PREDICTIVE PRECISION: AN IN-DEPTH EXAMINATION OF MACHINE LEARNING APPROACHES FOR HEART ATTACK PROGNOSIS.

BY

MD RAKIBUL ISLAM RAJ
ID: 201-15-3464

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MS AMATUL BUSHRA AKHI
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

NUSRAT KHAN
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2024

APPROVAL

This Project titled **"Predictive Precision: An In-Depth Examination of Machine Learning Approaches for Heart Attack Prognosis"**, submitted by MD. Rakibul Islam Raj, ID No: 201-15-3464 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25 January,2024.

BOARD OF EXAMINERS

______________________________ **Chairman**

Dr. Sheak Rashed Haider Noori
Professor & Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

______________________________ **Internal Examiner**

Md. Abbas Ali Khan
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

______________________________ **Internal Examiner**

Mohammad Monirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

______________________________ **External Examiner**

Dr. Md. Arshad Ali
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and
Technology University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **MS Amatul Bushra Akhi (Assistant Professor), Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Amatul Bushra 18.01.24

MS Amatul Bushra Akhi
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Nusrat Khan 20/01/24.

Nusrat Khan
Lecturer
Department of CSE
Daffodil International University

Submitted by:

RAJ

MD Rakibul Islam Raj
ID: 201-15-3464
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to **MS Amatul Bushra Akhi (Assistant professor)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*ML*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of CSE, for their kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Currently several diseases have become epidemic in Bangladesh, one of them is heart attack. Anyone can suffer a heart attack at any age. Generally, older people and men are more prone to it. However, women are also at increased risk of heart attack as they age. Individuals who smoke, have diabetes, high blood pressure, high cholesterol, or both are also more vulnerable. A family history of coronary artery disease or ischemic heart disease increases the risk for others in the family. By applying machine learning and deep learning, this paper proposes to automate heart attack diagnosis. This dataset consists of both heart attack patients and non-heart attack patients. Heart attacks were detected and its types determined in this study, taking the classification process a step further, since most previous studies only detected heart attacks or classified them into a few types. We can solve this by using artificial intelligence, for example, With the use of machine learning and deep learning algorithms, we are able to identify heart attacks and get alerts about them. In this paper I used machine learning and deep learning algorithms, through which we used 1 model of deep learning and 4 models of machine learning. CNN model of deep learning is used here, and Random Forest, KNN, Decision Tree, SVM model of machine learning is used. The accuracy of 76% using CNN model of deep learning, and accuracy of 94.9% using Random Forest model of machine learning, accuracy of 87.6% using KNN model, accuracy of 91% using Decision Tree model 63% using SVM model. All techniques demonstrate that the Random Forest model yields the best accuracy. Using the SVM, the lowest accuracy was attained.

TABLE OF CONTENTS

CONTENTS		PAGE
Board of examiners		i
Declaration		ii
Acknowledgements		iii
Abstract		iv
List of figures		vii
List of tables		viii
CHAPTER		
CHAPTER 1: INTRODUCTION		**1-5**
1.1.	Introduction	1
1.2.	Motivation	2
1.3.	Relational of the Study	2
1.4.	Research Questions	3
1.5.	Expected Outcome	4
1.6.	Report Layout	4
CHAPTER 2: BACKGROUND		**6-12**
2.1.	Terminologies	6
2.2.	Related work	5
2.3.	Scope of the Problem	11
2.4.	Challenges	12
		13-24

CHAPTER 3: RESEARCH METHODOLOGY

3.1.	Proposed Method	13
3.2.	Dataset Description	14
3.3.	Dataset Preprocessing	15
3.4.	Proposed Model	19

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **23-35**

4.1.	Discussion	25
4.2.	Experimental Results and Analysis	26

CHAPTER 5: IMPECT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **36-40**

5.1.	Impact on Society	36
5.2.	Impact on Environment	37
5.3.	Ethical Aspects	38
5.4.	Sustainability	39

CHAPTER 6: SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH **41-44**

6.1.	Summary of the Study	41
6.2.	Conclusions	42
6.3.	Implication for Further Study	43

REFERENCES **45**

LIST OF FIGURES

FIGURES	**PAGE NO**
Figure 1: Architecture of Working Process	13
Figure 2: Architecture of Working Process	16
Figure 3: Dataset Normalization	17
Figure 4: Correlation Matrix of Data Parameters	18
Figure 5: PCA for Feature Extraction	19
Figure 6: CNN Architecture	20
Figure 7: Random Forest Architecture	21
Figure 8: SVM Architecture	22
Figure 9: KNN Architecture	23
Figure 10: Dicision Tree classifier Architecture	24
Figure 11: AUC Score of each model	27
Figure 12: Cross Validation Score of each model	29
Figure 13: Mean Cross Validated Score of each model	30
Figure 14: Misclassification Error of each model	31
Figure 15: Jaccard Score of each model	33
Figure 16: Confusion Matrix of each model	33
Figure 17: Classification Report of each model	35

LIST OF TABLES

TABLES	PAGE NO
Table 1: Dataset Description	14
Table 2: Classifiers Description	26
Table 3: AUC Score for all Method	27
Table 4: Cross Validation Score for all Method	28
Table 5: Mean Cross Validation Score for all Method	30
Table 6: Misclassification Error for all Method	31
Table 7: Jaccard Score for all Method	32
Table-8: Classification Report	35

CHAPTER 1
INTRODUCTION

1.1. Introduction

Predicting heart attacks with deep learning and machine learning involves collecting patient data, selecting relevant features, training models to learn patterns, and using these models to predict the likelihood of a cardiac event. Machine learning employs algorithms like decision trees or support vector machines, while deep learning uses neural networks like CNN or RNN for complex pattern recognition [2]. Challenges include data quality, interpretability of deep learning models, and ethical considerations, but these methods hold promise for early detection and personalized healthcare. Predicting heart attacks using machine learning and deep learning techniques has become an important area of healthcare. The ability to predict the likelihood of cardiac events can significantly aid in early intervention and prevention strategies [4].

Relevant patient information is collected, including medical history, lifestyle factors, genetic information and clinical tests (such as ECG, cholesterol levels, etc.). Determination of key elements linked to heart disease, including age, blood pressure, cholesterol, smoking status, and so on. Using this data, algorithms are trained to find patterns and connections between characteristics and the incidence of heart attacks, such as logistic regression, decision trees, random forests, or support vector machines. Model performance is evaluated using validation techniques such as cross-validation or splitting the data into training and testing sets. As with machine learning, data preparation involves cleaning, normalization and feature extraction [1]. Deep learning often requires more complex preprocessing due to the nature of neural networks. Convolutional and recurrent neural network architectures are two examples of neural network designs that may be used to identify complicated patterns in data. Using large datasets, deep neural networks are trained using methods such as gradient descent and backpropagation to optimize network weights. Assessing the efficacy of the model in predicting heart attacks by utilizing criteria like accuracy, precision, recall, and F1-score [5]. Availability of comprehensive and high-quality data is crucial for accurate predictions. Interpreting deep learning models can be challenging due to their complexity.

Personalized medicine: moving toward personalized risk assessment considering individual genetic variation and lifestyle factors. Ensuring patient confidentiality and ethical use of sensitive medical data [7].

In summary, Heart attack prediction using a combination of deep learning and machine learning approaches shows great potential for early identification and individualized management options, contributing to improved patient outcomes and healthcare management.

1.2. Motivation

The pressing demand for preventative healthcare solutions is what drives the use of deep learning and machine learning in heart attack prediction. By leveraging these technologies, we aim to: - By identifying high-risk patients, predictive models can facilitate early intervention and lifestyle adjustments aimed at preventing heart attacks. Tailoring interventions based on individual patient data ensures personalized care and targeted preventive measures. Efficiently allocating medical resources by prioritizing high-risk patients for further examinations or treatments. Proactive prevention is more cost-effective than reactive treatments, potentially reducing the economic burden of heart-related healthcare. Early identification can lead to timely medical interventions, ultimately saving lives and improving patient outcomes. Analyzing vast amounts of patient data fosters a deeper understanding of heart disease, contributing to advancements in cardiovascular research. Early detection can prevent heart attacks, potentially saving lives. Predictive models can optimize healthcare resources by focusing on high-risk patients. Applying innovative technologies improves diagnostic capabilities and fosters medical advancements. Tailoring interventions based on individual data promotes more effective and personalized healthcare. Addressing cardiovascular diseases globally by employing proactive measures for prevention.

Overall, predicting cardiac attacks using deep learning and machine learning is driven by the pursuit of proactive, personalized, and effective healthcare strategies to reduce the incidence and severity of cardiac events.

1.3. Relational of the Study

Various diseases like heart attack and to improve their detection and awareness system. For example, proper assessment of the patient first, technology-based care for a patient, use of skilled

physicians to treat the patient, etc. Although there are skilled doctors at the district level, accurate diagnosis is not possible due to lack of modern equipment. It is true that there are competent doctors at the departmental level but our diagnostic machine does not give good results. As a result, the correct diagnosis cannot be made. It is never possible to cure the patient until the correct diagnosis is made. A number of models can be developed using technology that will play an important role in the diagnosis of diseases of doctors today. A heart attack patient's whole body is dealt with by proper diagnosis; They use review for different diagnoses like us and use different algorithms like DL and ML for different diagnoses.

1.4. Research Questions

This research may include a wide range of question kinds. As an illustration,

- What is the aim of this study?
- Which features in the database perform the highest impact on the accuracy of heart attack prediction?
- How accurately can machine learning models predict the heart attack?
- Which machine learning algorithm performs the best in terms of accuracy, precision, and generalizability in prediction?
- What are the benefits of prediction?
- What is the overall success rate in predicting heart attack?
- How will this thesis help us?

1.5. Expected Outcome

The combination of deep learning (DL) and machine learning (ML) methods for heart attack prediction is a big advancement in preventive healthcare. These models demonstrate the ability to accurately predict risk by utilizing a multitude of patient data, such as demographics, medical histories, and lifestyle variables. Important risk indicators can be identified thanks to the effective feature selection provided by machine learning algorithms like Random Forests and Support Vector Machines.

Deep learning models, particularly neural networks, excel in discerning intricate patterns within extensive datasets. The expected outcomes encompass heightened predictive accuracy, enabling the early identification of individuals at high risk of experiencing a heart attack. These models facilitate personalized risk assessments, tailoring predictions to individual patient profiles. The integration of ML and DL empowers healthcare professionals with actionable insights, fostering timely intervention strategies to mitigate cardiovascular risks.

Moreover, the continuous learning and adaptive capabilities inherent in deep learning models contribute to ongoing refinement, ensuring sustained accuracy as they encounter new data. The implementation of these advanced technologies holds the potential to revolutionize preventive cardiology by providing a proactive approach to heart health. This might therefore result in better patient outcomes, lower healthcare expenses connected to heart-related conditions, and a paradigm change in cardiovascular treatment towards preventative rather than reactive measures. In summary, the convergence of ML and DL in heart attack prediction offers a transformative path towards more effective, personalized, and preemptive cardiovascular healthcare.

1.6. Report Layout

Due to heart attack symptom, our body gets tired easily, fever and night sweats, sudden weight loss, pain in bones and joints, pain under left rib, bleeding from nose and gums, red blisters etc. has to be heart attack detection might be much improved if the condition could be automatically identified with machine learning (ML) and deep learning (DL) in a method that is quicker, more flexible, less expensive, and more dependable. The remaining portion of the essay is structured as follows: Section 2- Background, Section 3-Methodology, Section 4- Experimental Results and Discussion, Section 5- Impact on Society, Environment and Sustainability. Finally, Section 6 will conclude the paper.

