

Heart Failure Prediction Using Feature Selection Methods and Machine Learning Algorithms

Swan Mollick^{1*}, Najnine Afroz Estina¹, Anindya Mozumder²

¹Department of Computer Science and Engineering, North Western University, Khulna-9100, Bangladesh

²Department of Software Engineering, Hubei University of Technology, Hubei 430068, China

Email: swanmollick@gmail.com, najnineafrozestina@gmail.com, anindya5636@gmail.com

Abstract—When the heart is unable to pump blood as effectively as it ought to, heart failure occurs. Breathing difficulty is frequently brought on by this because blood frequently backs up and fluid can accumulate in the lungs. It is one of the major killers in the modern world at this moment. Heart failure prediction is a substantial challenge for clinical data analysis. Heart failure is a complex medical condition that might be difficult to predict using traditional risk assessment tools. Given the significant morbidity and mortality associated with heart failure, more precise methods of disease prediction are required. The first problem that has to be overcome is that one should be alerted far enough in advance to take precautions. As a result, we propose a comparative machine learning framework based on eight distinct classification algorithms, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB), and Bernoulli Naive Bayes (BNB). This study also emphasizes data visualization to spot outliers, trends, and patterns in a sizable data collection. For ML, Scikit-learn and Python are utilized. Model training is carried out using Python and Scikit-learn. Among all models, SVM achieved a test recall of 0.92 using 10-fold cross-validation and a test accuracy of 88.59% using the holdout method.

Index Terms—Heart failure, Chi-squared test, Predictive analysis, Machine learning, Feature extraction, SVM, RF.

I. INTRODUCTION

Heart is most crucial for human health. Yet, heart disease has emerged as a prominent killer of both men and women in the modern world. A healthy life requires the early identification of heart failure. Aging, genetic susceptibility, and previous heart disease all have an impact on the complex clinical condition known as heart failure [1]. Heart failure (HF) is a crippling illness in which the heart is unable to pump enough blood to satisfy the body's needs. New methodologies are needed for risk estimation and the creation of efficient ways to mitigate it. Modern methods include measuring cholesterol and blood pressure, evaluating symptoms, and conducting tests, but they are costly and ineffectual.

Every year approximately 287,000 people are dying because of HF [2]. According to studies, metabolic abnormalities and early life variables including diet, alcohol consumption, and smoking, all have a big influence on how likely someone is to develop heart failure. Heart failure can be predicted

and prevented by evaluating risk factors with simple and economical methods [3].

When it comes to forecasting cardiac failure, machine learning provides a number of benefits. By analyzing vast volumes of data, finding patterns, and making predictions relying on such patterns, machine learning algorithms can increase the accuracy of predictions. As a result, risk factors and symptoms for heart failure may be predicted more precisely. Healthcare professionals may create more efficient preventive and therapeutic plans by using it to examine and forecast heart failure possible risks and consequences across huge populations [4]. It might greatly boost our capacity to identify and treat heart failure, which would ultimately lead to higher patient survival and better overall community health. The following are the contributions we made to this paper:

- Employing the chi-square method, we extracted features and compared observed results with expected outcomes to enhance our understanding of the data.
- To evaluate the model's performance, we utilized two different dataset splitting methods: a stratified hold-out method (70% train, 30% test) and 10-fold cross-validation, enabling us to compare and identify the most effective approach for our heart failure prediction model.
- To improve heart failure prediction, we employed a suite of eight machine learning algorithms to optimize our heart failure prediction model for enhanced accuracy and reliability.
- Our model's performance was rigorously assessed using key metrics including accuracy, precision, recall, f1-score, receiver operating characteristic (ROC) curve analysis, Mean Square Error (MSE), and Average Precision(AP), ensuring robust validation and evaluation.

This paper is organized as follows: Section 1 demonstrates the introductory part. In section 2, we discuss the related works. Section 3 describes the proposed methodology for heart failure prediction. Experimental results and analysis are discussed in section 4. Finally, in Section 5 the study has been concluded with some future remarks.

II. RELATED WORKS

Some studies have been done by researchers on the prognosis of heart failure. Some of these recent studies are discussed here.

Sahoo P. K. et al. [3] developed a robust technique in 2020 to accurately predict heart failure using datasets from the UCI repository. They used 70% data for the training set and 30% for testing. They used various ML algorithms and achieved the highest accuracy of 85.2% using the Support Vector Machine.

