Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Predicting Heart Failure Survival with Machine Learning: Assessing My Risk

Md Nasiruddin[1] ✉ Shuvo Dutta[2], Rajesh Sikder[3], Md Rasibul Islam[4], Abdullah AL Mukaddim[5] and Mohammad Abir Hider[6]

[1]Department of Management Science and Quantitative Methods, Gannon University, USA
[2]Master of Arts in Physics, Western Michigan University, USA
[3]PhD Student in Information Technology, University of the Cumberlands, KY, USA
[4]Department of Management Science and Quantitative Methods, Gannon University, USA
[56]Masters of Science in Business Analytics, Grand Canyon University

Corresponding Author: Md Nasiruddin, **E-mail**: nasirudd001@gannon.edu

| ABSTRACT

This study investigates the application of machine learning techniques for heart disease prediction using a comprehensive dataset of 918 patients. The research employs multiple algorithms, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Networks, to develop predictive models based on 11 clinical features. The dataset, compiled from five independent sources, underwent thorough preprocessing and was split into training (70%) and test (30%) sets. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Results demonstrate consistently high performance across all models, with the SVM achieving the highest overall performance (accuracy: 88.41%, precision: 89.76%, recall: 90.85%, F1-score: 90.30%, ROC-AUC: 94.97%). Key predictors identified include age, maximum heart rate, and ST depression (Oldpeak). The study's findings have significant implications for clinical practice, offering the potential for rapid, objective heart disease risk assessment. The consistent performance across different model architectures provides flexibility for implementation in various healthcare settings. Limitations include potential data collection variability and gender imbalance in the dataset. Future research directions include developing more sophisticated neural networks, incorporating additional data types, and conducting prospective studies to validate model performance in real-world clinical settings. This research contributes to the growing body of evidence supporting the use of machine learning in medical diagnostics. The developed models enhance early detection and risk stratification of heart disease, potentially improving patient outcomes through timely interventions.

| KEYWORDS

Heart disease prediction, machine learning, Support Vector Machine (SVM), neural networks, clinical decision support, risk assessment, predictive modeling, feature importance, healthcare analytics, early detection, cardiovascular health, data-driven diagnostics, medical informatics, precision medicine, ROC-AUC.

| ARTICLE INFORMATION

ACCEPTED: 01 August 2024 **PUBLISHED:** 07 August 2024 **DOI:** 10.32996/jcsts.2024.6.3.5

1. Introduction

1.1 Background

Heart failure is an extremely severe global health concern that affects an estimated 64.3 million patients worldwide. This chronic, life-threatening condition—the inability of the heart to pump blood effectively—generates serious complications and is accompanied by a poor quality of life for patients. With aging populations, especially in the more developed parts of the world, the prevalence of heart failure will continue to increase, underscoring the rising need for accurate prognosis and survival prediction.

The ability to predict survival in heart failure is relevant to healthcare providers, patients, and health systems. On the one hand, accurate predictions provide an informed basis for which healthcare providers decide on treatment pathways, intensify intervention, apply advanced life-prolonging therapies, or plan for end-of-life care. Reliable survival predictions enable more efficient use of resources so that high-risk patients get appropriate care and follow-up.

Recently, growing interest has developed regarding the role that machine learning could play in health care. The power of machine learning algorithms in processing large amounts of data, detecting subtle patterns, and accounting for sleek interactions among variables opens new opportunities for examining intricate medical data and improving the accuracy of predictions. Such technological advancement now opens the possibility of changing the concept of heart failure treatment and can save many lives.

1.2 Problem Statement

Conventionally, clinical judgment and risk factor statistical models, primarily based on established risk factors, have been used to make a prognosis for heart failure. These methods are unsatisfactory for modeling all the complexities of the process: intricate interplays among multiple physiological systems, comorbidities, and treatment responses.

Because traditional methods have limitations, interest in more advanced prediction methods, such as machine learning and artificial intelligence, has recently increased. Such an approach may circumvent all the potential fallacies of conventional techniques to provide better and more granular predictions of heart failure outcomes.

The challenge is developing simultaneously accurate, efficient, interpretable, and clinically applicable models. Although machine learning offers many promises, careful validation is required for successful integration into clinical practice, given issues of practical implementation and ethical concerns.

1.3 Objectives

The main aims of this study are:

1. Development of predictive machine learning models for survival from heart failure using advanced algorithms and large datasets for a more accurate and efficient prediction tool.
2. Performance in effectiveness between different machine learning techniques will be compared to predict the outcome of heart failure. In this case, the objective setting targets spotting the most promising algorithms for this application.
3. The goal is to identify and analyze the key predictive factors for survival rates and measure their relative influence in the prediction models. Such an analysis will reveal how complex interactions among patient parameters, previous medical history, and clinical parameters affect HF outcomes.
4. The predicting power of these models will be assessed to determine their possibility of adoption into clinical decision-making based on their predicting power and practicality relevant within a healthcare environment.

1.4 Research Questions

Guided by the above objectives, the following research questions have been formulated to help the study in efforts aimed at addressing the set objectives:

1. Which machine learning algorithms show the most promise in predicting heart failure survival, and how do their predictions compare with those of the standard statistical models?
2. What relationships can be assessed between the prediction outcomes and input variables, such as patient demographics, history of medical conditions, and clinical parameters?
3. Can a reliable, interpretable model for decision-making in clinical care be developed with a significant improvement in patient care?
4. What challenges and issues might implementing machine learning prediction models into clinical practice face, and how are they addressed?

1.5 Scope of the Study

This study will develop and evaluate machine learning models for predicting heart failure survival based on a varied patient information dataset. Data will be derived from a multi-center heart failure registry, including demographic information, past medical history, laboratory reports, and clinical outcomes of patients with a diagnosis of heart failure.

It will involve only adult patients aged 18 years or older with a diagnosed case of heart failure. Patients will be followed prospectively for a defined period to capture outcome data. The research will focus on the patient population so that the models developed are most relevant to adults who have suffered from heart failure.

