A Predictive Analytics Approach for Proactive Heart Failure Diagnosis Using Machine Learning

OPEN ACCESS

Volume: 13

Special Issue: 3

Month: February

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Citation:
Devi, V., et al. "A Predictive Analytics Approach for Proactive Heart Failure Diagnosis Using Machine Learning." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 3, 2026, pp. 83–92.

DOI:
https://doiorg/10.34293/sijash.v13iS3-i1-Feb.10265

Dr. V. Devi
Professor, PG & Research Department of Computer Science

Mrs. S. Bhuvana
Research Scholars, PG & Research Department of Computer Science

Dr. A. Ambeth Raja
Associate Professor, PG & Research Department of Computer Science

D. R. Ashwin Kumar
Department of Computer Science (Data Science)
Manipal Academy of Higher Education, Bengaluru

Abstract

The early and precise detection of heart failure can lead to the mitigation of disease progression and decrease cardiovascular mortality. However, traditional diagnostic methods utilize subjective judgment of symptoms and isolated clinical factors, which limit the scope for early intervention. This paper suggests a framework of predictive analytics for early detection of heart failure through efficient machine learning models, trained on publicly available clinical data. Experiments were performed on the Heart Failure Clinical Records data set, collected from UCI Machine Learning Repository, consisting of demographic attributes, physical measurements and biochemical markers. We used a detailed preprocessing pipeline including normalization, outlier detection and class imbalance problem management. To increase the discriminative ability, a hybrid feature selection method involving mutual information ranking and evolutionary optimization was devised. Finally, diverse learners, including Random Forest, Support Vector Machine, Gradient Boosting and XGBoost were trained and merged through a weighted ensemble learning scheme to improve robustness and generalization. The learners and the hybrid feature selection method were assessed using accuracy, precision, recall, F1-score, RMSE and AUC metrics. Our proposed ensemble learning system obtained an accuracy of 97.1% and an AUC of 0.986. The proposed ensemble approach successfully outperformed the individual learning models and the standard feature selection technique.

Keywords: Hear Failure, Machine Learning, Prediction, Supervised Learning, Detection

Introduction

As a result of HF being the commonest reason for admission to hospital and a prominent cause of death from cardiovascular disease, it is still an immense clinical and financial burden upon individuals and society at large [1].This article concentrates on the underlying

progressive, pathologic inadequacy of cardiac perfusion of the body, which many times can be subclinical, developing over years unnoticed to the development of dramatic symptoms [2,3]. Early HF often goes un, diagnosed due to the non, specific clinical features and the usage of standard diagnostic and risk, stratification methods (echo, labs, general clinician/doctor risk assessment methods, etc) [4, 6], thus not prompting therapy until myocyte loss have occurred and the damage is irreversible, decreasing treatment effectiveness and prognosis.

As of recent digital health informatics progress, and early application of EHR systems, we now have an unparalleled capability to collect heterogeneous clinical data like demographic characteristics, physiologic measurements, lab values, and comorbidity indices from many patients at a scale which was previously unachievable. The growing data ecosystem presents an unprecedented opportunity to transition from the traditional, disease prediction approach of diagnosis to prediction through proactive data driven practice [7, 10]. Predictive analytics leveraging data mining and ML has revealed itself to have tremendous promise at discerning subtle patterns within complex, high, dimensional data sets which often cannot be found with classical statistics [10]. ML methods have the power to effectively model the complex, non, linear interrelationships of many risk factors in cardiovascular disease, to achieve early disease detection, and adaptable risk prediction for each patient [11, 13].

Also earlier research has examined the machine learning techniques to build heart disease predictors. Different model such as logistic regression, decision trees, support vector machine, random forests, neural networks were built and most of them have performed very good classification accuracy. Nevertheless, these models predict general heart disease and cannot predict the early stages of diseases for developing heart failure. Furthermore, problems such as weak feature selection, class bias problem, overfitting, and the poor clinical interpretability have reduced the effectiveness and buy in of these methods [13,15].

