Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

Systematic Optimization of Ensemble Learning for Heart Failure Survival Prediction using SHAP and Optuna

Bayu Setia[1], Umar Zaky*[2]

[1]Informatics, Universitas Teknologi Yogyakarta, Indonesia
[2]Information Systems, Universitas Teknologi Yogyakarta, Indonesia

Email: [2]umarzaky@uty.ac.id

Received : Sep 23, 2025; Revised : Sep 26, 2025; Accepted : Oct 29, 2025; Published : Oct 31, 2025

Abstract

Heart failure (HF) stands as a major global health problem where precise and early prediction of patient prognosis is essential for improving clinical management and patient care. A common obstacle for standard machine learning models in this domain is the prevalent issue of class imbalance within clinical datasets. To overcome this challenge, this study introduces a systematically optimized ensemble learning model for the accurate classification of patient survival. The methodology was applied to a publicly accessible clinical dataset of 299 heart failure patients. Its comprehensive framework included logarithmic transformation, stratified data splitting (80:20), SHAP-based selection of eight key features, and hyperparameter tuning with Optuna over 75 trials, with the specific objective of maximizing the F1-score using 10-fold cross-validation. The performance of three ensemble models (Random Forest, XGBoost, and LightGBM) was refined using decision threshold tuning. The results revealed that the fully optimized Random Forest model yielded superior outcomes, attaining an accuracy of 96.67%, an F1-score of 0.9474, and precision and recall values of 0.95, demonstrating high reliability with only a single instance of a False Negative and False Positive. The study concludes that the systematic application of SHAP, SMOTE, and Optuna within an ensemble framework substantially improves classification performance for imbalanced HF data, surpassing existing benchmarks. This work thus provides a replicable and systematic framework for developing reliable machine learning models from complex, imbalanced medical datasets, contributing a valuable methodology to the field of computational science.

Keywords : *Ensemble Learning, Heart Failure, Optuna, SHAP, SMOTE.*

This work is an open access article licensed under a Creative Commons Attribution 4.0 International License.

1. INTRODUCTION

Heart failure (HF) is a multifaceted clinical syndrome that poses a major challenge to public health worldwide. With a rising prevalence, better prognostic tools are essential for the efficient allocation of healthcare resources [1]. The condition is a primary driver of hospital admissions and severely impacts patients' quality of life by compromising both their physical and mental well-being [2]. Consequently, accurate survival prediction is of critical importance, as it empowers clinicians to tailor treatment strategies and optimize care intensity for better patient outcomes. Identifying high-risk patients at an early stage facilitates personalized interventions, a practice known to improve survival rates and quality of life [3]. In this context, computer-assisted predictive models act as essential tools for frontline clinicians, enabling early identification and intervention for at-risk patients [4]. The application of machine learning (ML) has emerged as a promising approach in this domain, utilizing extensive data from electronic health records (EHRs) to build predictive models [5]. Compared to traditional statistical methods, which are often limited to simpler data structures, ML models frequently provide more accurate risk predictions when applied to large and complex datasets [6]. This advantage stems from the ability of ML algorithms to automatically learn and map the intricate relationships between variables within large-scale data, often surpassing the performance of conventional models [7]. This capability

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

includes handling high-dimensional data, such as protein networks in cardiac remodeling [8], because clinical HF data is often structurally complex, containing interactive and non-linear information that is difficult for traditional analyses to process effectively [9]. Despite these capabilities, the performance of such models faces inherent challenges in medical data, most notably class imbalance, where one outcome class significantly outnumbers the other. Moreover, the opaque, "black-box" nature of many sophisticated algorithms can hinder their adoption in clinical practice, as the logic behind their predictions often lacks transparency [10].

Several previous studies have applied various methods to classify heart failure patient outcomes. For instance, a study focusing on imbalanced data handling using the Balanced Random Forest (BRF) method achieved an accuracy of 76.25% [11], while a standard Random Forest implementation on the same dataset yielded an accuracy of 86.62% [12]. Other advanced approaches have shown better results; a study focused on wrapper feature selection combining MLP and BPSO reached an accuracy of 91.11% [13], and the use of an Extra Tree Classifier was reported to obtain an accuracy of 92.62% [14]. Further optimization efforts, such as integrating Random Forest with a Genetic Algorithm (GA), successfully pushed performance further to 93.36% [15].

Despite these advancements, a significant research gap exists in the systematic and comprehensive optimization of these models. Many studies tend to apply one or two optimization techniques in isolation, such as focusing solely on imbalance handling or only on feature selection. This is a critical oversight, as effective feature selection not only simplifies the model and improves computational speed but can also enhance predictive performance by reducing the curse of dimensionality [16]. Furthermore, many existing predictive models are validated only on their original dataset, with a lack of independent external verification to truly assess their generalizability [17]. Very few have integrated a complete pipeline that includes feature engineering, robust feature selection, imbalance handling applied specifically to the training data, automated hyperparameter tuning, and decision threshold optimization. This lack of a holistic approach leaves potential performance improvements unexploited.

