks across various fields. As introduced by Kennedy and Eberhart (1995), PSO attempts to obtain the best outcomes by imitating the social behaviour of a flock of fish or a swarm of birds. A swarm of particles in PSO act as potential solutions that move throughout the search space to find the best options. The swarm's movements are guided by improved positions.

One of the key advantages of PSO is that it is easy to implement and requires minimal parameter adjustment. According to Raja and Pandian (2020), PSO consists of three fundamental steps, repeated until all particles converge to a certain point where optimal value is attained.

i. Determine the fitness of each particle.
ii. Continuously update the best individuals and global functionalities.
iii. Update the position and speed of each particle.

Every particle is situated with each swarm's best fit. The particle status update formula is given by

$$v_{ij}(t+1) = v_{ij}(t) + r_1\left(P_{best_{ij}} - x_{ij}(t)\right) + r_2\left(G_{best_j} - x_{ij}(t)\right) \quad (3.11)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (3.12)$$

where

r_1 and r_2 = random numbers within the interval $(0,1)$,

v_{ij} = velocity for the i^{th} particle in j^{th} dimension,

x_{ij} = position for the i^{th} particle in j^{th} dimension,

$P_{best_{ij}}$ = personal best solution,

G_{best} = global best solution.

Algorithm 1: Proposed PSO-SOM Algorithm

Initialise particles with the hyperparameters vector $\boldsymbol{h}$ = [map size, number of iterations, learning rate, radius] randomly
Define the accuracy as the objective function
do
  for each particle
    Train SOM with the assigned $\boldsymbol{h}$
    Calculate the fitness value (accuracy)
    If the accuracy is better than P_{best} in history
      Set the current value as the new P_{best}
  end for
  Choose the particle with the best accuracy from all particles as the G_{best}
  for each particle
    Calculate the particle velocity according to Equation (3.11)
    Update the particle position according to Equation (3.12)
  end for
while (maximum iterations or minimum criteria is not attained)
return the best $\boldsymbol{h}$ and corresponding accuracy

The proposed algorithm above integrates PSO with SOM to optimise key hyperparameters, with the objective of enhancing the accuracy of the SOM. In this approach, the fitness function of the PSO is defined as the accuracy of the SOM model, which serves as the primary metric for evaluating the

performance of each candidate solution. Each particle in the PSO swarm represents a unique configuration of SOM hyperparameters, specifically the map size, number of iterations, learning rate, and radius. These hyperparameters are encapsulated within a vector $\boldsymbol{h}$, where $\boldsymbol{h} =$ [map size, number of iterations, learning rate, radius]. In other words, in the proposed PSO-SOM algorithm, each particle represents a potential $\boldsymbol{h}$, while the accuracy of SOM serves as the fitness value or solution quality.

Through this process, the algorithm systematically converges on the hyperparameter set that yields the highest SOM accuracy. Finally, the outcome of the algorithm is the identification of the optimal hyperparameters—map size, number of iterations, learning rate, and radius—that produce the best-performing SOM model. This integration of PSO not only automates the hyperparameter tuning process but also enhances the overall model accuracy, making it a powerful tool for optimising SOM in medical applications.

3.6 Model Evaluation

At the final stage, a confusion matrix is exported as the model prediction outcome as illustrated in Figure 3.2 below.

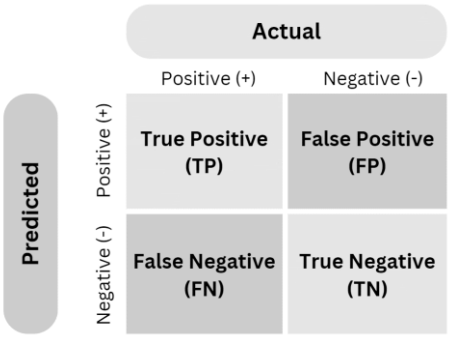

Figure 3.2: Layout of a Confusion Matrix.

The confusion matrix obtained is used to compute several evaluation metrics, including accuracy, sensitivity, specificity, precision and F1 score. The evaluation metrics to be computed are described as follows:

i. Accuracy (ACC) refers to the ratio of the total number of correctly classified predictions to the total number of instances.

$$ACC = \frac{TN + TP}{TN + TP + FP + FN} \times 100\% \quad (3.13)$$

ii. Sensitivity (SN), also known as recall, refers to as True Positive rate, is the fraction of true positive predictions among all actual positive predictions.

$$SN = \frac{TP}{TP + FN} \times 100\% \quad (3.14)$$

iii. Specificity (SP), also referred to as True Negative rate, is the fraction of true negative predictions among all actual negative predictions.

$$SP = \frac{TN}{TN + FP} \times 100\% \quad (3.15)$$

iv. Precision refers to the proportion of correctly predicted positive class to all positive class predictions.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3.15)$$

v. F1 score, also is referred to as F score, or F measure, is a measure of the harmonic mean of precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3.16)$$

While these basic metrics provide useful information, additional elements can also be considered for specific models. In the context of SOM, Quantisation Error (QE) serves as an important metric that evaluates how well the SOM model maps the input data onto the grid. The QE represents the average squared distance between each data point and its corresponding BMU on the map (Kohonen, 2001). It is mathematically represented as Equation (3.17) below.

$$QE = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \bar{w}_i\| \tag{3.17}$$

QE quantifies how well the map's prototypes represent the data points. A lower QE indicates that the map's prototypes are closer to the data points, reflecting better representation and organisation of the data within the SOM (Kohonen, 2001). Thus, minimising QE is crucial for enhancing the model's performance and ensuring accurate data mapping. Including such metrics offers a more comprehensive assessment of the model's performance and overall effectiveness.

On the other hand, visualisation plays a crucial role in providing insights into the performance of the SOM model. The unsupervised nature of the SOM allows for a few meaningful visualisations. Among these, heatmaps are valuable tools for analysing and uncovering relationships between variables. Heatmaps offer a detailed view of the data distribution across the SOM, illustrating the distribution of individual variables. Additionally, the Unified Distance Matrix (U-Matrix) is essential for visualising clustering patterns, highlighting the distances between data points and their respective clusters. These visualisations can collectively enhance the understanding of the SOM model's performance and the underlying data structure.

By incorporating these additional elements alongside numeric performance metrics, it provides a more comprehensive and insightful evaluation of the SOM model's performance.

3.7 Gantt Chart

The following Gantt charts describe the timeline of the study and the key milestones achieved for Project I and Project II.

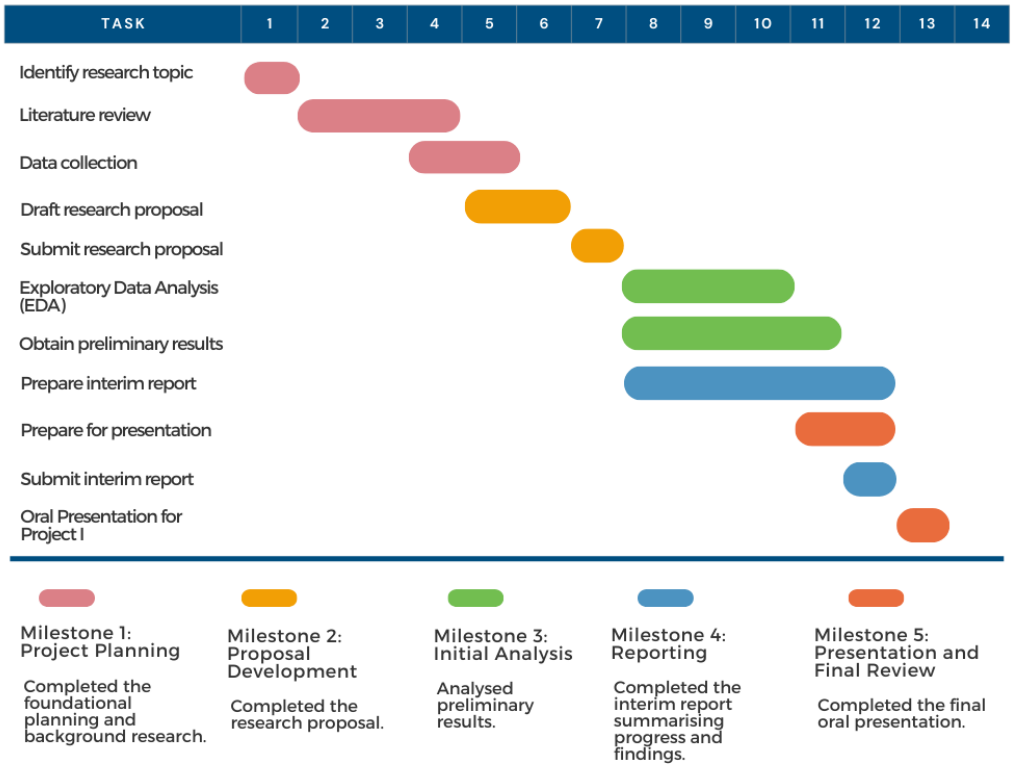

Figure 3.3: Project Milestone Gantt Chart for Project I.

Project I mainly focused on literature review on existing studies related as well as data exploration for better understanding of the dataset used. Practical implementation of methodology following EDA was carried out with the aim of developing an unsupervised model, SOM. Its output was presented as the preliminary result.

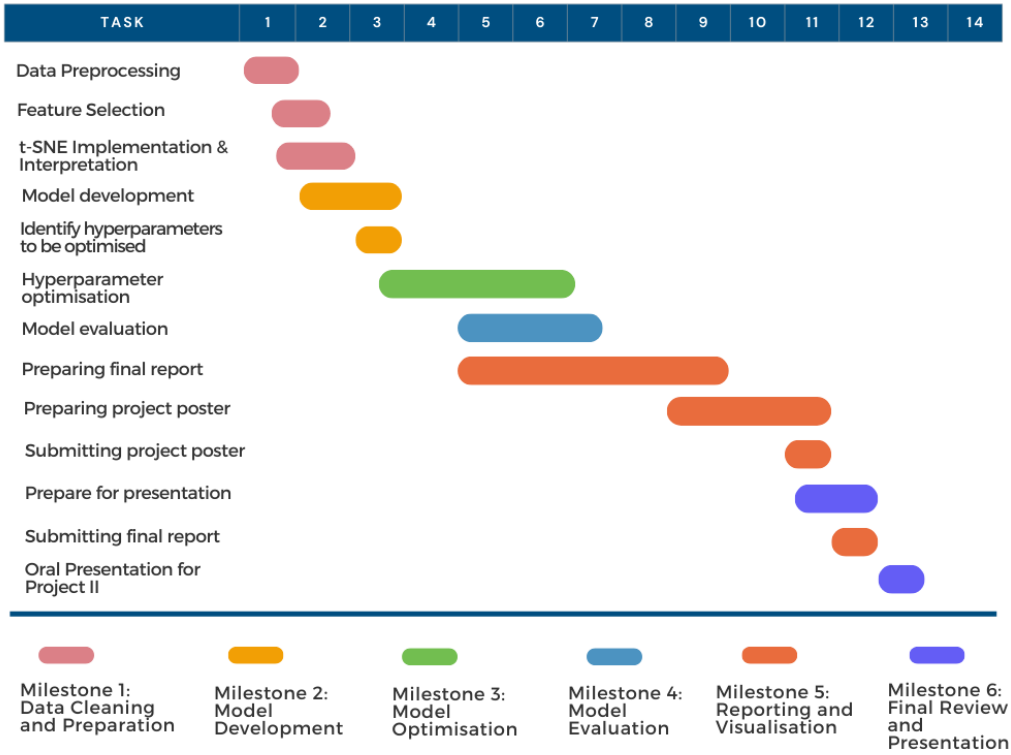

Figure 3.4: Project Milestone Gantt Chart for Project II.

