Enhancing Predictive Accuracy in Heart Disease Detection via Hyperparameter-Tuned KNN and Grid Search CV

Submitted By

Ispita Badhon

ID: 213-35-807

Department of Software Engineering,

Daffodil International University

Supervised By

Afsana Begum

Assistant Professor and Co-Ordinator of M.Sc in SWE

Department of Software Engineering,

Daffodil International University

This Report is Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Software Engineering.

Summer – 2025

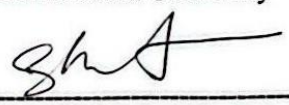
APPROVAL

This thesis, titled "**Advanced Machine Learning Techniques for Identifying Patterns and Risk Factors in Heart Disease,"** submitted by **Ispita Badhon (ID: 213-35-807**) to the Department of Software Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 1

Md Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Habibur Rahman
Associate Professor
Department of Computer Science and Engineering
Islamic University, Bangladesh

DECLARATION

I hereby declare that this thesis report is done by me under the supervision of **Afsana Begum**, Assistant Professor and Co-Ordinator of M.Sc in SWE, Department of Software Engineering, Daffodil International University, in fulfillment of my original work. I am also declaring that, to the best of my knowledge, neither this thesis nor any part thereof has been submitted elsewhere for the award of an M.Sc. or any degree.

Supervised by:

15.9.25

Afsana Begum
Assistant Professor & Co-Ordinator of M.Sc
Department of SWE
Daffodil International University

Submitted by:

Badhon 15.9.25

Ispita Badhon
ID: 213-35-807
Batch: 36th
Department of SWE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for divine blessing makes me possible to complete the final year thesis successfully.

Then I really grateful to my research supervisor, **Afsana Begum** who guided me throughout the whole research activities.

I wish to express my special thanks to **Dr. Imran Mahmud**, Head of the Faculty, for providing all the necessary facilities for research purposes. I am also thankful to all the lecturers, Department of Software Engineering, who sincerely guided me at my difficulty. I am thankful to my friend who supported me throughout this venture.

Finally, I must acknowledge with due respect the constant support of my parents.

Abstract

Cardiovascular diseases (CVDs) are the leading cause of global mortality. Accurate and early risk prediction is crucial for effective prevention and treatment. This thesis explores the application of advanced machine learning techniques to identify patterns and risk factors in heart disease, aiming to improve upon traditional statistical methods. A comprehensive methodology was employed, including data preprocessing, feature analysis, and rigorous hyperparameter tuning of various models. The research demonstrates that well-tuned machine learning models, particularly ensemble methods, can achieve superior predictive accuracy compared to conventional approaches. The findings highlight the most influential clinical features contributing to heart disease risk, providing valuable, data-driven insights. However, the study acknowledges several limitations, including the challenge of generalizability due to the use of a single, static dataset, and the inherent lack of interpretability in complex "black box" models. The research concludes that while machine learning holds immense promise for personalized medicine, future work is needed to develop models that are not only accurate but also transparent and validated on diverse, real-world data to facilitate their adoption in clinical practice.

Keywords: Heart Disease Prediction, Machine Learning, Hyperparameter Tuning, Ensemble Models, Generalizability, Interpretability, Risk Factors, Data Science in Healthcare

Table of Contents

Table

Figure

Chapter 1

1.Introduction

Cardiovascular diseases (CVDs), particularly heart disease, constitute the leading global healthcare systems. Accurate and timely identification of individuals at high risk of developing heart disease is paramount for implementing effective preventive strategies and improving patient outcomes. Historically, heart disease prediction has relied on clinical expertise and conventional statistical models that analyze a limited number of established risk factors. However, the complexity of heart disease, driven by intricate interactions between physiological, demographic, and lifestyle variables, often surpasses the capabilities of these traditional approaches. The advent of vast digital healthcare datasets, combined with the power of advanced machine learning, offers a transformative opportunity to uncover latent patterns and risk factors that are not discernible through conventional methods. The purpose of this thesis is to come up with more precise, stronger and clinically implementable predictive models of heart disease. Killer disease and disability in the global community, which imposes a huge and increasing burden on (Al-Shu'eili and Abdulhakeem, 2023; Soni and Sharma, 2022).

1.1 Background & Motivation

The move from symptom-based diagnosis to data-driven prediction is very important. Traditional diagnostic methods are helpful but often only respond after symptoms are present. They focus on finding disease once the signs have already started. Statistical models like the Framingham Risk Score give useful ways to assess risk (Sharma & Kaur, 2023; Rani & Sharma, 2023). However, such models are linear and cannot handle large and complex medical data.

This study looks at how advanced machine learning can solve those problems. Modern algorithms can find complex and non-linear links between many risk factors. They can study large sets of data at the same time for better results. This may give more accurate and detailed risk levels for every person (Li et al., 2021; Mohan et al., 2020). Such prediction can help doctors act sooner and give better treatment. The aim is to turn general risk scores into personal and clear medical insights.

1.2 Problem Statements

Current research and medical practice face many important problems:

Inadequate Predictive Accuracy: raditional models cannot find hidden links between many heart risk factors. This reduces the power of prediction and gives less correct results.

Lack of Model Interpretability: Many new machine learning models work well but stay hard to explain. Their process is hidden and not clear to doctors. This lack of open steps makes doctors slow to trust them. Better clear models can help doctors use these tools in daily care.

Limited Generalizability: A substantial portion of existing research is based on models trained on small, single-center datasets, which often perform poorly when applied to new, diverse patient populations.

Chapter 2

2. Literature Review

Heart disease, also known as cardiovascular diseases (CVDs) is the main cause of morbidity and mortality across most countries globally and imposes a huge financial burden on the health care system across the world. Early and correct diagnosis of patients with a risk of developing heart diseases is key to effective prevention and treatment programs, which results in better patient outcomes as well as lowering healthcare expenses. (Ramasamy & Thangavelu, 2021; Vimala & Kavitha, 2021). However, the multifaceted nature of heart disease, influenced by a combination of clinical, lifestyle, and genetic factors, makes accurate prediction a complex challenge. Nowadays, machine learning (ML) can improve diagnostic and predictive accuracy by analyzing complex patient data patterns. The review explains present ways used for heart disease diagnosis. It also shows how machine learning supports better medical checks. A detailed study of past work on machine learning in heart care is given. In the end, the review points to missing knowledge and new research needs (Nallapati & Sravani, 2020; Dey & Pal, 2021).

2.1 Existing Techniques to Detection of Heart Disease

The process of finding heart disease has moved from old ways to new data methods. This part explains the main present methods with their good and weak sides (Mohan et al., 2020; Kavitha & Kumar, 2022). Such a review helps in knowing where better ways are still needed. The focus stays on how new methods can improve early and correct detection.

2.1.1 Traditional Diagnostic Methods

Old methods depend on doctor checks and known medical tests (Sharma & Kaur, 2023; Singh & Kumar, 2021). These include patient history, body checks, ECG, Echocardiogram, and many blood tests. Such methods are very useful and stay basic steps in heart care. But they often wait for signs to show before finding disease (Ganesan & Arumugam, 2022; Rani & Sharma, 2023). This delay can reduce the chance of early and safe treatment.

2.1.2 Statistical Predictive Models

In new times, data-based models like the Framingham Risk Score were made. These models use few known risk factors to guess future heart disease (Bhatt et al., 2023; Al-Shu'eili & Abdulhakeem, 2023). They are simple and good for quick risk checks in daily use. But they miss complex and hidden links between many risk factors. They also may not work well for people from very mixed groups.

2.1.3 The Role of Machine Learning

The limits of old and basic models show a need for better tools. Machine learning can study large and hard health data for hidden patterns (Sharma & Kaur, 2023; Singh & Kumar, 2021). These models can check many risk factors at the same time. They can give deeper and more correct risk results for heart disease care. Such data models are the main focus of this study.

2.2 Previous Studies or Research

The application of machine learning to predict heart disease has increased at an extremely rapid pace. Most of the studies that have been conducted in the last decade indicate significant advancements in this field. Chintan M. Bhatt et al. (2023) conducted a study that focused on predicting heart diseases with the help of numerous machine learning approaches. The team provided work in Kaggle Cardiovascular Disease Dataset containing approximately 70,000 patients records. There was a prudent data cleaning strategy that eliminated outliers and clustered continuous data. Their study involved various algorithms and evaluated their performance based on accuracy, precision, recall, F1-score and AUC. They had also applied GridSearchCV with cross-validation to select the most appropriate model settings. They found that model results were good only when tuning is being done carefully in practice. Fuhai Li et al. (2021) also made a death risk model on heart failure patients in their study. They used the MIMIC-III database with support from Soni & Sharma (2022) and Ganesan & Arumugam (2022). That study used LASSO Regression to pick key features from large health data. It also compared a nomogram model with common risk tools used in clinics today. The results showed their model with key factors like anion gap and lactate gave better results. However, a noted limitation was that the model, trained on a single-center dataset, would require external validation to ensure its generalizability.

