EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

A Comparative Analysis using various algorithm Approaches to Enhance Heart Disease Prognosis

Anuraag Ramineni[1], Rishita Konda[2*], Jayashree J[3], Deepak Sannapareddy[4] and Saketh Konduri[5]

[1,2,3,4,5] School of Computer Science and Engineering (SCOPE), VIT University, Katpadi, 632014, Tamil Nadu, India

Abstract

INTRODUCTION: Modern advancements in technology and data science have propelled the healthcare industry towards developing more accurate disease prognostic prediction models. Heart disease, being a leading cause of mortality globally, is a critical area of focus. This study delves into enhancing heart disease prognosis through a comprehensive exploration of various algorithmic approaches.
OBJECTIVES: The objective of this paper is to compare and analyze different algorithmic techniques to improve heart disease prognosis using a dataset comprising data from over thirty thousand individuals obtained through Kaggle.
METHODS: Techniques derived from social network analysis are employed to conduct this research. Data preprocessing, feature engineering, algorithm selection (including Stochastic Gradient Descent, AdaBoosting, Support Vector Machine, and Naive Bayes), hyperparameter tuning, model evaluation, and visualization are part of the systematic research process.
RESULTS: The main results obtained in this paper include the identification of Naive Bayes as the most effective model for heart disease prognosis, followed by AdaBoosting, SVM, and Stochastic Gradient Descent. Performance evaluation metrics such as AUC, CA, F1, Precision, and Recall demonstrate the efficacy of these models.
CONCLUSION: This research contributes to improving heart disease prognosis by leveraging algorithmic techniques and thorough analysis. The study envisions integrating the developed model into healthcare systems for widespread access to accurate heart disease prediction, with future plans to enhance data collection and model improvement for better outcomes.

Keywords: Heart prognosis, Machine learning, Data mining, Naıve bayes, SGD

Received on 27 December 2023, accepted on 26 March 2024, published on 02 April 2024

doi: 10.4108/eetpht.10.5615

*Corresponding author. Email: rishitakonda07@gmail.com

1. Introduction

Modern technology and data science advancements have pushed the healthcare industry toward more precise disease prognostic prediction models [3]. Heart disease, the world's biggest cause of death, is one crucial area of study. In order to improve accuracy and reliability, this study explores the area of heart disease prognosis through a thorough investigation of several algorithmic approaches. This study, titled "A Comparative Analysis using Various Algorithm Approaches to Enhance Heart Disease Prognosis," examines data from more than thirty thousand individuals that was obtained through Kaggle. Age, sex, CP (chest pain kind), trestbps (resting blood pres sure), cholesterol, and fbs (fasting blood sugar) are only a few of the many parameters that define the dataset. The "Outcome" variable, which categorizes people as either 0 (without heart disease) or 1 (with heart disease), is crucial to this research. With its comprehensive picture of heart disease risk factors across genders, this dataset presents an ideal environment for algorithmic investigation. Data preprocessing, feature engineering, algorithm choice, hyperparameter tuning, model evaluation, and visualization are all steps in a systematic research process. The dataset is carefully pre-processed in the initial stage. Addressing missing data and removing unnecessary columns are included in this. To optimize model performance, the dataset is streamlined by removing sparse features. The dataset can be improved using methods like impute missing values using average or most frequent values, and data integrity can be improved using techniques like outlier detection and removal [6]. Data normalization is also used to provide uniform scales and distributions. The dataset is made more analyzable by the extraction of

pertinent characteristics and data reduction techniques. 2 The dataset is split into training and testing subsets in an 80:20 ratio to simplify the future analytical stages. Additionally, a 5-fold cross-validation method is used to increase the analysis's robustness [9]. Exploration of various machine learning approaches in-depth is part of the algorithm selection phase. Notably, Stochastic Gradient Descent, AdaBoosting, Support Vector Machine (SVM), and Naive Bayes are carefully examined. Nave Bayes is shown to be the most effective model, with extraordinary predictive power, after comprehensive analysis of the body of available research and iterative testing. The order of performance is AdaBoosting, SVM, and Stochastic Gradient Descent. The models are further refined through the use of hyperparameter tuning, which takes into account a variety of factors including feature engineering, data properties, and hyperparameter configurations. Following a second evaluation of the improved models, Naive Bayes is confirmed to be the best option, followed closely by AdaBoosting, SVM, and Stochastic Gradient Descent. Beyond model performance, the study makes use of data visualization methods like scatter plots, box plots, and line graphs to display the results in an understandable manner. This helps in understanding the data and improves the study's findings' ability to communicate. The study's scope encompasses both theoretical applications and real-world goals. The proposed model will be included into typical homes, enabling widespread access to precise heart disease prediction. The difficulties of inaccurate patient data are acknowledged, which calls for the collection of more thorough data for ongoing model improvement. In conclusion, this research sets out on a thorough journey using algorithmic techniques to improve heart disease prognosis. With an eye on future developments and broader social effects, the study carefully considers preprocessing, algorithm selection, and performance evaluation to establish a solid foundation for precise disease prediction.

