IDENTIFYING RISK FACTORS AND SURVIVAL PREDICTION OF HEART FAILURE PATIENTS USING MACHINE LEARNING

Minh Tuan Nguyen, Thang Le Nhat
Posts and Telecommunications Institute of Technology

Abstract – Heart failure (HF) is a prevalent and complex clinical syndrome with high mortality, making accurate survival prediction crucial for patient management. In addressing this challenge, our research introduces a method that integrates exploratory data analysis, feature selection, and machine learning (ML) models to identify significant risk factors for HF events accurately. Optimizing ML models through grid search and 5-fold cross-validation (CV) are used to improve their performance. Our approach identifies combinations of features that comprise important risk factors using four ML models with 5-fold CV. The results highlight important factors that impact HF events including Time, Serum Creatinine, Ejection Fraction, Age, Creatinine Phosphokinase, and Diabetes. Among four models, Random forest model stands out for its robustness in predicting HF mortality. This is demonstrated by the performance of model through validation and testing data. Specifically, the performance on the validation set achieves an accuracy of 84.9%, a precision of 81.71%, a recall 71.02%, and an F1-score 74.94%. On testing set, the performance of model achieves an accuracy of 86.67%, a precision of 81.82%, a recall of 69.23%, and an F1-score of 75%. This results confirm the performance of our proposed method to predict HF events with high accuracy and reliability.

Keywords— Heart failure, Machine learning, Feature selection, Survival prediction

I. INTRODUCTION

Heart failure (HF) is a condition in which the heart fails to pump blood effectively, disrupting the body's metabolic needs. It is a serious and potentially life-threatening condition that often arises from long-term health problems such as coronary heart disease and high blood pressure. With more than 64 million people affected worldwide and an estimated 8.5 million cases in the United States by 2030, HF is a major public health concern [1, 2]. Therefore, early prediction of mortality in HF patients plays a crucial role in reducing the mortality rate and supporting effective disease management.

Due to the various factors that contribute to its causes and the progression of underlying heart conditions, the prediction of HF in the past faced significant difficulties. However, recently, the development of artificial intelligence (AI) technologies has significantly improved the ability to diagnosis and manage HF [3]. Indeed, the field of medical diagnostics has witnessed rapid advancements by using Machine Learning (ML) [4]. Unlike many clinical procedures, ML approaches to diagnose heart disease require only a dataset of relevant information and features to achieve high accuracy [5-7]. ML offers a promising alternative by using advanced computational techniques to analyze complex, multidimensional data [8]. ML has the ability to analyze complex, nonlinear relationships among a multitude of clinical variables. Additionally, ML algorithms can dynamically learn from new data, allowing for real-time updates to predictions and treatment recommendations [9].

In the paper [10], ML models were employed to analyze and evaluate two HF survival prediction models using a dataset of 299 patients. The first model utilized survival analysis with death events and time as target features, while the second approached the problem as a classification task for predicting mortality. Optimization techniques were used to select the best ML algorithms and feature sets. Key findings revealed the Survival Gradient Boosting model and the Random Forest (RF) model as the most balanced for survival analysis and classification, respectively, achieving an accuracy of 0.74.

A thorough survival analysis and survival prediction were conducted in the research [11]. The survival analysis was performed using Kaplan-Meier (KM) estimates and Cox Proportional Hazard regression methods. KM plots were used to show survival estimates as a function of each

Contact author: Thang Le Nhat
Email: thangln@ptit.edu.vn
Manuscript received: 3/2025, revised: 4/2025, accepted: 5/2025.

clinical feature and their impact on survival over time. The Cox regression model analyzed the hazard of death related to clinical features. ML classification models for survival prediction were also built using significant variables identified from the survival analysis and employing algorithms. Subsequently, the Support Vector Machine (SVM) classifiers achieved the highest accuracy of 83.33%.

The authors [12] proposed a reliable decision-support system for predicting HF patients' survival by employing a sampling strategy within an ensemble learning framework, the paper developed a robust RF Classifier. This model addresses the data's imbalanced nature and enhances prediction accuracy through feature selection techniques such as the Chi-square test and Recursive Feature Elimination, achieving a maximum G-mean score of 76.83% and a sensitivity of 80.2%. The paper [13] introduced two hybrid ML methods, Boosting, SMOTE, and Tomek links (BOO-ST) and combining the best-performing conventional classifier with ensemble classifiers (CBCEC), to enhance early detection of HF mortality. BOO-ST addresses class imbalance, and CBCEC optimizes feature selection using feature importance and information gain techniques, aiming for improved prediction accuracy. These approaches demonstrated significant effectiveness, with CBCEC achieving a notable accuracy of 93.67% in predicting HF mortality, underscoring their potential to reduce the death rate and alleviate healthcare sector stress.

Several other studies have explored various approaches. In research [14], six ML classifiers were trained to develop a model for predicting hospital mortality in HF. The authors reported that RF achieved the highest accuracy of 88% during the test phase. Research [15] endeavored to predict changes in left ventricular ejection fraction in HF patients. XGBoost was identified as the highest-performing model, achieving an area under the ROC curve (AUC) of 88.6%. Additionally, utilizing feature importance-based selected features, study [16] achieved accuracy of 76.4% using the XGBoost classifier.