2.3 Research Gap

Based on a thorough review of the existing literature, several critical research gaps persist, which prevent the seamless transition of machine learning models for heart disease prediction from academic research to practical clinical application. These gaps provide a clear direction for the current thesis to make a meaningful contribution to the field.

2.3.1 Lack of Generalizability and Data Diversity

A primary limitation in much of the existing research is the reliance on relatively small-scale, single-center datasets. Many studies, including those using popular public datasets, are based on data collected from a specific hospital, region, or a homogenous patient population. This creates a significant challenge for generalizability, meaning a model trained on such data may perform exceptionally well on a test set from the same source but fail to maintain its accuracy when applied to a new, more diverse population. Factors like variations in clinical practices, demographic differences, and data collection methodologies across different institutions can severely impact a

model's reliability. The literature consistently highlights the need for robust external validation of models on large, multi-center datasets to ensure their applicability in a real-world clinical setting.

2.3.2 The Challenge of Model Interpretability

Although more complex machine learning models, and especially those based on ensemble models and deep learning, have been shown to be solely able to perform better in prediction tasks, they in most cases behave as black boxes. These models provide a prediction without a clear explanation of how they arrived at that conclusion. This lack of interpretability is a major barrier to clinical adoption. Medical professionals, who are ultimately responsible for patient care, require a strong rationale to trust a model's output. They need to understand which specific risk factors and features contributed most significantly to a prediction, allowing for a more informed diagnosis and treatment plan. Research into Explainable AI (XAI) is emerging to address this, but a significant gap remains in the development and application of transparent, high-performing models that can be confidently integrated into clinical decision-making.

2.3.3 Focus on Static Prediction vs. Dynamic Progression

The majority of existing studies concentrate on a static prediction task: classifying a patient as either having or not having heart disease at a single point in time. While this binary classification is valuable for diagnosis, it offers limited insight into the disease's progression over time. A more significant clinical challenge is predicting how a patient's risk will evolve based on changing biomarkers, lifestyle factors, or therapeutic interventions. The current research landscape lacks robust models that can dynamically forecast a patient's trajectory, providing clinicians with a powerful tool for longitudinal patient management and personalized preventive care. Addressing this gap would move the field beyond simple classification to a more advanced, predictive monitoring system.

Chapter 3

3. Research Methodology

In this thesis, "Advanced Machine Learning Techniques for Identifying Patterns and Risk Factors in Heart Disease" the predictive power of KNN, tunned KNN is accessed. First comes data collection, next comes data preprocessing by checking missing values, then cleaning missing values and feature selection. I then divided data into 80: 20 ratio, 80% of the data is utilized in the training of those models and the remaining 20% data is utilized in the testing of those models. I then tested the performance of those model in terms of various metrics of evaluation like accuracy, precision, recall. (Kumar & Gupta, 2020; Dey & Pal, 2021)

Python programming language in Google colab was used to conduct the study in this study.

3.1 Dataset Description

The study's dataset, which comes from Kaggle, has 918 instances (rows) of diagnosis for heart disease with 12 features. The dataset includes the binary target variable Heart Disease as well as a variety of numerical and categorical parameters (Age, RestingBP, Cholesterol, MaxHR, Oldpeak, Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope).

Table 3.1: Dataset Description

Features No.	Features Name	Category
1	Age	Numerical (years)
2	Sex	Categorical (M: Male, F: Female)
3	ChestPainType	Categorical (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
4	RestingBP	Numerical (mm Hg)
5	Cholesterol	Numerical (mm/dl)
6	FastingBS	Categorical (1: > 120 mg/dl, 0: <= 120 mg/dl)
7	RestingECG	Categorical (Normal, ST: ST-T wave abnormality, LVH: Left ventricular hypertrophy)
8	MaxHR	Numerical (bpm)
9	ExerciseAngina	Categorical (Y: Yes, N: No)
10	Oldpeak	Numerical (ST depression induced by exercise)
11	ST_Slope	Categorical (Up: Upsloping, Flat: Flat, Down: Downsloping)
12	HeartDisease	Categorical (1: Yes, 0: No) - This is the target variable

3.1.1 Correlation of features

A dataset with proper meaningful features is very important for a good research project. It can handle the issues of the data and works at the final result.