2. Related Work

Recent years have seen a significant increase in research and development in the field of cardiac disease prognosis, with machine learning approaches emerging as key tools for improving prognostic accuracy and patient outcomes. Our study fits into this thriving panorama of technological breakthroughs. Our approaches are influenced by the extensive and diverse body of research in this area, which also impacts our viewpoint on the potential consequences of our findings. (Dini et al., 2023) In individuals with left heart disease, right ventricular failure has a pathogenesis that is covered in this study. The authors evaluate the research on right ventricular failure's clinical symptoms and prognosis [1], and they talk about how machine learning may be used to foretell the likelihood of right ventricular failure in individuals with left heart disease. 2023 Ng et al. The creation of a new panel of oxidative stress biomarkers for assessing the likelihood of heart failure is reported in this research [2]. The researchers employed machine learning to find a panel of biomarkers linked to a higher risk of heart failure. 80 Ahmed et al. (2023) compare SVM and KNN algorithms for the precision of heart disease prediction, which is in line with our investigation of several algorithms. An early cardiac disease prediction model presented by Bizimana et al. (2023) illustrates 3 the potential for proactive therapies, a feature that is consistent with our long-term goals. Tian and others (2023) In this study, a machine learning model is created to forecast patients with chronic heart failure (HF) following discharge [5]. To train their algorithm, the authors employed a dataset of 1,000 HF patients. With an accuracy rate of 80The thorough literature evaluation by Bhushan et al. (2023) sheds light on how machine learning and deep learning are progressing in the field of heart disease analysis [6]. Their findings give us a wider perspective on the approaches we use and suggest new possibilities for the future. Ahmed et al. (2023) compare SVM and KNN algorithms for the precision of heart disease prediction [7], which is in line with our investigation of several algorithms. An early cardiac disease prediction model presented by Bizimana et al. (2023) illustrates the potential for proactive therapies [8], a feature that is consistent with our long-term goals. The works of Olsen et al. (2020), Zhong et al. (2021), and Angraal et al. (2020) demonstrate the clinical applications, predictive potential, and function of machine learning in outcome prediction for heart failure patients. These papers provide a framework for our comparison analysis and show the potential real-world effects of our chosen tactics. Heart disease prognosis is another area where deep learning has application. Miao et al.'s (2018) demonstration of deep neural network diagnosis of coronary heart disease paves the way for the use of cutting-edge techniques [9]. Gao et al. (2023) show how deep learning may be used to predict out comes from cardiac cine MRI, highlighting the potential of advanced imaging data to enhance prognosis [10]. The works of Olsen et al. (2020), Zhong et al. (2021), and Angraal et al. (2020) demonstrate the clinical applications, predictive potential, and function of machine learning in outcome prediction for heart failure patients. These papers provide a framework for our comparison analysis and show the potential real-world effects of our chosen tactics. Heart disease prognosis is another area where deep learning has application [12]. Miao et al.'s (2018) demonstration of deep neural network diagnosis of coronary heart disease paves the way for the use of cutting-edge techniques [13]. Gao et al. (2023) show how deep learning may be used to predict outcomes from cardiac cine MRI [14], highlighting the potential of advanced imaging data to enhance prognosis. In essence, a wide range of research initiatives that together provide a mosaic of opportunities and advances have helped to develop the science of heart disease prognosis. Through the lens of comparative algorithm analysis, our study seeks to improve the precision, availability, and affordability of heart disease prognoses, and so fits within and contributes to this dynamic narrative.