While developing robust predictive models is the aim of this study, one must recognize its limitations. Model performance may be data-specific; therefore, generalization to other patient populations would mandate further validation. Translating the application of these models into clinical workflows lies outside the scope of this study and, hence, would be evaluated separately based on practical and ethical considerations.

Data sources include electronic health records, clinical trial data, and registry information from contributing healthcare institutions. Such data sources will yield a rich and varied dataset for the development and testing of models. However, these sources are bound to result in varying levels and quality of data, which could have consequences for the performance and generalizability of this model.

The research will focus on adult patients with a confirmed diagnosis of heart failure to develop models tailored to this patient population. This may consequently limit the generalizability of findings to other age groups or to patients with different cardiovascular health conditions. Further studies are needed to generalize such models to a broader population of patients or to adapt them for application in other clinical contexts.

The development and evaluation of the predictive models will be the main focus, not their deployment in a clinical setting. While the potential for clinical application will be discussed, further investigation into the testing of such models in real-world settings to see how they work when put into general healthcare systems and subsequently seeing the effect of such models on the patients' outcomes will be required.

This study will help advance heart failure prediction using machine learning algorithms' power. If more accurate and efficient prediction models are developed, this research may profoundly impact patient care, resource allocation, and clinical decision-making. Therefore, results from the current study could pave the way for more personalized treatment approaches, enhancing care outcomes for heart failure patients. Such models should be developed and implemented carefully, with the benefits and limitations understood in a real clinical setting.

2. Literature Review

2.1 Overview of Heart Failure

Introduction Heart failure is a complex clinical syndrome in which the heart cannot pump blood effectively to meet the body's metabolic demands. It represents the failure of one or other organ to meet metabolic demands—both final cumulative and the result of several heterogeneous disorder states. However, it is a major cause of morbidity and mortality. Thus, understanding the pathophysiology, risk factors, and predictor indicators of prognosis in heart failure is requisite in developing effective prediction models.

"In reality, heart failure pathophysiology is complex, with influences that reach from neurohormonal to inflammatory and metabolic mechanisms. In the acute setting, these counterregulatory mechanisms, such as the renin-angiotensin-aldosterone and sympathetic nervous systems, provide powerful compensatory benefits for cardiac output. However, if these systems are activated chronically, they result in maladaptive cardiac remodeling and progressive worsening of heart function, as Kemp & Conte 2012 described.

2.2 Critical factors involved in the pathogenesis of heart failure

Etiology: Prognosis largely depends on the cause of heart failure. In the developed world, ischemic heart disease is the most common underlying etiology; however, hypertension and valvular heart disease are more common in developing countries (Dokainish et al., 2015). Classification according to etiology is necessary to decide on a treatment plan and prognosis. • Left Ventricular Ejection Fraction (LVEF): The prognosis critically depends on the LVEF. Classification of heart failure is often made into HF with reduced ejection fraction, HF with a preserved ejection fraction, and HF with a mid-range ejection fraction. Under these categories lie varied pathophysiologies and treatment methods, as stated by Lam et al. 2018. HFrEF, characterized by an LVEF of ≤40%, constitutes the vast arm of clinical trials and has laid down therapies. HFpEF, with LVEF ≥ 50%, leaves elusive therapeutic options associated with a different risk factor profile and outcomes.

Comorbidities: Conditions like diabetes, chronic kidney disease, and chronic obstructive pulmonary disease have a big influence on heart failure patients' outcomes (van Deursen et al., 2014). These conditions are more likely to affect the progress of heart failure and act as complicating factors with treatment strategies and prognosis. For instance, diabetes increases the risk of hospitalization and mortality in heart failure patients.

Biomarkers: Natriuretic peptides (BNP and NT-proBNP) are firmly established biomarkers periodically applied for diagnosis and prognosis of heart failure. Emerging biomarkers, including ST2, galectin-3, and high-sensitivity troponin, provide incremental

prognostic information (Chow et al., 2017). These biomarkers are the substrates of different pathophysiological processes in HF, including myocardial stress, fibrosis, and injury. Integrating multiple biomarkers can arrive at a more comprehensive risk profile for a patient.

Functional capacity: Well-established measures strongly predict outcomes, particularly the New York Heart Association functional classification and exercise capacity—measured as peak oxygen consumption. The NYHA classification grading consists of four classes: Class I: no limitation is experienced in any activities; Class II: slight limitations; Class III: marked limitations; and Class IV: severe limitations. It is a simple but powerful index for disease severity and prognosis. Objective measures of exercise capacity supplement the prognostic information available from functional class.

Age and Frailty: Advanced age and frailty are related to poor outcomes in patients with heart failure. Given an aging population, which mirrors the rising burden of heart failure globally, there has been a growth in the number of heart failure cases in elderly patients. Frailty, viewed as a syndrome of decreased reserve and resistance to stressors, is widely prevalent among elderly heart failure patients and, therefore, has enormous treatment and prognostic implications.

All these predicted factors are very important in developing comprehensive prediction models that can effectively assess the risk of adverse outcomes in heart failure subjects. The intricate interplay of these factors makes it essential to rely on the needful investment in sophisticated analytical approaches and an attraction towards machine learning as a way of making inroads into risk prediction in heart failure.

2.3 Machine Learning in Healthcare

Thus, ML in health care has grown exponentially in recent years, with new possibilities to improve diagnosis, prognosis, and treatment decisions. Vast amounts of highly complex medical data can be crunched using machine learning algorithms, which could point out subtle patterns and make predictions far beyond the capabilities of classical statistical methods. Machine learning algorithms have been developed that are designed to handle large volumes of complex medical data, recognize subtle patterns, and make the corresponding predictions in a way that could not be made otherwise using classical statistical methods. Applications in the general regard of medical predictions include:

Disease Diagnosis: Many ML models have shown encouraging performance in solving these problems in medical diagnosis, from diabetic retinopathy to skin cancer. Their performances mostly match or surpass those of human experts in the domain of interest. The models can also analyze complex data, considering imaging, laboratory tests, and clinical information, to enable early diagnosis.

Risk Stratification: ML algorithms can combine different risk factors to yield accurate risk assessment in different diseases, and cardiovascular events are no exception. By their very nature—able to accommodate a wide range of variables and their interactions simultaneously—ML models can derive individual risk predictions greater than those derived from traditional risk scores.