To predict HF survival, it is of crucial importance to analyze the risk factors. Though many studies have been performed on heart disease prediction, very limited work is conducted on the survival prediction of a HF patient. Besides, despite its potential, the integration of ML into clinical practice faces challenges. In particular, modeling the early identification of features associated with mortality remains complex and has yet to achieve consistently high classification accuracy. Motivated by the strengths of ML, in this paper, we employ ML methods to explore risk factors and predict survival outcomes for HF patients, with the goal of reducing HF mortality. A comprehensive analysis is conducted on an HF survival event dataset. After preprocessing the HF dataset, we implement statistical analysis, exploratory data analysis (EDA), identify risk factors based on feature selection, and predict mortality using ML models. Statistical analysis and EDA extract some important information related to HF mortality and survival events. ML methods are then applied to build a model to predict whether an HF patient will survive based on the risk factors identified by feature selection. Our main contributions include:

1) Optimization of ML models using grid search and 5-fold cross-validation to enhance and ensure model performance reliability.
2) Feature selection based on ranking derived from feature importance analysis.
3) Identifying significant risk factors for HF patients through feature selection that combines mutual information and ML models.

The rest of paper is organized as follows. Section II outlines the proposed methodology. Section III and IV present the simulation results and discussion, followed by the concluding remarks in Section V.

II. METHOD

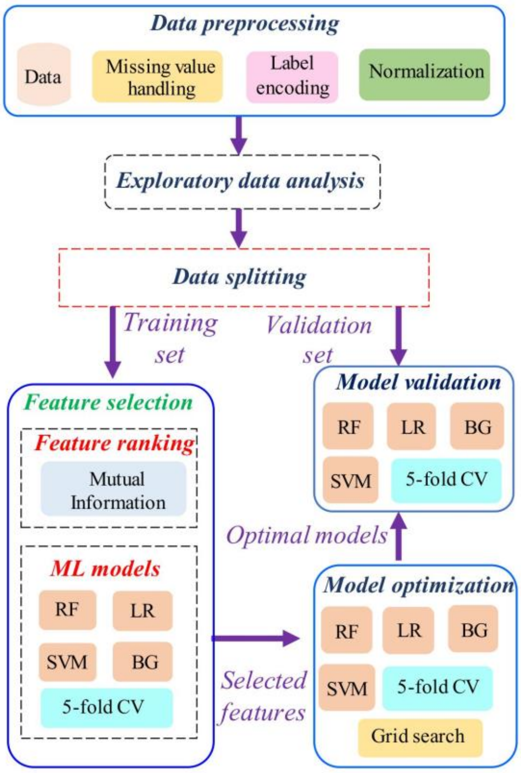

Figure 1: Methodology workflow diagram analysis for targeted research.

Three techniques including exploratory data analysis (EDA), mutual information (MI), and ML models, are applied in this research. EDA is used to explore the data distribution, identify patterns. MI is used to perform ranking of features based on their mutual dependency with the target variable, after which feature combinations are generated. Finally ML models are used to validate the predictive ability of various feature combinations on the training set, in order to identify the most effective feature sets for classification.

Our proposed method is illustrated in Figure 1, which consists of five main stages: data preprocessing, exploratory data analysis (EDA), feature selection, model optimization, and model validation. The process begins with data preprocessing, in which raw data undergoes cleaning procedures such as missing values handling, label encoding, and normalization to ensure consistency and improve data quality. This is followed by EDA, which aims to uncover hidden patterns and understand the distribution of variables associated with HF mortality. Next, the dataset is divided into 70% for training, 30% for testing, and 100% for validation. The third stage is feature selection, in which mutual information (MI) is used to rank features based on their relevance to the target variable. Four ML models, including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Bagging (BG) are used to identify significant features. Finally, in the model optimization and validation phase, hyperparameters are tuned using grid search and the optimized models are validated on the holdout validation set to identify the most accurate and generalizable models.

A. Data

In this study, we utilize a dataset of HF survival events for analysis and model prediction. The dataset, obtained from [17], consists of 299 patients, of whom 96 died and 203 survived. This dataset was collected at the Faisalabad Institute of Cardiology and at the University Allied Hospital in Faisalabad (Punjab, Pakistan).

Table 1: Data description.

Feature	Explanation
Age	Age of the participant
Anaemia	Decrease of red blood cells or hemoglobin
High Blood Pressure (HBP)	The patient has high blood pressure or not
Creatinine Phosphokinase (CPK)	The level of CPK enzyme in the blood in mcg/L
Diabetes	The patient has diabetes or not
Ejection Fraction (EF)	Percentage of blood leaving the heart at each contraction.
Sex	Sex of patient
Platelets	Platelets in the blood in kiloplatelets/mL
Serum Creatinine (SCR)	The level of creatinine in the blood in mg/dL
Serum Sodium (SS)	The level of sodium in the blood in mEq/L
Smoking	whether the patient has a smoking habit or not
Time	Follow-up period
Death Event	whether the patient survived or died during follow up period

B. Data preprocessing

Data preprocessing is an important task for ML analysis as it prepares a dataset for better analysis results. First, missing values in the dataset are handled based on their data types. Next, label encoding technique is applied to categorical features. Finally, the min-max scaling method is used to normalize the feature values so that they fall within the range of 0 to 1.