Subsequently, Project II was proceeded with the incorporation of PSO, a metaheuristic algorithm for enhancing the performance of the developed SOM model in predicting heart disease. The outcomes were evaluated and documented in the final report.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) is a critical and fundamental step to initiate a general understanding and insights of the dataset before implementation of ML. It involves statistical analysis, summarising main characteristics, identifying patterns and relationships using various graphical plots. In this study, the heart disease data is explored to provide a comprehensive overview.

In total, there are 303 instances in the dataset. It consists of 13 features, excluding the target variable. Five of them are numerical and eight are categorical, specifically three binary and five multi-categorical. All the values for each column are originally present in type integer except the column 'oldpeak' which has floats. However, it was observed that the values in the 'ca' and 'thal' columns are of the object type due to the presence of unknown values ('?'). There are six unknown values present in the 'ca' and 'thal' columns, accounting for less than 5% of the data. Hence, these rows were removed, resulting in 297 instances left. Apart from that, there were no null values detected in the dataset.

The target variable consists of unique values of 0, 1, 2, 3 and 4. Note that the values 1, 2, 3 and 4 convey the same meaning that the patient has heart disease. Hence, value reassigning was performed to group them together as a class, represented by the value 1, while value 0 indicates the absence of heart disease. The summary statistics were tabulated as shown in Table 4.1 below.

Table 4.1: Summary Statistics for Numerical Features.

Feature	Min	Max	Median	Mean	Standard Deviation
age	29.0	77.0	56.0	54.54	9.05
trestbps	94.0	200.0	130.0	131.69	17.76
chol	126.0	564.0	243.0	247.35	52.00
thalach	71.0	202.0	153.0	149.60	22.94
oldpeak	0.0	6.2	0.8	1.06	1.17

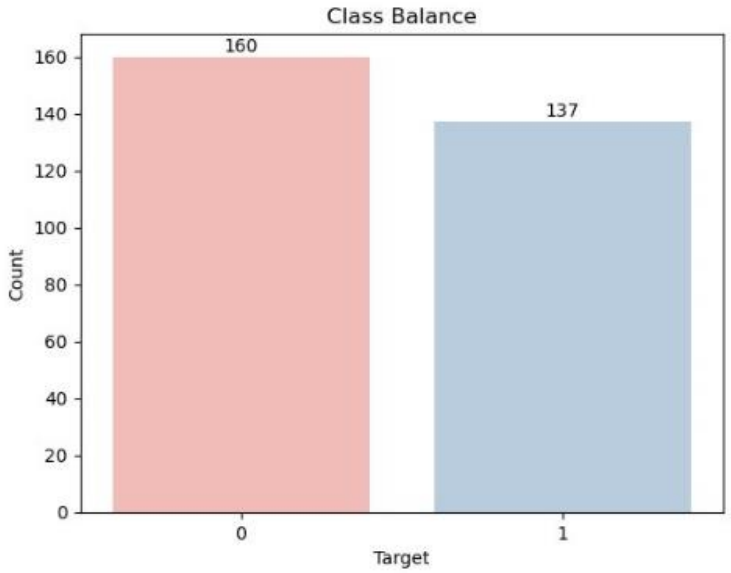

Figure 4.1: Class Balance.

Aside from the missing value issues, ensuring class balance for the predicted target labels is also a crucial factor as it can significantly impact the prediction performance of a given model. As depicted in Figure 2, the class distribution is fairly even. The total number of patients without heart disease is slightly higher than those with heart disease.

4.1.1 Univariate Analysis

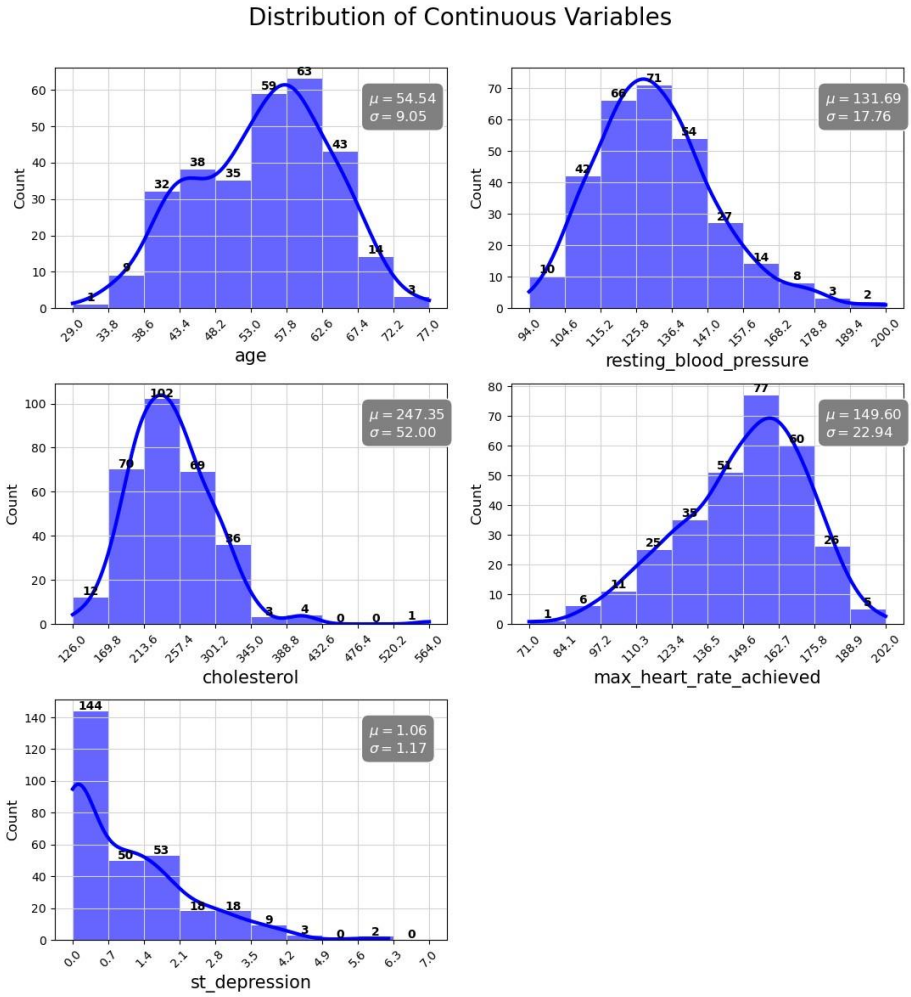

Figure 4.2: Distribution of Continuous Variables.

Based on Figure 4.2 above, Age and ST Depression are lack of variation as they have low standard deviation. It is observed that age follows a normal distribution. Resting Blood Pressure has a moderately right-skewed distribution while Cholesterol and ST Depression show highly right-skewed distributions. Conversely, the distribution of Maximum Heart Rate Achieved is moderately left-skewed. In addition, outliers are detected in the distribution of Cholesterol and ST Depression. Nevertheless, in medical data, it often exists rare but significant events where extreme conditions or unusual phenomena occur in certain patients, represented by outliers. To preserve the genuine information and maintain the integrity of the data, outliers are not discarded in this case.

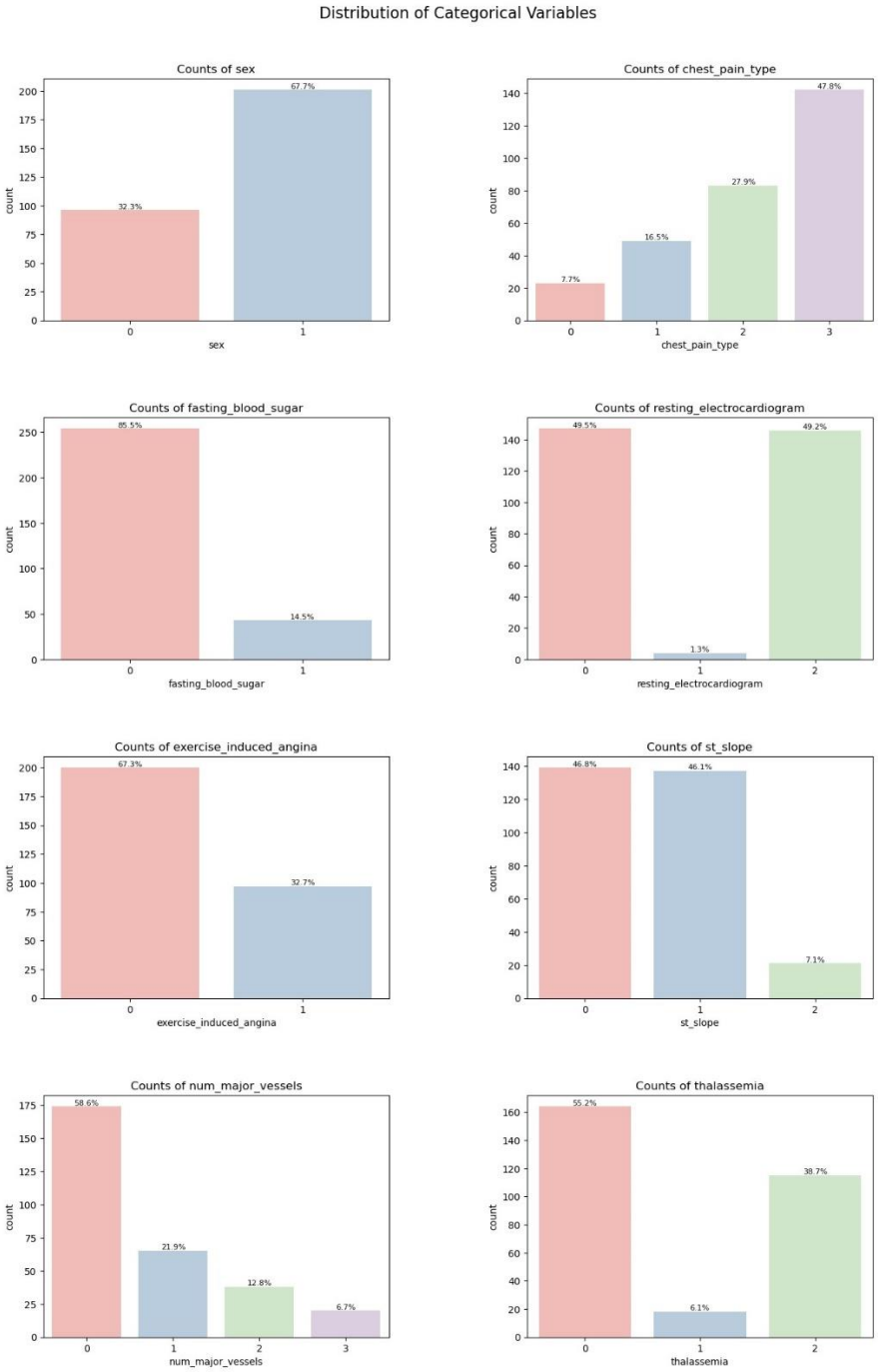

Figure 4.3: Distribution of Categorical Features.

According to the bar charts as shown in Figure 4.3 above, there are more male patients than female patients. Majority do not experience chest pain, also known as asymptotic (type 3) and have fasting blood pressure that is less than 120 mg/dl (value 0). The individuals having resting electrocardiogram of type 1 (having ST-T wave abnormality) are the least, barely any. Normal (type 0) and left ventricular hypertrophy (type 2) patients are of almost the same

amount. Besides, there is a greater portion of patients who do not experience exercise-induced angina. The number of patients with upsloping and flat ST slopes account for similar percentages and are higher than those with downsloping. Patients with 0 major vessels coloured by fluoroscopy are obviously much more numerous than those with major vessels. In terms of thalassemia, patients with normal blood flow (type 0) have the highest distribution, followed by reversible defect, whose blood flow is not normal (type 2) and then fixed defect (type 1), where there is no blood flow in a specific part of the heart.

