

A PROJECT REPORT

On

**“A Comparative Analysis of Heart Failure Prediction
Models: K-Nearest
Neighbours, Logistic Regression, and Naive Bayes”**

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

**BACHELOR’S DEGREE IN
INFORMATION TECHNOLOGY
BY**

| | |
|--------------------------|-----------------|
| Prajukta Dey | 21052263 |
| Subarna Sutradhar | 21052288 |
| Riya Singh | 21052269 |
| Snehasish Pradhan | 21052107 |
| Abhishek Anand | 21052216 |
| Aditya Singh | 21052220 |

UNDER THE GUIDANCE OF

**Prof, A. Ranjith
Faculty ID:106797**



**SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
May 2020**

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024



CERTIFICATE

This is certify that the project entitled
“A Comparative Analysis of Heart Failure Prediction Models:
K-Nearest
Neighbours, Logistic Regression, and Naive Bayes”
submitted by

| | |
|-------------------|----------|
| Prajukta Dey | 21052263 |
| Subarna Sutradhar | 21052288 |
| Riya Singh | 21052269 |
| Snehasish Pradhan | 21052107 |
| Abhishek Anand | 21052216 |
| Aditya Singh | 21052220 |

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2023-2024, under our guidance.

Date: / /

(Guide Name)
Prof. A. Ranjith

Acknowledgements

We are profoundly grateful to Prof. A. Ranjith of **KIIT** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

Prajukta Dey
Subarna Sutradhar
Riya Singh
Snehasish Pradhan
Abhishek Anand
Aditya Singh

ABSTRACT

Cardiovascular diseases (CVDs) remain the primary cause of global mortality, claiming an estimated 17.9 million lives annually, which represents 31% of all deaths worldwide. This project delves into the critical task of predicting heart failure, a significant component of CVDs, leveraging machine learning techniques. With four out of five CVD deaths attributed to heart attacks and strokes, early detection becomes imperative, particularly as one-third of these deaths occur prematurely in individuals under 70 years of age. Through an extensive dataset encompassing 11 crucial features, including age, sex, and physiological indicators such as blood pressure and cholesterol levels, this study endeavours to develop predictive models to aid in the early identification and management of cardiovascular diseases.

The methodology employed involves a systematic exploration of the dataset, incorporating comprehensive exploratory data analysis, preprocessing techniques, and feature selection. Rigorous training and evaluation of selected machine learning algorithms are conducted, considering metrics such as accuracy, precision, recall, F1 score, and ROC-AUC, to discern their effectiveness in predicting heart failure. The comparative analysis offers insights into the strengths and weaknesses of each model, providing valuable guidance for healthcare practitioners and researchers in selecting the most suitable approach for early detection and management of cardiovascular diseases.

This project contributes significantly to the broader context of cardiovascular health by offering a nuanced evaluation of machine learning algorithms in predicting heart failure. By unravelling the intricate relationships between various attributes such as age, sex, chest pain types, and physiological indicators, the study aims to inform healthcare professionals about the most effective strategies for mitigating the burden of cardiovascular diseases. Ultimately, the findings of this research endeavour to advance the global effort to reduce the prevalence and impact of CVDs, thereby improving public health outcomes worldwide.

Keywords:

- | | |
|-------------------------------|----------------------|
| • Cardiovascular diseases | KNN |
| • Machine learning algorithms | Logistic Regression |
| • Risk factors assessment | Predictive modelling |
| • Exploratory data analysis | Naïve Bayes |

Contents

| | | |
|---|----------------------|---|
| 1 | Introduction | 1 |
| 2 | Literature Review | 2 |
| 3 | Implementation | 3 |
| | 3.1 | |
| | Research Methodology | 3 |
| | 3.2 | |
| | Result Analysis | 6 |
| 4 | Conclusion | 8 |

Chapter 1

Introduction:

Cardiovascular diseases (CVDs) represent the leading cause of global mortality, claiming approximately 17.9 million lives annually, which accounts for 31% of all deaths worldwide. Among CVD-related fatalities, four out of 5 deaths are attributed to heart attacks and strokes, with a significant portion occurring prematurely in individuals under 70 years of age. Heart failure, a common consequence of CVDs, poses a substantial health threat. Utilizing datasets containing 11 predictive features, researchers have explored machine learning techniques to forecast heart failure. Despite advancements, there remains a notable gap in comprehensive model comparisons, including newer methodologies like Deep Forest, which could enhance predictive accuracy and inform preventative interventions. Addressing this gap is crucial given the escalating prevalence of heart failure driven by sedentary lifestyles, poor dietary habits, and environmental factors. Failure to mitigate these risks could exacerbate the global burden of cardiovascular diseases, underscoring the urgency for robust predictive models and proactive health strategies.

Machine learning, as defined by IBM, is a discipline of artificial intelligence (AI) that focuses on using data and algorithms to replicate the way humans learn, give some decisions. There are two types of machine learning technique: unsupervised learning and supervised learning. Unsupervised learning is involving of analyses and classifies unlabelled data sets and supervised learning is learning that works by labelled data.

This methodology entails a structured examination of the dataset, comprising in-depth exploratory data analysis, preprocessing steps, and feature selection procedures. Following this, the chosen algorithms undergo thorough training and evaluation, with careful consideration given to metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. Through a comparative analysis, the strengths and weaknesses of each model are discerned, offering valuable insights into their effectiveness for early heart failure prediction. This study meticulously explores various attributes like age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, Resting ECG, MaxHR, Exercise Angina, etc.

Chapter 2

Literature Review :

A total of 17.9 million cases of death are caused by constriction of blood vessels in the heart, this makes heart disease ranks first among the top ten causes of death in the world . Heart disease has symptoms such as an irregular heartbeat, blood vessel problems, or chest pain. There are several factors that can cause heart disease, including: age, gender, diabetes, cholesterol, hypertension, and diet . Over time, the number of deaths from heart disease is also increasing, which makes technology important in predicting heart disease as a preventative measure. By predicting heart disease early, it can help treat heart disease better and more accurately, so that it can have an impact on improving the patient's quality of life. There are a variety of methods that can be used to classify heart disease against data from the health industry about the factors that cause heart disease. Various AI methods have been applied to help predict heart disease, namely machine learning that can learn from datasets from various hospitals to produce accurate prediction

D.P.Yadav, Prabhav Saini, Pragya Mittal applied machine Learning techniques like K-Nearest Neighbour, Support Vector Machine (SVM), Naïve Bayes and Random Forest on the dataset to predict Heart Disease. Among these model Naïve bayes using 3-fold cross validation get the highest accuracy of 87.9%. A feature optimization technique Genetic algorithm is implemented for increasing the model performance. After applying optimization technique Naïve Bayes achieved accuracy 96%.

Surai Shinde, Juan Carlos Martinez-Ovando combines the features of recurrent neural network and convolutional neural network to create the deep learning based hybrid model, which helped to attain better accuracy.

Support Vector Mac Support Vector Machine, K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), Deep Residual Neural Network, Logistic Regression, and other algorithms could be used to identify heart failure. The supervised learning model of Logistic Regression and Support Vector Machine will be classified in the paper, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Deep Residual Neural Network, Logistic Regression, and other algorithms could be used to identify heart failure . The supervised learning model of Logistic Regression, K Nearest Neighbor and Naïve Bayes Classifier will be classified in the Pape

Chapter 3

Implementation :

3.1 Research Methodology:

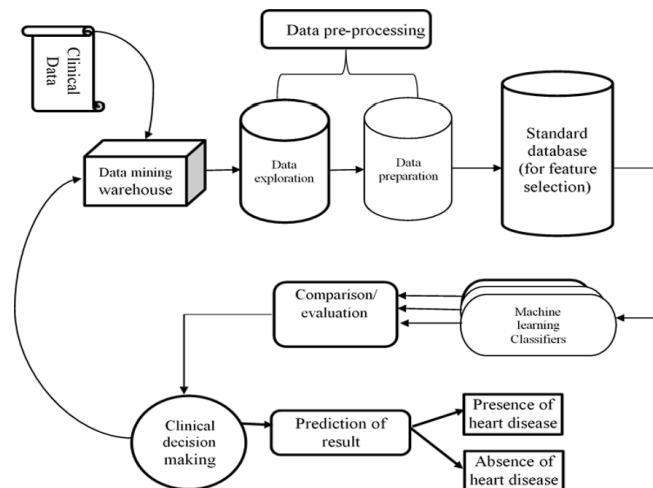
Dataset & Attributes:

The creation of this dataset involved the integration of multiple existing datasets that had not been previously combined. Currently, this dataset stands as the most extensive resource available for heart disease research, as it merges five distinct heart datasets (Statlog (Heart) Data Set: 270 records, Cleveland: 303 records, Hungarian: 294 records, Long Beach, VA: 200 records, Switzerland: 123 records) and shares 11 common features. The dataset is accessed from Kaggle named 'Heart Failure Prediction Dataset' [25]. The dataset contains 920 patient records, including 725 males and 195 females of different ages. Where 267 males are normal, and 458 males have heart disease, 145 females are normal, and 50 females have heart disease. The comprehensive depiction of every attribute, along with the corresponding count of values for each attribute, can be observed in the figure given below.



Data Preprocessing:

Data preprocessing plays a crucial role in machine learning , and its importance cannot be overstated. To enable the machine to learn from the data and generate the suitable model, it is crucial to convert the categorical feature values into numerical representations through a process known as an encoding` method, which is utilised here .



Data cleaning:

After understanding the dataset, it is necessary to do data cleaning such as handling null values and data duplication, normalizing data, removing irrelevant variables, and so on. The dataset used in this paper does not have null values and duplicate data, but categorical variables need to be converted into numeric variables because machine learning models are based on mathematical calculations. Since there are not many unique values of categorical variables in this dataset the ideal encoding method is the OneHot Encoding. OneHot encoding is the process of adding a new feature to each categorical variable whose values are 0 and 1.

| | Age | RestingBP | Cholesterol | MaxHR | Oldpeak | HeartDisease | Sex_F | Sex_M | ChestPainType_ASY | ChestPainType_ATA | ... | FastingBS_0 | FastingBS_1 | RestingECG_LVH | RestingECG_Normal | RestingECG_ST | ExerciseAngina_N | ExerciseAngina_Y | ST_Slope_Down | ST_Slope_Flat | ST_Slope_Up |
|-----|-----|-----------|-------------|-------|---------|--------------|-------|-------|-------------------|-------------------|-----|-------------|-------------|----------------|-------------------|---------------|------------------|------------------|---------------|---------------|-------------|
| 0 | 40 | 140 | 289 | 172 | 0 | 0 | 0 | 1 | 0 | 1 | ... | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 49 | 160 | 180 | 156 | 1 | 1 | 1 | 0 | 0 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 2 | 37 | 130 | 283 | 98 | 0 | 0 | 0 | 1 | 0 | 1 | ... | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 3 | 48 | 138 | 214 | 108 | 1 | 1 | 1 | 0 | 1 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 4 | 54 | 150 | 195 | 122 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 913 | 45 | 110 | 264 | 132 | 1 | 1 | 0 | 1 | 0 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 914 | 68 | 144 | 193 | 141 | 3 | 1 | 0 | 1 | 1 | 0 | ... | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 915 | 57 | 130 | 131 | 115 | 1 | 1 | 0 | 1 | 1 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 916 | 57 | 130 | 236 | 174 | 0 | 1 | 1 | 0 | 0 | 1 | ... | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 917 | 38 | 138 | 175 | 173 | 0 | 0 | 0 | 1 | 0 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

918 rows x 22 columns

Proposed model :

In this study we have used the following three classifier techniques to achieve the most efficient results.