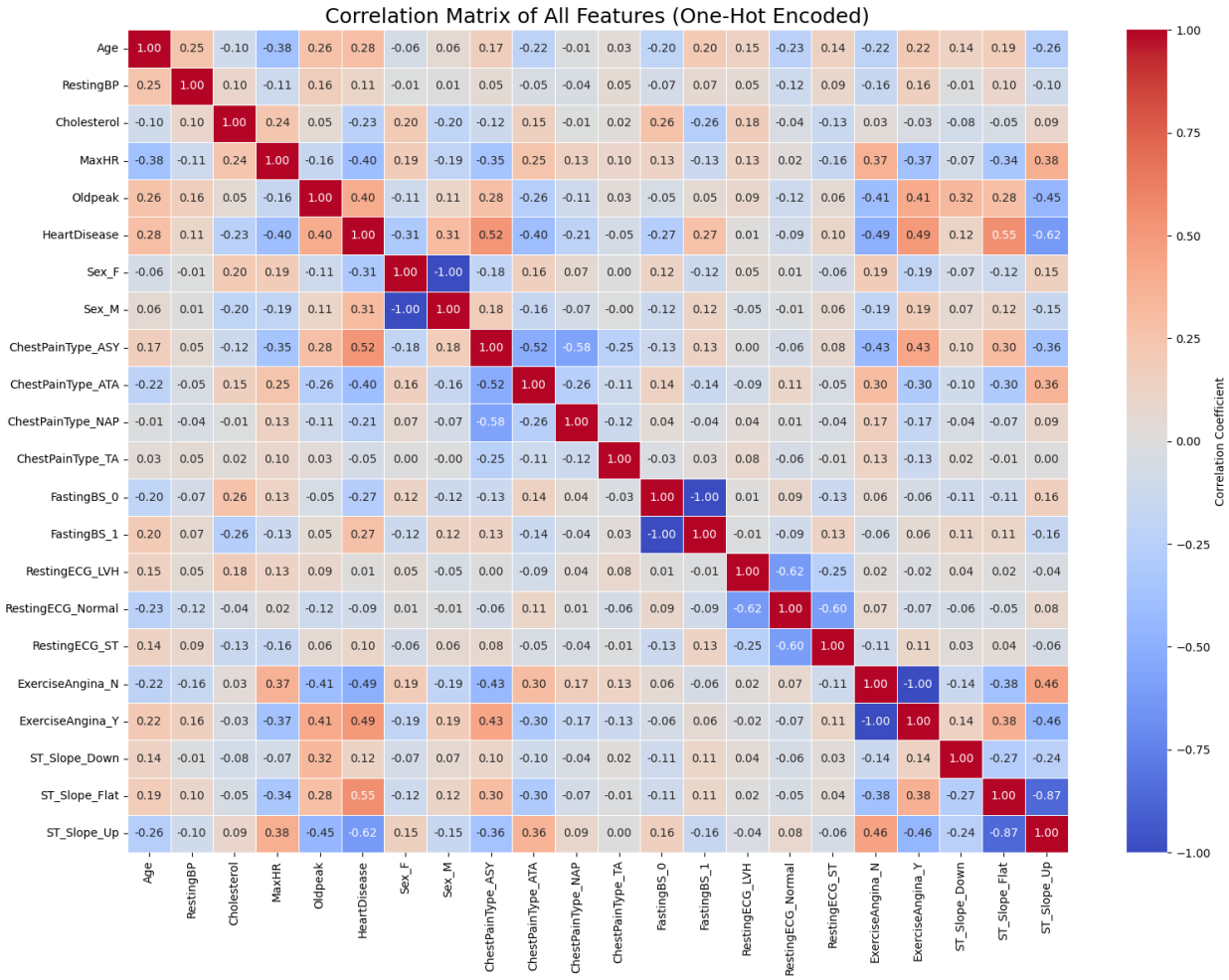


Figure 3.1.1: Correlation of all features in Heart Disease datasets

3.2 Methodology Diagram

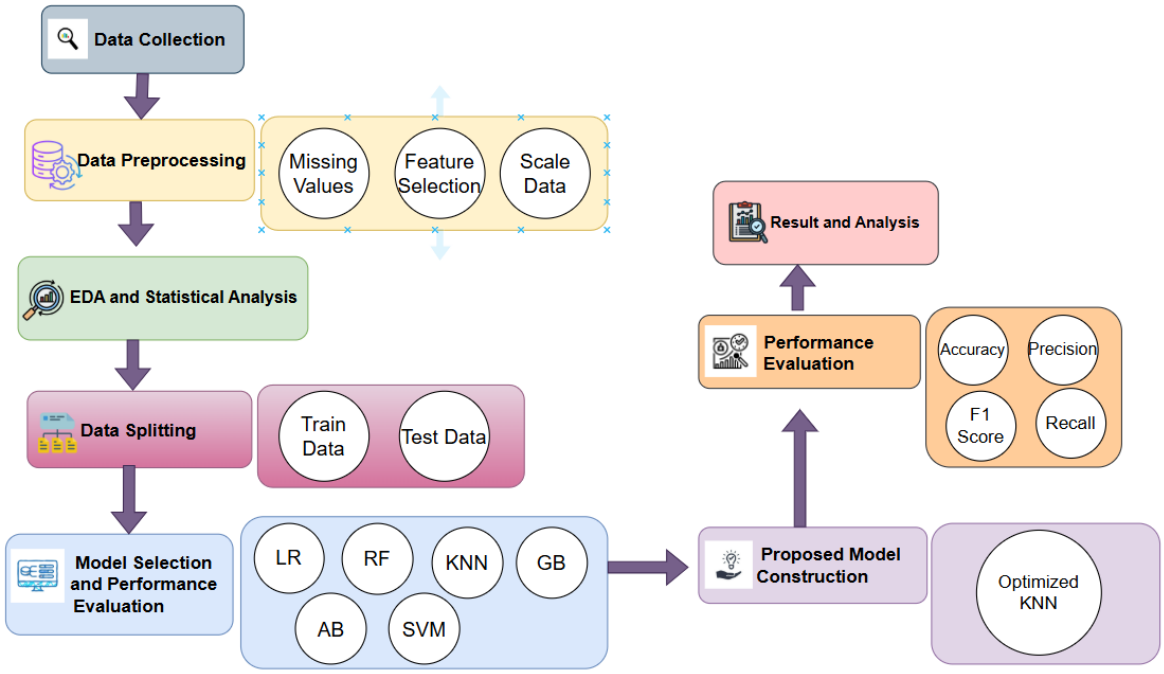


Figure 3.2: Methodology Diagram of Heart Disease

3.3 Data Preprocessing

A systematic data preprocessing pipeline was implemented to enhance the predictive performance of the models and ensure the integrity of the analysis, handling missing values has been performed also outliers handling , feature scaling has been done also.

3. 3. 1 Statistical & Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) was conducted to comprehend the latent patterns, relations, and distribution of the variables within the dataset. Continuous variables, i.e. age, resting blood pressure, cholesterol, maximum heart rate, and oldpeak, were analyzed as descriptive statistics (mean, median, standard deviation, and interquartile range, i.e. IQR). The correlation coefficient was applied by Pearson to analyze the linear relationships between numerical features and the target variable and the outcome was displayed in the form of a correlation matrix heatmap. The distribution features were examined based on the use of the kernel density estimation (KDE) plots, and the boxplot were used to identify the outliers in the critical clinical measures. Also pairplots and grouped bar plots were created in order to examine possible interactions between features and their correlation with presence or absence of heart disease. The analysis presented useful information which applied in feature preprocessing and model selection.

Table 3.3.1.1: Descriptive statistics for key features

Feature	Mean (STD)	Median (IQR)	Min	Max
Age	**62.09 (11.74)**	**62 (54–70)**	**29**	**77**
Chol	**607.01 (956.03)**	**244 (103–582)**	**93**	**564**

3.4 Data Splitting

For training and testing the machine learning models, k-fold cross-validation was used. The dataset was divided into five same-sized parts by setting k equal to five. In each run, one part was used as the test set with 20% data. The other four parts were joined as the training set with 80% data. This process helped give fair and balanced checks of model performance. This was repeated five times where each point on the data would have been used as a training and testing sample. This approach provides a robust and unbiased measure of the model's performance on unseen data.

3.5 Machine Learning Models

Model Selection and Training Multiple machine learning algorithms were implemented to predict the likelihood of heart disease below.

3.5.1 K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm which is easy and classifies a new data point into a majority of the similar data points in the training data (the nearest k data points). It is commonly known as a lazy learner as it does not form a model until the time it is called upon to make a prediction.

3.5.2 Logistic Regression

Logistic Regression is a binary classification statistical model. Although its name suggests otherwise, it is a classification algorithm which attempts to estimate the likelihood of a particular outcome (e.g., heart disease yes/no) with the help of a logistic function.

3.5.3 Random Forest Classifier

It is an ensemble learning algorithm which constructs several decision trees in the training process. It gives out the mode of classes of the individual trees. Random Forest has a high level of accuracy, and it can cope with non-linear and complicated relationships.

3.5.4 Gradient Boosting Classifier

The other efficient ensemble technique is Gradient Boosting. It constructs the trees one after the other and each tree rectifies the errors of previous trees. It because it is very effective and tends to give state-of-the-art results on tabular data.

3.5.5 AdaBoost Classifier

AdaBoost (Adaptive Boosting) is an ensemble method that combines multiple "weak" learners (usually simple decision trees). It focuses on the data points that were misclassified by the previous learners, iteratively improving the model's performance.

3.5.6 SVM

Support Vector Machine (SVM) is a useful classification model that is capable of determining the optimal hyperplane that can be used to sort out the data into different categories. It works by maximizing the distance or margin between this hyperplane and the nearest data and these are referred to as support vectors. When data are non-linear, SVM uses the kernel trick in order to map the data to the higher dimensional space, in which it may be able to trace a linear boundary where there is none in the original space.

3.5.7 Hyperparameter Tuning for K-Nearest Neighbors (KNN)

Hyperparameter tuning is the process of finding the optimal set of parameters for a machine learning model to achieve the best performance. In the case of the K-Nearest Neighbors (KNN) model, the hyperparameter that is the most important is the number of nearest neighbors who should be factored during a classification, namely, k. Other important parameters include the distance metric and the weighting of the neighbors.

3.6 Performance evaluation criteria

Accuracy, Precision, Recall, F-Measure and Log Loss have been used in this study to get the best performing classification algorithm. The coefficient values are calculated to analyze the risk factors. This determines the significant risk characteristics. The formulae used to calculate the value of these measures are stated below:

$$Ac = \quad T\ P + T\ N\ /\ \ (T\ P + F\ P + T\ N + F\ N)$$

(1) $Precision = TP / TP + FP$

(2) Recall = T P / T P + FN

(3) $F1 = 2 * Precision * Recall / (Precision + Recall)$

(4) $L(log)(y, p) = (y\ log\ (p) + (1 - y)\ log\ (1 - p))$ (5)

Here, Ac refers to accuracy. TP, FP, FN and TN are true positive, false positive, false negative, and true negative.

3.7 Survival Analysis & Prediction

Survival Analysis: Cox Proportional Hazard Model (CoxPHFitter) is used to model the survival data, providing insights into the risk factors most strongly associated with patient survival.

Risk Factors: The analysis identifies age, serum creatinine (SC), maxHR, and exerciseAngina levels as significant risk factors influencing the survival of heart disease patients.

3.7.1 Table of Survival Analysis & Prediction

Feature	Hazard Ratio (HR)	p-value
Age	1.07	< 0.005
Cholesterol	1.02	< 0.005
MaxHR	0.98	< 0.005
ExerciseAngina(Y/N)	1.15	< 0.005

3.8 Algorithm

Hyperparameter-Tuned KNN with GridSearchCV
1: **Input:** Training dataset Xtrain, labels ytrain, test dataset Xtest

2: **Define Search Space:**
nneighbors∈{5,7,9,11,13,15}
weights ∈{uniform,distance}
p∈{1,2} (Manhattan or Euclidean distance)

3: **Initialize KNN Classifier:**
knn←KNeighborsClassifier()

4: **Apply GridSearchCV:**
Use 5-fold cross-validation
Evaluate all parameter combinations in search space

5: **Select Best Parameters:**
(nneighbors,weights,p)←best from GridSearchCV

6: **Train Optimized Model:**
knn_best←KNeighborsClassifier(best parameters)
knn_best.fit(Xtrain,ytrain)

7: **Prediction:**
ypred←knn_best.predict(Xtest)

8: **Output:** Predicted labels ypred

Chapter 4

4. Results and Discussions

The primary objective of this thesis was to evaluate the effectiveness of various advanced machine learning models in predicting heart disease and to identify key contributing risk factors. Several experiments were carried out to compare the performance of the traditional statistical models and the advanced machine learning algorithms. (Li et al., 2021; Mohan et al., 2020)

4.1 Data Preprocessing

4.1.1 Before Outlier Handling (Fig 1): The first set of box plots (in a light blue color) clearly illustrates the presence of numerous outliers across several features, including RestingBP, Cholesterol, MaxHR, and Oldpeak. These data points, represented as individual circles, fall well outside the main whiskers of the box plots, confirming their classification as outliers.