The question of high, dimensionality and data heterogeneity raises numerous issues on cardiovascular data analysis, too. Irrelevant and possibly redundant features can overwhelm physiological signals and can lead to over, fitting and significantly reduced generalization of models. Methods for feature engineering and feature selection should be employed to reduce feature number such that high predictive accuracy can be reached without requiring excessive computational expenses. Moreover, single classifier models are known to work poorly on certain groups of patients which motivate us to apply ensemble learning approaches, to combine a number of weak or strong learners together.

This paper focuses on developing an optimal machine learning classifiers for early and accurate diagnosis of heart failure through the proposed predictive analytic system. The most outstanding part of the suggested technique relies on advanced data cleaning method, ensemble feature selection methods and ensemble classification methods, capable for disease diagnosis at an early stage with less false alarm while predicting the disease. Designed predictive model also increases the ability to use with the application of explainability diagnostic while in the middle of making decision.
The key contributions of this study are summarized as follows:

1. Machine learning-based overall prediction system for early detection of heart failure using clinical information.
2. Application of a hybrid feature optimization strategy to boost discriminative capacity and decrease dimensionality.
3. Proposed architecture for ensemble learning aiming at the generalized and robust diagnostic performance.
4. Integration of explainable AI techniques to yield clinically interpretable insights on risk factors for heart failure.

The remainder of this paper is structured as follows: Section 2 reviews related works in heart failure prediction using machine learning. Section 3 describes the dataset characteristics and proposed predictive analytics methodology. Section 4 presents experimental results and comparative performance analysis. Section 5 discusses clinical implications and limitations, followed by conclusions and future research directions in Section 6.

Related Work

Modern developments in machine learning and artificial intelligence have drastically changed cardiovascular disease prognostication and in specific the heart failure classification and survival prediction domains. Kokori et al. [1] showed a review on modern machine learning models for heart failure survival analysis and also illustrated the recent acceptance of ensemble learning, deep neural networks, and explainable models in attempt to improve clinical interpretability. From their conclusions, hybrid models appear to always achieve better performance compared to single learners but problems concerning the variation between datasets and models' interpretation are still to be answered.

Imaging-based AI methods have shown similar promises in pre-screening for heart failure. For heart failure with preserved ejection fraction (HFpEF), Chiou et al. [2] constructed an AI-prescreening tool for heart failure detection based on intra-beat cardiac dynamics by a deep learning model. The system was proven highly sensitive by modeling minor temporal changes of cardiac motion patterns. Such time dynamics is also considered important for cardiac prognosis. Duffy et al. [5] used high throughput deep learning tool to characterize the left ventricular hypertrophy precision phenotype in a scalable way on large patient populations, thus allow large scale cardiac risk stratifying.

Besides imaging techniques, the role of EHR-based predictive analytics is being studied much. Zhou et al. [3] introduced an ensemble deep learning framework combined with improved clinical risk factors for prediction of heart failure initiation using real world health records. It achieved decreasing of false prediction rates by hierachical features optimization and multi-model combination. Miranda et al. [7] improved clinical explainability by incorporating SHapley Additive exPlanations(SHAP) into machine learning models to predict arteriosclerotic heart disease and enables doctors to check for the biggest contributors and evaluate a model's predictions.

Multiple works have successfully pushed the realm of machine learning approaches towards the generalized context of cardiovascular as well as biomedical diagnosis. In ventricular volume asymmetry as an imaging biomarker, McCracken et al. [8] described a new measure of disease classification and prognostication and also found out that cardiac structural asymmetry is a powerful predictor of outcome in many cardiovascular diseases. In order to optimize prediction in myocarditis, Kasmaee et al. [6] have created an ensemble-reinforcement based deep learning architecture for diagnosis of myocarditis from cardiac MRI.

AI's evolving importance in personalized cardiovascular medicine has been underscored through thorough literature reviews. Udoy and Hassan [4] conduct a systematic review on recent advancements of AI technology in diagnosis of heart failure, highlighting key developments in pre-clinical diagnosis, predictive modeling of risks, hybrid intelligent systems, and explainable intelligence which will assist in clinical implementation. In their findings, several limitations still remain to be researched: imbalanced data, feature redundancy, and poor generalizability.