4.1.2 Correlation

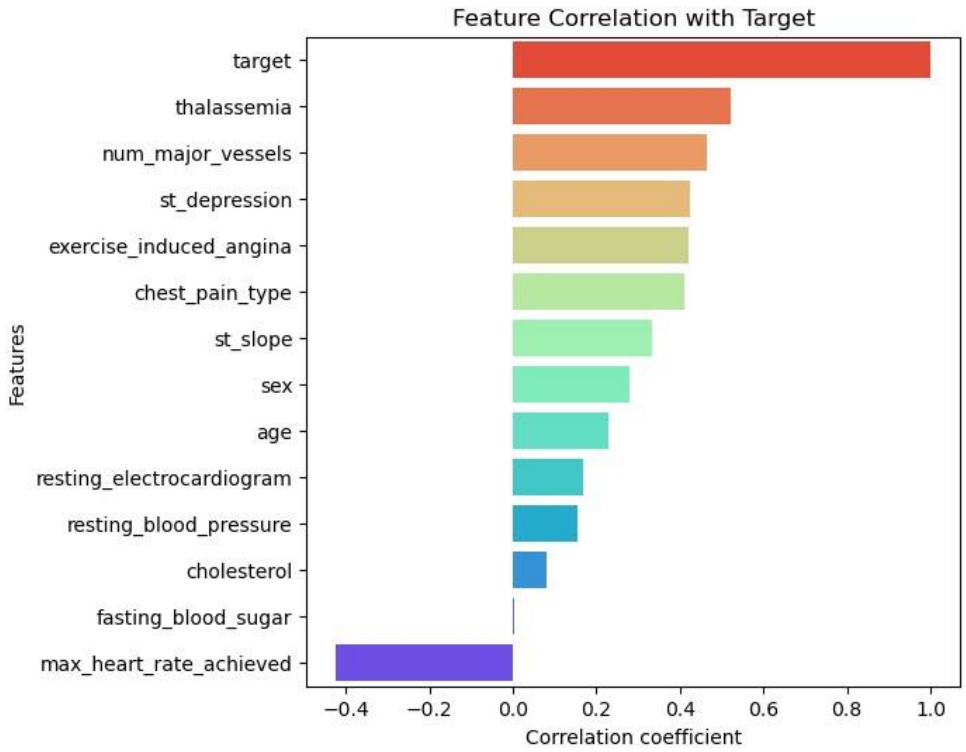

Figure 4.4: Correlation between Features and Target Variable.

Pearson's Correlation is a popular method to quantify the strength of linear relationship between variables using correlation coefficients. Ranging from -1 to 1, a correlation coefficient reflects the direction of the relationship, indicating whether they are positively correlated, negatively correlated, or not correlated at all. Figure 4.4 above displays the sorted correlation between each

feature and target variable. There are 6 features relatively being highly correlated with target variable, having absolute coefficients greater than 0.4. The high positive correlation involves features like Chest Pain Type, Exercise-induced Angina, ST Depression, Number of Major Vessels and Thalassemia. Meanwhile, Maximum Heart Rate Achieved is the only feature having a negative correlation with the target variable. In contrast, when correlation coefficients fall below 0.2, they signify extremely weak correlations (Akoglu, 2018). In light of this criterion, the 4 features with the lowest correlation coefficients, which are less than 0.2, are highlighted as having very weak correlations with the target variable.

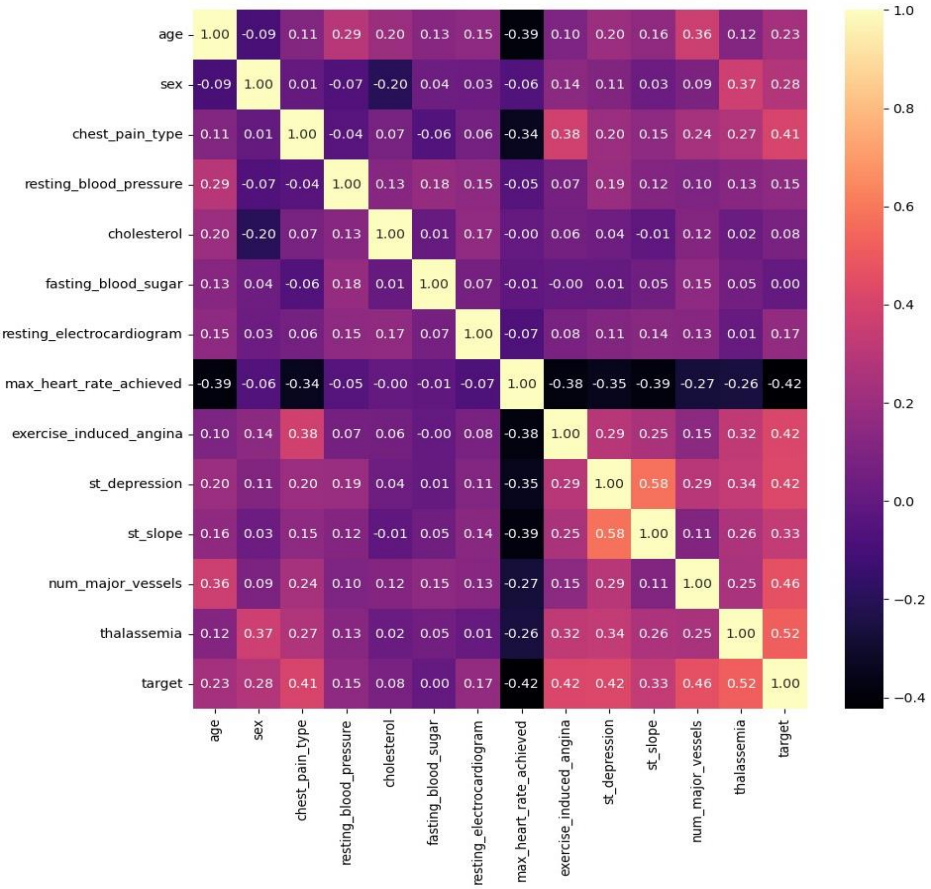

Figure 4.5: Correlation Matrix for All Features.

Aside from focusing on target variable, Figure 4.5 above shows an overview of the correlation between all features. As the dataset contains quite a number of features, the following bivariate analysis will focus selectively on variables having high correlation, enabling a more insightful exploration.

4.1.3 Bivariate Analysis

First and foremost, a bivariate analysis between each feature and target variable is performed. Continuous features and categorical features are explored separately.

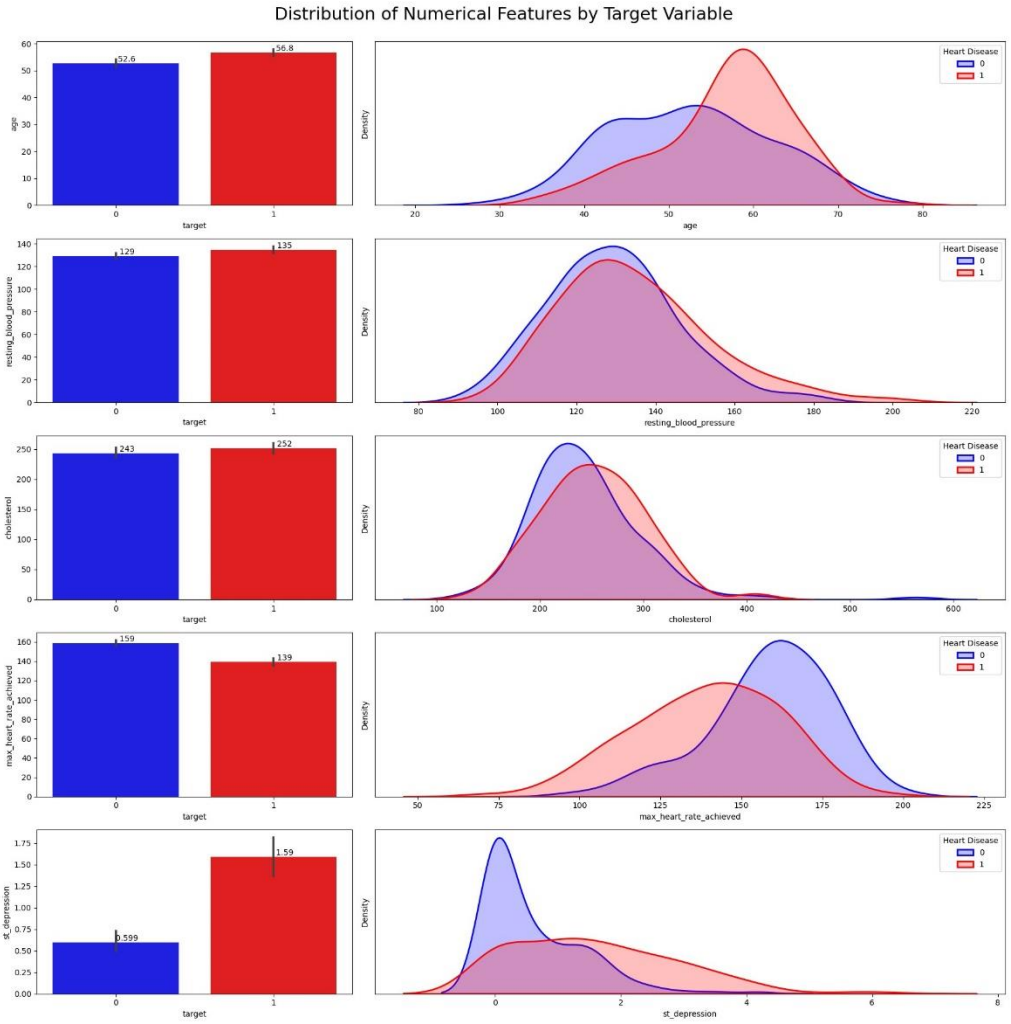

Figure 4.6: Distribution of Continuous Features by Target Variable.

Connecting the features with the target variable is crucial for uncovering their relationship. The distribution of each continuous feature by the target variable is visualised as shown in Figure 4.6 above. It is observed that most patients aged about 60 are diagnosed with heart disease. This finding aligns with existing studies indicating that adults aged around 65 are more likely to develop heart disease compared to younger individuals (National Institute on Aging, 2018).

Upon examining the plot, it becomes evident that Resting Blood Pressure and Cholesterol may not be significant factors in determining heart disease, as their distributions for both classes are similar. In addition, a normal range of blood pressure is less than 120 mm Hg, while blood pressure exceeding 140 mm Hg is considered high, also known as hypertension (CDC, 2021). In the case of cholesterol, measurements below 200 mg/dL are within the normal range, while levels above 240 mg/dL are considered high cholesterol (Johns Hopkins Medicine, n.d.).

Furthermore, the distribution reveals that populations with a higher maximum heart rate are less likely to suffer from heart disease. In fact, determining whether a heart rate is normal or abnormal is contingent on age, which is a topic that will be explored further in the discussion later with Figure 4.11. It is also evident that individuals with lower ST depression are at lower risk of getting heart disease, as the peak of the blue bell curve is pronounced and narrow at the value zero. Higher ST depression tend to appear among heart disease patients as existing studies have stated that values below 0.5 mm are generally considered normal, while values exceeding this threshold are considered pathological (Rawshani, 2021).