CHAPTER 2
BACKGROUND

2.1. Terminologies

Here are some key terminologies related to predicting heart attacks using machine learning and deep learning. Like Supervised Learning, Feature Selection, Model Training, Algorithm, Neural Networks, here are some key terminologies related to predicting heart attacks using machine learning and deep learning. A method of machine learning wherein labelled data is used to teach the model correlations and patterns, making predictions based on input-output pairs (e.g., patient data with known heart attack outcomes). The process of choosing the most relevant variables or features from the dataset that have a significant impact on predicting heart attacks. The phase where machine learning algorithms or neural networks adjust their parameters by learning from the provided dataset to make accurate predictions. a particular set of guidelines or rules that the machine learning model uses to interpret information and forecast outcomes. a kind of deep learning model made of linked layers of nodes that analyze and learn from input, and which is modelled after the architecture of the human brain. the act of adding new features or altering current ones in order to increase the model's capacity for prediction.

2.2. Related Work

The study created a model to forecast cardiac attacks using machine learning likelihood using a real-world dataset of patient features. Various algorithms were evaluated, including logistic regression, decision trees, and random forest. The algorithms' ability to accurately forecast heart attacks and identify people at high risk was demonstrated by the results. However, the XGBoost Classifier showed superior performance, with significant test accuracy and a high mean area under the curve [1].

In this research, we offer a machine learning-based heart attack prediction (ML-HAP) system that uses ML methodologies such as Support Vector Machines, Logistic Regression, Naïve Bayes, and XGBoost to analyse various risk variables and predict heart attacks. The UCI ML Repository provided the data on heart disease symptoms, which were then analysed using machine learning techniques. Out of the four, XGBoost offered the most accurate prediction. Logistic regression yields an area under the curve of 0.92 while XGBoost produces one of 0.94 [2].

Based on user input, the AI-driven health risk assessment application's main goal is to forecast the likelihood of heart attacks. Among these algorithms are Random Forest, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, SG Booster, Voting Ensemble (Logistic Regression, KNN, Decision Tree), and Logistic Regression. To ensure accuracy and effectiveness, a dataset containing risk factors data from 606 patients was collected. This dataset was used to thoroughly train and test the machine learning models, with the Random Forest algorithm emerging as the most accurate, achieving an impressive accuracy rate of 95.04% [3].

The objective of this research is to develop a machine learning model that can predict a person's risk of suffering a heart attack based on a variety of medical parameters. The data will be cleaned, pre-processed, and analysed using machine learning methods such as the naive Bayes classifier, SVM classifier, and KNN classifier. The model will be trained and evaluated on a dataset of individuals who are known to be at risk for heart attacks using measures like as precision, recall, and the F1-score. The completed model will be implemented as an online application where users may provide their medical details to receive a personalised heart attack risk score [4].

The UCI Machine Learning repository's dataset on heart disease is used in this work. This paper uses seven machine learning methods, including Naive Bayes, Decision Tree, Logistic Regression, KNN (K-Nearest Neighbours), SVM (Support Vector Machine), Gradient Boosting, and Random Forest, to predict the likelihood of heart disease. The results acquired from the classification report confirms that the KNN (K-Nearest Neighbors) algorithm achieved a very high accuracy of 85.18% compared to other ML algorithms used [5].

In this paper, we provide a method for predicting heart attacks that leverages machine learning to assess many risk variables and generate heart attack predictions. The study paper's main focus is on identifying people who are more prone to develop heart disease based on specific medical criteria. We created a mechanism to assess the likelihood of a heart disease diagnosis for a patient based on their medical history. According to experiments, the Naive Bayes algorithm has the best accuracy rate (88%), followed by ANN and KNN (87%) [6].

This work makes use of the machine learning repository's heart disease dataset. The proposed study classifies the patient's risk level and forecasts the patient's risk of heart disease by using KNN and SVM. Thus, in order to offer comparative study, this article looks at the efficacy of several machine learning algorithms. The test results demonstrate that the algorithm achieves a maximum accuracy of 94.3 percent when compared to other ML algorithms that were utilized [7].

The scenarios that arise from Heart Disease are calculated in this study. In this study, a robust machine learning approach called the Random Forest algorithm is used to construct a dependable cardiac disease prediction system. The algorithm reads a CSV file containing patient record data. Following the procedure and dataset access, an effective heart attack level is generated. Benefits of the suggested system include excellent rates of success are attained, along with excellent performance and accuracy rates as well as flexibility [8].

Many machine learning methods have been presented recently for the diagnosis and prediction of heart disorders. This study aims to evaluate the predictive power of a deep learning (DNN) model using a commonly used reference dataset for cardiovascular disease incidence. The deep learning model outperformed other machine learning models in terms of accuracy, F1-score, and precision (97.07%, 97.07%, and 97.08%) [9].

Many machine learning methods have been presented recently for the diagnosis and prediction of heart disorders. This study aims to evaluate the predictive power of a deep learning (DNN) model using a commonly used reference dataset for cardiovascular disease incidence. The deep learning model outperformed other machine learning models in terms of accuracy, F1-score, and precision (97.07%, 97.07%, and 97.08%) [10].

This research used Logistic Regression algorithm and 14 physical indicators from 302 patients to investigate heart attack. It found that because the correlation coefficient is greater than 0.4, the correlation that has to do with heart attack and what kind of chest pain it is, the maximum value of heart rate, whether exercise induced angina and the ST depression was strong. Correlations have to do with heart attack and cholesterol, the fasting blood sugar was the weakest. The accuracy of the prediction is about 85.95% [11].

Finding the most effective and precise machine learning models for illness prediction is the goal of this article. Heart disease was diagnosed and predicted using a number of supervised machine learning methods, including logistic regression, decision trees, random forests, and KNN. The methods are used on a dataset including 70,000 samples that was downloaded from the Kaggle website. The assessment shows that the random forest method performs better than other algorithms, with an F1 score of 92%, accuracy of 92%, and AUC ROC of 95% [12].

This project used SVM, KNN, LR, RF and DT algorithms to improve heart disease detection using UCI data. The goal of these programs is to be able to tell if a person has heart disease or not. It has 14 classification features. We found that among the five algorithms Random Forest performs best by giving 98.05% accuracy, Decision Tree performs next to Random Forest with 97.08% accuracy, Support Vector Machine gives 90.25% accuracy, K-Nearest Neighbor and Logistic Regression gives 81.82% accuracy [13].

The goal of this article is to identify the machine learning models for illness prediction that are the most effective and precise. A number of supervised machine learning techniques, including logistic regression, decision trees, random forests, and KNN, were used to diagnose and predict cardiac disease. A dataset including 70,000 samples is downloaded from the Kaggle website and subjected to the algorithms. The random forest method outperforms the other algorithms according to the evaluation, with an F1 score of 92%, accuracy of 92%, and AUC ROC of 95%. [14].

This project used SVM, KNN, LR, RF and DT algorithms to improve heart disease detection using UCI data. The goal of these programs is to be able to tell if a person has heart disease or not. It has 14 classification features. We found that among the five algorithms Random Forest performs best by giving 98.05% accuracy, Decision Tree performs next to Random Forest with 97.08% accuracy, Support Vector Machine gives 90.25% accuracy, K-Nearest Neighbor and Logistic Regression gives 81.82% accuracy [15].

This work contributes to the field by facilitating the diagnosis of cardiac issues through the use of machine learning techniques on the widely used Cleveland heart disease dataset. A variety of performance metrics are employed to assess the potency of any model. Thus, based on the computational complexity of each model and the experimental findings, an enhanced prediction of heart disease for an embedded platform is provided, whereby the benefits of many classifiers are aggregated [16].

Machine learning algorithms are applied to a wide range of illnesses. Additionally, it is a branch of AI that works with teaching a computer to think on its own. Nowadays, heart attacks afflict the majority of people, which causes doctors' heads to hurt. We must anticipate heart attacks in order to lower the mortality toll. In this research, machine learning is a key component in solving this challenge. A person's life is lifted out of danger by this forecast. This study employs the KNN algorithm and the Random Forest method to anticipate cardiac attacks [17].

This research aims to provide a machine learning-based, lightweight solution for the detection of cardiac disease. Many datasets are accessible, including those from Cleveland, Hungary, and Switzerland. This study and testing employ the Cleveland dataset, which consists of 303 patient records and 14 characteristics. Six distinct machine learning approaches were compared in this study using a range of performance criteria. According to the research, SVM yields the highest results with 89.34% among the six approaches [18].

The study employed the Recurrent Neural Network (RNN) classification algorithm with varying parameters to assess the risk of heart attack. The influence of appropriate parameter selection on classification accuracy was examined by taking into account a number of parameters that have an impact on the RNN model's performance. Accuracy, precision, recall, and F1 score were among the metrics used to assess the categorization performance [19].

The random forest and decision tree regression models are used in this study to assess the existing data set of patients with symptoms of heart abnormalities. The attribute's mean value is used to update the missing data. For the purpose of accuracy prediction, Python is employed. For the purpose of analysing the UCI Cleveland heart disease data set, three performance metrics are used. The random forest regression model and the decision tree regression model are used to assess key characteristics of the data set [20].

2.3. Scope of the Problem

Given the significance and prevalence of cardiovascular disease, the scope for using machine learning and deep learning to predict heart attacks is huge. The scope of heart attack prediction using machine learning and deep learning is broad and multifaceted. It involves addressing the global impact of cardiovascular diseases, integrating diverse risk factors and medical data, aiming for personalized and preventive healthcare solutions, while ensuring ethical use of data and resources. Here are the key aspects that determine the scope of the problem:

- Data Availability: Vast amounts of diverse patient data, including medical records, imaging, genetic information, lifestyle factors, and physiological measurements, provide a rich source for predictive modeling.
- Complications of heart disease: cardiovascular disease is multifaceted, influenced by numerous interrelated factors. Machine learning and deep learning can help decipher complex patterns within this complexity.
- Predictive power: These technologies can detect subtle correlations and non-linear relationships between different risk factors, increasing predictive accuracy beyond conventional risk assessment methods.
- Continuous Monitoring: Real-time data from wearable devices and health sensors enables continuous monitoring, facilitating early detection and intervention.
- Integration with healthcare systems: Implementing predictive models within healthcare systems allows for seamless integration into clinical workflows, supporting healthcare professionals in decision making.

- Challenges and Opportunities: Ethical considerations, data privacy, Challenges include the necessity for comprehensive validation and the interpretability of the model but also present opportunities for further research and development.
- Global impact: cardiovascular diseases, including heart attacks, are a primary global cause of sickness and death. Opportunities to address the global burden of these diseases and implement predictive solutions on a large scale expand.
- Various risk factors: Heart attack is affected by many factors, including age, genetics, lifestyle choices and medical history. The scope includes the challenge of integrating and analyzing different datasets to account for these complex risk factors.