Unlike previous studies that tended to apply optimization techniques in isolation, this study proposes and validates a comprehensive, holistic framework for optimizing ensemble learning models. This research represents one of the first efforts to systematically integrate a full pipeline, which includes feature engineering, SHAP-based feature selection, SMOTE for imbalance handling, automated hyperparameter tuning with Optuna, and decision threshold optimization. The focus is on ensemble learning, a method where several individual learners are combined into one stronger model. This approach is known for its ability to increase robustness, mitigate overfitting, and often lead to superior predictive outcomes [18]. Furthermore, relying on a combination of algorithms, rather than a solitary one, has been demonstrated to enhance both predictive accuracy and sensitivity [19]. Therefore, the primary contribution of this work is to show that this systematically optimized pipeline can yield a high-performance model for heart failure survival classification.

2. METHOD

This study presents a structured framework, fully developed in Python, to build and assess classification models for predicting the mortality risk in heart failure patients. The end-to-end process, from data acquisition to performance evaluation, relied on a suite of essential machine learning libraries. These included Scikit-learn for data preprocessing and evaluation metrics, imblearn for handling class imbalance with SMOTE, XGBoost and LightGBM for model construction, Optuna for hyperparameter optimization, and SHAP for feature selection.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

2.1. Dataset

This study makes use of a publicly available dataset [20]. The data consists of 299 records from heart failure patients, each described by 13 attributes. The patient population consists of individuals diagnosed with advanced stages of heart failure, originally collected in Faisalabad, Pakistan [21]. The target for classification is the binary DEATH_EVENT feature, where '1' denotes mortality and '0' denotes survival. The data is significantly imbalanced, containing 203 'survived' instances (67.9%) and 96 'deceased' instances (32.1%), a factor that heavily influenced the methodological design. A thorough explanation of the utilized features is summarized in Table 1.

Table 1. Summary of Dataset Features

Feature	Description	Variable Type	Unit / Value
age	Age of the patient	Integer	Years
anaemia	Presence of anaemia	Binary	0 = No, 1 = Yes
creatinine phosphokinase	Blood concentration of the CPK enzyme	Integer	mcg/L
diabetes	Presence of diabetes	Binary	0 = No, 1 = Yes
ejection fraction	Ventricular ejection percentage per heartbeat	Integer	Percentage (%)
high blood pressure	Presence of hypertension	Binary	0 = No, 1 = Yes
platelets	Blood platelet count	Float	kiloplatelet/mL
serum creatinine	Serum creatinine concentration	Float	mg/dL
serum sodium	Serum sodium concentration	Integer	mEq/L
sex	Gender of the patient	Binary	0 = Female, 1 = Male
smoking	Patient's smoking status	Binary	0 = No, 1 = Yes
time	Duration of follow-up	Integer	Days
DEATH EVENT	Target Variable: Patient mortality during follow-up	Binary	0 = No, 1 = Yes

2.2. Research Framework

The research methodology is designed as a structured and chronological workflow to ensure reproducibility and optimal predictive performance. This framework, as illustrated in Figure 1, is divided into two main pipelines.

1. Training Pipeline
   Focuses on training data preparation, feature selection, imbalanced class handling, and hyperparameter optimization to generate the most optimal model.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

2. Evaluation Pipeline
   To measure the robustness and generalization of the final model, a separate test set that was not involved in training is applied.

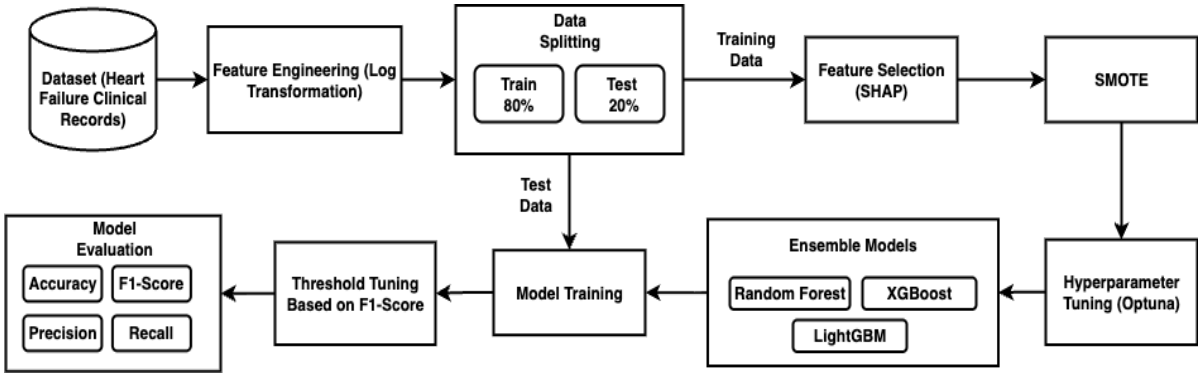

Figure 1. The proposed research workflow.