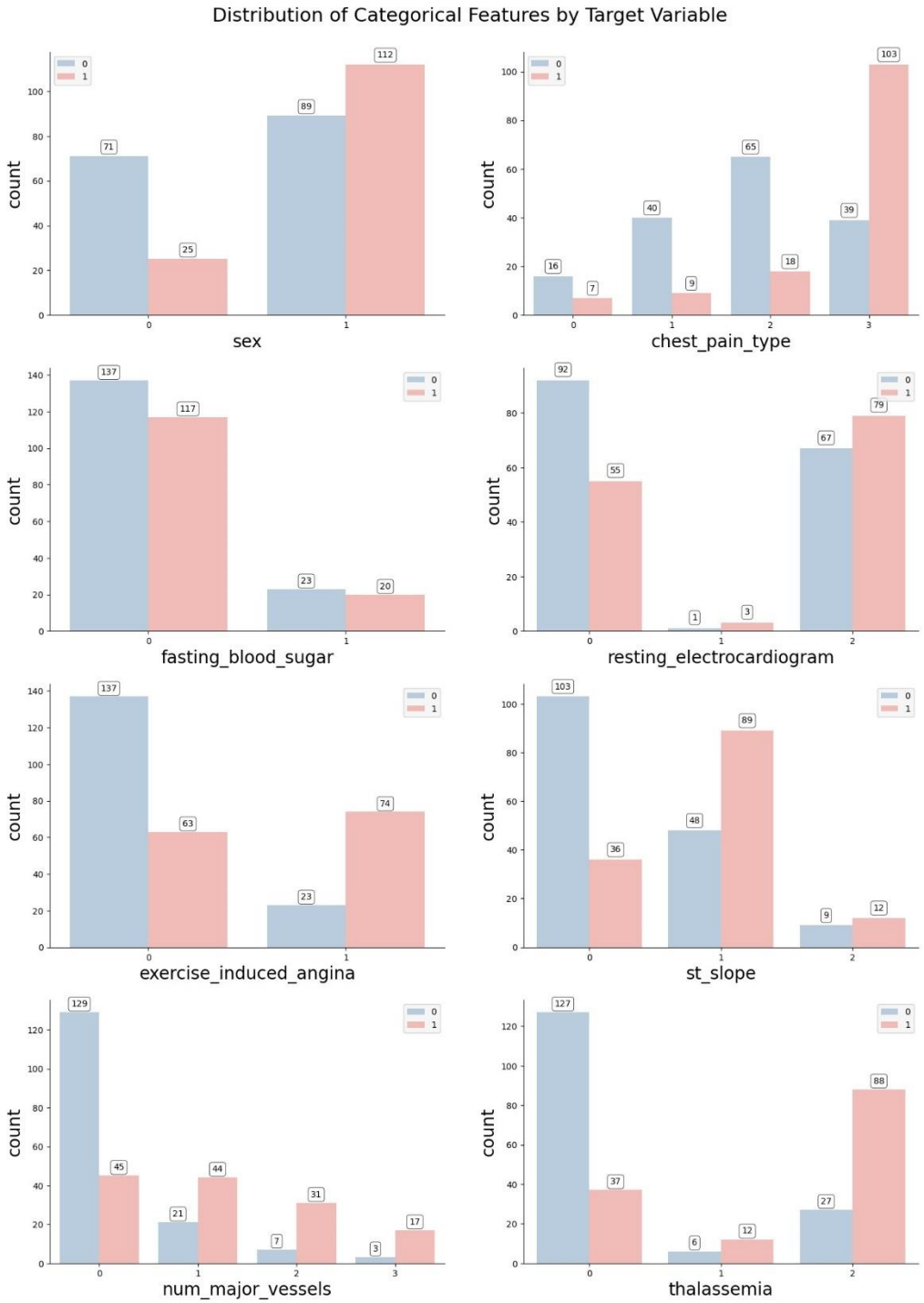

Figure 4.7: Distribution of Categorical Features by Target Variable.

The bar charts relating categorical features to the target variable are displayed in Figure 4.7 above. Chest pain, also known as angina, is a temporary discomfort that happens when the heart does not receive sufficient blood and oxygen supply. It could be one of the symptoms of heart disease. Still, many heart disease patients do not experience any chest pain symptoms, which is

evident from the plot showing chest pain type 3 (asymptomatic) is higher than the other types. Besides, it can be inferred that fasting blood sugar is not a good indicator of heart disease since healthy and unhealthy population distribution is almost the same for each category.

A resting electrocardiogram, commonly known as an ECG, is a graphical representation of the heart's electrical activity when one is at rest. This output is analysed to detect any abnormalities in the heart's rhythm or structure. A closer look at the plot reveals that normal patients (value 0) and patients with left ventricular hypertrophy (value 2) have a much higher incidence of heart disease. There are very less patients with ST-T wave abnormality (value 1).

Moreover, the relationship between exercise-induced angina and heart disease seems to be direct. Individuals without exercise-induced angina are less likely to develop heart disease while those experiencing exercise-induced angina tend to have a higher likelihood of heart disease. Additionally, ST slope refers to the direction in which the ST segment moves during peak exercise on ECG. The figure illustrates that individuals with flat sloping (value 1) have a higher incidence of heart disease. Upsloping ST segments (value 0) apparently suggest a lower risk of developing heart disease. This inference is explained by the National Centre for Biotechnology Information (2002) which states that flat and downsloping ST segment indicates a higher probability of heart disease.

Furthermore, the number of major vessels coloured by fluoroscopy with values 0-3 reflects the severity of heart disease based on the number of major vessels, specifically coronary arteries that are affected or blocked. The fluoroscopy is used to visualise the blood flow through major vessels with dye. As depicted from the figure, patients with no vessels coloured by fluoroscopy have a much lower incidence of heart disease because value 0 represents that there are probably no major vessels that show significant blockages or narrowing, indicating unobstructed blood flow to the heart. In terms of thalassemia, patients with normal condition (value 0) are more likely to be free from heart disease but reversible defect thalassemia (value 2) has a significantly higher incidence of heart disease.

Subsequently, the following section discusses about the bivariate analysis between one feature and another.

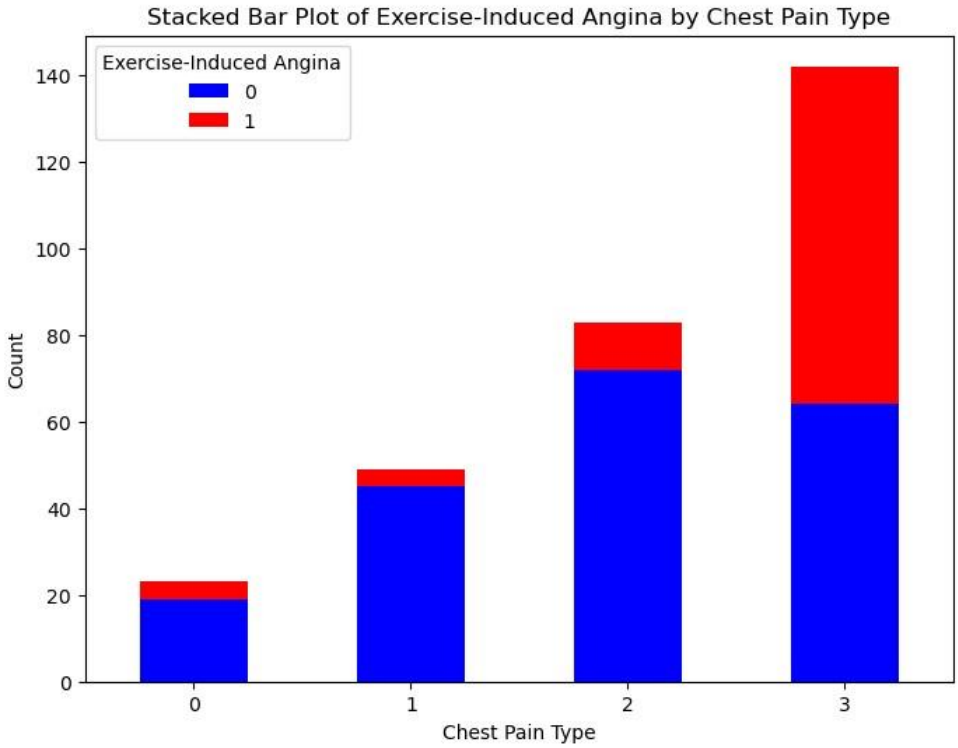

Figure 4.8: Exercise-Induced Angina by Chest Pain Type.

According to the previous analysis regarding the correlation in Figure 4.5, the correlation between exercise-induced angina and chest pain type is noticeably moderate. As mentioned before, angina is a temporary chest pain. Exercise-induced angina, also known as exertional angina or angina pectoris, is a chest pain that occurs during physical activity. Hence, they are put together for further exploration as they are relatively related. From the stacked bar plot, it can be seen that chest pain type 0 (typical angina), type 1 (atypical angina) and type 2 (non-angina pain) all shows very few instances of exercise-induced. On the other hand, chest pain type 3 (asymptomatic) has the highest count overall, with a significant major portion experiencing exercise-induced angina. Therefore, it suggests that patients with asymptomatic chest pain are more likely to experience exercise-induced angina. Meanwhile, typical angina and atypical angina patients have the least likelihood of suffering from exercise-induced angina.

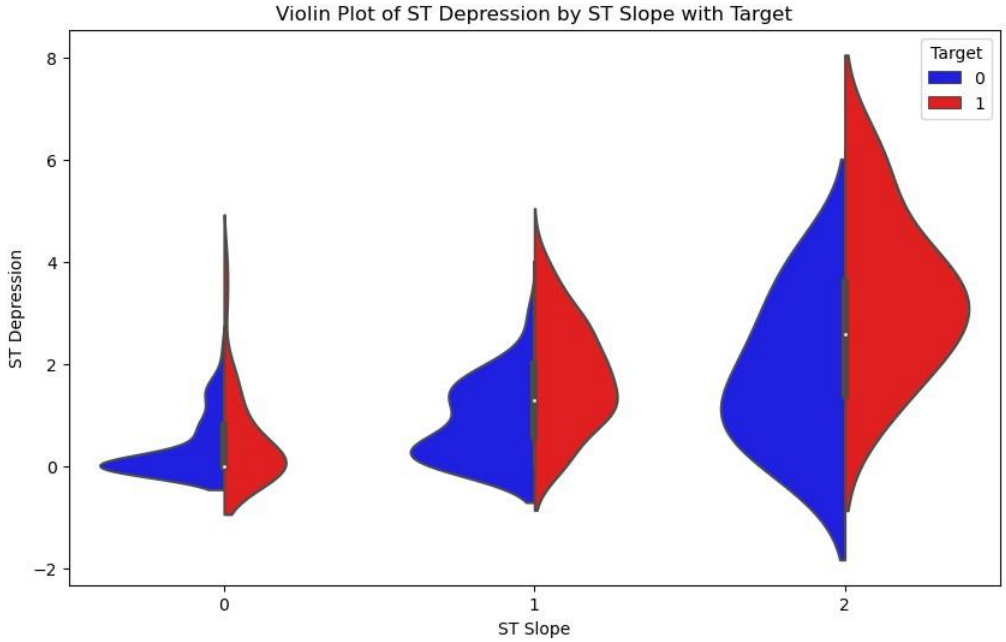

Figure 4.9: Violin Plot of ST Depression by ST Slope with Target.

Based on Figure 4.5 that displays the correlation matrix, the highest correlation is observed between the features ST depression and ST slope, with a coefficient of 0.58. Looking at the upsloping ST segment slope (value 0), the violin plot shows overlapping distributions. However, there is a significant shift towards lower ST depression values for individuals without heart disease. On top of that, there is less distinction in ST depression levels between the two groups for flat ST slope (value 1). The peak around the ST depression of 0.5 with a flat slope suggests common amounts for both target groups. Then, with downsloping ST slope (value 2), individuals with higher ST depression values tend to have heart disease. In contrast to the rest, this violin plot exhibits a more extensive range of ST depression values. In short, higher levels of ST depression within a downsloping ST segment may signal an increased probability of heart disease, whereas a lower degree of ST depression in an upsloping ST segment could imply a lower risk of heart disease.

