Heart Failure Prediction

1st Ajay Talele
ENTC
Vishwakarma Institute of Technology
Pune, India
ajay.talele@vit.edu

2nd Aum Kolekar
ComputerEngeneering
Vishwakarma Institute of Technology
Pune, India
aum.kolekar23@vit.edu

3rd Sanket Auti
ComputerEngeneering
Vishwakarma Institute of Technology
Pune, India
sanket.auti23@vit.edu

4th Azlan Ahemad
ComputerEngeneering
Vishwakarma Institute of Technology
Pune, India
ishteyaq.azlan23@vit.edu

5th Omkar Bankar
ComputerEngeneering
Vishwakarma Institute of Technology
Pune, India
omkar.bankar23@vit.edu

Abstract

Heart failure is a serious worldwide health concern that frequently results in death and serious consequences. Improving patient outcomes requires early diagnosis, but conventional approaches are resource-intensive and rely on clinical knowledge. This study investigates the prediction of heart failure events using machine learning models, particularly Random Forest, Support Vector Machine (SVM), and Logistic Regression. The models were trained using a dataset that included clinical and demographic data, including age, blood pressure, smoking, and health indicators including diabetes and anemia. The models were evaluated using key performance characteristics such as precision, specificity, sensitivity, and accuracy. Random Forest outperformed both SVM and Logistic Regression in detecting heart failure events, achieving the top scores on all metrics. The results show that by offering a useful tool for early detection and decision-making, machine learning has the potential to enhance patient outcomes and optimize healthcare resources.

Keywords:

Heart Failure, Machine Learning, Logistic Regression, Support Vector Machine (SVM), Random Forest, Predictive Modeling, Sensitivity, Specificity, Precision, Accuracy, Healthcare, Early Diagnosis, Data Preprocessing, Class Imbalance, Clinical Data

Introduction

Heart failure is a major global health issue that impacts millions of people globally. It happens when the heart cannot adequately pump blood to fulfill the body's demands, which can result in serious problems and, frequently, death. Improving patient outcomes and lowering healthcare costs need early diagnosis and prompt action. Traditional diagnostic techniques, however, can need a lot of resources and rely significantly on clinical knowledge, which isn't always accessible.

This initiative uses machine learning—more especially, a customized Logistic Regression model—to forecast patients' risk of developing heart failure. Key clinical and demographic information, including age, blood pressure, lifestyle choices like smoking, and physical health markers like anemia, are all included in the extensive dataset used to train the algorithm. The main objective is to develop an interpretable and accurate forecasting system that can be used practically in medical contexts.

To further assess the model's robustness and dependability, sophisticated performance indicators such as specificity, sensitivity, precision, and accuracy are used. Normalization, oversampling to correct for class imbalances, and hyperparameter adjustment to maximize the model's performance are further steps in the data preparation process.

In addition to proving that machine learning is feasible in the medical field, this study lays the groundwork for the creation of increasingly complex systems that will aid physicians in making well-informed choices.

Literature Survey

[1]Heart failure prediction has been extensively studied using various machine learning techniques to improve diagnostic accuracy and patient outcomes. Wang (2021) conducted a comprehensive comparative analysis of 18 machine learning models, demonstrating the effectiveness of methods like Random Forest, Support Vector Machines, and Logistic Regression in predicting heart failure events. The study emphasized the importance of data preprocessing techniques such as SMOTE for addressing class imbalance and normalization methods like z-score and min-max for feature scaling. Serum creatinine, ejection fraction, and time emerged as critical predictors across models, aligning with findings from prior works using datasets like the Cleveland Heart Disease Dataset and electronic health records (EHR). Other studies highlighted the potential of deep learning methods and ensemble techniques for enhanced predictive performance. Collectively, these approaches underscore the growing utility of machine learning in heart failure prediction, offering tools to assist clinicians in early diagnosis and targeted treatment strategies.