Logistic Regression:

Logistic Regression is a supervised classification algorithm. It's a probabilistic analysis algorithm that predicts outcomes. By estimating probabilities using the underlying logistic equation, it assists in measuring the relationship between the dependent variable (TenyearCHD) and one or more independent variables (risk factors) (sigmoid function). The following logistic function is used in the logistic regression algorithm:

$$p=1/(1+e^{-x}) \quad (1)$$

The logistic coefficients for each instance $x_1, x_2, x_3, \dots, x_n$ will be $b_0, b_1, b_2, \dots, b_n$ during the training stage. Stochastic gradient descent is used to estimate and update the coefficient values.

$$\text{Values} = b_0x_0 + b_1x_2 + \dots + b_nx_n \quad (2)$$

$$b = b + l * (y - p) * (1 - p) * p * x \quad (3)$$

Naïve Bayes :

NB algorithm is a probabilistic supervised ML technique used for Bayes theorem-based classification problems. Bayes Theorem is used for calculating conditional probability. The possibility of an event happening given that another action has already happened is called conditional probability.

Mathematical Representation of Bayes theorem:

$$P(AB) = (P(BA) * P(A)) / P(B)$$

KNN (K- Nearest Neighbor) :

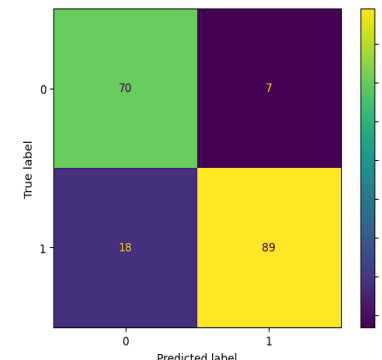
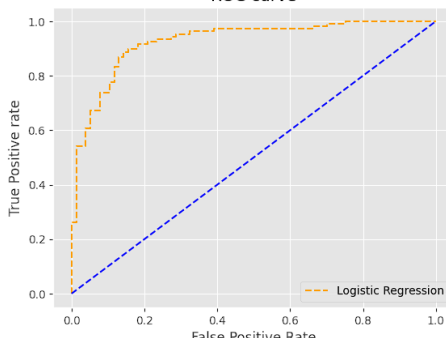
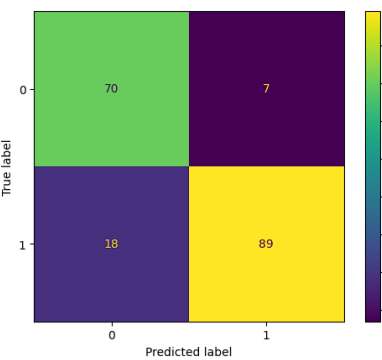
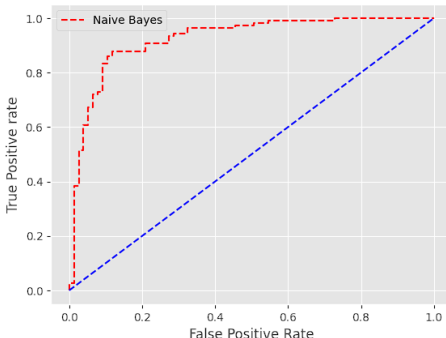
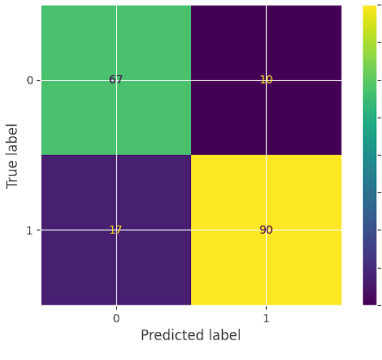
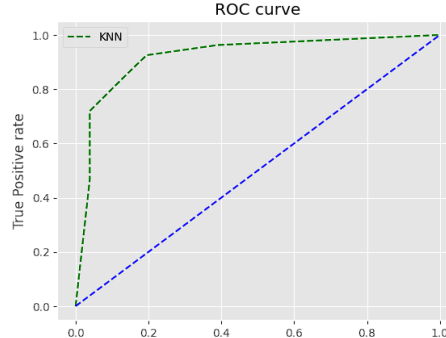
The K nearest neighbor algorithm is an example-based learning algorithm that is widely used in real-life scenarios. The K Nearest Neighbors algorithm can be used to solve both classification and regression problems. The K Nearest Neighbor algorithm involves preprocessing the dataset, training the model, and testing the model. Cleaning and removing erroneous and outlier values from a dataset is usually part of the preprocessing phase. In the algorithmic process, this is the most important step. It's also crucial to check the dataset's accuracy before running algorithmic tests on it. It is important to pay attention to missing values and outliers.

