BMC Cardiovascular Disorders

RESEARCH

Open Access

Interpretable artificial intelligence model for predicting heart failure severity after acute myocardial infarction

Chenglong Guo[1†], Binyu Gao[2,3†], Xuexue Han[4], Tianxing Zhang[4], Tianqi Tao[5*], Jinggang Xia[4*] and Honglei Liu[3,6*]

Abstract

Background Heart failure (HF) after acute myocardial infarction (AMI) is a leading cause of mortality and morbidity worldwide. Accurate prediction and early identification of HF severity are crucial for initiating preventive measures and optimizing treatment strategies. This study aimed to develop an interpretable artificial intelligence (AI) model for HF severity prediction using multidimensional clinical data.

Methods This study included data from 1574 AMI patients, including medical history, clinical features, physiological parameters, laboratory test, coronary angiography and echocardiography results. Both deep learning (TabNet, Multi-Layer Perceptron) and machine learning (Random Forest, XGboost) models were employed in constructing model. Additionally, the Shapley Additive Explanation (SHAP) method was used to elucidate clinical factors importance and enhance model interpretability. A web platform (https://prediction-killip-gby.streamlit.app/) was also developed to facilitate clinical application.

Results Among the models, TabNet demonstrated the best performance, achieving an AUROC of 0.827 for KILLIP four-class classification and 0.831 for KILLIP binary classification. Key clinical factors such as GRACE score, NT-pro BNP, and TIMI score were highly correlated with KILLIP classification, aligning with established clinical knowledge.

Conclusions By leveraging easily accessible multidimensional data, this model enables accurate early prediction and personalized diagnosis of HF risk and severity following AMI. It supports early clinical intervention and improves patient outcomes, offering significant clinical application value.

Clinical trial number Not applicable.

Keywords Artificial intelligence, Deep learning, Heart failure, Acute myocardial infarction

†Chenglong Guo and Binyu Gao contributed equally to this work.

*Correspondence:
Tianqi Tao
ttqtxt@163.com
Jinggang Xia
xiajinggang@sina.cn
Honglei Liu
liuhonglei@ccmu.edu.cn
[1]Pulmonary Vascular Disease Center, Beijing Anzhen Hospital, Capital Medical University, Beijing, China
[2]Biological Science & Medical Engineering, Southeast University, Nanjing 518000, China
[3]School of Biomedical Engineering, Capital Medical University, Beijing 100069, China
[4]Department of Cardiology, Xuanwu Hospital, Capital Medical University, Beijing 100053, China
[5]Department of Geriatrics, The Second Medical Center, National Clinical Research Center for Geriatric Diseases, Chinese PLA General Hospital, Beijing 100853, China
[6]Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing 100069, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Acute myocardial infarction (AMI), commonly known as a heart attack, remains a leading cause of mortality and morbidity worldwide [1]. AMI often leads to the development of heart failure (HF), a debilitating condition with high morbidity and mortality rates [2]. Early identification of patients at risk of developing HF after AMI is crucial for initiating preventive measures and optimizing treatment strategies. The Killip classification, a widely used bedside tool, assesses clinical signs of HF after AMI to stratify patients into different risk categories [3]. However, this classification relies on subjective assessment and may not fully capture the complex interplay of factors contributing to HF development. Leveraging multidimensional data for HF severity prediction could enhance early identification of high-risk patients and guide timely clinical interventions.

In recent years, artificial intelligence (AI) methods, particularly deep learning models, have emerged as powerful tools for analyzing large and complex clinical datasets. These models can identify subtle patterns and generate highly accurate predictions, often surpassing traditional methods in precision [4–6]. In the context of AMI, deep learning can leverage diverse patient data (e.g., clinical features, biomarkers, imaging data) to enhance outcome prediction accuracy [6–9]. By capturing intricate relationships within the data, deep learning approaches, such as artificial neural networks, offer insights that conventional statistical methods might overlook [10]. Training these models on large AMI cohorts allows them to uncover hidden patterns and interactions associated with HF risk [11]. Moreover, deep learning's ability to handle high-dimensional data allows for the incorporation of a wide range of variables, potentially leading to more comprehensive and accurate risk prediction models.

While recent studies have shown promising results in using machine learning to predict HF after AMI [7, 12], few studies have focused on the specific task of predicting HF combined in conjunction with Killip classifications. Such a focus could provide valuable insights into the early detection of HF after AMI and support personalized treatment strategies. Additionally, existing HF prediction models often lack interpretability, making it challenging to explain predictions and analyze feature importance. The recent introduction of TabNet [13], a deep learning algorithm specifically designed for tabular data, offers new possibilities for improving the performance of clinical tabular data processing.