Response Prediction to Treatment: ML models can predict patient responses to some treatments, which can be used to formulate much more personalized treatment strategies compared to those used now (Cho et al., 2018). Those applications are especially relevant in fields such as oncology and cardiology because treatment responses vary widely among different patients.

Healthcare Resource Utilization: ML can predict readmission rates and resource use, contributing to managing health systems. If healthcare providers identify patients more likely to be readmitted or have complications, they can provide targeted interventions to improve outcomes and reduce costs.

Various machine learning techniques have been applied to execute different functions for heart-related diseases.

Motwani et al. (2017) applied machine learning to coronary computed tomography angiography data and showed improved prediction of 5-year all-cause mortality than that possible with traditional clinical risk assessment. Their model included imaging features and clinical data, pointing to the potential of integrating multi-data sources in risk prediction.

Kwon et al. developed a deep learning-based model for predicting heart failure readmission with great superiority over performance in traditional risk scores using electronic health record data. The model employed a wide range of clinical variables, from vital signs to laboratory results and medication information, to deliver highly accurate predictions for readmission.

In addition, Shameer et al. (2018) employed an ensemble learning approach that was ascertainable in predicting heart failure readmission through the large integration of clinical, laboratory, and medication data. They then combined it using many machine learning algorithms; hence, eventually, it was robust and accurate in prediction.

From these reports, by exploiting diversified data types, machine learning can give a strong boost to traditional predictive models that calculate cardiovascular risk while at the same time ensuring better accuracy. The greater capacity of ML models to manage high-dimensional data and complex relationships between variables makes them particularly suitable for predicting heart failure from several factors interacting with heart function that affects prognosis.

Comparisons have been made within several studies for various machine learning models of the heart failure-prediction model and have given an insight into relative strength and weakness parameters.

Chicco and Jurman compared several machine learning algorithms using a heart failure dataset: logistic regression, random forests, support vector machines, and artificial neural networks. Random forests emerged with the highest accuracy, followed by F1 scores, while SVM came very close. Even logistic regression, as easy as it is, performed well, underlining feature selection and data preprocessing.

Random forests have been shown to be effective in preventing overfitting and handling weak nonlinear relationships. However, their interpretability could be stronger, especially compared to simple models like logistic regression. This highlights a point of great interest in many clinical applications: the trade-off between model complexity and interpretability.

As Angraal and colleagues did in their systematic review of heart failure with machine learning applications, the conventional statistical approaches were benchmarked against various ML algorithms. ML models perform better than conventional statistical models to analyze complex, high-dimensional data. However, substantial heterogeneity in the study designs and evaluation metrics impedes direct comparisons.

The ML models' strengths included their ability to capture complex interactions between variables and handle large datasets. Weaknesses included the danger of overfitting, particularly in smaller datasets, and some complexes are "black boxes," making the models poorly interpretable. This review has indicated a need for standardized reporting and evaluation metrics in ML studies to facilitate comparisons and clinical translation.

In their work, Adler et al. compared head-to-head deep learning models with the traditional risk scores commonly used in practice to predict the mortality outcome among patients already diagnosed with heart failure. They found better discriminative ability, especially towards high-risk patients, although external validation was considered, and interpretative model outputs for clinical adoption were needed.

The present study thus holds promise for deep learning to enhance the accuracy of risk prediction but raises challenges in clinically practicing these models. The general application of deep learning to heart failure prediction presents challenges, opportunities, and requirements of large datasets, computational resources, and interpretable outputs.

2.4 Gaps in Current Research

Among these, machine learning has held promise for the prediction of heart failure, though the study of heart failure still has several gaps so far.

Limited External Validation: Most studies have been based on single-center datasets, thus limiting generalizability. Therefore, this hypothesis should not be confirmed since external validation has yet to be conducted for this aspect in major populations and settings within healthcare. This will ensure that the performance of ML models is homogeneous across populations and, therefore, is reliably reproducible within divergent clinical environments.

Interpretability: In the next level, when models become more complex, as in deep neural network models, they often achieve high prediction accuracy but with little interpretability. How to make model predictions interpretable in clinically meaningful ways is still an area of active research (Ahmad et al., 2018). These factors make interpretable models in ML highly important for gaining the trust of the clinician for integration into the decision process.

Dynamic Risk Prediction: Most existing models have focused only on a static risk assessment. There is room for dynamic models whereby predictions are updated with the increased availability of data over time and the course of a patient's disease. Dynamic risk prediction models help personalize estimates and provide more timely risk estimates, ultimately improving patient management.

Multi-modal Data Integration: While some studies focus on applying multiple data from different sources, new sources like wearable devices, genomics, and social determinants related to health can be integrated to provide dense data sources for population and individual health (Johnson et al., 2021). This can provide richer information about a patient's health state and a risk factor information set for prediction personalization.

Clinical Implementation: Few studies have focused on the practical implementation of ML models in the clinic, including their integration with electronic health records, user interface design, and impacts on clinical decisions. Addressing such problems in implementing ML is necessary to effectively translate such models from the research setting into clinical practice.

Missing Data Handling: Any real clinical data most probably contains missing values. According to Beaulieu-Jones et al. (2017), much work must be done on robustly handling missing data in machine-learning models for heart failure prediction. Therefore, unfitting such missing data requires the development of ways to effectively handle such data to build robust machine-learning models for practical, real-world clinical applications.

Fairness and Bias: Ensure that ML models provide fair outcomes to all population groups without creating or reinforcing disparities in treatment among population groups. This is, however, essentially an area of ongoing research as the impact of models on clinical decisions grows, with bias and fairness of the model under parallel concern.

This work will fill these gaps by developing interpretable machine-learning models, incorporating diverse data types, and evaluating their performance across patient subgroups. It will contribute to the emerging field of machine learning applications for heart failure prediction and step closer to moving into clinical practice.

In this background, much hope lies in integrating machine learning in heart failure prediction to improve a more personal focus on patient care and outcomes. Given ML's enormous analytical capacity for such complex multidimensional data, it is supposed to provide rather exact and personalized risk assessment, followed by the corresponding treatment strategy and resource allocation for patients.