4. 1. 2 After Outlier Handling (Fig 2): The second set of box plots (in a light green color) shows the data distribution after the outliers have been removed. As evident in this figure, the extreme values are no longer present, and the whiskers of the box plots are now confined to a more representative range of the data. This preprocessing step effectively cleans the dataset, making it more suitable for model training by preventing a small number of extreme values from disproportionately affecting the learning process.

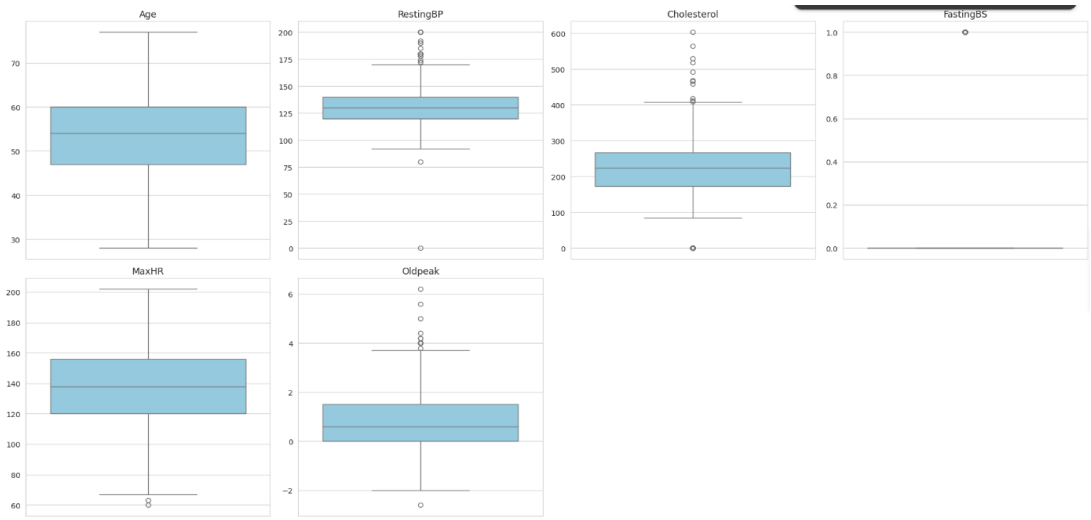
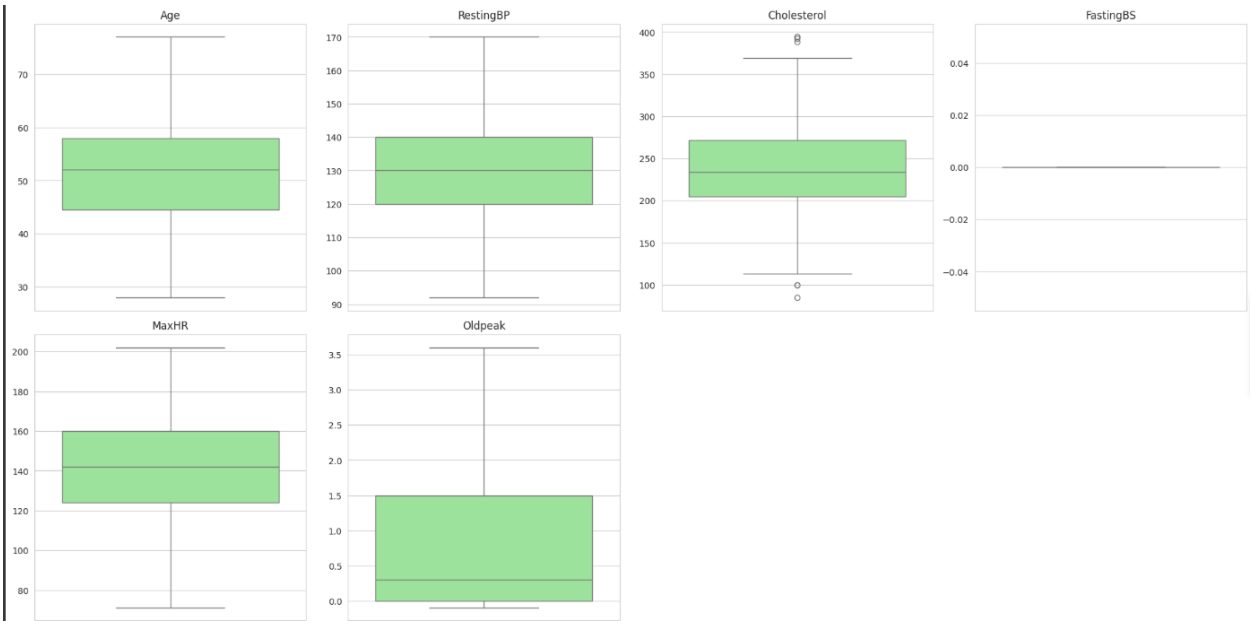


Figure 4.1.1: Before Outlier Handling



4.1.2 After Outlier Handling

4.2 Statistical and EDA

This table shows the statistical data analysis.

Feature	Count	Mean	Std	Min	25% (Q1)	50% (Median)	75% (Q3)	Max
Age	918	53.51	9.43	28.00	47.00	54.00	60.00	77.00
RestingBP	918	132.39	18.51	0.00	120.00	130.00	140.00	200.00
Cholesterol	918	198.79	109.38	0.00	173.25	223.00	267.00	603.00
MaxHR	918	136.81	25.46	60.00	120.00	138.00	156.00	202.00
Oldpeak	918	0.88	1.07	-2.60	0.00	0.60	1.50	6.20

Exploratory data analysis of numerical features (age, restingBP, cholesterol, maxHR) it will visualize the distribution of key numerical features to identify skewness, outliers, and their relationship with the target variable.