2.4. Challenges

i. Data Quality and Quantity: Obtaining comprehensive, high-quality datasets that encompass diverse patient information, including medical history, genetics, lifestyle factors, and imaging data, can be challenging. Limited or incomplete data can affect the accuracy of predictive models.

ii. Feature Selection: Identifying the most relevant features or risk factors associated with heart attacks among extensive datasets is complex. Determining which features are most predictive and informative for the models can be crucial for accurate predictions.

iii. Imbalanced Data: Heart attacks are relatively rare compared to non-events in datasets. Unbalanced data might result in models that are biassed in favour of the majority class, affecting the model's ability to accurately predict rare events.

iv. Interpretability: Neural networks in particular, which are deep learning models, are frequently regarded as "black boxes," making it difficult to understand how they make predictions. Interpretability is essential in the healthcare industry to build confidence and comprehend the rationale behind forecasts.

v. Generalizability: Ensuring that predictive models generalize well to unseen data or diverse populations is essential. Models trained on specific datasets may not perform well on different demographics or when applied in different healthcare settings.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Proposed Method

We employed both supervised and unsupervised learning techniques since the data presents a multivariate classification challenge. Here, a combination of machine learning and deep learning techniques have been applied. For every model, a confusion matrix, accuracy, performance prediction, and precision results are produced. Figure displays the overall system architecture:

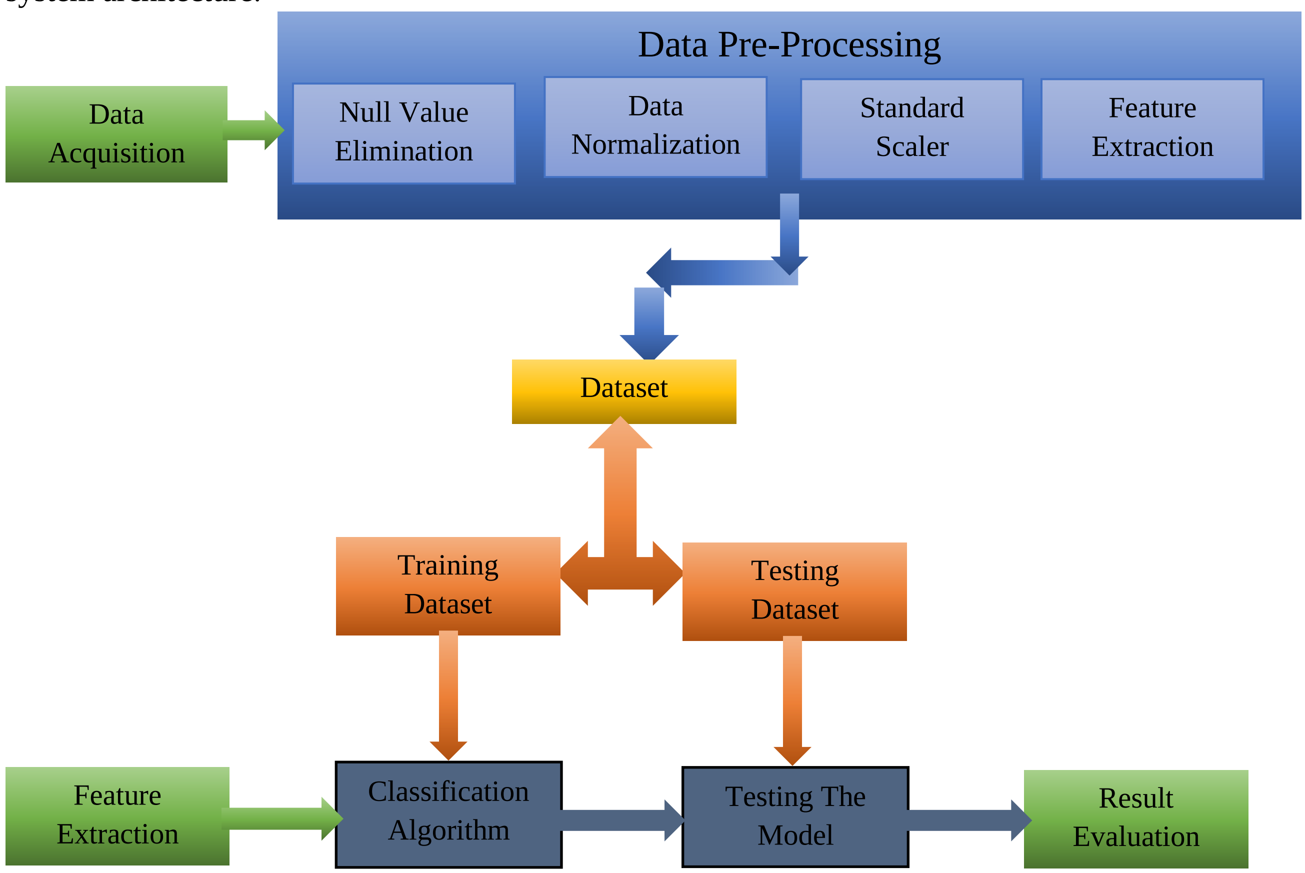


Figure-1: Architecture of Working Process

3.2. Data Description

Our dataset was collected from Kaggle and the dataset was not preprocessed. Our dataset is attack reports, through which we know whether or not a heart attack will occur. There are 253662 data points in all in the dataset. Twenty percent of the dataset's total was designated for testing, while the remaining eighty percent was used for training. The dataset has a total of 22 columns, out of which all columns are of numeric type. The dataset contains 21 independent variables and 1 dependent variable. That is, the value of 1 dependent variable depends on 21 independent variables. The dataset has 1 dependent variable column on whether to have a heart attack.

Table-1: Dataset Description

SL	Attribute Name	Full Form	Description
1	HeartDiseaseorAttack	Heart Disease or Attack	This column indicates whether heart disease is present, yes or no.
2	HighBP	High Blood Pressure	This column indicates whether there is high blood pressure.
3	HighChol	High Cholesterol	This column indicates whether there is high cholesterol.
4	CholCheck	Cholesterol Check	This column indicates whether cholesterol is present.
5	BMI	BMI	This column indicates what is the present BMI.
6	Smoker	Smoker	This column indicates whether one smokes or not.
7	Stroke	Stroke	This column indicates whether a person has had a previous stroke.
8	Diabetes	Diabetes	This column indicates whether diabetes is present.
9	PhysActivity	Physical Activity	Doing Physical activities or exercise during past 30 days other than their regular job.
10	Fruits	Fruits	Consume Fruit one or more times per day
11	Veggies	Veggies	Consume vegetables one or more times per day
12	HvyAlcoholConsump	Heavy Alcohol Consump	Heavy drinkers (Men having more than 14 times per week and women having more than 7 times per week)

13	AnyHealthcare	Any Health Care	This column describes the health care coverage such as health insurance, Medicare or health service.
14	NoDocbcCost	No Doctor Cost	This column indicates the doctor's cost in the past 12 months.
15	GenHlth	General Health	This column indicates the situation of the general health.
16	MentHlth	Mental Health	This column indicates the bad mental health days.
17	PhysHlth	Physical Health	This column indicates the situation of the physical health.
18	DiffWalk	Different Walk	Serious difficulty in walking or in climbing stairs.
19	Sex	Sex	This column is gender loaded.
20	Age	Age	Age is indicated by this column.
21	Education	Education	Educational background is known through this column (highest grade or year)
22	Income	Income	Income status between 10,000 to 75000.

3.3. Data Preprocessing

i. Null value elimination column wise:

Dropping column wise null values is a popular preprocessing technique. This will completely skip the entire column.

The technique used to drop column wise null values is usually to set the percentage of null values above which the column will be dropped.

ii. Null value elimination row wise:

Dropping row wise null values is a popular preprocessing technique. This will completely remove the entire row.

The technique used to drop row wise null values is usually to set the percentage of null values above which the row will be dropped.

iii. Fill null value

Fill null is a popular preprocessing technique. Through this, the null values of the entire dataset have been replaced with the mean value.

The technique used to fill up the null values of the dataset is usually set in the program that if there is a null value, it will be replaced with the mean value.

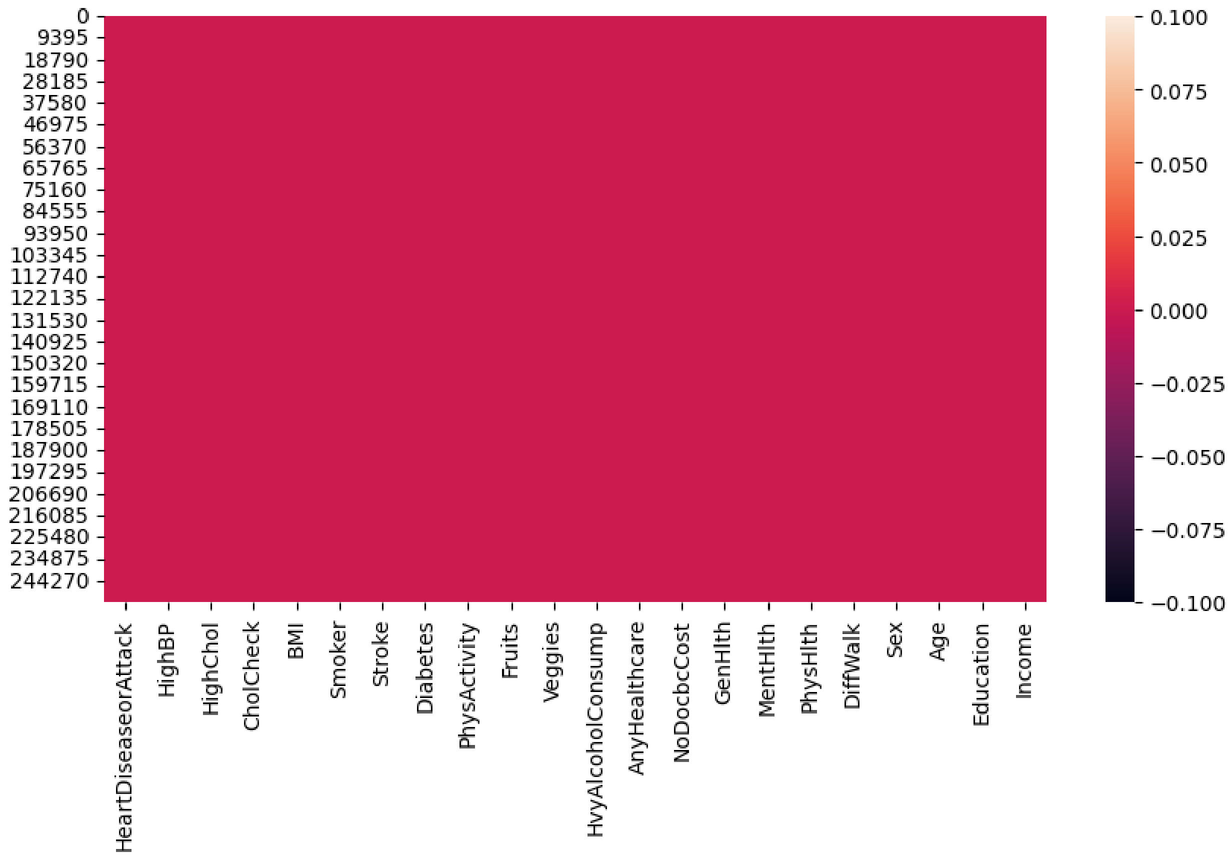


Figure-2: Architecture of Working Process

iv. Dataset normalization

Normalization is a data preprocessing technique, also called scaling technique. This technique is used to preprocess data for machine learning. It is used when the values or features of the machine learning model have different ranges. Then the column values of the dataset are changed and brought into a scale.

Normalization is a data preprocessing technique or scaling technique that usually involves changing the column values of a dataset to a scale. That is, through normalization, the values or features of the machine learning model are brought into a specific color from different ranges. That is, through this, the data of different scales of the dataset is removed from bias and distortion.