2.3. Data Preprocessing and Feature Selection

The next step in the process was data preparation. First, simple feature engineering was performed by applying a logarithmic transformation to three features (creatinine phosphokinase, platelets, and serum creatinine). This step aimed to normalize the highly skewed data distributions, which can often improve model performance. The need for this transformation is supported by previous analyses of this dataset, which noted that features like creatinine phosphokinase have distributions with a few extremely high values, characteristic of a skewed distribution [22]. To maintain class balance, the dataset was subjected to stratified partitioning, creating a training set with 80% of the data and a test set with the remaining 20%, which is a conventional split supported by prior studies [23]. This method was chosen to ensure that the proportion of the DEATH EVENT class in both the training and test sets remained the same as in the original dataset, a crucial step for valid evaluation on imbalanced data.

Feature selection was performed on the training set to optimize both interpretability and efficiency. Algorithms belonging to the tree-based ensemble family (e.g., Random Forest and XGBoost) inherently yield feature importance values, offering valuable insights into critical variables and informing model construction [24]. Building on this principle, this study utilized SHAP (Shapley Additive exPlanations), a novel approach designed to explain the outputs of complex "black-box" models [25]. A key advantage of SHAP is that it is a model-agnostic technique, making it universally applicable for interpreting a wide range of machine learning models [26]. SHAP was selected for its ability to provide accurate and consistent justifications for the predictive contribution of each feature. Specifically, SHAP values quantify the contribution of each feature to a given prediction, where the magnitude of the value indicates the influence's strength and its sign indicates the direction of the effect [27]. Moreover, SHAP is a favored method as its additive feature attribution approach provides explanations that are relatively consistent with human intuition [28]. This allows for interpretation at a local level, meaning the impact of the features on each individual prediction can be precisely calculated [29]. From this process, the eight features with the most significant contributions were identified and subsequently used in all following modeling stages.

2.4. Model Training and Optimization

To ensure consistent and reproducible results, a random_state of 123 was used throughout all experiments involving stochastic processes, such as in data splitting, SMOTE, and model initialization. The entire training and optimization process was performed exclusively on the training data. To mitigate unequal class representation, Synthetic Minority Oversampling Technique (SMOTE) was employed.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

Through this approach, new artificial instances of the minority outcome (deceased patients) were generated, resulting in a more balanced dataset for learning. Employing SMOTE is an effective strategy to prevent models from developing a bias towards the majority class. Notably, this method has been shown to be highly beneficial for tree-based ensemble classifiers like Random Forest, as it can substantially elevate their predictive performance on this type of clinical data [30].

Modern ensemble models such as LightGBM and XGBoost contain numerous hyperparameters, and since manual adjustment is often cumbersome and inefficient, an automated optimization framework is preferable [31]. After the training data was balanced, automated hyperparameter tuning was performed using the Optuna framework. Optuna is an advanced framework that efficiently searches for optimal parameters by dynamically adjusting its search space based on the results of previous trials, which improves both efficiency and model performance [32]; this approach utilizes Bayesian Optimization and is more efficient than traditional methods like GridSearch [33]. This is achieved through two main components: a sampling algorithm that intelligently selects the next hyperparameters to test based on historical trial data, and a pruning algorithm that can terminate unpromising trials early to save computational time [34]. Finding the optimal set of hyperparameters is a critical step, as ideal parameters for models like Random Forest and XGBoost can vary significantly depending on the optimization search strategy employed [35]. This process aimed to find the optimal hyperparameter combination for three ensemble models (Random Forest, XGBoost, and LightGBM), which have demonstrated high performance in various clinical studies. These models are all powerful ensemble techniques; Random Forest operates by combining decision tree results through majority voting, while XGBoost and LightGBM are gradient boosting methods that iteratively build a strong predictive model from a series of weaker ones [36]. The optimization was conducted for 75 trials with the primary objective of maximizing the F1-score. For this process, the Optuna framework was configured to use a Tree-structured Parzen Estimator (TPE) sampler and a Successive Halving pruner to efficiently search the hyperparameter space. A 10-fold cross-validation scheme was employed within each trial to ensure that the performance evaluation was robust and to mitigate the risk of overfitting. This use of k-fold cross-validation is a robust method for reliably assessing model performance during the hyperparameter tuning process [37]. Once the optimal hyperparameter configuration was found, the three models were retrained using the entire SMOTE-processed training dataset.

2.5. Model Testing and Performance Evaluation

In the final stage, the trained models were evaluated against the unseen test set to assess their performance. Initially, decision threshold tuning was applied to the model's output probabilities. This was a necessary step, as the standard 0.5 threshold can be ineffective for imbalanced data; the objective was to identify a threshold that optimized the F1-score. Model performance was ultimately quantified using four standard metrics. The metrics are determined using the standard outputs of a confusion matrix (TP, TN, FP, FN). Choosing these measures aligns with prior work in this field [38].

1. Accuracy
   Accuracy measures the overall correctness of the model, calculated as the ratio of all correct predictions to the total number of samples, as defined in Equation (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. Precision
   Precision evaluates the accuracy of the positive predictions. It is the ratio of true positives to the total number of instances predicted as positive, as detailed in Equation (2).