In this study, we utilized easily accessible multidimensional data obtained during hospitalization-including medical history, clinical features, physiological parameters, and results from laboratory test, coronary angiography combined with echocardiography. We leveraged both deep learning (TabNet, Multi-Layer Perceptron [14–16](MLP)) and machine learning models (Random Forest (RF) [17], XGboost [18]) to predict HF severity in patient after AMI, enabling accurate and personalized identification of HF severity. Additionally, we employed Shapley Additive Explanation [19] (SHAP) to elucidate risk factor importance and improve model interpretability. We also developed a web platform to facilitate clinical application.

Method

Dataset

A retrospective study was conducted on 2993 patients diagnosed with type I AMI at Xuanwu Hospital, Capital Medical University, between January 2017 and December 2022. It was authorized by the Ethics Committee of Xuanwu Hospital, Capital Medical University with the approval document number (2022–129) and was processed according to the principles of the Declaration of Helsinki. All enrolled patients signed informed consent forms.

We selected several factors that could potentially influence the development of heart failure in AMI patients. These factors included demographic characteristics such as age, sex, and body mass index (BMI); clinical scores like the GRACE and TIMI risk scores; and medical history, including hypertension, atrial fibrillation (AF), diabetes, hyperlipidemia, cerebrovascular disease (CVD), peptic ulcer (PU), previous myocardial infarction, stent implantation, and coronary artery bypass grafting (CABG). Smoking status was also considered, including whether patients had quit smoking. In addition, we analyzed post-admission blood test results, which included white blood cell count, neutrophils, lymphocytes, monocytes, hemoglobin, platelet count, blood glucose, alanine aminotransferase (ALT), aspartate aminotransferase (AST), creatinine clearance rate (CCR), total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), uric acid (UA), hemoglobin A1c (HbA1c), high-sensitivity C-reactive protein (hs-CRP), interleukin-6 (IL-6), N-terminal pro-brain natriuretic peptide (NT-pro BNP), and peak troponin I (TNI). Echocardiographic indicators such as left ventricular ejection fraction (LVEF), left atrial diameter, and left ventricular end-diastolic diameter (LVEDD) were also assessed. Additionally, the length of hospitalization was recorded. Patients missing critical features such as NT-proBNP, LVEF, or GRACE/TIMI scores were excluded from the database, resulting in a final cohort of 1,574 patients.

Based on the Killip classification for HF in AMI, patients were categorized into four groups. This classification system is widely used in clinical cardiology and provides a rapid bedside assessment of HF severity. The

Table 1 The Killip class definitions in AMI patients

Killip Class	Clinical Description
Class 1	No clinical signs of heart failure
Class 2	Signs of mild to moderate HF (e.g., S3 gallop, rales < 50% lung field)
Class 3	Acute pulmonary edema (rales > 50% lung field, severe respiratory distress)
Class 4	Cardiogenic shock

Killip class definitions used in this study were summarized in Table 1.

Model construction and comparison

Both machine learning and deep learning models were applied to the AMI data. Considering the property of medical tabular data, the machine learning models included RF and XGBoost, while the deep learning models comprised a MLP and TabNet. The RF and MLP models were implemented using the Python scikit-learn package, XGBoost using the XGBoost package, and TabNet using PyTorch library. The flow chart of the study design was shown in Fig. 1.

RF and XGBoost are commonly used tree-based machine learning models. MLP is a type of feedforward artificial neural network consisting of multiple layers of nodes, where each node is fully connected to the next