While advancements have been made in imaging, EMR analysis, and ensemble deep learning systems, numerous issues have yet to be overcome. Most current systems heavily depend on unimodal datasets, preventing successful generalization across datasets. Furthermore, feature selection methods are often underestimated and can lead to poor performance in clinical practice.

While an increasingly investigated subject, explainability is not equally incorporated in each predictive pipeline.

Different from the previous studies, this paper presents a comprehensive predictive analytics framework by integrating hybrid feature selection, ensemble machine learning and explainable AI in a synergistic manner for preemptive heart failure prediction from clinical data. Through tackling issues on dimensional redundancy, robustness, and interpretability simultaneously, it tries to provide a clinically trustworthy and scalable early diagnosis system.

Methodology

In this paper, a new predictive analytics framework to proactively identify heart failure was constructed by integrating autonomic clinical data processing, ensemble feature optimization, ensemble classification, and interpretation models. This general framework endeavors to provide high accuracy at early stages while ensuring clinical applicability and stability.

Dataset Description

For the experimental validation of the predictive analytics framework, the data set "Heart Failure Clinical Records Dataset" from the UCI Machine Learning Repository has been used. This dataset is quite a popular dataset on machine learning for heart failure prediction and is available for free for research purposes provided that the dataset is cited.

Heart Failure Clinical Records Dataset — UCI Machine Learning Repository

available from : https://archive.ics.uci.edu/dataset/519/heart%2Bfailure%2Bclinical%2Brecords

- Instances: 299 patients
- Features: 13 clinical attributes per patient
- Target Variable: Death event (binary flag indicating whether a patient died during the follow-up period)
- Data Types: Integer and Continuous clinical features
- Tasks Supported: Classification, Regression, Clustering

The dataset incorporates both demographic and clinical measurements relevant to heart failure. Table 1 summarizes the key features used for model training:

Table 1. Dataset parameters

Feature Name	Description	Type
age	Patient age (years)	Integer
anaemia	Reduction in red blood cells / haemoglobin (Boolean)	Binary
creatinine_phosphokinase (CPK)	Level of CPK enzyme in blood (mcg/L)	Continuous
diabetes	Indicates if patient has diabetes	Binary
ejection_fraction	Percentage of blood leaving heart at each contraction (%)	Continuous
high_blood_pressure	Hypertension status	Binary
platelets	Platelet count (kiloplatelets/mL)	Continuous
serum_creatinine	Serum creatinine level (mg/dL)	Continuous
serum_sodium	Serum sodium level (mEq/L)	Continuous

sex	Gender (female/male)	Binary
smoking	Smoking status	Binary
time	Follow-up period duration (days)	Integer
death_event	Target: Death during follow-up	Binary

Median imputation was adopted to handle missing values to make model generalization, continuous attribute were normalized using min-max scaling:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Class imbalance was addressed using adaptive synthetic sampling, which avoid biased learning to the majority class.

Data Preprocessing and Feature Engineering

Noise, redundant features, and variables with weak discriminative power is well-established in the clinical data. Such data can degrade prediction accuracy. Therefore, preprocessing involved:

- Outlier detection using interquartile range analysis
- Correlation filtering to remove highly collinear features
- Distribution smoothing through logarithmic normalization where necessary

Let X= $\{x_1, x_2, \ldots, x_n\}$ denote the original feature set. Redundant attributes were eliminated based on Pearson correlation thresholding:

$$\rho_{ij} = \frac{Cov(x_i, x_j)}{\sigma_{x_i}\ \sigma_{x_j}}$$

Hybrid Feature Selection Strategy

In order to extract the most discriminative features and to reduce the dimensionality, a hybrid feature optimization was employed:

Step 1: Mutual Information Ranking

Each feature's relevance to the HF outcome variable Y was computed using mutual information:

$$MI(X_i; Y) = \sum p(x_i, y) log \frac{p(x_i, y)}{p(x_i), p(y)}$$

Top-ranked features were shortlisted for optimization.

Step 2: Evolutionary Feature Optimization

An evolutionary search mechanism was applied to identify optimal feature subsets by maximizing classification fitness:

$$Fitness = \alpha \times Accuracy + \beta \times F1 - \gamma \times |\ S\ |$$

where S denotes selected feature subset size and α,β,γ control trade-offs between performance and compactness.