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

3. Recall (Sensitivity)
   Recall, also known as Sensitivity, determines the model's ability to identify all relevant instances. It is calculated as the ratio of true positives to the total number of actual positive instances, as shown in Equation (3).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

4. F1-Score
   The F1-Score provides a single metric that balances Precision and Recall by calculating their harmonic mean, as detailed in Equation (4).

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The F1-score was selected as the primary evaluation metric for this study. Its choice is justified by its capacity to offer a more balanced and dependable evaluation than accuracy on imbalanced datasets, especially in a clinical context where the consequences of misclassifying the minority (death) class are more severe.

3. RESULT

The primary quantitative outcomes of this research, which fulfill the study's objective of creating a high-performance classification framework, are presented in this chapter. The findings are organized to mirror the methodological workflow, commencing with the identification of the most predictive features, proceeding to an evaluation of model performance in incremental scenarios, and culminating in a detailed analysis of the top-performing model. The results detailed herein serve to validate the contribution of the proposed optimization pipeline.

3.1. Results of Feature Selection

The optimization pipeline commenced with identifying the most predictive features via SHAP analysis on the training data. Figure 2 presents the SHAP feature importance plot, which establishes a clear hierarchy by ordering features based on their mean absolute SHAP values. Such a ranking is instrumental in discerning the most influential clinical factors for the model's predictive process [39].

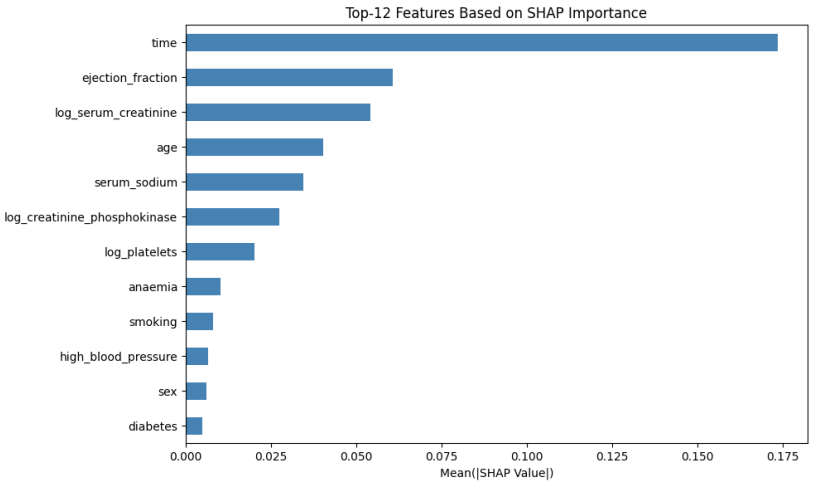

Figure 2. SHAP feature importance ranking

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

The SHAP analysis pinpointed time, ejection fraction, and log serum creatinine as the three most influential predictors. Following this, the essential task was to ascertain the optimal quantity of top features (top-k) for the model. An experiment was therefore conducted to determine this, with its results summarized in Table 2. In this study, attention was directed toward Random Forest, which was selected because of its favorable early outcomes.

Table 2. Impact of Top-k Feature Selection on Random Forest Performance

Number of Features	Accuracy	F1-Score	Precision	Recall
12	0.9500	0.9143	1.00	0.84
11	0.9333	0.8947	0.89	0.89
10	0.9333	0.8947	0.89	0.89
9	0.9333	0.8889	0.94	0.84
8	**0.9667**	**0.9474**	**0.95**	**0.95**
7	0.9500	0.9189	0.94	0.89
6	0.9500	0.9143	1.00	0.84
5	0.9333	0.8889	0.94	0.84
4	0.9333	0.8947	0.89	0.89
3	0.9167	0.8718	0.85	0.89

The data in Table 2 shows that the model's peak performance was achieved when using the top 8 features, with a peak F1-score of 0.9474 and an accuracy of 0.9667. Using more than eight features did not yield a significant improvement and instead posed a risk of overfitting. Therefore, the following eight features were established as the final feature set: time, ejection fraction, log serum creatinine, age, serum sodium, log creatinine phosphokinase, log platelets, and anaemia.

3.2. Incremental Model Performance Analysis

In order to assess how each optimization stage influenced the results, five successive experimental scenarios were designed. Table 3 provides a comparative overview of the three ensemble algorithms, demonstrating how every optimization step contributed to the overall performance.