A radical solution based on assembling 10 individual classification algorithms proposed by Agarwal H. et al. [5] in 2021. They split the dataset into train and test data and tested different algorithms from basic classifiers to complex boosting and bagging algorithms. Their ensemble model acquired 85.2% test accuracy and 87.50% test recall.

In 2021, Krishna P. G. et al. [6] designed a web application and automated prediction system utilizing IBM auto AI service to assess the probability of cardiac failure. They used datasets and automatic AI analysis to train their model. With 87% accuracy, Gradient Boosting algorithm provided the best results. Multiple ML models with 30-day HF readmission or mortality using a linked administrative health dataset were compared by Awan S. E. et al. in 2019 [7]. They postulated that they might improve HF prediction using the conventional regression approach. Their most accurate classifier, weighted random forests, achieved an accuracy of 76.39%.

In 2020, Chicco D. et al. [8] found that ejection fraction and serum creatinine alone may predict a patient's survival in heart failure, based on 299 patients' medical records from the Faisalabad Institute of Cardiology and Allied Hospital. The best-performing ML method was the Random Forest classifier, which achieved an accuracy of 73%.

Machine learning classifiers and a correlation matrix to identify the major risk factors and produced prediction models for heart failure survival were used by Mamun M. et al. in 2022 [?]. LR, DT, SVM, XGBoost, LightGBM, RF, KNN, and Bagging were employed. LightGBM classifier achieved highest accuracy 85%.

TABLE I offers an overview of the summary of prior works.

The proposed work distinguishes itself by evaluating a wider range of classifiers, including three Naive Bayes variants, on a large, merged dataset. Unlike prior studies that employed fewer algorithms or smaller datasets, this approach leverages diverse models, 10-fold cross-validation, and multiple evaluation metrics to enable a more robust and comprehensive comparison.

III. PROPOSED METHODOLOGY

Figure 1 depicts the suggested framework of the proposed Heart failure prediction model. For any suggested system to function well, system design is a crucial development. The system's elevated degree of perspective and functionality is provided by a noteworthy level arrangement.

The working process starts with data collection. For accurate data analysis, the data flow is largely planned. The data is loaded for correlation analysis, and after analyzing the relationships among various features, a correlation score is established. The features are then prioritized according to the correlation score. The correlation matrix of the dataset is shown in Figure 2.

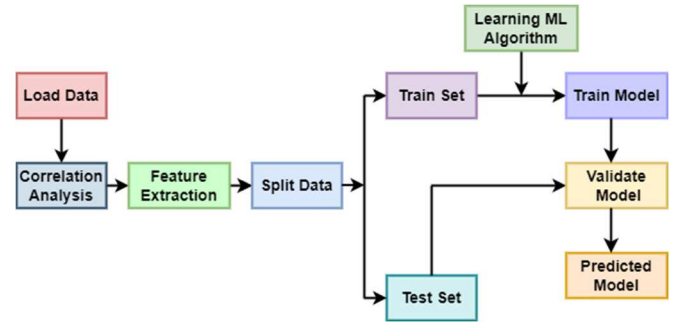


Fig. 1. The proposed system architecture for heart failure prediction, showing stages from data acquisition to model evaluation.

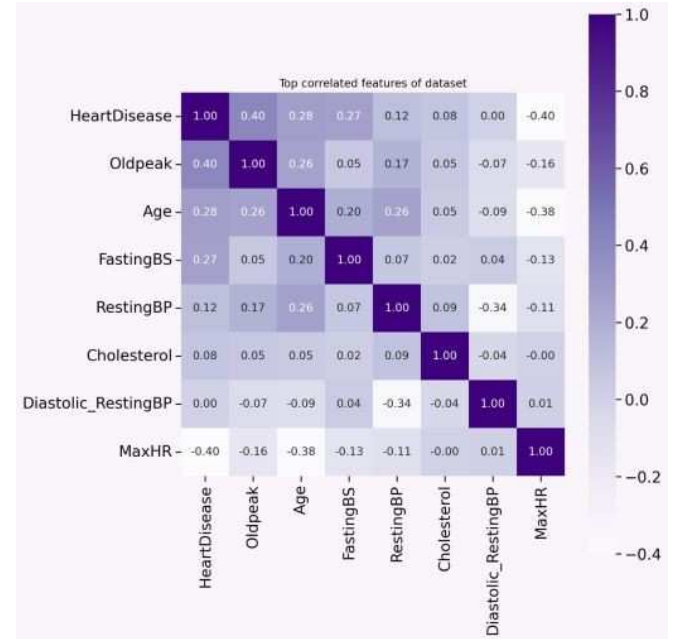


Fig. 2. Correlation matrix illustrating the relationships between dataset features, aiding in feature selection.