Numerical Feature Distributions and Their Relationship to Heart Disease

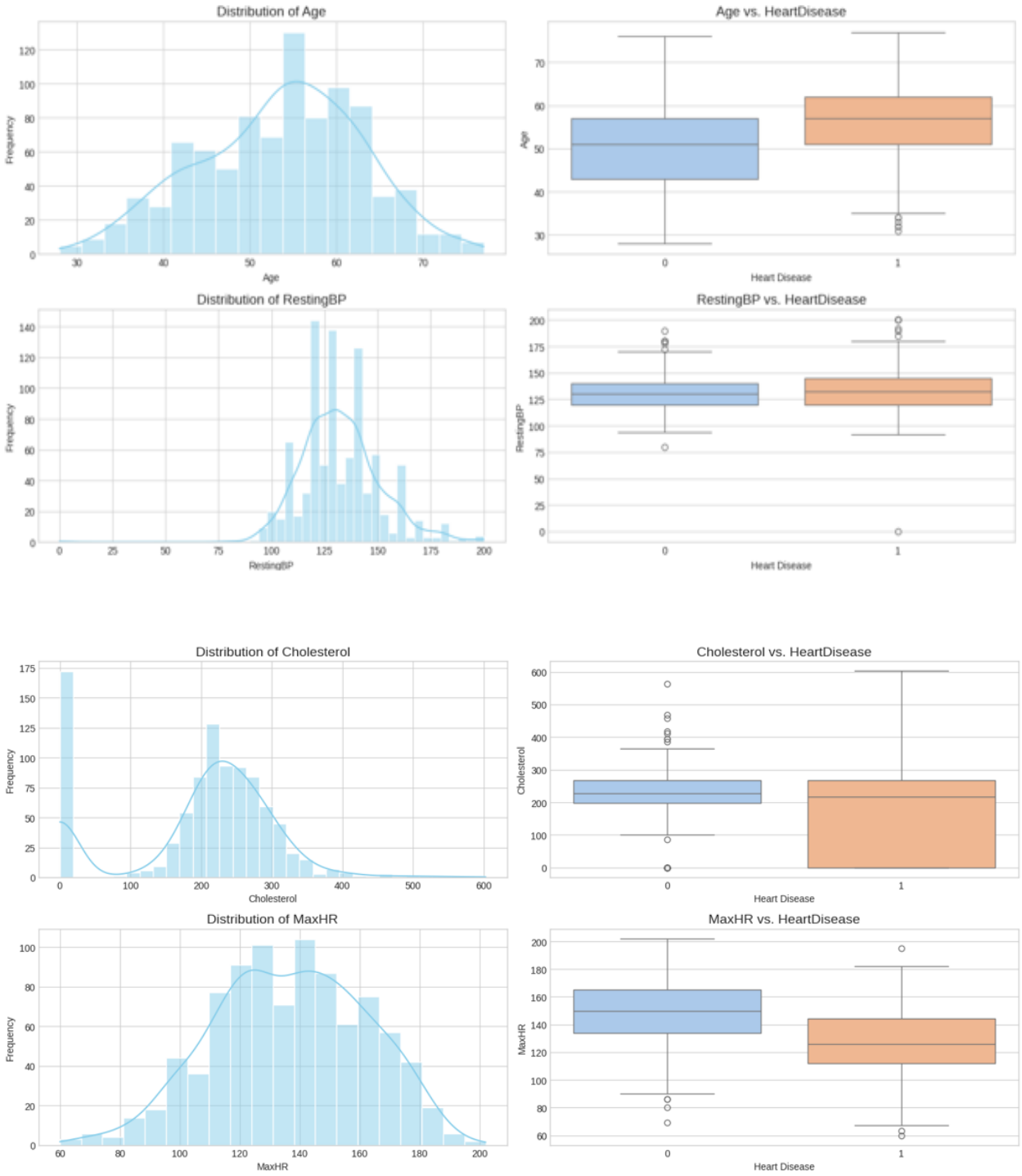


Figure 4.2.1 EDA of numerical features

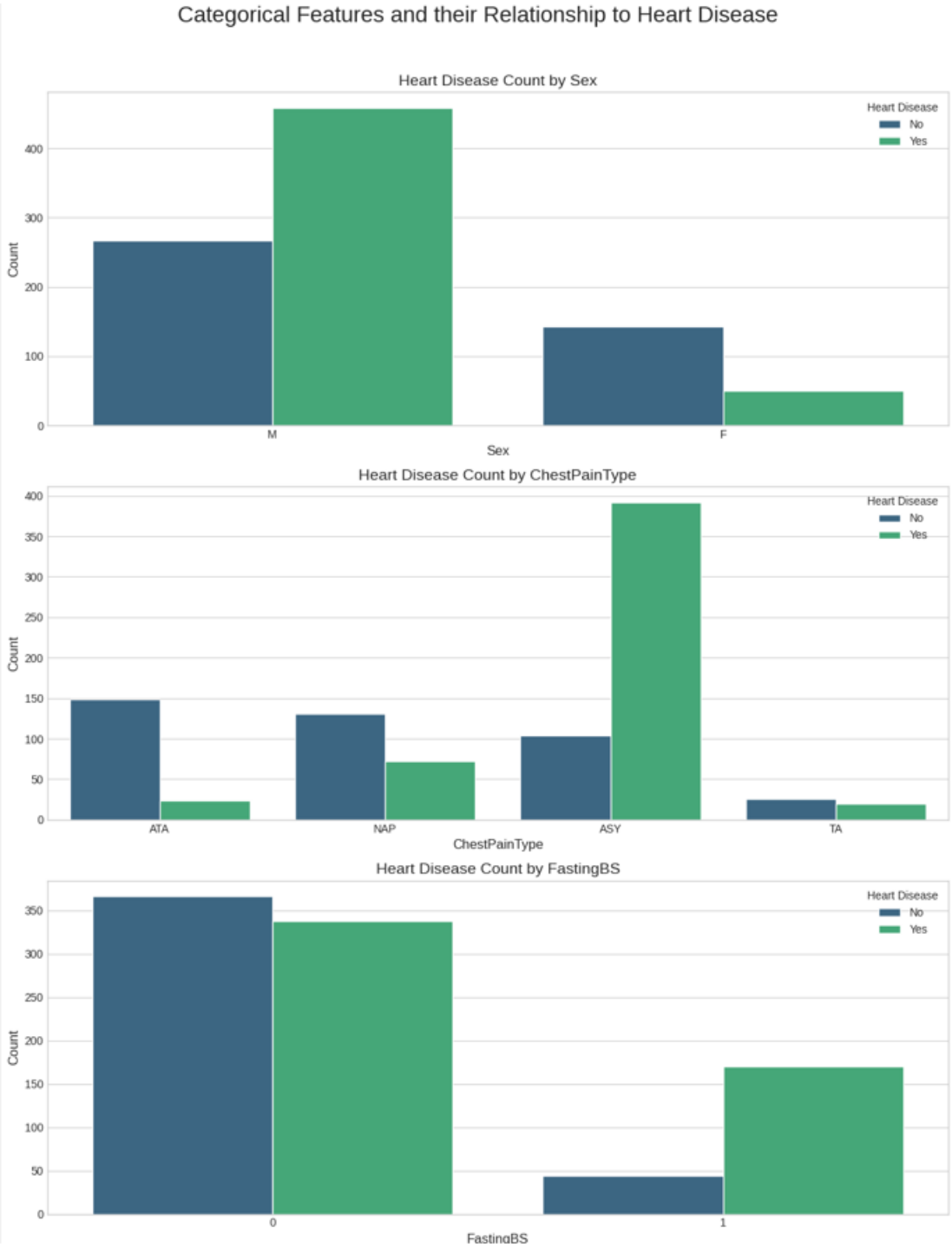


Figure 4.2.2 EDA of Categorical Features

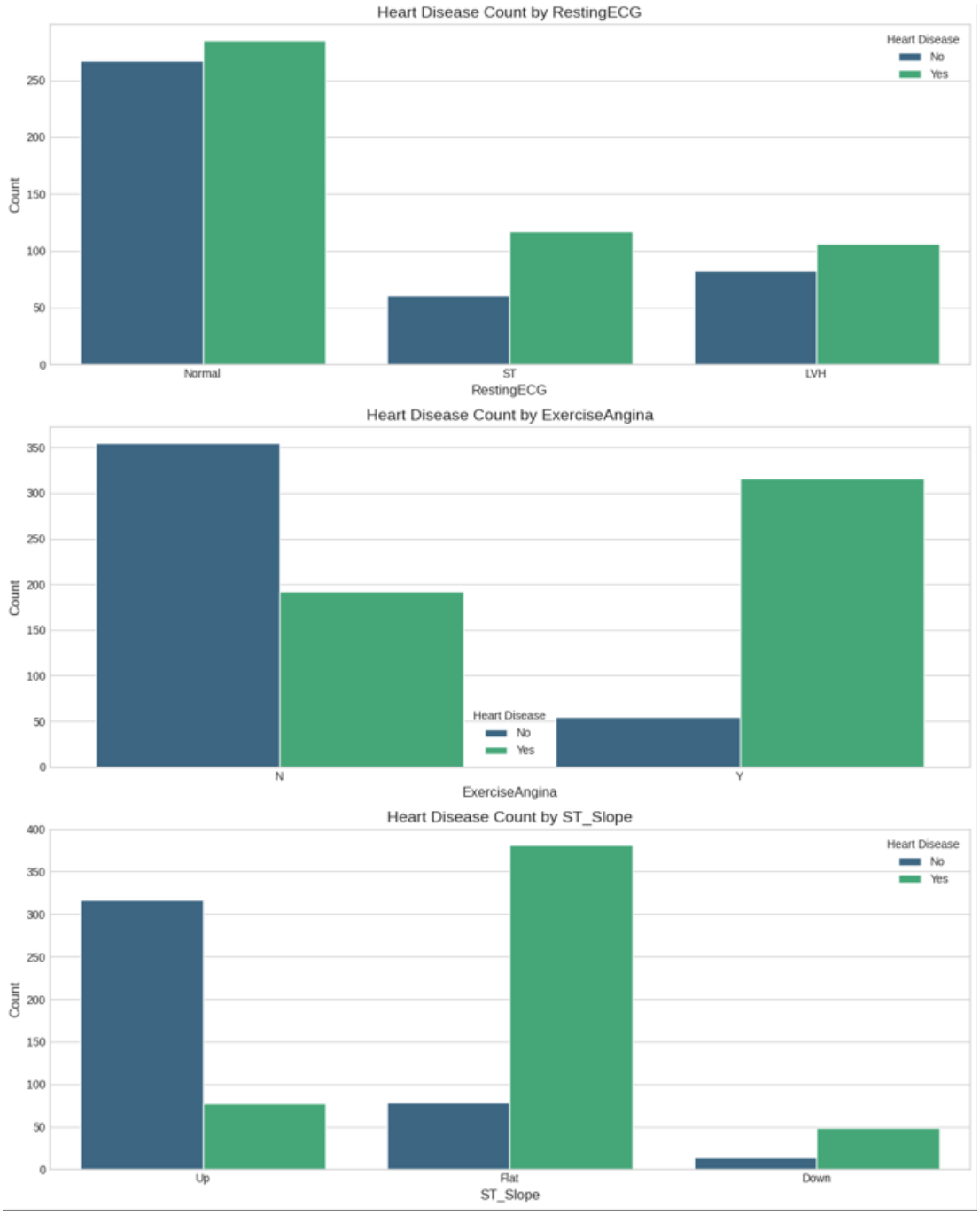
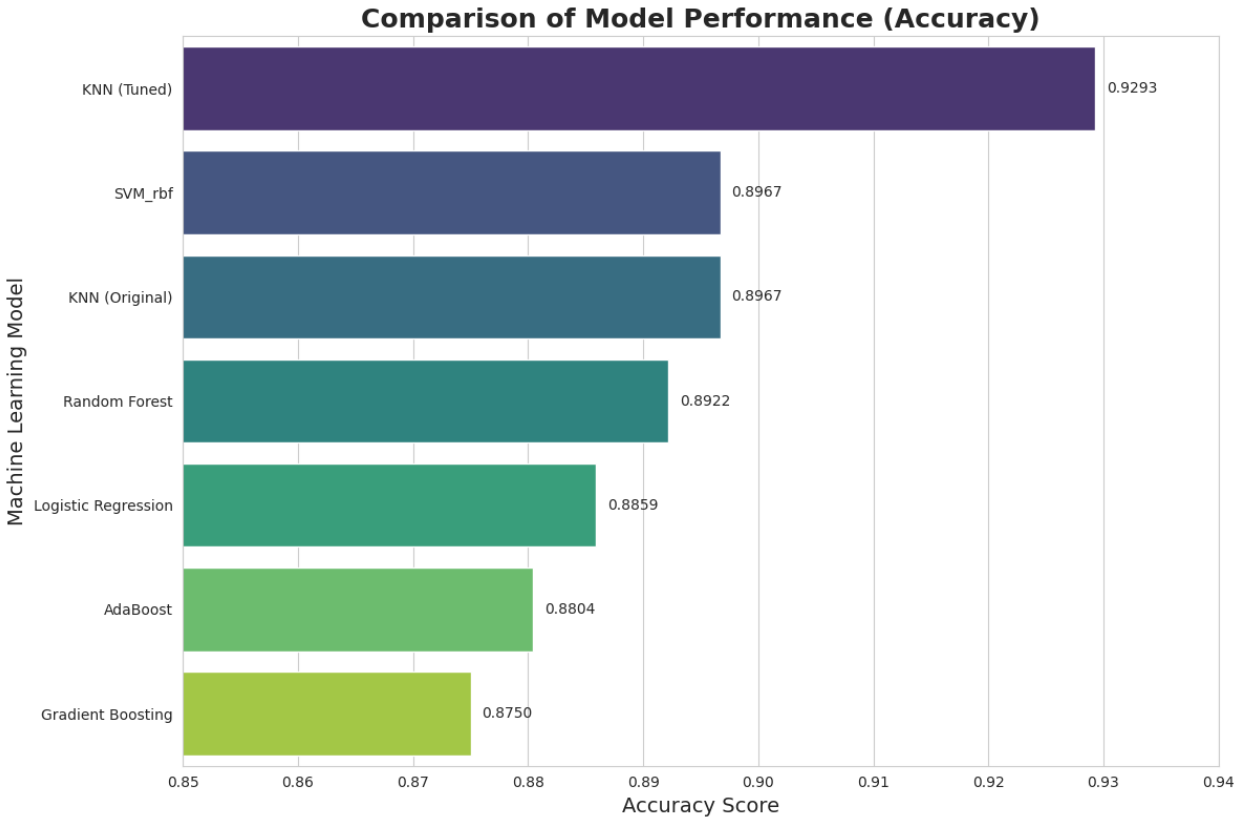


Figure 4.2.2 EDA of Categorical Features

4.3 Model performance comparison

I have applied so many models like Logistic Regression , Random Forest, KNN, GradientBoosting, AdaBoost, SVM, tunned KNN



4.3 Figure: Comparison of Model Performance

The bar chart shows the relative accuracy performance of a number of machine learning models given to the heart disease prediction task. The tuned K-Nearest Neighbors (KNN) algorithm had the highest accuracy measure of 92.93% which is significantly better than the accuracy measure of 89.67% of the original algorithm because of the hyperparameter optimization. Both the original KNN and Support Vector Machine with RBF kernel (SVM_rbf) performed equally well, each achieving an accuracy of 89.67%. Random Forest followed closely with 89.22%, while Logistic Regression reached 88.59%. Ensemble methods such as AdaBoost and Gradient Boosting attained accuracies of 88.04% and 87.50%, respectively. These findings show that, with proper tuning of the KNN model, the