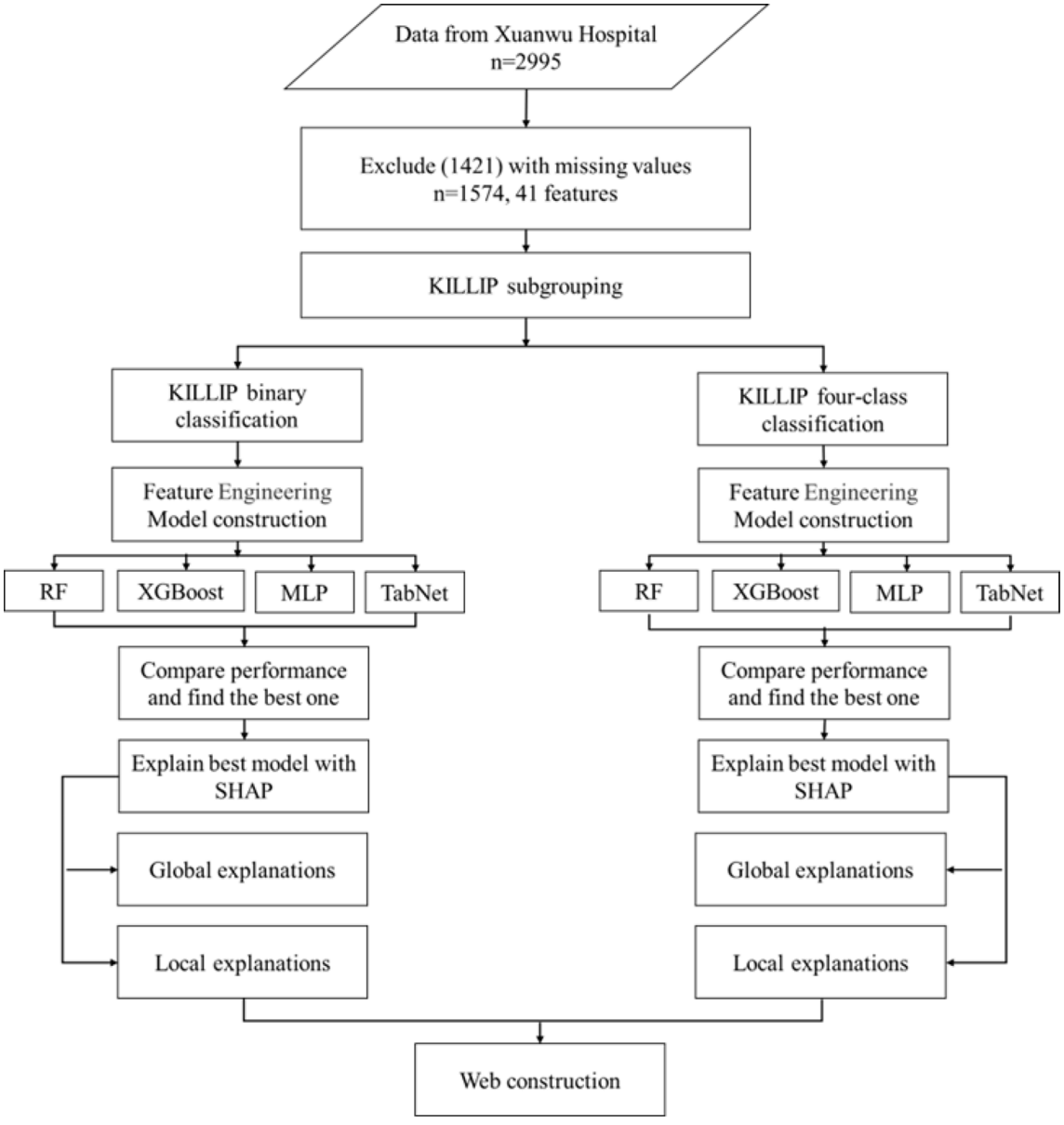

Fig. 1 Flow chart of the study design

layer, allowing for non-linear modeling of complex relationships. The Attentive Interpretable Tabular Learning (TabNet) model is a multi-stage deep learning model introduced by Google Cloud AI and applies a sequential instance-wise attention mechanism allowing it to inherently select the most salient set of radiomic features at different decision steps within its architecture. TabNet specifically designed for tabular data, combining interpretability with state-of-the-art performance through sequential attention mechanisms that decide which features to utilize at each decision step.

A total of 41 features were utilized to develop the prediction models. To build a standardized feature space across all models, continuous variables were normalized using a standard scaler such that they have a mean of 0 and a variance of 1. Given the imbalance in KILLIP classes, we employed a stratified partitioning method to split the dataset into two subsets: a training dataset (80%) and a test dataset (20%). The distribution of KILLIP classes in each subset was consistent with that in the original dataset. During training, to enhance the model's ability to identify minority class samples, we applied the synthetic minority over-sampling technique (SMOTE) to interpolate the minority KILLIP classes (KILLIP 2, KILLIP 3, KILLIP 4).