However, realizing this potential requires addressing the identified gaps in the current state of research. Future studies should focus on developing ML models that are highly interpretable, can be easily integrated into clinical workflows, and allow for validation across different populations to ensure fairness and reliability.

As ML in healthcare unfolds, collaboration between data scientists, clinicians, and healthcare administrators is imperative. This collaboration will be important in overcoming barriers to implementation and translating research findings into practice. By addressing these challenges, we may optimize machine learning for the prediction of heart failure and improve patient care and outcomes.

3. Methodology

3.1 Data Collection

The database used in this research consolidates information on heart disease from five independent sources: Cleveland Clinic Foundation, Ohio; Hungarian Institute of Cardiology, Budapest; University Hospital, Zurich; V.A. Medical Center, Long Beach, California; and Stalog Heart Disease Dataset. This puts together the largest publicly available heart disease dataset for research, with 918 unique observations after removing duplicates.

The inclusion criteria in the dataset ensured that all patients had records for the following features: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, old peak, ST slope, and a definite diagnosis about the presence or absence of heart disease. The exclusion criteria included duplicate entries, incomplete records, and ambiguous or inconclusive diagnoses.

Data collection aimed to obtain a diverse and representative sample of patients with various risk factors and outcomes. This dataset includes broad demographic and geographic diversity because it was combined from multiple sources. This enhances the generalizability of any predictive models developed on this data. However, one should be aware that possible discrepancies in data collection or diagnostic criteria might appear between the sources.

3.2 Data preprocessing

Some vital steps that the data preprocessing went through to make this dataset ready to be fed into the machine learning algorithms are described here. First, it checked for missing values using the panda's function data. isnull().sum(). This indicated a

complete dataset with no missing entries. The second involved separating feature X from the target variable of heart disease. In the models, all available features would be used as predictors.

The data set was then split into a train and test set in the ratio 70:30. This was done using the 'train_test_split' function from the sklearn library. The Features were then classified into numerical and categorical types. Numerical consisted of Age, RestingBP, Cholesterol, FastingBS, MaxHR, and Oldpeak, while categorical consisted of Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope. This was done to ensure the respective preprocessing techniques were applied for these feature types.

Two different preprocessing pipelines had to be created: one for numerical data and the other for categorical. The former included imputation with the median in case there were missing values, while the latter was followed by standardization. The latter pipeline imputes missing values with the most frequent value, followed by one-hot encoding. Then, these pipelines were merged using ColumnTransformer.

3.3 Machine Learning Algorithms

The authors used four different machine-learning algorithms in this work. All of these algorithms were chosen for the different strengths and insights they can provide. Logistic regression, a basic algorithm for binary classification problems, was chosen because of its interpretability and ease of implementation. This algorithm is a good starting point for determining the relationship between the feature and target variables.

The second option was that of Random Forest, an ensemble learning method. In this case, the algorithm constructs many decision trees and merges their predictions to provide strong performance in nonlinear relationship handling and further valuable feature importance information. Because it is possible to capture complex interactions between features, Random Forest became a strong candidate for heart disease prediction.

It included a Support Vector Machine (SVM) for its effectiveness in high-dimensional spaces, and different kernel functions ensure versatility. An SVM has to be set for probability estimation to compare results with other models. Due to its ability to detect optimal decision boundaries in complex feature spaces, SVM has the potential to become a robust tool for the task of classification at hand.

Finally, the neural network implementation using sci-kit-learn's MLPClassifier was used. It can handle complicated non-linear relationships in the data and find patterns that simpler models cannot. A maximum iteration value of 1000 was used to ensure the model had enough iterations to converge while training.

Using such a diverse set of algorithms, the aim was to capture different aspects of the dataset to establish which one effectively predicts heart disease with the given features. These algorithms contribute individually to such analysis with their different strengths, making the understanding of the data complete and probably bringing out some insights that no single approach might give.

3.4 Model Development

In the model development phase, five various machine learning algorithms connected with heart disease prediction have been implemented and trained. The development process started with preparing the dataset and was divided into two parts: 70% for training and 30% for testing, both of which are used for unbiased evaluation.

In these premises, four initial models were developed using scikit-learn: Logistic Regression, Random Forest, Support Vector Machine, and a Neural Network—the Multi-Layer Perceptron. All the above models were chosen because of their radically different ways of treating a classification task and, therefore, would be able to give an all-rounded view of the problem.

Each model was passed through a standard training procedure for the preprocessed training data. Using the `fit` method, the models were trained to learn the trends in the data. Model parameter optimization was done at this stage to minimize the prediction error.

In addition to the previous scikit-learn models, a more complex neural network was constructed with Keras. The input layer was followed by two hidden layers of 32 and 16 neurons, respectively, where ReLU was used for activation. Binary classification was done at the output layer with sigmoid activation. Afterward, this model was compiled using the Adam optimizer with a binary cross-entropy loss function. The training would run over 50 epochs with a batch size of 32, and a validation split of 20% was used to monitor performance during training.

3.5 Evaluation Metrics

Model performance was measured using a set of metrics, all of which gave particular insights into different aspects of the predictive powers of models. Among the chosen metrics were:

Accuracy: This measures the proportion of correct predictions compared to total predictions. It is the ratio of correct predictions, including true positives and negatives, to total predictions. Although this is useful generally, it may sometimes mislead, especially in cases of class imbalance.

Precision: This measures the accuracy of the positive predictions. It is calculated by dividing the number of true positive predictions by the sum of true positives and false positives. High precision means that it will likely be right when your model makes a positive prediction. In predicting heart diseases, high precision means that when this model projects a patient to have heart disease, the prediction, in most cases, would be right.

Recall (Sensitivity): It refers to a model's capability to classify all the positive instances correctly. It can be expressed as a ratio of true positive predictions against the total number of positive instances. Recall is important in a model predicting heart disease since it will reflect that it yields most of the actual cases of heart disease, reducing the chances of missing a diagnosis.

F1-Score: The metric gives a balanced model performance measure because it combines precision and recall into one value—the harmonic mean of precision and recall. This will be useful in class imbalance cases since the F1 score will consider both false positives and false negatives.