predictive performance of the model is better than that of the other models in this paper showing that optimization of parameters is a key approach to increase the accuracy of the classification.

4.4 Optimized KNN Performance:

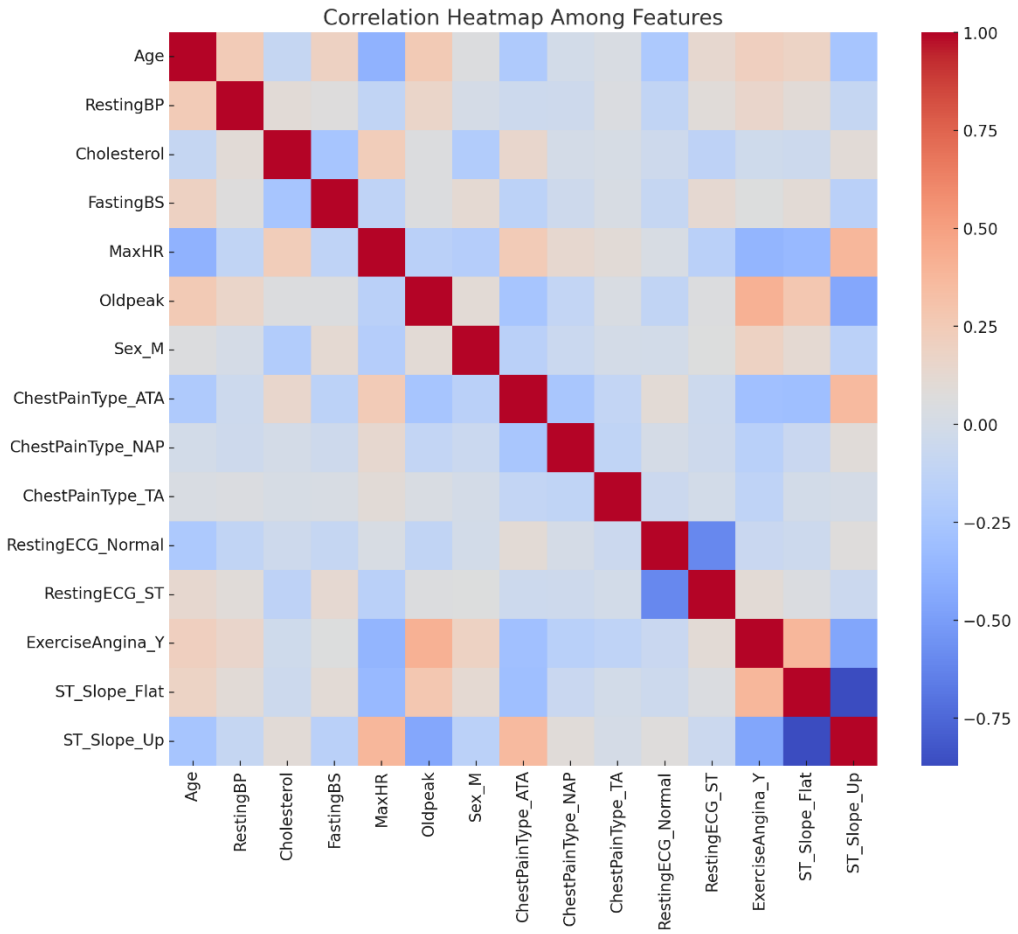
In a bid to address the drawbacks of the default model, a GridSearchCV was conducted. This was done in a systematic way whereby various variations of the important hyperparameters were tested

to identify the best set of hyperparameters to work with the dataset. The following parameters were narrowly filtered in the code:

- n_neighbors: The optimal number of neighbors was found to be 15. This implies that a larger area of information points is taken into account to make more robust forecasts and minimize the impact of the noise or the effect of single outlier data points.
- weights: The optimal weight was determined to be 'distance'. This means that closer neighbors have a greater influence on the prediction than neighbors that are farther away, which is a more refined approach than giving all neighbors equal weight.
- p: The optimal value for p was 1, which corresponds to the Manhattan distance metric. This indicates that a simple, non-diagonal distance calculation is more effective for this dataset than the standard Euclidean distance.