For each model, we employed a grid search approach for parameter tuning until the optimal parameter combination was identified. At each tuning step, fivefold cross-validation was performed using stratified shuffles of the training dataset to estimate the parameters of each model and evaluate their predictive performance. To consider class imbalance and classification accuracy, with specific justification, the evaluation metrics included the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), precision, and F1 scores. For each complication, the model with the highest mean AUROC was selected as the best-performing model.

Model interpretation

Model predictions were interpreted using SHAP, a model-agnostic explanation technique. SHAP values for each feature were calculated to represent the contribution of each feature to the predicted risk of a complication. The SHAP method provided both global and local explanations. Global explanations offered consistent and accurate attribution values for each feature, demonstrating the relationships between input features and KILLIP classification. Local explanations provided insights into specific predictions for individual cases by inputting the corresponding data.

Results

Patient characteristics

A total of 1574 patients were identified during the study period. Among them, 1005 patients (63.8%) were classified as KILLIP 1, 468 patients (29.7%) as KILLIP 2, 72 patients (4.6%) as KILLIP 3 and 29 patients (1.8%) as KILLIP 4. The demographic and clinical characteristics across KILLIP classes are summarized in Table 2. The results indicated that compared to patients in KILLIP class 1, those in KILLIP classes 2–4 were more likely to be female, older, and have a history of hypertension, diabetes, atrial fibrillation (AF), myocardial infarction, and stent placement. Additionally, these patients exhibited higher GRACE and TIMI risk scores, elevated white blood cell, neutrophil, and monocyte counts, as well as increased levels of HbA1c, hs-CRP, IL-6, ALT, AST, UA, NT-pro BNP, left atrial diameter, and LVEDD. Conversely, patients in KILLIP classes 2–4 had lower CCR, LVEF, TC, TG, and hemoglobin levels. These findings demonstrate that female gender, advanced age, elevated inflammation markers, pre-existing cardiac conditions, diabetes, and renal dysfunction are significantly associated with an increased risk of HF following AMI. Moreover, relatively low levels of TC, TG, and hemoglobin were also found to contribute to the increased risk of HF after AMI.

Model performance

To gain a more comprehensive understanding and improve the prediction of HF severity following AMI, we employed machine learning models (RF and XGBoost) and deep learning models (TabNet and MLP) for two tasks: four-class classification (KILLIP 1, 2, 3, 4) and binary classification (KILLIP 1 vs. KILLIP 2, 3, 4). The discriminative performance results are presented in Tables 3 and 4, respectively, with ROC curves illustrated in Fig. 1. Performance metrics of different models for each KILLIP class in four-class classification and binary classification were presented in Appendix Tables 2, 3, 4 and 5 and Appendix Tables 6, 7, 8 and 9, respectively. Appendix Table 10 shown the final hyperparameters of all the models.

As shown in Table 3; Fig. 2, among the models considered, the TabNet model achieved the highest predictive performance for KILLIP classification (four-class classification) with an AUROC of 0.827, followed by the MLP model and the RF model. The XGBoost model demonstrated the lowest performance. Overall, the deep learning models (MLP, TabNet) outperformed the machine learning models (RF, XGBoost).

Similarly, in the binary classification task (Table 4), the same trend was observed, with the TabNet model again delivering the best performance with an AUROC of 0.831.

Table 2 Comparison of demographic, clinical characteristics, and outcomes across KILLIP classes in the dataset