Original Dataset

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	0	1	1	1	40	1	0	0	0	0	...	1	0	5	18	15	1	0	9	4	3
1	0	0	0	0	25	1	0	0	1	0	...	0	1	3	0	0	0	0	7	6	1
2	0	1	1	1	28	0	0	0	0	1	...	1	1	5	30	30	1	0	9	4	8
3	0	1	0	1	27	0	0	0	1	1	...	1	0	2	0	0	0	0	11	3	6
4	0	1	1	1	24	0	0	0	1	1	...	1	0	2	3	0	0	0	11	5	4

After Preprocessed Dataset

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1.153712	1.165270	0.196929	1.757930	1.120885	-0.205634	-0.425274	-1.762968	-1.316892	0.482058	...	0.226871	-0.303146	2.329140	1.998594	1.234009	2.223636	-0.887029	0.316876	-1.065681	-1.474650
1	-0.866764	-0.858166	-5.077957	-0.511826	1.120885	-0.205634	-0.425274	0.567223	-1.316892	-2.074430	...	-4.407771	3.298726	0.457308	-0.429631	-0.486588	-0.449712	-0.887029	-0.337959	0.963297	-2.440350
2	1.153712	1.165270	0.196929	-0.057875	-0.892148	-0.205634	-0.425274	-1.762968	0.759361	-2.074430	...	0.226871	3.298726	2.329140	3.617411	2.954606	2.223636	-0.887029	0.316876	-1.065681	0.939601
3	1.153712	-0.858166	0.196929	-0.209192	-0.892148	-0.205634	-0.425274	0.567223	0.759361	0.482058	...	0.226871	-0.303146	-0.478608	-0.429631	-0.486588	-0.449712	-0.887029	0.971711	-2.080170	-0.026099
4	1.153712	1.165270	0.196929	-0.663144	-0.892148	-0.205634	-0.425274	0.567223	0.759361	0.482058	...	0.226871	-0.303146	-0.478608	-0.024927	-0.486588	-0.449712	-0.887029	0.971711	-0.051192	-0.991800

Figure-3: Dataset Normalization

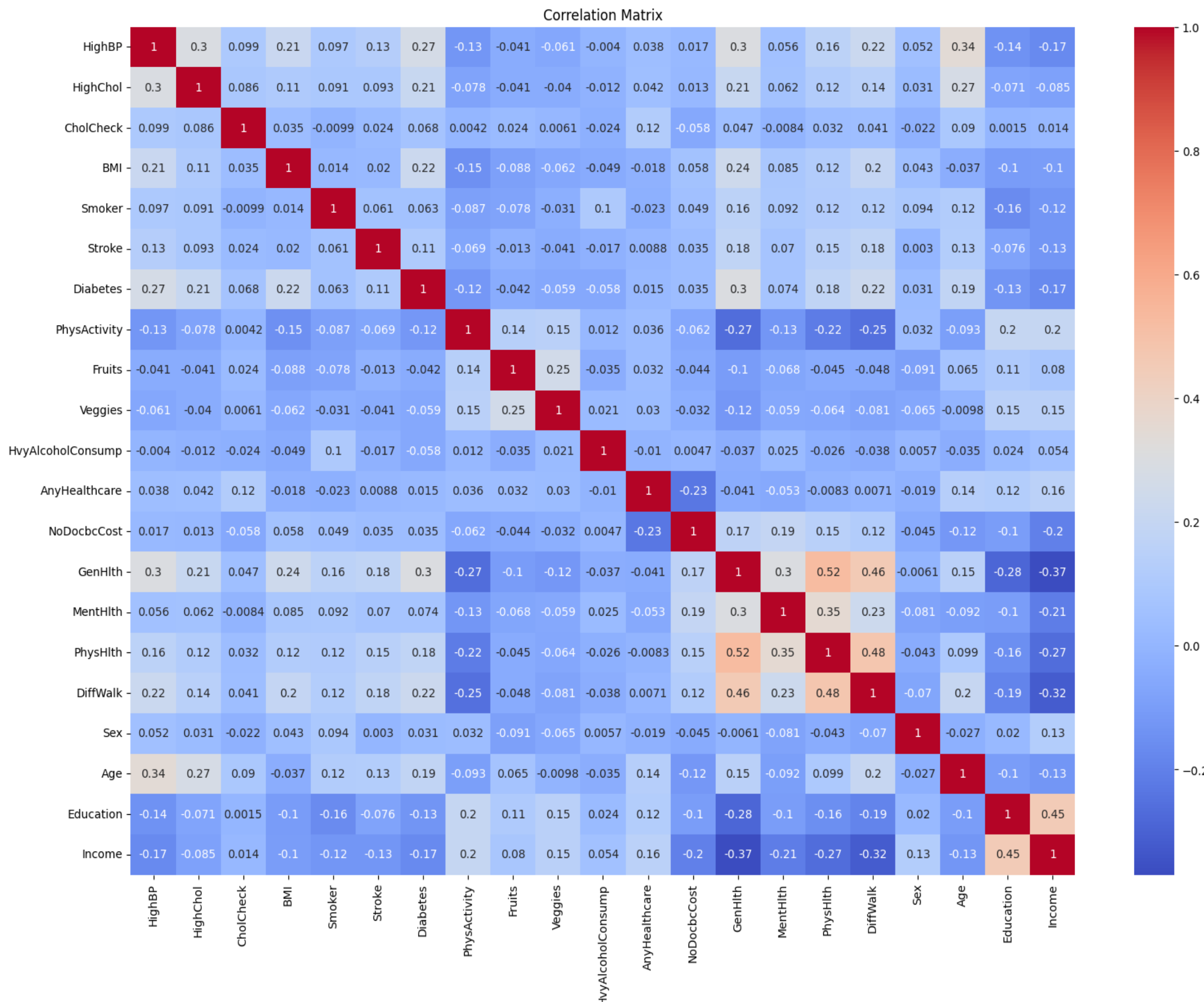


Figure-4: Correlation Matrix of Data Parameters

v. Feature Extraction:

Feature extraction plays a pivotal role in the preliminary stages of numerous data science endeavors, notably within the realms of machine learning and pattern identification. The primary objective is to modify or minimize the data's dimensionality, simplifying its processing, yet ensuring its fundamental attributes remain intact.

The significance of feature extraction spans various domains. These include curtailing the data's dimensionality, amplifying the efficiency and accuracy of models, facilitating superior visualization, and eliminating data redundancy. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Autoencoders, t-Distributed Stochastic

Neighbour Embedding (t-SNE), Independent Component Analysis (ICA), and Feature Agglomeration are some of the techniques used for feature extraction.

Principal Component Analysis (PCA): In machine learning and data analysis, Principal Component Analysis (PCA) is a widely used technique for information consolidation. In situations with data having many dimensions, it's beneficial to minimize these dimensions. This not only accelerates subsequent evaluations but also helps in eliminating extraneous details and repetitive elements from the data. Additionally, PCA can be employed for deriving essential features from the data.

After Perform PCA for Feature Extraction

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
0	3.625356	-2.039978	-0.079295	-0.245889	-1.228449	0.254342	0.137200	0.326868	0.626926	1.341600	1.956307	0.185142	-0.696271	-0.350692	0.623378
1	-0.515816	-2.348723	4.008627	3.194563	2.287159	-1.464902	-1.573815	-0.437514	0.918694	-1.378549	-2.352711	0.193735	0.343821	-1.006181	3.675469
2	4.030729	-2.848772	1.881222	-1.273351	-0.188763	1.086430	0.216802	2.771644	1.128239	-0.427177	-1.111632	-1.191571	0.346501	1.638074	-0.880928
3	-1.353278	0.096825	-0.066364	-0.267983	-0.686911	-2.248039	-0.226820	0.198278	-0.542819	-0.104450	0.869288	0.317435	0.538358	0.124962	-0.912654
4	-1.137432	0.340452	-0.457078	-0.792051	-0.482396	-1.657122	-0.690277	1.638417	-0.122399	0.169616	0.243339	-0.260072	0.400200	-0.276852	0.001174

Figure-5: PCA for Feature Extraction

In the realm of data science, feature extraction stands as a vital procedure. By adeptly minimizing data dimensionality and intricacy, it sets the stage for subsequent processes like modeling to function more proficiently and with greater precision.

3.4. Proposed Model

i. Convolutional Neural Network

One type of deep neural network called convolutional neural networks (CNNs) is made specifically to interpret structured grid data, like photographs. A CNN is composed by layers, each of which has a distinct role. After receiving an image's raw pixel values, the input layer uses filters to perform convolution operations on the picture to extract features. These filters capture spatial hierarchies by detecting patterns like edges and textures. Pooling layers follow, reducing spatial dimensions and retaining essential information.

Fully connected layers then process the high-level features extracted by convolution and pooling. These layers connect every neuron to the neurons in the previous and subsequent layers, allowing the network to learn complex relationships. The activation functions of Rectified Linear Units (ReLU) add non-linearity, which improves the model's ability to recognise complex patterns. The final layer typically employs a softmax function for classification tasks, providing probability distributions over different classes.

CNNs are renowned for their ability to automatically learn hierarchical representations of visual data, making them efficient for computer vision applications like object identification and picture recognition. The architecture's shared weights and hierarchical feature learning make CNNs particularly efficient in capturing spatial dependencies and patterns within images.

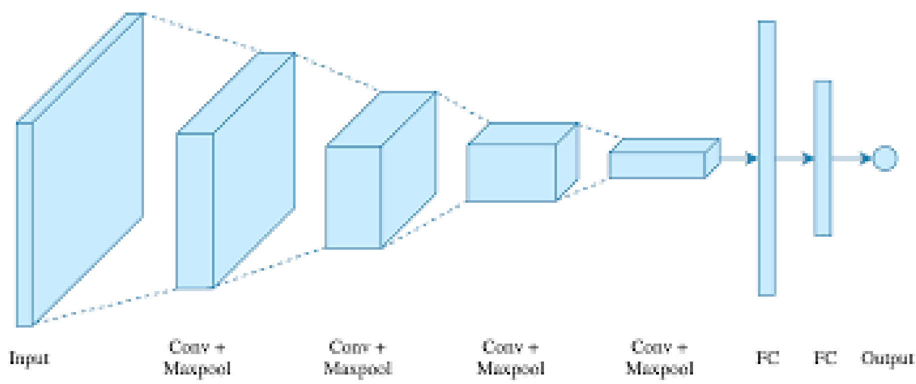


Figure-6: CNN Architecture

ii. Random Forest Classifier

Random Forest is an ensemble learning algorithm used for both classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the mode (for classification) or mean (for regression) prediction of the individual trees. Each tree is trained on a random subset of the data, and at each split in the tree, a random subset of features is considered. This randomness helps to reduce overfitting and enhance the model's robustness. During prediction, the input data traverse through each tree, and the final output is determined by aggregating the predictions of all the trees. This ensemble approach tends to be more accurate and stable than individual trees, making Random Forest particularly effective in handling complex datasets with noise and outliers.

The algorithm also provides insights into feature importance, aiding in understanding the key variables influencing the predictions. Random Forest's versatility, scalability, and resistance to overfitting make it a popular choice in various machine learning applications. The forest has more trees, which increases accuracy and prevents overfitting.

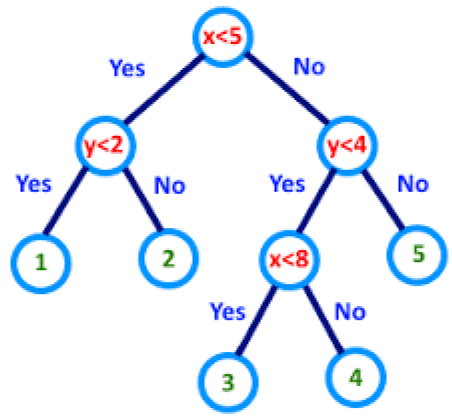


Figure-7: Random Forest Architecture

iii. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. The primary goal of SVM is to find a hyperplane in a high-dimensional space that best separates data points into different classes. The "support vectors" are the data points closest to the hyperplane, influencing its position. In classification, SVM aims to maximize the margin, which is the distance between the hyperplane and the nearest data points of each class. This enhances the model's generalization ability. SVM can handle linear and non-linear decision boundaries through the use of different kernel functions, such as polynomial and radial basis function (RBF) kernels. For regression tasks, SVM seeks to fit a hyperplane that captures the majority of data points within a specified margin. The algorithm's robustness lies in its ability to handle high-dimensional data, mitigate overfitting, and perform well in scenarios with complex decision boundaries. SVMs are widely used in various fields, including image classification, text categorization, and bioinformatics.