C. Exploratory data analysis

Generally, a dataset contains various types of important information that are not easily found. By analyzing the dataset, this information can be extracted. In this study, EDA is used on the dataset to discover hidden patterns and trends. EDA is a method for analyzing datasets and describing their main properties, mainly using graphical methods. Before the modeling process, EDA is used to examine the data.

D. Feature selection

First, mutual information (MI) is used for ranking the 12 features. Then, 12 feature combinations are generated: combination 1 contains the single highest-ranked feature, combination 2 contains the two highest-ranked features, and so on until combination 12, which includes all 12 features. Finally, ML models are applied to evaluate these 12 combinations and identify the most significant feature sets based on accuracy (Acc) metric. Therefore, 4 optimal feature combination sets are selected, each corresponding to one of the 4 ML models. The role of MI and ML models here is that MI provides a shared, interpretable foundation for ranking features, while the ML models determine the final feature subset from empirical outcomes. This maintains fairness and model-specific optimization in selection.

MI [19] emerges as a pivotal metric, quantifying the dependency between feature variables and the target variable. It calculates the extent to which the knowledge of a feature variable decreases uncertainty about the target variable. In this research, MI is employed to evaluate and rank features according to their individual contributions to the accurate prediction of the target event. A more substantial MI score indicates a more pronounced dependency between a feature and the target outcome. The formula for calculating the MI between a feature variable X and the target variable Y is written by

$$MI(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (1)$$

Where $p(x, y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively. Features exhibiting high mutual information with the target variable are deemed influential, as they indicate strong statistical dependence.

E. Model optimization

To enhance performance in predicting HF events, 4 ML models (LR, RF, SVM, and BG) are optimized using grid search combined with 5-fold CV to find optimal configuration for each model. The ML models are described as follows:

Logistic regression (LR) [20]: A linear classification algorithm that models the probability of a binary outcome using a logistic function. It is widely used for baseline comparisons in medical prediction tasks.

Random forest (RF) [21]: An ensemble learning method based on decision trees, where multiple trees are trained on bootstrapped samples and their predictions are aggregated to improve accuracy and reduce overfitting.

Support Vector Machine (SVM) [22]: A powerful classifier that constructs an optimal hyperplane in a high-dimensional space to separate classes, effective in handling non-linear relationships through the use of kernel functions

Bagging (BG) [23]: An ensemble technique that combines predictions from multiple base estimators trained on different subsets of the training data, aiming to reduce variance and improve generalization.

F. Model validation

The optimal models are evaluated using 4 optimal feature combination sets identified the previous step and the full feature set. The evaluation is conducted on the validation set using 5-fold CV. In this procedure, the validation set is randomly divided into 5 folds, where each fold is used once as the test set while the remaining folds are used for training. This process is repeated until every fold has served as the test set, enabling the computation of the mean and standard deviation of model performance metrics. Among the tested feature combinations, the one that yields the highest prediction accuracy is proposed as the final configuration for mortality recognition in HF patients.

III. SIMULATION RESULTS

A. Measurement

The performance of ML models are evaluated by different measured parameters namely accuracy (Acc), precision (Pre), Recall, and F1-score. Acc measures the percentage of patients who are correctly diagnosed. Pre and Recall show the percentage of patients predicted as mortality who are truly dead out of the total predicted as HF and the fraction of patients predicted as mortality out of the total of actual mortality, respectively. F1-score is calculated by the harmonic mean of Pre and Recall, which provides an overall view of the performance of the models in predicting HF events.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Pre = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}score = 2\frac{Pre \times Recall}{Pre + Recall} \tag{5}$$

B. Explore data analysis

Figure 2 presents an exploratory data analysis of the dataset using bar plots for categorical variables and Kernel Density Estimation (KDE) plots for numerical variables. In general, the number of male patients is higher than that of female patients, and the majority of patients in the dataset