Table 3. Performance comparison of models in each experimental scenario

Experimental Scenario	Model	Accuracy	F1-Score	Precision	Recall	Top-k
Baseline	RF	0.9167	0.8649	0.89	0.84	12
	XGB	0.8667	0.7778	0.82	0.74	12
	LGBM	0.9000	0.8421	0.84	0.84	12
FE (Log Trans)	RF	0.9333	0.8889	0.94	0.84	12
	XGB	0.9167	0.8649	0.89	0.84	12
	LGBM	0.9167	0.8718	0.85	0.89	12
SHAP Selection	RF	0.9167	0.8649	0.89	0.84	3
	XGB	0.9167	0.8649	0.89	0.84	9
	LGBM	0.9333	0.8947	0.89	0.89	9
Threshold Tuning	RF	0.9333	0.8889	0.94	0.84	3
	XGB	0.9333	0.9000	0.86	0.95	9
	LGBM	0.9500	0.9189	0.94	0.89	9
HPO (Optuna)	**RF**	**0.9667**	**0.9474**	**0.95**	**0.95**	**8**
	XGB	0.9500	0.9143	1.00	0.84	10
	LGBM	0.9500	0.9143	1.00	0.84	6

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

An in-depth analysis of Table 3 reveals a systematic improvement in model performance at each stage:

1. Baseline Performance
   As a baseline, the models were first tested using the full set of 12 features, where the Random Forest (RF) achieved an F1 score of 0.8649.
2. Impact of Feature Engineering
   Applying a logarithmic transformation to handle skewed data provided a notable improvement, increasing the RF model's F1-Score to 0.8889.
3. Significance of Threshold Tuning
   This stage yielded a substantial performance boost across the models, with the XGBoost F1-Score, for example, jumping to 0.9000. This highlights that optimizing the decision threshold is a critical step for imbalanced datasets.
4. Final Optimization and Peak Performance
   In the final optimization stage using HPO with Optuna, the Random Forest model attained its highest performance level. The fully optimized model yielded an Accuracy of 0.9667 and an F1-score of 0.9474 using an optimal subset of 8 features, which confirmed its status as the superior model.

3.3. Evaluation and Performance Analysis of the Best Model

The fully optimized Random Forest model, which emerged as the superior performer, underwent a detailed performance analysis on the test set using the confusion matrix presented in Figure 3.

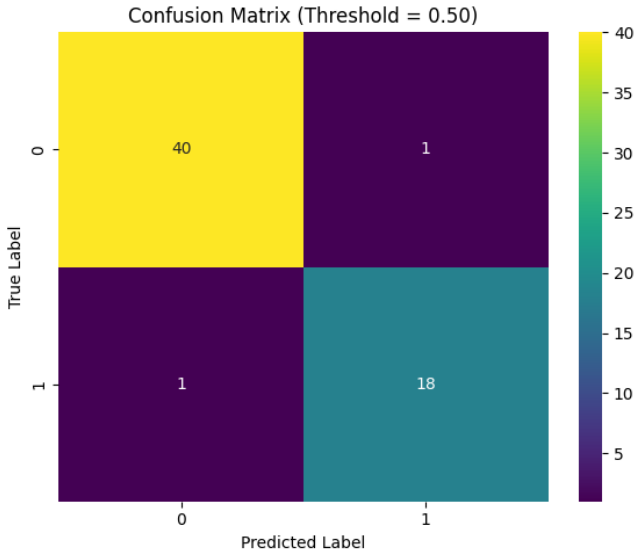

Figure 3. Confusion Matrix for the Optimized Random Forest (Test Set)

An examination of Figure 3 reveals the model's strong capacity for generalization and leads to the subsequent conclusions:

1. High Accuracy
   An accuracy of 96.67% was attained by the model (58 out of 60 correct predictions) on the previously unseen test data.
2. Superior Clinical Sensitivity
   The model achieved a Recall (Sensitivity) of 94.7% by correctly identifying 18 out of 19 death cases. This high True Positive Rate is vital in a medical context, where the ability to detect high-risk patients is a top priority.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

3. Balanced Performance
   An F1-score of 0.9474 indicates that the model maintains a strong equilibrium between Precision and Recall, which confirms its robust performance on the imbalanced dataset.

To further assess the model's discriminative ability, especially concerning the imbalanced nature of the dataset, the Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve were analyzed, as shown in Figure 4. The model demonstrated excellent classification capability with an Area Under the ROC Curve (AUC) of 0.96. Furthermore, the PR curve, which is particularly informative for imbalanced data, showed a high Area Under the PR Curve (AUPRC) of 0.96, confirming that the model maintains high precision even at high recall rates. These results provide strong evidence of the model's robust generalization performance.