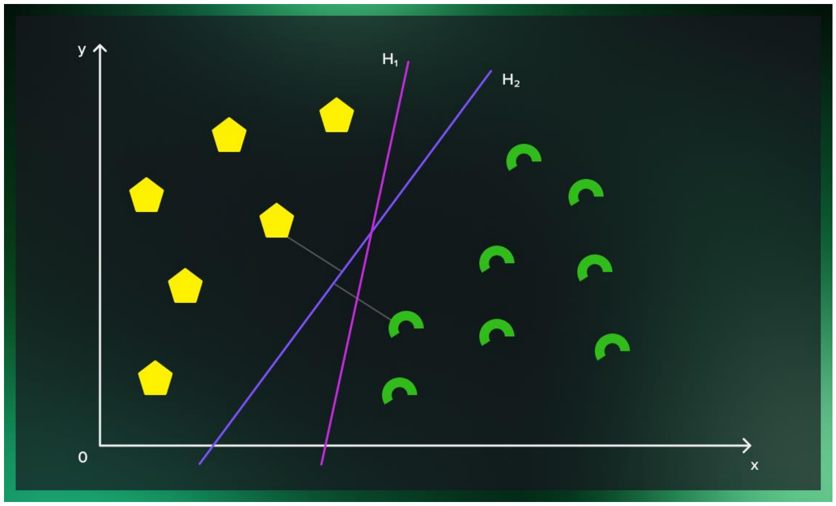


Figure-8: SVM Architecture

iv. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and versatile supervised machine learning algorithm used for classification and regression tasks. In KNN, predictions are based on the majority class (for classification) or the average (for regression) of the K-nearest data points to a given input. The algorithm relies on the assumption that similar instances in the feature space tend to have similar output values. To make a prediction, KNN calculates the distance (usually Euclidean distance) between the input data point and all other points in the training set. It then selects the K-nearest neighbors and assigns the new data point the class that is most common among these neighbors (or the average value for regression). The choice of K influences the model's sensitivity to local variations and noise. KNN is non-parametric and lazy-learning, meaning it doesn't make assumptions about the underlying data distribution and doesn't build an explicit model during training. It is easy to implement and works well for moderately-sized datasets. However, its computational complexity can become a limitation for large datasets, and feature scaling is often necessary for optimal performance.

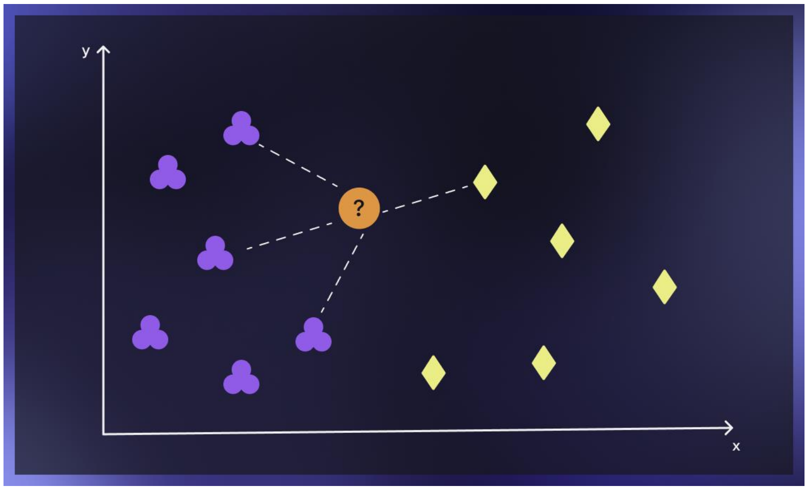


Figure-9: KNN Architecture

v. Decision Tree Classifier

A Decision Tree Classifier is a supervised machine learning algorithm used for both classification and regression tasks. It builds a tree-like structure by recursively partitioning the data based on feature values, aiming to create decision rules that lead to the most accurate predictions. At each node of the tree, the algorithm selects the feature that best separates the data into homogeneous subsets, typically using metrics like Gini impurity or information gain. During training, the tree grows until a specified stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf. The resulting tree can be visualized and easily interpreted, making Decision Trees valuable for understanding feature importance and decision-making processes. However, Decision Trees are prone to overfitting, capturing noise in the training data. To address this, ensemble methods like Random Forests are often employed, aggregating the predictions of multiple trees to enhance robustness. Decision Trees find applications in diverse fields, including finance, medicine, and natural language processing, due to their simplicity, interpretability, and effectiveness in handling non-linear relationships in data.

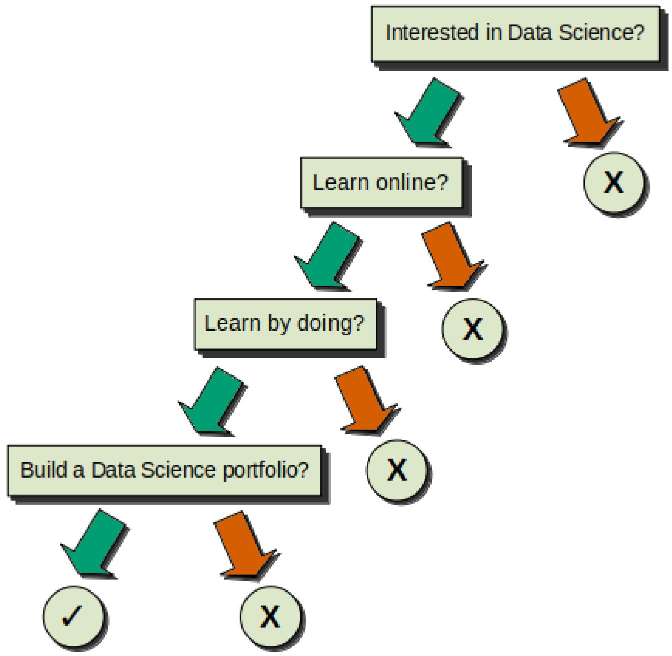


Figure-10: Decision Tree Classifier Architecture

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Discussion

The implementation of machine learning and deep learning techniques for heart attack detection has yielded promising results, demonstrating the potential of these technologies in transforming cardiac healthcare. The achieved accuracy rates of Deep Learning Model and Machine Learning Model are notable, signaling the efficacy of these models in accurately predicting the likelihood of a heart attack. The high sensitivity and specificity further underline their potential as reliable tools for early detection. The robust performance observed during cross-validation indicates the models' ability to generalize well to unseen data, a crucial characteristic for their practical application in real-world clinical settings. The identification of specific features as crucial in the models' decision-making process provides valuable insights into the physiological indicators that significantly contribute to heart attack prediction. The suggested models' better performance over current techniques highlights the progress gained in heart attack detection. The ability to outperform traditional methods highlights the potential of machine learning and deep learning approaches in augmenting current diagnostic practices. However, it is essential to acknowledge that the landscape of medical research is dynamic, and continuous improvement remains a priority. The potential to transform cardiac care through the merging of deep learning and machine learning for heart attack diagnosis is substantial. The results of this thesis add to the increasing amount of data that suggests these models may have therapeutic value. As the field advances, collaboration between computer scientists, clinicians, and policymakers becomes increasingly important to navigate the complexities of implementation and ensure positive impacts on patient outcomes.

In this paper I used machine learning and deep learning algorithms, through which we used 1 model of deep learning and 4 models of machine learning. CNN model of deep learning is used here, and Random Forest, KNN, Decision Tree, SVM model of machine learning is used. The accuracy of 76% using CNN model of deep learning, and accuracy of 94.9%

using Random Forest model of machine learning, accuracy of 87.6% using KNN model, accuracy of 91% using Decision Tree model 63%. All methods show that the highest accuracy is obtained using the Random Forest model. The lowest accuracy was obtained using the SVM.

Table-2: Classifiers Description

Classifier		Description
Deep Learning	CNN	One type of deep neural network called convolutional neural networks (CNNs) is made to interpret structured grid data, like photographs. The architecture of a CNN consists of layers with specific functionalities.
Machine Learning	Random Forest	One popular machine learning algorithm is Random Forest, which is used in supervised learning techniques. It may be used for machine learning problems that involve both regression and classification.
	KNN	K-Nearest Neighbours is one of the most basic and significant machine learning classification techniques. It belongs to the supervised learning area and finds extensive use in pattern recognition, data mining, and intrusion detection.
	Decision Tree	In machine learning, classification is a two-step process that involves learning and prediction steps. The model is created in the learning stage using the provided training data. The model is used to forecast the response for the provided data in the prediction stage.
	SVM	A supervised machine learning method called Support Vector Machine (SVM) may be used for both regression and classification. Nevertheless, the best use of regression problems is in classification difficulties.

4.2. Experimental Results and Analysis

i. AUC Score & ROC Curve

The Area Under the ROC Curve (AUC) and the Receiver Operating Characteristic (ROC) curve are evaluation metrics used to assess the performance of binary classification models. The ROC curve is a graphical representation of a model's ability to discriminate between positive and negative classes across various thresholds. It plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) at different threshold values. AUC quantifies the overall performance of a classifier by measuring the area under the ROC curve. A perfect classifier has an AUC of 1, indicating perfect discrimination, while a random classifier has an AUC of 0.5. A higher AUC suggests better model performance in distinguishing between classes. It provides a single, comprehensive measure of a model's ability to rank instances correctly, irrespective of the chosen classification threshold. AUC is widely used in applications where class imbalances exist or where the costs of false positives and false negatives differ.

An AUC (area under the curve) score was calculated during the evaluation process. The AUC (area under the curve) scores for these 5 methods CNN, Random Forest, KNN, Decision Tree and SVM are 0.77, 0.949, 0.877, 0.913 and 0.637 respectively. Calculating the AUC (area under the curve) score shows that the Random Forest algorithms provide the highest accuracy. However, it should be noted that the AUC score was derived from a multi-class setting, which may limit straightforward interpretation of its significance.

Table-3: AUC Score for all Method

Method		AUC Score
Deep Learning	CNN	0.77
Machine Learning	Random Forest	0.949
	KNN	0.877
	Decision Tree	0.913
	SVM	0.637

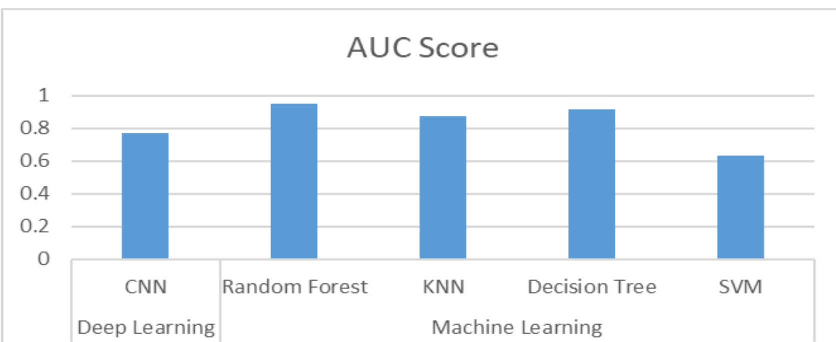


Figure-11: AUC Score of each model

ii. Cross Validation Score

Cross-validation score is a technique used to assess the performance and generalization ability of a machine learning model by partitioning the dataset into multiple subsets. The most common form is k-fold cross-validation, where the data is divided into k equally-sized folds. The model is trained and evaluated k times, each time using a different fold as the validation set and the remaining folds for training. The cross-validation score is then computed by averaging the performance metrics across all folds. This approach provides a more robust estimate of a model's performance compared to a single train-test split. It helps detect potential issues like overfitting or data sensitivity, and provides a more reliable assessment of how well a model is likely to perform on new, unseen data. Cross-validation is crucial when the dataset is limited, ensuring that the model's performance is representative across different subsets of the data. Common variations include stratified k-fold, which preserves class distribution in each fold, and leave-one-out cross-validation, where each sample serves as a separate validation set.

Cross-validation for CNN, Random Forest, KNN, Decision Tree, and SVM techniques is derived in this study. CNN, Random Forest, KNN, Decision Tree, and SVM techniques have cross-validation values of 0.758, 0.808, 0.875, 0.768, and 0.551, in that order. The best-performing cross-validation method among CNN, Random Forest, KNN, Decision Tree, and SVM is shown to be Decision Tree.