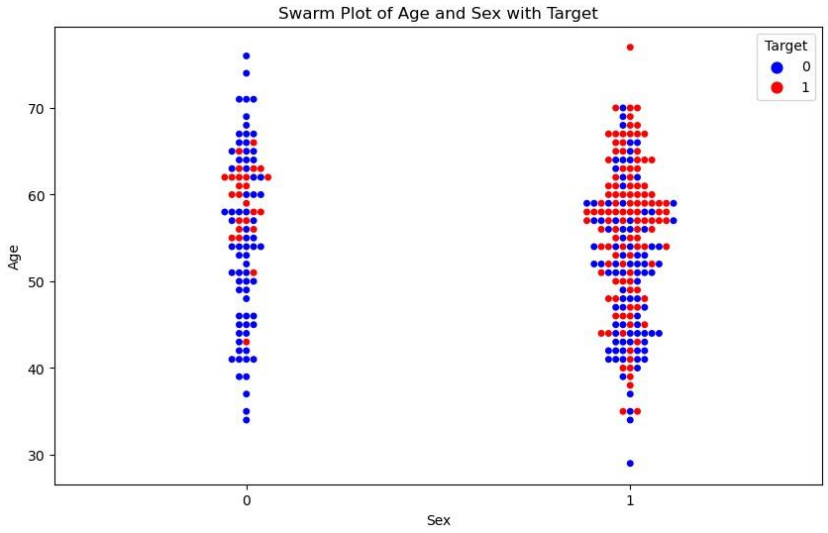

Figure 4.10: Swarm Plot of Age by Sex with Target.

To provide a comprehensive view of the demographics of the dataset, a swarm plot illustrating age and sex alongside the target variable has been generated, as shown in Figure 4.10 above. As illustrated in the figure, a male with younger age tends to be exposed to a higher risk of heart disease compared to a young female. The risk tends to increase significantly as male individuals start entering the age of 60 and above. On the other hand, females tend to develop heart disease with age ranging approximately from 55 to 65, and that is the period when the menopause hits.

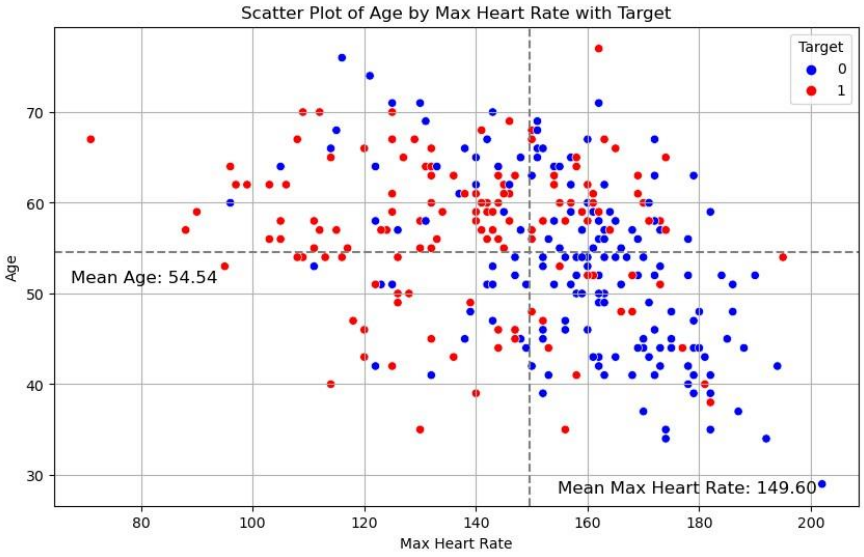

Figure 4.11: Scatter Plot of Age by Maximum Heart Rate with Target.

On another note, a moderate correlation between age and maximum heart rate achieved is observed, with a coefficient of -0.39 as shown in Figure 4.5. The scatter plot in Figure 4.11 above illustrates that patients with and without heart disease are aged between 40 to 70 years old. The spread of maximum heart rate of individuals without heart disease ranges from 140 to 180. From the plot, it appears that patients who tend to be less prone to heart disease have achieved maximum heart rate over 149 and under the age of 54. It can also be seen overall that age and maximum heart rate has a negative correlation.

Further to this, conditions like heart disease may actually lead to lowering the maximum heart rate (Harvard Medical School, 2023). In addition, age is required to determine whether the maximum heart rate achieved is acceptable. To be more precise, during moderate-intensity activities, a normal maximum heart rate typically falls within the range of 50-70% of the maximum heart rate, while during vigorous physical activity, it is typically ranges from 70-85% of the maximum (American Heart Association, 2021). For instance, according to studies by American Heart Association, 60-year-old individuals are expected to have a maximum heart rate of 160 beats per minute (bpm). However, upon the review of the plot, it becomes evident that a subset of heart disease patients around this age have achieved a maximum heart rate lower than 160 bpm, in contrast to those without heart disease. Therefore, the presence of heart disease is likely to result in a reduced maximum heart rate achieved by a patient.

4.2 Feature Selection

Before proceeding to SOM model development, by removing redundant or irrelevant features and applying dimensionality reduction can help improve the performance of SOM. As such the model can focus better on the significant features and learn more robust patterns from the data.

From Figure 4.5 displaying the correlation matrix, it becomes evident that there exists at least a feature that is very less correlated with the rest of the features. Such variables include resting blood pressure, cholesterol, fasting blood sugar and resting electrocardiogram. Their coefficients are extremely low, that is, less than 0.2. This implies very weak correlations, as described by

Akoglu in 2018. The fasting blood sugar even shows no correlation with the target variable, with a Pearson's correlation coefficient of 0. Hence, these 4 features become strong candidates to be removed. Nevertheless, correlation measures linear relationships between variables. In other words, it might not be able to capture nonlinear relationships effectively. Hence, t-SNE was implemented as it is a powerful tool for visualising high-dimensional data in low dimension. It excels in capturing nonlinear relationships, thereby preserving the local structure of the dataset.

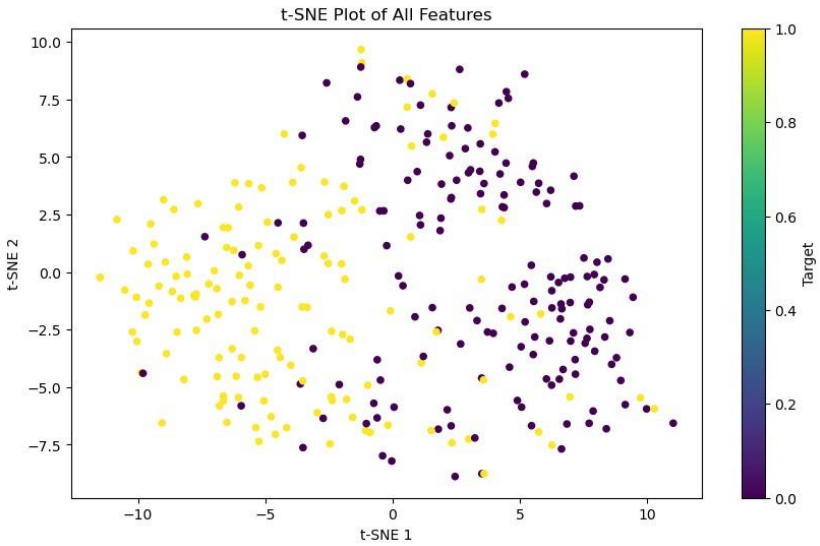

Figure 4.12: t-SNE Plot for All Features.

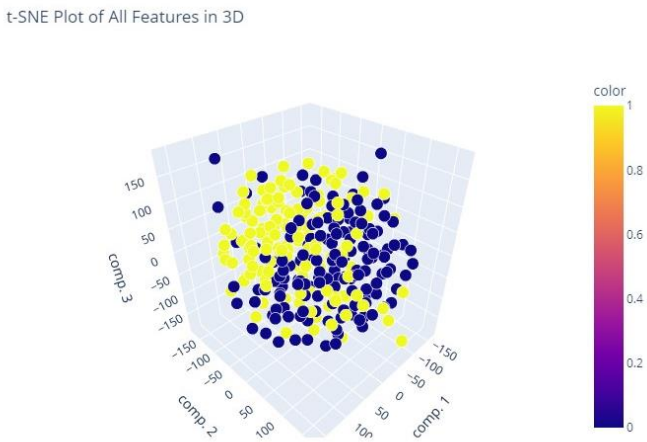

Figure 4.13: t-SNE Plot for All Features in 3D.

The t-SNE plot in Figure 4.12 above was generated by including all features in the dataset. One cluster is predominantly yellow, suggesting a grouping of individuals without heart disease. The other cluster is predominantly purple, indicating individuals with heart disease. From the plot, it is obvious that there is some degree of overlap. The area where yellow and purple points mix represents individuals whose feature values are similar to both groups. This could indicate that borderline cases or noise existing in the data. This t-SNE output was visualised in three dimensions as well, as shown in Figure 4.13. Similarly, there exists some obvious overlapping areas. Hence, a plot in Figure 4.14 below was further generated by excluding the four features that have the least correlation.

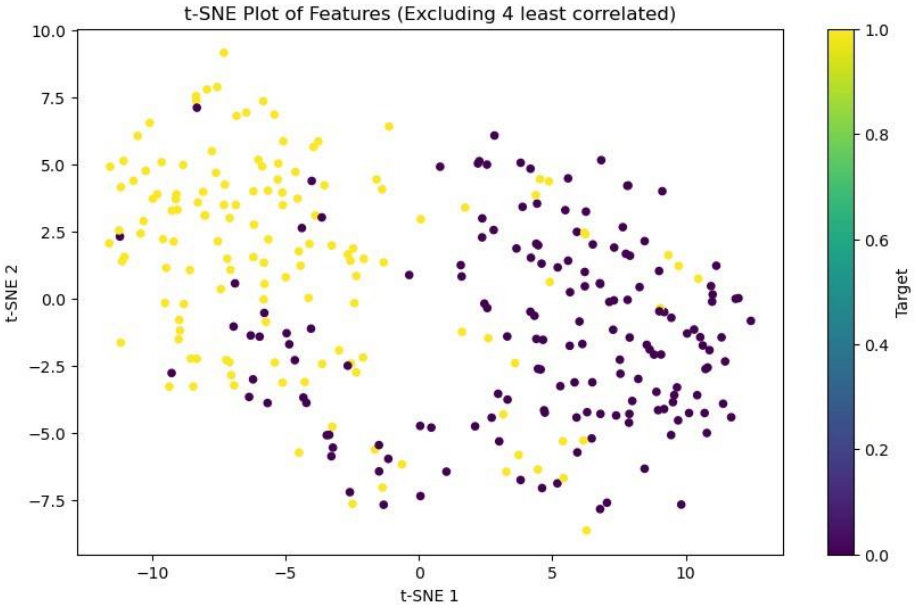

Figure 4.14: t-SNE Plot Excluding Least Correlated Features.

Comparing with Figure 4.12, it is evident that two clusters are fairly well-separated, which suggests that the features used to generate the t-SNE plot perform a good job of distinguishing between the two groups. The separation between the two groups became clearer, which could be indicative of strong predictive features that could be useful for building a classification model to predict heart disease. In a nutshell, the 4 features having the least correlation

were to be removed. They are resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs) and resting electrocardiogram (restecg).

4.3 Model Development

Prior to developing an unsupervised SOM model, some preprocessing steps were implemented. Categorical features underwent encoding, while numerical features were processed using min-max scaling. This scaling procedure ensures that numerical values are transformed to a standardised range between 0 and 1. Besides, the dataset was divided into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing.

Leveraging the *kohonen* package in the R programming environment, the development of a SOM model becomes both accessible and efficient. The *somgrid* function is critical in structuring the SOM by determining the grid dimensions and topology (Wehrens and Buydens, 2007). On this grid, each neuron or unit represents a cluster of similar observations, grouping data points that share common characteristics.