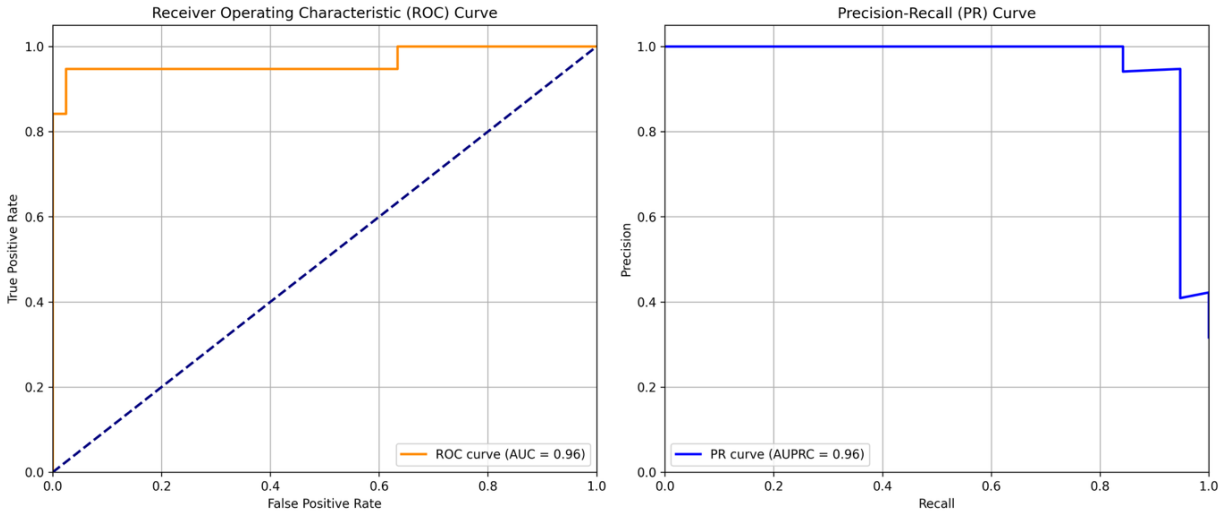

Figure 4. ROC and PR Curves for the Optimized RF Model (Test Set).

The success of the proposed framework is demonstrated by its ability to produce a model with both high overall accuracy and crucial sensitivity to the minority (death) class, which represents the key priority in this clinical context.

4. DISCUSSIONS

This study demonstrates that the systematic optimization framework successfully achieved an accuracy of 96.67%, a performance that significantly surpasses previous studies, including the prior best method, which achieved 93.36% (Table 4). This success is attributed to the effective and comprehensive pipeline: SHAP feature selection successfully identified clinically relevant predictors (such as ejection fraction and serum creatinine), while SMOTE and Optuna ensured the model could optimally recognize patterns in the minority class (death), as reflected by the high F1-Score.

Beyond achieving high predictive accuracy, the primary contribution of this research to the field of computer science lies in two key areas. First, by integrating SHAP as a core component, this study contributes to the growing body of work on Explainable AI (XAI) in medicine. It demonstrates how interpretability can be built into the optimization pipeline to address the critical 'black-box' problem, thereby increasing the trustworthiness of complex models for clinical applications. Second, the proposed holistic pipeline serves as a replicable and systematic framework that can be adapted for developing high-performance predictive models in other medical domains that face similar challenges with complex and imbalanced data.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

Table 4. Comparison of accuracy with previous studies

Reference Sources	Method	Accuracy (%)
[14]	Extra Tree Classifier	92.62
[11]	Balanced Random Forest (BRF)+Chi2	76.25
[15]	Random Forest+GA	93.36
[12]	Random Forest	86.62
[13]	Multilayer Perceptron+BPSO	91.11
This study	**Random Forest+SHAP+Optuna**	**96.67**

Despite the model's excellent performance, a key limitation of the study is the modest size of the dataset, a factor that may reduce the model's generalization capacity when applied to broader patient populations. The occurrence of one False Negative prediction, when considered alongside this limitation, reinforces the model's intended role as a clinical decision support tool rather than a substitute for professional medical judgment. Therefore, the crucial next step involves validating the framework on more extensive and varied external datasets before considering further implementation.

5. CONCLUSION

This study concludes that a holistic optimization framework, which integrates feature engineering, SHAP-based feature selection, data balancing with SMOTE, and hyperparameter tuning with Optuna, is a highly effective strategy for enhancing predictive performance in heart failure survival classification. Of the models evaluated, the fully optimized Random Forest model demonstrated superior performance. The principal contribution of this work is the validation of this systematic framework. As a contribution to the field of medical informatics, this framework demonstrates how a systematic optimization approach can advance the development of high-performing and interpretable machine learning models. This is evidenced by the final model, which achieved a highly competitive accuracy of 96.67% and an F1-Score of 0.9474, outperforming previously established benchmarks. This result underscores that a comprehensive and systematic optimization approach is crucial for unlocking the full potential of machine learning on complex clinical data. Although these findings are promising, future investigations should prioritize assessing the framework's performance on larger and more diverse patient cohorts to establish its generalizability before its potential for clinical application is fully realized.

REFERENCES

[1] S. König *et al.*, "From population-to patient-based prediction of in-hospital mortality in heart failure using machine learning," *European Heart Journal - Digital Health*, vol. 3, no. 2, pp. 307–310, Jun. 2022, doi: 10.1093/ehjdh/ztac012.

[2] C. Zheng *et al.*, "Time-to-event prediction analysis of patients with chronic heart failure comorbid with atrial fibrillation: a LightGBM model," *BMC Cardiovascular Disorders*, vol. 21, no. 1, p. 379, 2021, doi: 10.1186/s12872-021-02188-y.