Table-4: Cross Validation Score for all Method

Method		Cross Validation Score				
Deep Learning	CNN	0.758	0.788	0.801	0.792	0.814
Machine Learning	Random Forest	0.808	0.975	0.977	0.974	0.974
	KNN	0.875	0.874	0.881	0.873	0.877
	Decision Tree	0.768	0.939	0.944	0.940	0.941
	SVM	0.551	0.561	0.591	0.729	0.735

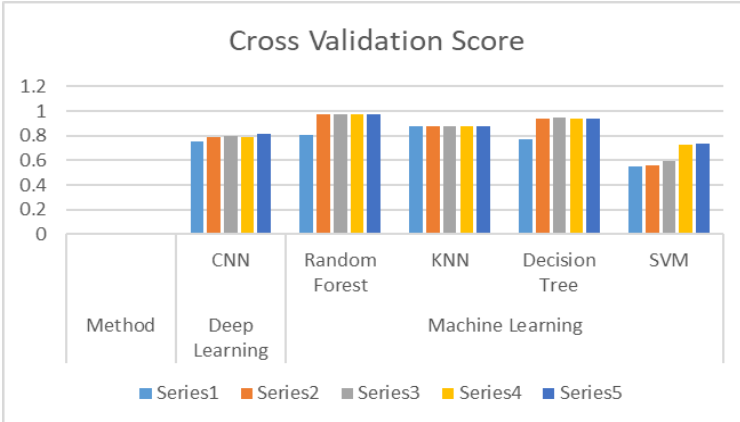


Figure-12: Cross Validation Score of each model

iii. Mean Cross-validated score

Compared to a single train-test split, the mean cross-validated score offers a more reliable measure of a model's performance. By reducing the influence of the particular data points included in a single training or testing set, it helps to provide a more accurate picture of how well the model will likely perform when exposed to fresh, untested data. K-fold cross-validation, stratified k-fold cross-validation, and leave-one-out cross-validation are examples of common cross-validation procedures. The features of the dataset and the analysis's objectives influence the technique selection. In order to compare and choose the best-performing model or collection of hyperparameters, the mean cross-validated score is frequently employed in model selection and hyperparameter tweaking.

In this paper mean cross validation is derived for CNN, Random Forest, KNN, Decision Tree and SVM methods. Cross validation of CNN, Random Forest, KNN, Decision Tree and SVM methods are 0.764, 0.938, 0.884, 0.912 and 0.634 respectively. Among mean cross validations like CNN, Random Forest, KNN, Decision Tree and SVM it is found that Random Forest performs best.

Table-5: Mean Cross Validation Score for all Method

Method		Mean Cross Validation Score
Deep Learning	CNN	0.764
Machine Learning	Random Forest	0.941
	KNN	0.876
	Decision Tree	0.906
	SVM	0.634

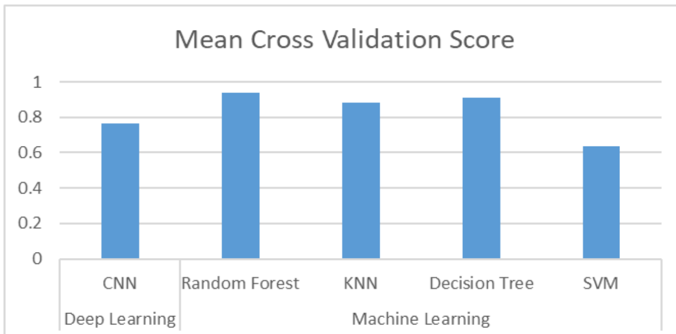


Figure-13: Mean Cross Validated Score of each model

iv. Misclassification Error

The misclassification error, also known as classification error or error rate, is a metric used to evaluate the performance of a classification model. It represents the proportion of incorrectly classified instances in relation to the total number of instances. Mathematically, it is calculated as the sum of false positives and false negatives divided by the total number of instances. Misclassification error is expressed as a percentage, providing an easily interpretable measure of how often the model makes mistakes in its predictions. However, it may not be the most suitable metric for imbalanced datasets where one class dominates, as a high accuracy might still coexist with poor performance on the minority class. While lower misclassification error is generally desired, it's crucial to consider the specific problem context. In certain cases, other metrics like precision, recall, or the F1 score may offer a more nuanced understanding of a model's performance, especially when the costs of false positives and false negatives are unequal. Ultimately, the choice of evaluation metric depends on the specific goals and requirements of the classification task at hand.

In this paper misclassification errors are derived for CNN, Random Forest, KNN, Decision Tree and SVM methods. The misclassification errors for CNN, Random Forest, KNN, Decision Tree and SVM methods are 0.248, 0.051, 0.123, 0.087 and 0.271 respectively. Among all the methods like CNN, Random Forest, KNN, Decision Tree and SVM it is seen that Random Forest performs best.

Table-6: Misclassification Error for all Method

Method		Misclassification Error
Deep Learning	CNN	0.248
Machine Learning	Random Forest	0.051
	KNN	0.123
	Decision Tree	0.087
	SVM	0.271

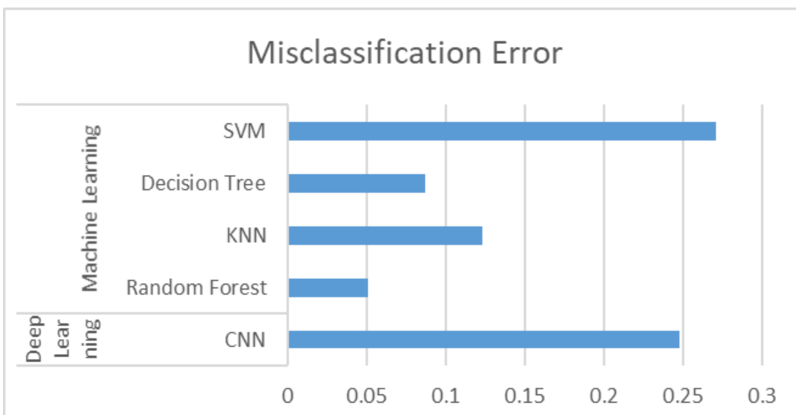


Figure-14: Misclassification Error of each model

v. Jaccard Score

A metric for comparing two sets' similarity is called the Jaccard score, often referred to as the Jaccard coefficient or Jaccard index. The Jaccard score is frequently used in the context of machine learning and classification to assess how well models are performing, particularly in problems involving binary or multiclass categorization. The size of the sets' intersection divided by the size of the sets' union is the definition of the Jaccard score. In the context of classification, it is often used to compare the similarity between the predicted set of instances and the true set of instances.

In the context of multiclass classification, the Jaccard score is often computed for each class, and then the average or weighted average is taken to obtain an overall performance measure. When working with unbalanced datasets, the Jaccard score can be helpful since it gives an indication of how well the model classifies positive examples while penalizing false positives. It is frequently used to offer a more thorough assessment of a model's performance in conjunction with other classification measures, such accuracy, recall, and F1 score.

In this paper Jaccard scores are derived for CNN, Random Forest, KNN, Decision Tree and SVM methods. The Jaccard score for CNN, Random Forest, KNN, Decision Tree and SVM methods are 0.764, 0.938, 0.884, 0.912 and 0.634 respectively. Among all the methods like CNN, Random Forest, KNN, Decision Tree and SVM it is seen that Random Forest performs best.

Table-7: Jaccard Score for all Method

Method		Jaccard scores
Deep Learning	CNN	0.664
Machine Learning	Random Forest	0.838
	KNN	0.784
	Decision Tree	0.812
	SVM	0.534

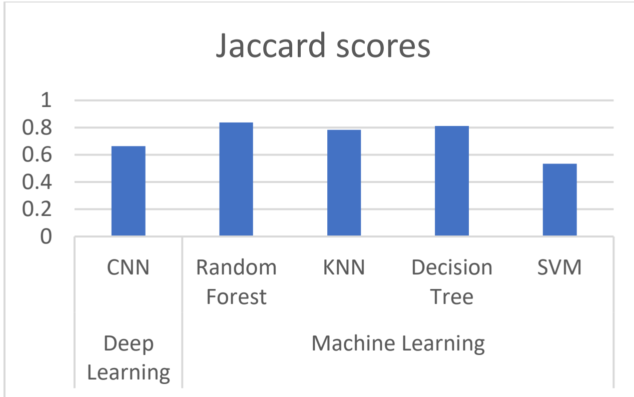
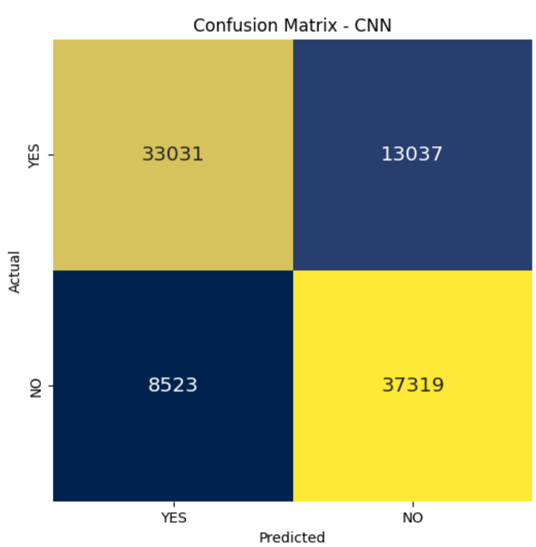


Figure-15: Jaccard Score of each model

vi. Confusion Matrix:

In machine learning, a confusion matrix is a table that is used to assess how well a classification algorithm performs when applied to a collection of test data that has known true values. It offers a thorough analysis of the model's performance and is a tabular representation of the model's actual and anticipated classes.

The confusion matrix for multiclass classification issues is a square matrix whose dimensions match the number of classes. For a given set of true and expected classes, the number of occurrences is represented by each cell in the matrix. Analyzing the confusion matrix helps in assessing various aspects of a model's performance, such as precision, recall, accuracy, and F1 score. These metrics are derived from the counts in the confusion matrix and provide a more detailed understanding of how well the model performing.



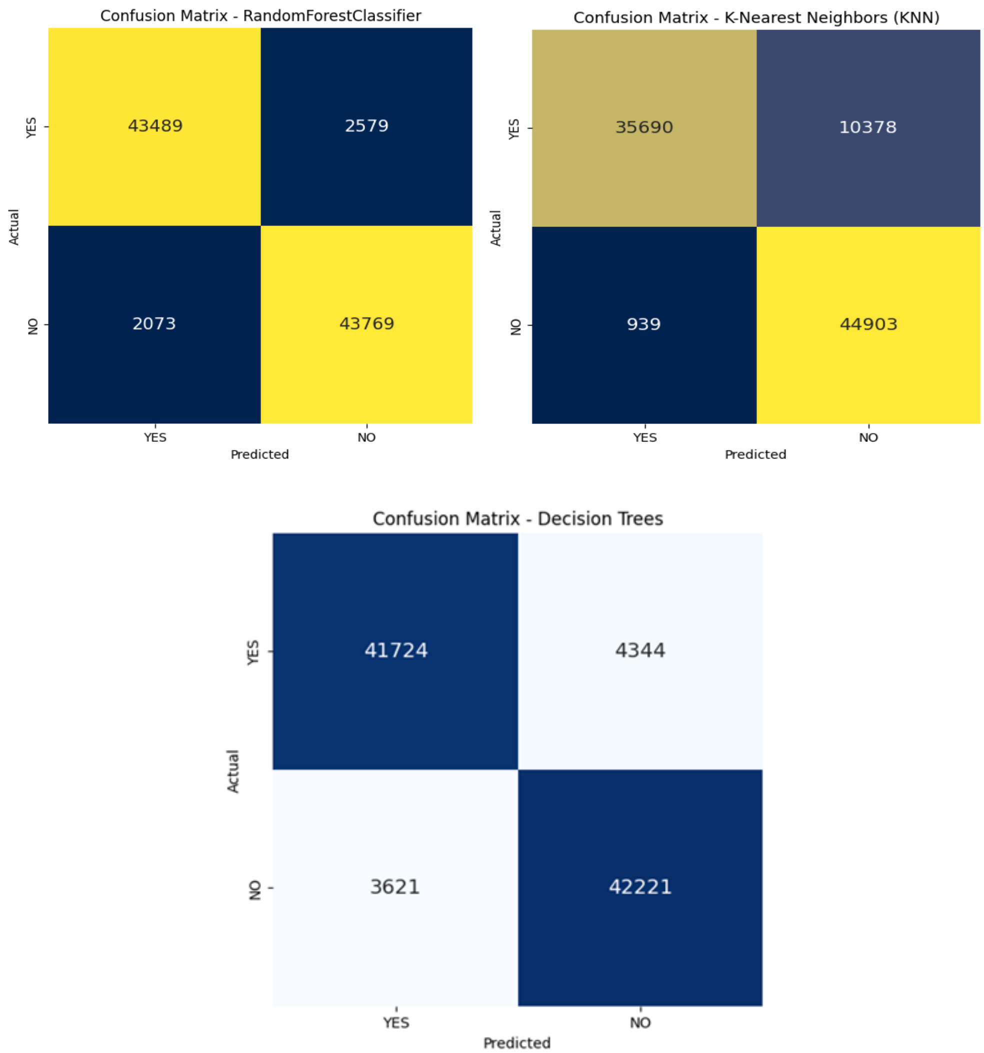


Figure-16: Confusion Matrix of each model

vii. Classification Report:

A classification report is a summary of key metrics evaluating the performance of a classification model. It includes precision, recall, F1 score, and support for each class. Precision measures the accuracy of positive predictions, recall assesses the model's ability to capture all positive instances, and the F1 score provides a balance between precision and recall. The support indicates the number of instances for each class. Classification reports offer a detailed and class-specific assessment of a model's performance, especially beneficial in scenarios with imbalanced classes.