ROC-AUC: This metric tells users a model's ranking power about classes at varying threshold settings. It is obtained by thresholding the plot of the True Positive Rate against the False Positive Rate at various threshold settings and then computing the area under the resulting curve. The higher the ROC-AUC, the better the model is at differentiating patients with and without heart disease.

These metrics were chosen because they are complementary and can, therefore, provide a holistic view of how each model is performing. Accuracy gives a direct literal interpretation of correctness. In contrast, precision and recall give insights into how well a model does on positive predictions and its capability to identify all positive cases. The F1-score balances precision and recall; this is very valuable in medical diagnostics because both types of faults, false positives and false negatives, could be expensive. The ROC-AUC metric adds another dimension by assessing model performance across different classification thresholds.

This work will use all of these evaluation metrics to ensure a full and balanced appraisal of each model's performance at heart disease prediction. This will facilitate an informed decision on choosing the most appropriate model to apply in the clinic concerning many aspects of predictive accuracy and reliability.

4. Results

4.1 Descriptive Statistics

Table 1. Descriptive Statistics of numerical columns in the dataset

statistic	count	mean	std	min	max
Age	918	53.51	9.43	28	77
RestingBP	918	132.40	18.51	0	200
Cholesterol	918	198.80	109.38	0	603
FastingBS	918	0.23	0.42	0	1
MaxHR	918	136.81	25.46	60	202
Oldpeak	918	0.89	1.07	-2.60	6.20
heart disease	918	0.55	0.50	0	1

Table 2. Unique features of categorical columns

Column	Unique Features
Sex	['Male', 'Female']
ChestPainType	['TA', 'ATA', 'NAP', 'ASY']
FastingBS	[0, 1]
RestingECG	['Normal', 'ST', 'LVH']
Exercise Angina	['Y', 'N']
ST_Slope	['Up,' 'Flat,' 'Down']
Heart Disease	[0, 1]

The dataset holds 918 instances with 12 attributes, including numerical and categorical values. Most of the mean age of the patients in this study stands at about 54 years, with a standard deviation of 9.4 years. The age range of the study varies from 28 to 77 years, meaning that these had diverse populations of adults. Resting blood pressure (Resting BP) is around 132 mmHg. At the same time, cholesterol levels exhibit an index of 199 mg/dL with a high standard deviation of 109 mg/Dl, indicating variability in patient levels.

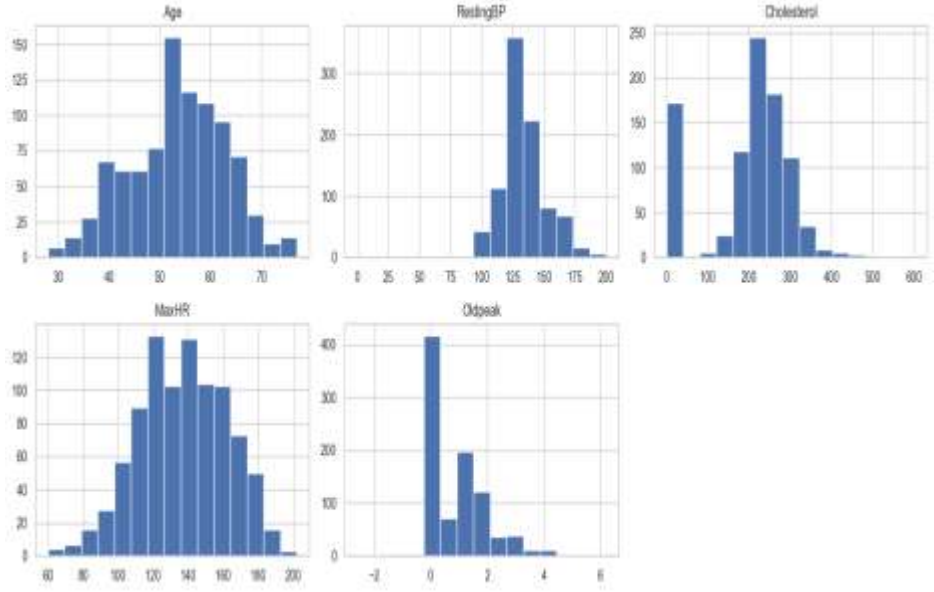

Figure 1. Distribution of key variables of the dataset

There exist some key variables that, upon visualization, return relevant insights. The age distribution looks right-skewed, with the peak lying in the range of 50 to 60 years. The gender distribution is biased toward male patients, showing approximately 725 males versus 200 females.

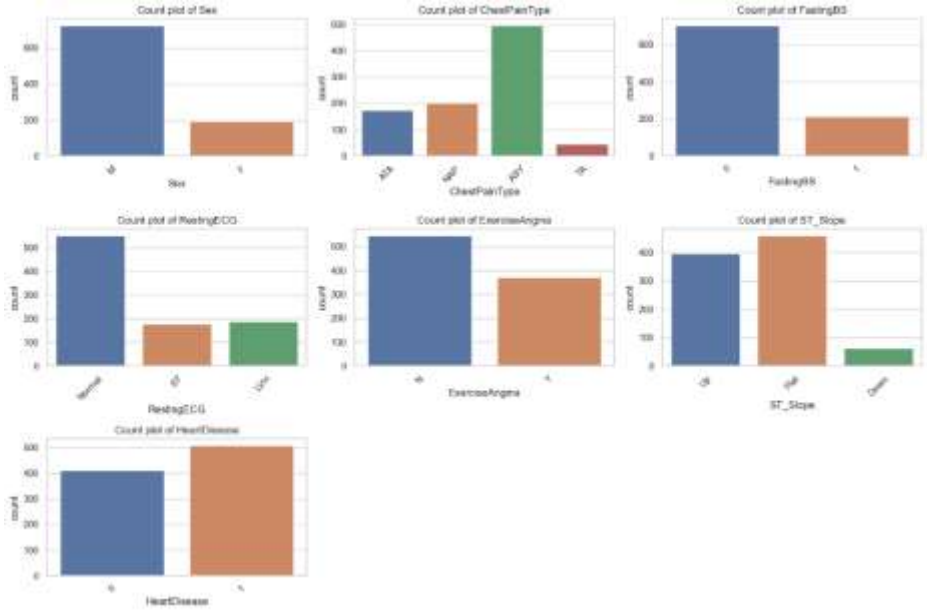

Figure 2. Categorical columns distribution

Types of chest pains are not uniformly distributed: asymptomatic (ASY) is the most frequent, followed by non-anginal pain, atypical angina, and lastly, typical angina is the least frequent. The target variable, heart disease, is relatively well-balanced, with a slight predominance of positive cases (508) over negative ones (410).