A. Dataset Description

We used a publicly available heart disease dataset from Kaggle [9], provided in comma-separated values (CSV) format. It was curated by the dataset publisher by merging five well-known heart disease datasets based on 12 common features, making it one of the most comprehensive resources currently available for heart failure prediction research. The dataset is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing academic use and redistribution with proper attribution. The five datasets included are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

This dataset consists of 12 attributes. They are Age,

TABLE I
SUMMARY OF PRIOR WORKS

Name	Year	Dataset Used	Method	Accuracy (%)	Advantages	Disadvantages
Sahoo P.K. et al. [3]	2020	UCI Repository	SVM, Naive Bayes, LR, DT, KNN.	85.2% (SVM)	Utilized five distinct algorithms.	The dataset used was very small.
Agrawal H. et al. [5]	2021	Kaggle	AdaBoost, DT, KNN, LR, GNB, RF, SVM.	86.40% (LR)	Applied a practical and efficient model.	Potential for further accuracy improvement.
Krishna P.G. et al. [6]	2021	IBM	GB, XGB Classifier, Extra Trees Classifier.	87.00% (GB)	Reduced time for result construction.	Accuracy may grow further.
Awan S.E. et al. [7]	2019	Hospital Morbidity Dataset	LR, RF, DT, SVM, Multilayer perceptron.	76.39% (RF)	Used a high-quality but imbalanced dataset.	Patients younger than 65 are not included.
Chicco D. et al. [8]	2020	HF patient's medical records	RF, ANN, NB, DT, GB, LR, SVM, KNN.	83.00% (LR)	Serum creatinine alone can predict a patient's outcome.	Insufficient dataset size.
Mamun M. et al. [2]	2022	UCI Repository	LightGBM, LR, XGBoost, SVM, DT.	85.00% (LightGBM)	Six different algorithms were devised.	A modest-sized dataset.

Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST-Slope, and Heart Disease. Details on these attributes with Chi-square Score are provided in TABLE II. Here *Heart Disease* is the target feature, while the remaining 11 are considered input attributes. Details on these attributes with their Chi-square scores are shown in TABLE II.

TABLE II
DESCRIPTION OF THE ATTRIBUTES OF DATASET

Feature	Description	Data Type	Chi-square Score
Age	Age of the patient	Numerical	125.37
Sex	Sex of the patient	Categorical	84.15
ChestPain Type	Chest pain type[TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	Categorical	268.07
Resting BP	Resting blood pressure [mm Hg]	Numerical	93.64
Cholesterol	Serum cholesterol [mm/dl]	Numerical	332.19
Fasting BS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	Numerical	64.32
Resting ECG	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality, LVH: showing definite left ventricular hypertrophy by Estes' criteria]	Categorical	10.93
MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]	Numerical	241.32
Exercise Angina	Exercise-induced angina [Y: Yes, N: No]	Categorical	222.26
Oldpeak: oldpeak	ST [Numeric value measured in depression]	Numerical	230.51
ST-Slope	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]	Categorical	355.92
Heart Disease	Output class [1: heart disease, 0: Normal]	Numerical	Target Feature

B. Data Preprocessing

The dataset has been pre-processed using a variety of techniques, including feature extraction and modification of categorical variables. We utilized two methods to convert categorized attributes into ordinal numeric values: OrdinalEncoder and OneHotEncoder. OneHotEncoder creates new columns for each unique group in each feature and assigns a number of 1, whereas OrdinalEncoder converts categories like ChestPainType, RestingECG, ST Slope, Cholesterol, and RestingBP into ordinal numeric values. Ordinal encoding was preferred for classifiers like Naive Bayes and SVM, which benefit from a compact representation of ordinal relationships, while one-hot encoding was used in exploratory tests where models required non-ordinal categorical treatment, to align with the strengths of each algorithm and improve compatibility and performance.