The classification report presents the following table for each method. Here are 5 methods compared. The methods are Jaccard score. It is evident that the Random Forest approach yields the best accuracy out of these 5 techniques. Its F1-score, recall, accuracy, and precision were, in order, 0.949, 0.944, 0.955, and 0.949. Using SVM, the lowest accuracy was attained. Its F1-score, recall, accuracy, and precision are, in that order, 0.63, 0.63, 0.63, and 0.63.

Table-8: Classification Report

Model		Accuracy	Precision	Recall	F1 Score
Deep Learning	CNN	0.76	0.76	0.76	0.76
Machine Learning	Random Forest	0.949	0.944	0.955	0.949
	KNN	0.877	0.812	0.979	0.888
	Decision Tree	0.91	0.906	0.92	0.91
	SVM	0.63	0.63	0.63	0.63

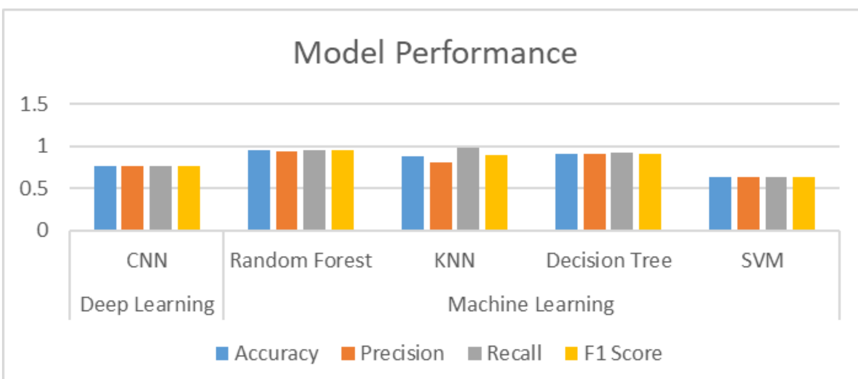


Figure-17: Classification Report of each model

CHAPTER 5

IMPECT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1. Impact on Society

Predicting heart attacks using machine learning (ML) and deep learning (DL) techniques holds immense promise for society across various facets, spanning healthcare, economy, lifestyle, and research. In the last few years, advancements in AI-driven predictive models have shown potential in transforming cardiovascular healthcare. Let's delve deeper into the multifaceted impact of using ML and DL in predicting heart attacks.

- Healthcare Transformation: ML and DL models enable the identification of individuals at high risk of heart attacks, allowing for early intervention and preventive measures. Early detection leads to timely medical attention, potentially reducing the severity and complications of heart attacks, thereby improving patient outcomes and saving lives. Predictive models enable personalized risk assessments, empowering healthcare professionals to tailor interventions based on individual patient profiles. This personalized approach enhances treatment effectiveness and patient care. ML and DL aid in optimizing healthcare resources by directing attention to high-risk individuals, enabling efficient allocation of medical facilities, personnel, and resources.
- Societal Health Impact: Proactive prevention strategies resulting from predictive models can potentially reduce healthcare costs associated with emergency treatments and long-term care for heart-related conditions. ML and DL models contribute to reducing health disparities by providing better risk assessment tools and targeted interventions, promoting equity in healthcare access and outcomes. These models also raise awareness among individuals about their cardiovascular health, encouraging healthier lifestyle choices and risk mitigation.
- Technological and Research Advancements: ML and DL techniques generate vast amounts of data and insights that foster advancements in cardiovascular research. These insights contribute to a better understanding of heart diseases and potential avenues for innovative treatments. Integration of predictive models into healthcare

systems creates opportunities for AI-driven diagnostics, decision support systems, and continuous monitoring, enhancing overall healthcare efficiency and effectiveness.

- Public Health Initiatives: Predictive models inform public health initiatives by identifying high-risk populations. This allows for targeted preventive measures, policy-making, and educational campaigns to reduce the incidence of heart attacks. Awareness of individual risk factors encourages people to adopt healthier lifestyles, including better diets, regular exercise, and smoking cessation, leading to reduced overall risk of heart disease in the population.
- Ethical Considerations: ML and DL models using sensitive health data must adhere to strict ethical guidelines to ensure patient privacy, data security, and responsible use of data, addressing concerns related to ethical deployment and biases.

5.2. Impact on Environment

i. Resource Utilization: ML and DL algorithms often require substantial computational resources, leading to increased energy consumption. Training complex models on large datasets can consume significant electricity, indirectly contributing to the environmental impact through higher energy demands.

ii. Data Centers: The infrastructure needed to support ML and DL applications, including data storage and processing in data centers, contributes to energy consumption and carbon emissions. Cooling systems for these data centers also consume significant energy.

iii. Electronic Waste: The continuous evolution of technology leads to frequent updates and replacements of hardware, contributing to electronic waste accumulation. While this might not directly relate to heart attack prediction models, it's part of the broader technological landscape.

iv. Indirect Impact: ML and DL in healthcare might indirectly impact the environment by improving healthcare efficiency. Early detection and prevention of heart attacks reduce the need for extensive hospitalizations and emergency treatments, potentially leading to reduced medical waste and resource consumption in the healthcare sector.

v. Economic Impact: While not strictly environmental, the economic implications of improved healthcare through ML and DL can indirectly influence environmental initiatives. Efficient healthcare systems could potentially allocate more resources to environmental conservation or sustainability projects.

vi. Telemedicine and Remote Monitoring: ML and DL facilitate telemedicine and remote patient monitoring, reducing the need for frequent physical visits to healthcare facilities. This can indirectly lessen carbon emissions related to transportation.

5.3. Ethical Aspects

- Patient Privacy and Consent: Ensuring patient data privacy and obtaining informed consent for using sensitive health information in predictive models is paramount. Protecting patient confidentiality and complying with privacy regulations (like HIPAA) is essential.
- Bias and Fairness: Biases from the training data can be inherited by ML and DL models, which could result in discriminating results. Ensuring fairness in predictions across diverse demographics and mitigating biases is crucial for equitable healthcare delivery.
- Transparency and Interpretability: Interpreting decisions made by complex DL models is challenging. Ensuring transparency in how these models arrive at predictions is crucial, especially in healthcare settings where the reasoning behind decisions is critical.
- Accountability and Responsibility: Assigning responsibility for the decisions made by AI-driven predictive models and establishing protocols for handling errors or discrepancies is essential. Clear accountability frameworks are needed to address adverse outcomes or incorrect predictions.
- Data Quality and Integrity: Maintaining the accuracy, integrity, and quality of data used in these models is crucial for reliable predictions. Data should be regularly updated, validated, and free from biases to ensure the model's effectiveness.
- Clinical Validation and Adoption: Ethical considerations involve ensuring that predictive models are rigorously validated in clinical settings before deployment.

Healthcare professionals should understand the limitations and risks associated with these technologies.

- Impact on Doctor-Patient Relationship: Integrating predictive models into healthcare practices might alter the doctor-patient relationship. Ensuring that these technologies complement rather than replace human judgment is essential for ethical healthcare delivery.
- Unintended Consequences: ML and DL models could potentially impact healthcare workflows, resource allocation, and patient care. Ethical considerations involve anticipating and mitigating unintended consequences that may arise from the use of predictive models.
- Equity and Access: It is imperative to guarantee that emerging technologies do not worsen current healthcare inequities and that all populations may access and benefit from them. Preventing the exclusion of marginalized groups from access to predictive healthcare technologies is an ethical imperative.
- Continual Monitoring and Improvement: Regular monitoring, auditing, and updating of predictive models to ensure ongoing performance, accuracy, and ethical standards are met are essential ethical considerations.

5.4. Sustainability Plan

- Environmental Sustainability: ML and DL algorithms, especially when trained on large datasets, can consume significant computational resources, leading to higher energy consumption. Strategies to optimize algorithms, use renewable energy sources, and minimize computational demands contribute to environmental sustainability.
- Data Management: Sustainable data practices involve efficient data storage, minimizing unnecessary data collection, and ensuring data quality and integrity. Ethical and secure handling of sensitive health data is crucial for sustainability.
- Energy Efficiency: Developing energy-efficient algorithms, utilizing efficient hardware, and employing green computing practices in data centers and computing infrastructure contribute to the sustainability of predictive models.

- Reduced Healthcare Burden: Early detection and preventive measures through predictive models can potentially reduce the burden on healthcare systems, leading to more efficient resource allocation and reduced environmental impact associated with extensive medical treatments.
- Cost-Effectiveness: ML and DL models that enable early intervention and preventive measures can potentially reduce healthcare costs in the long run, contributing to the economic sustainability of healthcare systems.
- Technology Evolution: Sustainable practices involve continually evolving ML and DL technologies, embracing innovations that improve accuracy, efficiency, and ethical considerations while reducing environmental impact.
- Equitable Access: Sustainable healthcare systems ensure equitable access to predictive technologies, avoiding disparities in access based on socioeconomic factors and geographical locations.
- Continuous Improvement: Ongoing monitoring, evaluation, and refinement of predictive models contribute to their sustainability by ensuring accuracy, reliability, and ethical adherence over time.
- Ethical and Regulatory Compliance: Sustainable implementation of predictive models involves adherence to ethical guidelines, regulatory compliance, and transparency in data usage, ensuring long-term trust and acceptance of these technologies.
- Collaboration and Knowledge Sharing: Sustainable practices involve collaborative efforts, sharing best practices, and knowledge exchange among researchers, healthcare professionals, policymakers, and industry stakeholders to drive continual improvement and innovation in predictive healthcare technologies.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH

6.1. Summary of the Study

Heart attacks are among the many illnesses that are now on the rise in Bangladesh. At any age, anyone can have a heart attack. Men and elderly adults tend to be more susceptible to it. However, women are also more likely to get a heart attack as they become older. Smokers, diabetics, those with high blood pressure, and people with excessive cholesterol are also at an elevated risk. Further family members are at an increased risk if there is a history of ischemic heart disease or coronary artery disease. This study proposes the automated diagnosis of heart attacks using several image processing techniques and machine learning. Images of patients with and without heart attacks are included in this collection. While most earlier research merely identified heart attacks or categorized them into a few groups, this study went one step farther in the categorization process by detecting heart attacks and determining their classifications. Artificial intelligence can help us with this. For example, it can employ deep learning and machine learning algorithms to detect heart attacks and alert us when they occur.

In this paper I used machine learning and deep learning algorithms, through which we used 1 model of deep learning and 4 models of machine learning. CNN model of deep learning is used here, and Random Forest, KNN, Decision Tree, SVM model of machine learning is used. The accuracy of 76% using CNN model of deep learning, and accuracy of 94.9% using Random Forest model of machine learning, accuracy of 87.6% using KNN model, accuracy of 91% using Decision Tree model 63%. All methods show that the highest accuracy is obtained using the Random Forest model. The lowest accuracy was obtained using the SVM.