[2]Cardiovascular diseases (CVDs) represent a global health challenge, with 17.9 million deaths reported annually. The growing emphasis on predictive healthcare has led to advanced models utilizing machine learning to assess risk and mortality in heart failure (HF) patients. A notable approach involves employing Support Vector Machines (SVMs) with random undersampling to enhance recall and prioritize high-risk individuals. Such models integrate diverse variables, including demographic, clinical, and behavioral data, selected through iterative optimization. The literature highlights disparities in CVD outcomes across demographics and emphasizes the need for robust, scalable models. Prior studies often relied on statistical methods unable to capture the multidimensional nature of CVD. Recent advancements, including the S-HF model, address these limitations, achieving improved recall and discriminatory capabilities across subgroups. Despite promising developments, challenges remain, such as demographic bias and dynamic shifts in risk factors, necessitating ongoing validation for broader applicability.

[3]Heart disease remains a leading cause of mortality globally, emphasizing the need for accurate and early detection methods. Several machine learning-based systems have been developed to predict heart diseases by leveraging clinical datasets and various algorithms, such as Support Vector Machines (SVM), Naïve Bayes, Decision Trees, and Random Forests. Studies reveal that SVM often outperforms other methods in accuracy, achieving up to 85.2%. These models utilize diverse attributes like age, blood pressure, cholesterol levels, and chest pain type, often sourced from repositories like UCI. Advanced methods, including hybrid neural networks and feature selection techniques, further enhance prediction reliability. However, challenges such as data sparsity, unstructured inputs, and demographic biases persist, necessitating continued optimization and validation for robust, scalable implementations in healthcare. The integration of machine learning in this domain shows promise in reducing diagnostic costs, improving resource allocation, and potentially saving lives through timely interventions.

[4]Heart disease is a leading cause of mortality, necessitating advancements in early detection and prediction. Machine learning (ML) has emerged as a transformative tool for addressing these challenges by leveraging extensive clinical and electrocardiogram (ECG) datasets. Algorithms like Support Vector Machines (SVM), XGBoost, and neural networks have been widely adopted for their predictive

accuracy, achieving F1 scores above 0.93 in various contexts. Techniques such as real-time feature extraction from ECG data and model calibration enhance their applicability across different populations and settings. Recent studies have demonstrated the effectiveness of ML models in identifying arrhythmias, atrial fibrillation, and other cardiac abnormalities in both clinical and wearable device scenarios. Despite these advances, challenges like dataset variability, demographic biases, and real-time implementation persist, emphasizing the need for continuous optimization and cross-validation across diverse patient groups to improve reliability and scalability.

[5]Heart failure research is a dynamic field marked by significant advancements in understanding, diagnosing, and treating this multifaceted condition. The NHLBI (National Heart, Lung, and Blood Institute) has been pivotal in driving research through basic science, clinical trials, and longitudinal studies. Notable efforts include studies like HF-ACTION, which demonstrated the benefits of supervised aerobic exercise for improving outcomes in heart failure patients, and CONCERT-HF, which explored innovative cellular therapies. The institute has also focused on the disparities in diagnosis and treatment, aiming to mitigate the underrepresentation of minority populations in clinical trials. Recent initiatives emphasize precision medicine, such as using artificial intelligence to identify specific heart failure subtypes and novel interventions like man-made cell patches to repair heart muscle. Through its Heart Failure Network (HFN), the NHLBI has supported research on novel drugs, diagnostic methods, and management strategies, which have informed clinical guidelines and enhanced patient care quality .

[6]The uploaded document explores the development of a predictive model using machine learning, specifically the XGBoost algorithm, to assess in-hospital mortality risks for patients with heart failure (HF) and atrial fibrillation (AF) in the ICU. Using data from the MIMIC-III and eICU databases, the study identifies 19 critical variables influencing mortality risk, including APS III, OASIS scores, age, and renal function indicators. These variables were selected through LASSO regression, and the model outperformed traditional scoring systems like SOFA and OASIS in predictive accuracy. The findings highlight the model's utility in personalized risk assessment and potential guidance for clinical decision-making, emphasizing machine learning's role in critical care. The research was validated internally and externally to confirm reliability, showing promise for future application despite limitations like lack of subgroup analyses for HF subtypes.