Our dataset contains a total of 1190 patient records, where 272 records are duplicates. Those 272 records have been removed from the dataset and the remaining 918 patient records are used. Our dataset has some minor imbalances. Additionally, this dataset is error-free and has no missing values.

The target variable has two classes: 0 (no heart disease) and 1 (heart disease), with a minor imbalance favoring class 1. To preserve this distribution during evaluation, stratified sampling was applied during train-test splitting. Although no resampling was used, this approach helped mitigate bias and ensured fair computation of metrics such as precision and recall.

C. Feature Extraction

The technique of turning raw data into numerical features that can be handled while keeping the information in the original data set is known as feature extraction. Compared to using machine learning on the raw data directly, it produces superior outcomes [10]. Categorical characteristics in a dataset are tested using chi-square. We compute the Chi-square between each feature and the intended outcome, then choose the desired number of features with the highest Chi-square values. It examines whether the connection between

two categorical variables in the sample accurately reflects that association. Chi-square score is calculated using:

$$\chi^2 = \frac{\sum (B_n - Q_n)^2}{Q_n} \quad (1)$$

The test statistic generated from the data has a χ^2 frequency distribution if the null hypothesis, according to which there are no differences between the classes in the population, is true. The test's goal is to establish the likelihood of the frequency components if the null hypothesis is true. When the data are uncorrelated, test statistics with a χ^2 distribution are produced. A contingency table's two parameters are assessed to see if they are related using a chi-square test for independence. It looks into the possibility of wider-ranging differences between categorical variable distributions [11].

The objective of this study was to generate a prognosis of an individual's risk of heart failure based on the provided records. These techniques comprised classifiers for DT, RF, MNB, BNB, GNB, KNN, LR and SVM. The suitable workflow of the proposed model is shown in Figure 3.

Among the evaluated features, ST-Slope, Cholesterol, MaxHR, and Oldpeak received the highest Chi-square scores as presented in Table II, indicating a strong statistical association with heart disease. These features were prioritized during model training.

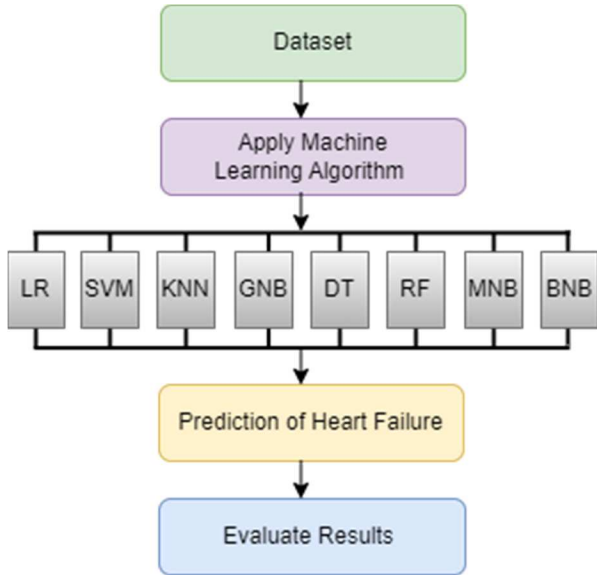


Fig. 3. End-to-end workflow of the heart failure prediction pipeline using eight machine learning algorithms.

D. Apply Machine Learning Algorithms

A classification issue in machine learning is one in which, a class label is anticipated for a specific sample of input data. For heart failure prediction, we tested and applied various machine-learning algorithms in this paper. Eight classifiers were selected to compare diverse learning strategies—linear, kernel-based, ensemble, and probabilistic. The three Naive

Bayes variants were included due to their differing assumptions: Gaussian for continuous features, Multinomial for count-based features, and Bernoulli for binary input data. The likelihood of a binary (yes/no) event is chosen using logistic regression [12]. SVM uses statistical learning frameworks and kernel approaches for regression and classification. KNN categorizes new data points based on similarity and is mostly employed for classification issues. GNB predicts output variables using probabilistic and Gaussian distributions, with each parameter having a separate capacity. With leaf nodes representing all outcomes, the decision tree is a non-parametric supervised learning technique for classification and regression. Combining decision trees with regression using the mean of all tree outputs, Random Forest generates a forest. Machine learning techniques like MNB are used to forecast a text's label. With discrete data, the BNB classifier takes binary values such as true or false, yes or no, success or failure, and 0 or 1.