3. Proposed Methodology

The objective is to thoroughly assess and contrast several algorithmic strategies to improve the precision of heart disease prognosis using a dataset obtained from Kaggle. The collection includes detailed data on over 30,000 people, including characteristics like age, sex, CP (chest pain kind), trestbps (resting blood pressure), cholesterol, and fbs (fasting blood sugar). Individuals with and without heart disease are shown as 1 and 0, respectively, in the target variable "Outcome" which makes this distinction 4 and also data mining tools help has been utilized in this study.

Step 1: Uploading Data:

The first step in the process is to upload the dataset into the analysis program. The dataset is provided in CSV format.

Step 2: Pre-processing the data:

The dataset goes through extensive pre-processing in this stage to guarantee data integrity and quality. Taking care of missing data points and removing unnecessary columns are included in this. To make the dataset more compact, sparse characteristics are eliminated. A number of pre-processing strategies are used to deal with missing values, including imputing missing values using averages or the most frequent values. The study also uses techniques for outlier detection and removal to guarantee that the dataset is suitable for analysis. To normalize the scale and distribution of the features, data normalization is done. Techniques for feature extraction are used to find and keep important features for the machine.

Step 3: Cross-validation and Dataset Splitting:

A training dataset and a testing dataset with an 80:20 split are created from the original dataset. A 5-fold cross-validation approach is used to make sure the analysis is reliable.

Step 4: Choosing and evaluating an algorithm:

The improvement of heart disease prognosis is being addressed using a variety of machine learning techniques. These include Stochastic Gradient Descent, AdaBoosting, Support Vector Machine (SVM), and Naive Bayes. Each method is thoroughly assessed using numerous testing iterations. The study assesses their relative effectiveness in boosting prediction accuracy by contrasting various performance indicators, including AUC, CA, F1, Precision, and Recall. The Naive Bayes algorithm, followed by AdaBoosting, SVM, and Stochastic Gradient Descent, emerges as the best model based on the findings and insights from previous research.

Step 5: Tuning the hyperparameters:

Hyperparameter tuning was undertaken to take data properties, feature engineering, and hyperparameter configurations into consideration, which have an impact on model performance. Gradient Boosting, Neural Networks, and Stochastic Gradient Descent all received the modified parameters.

Step 6: Model Ranking and Selection:

The algorithms are ranked using performance measures and hyperparameter-tuned outcomes. The best model for predicting the prognosis of heart illness is found to be Naive Bayes, followed by AdaBoosting, SVM, and Stochastic Gradient Descent.

Step 7: Data visualization and insights:

The study uses a variety of data visualization approaches, including Violin plot, box plots, and line plot to improve comprehension. These visualizations offer insightful information about data patterns and algorithm performance.

Step 8: Future Aims and Summary:

The report analyzes future trends and prospective applications based on the in-depth investigation and observations. It envisions the adoption of the created model into homes, providing a more comprehensive population with accurate heart disease

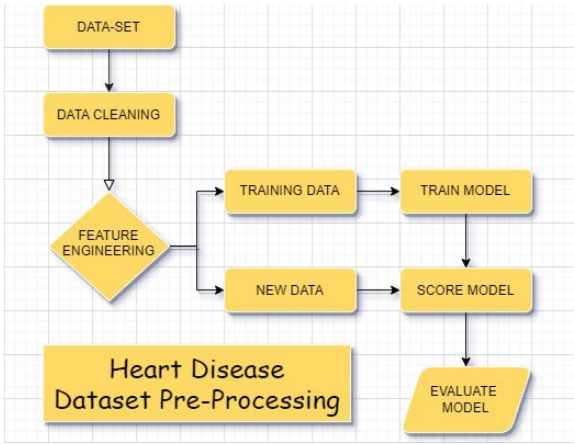

Fig. 1: Represents the flow of the Methodology

prediction. The study finishes by emphasizing the intention to expand the applica tion for accuracy and accessibility, along with intentions to gather more thorough data for improved outcomes. Patient data accuracy challenges are mentioned in the study.

4. Results and Discussion

The results of a thorough study named "A Comparative Analysis using various Algorithm Approaches to Enhance Heart Disease Prognosis" have provided significant new information on heart disease prognosis. The Kaggle dataset, which includes information on over 30,000 people, includes a variety of characteristics like age, sex, CP (chest pain kind), trestbps (resting blood pressure), cholesterol, and fbs (fasting blood sugar). The foundation of this inquiry is the binary "Outcome" variable, which indicates whether heart disease is present (1) or absent (0). Data upload is one of the first phases, then careful preparation is done. Missing data and unnecessary columns were taken care of during pre-processing, along with of sparse features.