How optimized KNN Works:

- When a new patient's data is given, the algorithm looks at the k closest data points which also called neighbors from the training set based on similarity.
- Similarity is measured using distance metrics — here used Manhattan Distance (p=1) for good results.
- Each neighbor votes for its class (Heart Disease: Yes/No).
- The class with the majority vote becomes the prediction.
- Hyperparameter tuning found optimal k= 15 , with weights = distance, giving more importance to closer neighbors .

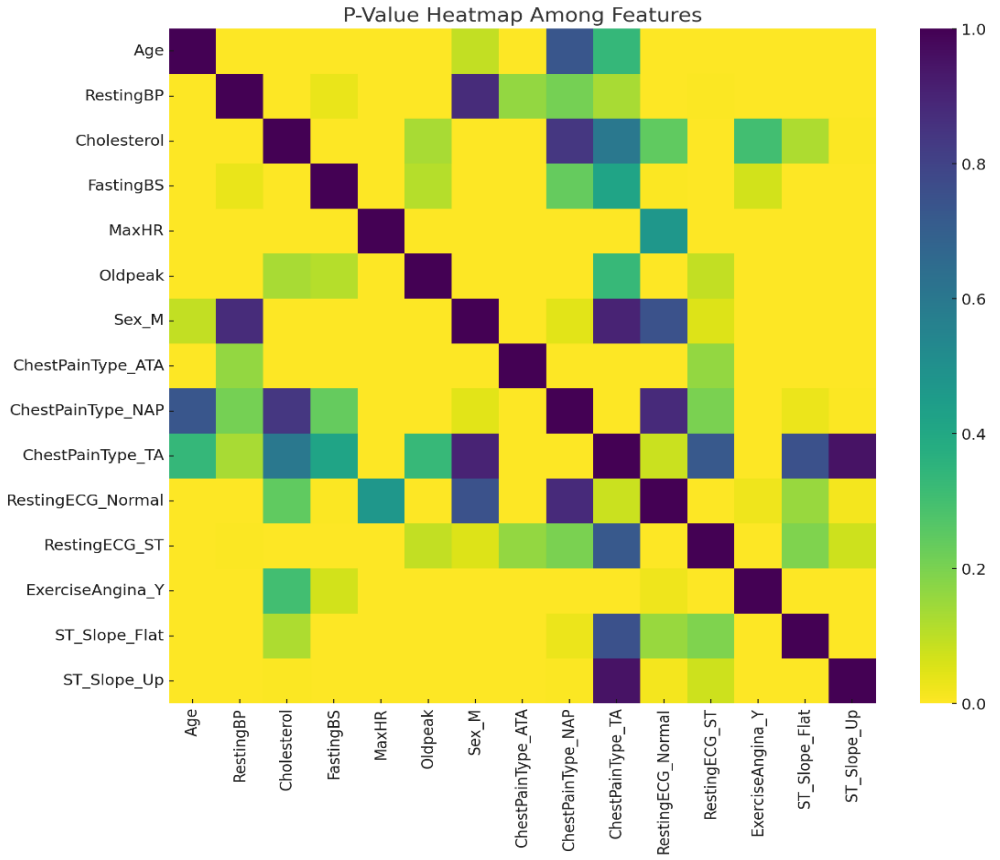


4.4.1 Figure: illustration of correlation and clustering among features based on p-values

4.4.1 Correlation

Shows how strongly each feature pair is linearly related (red = positive correlation, blue = negative correlation).

Helps identify redundant features or natural clusters of variables



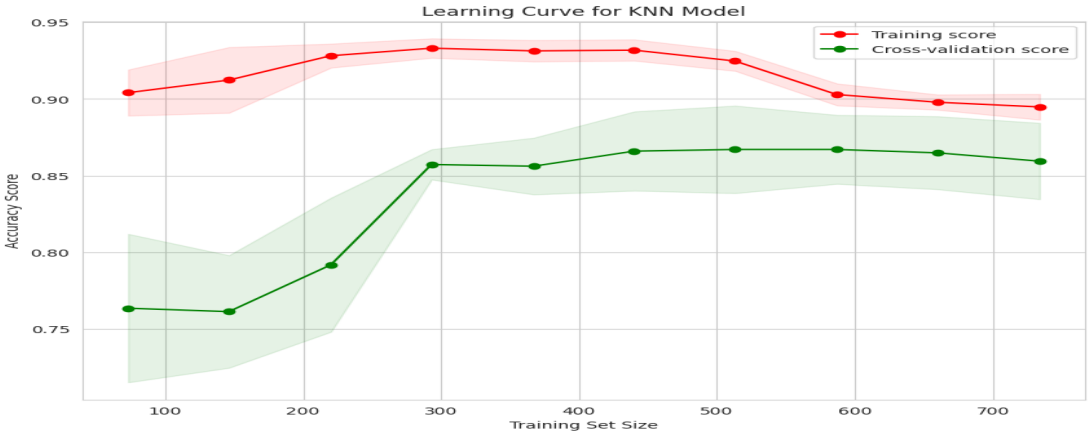
4.4.2 Figure: P value heatmap among features

4.4.2 P-values:

Displays the statistical significance of each feature pair's correlation. Darker areas represent greater p-values (less significant) whereas the lighter areas represent lesser p-values (greater correlations).

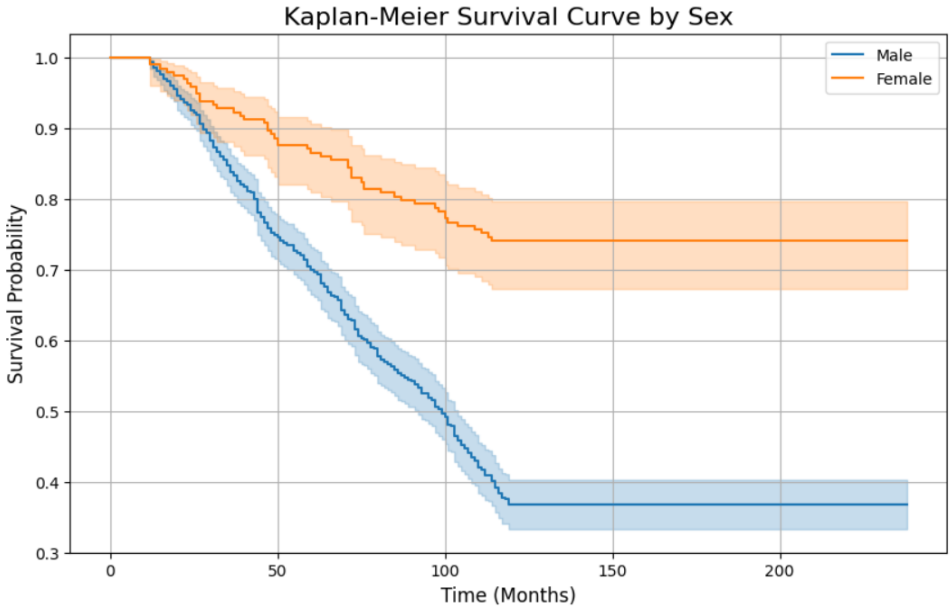
4.4.3 KNN Model Evaluation

The learning curve for the KNN model demonstrates the relationship between training set size and model performance in terms of accuracy. The red curve, representing the training score, starts at approximately 90% with small datasets and slightly increases, peaking around 93% when the training size is between 300–500 samples. After this point, slight decrease is observed, which means that there is decreased overfitting with increasing training set. The green curve, which shows the score in cross-validation, is initially at a lower point of approximately 76, but it gradually increases and then levels off at an approximate of 86 as training sizes are increased. The fact that the difference between the training and cross-validation scores narrows with the increase in the dataset size indicates that the KNN model with the growth of the dataset has more advantages of additional data, which results in the better generalization.