Subsequently, the *xyf* function was utilised as it is suited for scenarios where labelled data is available, as exemplified in this study. This function is a variant of the Kohonen SOM, combining unsupervised clustering and supervised classification. It is particularly useful for supervised mapping, enabling the mapping of high-dimensional data to a lower-dimensional grid while considering a target variable that guides the training process (Wehrens and Buydens, 2007). In this study, heart disease status (yes or no) served as the target variable, providing a supervised learning component to the analysis. By synergising the SOM's unsupervised clustering capabilities with supervised learning, the *xyf* function ensured that the SOM not only organises data based on similarities but also aligns the map with the relationship between features and heart disease outcomes (Wehrens and Buydens, 2007). This resulted in a more structured map, organising the data into distinct regions corresponding to specific heart disease outcomes.

Moreover, in the context of SOM topology, the two tables below, Table 4.2 and Table 4.3, present a comparison of performance metrics for rectangular and hexagonal SOM topologies across various grid dimensions.

Table 4.2: Performance Metrics of Rectangular Topology across Various Grid Dimensions of SOM.

Dimension	Accuracy	Precision	Recall	F1 Score	QE
(5, 5)	0.5833	0.7333	0.6471	0.6875	0.0476
(6, 6)	0.6531	0.7250	0.8286	0.7733	0.0425
(7, 7)	0.6250	0.7407	0.7143	0.7273	0.0368
(8, 8)	0.6190	0.6970	0.7931	0.7419	0.0301
(9, 9)	0.6047	0.7419	0.7188	0.7302	0.0256
(10, 10)	0.6279	0.7500	0.7500	0.7500	0.0252
(11, 11)	0.6757	0.8400	0.7241	0.7778	0.0238
(12, 12)	0.5349	0.6786	0.6333	0.6552	0.0247
(13, 13)	0.6000	0.8000	0.6667	0.7273	0.0217
(14, 14)	0.5952	0.7500	0.6774	0.7119	0.0230
(15, 15)	0.5333	0.7647	0.5652	0.6500	0.0212

Table 4.3: Performance Metrics of Hexagonal Topology across Various Grid Dimensions of SOM.

Dimension	Accuracy	Precision	Recall	F1 Score	QE
(5, 5)	0.5625	0.7692	0.5714	0.5385	0.0498
(6, 6)	0.5306	0.6765	0.6571	0.6667	0.0431
(7, 7)	0.6053	0.7917	0.6552	0.7170	0.0373
(8, 8)	0.6522	0.7813	0.7353	0.7576	0.0293
(9, 9)	0.5778	0.6667	0.7742	0.7164	0.0256
(10, 10)	0.6122	0.7105	0.7714	0.7397	0.0247
(11, 11)	0.6190	0.7059	0.8000	0.7500	0.0247
(12, 12)	0.5556	0.6071	0.7727	0.6800	0.0227
(13, 13)	0.5588	0.7368	0.5833	0.6512	0.0219
(14, 14)	0.5455	0.7500	0.5625	0.6429	0.0197
(15, 15)	0.5357	0.6842	0.6500	0.6667	0.0193

To determine the suitable topology type, it is essential to ensure a fair comparison between hexagonal and rectangular topologies. This requires using

a constant grid dimension for both topologies, ensuring that the grid size and number of neurons are identical, However, for a thorough and reliable comparison, using just one constant grid dimension is typically insufficient. Hence, by varying the grid dimension from 5 x 5 to 15 x 15, it becomes possible to capture the impact of topology on performance metrics. This range was chosen to cover a variety of grid sizes ranging from small to large, providing a more comprehensive evaluation.

Based on the comparative results of both topologies above, it is evident that the rectangular topology generally outperforms the hexagonal topology across all metrics. Consequently, the subsequent SOM model development was conducted using the rectangular topology.

Furthermore, as discussed in Section 3.6, key hyperparameters of SOM include grid size, number of iterations, learning rate, and neighbourhood radius. These hyperparameters must be predefined before training, as they remain constant throughout the training process. In fact, SOM is highly sensitive to their hyperparameters, which can significantly impact the model performance (Astudillo and Oommen, 2014). For instance, small changes in grid size can alter the SOM's topology, while variations in learning rate can affect convergence speed.

Hence, in this study, an initial guess followed by a trial-and-error approach was first employed to set the required hyperparameter for the SOM model development. While initial guess and trial-and-error methods are common approaches for setting these hyperparameters, they can be time-consuming and may lead to suboptimal results. This is often a trade-off between accuracy and development time. In some cases, a suboptimal hyperparameter configuration might be acceptable if it allows for a quick and reasonably accurate model, especially under time constraints.

To assess the performance of the developed SOM model, several performance metrics were employed. By varying the grid dimension while keeping other hyperparameters constant, the impact of these hyperparameters on the model's behaviour was explored. The Table 4.4 below presents some of the results obtained with varying grid dimensions through trial and error. For

clarity, the Table 4.4 displays accuracy metrics sorted in ascending order to facilitate easier comparison.

Table 4.4: Performance Metrics of SOM Model with Varying Grid Dimensions through Trial and Error.

Grid x	Grid y	Accuracy	Precision	Recall	F1 Score	QE
5	6	0.5510	0.7600	0.5429	0.6333	0.0444
9	13	0.6000	0.7619	0.6957	0.7273	0.0242
10	14	0.6053	0.7407	0.7143	0.7273	0.0228
15	12	0.6053	0.7143	0.7407	0.7273	0.0224
12	8	0.6111	0.7500	0.6923	0.7200	0.0264
8	12	0.6136	0.7273	0.7500	0.7385	0.0258
14	7	0.6154	0.8182	0.6207	0.7059	0.0264
9	12	0.6190	0.7500	0.7500	0.7500	0.0238
5	10	0.6364	0.7222	0.8125	0.7647	0.0324
7	6	0.6364	0.7576	0.7576	0.7576	0.0393
11	13	0.6512	0.7500	0.7742	0.7619	0.0252
13	15	0.6552	0.7222	0.7222	0.7222	0.0222
7	8	0.6585	0.7931	0.7419	0.7667	0.0336
11	7	0.6585	0.8000	0.6897	0.7407	0.0255
6	14	0.6596	0.7368	0.8235	0.7778	0.0280
6	9	0.6667	0.8077	0.7241	0.7636	0.0288
14	15	0.6800	0.7500	0.8333	0.7895	0.0207
13	5	0.6818	0.8214	0.7188	0.7667	0.0289
8	13	0.6944	0.7857	0.8148	0.8000	0.0252
10	5	0.6977	0.8276	0.7500	0.7869	0.0350
15	9	0.7143	0.8148	0.8148	0.8148	0.0244
12	9	0.7179	0.7500	0.8889	0.8136	0.0256

Through trial-and-error, the highest accuracy of SOM that could be obtained was 71.79% with grid dimensions of 12 x 9. The remaining metrics were well-balanced, demonstrating the SOM's capability of organising the data and predicting heart disease. Before optimisation, such trial-and-error approach was used to explore different grid sizes for the SOM model. The best configuration found through this method served as a reference for comparing performance improvements after applying PSO optimisation. This comparison

would help evaluate the effectiveness of the optimisation process in enhancing the SOM model's performance.

Additionally, from this approach, one can infer that determining the optimal set of hyperparameters to achieve outstanding model performance can be both challenging and time-consuming when relying solely on trial-and-error methods. This approach may not only fail to yield satisfactory results but also makes it difficult to confirm whether the chosen hyperparameters represent the best possible configuration.

To address these limitations, additional hyperparameter tuning techniques are necessary to be employed. These techniques can often help achieve better model performance while reducing the time and effort required for hyperparameter tuning. In this study, the metaheuristic optimisation algorithm, PSO, is proposed to enhance the performance of the SOM model, with the outcomes discussed in the following section.

4.4 Hyperparameter Optimisation

In an effort to further enhance the SOM model’s effectiveness, PSO was employed after the SOM model development. Utilising the *pso* package in R environment, the *psoptim* function, serving as the Particle Swarm Optimiser, is introduced in this study for the general implementation of PSO.

The *psoptim* function is configured with several parameters (Bendtsen, 2022). The *par* parameter represents the initial parameter vector. The *fn* parameter specifies the fitness function, which was designed to be maximised to improve the accuracy of SOM. The *lower* and *upper* parameters define the lower and upper bounds for the hyperparameters, respectively. The *control* parameter is a list of control parameters, including *maxit* set to 1500, which represents the maximum number of iterations for the PSO algorithm.

Five hyperparameters of the SOM are optimised in this process. Table 4.5 below outlines the settings for the lower and upper bounds of these SOM hyperparameters, which are to be optimised through the PSO-SOM algorithm.

Table 4.5: Bounds for SOM Hyperparameters Optimised by the PSO-SOM Algorithm.

SOM Hyperparameter	Lower Bound	Upper Bound
Grid x	5	15
Grid y	5	15
Number of Iterations	500	2000
Maximum Number of Iterations	0.01	0.5
Neighbourhood Radius	1	10

This setup ensures that the PSO-SOM algorithm effectively searches for the optimal hyperparameters to maximise the accuracy of the SOM. By incorporating PSO into the SOM model, an optimal solution was obtained. The optimal hyperparameter settings for SOM are presented in Table 4.6 below.

Table 4.6: Optimal Hyperparameter Settings of SOM.

Hyperparameter	Optimal Value
Grid x	15
Grid y	13
Number of Iterations	776
Learning Rate	[0.2572, 0.01]
Neighbourhood Radius	[4.5304, 1]

With these optimal hyperparameter settings, the SOM model was trained to achieve improved performance.

4.5 Model Evaluation

Evaluating the performance of the developed SOM model involves assessing both quantitative metrics and qualitative visualisations to ensure reliable and interpretable outcomes. This section discusses the numerical performance metrics used to evaluate the SOM model. Additionally, visualisations such as heatmaps, U-Matrix, and training progress plots are utilised to provide insights into the model performance, which are essential for interpreting the underlying structure and the learning process of the SOM model.

4.5.1 Performance Metrics

The performance of the developed SOM model was reassessed following the integration of PSO. Table 4.7 below presents a comparison of the SOM model's performance before and after the PSO-SOM algorithm.

Table 4.7: Comparative Results of the SOM Model Before and After the PSO-SOM Algorithm.

Metric	Before PSO	After PSO
Accuracy	0.7179	0.9444
Precision	0.7500	1.0000
Recall	0.8889	0.9286
Specificity	0.3333	1.0000
F1 Score	0.8136	0.9630
QE	0.0256	0.024

Based on the result, it is evident that the incorporation of PSO led to significant improvements across all evaluated metrics. The optimised SOM model achieved a maximum accuracy of 94.44%, an improvement from the previous accuracy of 71.79%, reflecting a significant enhancement in its overall classification performance. Besides, precision improved from 0.75 to 1, demonstrating that the optimised model now identifies true positive cases with barely any false positives.

In heart disease prediction, while accuracy is crucial, recall and specificity are particularly important to avoid misclassifications that could impact patient outcomes. Recall rose from 0.8889 to 0.9286, indicating a slight reduction in missed positive cases. This is crucial in heart disease prediction as missing a diagnosis (false negatives) can have serious consequences for patients. Specificity experienced a dramatic increase from 0.3333 to 1, showing the model’s capability to correctly identify negative cases without false positives. This is important to avoid false alarms where healthy individuals are incorrectly classified as having heart disease, thereby reducing unnecessary treatments and ensuring that only those who truly need intervention are identified.