[7]The literature on heart disease prediction emphasizes the growing significance of machine learning (ML) in healthcare for enhancing diagnostic accuracy and identifying high-risk individuals. Techniques like Logistic Regression, Support Vector Machines (SVM), and Random Forests are widely used, demonstrating varying predictive capabilities across datasets such as the Framingham Heart Study and Heart Failure Clinical Records. Recent studies include the development of neural network models for sequential healthcare data modeling to predict heart failure and the use of evolutionary rule learning for efficient feature selection. Optimized approaches like Random Search Algorithms combined with Random Forests have achieved improved accuracy in diagnosing cardiovascular diseases. These efforts underline the integration of advanced algorithms with medical datasets for actionable insights, contributing to early diagnosis and effective management of heart diseases

[8]The paper conducts a comparative study to optimize heart disease diagnosis using various machine learning classifiers, including Random Forest (RF), Decision Tree (DT), Gradient Boosting Classifier (GBC), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), with and without Sequential Feature Selection (SFS). The study employs multiple datasets, such as the Cleveland, Hungary, Switzerland, Long Beach V, and Heart Statlog Cleveland Hungary datasets. Using K-fold cross-validation, it evaluates the performance of these algorithms based on accuracy, sensitivity, specificity, and other metrics.

The study finds that Random Forest and Decision Tree with SFS achieve the highest accuracy (100%) across multiple datasets, emphasizing the importance of feature selection in enhancing model performance. Gradient Boosting Classifier also shows strong results with a high average ROC AUC score. The results highlight the impact of removing redundant features through SFS in reducing computation time and improving classifier accuracy. Future directions include applying diverse feature selection techniques and extending the methodology for broader applications in real-world medical scenarios.

Methodology

Based on patient clinical data, we created a machine learning model in this work to forecast heart failure episodes. The following crucial steps make up the methodology: data gathering, preprocessing, model creation, assessment, and implementation.

1. **Data Collection:**

   The project's training and assessment datasets were from publicly accessible datasets that included heart failure patients' medical information. The target variable, which indicates whether the patient experienced a heart failure event (1 for "Death", 0 for "Survival"), is included in the dataset along with a number of clinical features, including age, sex, blood pressure, cholesterol levels, lifestyle factors like smoking, and physical health indicators (such as the presence of anemia and diabetes).

2. **Data Preprocessing:**

   Data preprocessing plays a crucial role in ensuring the quality and reliability of the model. The following steps were applied to the dataset:

   - **Normalization**: All numerical features (e.g., age, blood pressure) were normalized using the MinMaxScaler to scale the values between 0 and 1, aligning with the assumption that all features are on a comparable scale.
   - **Handling Missing Data**: The dataset did not contain significant missing values; however, imputation techniques were used where necessary to maintain dataset completeness.
   - **Resampling**: To address the class imbalance (with more instances of "Survival"), RandomOverSampler was employed to balance the dataset by oversampling the minority class, ensuring the model is not biased towards the majority class.

3. **Model Development:**

   A **Logistic Regression** model was chosen for its simplicity, interpretability, and effectiveness in binary classification tasks. Hyperparameter tuning was performed to optimize the model's performance, with the following key parameters considered:

   - **C** (regularization strength)
   - **Solver** (e.g., 'liblinear', suitable for small datasets)
   - **Penalty** (L2 regularization)

   The dataset was split into training (80%) and testing (20%) sets to train the model and evaluate its performance. The logistic regression model was trained on the training set, using the resampled dataset to ensure balanced class distribution.

4. **Model Evaluation:**

   To evaluate the performance of the model, the following metrics were computed:

   - **Accuracy**: The proportion of correct predictions among all predictions.
   - **Sensitivity (Recall)**: The model's ability to correctly identify patients with heart failure events.
   - **Specificity**: The model's ability to correctly identify patients without heart failure events.

- **Precision**: The proportion of positive predictions that are correct, indicating the model's reliability in detecting heart failure events.