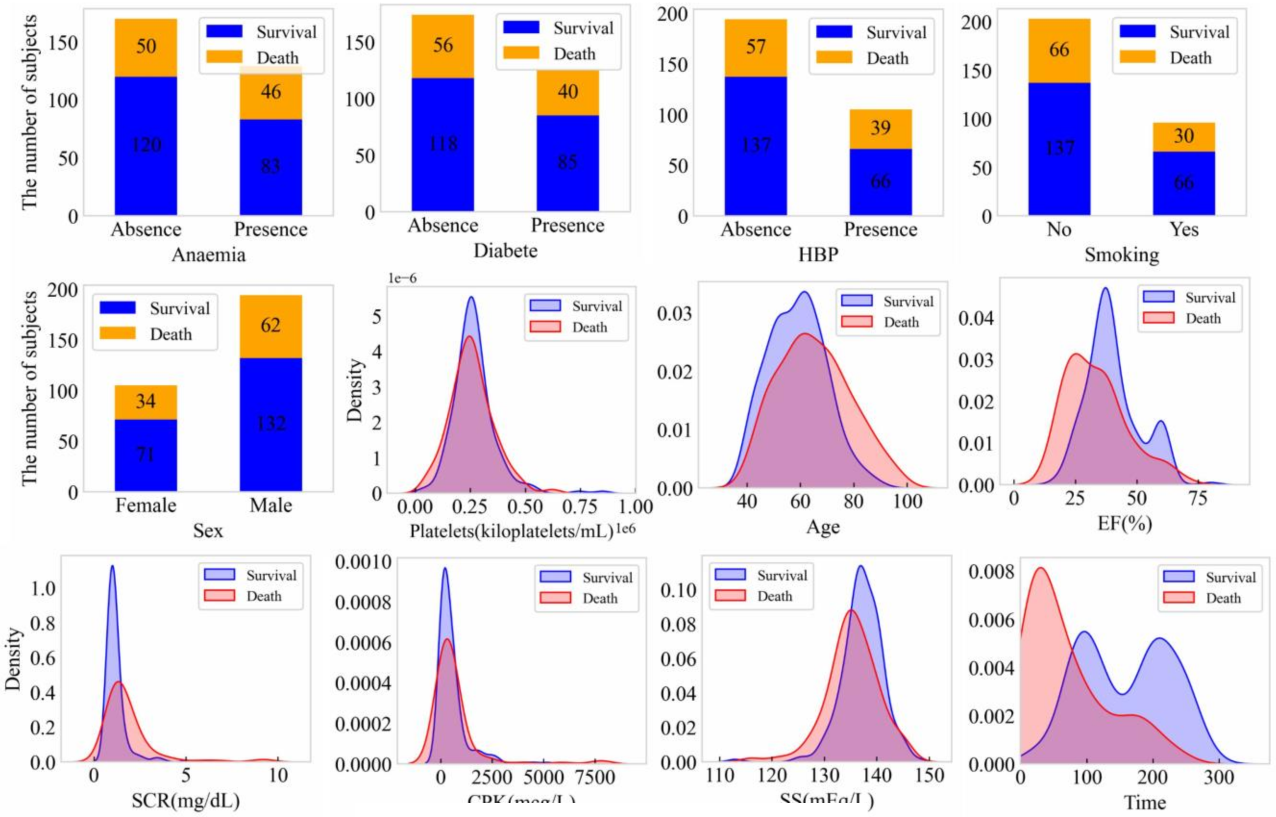

Fig 2: The distribution of the dataset

are non-smokers. The bar plot findings show higher mortality among patients with pre-existing conditions such as Anaemia, Diabetes, and HBP.

The KDE plot for each of the numeric attributes gives a clear overview of the ranges of values for attributes which are risk factors for HF patients. The results show that HF is very risky for people aged 50 years or more. EF below 30% is considered risky for patients. SCR levels above 2.0 mg/dL, CPK levels greater than 1500 mcg/L, and SS levels below 130 mEq/L are also associated with higher mortality. Platelet counts between 180,000 and 360,000 kiloplatelets/mL appear to be safe for HF patients. The Time variable confirms that longer follow-up durations are typically observed in surviving patients, whereas premature deaths are more common among those with shorter follow-up periods.

Overall, variables such as age, EF, SCR, CPK, SS, and Time are strong predictors of mortality risk in HF patients and hold significant potential for predictive modeling in clinical decision support systems.

C. Feature selection

Table 2 presents the feature importance scores for all features based on the MI method. The values indicate the contribution of each feature to predicting HF survival. MI reflects the amount of information a feature shares with the target variable, making it a key criterion for evaluating the relevance of individual features in the prediction task.

Table 2: Feature importance based feature ranking using MI method

	Feature	Importance value
1	Time	0.2254
2	SCR	0.0852
3	EF	0.0659
4	Age	0.0609
5	CPK	0.0251
6	Diabetes	0.0221
7	Sex	0.0125
8	Platelets	0.0026
9	Smoking	10^{-4}
10	SS	10^{-4}
11	HBP	10^{-4}
12	Anaemia	10^{-4}

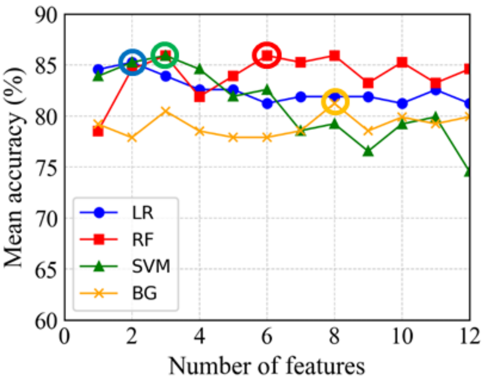

Fig 3: Evaluation of feature combinations based on average accuracy of 4 ML models.