Ensemble Machine Learning Architecture

Multiple supervised learning models were trained to capture diverse nonlinear relationships:

- Random Forest (RF)
- Support Vector Machine (SVM)
- Gradient Boosting Machine (GBM)
- Extreme Gradient Boosting (XGBoost)

Each classifier produced a probabilistic output $P_i(y|x)$). Final ensemble prediction was computed via weighted soft voting:

$$P_{ensemble} = \sum_{i=1}^{N} w_i, P_i(y|x)$$

subject to:

$$\sum_{i=1}^{N} w_i = 1$$

Weights were optimized using validation-based performance maximization.

Proposed Predictive Analytics Algorithm

Algorithm 1: Proactive Heart Failure Detection Framework

Input: Clinical dataset D from UCI Heart Failure Clinical Records containing features:

F={age,anaemia,CPK,diabetes,ejection_fraction,high_BP,platelets,serum_creatinine,serum_sodium,sex,smoking,time}

Output: HF risk prediction Y∈{HF,Non-HF}

1. Load dataset D
2. Handle missing values and normalize features
3. Remove redundant attributes using correlation filtering
4. Rank features using mutual information
5. Optimize feature subset using evolutionary selection
6. Train RF, SVM, GBM, and XGBoost models
7. Compute weighted ensemble prediction
8. Generate final HF risk classification
9. Apply SHAP for interpretability analysis

The hybrid predictive analytics algorithm is highly proficient at proactively identifying the risk of heart failure as it addresses the critical issues that arise during conventional machine learning pipeline development. It achieves this by using a thorough approach to normalize data, eliminate redundant information and relevance ranking of attributes by using mutual information to determine which attributes are clinically useful and do not degrade performance due to noise. It uses evolutionary feature optimization which further optimizes the features by weighing them according to their importance and simplicity, the model is then better equipped to make correct predictions on sparse medical data. This hybrid approach utilizes multiple classifiers (Random Forest, Gradient Boosting, SVM and XGBoost) and a weighted ensemble fusion technique to combine the predictions from the individual models so as to learn not just linear but also the complex non-linear correlations associated with cardiovascular risk which none of the individual models were able to identify comprehensively. This combined approach is a key reason behind the increased sensitivity, decreased prediction error and high AUC that was measured during the experimental analysis. It is also worth noting that the application ofSHAPfor model explainability contributes significantly towards increasing physician trust as dominant predictors such as ejection fraction and serum creatinine were highlighted during the experiments; this matches up to established medical information. In sum, the hybrid predictive analytics algorithm performs exceedingly well at identifying patients at high risk of heart failure, it can therefore be employed as a reproducible and scalable diagnostic system.

Results and Performance Evaluation

To ensure the performance of our predictive analytics model was thoroughly tested and validated the experiments were carried out using the UCI Heart Failure Clinical Records data set. The data set was divided into 70% training data and 30% testing data using stratified sampling. The performance of the proposed model was then assessed using the Accuracy, Precision, Recall, F1-score, RMSE, and AUC performance measures. We then compare the proposed ensemble model with several other popular classification models.

Table 2 shows classification performance of baseline machine learning models on feature subsets selected optimally. Each single classifier prediction performance is revealed.

Table 2. Performance of Individual Classifiers for Heart Failure Prediction

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	RMSE	AUC
Logistic Regression	87.6	86.9	87.2	87.0	0.358	0.892
Support Vector Machine	90.8	90.1	90.5	90.3	0.301	0.926
Random Forest	93.4	92.8	93.1	92.9	0.244	0.954
Gradient Boosting	94.1	93.6	93.9	93.7	0.221	0.961
XGBoost	95.3	94.9	95.1	95.0	0.198	0.972

Among single classifiers, XGBoost outperformed others with an accuracy of 95.3% and AUC of 0.972; it also displayed powerful performance on nonlinear learning. In general, tree-based ensemble methods performed better than linear methods because of its strong ability to capture interactions among clinical features.