E. Evaluate Results on Different Metrics

Evaluation metrics can be used to assess a machine learning model's performance. We analyze our ML models using a variety of measures. These measures aid us in assessing the correctness of our model. Utilizing diverse machine learning methods, we evaluated the efficiency of our suggested approach using a variety of metrics, including accuracy, precision, recall, f1-score, mean squared error value, average precision value, area under curve value, confusion matrix, ROC curve, and Precision-Recall curve etc. Section IV has a clear illustration of these metrics.

IV. RESULTS ANALYSIS AND DISCUSSION

Table III shows the performance of all eight classifiers using both the holdout and 10-fold cross-validation methods. SVM achieved the best overall performance, with the highest accuracy (88.59%), precision (0.89), recall (0.88), and F1-score (0.89) under the holdout method, and also led in recall (0.92) and F1-score (0.89) in cross-validation. RF showed the highest precision (0.88) in cross-validation and performed consistently well overall. KNN improved notably in cross-validation, while the three Naive Bayes models (GNB, MNB, BNB) delivered moderate results. DT had the lowest scores across both settings. These results confirm the strong predictive capability of SVM and the stability of RF across evaluation strategies.

The performance results for various machine learning classifiers are shown in TABLE V. This table displays three key metrics—Mean Squared Error (MSE), Average Precision (AP), and Area Under the Curve (AUC). MSE measures how closely the predicted values match the actual outcomes, with zero indicating a perfect model. The SVM algorithm achieved the lowest MSE of 0.11. Average Precision is calculated as the weighted mean of precisions at different threshold levels. The highest AP score of 0.93 was achieved by the LR, SVM, GNB, RF, and MNB algorithms. AUC represents the area under the ROC curve and reflects a model's ability to distinguish between classes. It is computed from the curve defined by

TABLE III
PERFORMANCE COMPARISON OF CLASSIFIERS USING HOLDOUT AND 10-FOLD CROSS-VALIDATION

Algorithm	Accuracy (%)		Precision		Recall		F1-Score	
	Holdout	Cross Validation	Holdout	Cross Validation	Holdout	Cross Validation	Holdout	Cross Validation
LR	83.70	85.39	0.83	0.85	0.83	0.88	0.84	0.87
SVM	88.59	87.46	0.89	0.85	0.88	0.92	0.89	0.89
KNN	83.15	87.35	0.83	0.86	0.82	0.91	0.83	0.88
GNB	82.61	83.42	0.82	0.86	0.83	0.82	0.83	0.84
DT	78.80	79.14	0.78	0.82	0.77	0.80	0.79	0.80
RF	85.87	84.78	0.85	0.88	0.85	0.85	0.86	0.87
MNB	82.61	80.48	0.82	0.83	0.83	0.80	0.83	0.82
BNB	82.61	82.88	0.82	0.85	0.83	0.83	0.83	0.84

the true and false positive rates. The LR model achieved the highest AUC value of 0.93. These results show that different models excelled in different metrics, with SVM performing best in error reduction.

TABLE IV
COMPARISON OF ML ALGORITHMS BASED ON MSE, AP, AND AUC

Algorithm	Mean Squared Error (MSE)	Average Precision (AP)	Area Under Curve (AUC)
LR	0.16	0.93	0.91
SVM	0.11	0.93	0.92
KNN	0.16	0.88	0.89
GNB	0.17	0.93	0.91
DT	0.21	0.77	0.77
RF	0.15	0.93	0.93
MNB	0.17	0.93	0.90
GNB	0.17	0.92	0.91

Figure 4 depicts the SVM Confusion Matrix, which summarizes the classifier's prediction performance on the test data. In a confusion matrix, the predicted and actual outcomes are compared and displayed in tabular form [?].

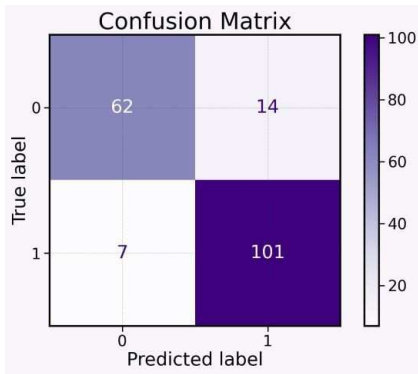


Fig. 4. Confusion matrix for the SVM classifier showing actual vs. predicted labels.