[3] X. Li, C. Shang, C. Xu, Y. Wang, J. Xu, and Q. Zhou, "Development and comparison of machine learning-based models for predicting heart failure after acute myocardial infarction," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 165, 2023, doi: 10.1186/s12911-023-02240-1.

[4] Q. Lin *et al.*, "Predicting the risk of heart failure after acute myocardial infarction using an

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

interpretable machine learning model," *Frontiers in Cardiovascular Medicine, vol. 12, 2025*, doi: 10.3389/fcvm.2025.1444323.

[5] B. Zheng *et al.*, "Prediction of 90 day readmission in heart failure with preserved ejection fraction by interpretable machine learning," *ESC Heart Failure*, vol. 11, no. 6, pp. 4267–4276, Dec. 2024, doi: 10.1002/ehf2.15033.

[6] X. Hou *et al.*, "Prediction of Acute Kidney Injury Following Isolated Coronary Artery Bypass Grafting in Heart Failure Patients with Preserved Ejection Fraction Using Machine Leaning with a Novel Nomogram," *Reviews in Cardiovascular Medicine*, vol. 25, no. 2, p. 43, doi: 10.31083/j.rcm2502043.

[7] F. Li, H. Xin, J. Zhang, M. Fu, J. Zhou, and Z. Lian, "Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database," *BMJ Open*, vol. 11, no. 7, p. e044779, Jul. 2021, doi: 10.1136/bmjopen-2020-044779.

[8] N. Cauwenberghs, F. Sabovčik, A. Magnus, F. Haddad, and T. Kuznetsova, "Proteomic profiling for detection of early-stage heart failure in the community," *ESC Heart Failure*, vol. 8, no. 4, pp. 2928–2939, Aug. 2021, doi: 10.1002/ehf2.13375.

[9] Q. Wang *et al.*, "Machine learning-based risk prediction of malignant arrhythmia in hospitalized patients with heart failure," *ESC Heart Failure*, vol. 8, no. 6, pp. 5363–5371, Dec. 2021, doi: 10.1002/ehf2.13627.

[10] K. Wang *et al.*, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP," *Computers in Biology and Medicine*, vol. 137, p. 104813, 2021, doi: 10.1016/j.compbiomed.2021.104813.

[11] A. Newaz, N. Ahmed, and F. Shahriyar Haq, "Survival prediction of heart failure patients using machine learning techniques," *Informatics in Medicine Unlocked*, vol. 26, p. 100772, 2021, doi: 10.1016/j.imu.2021.100772.

[12] T. A. Assegie, V. Elanangai, J. S. Paulraj, M. Velmurugan, and D. F. Devesan, "Evaluation of feature scaling for improving the performance of supervised learning methods," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1833–1838, Jun. 2023, doi: 10.11591/eei.v12i3.5170.

[13] S. Sutikno, "Combination of Binary Particle Swarm Optimization (BPSO) and Multilayer Perceptron (MLP) for Survival Prediction of Heart Failure Patients", *INFOTEL*, vol. 16, no. 1, pp. 96-104, Feb. 2024, doi: 10.20895/infotel.v16i1.974

[14] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[15] Y. Ramdhani, C. M. Putra, and D. P. Alamsyah, "Heart failure prediction based on random forest algorithm using genetic algorithm for feature selection," *International Journal of Reconfigurable and Embedded Systems*, vol. 12, no. 2, pp. 205–214, Jul. 2023, doi: 10.11591/ijres.v12.i2.pp205-214.

[16] N. Tasnim, S. Al Mamun, M. Shahidul Islam, M. S. Kaiser, and M. Mahmud, "Explainable Mortality Prediction Model for Congestive Heart Failure with Nature-Based Feature Selection Method," *Applied Sciences*, vol. 13, no. 10, p. 6138, 2023, doi: 10.3390/app13106138.

[17] Z. Chen, T. Li, S. Guo, D. Zeng, and K. Wang, "Machine learning-based in-hospital mortality risk prediction tool for intensive care unit patients with heart failure," *Frontiers in Cardiovascular Medicine*, vol. 10, 2023, doi: 10.3389/fcvm.2023.1119699.

[18] M. Tanaka *et al.*, "Development of interpretable machine learning models to predict in-hospital prognosis of acute heart failure patients," *ESC Heart Failure*, vol. 11, no. 5, pp. 2798–2812, Oct. 2024, doi: 10.1002/ehf2.14834.

[19] D. Tu, Q. Xu, Y. Luan, J. Sun, X. Zuo, and C. Ma, "Integrative analysis of bioinformatics and machine learning to identify cuprotosis-related biomarkers and immunological characteristics in heart failure," *Frontiers in Cardiovascular Medicine*, vol. 11, 2024, doi: 10.3389/fcvm.2024.1349363.

[20] "Heart Failure Clinical Records," *UCI Machine Learning Repository*, 2020. doi: 10.24432/C5Z89R.

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

[21] 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, 2020, doi: 10.1186/s12911-020-1023-5.