3.2 Result Analysis:

Three different classifying techniques have been applied in the proposed model with Logistic Regression, Naïve Bayes & and K Nearest Neighbour.

The following table depicts the results implementing various metric evaluation criteria's which are as follows - confusion matrix, cross validation scores , ROC curve & AUC score.

Confusion Matrix & ROC curve:

| Logistic Regression |  <p>Confusion Matrix for Logistic Regression:</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>0</th><th>1</th></tr> </thead> <tbody> <tr> <th>0</th><td>70</td><td>7</td></tr> <tr> <th>1</th><td>18</td><td>89</td></tr> </tbody> </table> | True label \ Predicted label | 0 | 1 | 0 | 70 | 7 | 1 | 18 | 89 |  <p>ROC curve for Logistic Regression:</p> <p>The ROC curve shows a True Positive Rate (Y-axis) versus False Positive Rate (X-axis). The curve is a dashed orange line, indicating good performance. The area under the curve is approximately 0.85.</p> |
|------------------------------|--|------------------------------|---|---|---|----|----|---|----|----|---|
| True label \ Predicted label | 0 | 1 | | | | | | | | | |
| 0 | 70 | 7 | | | | | | | | | |
| 1 | 18 | 89 | | | | | | | | | |
| Naïve Bayes |  <p>Confusion Matrix for Naïve Bayes:</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>0</th><th>1</th></tr> </thead> <tbody> <tr> <th>0</th><td>70</td><td>7</td></tr> <tr> <th>1</th><td>18</td><td>89</td></tr> </tbody> </table> | True label \ Predicted label | 0 | 1 | 0 | 70 | 7 | 1 | 18 | 89 |  <p>ROC curve for Naïve Bayes:</p> <p>The ROC curve shows a True Positive Rate (Y-axis) versus False Positive Rate (X-axis). The curve is a dashed red line, indicating good performance. The area under the curve is approximately 0.85.</p> |
| True label \ Predicted label | 0 | 1 | | | | | | | | | |
| 0 | 70 | 7 | | | | | | | | | |
| 1 | 18 | 89 | | | | | | | | | |
| K Nearest Neighbour |  <p>Confusion Matrix for K Nearest Neighbour:</p> <table border="1"> <thead> <tr> <th>True label \ Predicted label</th><th>0</th><th>1</th></tr> </thead> <tbody> <tr> <th>0</th><td>67</td><td>19</td></tr> <tr> <th>1</th><td>17</td><td>90</td></tr> </tbody> </table> | True label \ Predicted label | 0 | 1 | 0 | 67 | 19 | 1 | 17 | 90 |  <p>ROC curve for K Nearest Neighbour:</p> <p>The ROC curve shows a True Positive Rate (Y-axis) versus False Positive Rate (X-axis). The curve is a dashed green line, indicating good performance. The area under the curve is approximately 0.85.</p> |
| True label \ Predicted label | 0 | 1 | | | | | | | | | |
| 0 | 67 | 19 | | | | | | | | | |
| 1 | 17 | 90 | | | | | | | | | |