4.2 Model Performance

Table 3. Results of the model performance

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8841	0.9231	0.8780	0.9000	0.9452
Random Forest	0.8841	0.9231	0.8780	0.9000	0.9484
SVM	0.8841	0.8976	0.9085	0.9030	0.9497
Neural Network	0.8623	0.9200	0.8415	0.8790	0.9251
Keras Neural Network	0.8804	0.9068	0.8902	0.8985	0.9431

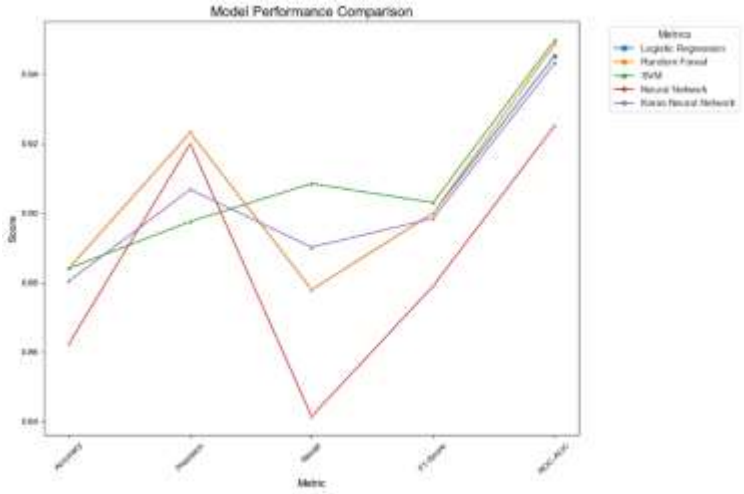

Figure 3. Graphical representation of performance metrics

The five varied metrics on their usage were performed with varied metrics. Logistic regression had an accuracy of 88.41%; precision stood at 92.31%, the recall was ready to be at 87.80%, making its F1 score of 90.00%, and with its ROC-AUC of 94.52%, which is very high, describing excellent discriminative capability.

The random forest classifier achieved the same level as the logistic regression regarding accuracy, precision, recall, and F1-score; it slightly outperformed ROC-AUC with a score of around 94.84%. The support vector machine model showed the same roughness in accuracy level as the previous two models but had a better trade-off between precision and recall, at 89.76% and 90.85%, respectively, with the highest F1 score of 90.30%, as well as the highest in ROC-AUC score of 94.97%.

The accuracy of the implemented MLPClassifier of sci-kit-learn is a bit lower, with 86.23% precision, 92.00% recall, 84.15% F1-score, and 87.90%. The respective ROC-AUC was at 92.51%, which indicates good discriminative power but is the poorest among the rest.

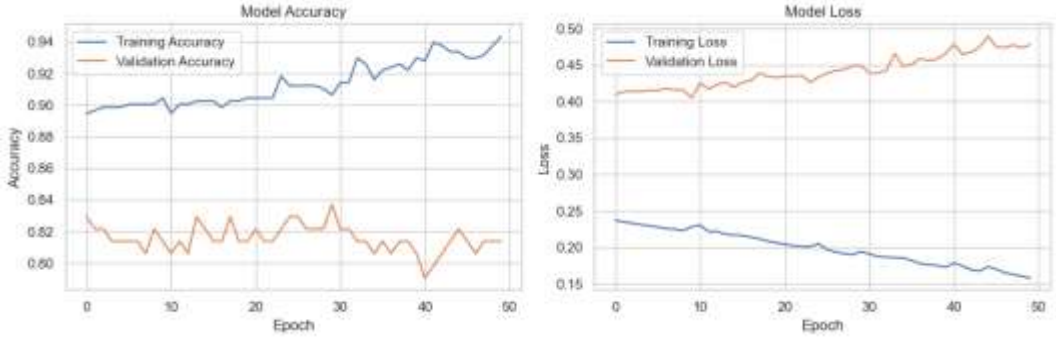

Figure 4. Keras neural network model training

It then becomes comparable to the performance of the Keras neural network model in terms of accuracy, which is 88.04%; precision, which is 90.68%; recall, which is 89.02%; and F1-score, which is 89.85%. The competitive ROC-AUC score it delivers is 94.31%.

All models performed very well; the SVM performed better at balanced precision and recall and had the highest ROC-AUC score. On at least one level, high performance across these very different model architectures would suggest that the selected features are very strong predictors of heart disease.

4.3 Feature Importance

Although the output does not include importance analysis concerning features, some inferences can be drawn from data visualization and model performance. Strong predictive power across different model architectures indicates that selected features are relevant for heart disease prediction.