With a recall value of 0.93, Figure 6 illustrates the Precision-Recall Curve for the SVM classifier. Figure 5 shows the ROC Curve of SVM, where the True Positive Rate appears on the Y-axis and the False Positive Rate on the X-axis. This curve demonstrates that the SVM classifier achieved an AUC of 0.92.

Figure 7 presents the accuracy overview of the proposed heart failure prediction system. It compares the test accuracy

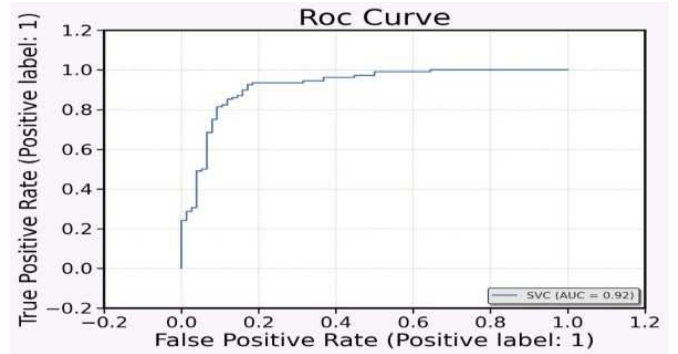


Fig. 5. ROC curve for the SVM classifier illustrating sensitivity and specificity.

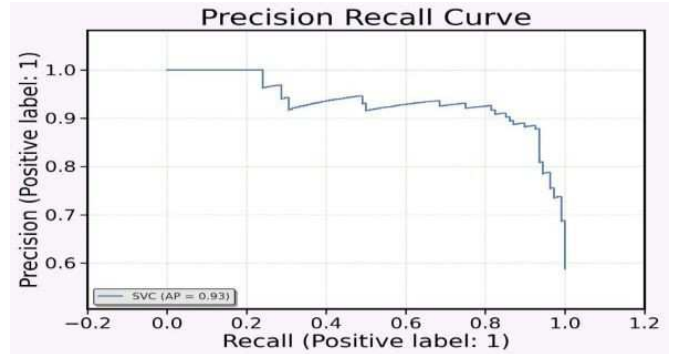


Fig. 6. Precision-recall curve for the SVM classifier showing trade-off between metrics.

of all eight classifiers used in this study. As shown in the figure, the DT method produced the lowest accuracy, while the SVM algorithm achieved the highest.

TABLE VI presents a comparison between our proposed system and previously published works. It confirms that the SVM model in our study achieved the highest accuracy of 88.59%. In comparison, the best result from related work was 85% using the LightGBM algorithm.

While the SVM classifier achieved the highest accuracy, it may scale poorly with larger datasets due to its high training complexity. In contrast, Naive Bayes and Logistic Regression models are computationally lightweight. KNN has low training cost but higher prediction time, while Random Forest requires more memory due to its ensemble nature.

TABLE V
COMPARISON AMONG EXISTING WORKS AND OUR PROPOSED WORK

Name	Year	Dataset	Learning Model	Accuracy (%)	Precision	Recall	F1-score
Sahoo P.K. et al. [?]	2020	UCI Repository	SVM	85.20	N/A	N/A	N/A
Agrawal H. et al. [?]	2021	Kaggle	LR	86.40	0.96	0.87	0.91
Krishna P.G. et al. [?]	2021	IBM	Gradient Boosting	87.40	N/A	N/A	N/A
Awan S.E. et al. [?]	2019	Hospital Morbidity Dataset	RF	76.39	N/A	N/A	N/A
Chicco D. et al. [?]	2020	HF patient's medical records	LR	83.30	N/A	0.532	71
Mamun M. et al. [?]	2022	UCI Repository	LightGBM	85.00	0.87	0.82	0.84
Our Proposed Model	2023	Kaggle	SVM	88.59	0.88	0.88	0.89

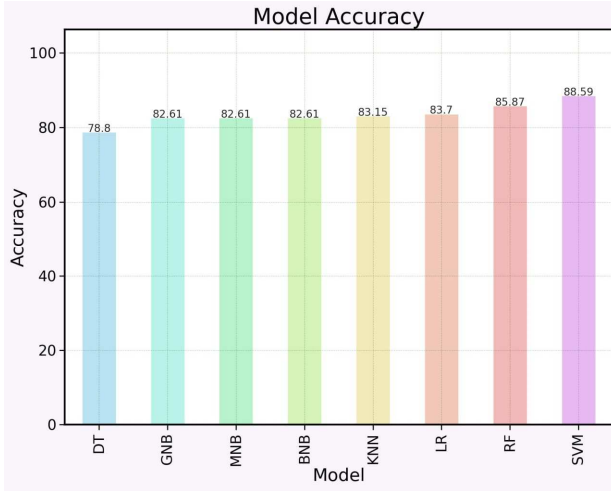


Fig. 7. Overview of classifier accuracy scores, highlighting SVM as the top performer.