The results presented in Figure 3 illustrate the performance of four ML models evaluated on 12 predefined feature combinations generated using the MI method. For each model, these combinations are used as inputs during training, and model performance is evaluated based on accuracy metrics. The optimal feature set (referred to as FS) for each ML model is determined as the set that provides the highest classification accuracy on the training data. Specifically, the LR model identifies FS1 (Time and SCR) as the optimal combination. The RF model selects FS2, which includes six important features: Time, SCR, EF, Age, CPK, and Diabetes. Similarly, the SVM model identifies FS3, which includes Time, SCR, and EF, as significant features. Finally, the BG model identifies FS4 as the best performing feature set, which includes Time, SCR, EF, Age, CPK, Diabetes, Sex, and Platelets.

D. Model optimization

There are 20 optimal models corresponding to LR, RF, SVM, and BG, selected using grid search with 5-fold CV across five feature sets: FS1, FS2, FS3, FS4, and the full feature set. The results are presented in Table 3.

Table 3: Hyperparameters tuned for all applied models using grid search

Models	Input	Hyper tuned parameters
LR	FS1	C=1.0, penalty='l2'
	FS2	C=0.5, penalty='l2
	FS3	C=1.5, penalty='l1'
	FS4	C=1.0, penalty='l2
	All	C=0.5, penalty='l2'
RF	FS1	n_estimators=100, max_depth=5
	FS2	n_estimators=120, max_depth=10
	FS3	n_estimators=85, max_depth=8
	FS4	n_estimators=95, max_depth=6
	All	n_estimators=100, max_depth=10
SVM	FS1	C=1.0, kernel='rbf'
	FS2	C=1.5, kernel='linear'
	FS3	C=1.5, kernel='rbf'
	FS4	C=2.0, kernel=’rbf '
	All	C=2.0, kernel= 'linear'
BG	FS1	base_estimator=DecisionTreeClassifier(), n_estimators=10,
	FS2	base_estimator=DecisionTreeClassifier(), n_estimators=10,
	FS3	base_estimator=DecisionTreeClassifier(), n_estimators=15,
	FS4	base_estimator=DecisionTreeClassifier(), n_estimators=10,
	All	base_estimator=DecisionTreeClassifier(), n_estimators=20,

E. Model validation

Table 4 illustrates the validation performance of various optimal models using different input feature combinations. The highest accuracy of 84.9% is generated by RF model using FS2 features, including 6 significant features, namely Time, SCR, EF, Age, CPK, and Diabetes. Therefore, we propose RF with these 6 features for predicting HF mortality.

IV. DISCUSSION

To address the important task of predicting mortality in HF patients, this study presents a comprehensive investigation of risk factor identification and prediction models using clinical variables. The integration of EDA, MI-based feature selection, and ML techniques enables the creation of a data-driven framework to explore key factors of survival outcomes in HF patients.

Feature selection is an important step in identifying the most relevant risk factors associated with HF mortality. MI is employed to rank the importance of 12 features based on their relevance to survival outcomes shown in Table 2. Time shows the highest relevance in predicting HF mortality, with a significance value of 0.2254, indicating its dominant influence on the outcome variable. This is followed by SCR and EF, which ranked second and third, with significance values of 0.0852 and 0.0659, respectively. This features are also identified during the EDA, reinforcing the reliability of the MI-based feature assessment procedure. Other significant features included Age (0.0609), CPK (0.0251), and Diabetes (0.0221), all of which are known to be strongly associated with HF progression and outcomes. In contrast, Sex, Platelets, Smoking, SS, HBP, and Anemia exhibited relatively low MI scores, suggesting limited predictive performance.