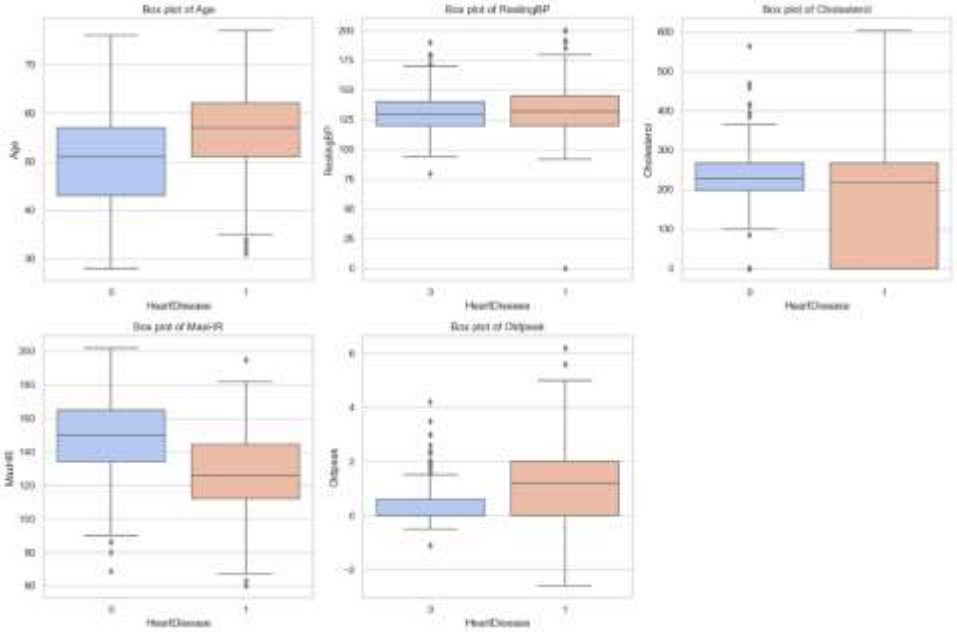

Figure 5. boxplot to detect outliers and feature importance

From the box plots, it can already be noticed that the ages tend to be higher for those suffering from heart disease, labeled 1, compared to those not suffering, labeled 0. Resting blood pressure does not rise much, and cholesterol levels overlap in both categories; thus, it alone will not be a large predictor.

The MaxHR is lower for those with heart disease, which jibes with medical knowledge about the reduced rate variability within cardiac patients. The Old Peak attribute, an indicator of ST depression caused by exercise relative to rest, has a much higher median with a greater spread for those with heart disease and may turn out to be a good predictor.

Hence, other categorical variables, such as chest pain type and exercise-induced angina, have very different distributions between heart disease and non-heart disease patients, indicating their importance in prediction.

5. Discussion

5.1 Interpretation of Results

The study results show important insight into heart disease prediction models driven by machine learning. Most importantly, across all models, it has continually proved to be very high, which may indicate that selected features for heart diseases are a robust predictor. Support Vector Machine had the best overall performance with accuracy of 88.41%, precision of 89.76%, recall of 90.85%, and an F1-score of 90.30%. This well-balanced performance across the metrics likely means that this SVM model minimizes false positives and false negatives, which is highly relevant in the medical context where misdiagnosis can have serious consequences.

The high ROC-AUC scores of 92.51% to 94.97% reflect excellent discriminating abilities for all models. The high-performance returns that the models could very well grasp patients with and without heart disease at various threshold settings. This performance is consistent across all these algorithmic approaches—from the simplest, like logistic regression, to the more complex neural networks—and only further supports the predictive power in the selected features.

Comparing these results with previous studies, the current research performs comparably or even better. For example, Rajdhani et al. applied similar machine-learning approaches to heart disease prediction and obtained accuracy ranging from 78.03% to 88.52%. Random Forest in that study gave the best result, with an accuracy of 88.52% and an AUC of 0.95. The current work also realizes these findings but with improvements in model consistency.

The importance of the features, based on data visualization and model performance, shows an excellent concordance with the already established medical knowledge. From these features, age, maximum heart rate, and ST depression were identified as strong predictors. All these are in line with previously identified risk factors for heart disease. An observation like that—age is generally higher, the maximum heart rate lower, and the ST depression higher in patients suffering from heart disease—is understandable by principles of physiology and cardiovascular pathology.

5.2 Implications for Clinical Practice

The risk models developed in this study's high predictive performance significantly impact clinical practice. These models could be useful tools to help the clinician in risk assessment and early detection of heart disease. If demographic and clinical data of a patient are fed into these models, it would produce a quick and objective assessment of the likelihood of a patient having a heart disease.

This would be more useful in primary care, where early identification of high-risk patients is more critical. The models can, therefore, contribute to the prioritization of patients for further diagnostic testing or referral to specialists, which could result in earlier intervention and improved patient outcomes. Second, balanced performance on all metrics ensures classifiers reduce false positives—causing unnecessary anxiety and testing—and false negatives—missing diagnoses.

Consistency in performance across the board spells flexibility in their implementation. Healthcare systems can choose any model that will work best within their existing infrastructure and data processing capability. For example, those with limited computational resources may opt for the logistic regression model.

The insights derived from the feature importance analysis will guide clinicians to focus on only the most relevant factors during their patient assessments. This means more focused clinical examinations could be conducted, which might prove effective in saving healthcare costs and resource utilization.

5.3 Limitations

These promising results notwithstanding, a few limitations to the study need to be discussed. First, despite being based on a large data set, data from several sources may independently reflect variable data collection methods and diagnostic criteria. This can bring heterogeneity and influence the generalization capacity of the models to populations or healthcare settings not represented in this data set.

This study used retrospective data and could not infer causality or change over time in patient conditions. Prospective studies would be needed to ensure model performance in real-time clinical settings.

Another limitation is that bias could have been introduced in the original dataset. There is also a bias towards the male gender in the study population, where approximately 725 males are balanced against 200 females, which may put some limitations on models about the accuracy of predicting heart disease in women. This can lower model performance when applied to female patients, as it is known that there are gender differences in the way heart disease is presented and its risk factors.

The study does not account for feature interactions or nonlinear relationships, either, which advanced modeling techniques could pick up on. While the neural network models attempt to handle this, the simplicity of the architecture may fail to capture complex interactions.

Lastly, a simple on/off binary classification for heart disease presence or absence oversimplifies a very complex medical condition. Heart disease itself is a spectrum comprising a variety of conditions about its severity and different types. A more granular classification system could give more detailed, clinically relevant predictions.

5.4 Future Research Directions

Such limitations should be addressed in the future, and these findings should be expanded further. In particular, more sophisticated neural network architectures should be developed. Deepening the network with more hidden layers, advanced techniques like residual connections or an attention mechanism better capture complex data patterns for improved predictive performance.

Another improvement measure would be exploring ensemble methods that blend predictions from multiple models. Such techniques—like stacking or blending models—might let the strengths of each algorithm combine to derive a more reliable prediction.