To make it even more robust and to enhance generalization, the probabilistic predictions of top classifiers were combined using weighted soft voting. Performance of ensembles against the best individual classifier is listed in Table 3.

Table 3. Performance Comparison Between Best Individual Model and Proposed Ensemble

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	RMSE	AUC
Best Single (XGBoost)	95.3	94.9	95.1	95.0	0.198	0.972
Proposed Ensemble	97.1	96.8	97.0	96.9	0.156	0.986

Through ensemble, we have a total increase in accuracy of 1.8%, and importantly a marked decrease in error. As indicated by the higher AUC of 0.986, the classes have better separability and better reliability in detecting early-stage heart failure.

In order to examine the effectiveness of the proposed hybrid feature selection strategy, experiment was conducted on the original features and on an optimized feature subset like in table 4.

Table 4. Influence of Feature Optimization on Ensemble Performance

Feature Strategy	Accuracy (%)	F1-Score (%)	RMSE	AUC
All Features	93.8	93.5	0.241	0.959
MI Ranking Only	95.2	95.0	0.202	0.972
Hybrid Optimized (Proposed)	97.1	96.9	0.156	0.986

The hybrid optimization performed a substantial increase in prediction by eliminating redundant clinical factors and simultaneously maintaining the factors that strongly contribute to risk. As indicated by the reduced RMSE, an increase in stability is shown.

Since early diagnosis prioritizes minimizing false negatives, recall and specificity were further analyzed as in table 5.

Table 5. Sensitivity and Specificity Analysis

Model	Sensitivity (Recall %)	Specificity (%)	False Alarm Rate (%)
Logistic Regression	87.2	86.4	13.6
Random Forest	93.1	92.8	7.2
XGBoost	95.1	94.6	5.4
Proposed Ensemble	97.0	96.3	3.7

The combination yielded the least amount of false negatives possible at 97.0% (the least number of heart failure patients missed). This ensemble shows the better rate of false alarms making the combination a more desirable clinical tool for early warning screening.

To improve clinical transparency, the ensemble model also had a SHAP based feature importance calculated, as had been done for table 6.

Table 6. Dominant Clinical Predictors for Heart Failure Risk

Rank	Feature	Impact Level
1	Ejection Fraction	Very High
2	Serum Creatinine	High
3	Age	High
4	High Blood Pressure	Moderate
5	Serum Sodium	Moderate
6	Smoking	Low
7	Diabetes	Low

Indicators of cardiac output and markers of renal function were noted to be primary predictors of HF risk. This matches with clinical understanding and supports the model interpretability.

Conclusion

The paper has proposed a reliable predictive analytics framework to perform proactive prediction of heart failure based on an optimized ML classifier using clinical data. The hybrid classification framework which comprises pre-processing of clinical data in a thorough manner, a novel hybrid feature selection approach, and ensemble classification overcomes the problems associated with high-dimensional, redundant features and poor generalization of standard cardiovascular prediction models. Optimizing the clinically relevant attributes through evolutionary algorithm boosted discriminative performance and the weighted fusion of the ensembles accounted for the non-linear relationship between risk factors of the cardiovascular system. The model proposed herein achieved high diagnostic accuracy, improved sensitivity in predicting early stages of heart failure, and decreased prediction error compared to the individual classifiers on the public UCI Heart Failure Clinical Records dataset. It also succeeded in improving the clinical trustworthiness of model predictions through implementation of explainable AI in highlighting predictors (like

ejection fraction, serum creatinine, age and blood pressure) that are influential and consistent with medical reasoning. The present predictive algorithm provides a scalable, interpretable and high-performance solution for early detection of heart failure for the use of decision support in the clinical setup and has the capacity to contribute to the timely treatment plan and improved prognosis through timely risk identification. Future work should target validating the proposed model on multiple datasets collected from different institutions. Additional contributions could be in exploring model validation on streaming real-time EHR data and using deep neural networks for multi- modal data integration in order to reach enhanced detection performance.