Table 4: Performance comparisons of different ML models on validation set

Model	Feature	Acc (%)	Pre (%)	Recall (%)	F1-score (%)
LR	FS1	83.29 ±3.10	82.33 ±6.69	62.85 ±8.09	70.5 ±3.77
	FS2	82.96 ±3.80	77.33 ±5.70	67.62 ±4.48	71.94 ±3.28
	FS3	83.28 ±2.95	77.14 ±9.45	67.22 ±6.79	71.42 ±5.40
	FS4	82.95 ±3.35	76.66 ±9.74	67.98 ±8.41	71.36 ±5.25
	All	82.96 ±4.09	76.98 ±9.00	67.72 ±10.1	71.24 ±6.64
RF	FS1	83.62 ±3.01	78.32 ±7.89	70.66 ±10.81	73.17 ±3.87
	FS2	**84.9 ±2.55**	**81.71 ±9.71**	**71.02 ±9.12**	**74.94 ±3.16**
	FS3	82.63 ±4.50	75.43 ±10.8	70.29 ±7.35	71.95 ±5.24
	FS4	84.31 ±3.98	77.75 ±9.38	71.54 ±8.49	74.05 ±6.86
	All	83.98 ±6.17	79.21 ±12.1	70.79 ±11.4	73.83 ±8.53
SVM	FS1	84.29 ±3.39	89.69 ±6.94	58.89 ±10.40	70.0 ±7.54
	FS2	82.62 ±5.09	82.59 ±9.86	60.66 ±8.25	69.30 ±5.95
	FS3	83.96 ±3.04	80.73 ±8.50	66.61 ±9.60	72.19 ±4.51
	FS4	79.94 ±4.45	78.12 ±7.88	53.83 ±9.50	62.71 ±4.94
	All	76.60 ±8.40	80.24 ±12.2	43.34 ±15.8	52.97 ±14.1
BG	FS1	79.29 ±4.22	72.14 ±6.96	60.34 ±12.86	64.6 ±7.67
	FS2	81.59 ±1.96	74.47 ±9.53	63.38 ±9.67	68.03 ±7.80
	FS3	78.94 ±3.68	70.58 ±9.31	60.40 ±9.95	64.04 ±5.89
	FS4	78.28 ±5.43	68.87 ±11.5	59.01 ±8.80	62.94 ±8.02
	All	75.93 ±4.73	64.03 ±6.25	57.86 ±4.80	60.46 ±3.29

To identify the risk factors for HF mortality prediction, we use 4 ML models including LR, RF, SVM, and BG. Each model selects its own optimal subset of features, resulting in 4 distinct feature sets (FS1 to FS4), as illustrated in Figure 3. Specifically, FS1 includes Time and SCR; FS2 includes Time, SCR, EF, Age, CPK, and Diabetes; FS3 includes Time, SCR, and EF; and FS4, extends FS2 by including Sex and Platelets. This multi-model feature selection process ensures that the chosen subsets reflect both statistical relevance and model-specific learning characteristics. Next, these four feature sets serve as inputs to each of the four ML models to evaluate their predictive performance. Using 5-fold cross-validation, the study compares all combinations of models and feature sets to identify the most effective configuration. As shown in Table 4, RF model with FS2 (Time, SCR, EF, Age, CPK, and Diabetes) as input achieves the highest performance, with an accuracy of 84.9%, precision of 81.71%, recall of 71.02%, and an F1-score of 74.94%. In addition, we evaluate RF model with 6 selected features (Time, SCR, EF, Age, CPK, and Diabetes) on an independent testing set to assess it generalization capability. The results shown in Figure 4 indicate that the model maintains stable and reliable performance on unseen data, thereby confirming its robustness and applicability in real-world clinical scenarios. The consistency between validation and testing performance underscores the effectiveness of the selected feature set and the potential of model as a decision-support tool for HF mortality prediction.

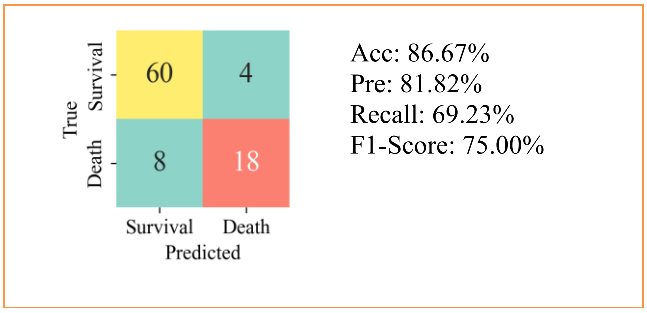

Fig 4: Confusion matrix on testing set

Each of the selected features has important physiological implications for patients with HF disease. In addition to their role in predicting mortality, these factors can assist clinicians in assessing patient status and developing individualized treatment strategies. The six risk factors associated with HF are as follows: Time, which stands for the length of follow-up, gives information about patient stability and the course of the disease and represents the temporal progression of HF. Early mortality is frequently associated with shorter follow-up periods, highlighting the importance of careful monitoring during the early phases of HF treatment. Thus, Time serves as both a clinical indicator of patient outcome and a proxy for disease severity. As people age, wear and tear on the cardiovascular system builds up over time, and this is a significant factor. The necessity of age-specific risk assessment techniques in clinical practice is reaffirmed by this study. Future cardiovascular medicine and economic needs will be challenged by the aging process, which is regarded as a major non-modifiable risk factor for heart health [24]. Elevated SCR levels, indicative of reduced renal function, have been identified as a critical predictor of HF. This relationship highlights the interconnectedness of renal and cardiovascular health, suggesting that monitoring SCR levels could be integral to HF risk management [25]. The inverse relationship between EF and HF risk reflects the fundamental role of cardiac function efficiency in maintaining cardiovascular health. Low EF values signal reduced heart pumping efficiency, underscoring the importance of cardiac function assessments in the early detection of HF risk [26]. Increased levels of the CPK, which is released during muscle breakdown, can be a sign of systemic muscle damage or myocardial injury, both of which are associated with a worse prognosis for HF [27]. A known comorbidity that significantly exacerbates the progression of HF is diabetes. This disease accelerates the progression of HF by contributing to endothelial dysfunction, myocardial remodeling, and vascular inflammation [28].