Additionally, the F1 score, which balances precision and recall, improved from 0.8136 to 0.9630, highlighting the model's improved

performance and effectiveness. Furthermore, the QE decreased marginally from 0.0256 to 0.0240, suggesting a slight reduction in the error associated with mapping input vectors to the nearest neurons, thus indicating better accuracy and consistency of the model. Overall, these results highlight the effectiveness of PSO in optimising the SOM model's hyperparameters, leading to great improvement in predictive accuracy and reliability.

4.5.2 Kohonen Map Visualisation

Upon developing the optimised SOM model, the Kohonen Map can be visualised using a variety of plots to interpret clusters and relationships within the data. Below are some common plots used for visualising SOM in R, along with explanations for each.

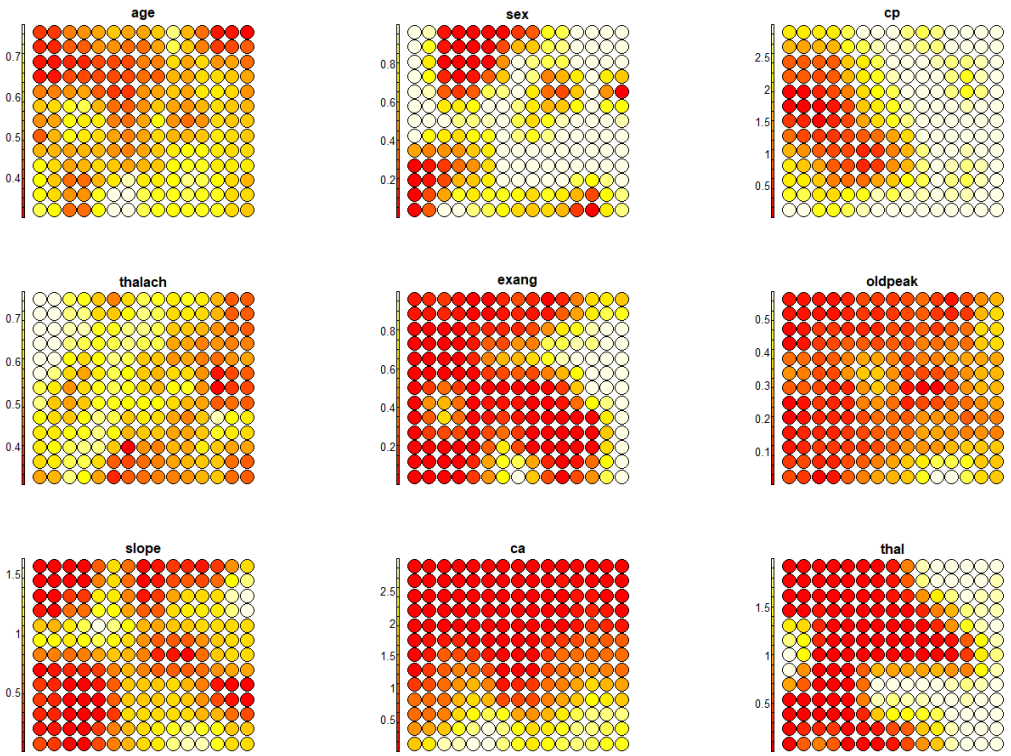

Figure 4.15: Feature Heatmaps of SOM.

Figure 4.16 above illustrates SOM's heatmap visualisations. These component planes show the distribution of data for each feature across the SOM grid. They are useful for identifying and understanding the overall structure of

the data. Red indicates low values and white indicates high values. By comparing the heatmaps of different features, the distribution of various variables can be analysed to determine whether they exhibit similar patterns. If two features have similar patterns, this suggests they are correlated. For instance, by observing the heatmaps of ‘oldpeak’ (ST depression) and ‘slope’ (slope of peak exercise ST segment), both features display similar patterns, suggesting a potential significant correlation between them.

On the other hand, the heatmaps for ‘thalach’ (maximum heart rate achieved) and ‘exang’ (exercise-induced angina) shows an inverse relationship between them, as it is noticed that regions with higher ‘thalach’ values (light yellow areas) correspond to regions with lower ‘exang’ values (red areas). The inverse relationship between "thalach" and "exang" as observed suggests that individuals with higher maximum heart rates are less likely to have angina during exercise. This finding aligns with medical knowledge, as a higher heart rate capacity generally indicates better cardiovascular health and a reduced risk of angina (Wong, et al., 2015).

When looking at the ‘sex’ variable alone, the heatmap reveals a distinct pattern that differs from those observed in other variables. This observation suggests that 'sex' may not exhibit a strong or direct relationship with the rest of the variables within the context of this SOM model. The lack of pattern similarity could also imply that 'sex' possibly interacts with these variables in a more complex manner that may not be immediately apparent from the heatmap alone.

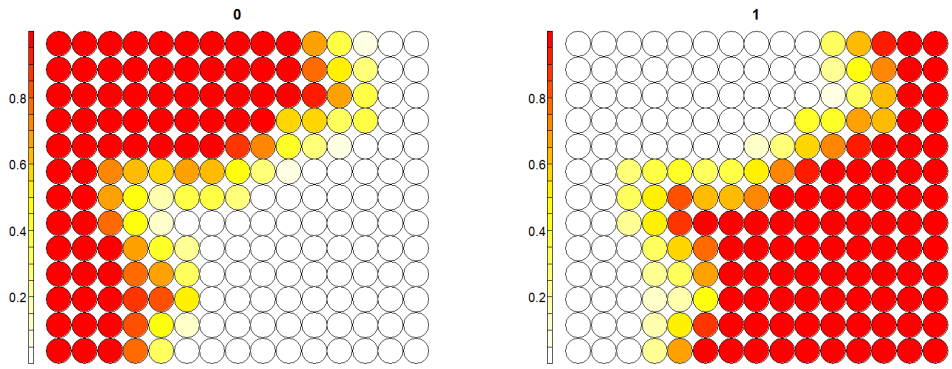

Figure 4.16 Heatmaps of SOM for the Target Variable.

The heatmaps in Figure 4.17 above illustrating the distribution of neurons in the SOM for the target variable, with separate plots for cases of no heart disease (0) and heart disease (1). This visualisation helps to identify regions of the SOM grid associated with higher or lower frequencies of the target variable. Red indicates the strong presence of each case of the target condition. Specifically, in the heatmap representing no heart disease (0), red indicates the frequency of cases with no heart disease and in the heatmap representing heart disease (1), red indicates the frequency of cases with heart disease. In short, the formation of distinct patterns indicates that SOM has identified meaningful patterns and relationships within the data.

In addition, by examining the heatmaps of features and the target variable concurrently, it allows for the inference of the importance of certain features indirectly. For instance, examining the heatmap of the feature ‘exang’ (exercise-induced angina) in Figure 4.15 alongside the heatmap of no heart disease (0) in Figure 4.16, reveals significant overlap in the red regions at the top left of both heatmaps., This overlap suggests that the region associated with the absence of heart disease is significantly composed of patients without exercise-induced angina, thereby indicating that 'exang' may be a critical predictor of heart disease.

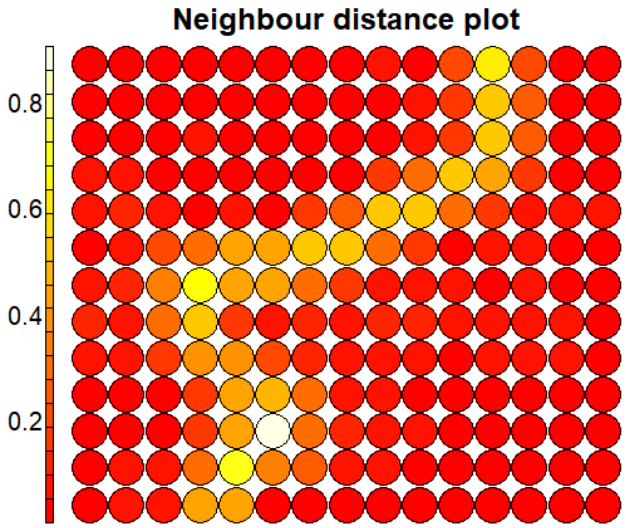

Figure 4.17: U-Matrix of SOM.

The Neighbour Distance Plot of the developed SOM as shown in Figure 4.15 above, also referred to as U-Matrix, reveals significant insights into the clustering performance of SOM. This plot visualises the distances between neighbouring neurons, thereby identifying cluster boundaries and the relative distances between clusters. The lighter areas, indicating larger distances between neighbouring nodes, are regions where the data points are significantly different from each other. These light regions serve as boundaries between distinct clusters, suggesting effective separation of two groups within the data. Conversely, the red circles, representing smaller distances, highlight regions of high homogeneity, implying that the data points within each cluster share substantial similarities. The upper left red areas indicate clusters of patients without heart disease, while the bottom right red areas indicate clusters of patients with heart disease. The presence of clear boundaries and two clusters highlights the SOM model's efficacy in organising the data.

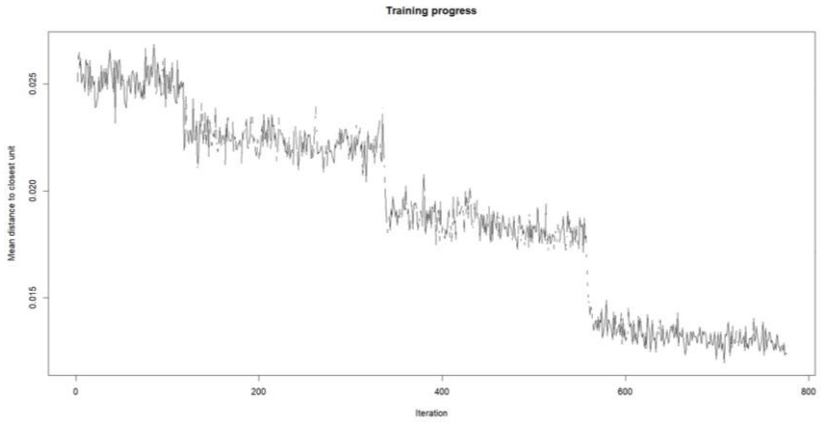

Figure 4.18: Training Progress of SOM.

This plot illustrates the training progress of SOM across 776 iterations. The y-axis shows the mean distance to the closest unit, while the x-axis represents the number of iterations. The plot shows a general decreasing trend in mean distance over time, indicating that the nodes' weights are becoming closer to the samples. As training progresses, the SOM's neurons are effectively adjusting and becoming better at representing the input data. A lower mean distance to the closest unit suggests that the SOM is improving in terms of how well it fits the data. However, there are fluctuations in the line, which is normal during SOM training as the map adjusts to the data. The overall downward trend is what indicates convergence. Toward the end of the iterations, the line appears to level off, suggesting that the training is stabilising, and the SOM has largely converged.

4.6 Summary

The implementation of PSO has led to substantial enhancements across all metrics, indicating that the optimisation technique has significantly boosted the model's performance. The SOM model is now more accurate and precise, with improved capability in correctly identifying both positive and negative cases of heart disease. Additionally, the visualisations of SOM including feature heatmaps and U-matrix as well offer valuable insights into the relationships among cardiovascular conditions—insights that are uniquely afforded by unsupervised learning models. In fact, it becomes evident that SOM, as an

unsupervised ML model, is capable of making predictions with high accuracy and effectiveness, potentially even outperforming supervised ML models.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Heart disease remains a chronic and persistent health issue, necessitating reliable and robust methods for early prediction and intervention. Considering the limitations of conventional models, this study highlights the efficacy of the SOM as an innovative unsupervised ML approach for predicting heart disease. Following the selection of a robust feature set, based on the correlation analysis and using the unsupervised t-SNE algorithm,, the SOM model was developed. The proposed PSO-SOM model then further contributes to resolving the challenges of predictive tasks within the medical field.