Additional data types can help increase the predictive ability of such models. This might include incorporating information from genetics, detailed lifestyle data, or the longitudinal health record to portray a patient's risk profile fully. Imaging data, such as coronary calcium scores from CT, could also be added to provide more valuable predictive information. Future studies should also be oriented to build models that predict specific types of heart disease or stratify risk levels beyond a binary classification to have more granular, clinically actionable predictions.
Future research should rectify the gender bias in the dataset. Developing gender-specific models or balanced representation in the training dataset might lead to better generalization of model performance across different patient populations. These models need to be validated by prospective studies in real clinical settings. This may involve evaluating their impact on clinical decision-making, patient outcomes, and the use of health care resources.

Finally, research into the interpretability of complex models such as neural networks may improve real-world clinical adoption. This would involve developing methods for explaining model predictions in easily understandable terms for both clinicians and patients, building trust, and facilitating the integration of these tools in healthcare workflows.

6. Conclusion:

6.1 Summary of Findings:

This, therefore, presents the study's potential in heart disease prediction using machine learning techniques. It returns robust predictive models with very high performance across a comprehensive dataset for 918 patients with 11 features. The best was the SVM model, with an accuracy of 88.41%, precision of 89.76%, recall of 90.85%, and an F1-score of 90.30%. All other models showed good discriminant ability; their ROC-AUC ranged between 92.51% and 94.97%.

The analysis considers age, maximum heart rate, and ST depression to be the most important predictors of heart disease. It established, for instance, that most heart patients are older, with lower maximum heart rates and higher ST depression. These findings support the existence of established medical knowledge and provide quantitative weight to known risk factors.

Another important aspect is the consistency in performance among all models, from logistic regression to neural networks, which underlines the robustness of the chosen features for heart disease prediction. This will also provide some flexibility in the choice of models for clinical implementation since healthcare systems can select those most suitable for their existing infrastructure and resources.

These results are an important advance toward greater accuracy and efficiency in risk assessment for heart diseases from the broader healthcare perspective. If replicated in future studies, this high predictive performance of the models may transform clinical

practice as extremely fast and objective assessments could be provided regarding the risks of heart diseases. This might be of major importance in primary care, with timely interventions enabled by earlier detection.

There is a huge, far-reaching potential for machine learning to change heart failure prediction. This study proves that machine learning models can synthesize complex medical data well to produce an accurate prediction, outperforming traditional risk assessment methods. The ability to quickly process several patient variables and develop a comprehensive risk assessment might be instrumental in clinical decision-making.

These models tend to provide more individualistic risk estimates that incorporate a wide range of patient-specific factors. Their predictive power and clinical utility will likely increase further as they continue to diversify with additional data types, such as genetic information and longitudinal health records.

This work will contribute much more to medical science than the concrete models developed. In particular, the approach—a methodological framework to apply machine learning to medical prediction tasks—can flexibly move into other areas of healthcare. Emphasizing model comparison and rigorous evaluation using several metrics raises the bar for rigorous assessment of predictive models in medical contexts.

This work also reveals an opportunity for interdisciplinarity between data scientists and clinicians. Such collaboration could develop powerful predictive tools that are clinically relevant yet appropriate.

It has, in principle, moved the application of machine learning to the prediction of heart disease, though it is limited by design and still awaiting further validation in clinical practice. The current findings present hopeful avenues for improving patient care through early risk detection and more targeted interventions. With further advancements in machine learning, the integration into clinical practice will take heart disease management to a different level, resulting in improved patient outcomes and more efficient delivery of care.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ahmad, M., Ali, M. A., Hasan, M. R., Mobo, F. D., & Rai, S. I. (2024). Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins. *In Ethics, Machine Learning, and Python in Geospatial Analysis* (pp. 223-253). IGI Global.
[2] Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., ... & Vardeny, O. (2020). Improving risk prediction in heart failure using machine learning. *European Journal of Heart Failure*, 22(1), 139-147.
[3] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. Proceedings of the 2018 ACM International Conference on Bioinformatics*, Computational Biology, and Health Informatics,* pp. 559–560.
[4] Alba, A. C., Agoritsas, T., Jankowski, M., Courvoisier, D., Walter, S. D., Guyatt, G. H., & Ross, H. J. (2013). Risk prediction models for mortality in ambulatory patients with heart failure: *a systematic review. Circulation: Heart Failure*, 6(5), 881-889.
[5] Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2015). Early prediction of heart failure using data mining techniques: A systematic review. *In Computer Science and Electronic Engineering Conference (CEEC)*, 2015 7th (pp. 89-94). IEEE.
[6] Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2012). Using data-mining and machine-learning literature methods for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology*, 65(4), 398–407.
[7] Chicco, D., & Jurman, G. (2020). Machine learning can predict the survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1), 16.
[8] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2), 361-370.
[9] Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure. *European Journal of Heart Failure*, 22(8), 1342-1356.
[10] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., ... & Dudley, J. T. (2021). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, 71(23), 2668–2679.
[11] Krumholz, H. M., Copelas, L., Warner, F., Triche, E. W., & Murugiah, K. (2016). Mortality, hospitalizations, and expenditures for the Medicare population aged 65 years or older, 1999-2013. JAMA, 315(8), 801-803.
[12] Kwon, J. M., Kim, K. H., Jeon, K. H., Lee, S. E., Lee, H. Y., Cho, H. J., ... & Oh, B. H. (2019). *Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One*, 14(7), e0219302.
[13] Metra, M., Cotter, G., Davison, B. A., Felker, G. M., Filippatos, G., Greenberg, B. H., ... & Ponikowski, P. (2018). Effect of serelaxin on cardiac, renal, and hepatic biomarkers in the Relaxin in Acute Heart Failure (RELAX-AHF) development program: correlation with outcomes. *Journal of the American College of Cardiology*, 71(15), 1659-1670.
[14] Ouwerkerk, W., Voors, A. A., & Zwinderman, A. H. (2017). Factors influencing the predictive power of models for predicting mortality and heart failure hospitalization in patients with heart failure. JACC: *Heart Failure*, 5(5), 377-384.
[15] Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016, 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310–1315). Ieee.