Finally, we compare our results with previous studies to demonstrate the effectiveness of the proposed method, which is shown in Table 5. Although [11] reports higher precision, recall, and F1-score than ours, this improvement is due to their use of the SMOTE technique to balance the dataset. In contrast, our method use imbalanced data. SMOTE can improve performance of model, however it may also introduce synthetic bias. Therefore, our approach emphasizes performance under real-world conditions without artificial data augmentation. Overall, our model achieves competitive and stable performance on both validation and test sets, confirming its reliability and practical applicability.

Table 5: Comparison with existing works

Ref	Validation method	Acc	Pre	Recall	F1-score
[11]	85%-15%	83.33	86.36	90.48	88.37
[10]	5-fold CV	78	66	64	65
Our	70%-30%	86.67	81.82	69.23	75.00
	5-fold CV	84.90	81.71	71.02	74.94

However, the limitations of this study are the small dataset. This may affect the generalizability of the model. These issues will be addressed in future work by collecting a larger and more diverse dataset, including other types of data such as image or time-series data. In addition, more advanced AI techniques will be explored to better handle complex data and further improve the performance of model and applicability in real-world clinical settings.

V. CONCLUSIONS

Identifying risk factors and accurately diagnosing HF events is critically important for enabling clinicians to take timely and appropriate interventions to prevent mortality. Therefore, in this paper, we proposed a method that combines exploratory data analysis, feature selection, and ML models to identify key risk factors for HF. Four ML models were optimized using grid search and 5-fold CV to enhance their predictive performance. By evaluating feature importance using MI method, followed by ML models, key risk factors associated with HF mortality were identified. The results illustrate that RF model with 6 six risk factors including Time, SCR, EF, Age, CPK, and Diabetes achieves the highest accuracy of 84.9%, precision of 81.71%, recall 71.02%, and F1-score 74.94% on validation set. These findings indicate that the models is capable of effectively predicting survival outcomes in HF patients. The proposed RF model has potential utility in clinical settings for assisting clinicians in screening and risk stratification of HF patients.

REFERENCES

[1] Savarese, Gianluigi, et al. "Global burden of heart failure: a comprehensive and updated review of epidemiology." *Cardiovascular research* 118.17 (2022): 3272-3287.

[2] MEMBERS, WRITING COMMITTEE, et al. "Heart failure epidemiology and outcomes statistics: a report of the Heart Failure Society of America." *Journal of cardiac failure* 29.10 (2023): 1412.

[3] Ahmad, A. A., and H. Polat. "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. Diagnostics, 13 (14), 2392." 2023.

[4] Ali, Md Mamun, et al. "A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients." *Healthcare Analytics* 3 (2023): 100182.

[5] Li, Xinmu, et al. "Clinical applications of machine learning in heart failure." *State of the Art in Neural Networks and Their Applications* (2023): 217-233.

[6] Banerjee, Amitava, et al. "Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study." *The Lancet Digital Health* 5.6 (2023): e370-e379.

[7] Srinivasan, S., S. Gunasekaran, and S. K. Mathivanan. "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database." *Sci Rep* 13, 13588 (2023).

[8] Newaz, Asif, Nadim Ahmed, and Farhan Shahriyar Haq. "Survival prediction of heart failure patients using machine learning techniques." *Informatics in Medicine Unlocked* 26 (2021): 100772.

[9] Kokori, Emmanuel, et al. "Machine learning in predicting heart failure survival: a review of current models and future prospects." *Heart Failure Reviews* (2024): 1-12.

[10] Moreno-Sánchez, Pedro A. "Improvement of a prediction model for heart failure survival through explainable artificial intelligence." *Frontiers in cardiovascular medicine* 10 (2023): 1219586.

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

[12] Newaz, Asif, Nadim Ahmed, and Farhan Shahriyar Haq. "Survival prediction of heart failure patients using machine learning techniques." *Informatics in Medicine Unlocked* 26 (2021): 100772.

[13] Sutradhar, Ananda, et al. "BOO-ST and CBCEC: two novel hybrid machine learning methods aim to reduce the mortality of heart failure patients." *Scientific Reports* 13.1 (2023): 22874.

[14] Mpanya, Dineo, et al. "Predicting in-hospital all-cause mortality in heart failure using machine learning." *Frontiers in cardiovascular medicine* 9 (2023): 1032524.

[15] Adekkanattu, Prakash, et al. "Prediction of left ventricular ejection fraction changes in heart failure patients using machine learning and electronic health records: a multi-site study." *Scientific reports* 13.1 (2023): 294.

[16] Sabahi, H., M. Vali, and D. Shafie. "In-hospital mortality prediction model of heart failure patients using imbalanced registry data: A machine learning approach." *Scientia Iranica* (2023).

[17] Ahmad, Tanvir, et al. "Survival analysis of heart failure patients: A case study." *PloS one* 12.7 (2017): e0181001.