[22] P. Rahman, A. Rifat, M. IftehadAmjad Chy, M. Monirujjaman Khan, M. Masud, and S. Aljahdali, "Machine Learning and Artificial Neural Network for Predicting Heart Failure Risk," *Computer Systems Science and Engineering*, vol. 44, no. 1, pp. 757–775, 2023. doi: 10.32604/csse.2023.021469

[23] K. Yongcharoenchaiyasit, S. Arwatchananukul, P. Temdee, and R. Prasad, "Gradient Boosting Based Model for Elderly Heart Failure, Aortic Stenosis, and Dementia Classification," *IEEE Access*, vol. 11, pp. 48677–48696, 2023, doi: 10.1109/ACCESS.2023.3276468.

[24] C.-Y. Guo, M.-Y. Wu, and H.-M. Cheng, "The Comprehensive Machine Learning Analytics for Heart Failure," *International Journal of Environmental Research and Public Health*, vol. 18, no. 9, p. 4943, 2021, doi: 10.3390/ijerph18094943.

[25] J. Tian *et al.*, "Machine learning prognosis model based on patient-reported outcomes for chronic heart failure patients after discharge," *Health and Quality of Life Outcomes*, vol. 21, no. 1, p. 31, 2023, doi: 10.1186/s12955-023-02109-x.

[26] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data*, vol. 11, no. 1, p. 44, 2024, doi: 10.1186/s40537-024-00905-w.

[27] Y. Wang *et al.*, "Clinical Prediction of Heart Failure in Hemodialysis Patients: Based on the Extreme Gradient Boosting Method," *Frontiers in Genetics*, vol. 13, 2022, doi: 10.3389/fgene.2022.889378.

[28] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," *Risk Management and Healthcare Policy*, vol. 14, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.

[29] J. Li, S. Liu, Y. Hu, L. Zhu, Y. Mao, and J. Liu, "Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study," *Journal of Medical Internet Research*, vol. 24, no. 8, p. e38082, 2022, doi: 10.2196/38082.

[30] L. T. Ravulapalli, R. K. Paladugu, V. K. Rao Likki, R. Mothukuri, N. Mukkapati, and S. Kilaru, "Evaluative Study of Machine Learning Classifiers in Predicting Heart Failure: A Focus on Imbalanced Datasets," *Ingenierie des Systemes d'Information*, vol. 28, no. 3, pp. 717–724, Jun. 2023, doi: 10.18280/isi.280322.

[31] J. I. E. Yang, J. Yan, Z. Pei, A. Hu, and Y. Zhang, "Prediction Model for In-Hospital Mortality of Patients with Heart Failure Based on Optuna and Light Gradient Boosting Machine," *Journal of Mechanics in Medicine and Biology*, vol. 22, no. 09, p. 2240059, Sep. 2022, doi: 10.1142/S0219519422400590.

[32] P. K. Sahu and T. Fatma, "Optimized Breast Cancer Classification Using PCA-LASSO Feature Selection and Ensemble Learning Strategies With Optuna Optimization," *IEEE Access*, vol. 13, pp. 35645–35661, 2025, doi: 10.1109/ACCESS.2025.3539746.

[33] G. Riski, D. Hartama, and Solikhun, "Optimizing Multilayer Perceptron for Car Purchase Prediction with GridSearch and Optuna," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 2, pp. 266–275, Apr. 2025, doi: 10.29207/resti.v9i2.6328.

[34] R. D. a. Abdu-Aljabar and O. A. Awad, "Improving Lung Cancer Relapse Prediction Using the Developed Optuna_XGB Classification Model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 1, pp. 131–141, 2023, doi: 10.22266/ijies2023.0228.12.

[35] Q. A. Hidayaturrohman and E. Hanada, "A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes," *Applied Sciences*, vol. 15, no. 6, p. 3393, 2025, doi: 10.3390/app15063393.

[36] P. Chen, J. Sun, Y. Chu, and Y. Zhao, "Predicting in-hospital mortality in patients with heart failure combined with atrial fibrillation using stacking ensemble model: an analysis of the medical information mart for intensive care IV (MIMIC-IV)," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 402, 2024, doi: 10.1186/s12911-024-02829-0.

[37] S. M. Al Younis *et al.*, "Investigating automated regression models for estimating left ventricular

Vol. 6, No. 5, October 2025, Page. 5320-5332
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5324

ejection fraction levels in heart failure patients using circadian ECG features," *PLoS ONE*, vol. 18, no. 12, p. e0295653, Dec. 2023, doi: 10.1371/journal.pone.0295653

[38] A. A. Almazroi, "Survival prediction among heart patients using machine learning techniques," *Mathematical Biosciences and Engineering*, vol. 19, no. 1, pp. 134–145, 2022, doi: 10.3934/mbe.2022007.

[39] S. M. Dalhatu and M. A. A. Murad, "A model for enhancing pattern recognition in clinical narrative datasets through text-based feature selection and SHAP technique," *International Journal on Informatics Visualization (JOIV)*, vol. 8, no. 4, pp. 2287–2296, Dec. 2024, doi: 10.62527/joiv.8.4.3664.