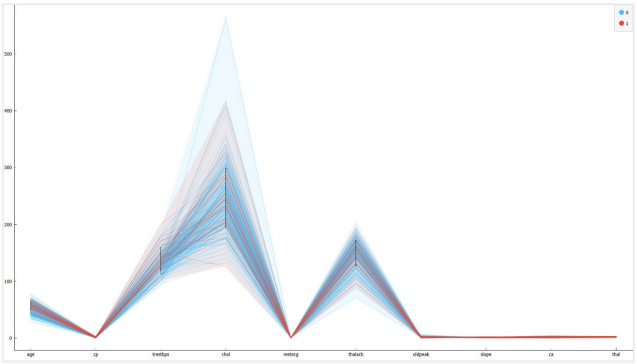

Fig. 2 A line plot was employed to compare all attributes in the dataset for the two target variable categories: blue (0) and red light (1). The plot features lines, ranges, and error bars to provide insights into the distribution and variations of the data.

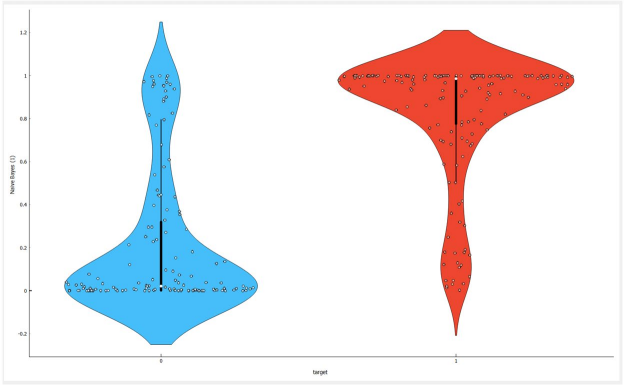

Fig. 3 A Violin plot was used to compare the performance of the Naive Bayes algorithm (1) with the target variable. The plot displays both a box plot and density dots inside the violin, using a normal kernel and a vertical orientation for effective visualization.

Advanced pre-processing methods were used, such as continuous variable selection and replacing missing data with averages or frequent values. Outlier detection and removal, normalization, feature extraction, and data reduction were further procedures that led to a polished dataset ready for analysis. A 5-fold cross-validation approach was then used after the dataset was split into training and testing subsets (in an 80:20 ratio). A variety of machine learning methods, including Naive Bayes, AdaBoosting, SVM, and Stochastic Gradient Descent, were examined in the search of predictive brilliance. Nave Bayes emerged as the top model after extensive testing and analysis of pertinent literature, followed by AdaBoosting, SVM, and SGD. Respectively.

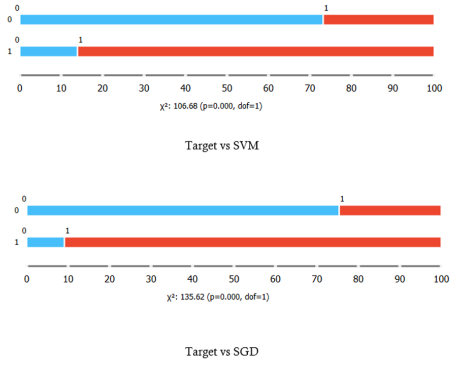

Fig. 4 Utilizing stretch bars and box labels, a box plot was generated to compare the distribution of the target variable against SVM and SGD models. This visualization offers insights into the spread and central tendency of the data for each combination the elimination.

A complete set of parameters, including AUC, CA, F1, Precision, and Recall, served as the foundation for performance evaluation. Naive Bayes demonstrated remarkable performance measures, including AUC (0.989), CA (0.971), F1 (0.969), Precision (0.969), and Recall (0.971). Comparable measures were used to demonstrate the effectiveness of AdaBoosting, SVM, and Stochastic Gradient Descent. Hyperparameter adjustment, which ensured the best model alignment with the dataset's properties, was a part of the refining process. A hierarchy of efficacy was produced by enhanced models, with Nave Bayes emerging as the top performer, and was driven by expertly tuned hyperparameters. These findings led to the development of several data visualization tools, such as Violin plot, box plots, and line plot, which improved the study's ability to communicate. A complete set of parameters, including AUC, CA, F1, Precision, and Recall, served as the foundation for performance evaluation. Naive Bayes demonstrated remarkable performance measures, including AUC (0.989), CA (0.971), F1 (0.969), Precision