[18] Veyrat-Charvillon, Nicolas, and François-Xavier Standaert. "Mutual information analysis: how, when and why?." *International workshop on cryptographic hardware and embedded systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[19] Veyrat-Charvillon, Nicolas, and François-Xavier Standaert. "Mutual information analysis: how, when and why?." *International workshop on cryptographic hardware and embedded systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[20] LaValley, Michael P. "Logistic regression." *Circulation* 117.18 (2008): 2395-2399.

[21] Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32.

[22] Pisner, Derek A., and David M. Schnyer. "Support vector machine." *Machine learning*. Academic Press, 2020. 101-121.

[23] Breiman, Leo. "Bagging predictors." *Machine learning* 24 (1996): 123-140.

[24] Grassow, Leonhard, et al. "Sex-specific structural and functional cardiac remodeling during healthy aging assessed by cardiovascular magnetic resonance." *Clinical Research in Cardiology* (2024): 1-12.

[25] Chávez-Íñiguez, Jonathan S., et al. "How to interpret serum creatinine increases during decongestion." *Frontiers in Cardiovascular Medicine* 9 (2023): 1098553.

[26] Hajouli, Said, and Dipesh Ludhwani. "Heart failure and ejection fraction." (2020).

[27] Crowley, Leonard V. "Creatine phosphokinase activity in myocardial infarction, heart failure, and following various diagnostic and therapeutic procedures." *Clinical Chemistry* 14.12 (1968): 1185-1196.

[28] Bell, David SH. "Heart failure: the frequent, forgotten, and often fatal complication of diabetes." *Diabetes care* 26.8 (2003): 2433-2441.

DỰ ĐOÁN SỰ SỐNG SÓT CỦA BỆNH NHÂN SUY TIM BẰNG HỌC MÁY

Tóm tắt: Suy tim (HF) là một hội chứng lâm sàng phổ biến và phức tạp với tỷ lệ tử vong cao, khiến việc dự đoán chính xác khả năng sống sót trở nên vô cùng quan trọng đối với việc quản lý bệnh nhân. Để giải quyết thách thức này, nghiên cứu của chúng tôi giới thiệu một phương pháp tích hợp phân tích dữ liệu khám phá, lựa chọn tính năng và các mô hình học máy (ML) để xác định chính xác các yếu tố rủi ro quan trọng đối với các biến cố HF. Tối ưu hóa các mô hình ML thông qua tìm kiếm lưới và xác thực chéo 5 lần (CV) được sử dụng để tăng cường hiệu suất của chúng. Phương pháp tiếp cận của chúng tôi xác định các kết hợp tính năng bao gồm các yếu tố rủi ro quan trọng bằng cách sử dụng bốn mô hình ML với CV 5 lần. Kết quả làm nổi bật các yếu tố quan trọng tác động đến các biến cố HF bao gồm thời gian, creatinin huyết thanh, phân suất tống máu, tuổi, creatinine phosphokinase và bệnh tiểu đường. Trong số bốn mô hình, mô hình rừng ngẫu nhiên nổi bật vì tính mạnh mẽ của nó trong việc dự đoán tỷ lệ tử vong do HF. Điều này được chứng minh bằng hiệu suất của mô hình thông qua dữ liệu xác thực và thử nghiệm. Cụ thể, hiệu suất trên bộ xác thực đạt độ chính xác Acc là 84,9%, độ chính xác Pre là 81,71%, khả năng thu hồi là 71,02% và điểm F1 là 74,94%. Trên bộ thử nghiệm, hiệu suất của mô hình đạt độ chính xác Acc là 86,67%, độ chính xác Pre là 81,82%, độ thu hồi 69,23% và điểm F1 là 75%. Kết quả này khẳng định hiệu suất của phương pháp chúng tôi đề xuất để dự đoán các biến cố suy tim với độ chính xác và độ tin cậy cao.

Từ khóa: Suy tim, Học máy, Lựa chọn tính năng, Dự đoán khả năng sống sót.

Minh Tuan Nguyen received the B.S. degree from the Post & Telecommunications Institute of Technology, Hanoi, Vietnam, in 2004, the M.S. degree from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2008, both in electronics and telecommunications engineering, and the Ph.D. degree at the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2018. He is with Posts and Telecommunications Institute of Technology. His research interests include network security, internet of things, biomedical signal processing, gene analysis, sentiment analysis, brain computer interface, machine learning, deep learning, optimization, and biomedical application design.

Le Nhat Thang received the B.Eng degree in Radio-Electronics and Communication from Hanoi University of Science and Technology (HUST), Vietnam, in 1995, the M.Eng degree in Telecommunications from Asian Institute of Technology (AIT), Bangkok, Thailand, in 2000 and Ph.D. degree in Information and Communication Technology (ICT) from the Department of Computer Science and Telecommunications (DIT), University of Trento, Italy in 2006. He is currently an Assoc. Professor and the Dean of Postgraduate Studies Faculty, Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam. His research interests now are performance analysis, modeling and simulations, wireless communications systems, physical layer security, queueing theory and applications, machine learning, deep learning, optimization, biomedical applications.