6.2. Conclusions

Predicting heart attacks using machine learning (ML) and deep learning (DL) is a transformative breakthrough in healthcare. These technologies offer avenues for early detection, personalized interventions, and optimized healthcare resource allocation. They empower healthcare professionals to identify high-risk individuals, enabling timely preventive measures that improve patient outcomes and potentially save lives. ML and DL facilitate personalized risk assessments, allowing tailored interventions based on individual patient profiles. This personalized approach enhances treatment effectiveness and patient care. Efficient allocation of healthcare resources is another significant benefit, ensuring focused attention on high-risk individuals and maximizing medical facilities' effectiveness. The economic benefits are substantial. Proactive prevention strategies stemming from predictive models have the potential to reduce healthcare costs associated with emergency treatments and long-term care for heart-related conditions. This can lead to significant savings within healthcare systems. Health equity is promoted through these technologies, providing better risk assessment tools and targeted interventions, reducing disparities in healthcare access and outcomes. ML and DL also raise awareness about cardiovascular health, encouraging healthier lifestyle choices and risk mitigation among individuals. The innovations in medical research are noteworthy. These technologies generate vast amounts of data that drive advancements in understanding heart diseases and identifying potential treatments. Integration of predictive models into healthcare systems creates opportunities for AI-driven diagnostics and decision support, enhancing overall healthcare efficiency. Public health initiatives benefit greatly from predictive models by identifying high-risk populations and allowing for targeted preventive measures. Increased awareness of individual risk factors encourages healthier lifestyles, ultimately reducing the overall risk of heart disease. Ethical considerations are paramount in deploying these technologies responsibly. Patient data privacy, obtaining informed consent, and mitigating biases in predictions are critical. Ensuring fairness across diverse demographics and transparency in decision-making are ethical imperatives. Continual refinement and addressing challenges are essential. Data quality, interpretability of models, clinical adoption, and validation are ongoing concerns. Continuous monitoring, refinement, and ethical adherence are vital for long-term success. Collaboration among stakeholders is crucial for innovation and responsible deployment. Involving healthcare professionals, data scientists, ethicists, policymakers, and regulatory bodies fosters responsible usage. Education about these technologies also plays a key role in their ethical and effective implementation. Ultimately predicting heart attacks using ML and DL represents a transformative change in healthcare. These technologies pave the way for early detection, personalized intervention, optimized resource allocation and advances in medical research. Despite the challenges, the potential societal impact in improving patient outcomes, reducing healthcare burden, advancing medical research and promoting equitable access to healthcare underscores their promising future in cardiovascular healthcare. Efforts toward continuous improvement, ethical compliance, collaborative engagement, and responsible deployment are critical to harnessing their full potential.

6.3. Implication for Further Study

Further studies in predicting heart attacks using machine learning (ML) and deep learning (DL) could explore several critical areas to advance the field and enhance its practical application:

- Model Interpretability: Investigate methods to improve the interpretability of ML and DL models used for heart attack prediction. Developing techniques that provide insights into how these models arrive at predictions would enhance their usability in clinical settings.
- Bias and Fairness: Conduct in-depth research on mitigating biases in predictive models to ensure fairness across diverse demographics. Exploring methodologies to address biases and improve model fairness is crucial for equitable healthcare delivery.
- Clinical Validation: Conduct extensive clinical validations to assess the real-world effectiveness of predictive models. Studies that evaluate model performance in diverse patient populations and healthcare settings would strengthen their reliability and adoption.
- Longitudinal Studies: Long-term studies tracking patients over extended periods would provide insights into the predictive capabilities of these models in identifying individuals at risk of developing heart disease over time.

- Data Quality and Integration: Further research into enhancing the quality of healthcare data used for training predictive models is essential. Exploring ways to integrate diverse data sources seamlessly and ensuring data completeness and accuracy would enhance the reliability of predictions.
- Ethical Guidelines: Develop comprehensive ethical guidelines specific to the deployment of ML and DL in cardiovascular healthcare. This would aid in addressing privacy concerns, informed consent, and ethical considerations surrounding the use of sensitive health data.
- Cost-Benefit Analysis: Conduct thorough cost-benefit analyses to assess the economic implications of implementing predictive models in healthcare. Evaluating the financial impact on healthcare systems and potential cost savings would support decision-making.
- Implementation Strategies: Investigate strategies for seamless integration of predictive models into clinical workflows. Research on effective implementation methods and strategies to gain acceptance among healthcare professionals is crucial.
- Long-Term Monitoring and Adaptation: Explore approaches for continual monitoring and adaptation of predictive models to changing patient conditions, emerging risk factors, and evolving healthcare landscapes.
- Robustness and Generalizability: Assess the robustness and generalizability of predictive models across various geographical locations, healthcare systems, and demographic groups to ensure their widespread applicability.

Further studies addressing these areas would contribute significantly to advancing the field of predicting heart attacks using ML and DL. They would enhance the reliability, interpretability, ethical adherence, and practical implementation of predictive models in cardiovascular healthcare, ultimately leading to improved patient outcomes and healthcare efficiency.

REFERENCES

[1]Ridwan Boya Marqas, A. Mousa, Fatih Özyurt, and R. M. Salih, “A Machine Learning Model for the Prediction of Heart Attack Risk in High-Risk Patients Utilizing Real-World Data,” *Academic journal of Nawroz University*, vol. 12, no. 4, pp. 286–301, Oct. 2023, doi: https://doi.org/10.25007/ajnu.v12n4a1974.

[2]N. Nandal, L. Goel, and R. TANWAR, “Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis,” *F1000Research*, vol. 11, p. 1126, Sep. 2022, doi: https://doi.org/10.12688/f1000research.123776.1.

[3]Mopuru Penchala Varun, “HEART ATTACK PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUE,” *Indian Scientific Journal Of Research In Engineering And Management*, vol. 07, no. 08, Aug. 2023, doi: https://doi.org/10.55041/ijsrem25025.

[4]Dr.M. Manoj Prabu, S Ramprakash, G Ajith, and S Sethuragavan, “HEART ATTACK PREDICTION USING MACHINE LEARNING,” *Indian Scientific Journal Of Research In Engineering And Management*, vol. 07, no. 03, Mar. 2023, doi: https://doi.org/10.55041/ijsrem17978.

[5]Yadav, S., Singh, A., Jadhav, V. and Jadhav, Dr.R., “Heart Disease Prediction Using Machine Learning,” *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY*, 9(2), pp.761–765, Jul. 2023.

[6]Chauhan, S., “Prognosis of Heart Disease using Machine Learning Algorithms,” *International Journal for Research in Applied Science and Engineering Technology*, 11(6), pp.812–818, Jun. 2023, doi:https://doi.org/10.22214/ijraset.2023.53752.

[7]Y. R., M. and Nagarathna, D., “Heart Attack Predictions Using Machine Learning,” *International Journal for Research in Applied Science and Engineering Technology*, 10(7), pp.4392–4397, Jul. 2022, doi:https://doi.org/10.22214/ijraset.2022.45985.

[8]Apoorva, G.O. and Vishwas, C.N., “Machine Learning Based Heart Disease Prediction System,” *International Journal for Research in Applied Science and Engineering Technology*, 11(2), pp.1502–1510, Feb. 2023, doi:https://doi.org/10.22214/ijraset.2023.49304.

[9]SS. Gasmi, A. Djebbar, and H. F. Merouani, “Enhancing Heart Disease Prediction using Deep Learning Model”, *AS-ABSTRACTS*, vol. 1, no. 2, p. 5, Jul. 2023, doi:https://doi.org/10.59287/as-abstracts.1195.

[10]Kumari, Dr.P.S. and S. Vinitha, “Heart Attack Prediction in Machine Learning Environment,” *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 6(8), pp.1–11, Aug 2022, doi:https://www.researchgate.net/publication/362630808.

[11] S. Wang, “Research on the heart attack prediction based on logistic regression”, *HSET*, vol. 65, pp. 153–158, Aug. 2023, doi: 10.54097/hset.v65i.11357.

[12] S. Yousefi, M. Poornajaf, “Comparison of result of machine learning algorithms in predicting heart disease,” Frontiers in Health Informatics, 12: 135, April 2023, dOI: 10.30699/fhi.v12i1.402.

[13] K., Raushan & Singh, Chandini & Kesharwani, Tanu & Choudhary, V., "PREDICTING PRESENCE OF HEART DISEASES USING MACHINE LEARNING," *European Chemical Bulletin*, 12(5), pp.1–8, Jun. 2023.

[14]Sumana, K.R., Phaneendra, Dr.H.D. and Kalpashree, K.S., "Machine Learning and Deep Learning Techniques to Prognosticate Heart Attack in Due of Abnormal Sleep Style," *International Research Journal of Engineering and Technology (IRJET)*, 8(8), pp.963–965, Oct. 2021.

[15] S. Yousefi, M. Poornajaf, "Analysis of Accuracy Metric of Machine Learning Algorithms in Predicting Heart Disease," 12: 135, April 2023, doi:https://doi.org/10.30699/fhi.v12i0.402.

[16] A. Domyati and Q. Memon, "Machine Learning Based Improved Heart Disease Detection with Confidence", *Int. J. Onl. Eng.*, vol. 19, no. 08, pp. pp. 130–143, Jun. 2023, doi:https://doi.org/10.3991/ijoe.v19i08.37417.

[17] P. . et. al., "A Survey on Heart Attack Prediction Using Machine Learning", Turkish Journal of Computer and Mathmatics Education (TURCOMAT), vol. 12, no. 2, pp. 2303 –, Apr. 2021.doi:https://doi.org/10.17762/turcomat.v12i2.1955.

[18] S. ware*, S. K. Rakesh, and B. Choudhary, "Heart Attack Prediction by using Machine Learning Techniques," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 5. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1577–1580, Jan. 30, 2020. doi:https://doi.org/10.35940/ijrte.d9439.018520.

[19] P. Cihan, "Effect of Parameter Selection on Heart Attack Risk Prediction in an RNN Model", *ICAENS*, vol. 1, no. 1, pp. 56–60, Jul. 2023. doi: https://doi.org/10.59287/icaens.964.

[20]Lutimath, N., Sharma, N. and K, B., "Prediction of Heart Disease using Biomedical Data through Machine Learning Techniques," *EAI Endorsed Transactions on Pervasive Health and Technology*, p.170881, Aug. 2021, doi:https://doi.org/10.4108/eai.30-8-2021.170881.

[21]Multirotor Research – Convolutional Neural Networks, available at <<https://multirotorresearch.com/career/>>, last accessed on 21-01-2024 at 8.00 PM.

[22]C# Corner - Random Forest Machine Learning Technique, available at <<https://www.c-sharpcorner.com/article/random-forest-machine-learning-technique/>>, last accessed on 30-12-2023 at 2.00 AM.

[23] Serokell Blog - Support Vector Machine Algorithm, available at <<https://serokell.io/blog/support-vector-machine-algorithm>>, last accessed on 27-12-2023 at 5.22 PM.

[24] SLA Jobs - KNN Algorithm in Machine Learning, available at <<https://www.slajobs.com/knn-algorithm-in-machine-learning/>>, last accessed on 05-01-2024 at 12.20 AM.

[25] Dataquest Blog - Tutorial: Decision Trees in Machine Learning, available at <<https://www.dataquest.io/blog/tutorial-decision-trees-in-machine-learning/>>, last accessed on 01-01-2024 at 2.01 PM.

PLAGARIASM REPORT

Predictive Precision: An In-Depth Examination of Machine Learning Approaches for Heart Attack Prognosis

ORIGINALITY REPORT

22%	17%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	fastercapital.com Internet Source	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Submitted to Coventry University Student Paper	1%
5	www.mdpi.com Internet Source	1%
6	Zhipeng Zhou, Lixuan Wei, Jingfeng Yuan, Jianqiang Cui, Ziyao Zhang, Wen Zhuo, Dong Lin. "Construction safety management in the data-rich era: A hybrid review based upon three perspectives of nature of dataset, machine learning approach, and research topic", Advanced Engineering Informatics, 2023 Publication	1%