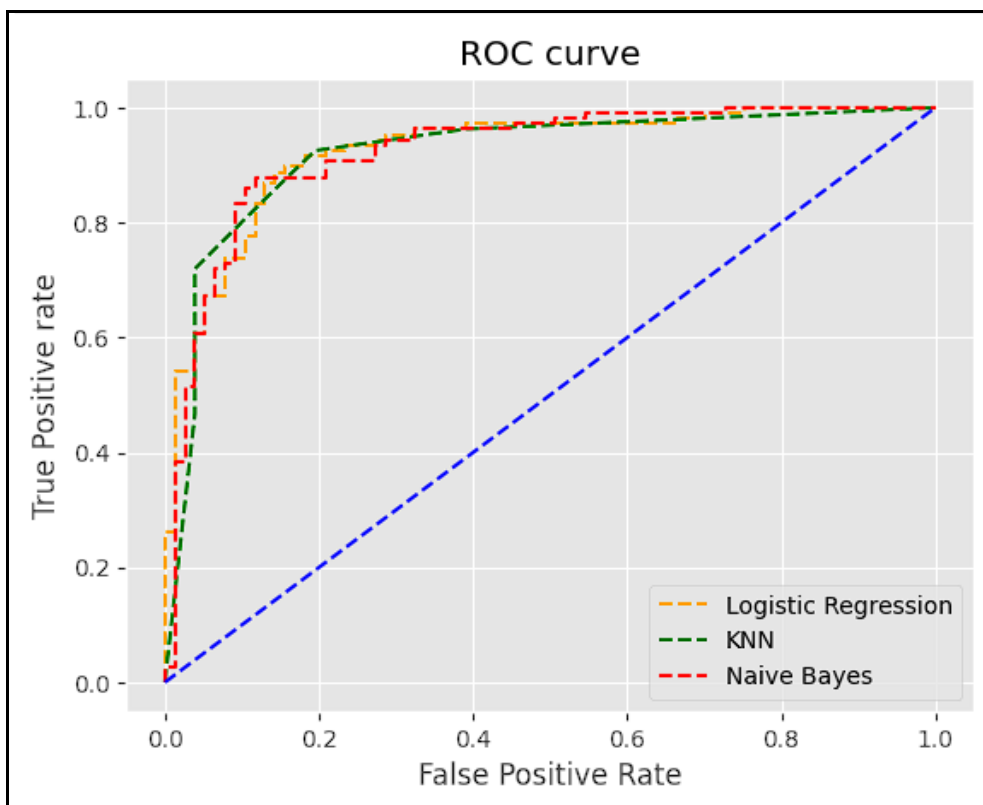
(Table-1)

Accuracy & AUC score :

| Model | Accuracy (after Tuning) | AUC score |
|---------------------|----------------------------|--------------------|
| Logistic Regression | 0.871 | 0.926690132297609 |
| Naïve Bayes | 0.8641304347826086 | 0.9236557834688677 |
| K Nearest Neighbour | 0.875 | 0.9202573127806772 |

(Table-2)

Combined ROC curve :



Chapter 4

Conclusion :

Machine learning based solutions are widely used in healthcare sector for analyzing patients' data, predicting diseases and suggesting possible treatments. With a number of machine learning techniques available today, it is important to identify the most efficient and accurate technique especially in critical domains like healthcare. A comparative analysis of the various Machine learning algorithms used in the heart failure prediction is presented. Three of the most widely used techniques, KNN, Logistic regression and Naive Bayes are discussed and compared to identify the best suited classifier for heart failure prediction. Many previous researches and studies related to heart failure prediction were identified and analyzed. Based on the identified researches and studies a performance analysis of various machine learning algorithms along with their accuracies used for prediction of heart failure was done as presented in table 2. The findings shows that in most cases machine learning based approaches have shown significant potential to transform the healthcare sector and improve the entire process of disease predictions and suggesting treatments.

References

1. J. C. Martinez-Ovando and S. Shinde, "Heart Disease Detection with Deep Learning Using a Combination of Multiple Input Sources", *ETCM 2021 — 5th Ecuador Tech. Chapters Meet.*, pp. 2021-2023, 2021.
2. Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access*, 8, 107562-107582.
3. Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018, March). Prediction of heart disease using machine learning. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1275-1278). IEEE.
4. Lafta R, YanLi, Tseng VS. An Intelligent Recommender System based on Short Term Risk Prediction for Heart Disease patients. *IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Singapore: IEEE; (2015).