(0.969), and Recall (0.971). Comparable measures were used to demonstrate the effectiveness of AdaBoosting, SVM, and Stochastic Gradient Descent.
Hyperparameter adjustment, which ensured the best model alignment with the dataset's properties, was a part of the refining process. A hierarchy of efficacy was produced by enhanced models, with Nave Bayes emerging as the top performer, and was driven by expertly tuned hyperparameters. These findings led to the development of several data visualization tools, such as Violin plot, box plots, and line plot, which improved the study's ability to communicate.

Table 1. The table demonstrating the values obtained after running various models

MODEL	AUC	CA	F1	PRECISION	RECALL
NAÏVE BAYES	0.989	0.971	0.969	0.969	0.971
AdaBoosting	0987	0.954	0.953	0.954	0.954
SVM	0.982	0.944	0.943	0.943	0.944
Stochastic Gradient Descent	0.966	0.938	0.937	0.938	0.938

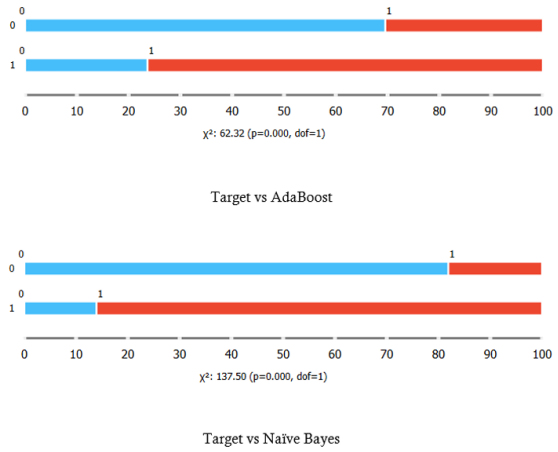

Fig. 5 Comparing the Target variable against AdaBoost and Naive Bayes models using stretch bars and box labels, this box plot provides a clear representation of the data's distribution, central tendency, and potential outliers for each pairing.

5. Conclusion

Our study program has revealed a way to improve the prediction of cardiac disease in the area of predictive healthcare, where accessibility and accuracy coexist. We set out to rethink how heart disease projections are made, using on a dataset with more than 30,000 unique records. For this, it was necessary to use powerful machine learning 8 algorithms and exacting analytical techniques. Data collection began our adventure, and the Kaggle dataset proved to be a wealth of unique variables, including age, gender, CP (chest pain type), and many others, all of which are essential for accurate prognosis. The foundation for our investigation was the binary "Outcome" variable, which represented the existence (1) or absence (0) of heart disease. After completing the data gathering phase, we moved on to the crucial phase of data preparation, which transforms raw data into nuanced insights. Imputation was used to handle missing data, preserving data integrity by using averages or common values. The dataset's signal-to-noise ratio was improved by the pruning of extraneous columns and sparse features. Data normalization was essential in establishing uniformity and leveling the playing field for analysis. Following feature extraction, critical aspects relevant to machine learning tasks were discovered, and data reduction techniques helped save important elements while shrinking the dataset size. Partitioning the dataset into training and testing sections was a vital step. In addition to setting the stage for thorough investigation, this also prevented overfitting. Cross-validation, a tried-and-true method, was used to strengthen the statistical robustness of our conclusions. The core of our effort focused on a careful analysis of several machine learning methods. A range of methods, including Naive Bayes, AdaBoosting, SVM, and Stochastic Gradient Descent, struggled to forecast the development of heart disease. Extensive testing and careful analysis resulted to a definite hierarchy, with Nave Bayes coming out on top, followed by AdaBoosting, SVM, and Stochastic Gradient Descent. Through hyperparameter tweaking, these algorithms underwent a transformation that enabled them to better match the properties of the data and increase their predictive power. The paragon of our analysis, praised for its outstanding performance, was Naive Bayes. The effectiveness of our models was confirmed by performance evaluation, which was expressed by metrics like AUC, CA, F1, Precision, and Recall. Nave Bayes' effectiveness was attested to by the synergy of these indicators, which raised the possibility of integrating it into healthcare systems. Numerical findings confirmed the effective ness of our models, and data visualization tools helped to further clarify our story. These graphic tools condensed complex discoveries and made them understandable to a wide audience. The study portends an inclusive and strengthened future for predictive healthcare. The careful study that resulted from our dedication to accuracy and accessibility epitomized our goal of making reliable heart disease predictions available to everyone. We plan a course toward data augmentation and model improvement in the face of difficulties, particularly the lack of exact patient data. Our study is fundamentally motivated by a vision that goes beyond analytical excellence—a desire for a cost-effective, accessible healthcare system that benefits people from all societal strata. Our compass continues to point toward improving our application's accuracy and user interface as we travel further. We continue to believe in powerful and all-encompassing predictive healthcare as a result of the