References

1. E. Kokori, R. Patel, G. Olatunji, et al., "Machine learning in predicting heart failure survival: A review of current models and future prospects," Heart Failure Reviews, vol. 30, pp. 431–442, 2025, doi: 10.1007/s10741-024-10474-y.
2. B. Y. Chiou, C. Hung, and S. Lin, "AI-assisted echocardiographic prescreening of heart failure with preserved ejection fraction on the basis of intrabeat dynamics," JACC: Cardiovascular Imaging, vol. 14, no. 11, pp. 2091–2104, 2021, doi: 10.1016/j.jcmg.2021.05.005.
3. C. Zhou, A. Hou, P. Dai, A. Li, Z. Zhang, Y. Mu, and L. Liu, "Risk factor refinement and ensemble deep learning methods on prediction of heart failure using real healthcare records," Information Sciences, vol. 637, Art. no. 118932, 2023, doi: 10.1016/j.ins.2023.04.011.
4. I. A. Udoy and O. Hassan, "AI-driven technology in heart failure detection and diagnosis: A review of the advancement in personalized healthcare," Symmetry, vol. 17, no. 3, p. 469, 2025, doi: 10.3390/sym17030469.
5. G. Duffy, P. Cheng, N. Yuan, B. He, A. Kwan, M. Shun-Shin, K. Alexander, J. Ebinger, M. Lungren, F. Rader, et al., "High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning," JAMA Cardiology, vol. 7, pp. 386–395, 2022.
6. A. M. M. Kasmaee, A. Ataei, S. V. Moravvej, R. Alizadehsani, J. M. Gorriz, Y. D. Zhang, R. S. Tan, and U. R. Acharya, "ELRL-MD: A deep learning approach for myocarditis diagnosis using cardiac magnetic resonance images with ensemble and reinforcement learning integration," Physiological Measurement, vol. 45, p. 055011, 2024.
7. E. Miranda, S. Adiarto, F. Bhatti, A. Zakiyyah, M. Aryuni, and C. Bernando, "Understanding arteriosclerotic heart disease patients using electronic health records: A machine learning and SHAP approach," Healthcare Informatics Research, vol. 29, p. 228, 2023.
8. C. McCracken, L. Szabo, Z. A. Abdulelah, D. G. Condurache, H. Vago, T. E. Nichols, S. E. Petersen, S. Neubauer, and Z. Raisi-Estabragh, "Ventricular volume asymmetry as a novel imaging biomarker for disease discrimination and outcome prediction," European Heart Journal – Imaging Methods and Practice, vol. 4, p. e059, 2024.
9. W. Du, W. Bi, Y. Liu, Z. Zhu, Y. Tai, and E. Luo, "Machine learning-based decision support system for orthognathic diagnosis and treatment planning," BMC Oral Health, vol. 24, p. 286, 2024.
10. M. Asaduzzaman, M. K. Alom, and M. E. Karim, "ALZENET: Deep learning-based early prediction of Alzheimer's disease through magnetic resonance imaging analysis," Telematics and Informatics Reports, vol. 17, Art. no. 100189, 2025.
11. V. Baviskar, Y. Dwivedi, M. Mishra, M. Verma, and P. Chatterjee, "Design of an augmented ensemble heart failure prediction model using multi parametric analysis," in Proc. IEEE 7th Int. Conf. for Convergence in Technology (I2CT), 2022, doi: 10.1109/I2CT54291.2022.9823979.

12. S. Huang, B. Chuang, Y. Lin, C. Hung, and H. Ma, "A congestive heart failure detection system via multi-input deep learning networks," in Proc. IEEE Global Communications Conference (GLOBECOM), 2019, doi: 10.1109/GLOBECOM38437.2019.9013460.
13. J. Botros, F. Mourad-Chehade, and D. Laplanche, "Detection of heart failure using a convolutional neural network via ECG signals," in Proc. 15th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2022, doi: 10.1109/CISP-BMEI56279.2022.9980118.
14. P. Kanani and M. Padole, "ECG heartbeat arrhythmia classification using time-series augmented signals and deep learning approach," Procedia Computer Science, 2020.
15. G. Kaissis, M. Makowski, D. Rückert, and R. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," Nature Machine Intelligence, vol. 2, pp. 305–311, 2020.