In cases where a predictive model like SOM is sensitive to hyperparameters, PSO is particularly advantageous as it can efficiently search through the possible hyperparameter space to identify the optimal configuration. Moreover, PSO's capability to handle non-differentiable spaces makes it well-suited for optimising SOM, an unsupervised ML model that does not provide a smooth gradient for optimisation. In fact, the effectiveness of the proposed PSO-SOM algorithm is clearly demonstrated by its ability to enhance the accuracy of the SOM model to 94.4%. This improvement significantly bolsters the model's performance in predicting heart disease, underscoring the potential of the PSO-SOM approach in this domain.

Furthermore, this study also highlights that unsupervised SOMs offer a compelling advantage over supervised methods due to their ability to preserve topology, reduce dimensionality, and utilise various visualisation techniques. For instance, visualisations such as heatmaps and U-matrix of SOM provide a rich visual representation of the underlying data structure, revealing hidden patterns and relationships that may be obscured by supervised approaches. This enhanced interpretability is crucial for understanding complex cardiovascular conditions and optimising the development of effective predictive models.

To conclude, the unsupervised SOM model has demonstrated highly effective prediction results, achieving strong performance across all assessment

metrics. It is crucial for predictive models, particularly in the medical field, to have a high recall rate alongside accuracy due to the significant costs associated with medical diagnoses. The optimal performance metrics of the SOM model indicate not only high accuracy but also effective identification of heart disease, minimising instances of missed actual heart disease cases. Hence, integrating the SOM model into clinical workflows could significantly enhance early detection and personalised treatment plans for heart disease, enabling healthcare professionals to identify at-risk individuals through data patterns and reducing disease prevalence.

In addition, this study closely aligns with and can contribute to the Sustainable Development Goals (SDGs), specifically SDG 3, as illustrated in Figure 5.1 below. SDG 3 emphasises on promoting healthy lives and well-being for people of all ages (United Nations, 2024).

Figure 5.1: SDG 3: Good Health and Well-Being.

Heart disease is one of the major non-communicable diseases (NCDs). By improving predictive models for heart disease, this study contributes to early detection and better management of cardiovascular conditions, which aligns with the goal of reducing premature mortality from NCDs including cardiovascular disease, diabetes, cancer, and chronic respiratory diseases, as outlined in Target 3.4.1 of SDG 3 (United Nations, 2024).

Besides, Target 3.d of SDG 3 is also relevant to this study, as it emphasises strengthening the capacity of countries, particularly developing ones, for early warning, risk reduction, and management of health risks (United Nations, 2024). The SOM model for heart disease prediction plays a crucial role in this objective by enhancing early warning systems and improving risk management strategies. By offering predictive insights into heart disease, this

study supports global health systems, including those in developing regions, thereby bolstering their ability to manage and respond to health risks more effectively.

5.2 Recommendations for Future Work

Upon achieving promising results with the developed SOM model, several limitations and areas for improvement can be acknowledged and addressed for future research.

In this study, the generalisability of the model may be constrained by the relatively small sample size of the heart disease data used. To strengthen the model's applicability across diverse populations, future research should validate the SOM model using larger, more representative datasets from reliable sources. These datasets should encompass a wider range of demographic, clinical, and sociocultural factors to ensure comprehensive evaluation and broader applicability.

While PSO was a useful tool for hyperparameter tuning in this study, investigating other optimisation algorithms such as GA may offer alternative perspectives on the model's performance. By comparing these methods, researchers can identify the most effective technique for fine-tuning the SOM model, enhancing the optimisation process. This would ultimately improve the model's overall effectiveness and ensure its robustness in real-world applications, across various demographic and clinical contexts.

In short, addressing these factors will contribute to the development of a more reliable and applicable SOM model, capable of effectively capturing the complexity of medical data and predicting heart disease. This will ultimately facilitate better clinical decision-making and lead to enhanced patient outcomes.

REFERENCES

Akoglu, H., 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3).

American Heart Association, 2021. *Target Heart Rates Chart*. [online] Available through: <https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates> [Accessed 8 April 2024].

Amin, S. U., Agarwal, K. and Beg, R., 2013. Genetic neural network based data mining in prediction of heart disease using risk factors. *2013 IEEE Conference on Information and Communication Technologies, ICT 2013*.

Arnaout, R., Curran, L., Zhao, Y., Levine, J. C., Chinn, E. and Moon-Grady, A. J., 2021. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nature Medicine*, 27(5).

Astudillo, C. A. and Oommen, B. J., 2014. Topology-oriented self-organizing maps: A survey. Pattern Analysis and Applications, 17(2).

Barfungpa, S. P., Samantaray, L., Sarma, H. K. D., Panda, R. and Abraham, A., 2023. D-t-SNE: Predicting heart disease based on hyper parameter tuned MLP. *Biomedical Signal Processing and Control*, 86.

Bendtsen, C., 2022. *pso: Particle Swarm Optimization.* [online] CRAN. Available through: <https://cran.r-project.org/web/packages/pso/pso.pdf> [Accessed 22 June 2024].

CDC, 2021. *About High Blood Pressure*. [online] Available through: <https://www.cdc.gov/high-blood-pressure/about/index.html> [Accessed 8 April 2024].

Chaudhary, V., Bhatia, R. S. and Ahlawat, A. K., 2014. An efficient self-organizing map (E-SOM) learning algorithm using group of neurons. International Journal of Computational Intelligence Systems, 7(5).

Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J. Z., Ni, D. and Heng, P-A., 2017. Ultrasound Standard Plane Detection Using a Composite Neural Network Framework. *IEEE Transactions on Cybernetics*, 47(6).

Harvard Medical School, 2023. *What is a normal heart rate?*. [online] Available through: <https://www.health.harvard.edu/heart-health/what-your-heart-rate-is-telling-you> [Accessed 8 April 2024].

Jiang, Z., Bo, L., Wang, L., Xie, Y., Cao, J., Yao, Y., Lu, W., Deng, X., Yang, T. and Bian, J., 2023. Interpretable machine-learning model for real-time, clustered risk factor analysis of sepsis and septic death in critical care. *Computer Methods and Programs in Biomedicine*, 241.

Johns Hopkins Medicine, n.d. *Lipid Panel*. [online] Available through: <https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel> [Accessed 8 April 2024].

Kapila, R. and Saleti, S., 2023. An efficient ensemble-based Machine Learning for breast cancer detection. *Biomedical Signal Processing and Control*, 86.

Kaur, I. and Ahmad, T., 2024. A cluster-based ensemble approach for congenital heart disease prediction. Computer Methods and Programs in Biomedicine, 243.

Kennedy, J. and Eberhart, R., 1942. Particle Swarm Optimization. Purdue School of Engineering and Technology.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1).

Kohonen, T., 2001. Self-Organizing Maps. Springer-Verlag, New York, Berlin, Heidelberg. Available through: <https://doi.org/10.1007/978-3-642-56927-2> [Accessed 2 September 2024].

Lin, K. C. and Hsieh, Y. H., 2015. Classification of Medical Datasets Using SVMs with Hybrid Evolutionary Algorithms Based on Endocrine-Based Particle Swarm Optimization and Artificial Bee Colony Algorithms. *Journal of Medical Systems*, 39(10).

Loh, W. S., Chin, R. J., Ling, L., Lai, S. H. and Soo, E. Z. X., 2021. Application of Machine Learning Model for the Prediction of Settling Velocity of Fine Sediments. Mathematics, 9(23), 3141.

Loh, W. S., Tan, W. L., Chin, R. J., Ling, L., Phoon, S. W. and Seah, C. S., 2024. An Unsupervised Machine Learning Approach for Estimating Missing Daily Rainfall Data in Peninsular Malaysia. The 19th IMT-GT International Conference on Mathematics, Statistics and their Applications. ITM Web of Conferences.

Maaten, V. D. L. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9.

Malakar, S., Sen, S., Romanov, S., Kaplun, D. and Sarkar, R., 2023. Role of transfer functions in PSO to select diagnostic attributes for chronic disease prediction: An experimental study. *Journal of King Saud University - Computer and Information Sciences*, 35(9).

Mohapatra, S., Maneesha, S., Patra, P. K. and Mohanty, S., 2022. Heart Diseases Prediction based on Stacking Classifiers Model. *Procedia Computer Science*.

National Centre for Biotechnology Information, 2002. *Exercise tolerance testing*. [online] Available through: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1123032/> [Accessed 8 April 2024].

National Institute on Aging, 2018. *Heart Health and Aging*. [online] Available through: <https://www.nia.nih.gov/health/heart-health/heart-health-and-aging> [Accessed 8 April 2024].

Nguyen, H. N., Hoang, T. H., Cao, V. L. and Tran, N. N., 2023. Tuning Hyperparameters of Self-Organising Maps In Combination with K-Nearest Neighbours For IoT Malwarre Detection. Journal of Science and Technique.

Nssibi, M., Manita, G. and Korbaa, O., 2023. Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey. *Computer Science Review*, 49.

Osman, A. H. and Alzahrani, A. A., 2019. New Approach for Automated Epileptic Disease Diagnosis Using an Integrated Self-Organization Map and Radial Basis Function Neural Network Algorithm. *IEEE Access*, 7.

Raja, J. B. and Pandian, S. C., 2020. PSO-FCM based data mining model to predict diabetic disease. *Computer Methods and Programs in Biomedicine*, 196.

Rankovic, N., Rankovic, D., Lukic, I., Savic, N. and Jovanovic, V., 2023. Unveiling the Comorbidities of Chronic Diseases in Serbia Using ML Algorithms and Kohonen Self-Organizing Maps for Personalized Healthcare Frameworks. *Journal of Personalized Medicine*, 13(7).

Rath, A., Mishra, D., Panda, G., Satapathy, S. C. and Xia, K., 2022. Improved heart disease detection from ECG signal using deep learning based ensemble model. *Sustainable Computing: Informatics and Systems*, 35.

Rawshani, A., 2021. *The ST segment: physiology, normal appearance, ST depression & ST elevation*. [online] Available through: <https://ecgwaves.com/st-segment-normal-abnormal-depression-elevation-causes/> [Accessed 8 April 2024].

Sheeba, P. T., Roy, D. and Syed, M. H., 2022. A metaheuristic-enabled training system for ensemble classification technique for heart disease prediction. *Advances in Engineering Software*, 174.

Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 17(8).

Subanya, B. and Rajalaxmi, R. R., 2014. Feature selection using artificial bee colony for cardiovascular disease classification. *2014 International Conference on Electronics and Communication Systems, ICECS 2014*.

UCI Machine Learning Repository, 1988. *Heart Disease*. [dataset] Available through: <https://archive.ics.uci.edu/dataset/45/heart+disease> [Accessed 20 March 2024].

United Nations, 2024. Goal 3 *Ensure healthy lives and promote well-being for all at all ages*. [online] Available through: <https://sdgs.un.org/goals/goal3#targets_and_indicators> [Accessed 18 August 2024].

Verma, L., Srivastava, S. and Negi, P. C., 2016. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of Medical Systems*, 40(7).

Wehrens, R. and Buydens, L. M. C., 2007. Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software, 21(5).

WHO, 2021. *Cardiovascular diseases (CVDs)*. [online] Available through: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)> [Accessed 22 February 2024].

Wong, Y. K., Stearn, S., Moore, S. and Hale, B., 2015. Angina at low heart rate and risk of imminent myocardial infarction (the ALARM study): A prospective, observational proof-of-concept study. BMC Cardiovascular Disorders, 15(1).

Yang, H. C. and Lee, C. H. 2012. A Novel Self-Organizing Map for Text Document Organization. Third International Conference on Innovations in Bio-Inspired Computing and Applications, Kaohsiung, Taiwan. 39-44.

Zarkogianni, K., Athanasiou, M. and Thanopoulou, A. C., 2018. Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication. *IEEE Journal of Biomedical and Health Informatics*, 22(5).