V. CONCLUSION

Experts must develop more reliable mechanisms to diagnose heart failure, given the significant mortality and morbidity associated with the disease. Predictive technology is essential to reduce the risk of heart failure fatalities, especially with the rising number of patients affected globally. In this study, we evaluated eight classification algorithms for heart failure prediction. Among them, SVM achieved the highest accuracy of 88%, while the RF classifier yielded the best AUC score of 0.93. These results highlight the robustness of our comparative machine learning framework and demonstrate notable performance gains over prior studies.

One limitation of our work is the modest size of the dataset used. To enhance our model further, we plan to explore larger and more diverse datasets in future studies. We also aim to implement a feedback mechanism to support continual model improvement as new data becomes available. Additionally, while the SVM classifier delivered the strongest results, its limited interpretability and scalability could hinder real-world deployment in clinical settings.

To address these challenges and unlock greater potential, we additionally envision incorporating deep learning techniques

into our framework. In particular, convolutional neural networks (CNNs) can learn hierarchical feature representations from data automatically and may capture more complex relationships that traditional machine learning models overlook.

REFERENCES

- [1] J. Wang, "Heart failure prediction with machine learning: a comparative study," in *Journal of Physics: Conference Series*, vol. 2031, no. 1. IOP Publishing, 2021, p. 012068.
- [2] T. Tong, Y. Guo, Q. Wang, X. Sun, Z. Sun, Y. Yang, X. Zhang, and K. Yao, "Development and validation of a nomogram to predict survival in septic patients with heart failure in the intensive care unit," *Scientific Reports*, vol. 15, no. 1, p. 909, 2025.
- [3] P. K. Sahoo and P. Jeripothula, "Heart failure prediction using machine learning techniques," *Available at SSRN 3759562*, 2020.
- [4] G. Quer, R. Arnaout, M. Henne, and R. Arnaout, "Machine learning and the future of cardiovascular care: Jacc state-of-the-art review," *Journal of the American College of Cardiology*, vol. 77, no. 3, pp. 300–313, 2021.
- [5] E. J. Topol, "Machine learning and the future of cardiovascular care," *Journal of the American College of Cardiology (JACC)*, vol. 77, no. 3, pp. 300–313, 2021. [Online]. Available: <https://www.jacc.org/doi/abs/10.1016/j.jacc.2020.11.030>
- [6] S. Suganthi, G. Vijipriya, N. Madian *et al.*, "An approach for predicting heart failure rate using ibm auto ai service," in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. IEEE, 2021, pp. 203–207.
- [7] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, and G. Dwivedi, "Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics," *ESC heart failure*, vol. 6, no. 2, pp. 428–435, 2019.
- [8] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, Feb. 2020. [Online]. Available: <https://doi.org/10.1186/s12911-020-1023-5>
- [9] F. Soriano, "Heart failure prediction dataset," Kaggle dataset, 2021, accessed June 25, 2025. [Online]. Available: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
- [10] M. Abid, S. Bukhari, and M. Saqlain, "Enhancing software effort estimation in healthcare informatics: A comparative analysis of machine learning models with correlation-based feature selection," *Sustainable Machine Intelligence Journal*, vol. 10, pp. 50–66, 2025.
- [11] B. Ahmad, J. Chen, and H. Chen, "Feature selection strategies for optimized heart disease diagnosis using ml and dl models," *arXiv preprint arXiv:2503.16577*, 2025.
- [12] Z. Xu, Y. Hu, X. Shao, T. Shi, J. Yang, Q. Wan, and Y. Liu, "The efficacy of machine learning models for predicting the prognosis of heart failure: A systematic review and meta-analysis," *Cardiology*, vol. 150, no. 1, pp. 79–97, 2025.