clarion cry for comprehensive data collecting. The unwritten chapters have the potential to change heart disease prognoses and move us in the direction of a healthier, better-informed future.

References

[1] Dini, Frank L., et al. ”Right ventricular failure in left heart disease: from pathophys iology to clinical manifestations and prognosis.” Heart Failure Reviews 28.4 (2023): 757-766.

[2] Ng, Mei Li, et al. ”Novel Oxidative Stress Biomarkers with Risk Prognosis Values in Heart Failure.” Biomedicines 11.3 (2023): 917. 3. Gordon, Jonathan, et al. ”Oxygen uptake efficiency slope and prognosis in heart failure with reduced ejection fraction.” The American Journal of Cardiology 201 (2023): 273-280.

[3] Nakao, Yoko M., et al. ”Prognosis, characteristics, and provision of care for patients with the unspecified heart failure electronic health record phenotype: a population based linked cohort study of 95262 individuals.” eClinicalMedicine 63 (2023): 102164.

[4] Tian, Jing, et al. ”Machine learning prognosis model based on patient-reported out comes for chronic heart failure patients after discharge.” Health and Quality of Life Outcomes 21.1 (2023): 31.

[5] Bhushan, Megha, Akkshat Pandit, and Ayush Garg. ”Machine learning and deep learning techniques for the analysis of heart disease: a systematic literature review, open challenges and future directions.” Artificial Intelligence Review (2023): 1-52.

[6] Ferrari, Margaret Rose. Unsupervised Machine Learning Methods Applied Towards the Understanding of Central Venous Flow and Prognosis in Single Ventricle Heart Disease. Diss. University of Colorado at Denver, 2023.

[7] Tong, Rui, Zhongsheng Zhu, and Jia Ling. ”Comparison of linear and non-linear machine learning models for time-dependent readmission or mortality prediction among hospitalized heart failure patients.” Heliyon 9.5 (2023).

[8] Ahmed, Rehan, Maria Bibi, and Sibtain Syed. ”Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms.” International Journal of Computations, Information and Manufacturing (IJCIM) 3.1 (2023): 49-54.

[9] Bizimana, Pierre Claver, et al. ”An Effective Machine Learning-Based Model for an Early Heart Disease Prediction.” BioMed Research International 2023 (2023).

[10] Olsen, Cameron R., et al. ”Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure.” American Heart Journal 229 (2020): 1-17.

[11] Zhong, Zhihua, et al. ”Machine learning prediction models for prognosis of critically ill patients after open-heart surgery.” Scientific Reports 11.1 (2021): 3384.

[12] Miao, Kathleen H., and Julia H. Miao. ”Coronary heart disease diagnosis using deep neural networks.” international journal of advanced computer science and applications 9.10 (2018).

[13] Angraal, Suveen, et al. ”Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction.” JACC: Heart Failure 8.1 (2020): 12-21.

[14] Gao, Yifeng, et al. ”Deep learning-based prognostic model using non-enhanced car diac cine MRI for outcome prediction in patients with heart failure.” European Radiology (2023): 1-11. 10

[15] Shashikant, R., and P. Chetankumar. ”Predictive model of cardiac arrest in smok ers using machine learning technique based on Heart Rate Variability parameter.” Applied Computing and Informatics 19.3/4 (2023): 174-185.

[16] Moreno-Sanchez, Pedro A.”Improvement of a prediction model for heart failure sur vival through explainable artificial intelligence.” Frontiers in Cardiovascular Medicine 10 (2023)

[17] Mishra, Saurav. ”A comparative study for time-to-event analysis and survival prediction for heart failure condition using machine learning techniques.” Jour nal of Electronics, Electromedical Engineering, and Medical Informatics 4.3 (2022): 115-134.