5. **Deployment:**
6. The model was saved for deployment using the joblib library once it had been trained and assessed. Using a straightforward graphical user interface (GUI) created using Streamlit, the finished model may be used to forecast heart failure occurrences depending on user input. Users provide pertinent clinical data, which the model processes and normalizes to produce a "Death" or "Survival" forecast.

Results

Model	Accuracy	Sencitivity	Specificity	Precision
Logistic Rigration	0.8500	0.7389	0.9009	0.7733
SVM	0.9740	0.9586	0.9810	0.9586
Random Forest	0.9910	0.9809	0.9956	0.9904

Random Forest achieved the greatest accuracy, sensitivity, specificity, and precision, outperforming both SVM and Logistic Regression on all criteria. This suggests that Random Forest is very good at reducing false positives (specificity and accuracy) and detecting actual heart failure cases (sensitivity). Additionally, SVM performed well, particularly in sensitivity and specificity, demonstrating its great efficacy in detecting heart failure events as well as non-events. Despite its strong performance, logistic regression's sensitivity and accuracy were inferior to those of the other models. Its ability to detect true positives was lower than that of SVM and Random Forest.

All things considered, Random Forest is the best accurate model for forecasting heart failure occurrences, with SVM coming in second. Despite its effectiveness, logistic regression might not be as reliable in this application, especially when it comes to precisely identifying heart failure episodes.

Conclusion

This study showed how well machine learning methods—specifically, Random Forest, Support Vector Machine (SVM), and Logistic Regression—predict heart failure occurrences. Early diagnosis and decision-making were aided by the models' ability to produce accurate predictions by utilizing clinical data and sophisticated preprocessing methods like normalization and resampling. Random Forest was the best model for predicting heart failure episodes in this study, outperforming both Logistic Regression and SVM on all important measures, including accuracy, sensitivity, specificity, and precision. While Logistic Regression was successful, it did not do as well in recognizing true positives and false positives as SVM, especially when it came to sensitivity and specificity. The findings underscore the potential of machine learning models in healthcare, particularly in the context of heart failure prediction, where early and accurate detection is critical for improving patient outcomes. While **Random Forest** emerged as the best-performing model in this study, further work could explore ensemble approaches or deep learning techniques to further improve performance and applicability in diverse patient populations.

By incorporating these models into clinical practice, medical professionals may be able to make better judgments, which might save diagnostic expenses, optimize resource allocation, and ultimately save lives through prompt treatments.

References

[1] J. Wang, “Heart failure prediction with machine learning: A comparative study,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Sep. 2021. doi: 10.1088/1742-6596/2031/1/012068.

[2] S. Ghansah, “Cardiovascular Disease Risk Prediction Using Machine Learning Using Machine Learning to improve CVD Risk Prediction.” [Online]. Available: https://ssrn.com/abstract=4726955

[3] * Prasanta, K. Sahoo, and P. Jeripothula, “HEART FAILURE PREDICTION USING MACHINE LEARNING TECHNIQUES.” [Online]. Available: https://ssrn.com/abstract=3759562

[4] D. Bertsimas, L. Mingardi, and B. Stellato, “Machine Learning for Real-Time Heart Disease Prediction,” *IEEE J Biomed Health Inform*, vol. 25, no. 9, pp. 3627–3637, Sep. 2021, doi: 10.1109/JBHI.2021.3066347.

[5] “Heart Failure Research _ NHLBI, NIH”.

[6] C. Wang *et al.*, “Development of a predictive model for in-hospital mortality risk in patients with heart failure and atrial fibrillation in the ICU: a machine learning study based on the MIMIC-III and eICU databases.” [Online]. Available: https://ssrn.com/abstract=4485609

[7] R. Baxani and M. E. Student, “Heart Disease Prediction Using Machine Learning Algorithms Logistic Regression, Support Vector Machine and Random Forest Classification Techniques.” [Online]. Available: https://ssrn.com/abstract=4151423

[8] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, “Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection,” *IEEE Access*, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.