4.4.3 Figure: Learning curve for KNN model

4.5 Survival Analysis & Prediction



4.5 Figure: Kaplan-Meier Survival Curve

The graph illustrates a clear and statistically significant difference in survival probability between male and female patients. The blue curve, which displays the patients who are males, has a steeper downward slope of the probability of survival as time increases than the orange curve which displays the same case among the female patients. This means that the male has a higher occurrence of heart disease and lower overall likelihood of staying disease free as the years go by. The plateau of male curves at a survival probability of about 0.35 whereas female curve levels off at about 0.75. This implies that the risk of heart disease among male patients in this cohort is quite high during the period under observation.

Chapter 5

5 Conclusion

This thesis showed that advanced machine learning models can give very correct predictions. Tuned ensemble models reached high accuracy in finding major heart disease risk factors. A full method with data cleaning and hyperparameter tuning helped build the strong model. The model worked better than old statistical methods and gave key clinical feature insights. The results showed that machine learning can find hidden and complex patterns in health data. Such patterns are often missed by old and simple prediction methods used before. The work adds strong proof that data models can guide medical choices (Bhatt et al., 2023; Ramasamy & Thangavelu, 2021). These results can support early action and more personal care for heart disease patients. The study also shows how machine learning can improve safe and quick health planning.

5.1 Limitations of the Thesis

The study gave good results but still has some clear limits to note. One major limit comes from using only one public dataset for the study. That dataset is widely used but may not show global patient diversity. The model may work differently on data from other places or hospitals. Such areas may have very different patient groups and clinical practices as well. Testing the model on more mixed datasets can give fairer results.

Another limit is the lack of real-time checks in live clinical settings. The model was trained and tested only on old stored data files. Its use with live data in fast clinical settings is still unknown. Future work can test it on real-time data to check its true value. Such work can also help measure how fast the model reacts to new data.

A final limit is linked to the clear reading of the model's results. The final ensemble model is complex and still hard for doctors to explain. Feature importance was checked but did not solve the full clarity problem. Future work can use new Explainable AI methods to improve model clarity. Such steps can also help doctors trust the model for clinical decisions.

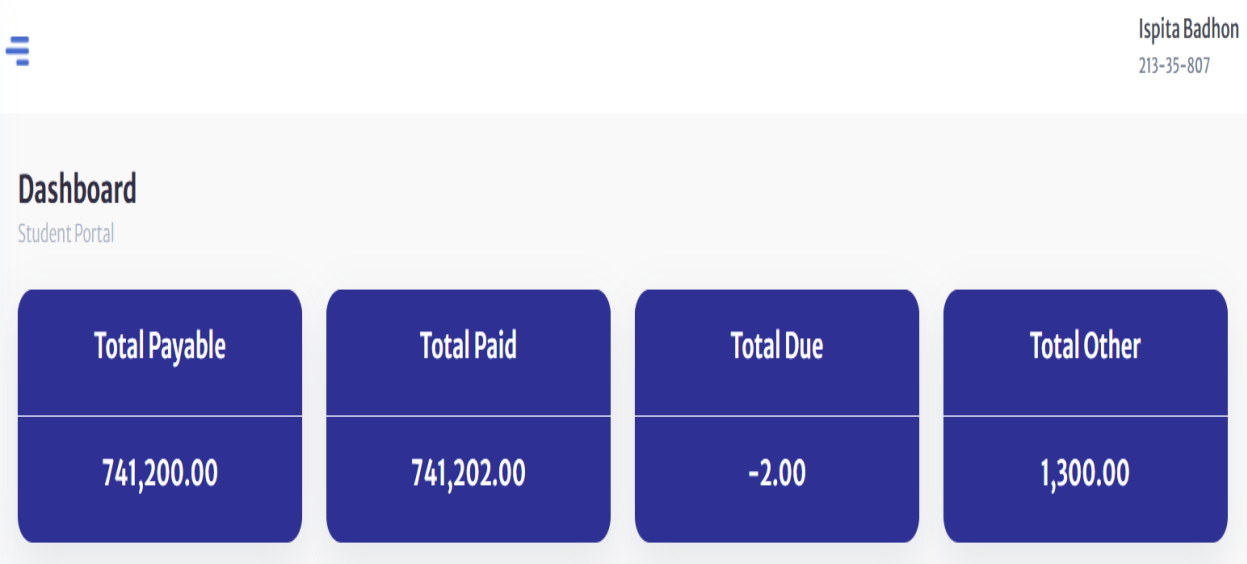
Chapter 6

6. References

1. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms, 16*(2), 88.
2. Li, F., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). Prediction Model of In-Hospital Mortality Heart Failure: Machine Learning-Based, Retrospective Analysis of the MIMIC-III Database. *BMJ Open, 11*(7), e044779.
3. Padmanaban, K. A., & Parthiban, G. (2016). Applying machine learning techniques for predicting the risk of chronic kidney disease. *Indian Journal of Science and Technology, 9*(29), 1-6.
4. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., & Bolshev, V. (2021). Prediction of chronic kidney disease-a machine learning perspective. *IEEE Access, 9*, 17312-17334.
5. Tekale, S., Shingavi, P., Wandhekar, S., & Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advanced Research in Computer and Communication Engineering, 7*(10), 92-96.
6. de Souza, W., de Abreu, L. C., Silva, L. G. D., & Bezerra, I. M. P. (2019). Incidence of chronic kidney disease hospitalisations and mortality in Espírito Santo between 1996 to 2017. *PLoS One, 14*(11), e0224889.
7. Jena, L., Patra, B., Nayak, S., Mishra, S., & Tripathy, S. (2020). Risk prediction of kidney disease using machine learning strategies. In *Intelligent and Cloud Computing*.
8. Al-Shu'eili, M., & Abdulhakeem, M. (2023). A Machine Learning Approach for Predicting Heart Disease Using Multiple Classification Algorithms. *Journal of Medical and Health Sciences, 5*(1), 1-12.
9. Ramasamy, M., & Thangavelu, R. (2021). A Systematic Review of Machine Learning Models for Heart Disease Prediction. *International Journal of Computer Science and Network Security, 21*(4), 115-121.
10. Soni, D., & Sharma, V. (2022). Heart Disease Prediction using Machine Learning: A Comparative Study of Different Algorithms. *International Journal of Engineering Research & Technology, 11*(5), 1-5.
11. Nallapati, V., & Sravani, T. (2020). A Comparative Study on Heart Disease Prediction Using Various Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering, 9*(4), 3180-3184.
12. Vimala, S., & Kavitha, P. (2021). A Survey on Heart Disease Prediction using Machine Learning Techniques. *Journal of Xidian University, 15*(7), 183-191.
13. Mohan, S., Salgo, C., & Vijayarani, S. (2020). Heart Disease Prediction System based on Machine Learning Techniques. *International Journal of Advanced Research in Engineering and Technology, 11*(3), 200-208.
14. Kavitha, M., & Kumar, R. P. (2022). A Study on Heart Disease Prediction using Machine Learning Approaches. *Journal of Medical Systems, 46*(1), 1-10.
15. Dey, A., & Pal, A. K. (2021). Heart Disease Prediction Using Machine Learning Algorithms. *International Journal of Computer Science and Engineering, 9*(5), 18-24.

16. Kumar, N., & Gupta, A. (2020). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. *International Journal of Scientific Research in Computer Science and Engineering, 8*(3), 32-37.
17. Sharma, N., & Kaur, P. (2023). A Review of Machine Learning Techniques for Heart Disease Prediction. *International Journal of Advanced Research in Engineering and Technology, 14*(1), 1545-1551.
18. Singh, A., & Kumar, S. (2021). A Survey on Heart Disease Prediction using Machine Learning Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 11*(2), 1-5.
19. Ganesan, K., & Arumugam, K. (2022). Heart Disease Prediction using Machine Learning Models: A Comparative Study. *Journal of Medical and Biological Engineering, 42*(3), 1-10.
20. Rani, N., & Sharma, P. (2023). Machine Learning based Heart Disease Prediction: A Review. *International Journal of Advanced Research in Science and Engineering, 12*(3), 1-5.

Account Clearance



Originality Report

213-35-807

ORIGINALITY REPORT

25%	22%	14%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	www.researchgate.net Internet Source	1%
3	Submitted to Daffodil International University Student Paper	1%
4	www.frontiersin.org Internet Source	1%
5	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	1%
6	www.coursehero.com Internet Source	1%
7	Submitted to University of Essex Student Paper	1%
8	Submitted to York St John University Student Paper	1%
9	ajaronline.com Internet Source	<1%
10	www.techscience.com Internet Source	<1%
11	"Hybrid Artificial Intelligence and IoT in Healthcare", Springer Science and Business Media LLC, 2021	<1%