Variable	Killip classification				*P* value
Case number	1 (*n* = 1005)	2 (*n* = 468)	3 (*n* = 72)	4 (*n* = 29)	
Sex, Male, n (%)	816(81.2)	356(76.1)	50(69.4)	22(75.9)	0.023
Age(year)	60.5 ± 12.3	68.1 ± 11.5	70.4 ± 10.4	65.4 ± 11.2	< 0.001
BMI, kg/m^2	25.7 ± 3.6	25.5 ± 3.6	25.1 ± 3.5	24.5 ± 2.9	0.212
History of hypertension, n (%)	575(57.2)	309(66)	51(70.8)	15(51.7)	0.002
History of hyperlipidemia, n (%)	183(18.2)	63(13.5)	9(12.5)	2(6.9)	0.044
History of diabetes, n (%)	339(33.7)	219(46.8)	36(50)	14(48.3)	< 0.001
History of AF, n (%)	28(2.8)	23(4.9)	36(50)	14(48.3)	< 0.001
Old myocardial infarction, n (%)	101(10)	92(19.7)	3(4.2)	5(17.2)	< 0.001
History of stent implantation, n (%)	105(10.4)	90(19.2)	13(18.1)	4(13.8)	< 0.001
History of CABG, n (%)	9(0.9)	11(2.4)	1(1.4)	1(3.4)	0.121
History of smoking, n (%)	627(62.6)	256(54.7)	36(50)	15(51.7)	0.008
No quitting smoking, n (%)	511(50.8)	189(40.4)	23(31.9)	11(37.9)	0.034
History of CVD, n (%)	101(2.1)	1(5.9)	1(8.3)	2(22.2)	0.123
History of PU, n (%)	29(2.9)	11(2.6)	4(5.6)	0(0)	0.164
GRACE risk score	138.0 ± 29.0	170.1 ± 30.1	203.8 ± 35.8	231.5 ± 33.0	< 0.001
TIMI risk score	3.4 ± 1.7	5.3 ± 2.1	5.6 ± 2.2	6.1 ± 2.7	< 0.001
Blood glucose, mmol/L	6.8 ± 4.3	10.0 ± 47.4	8.6 ± 4.8	9.0 ± 4.3	0.177
HbA1c, %	6.5 ± 1.5	6.9 ± 1.7	7.1 ± 1.6	6.9 ± 1.8	< 0.001
Creatinine, umol/L	74.4 ± 20.3	88.8 ± 56.0	121.0 ± 54.9	114.9 ± 50.6	< 0.001
CCR, mL/min	90.1 ± 45.1	74.1 ± 72.7	51.7 ± 24.4	53.6 ± 26.5	< 0.001
TC, mmol/L	4.2 ± 1.0	4.1 ± 1.0	3.9 ± 1.4	3.9 ± 1.2	0.023
LDL-C, mmol/L	2.6 ± 0.8	2.7 ± 3.4	2.5 ± 1.1	2.2 ± 0.9	0.698
HDL-C, mmol/L	1.0 ± 0.3	1.0 ± 0.3	1.1 ± 0.3	1.1 ± 0.4	0.647
TG, mmol/L	1.9 ± 1.3	1.7 ± 1.0	1.5 ± 0.7	1.3 ± 0.6	< 0.001
UA, mmol/L	348.9 ± 94.1	363.0 ± 116.1	405.0 ± 158.7	416.8 ± 177.1	< 0.001
ALT, IU/L	43.8 ± 59.9	40.5 ± 35.2	70.8 ± 148.1	95.1 ± 123.0	< 0.001
AST, IU/L	95.1 ± 109.9	96.2 ± 122.8	119.4 ± 183.4	190.7 ± 338.1	< 0.001
Peak troponin I, ng/mL	15.5 ± 17.9	36.8 ± 331.1	15.8 ± 19.1	28.6 ± 19.4	0.211
NT-pro BNP, pg/mL	998.0 ± 2257.5	3550.4 ± 5890.4	10567.0 ± 9999.6	9448.9 ± 10009.3	< 0.001
hs-CRP, mg/L	10.9 ± 14.2	16.2 ± 17.0	28.4 ± 18.6	25.2 ± 19.0	< 0.001
IL-6, pg/mL	32.2 ± 82.1	49.3 ± 126.6	79.1 ± 117.3	1874.4 ± 9257.9	< 0.001
Leukocytes, $n\times10^3$/μL	9.5 ± 3.1	9.9 ± 3.3	11.0 ± 4.5	12.9 ± 4.4	< 0.001
Neutrophils, $n\times10^3$/μL	7.3 ± 3.6	7.7 ± 3.3	8.7 ± 4.3	10.6 ± 4.2	< 0.001
Lymphocytes, $n\times10^3$/μL	1.7 ± 0.8	1.6 ± 0.8	1.5 ± 0.9	1.5 ± 0.8	0.006
Monocytes, $n\times10^3$/μL	0.5 ± 0.2	0.6 ± 0.3	0.7 ± 0.4	0.7 ± 0.5	< 0.001
Platelet, $n\times10^3$/μL	228.2 ± 63.2	226.7 ± 77.1	236.1 ± 103.1	247.6 ± 115.6	0.357
Hemoglobin, g/L	138.1 ± 17.0	132.7 ± 19.1	98.8 ± 20.8	121.6 ± 19.8	< 0.001
Left atrial diameter, mm	37.5 ± 4.7	38.6 ± 5.8	40.1 ± 5.9	39.6 ± 7.1	< 0.001
LVEDD, mm	51.5 ± 5.1	52.6 ± 6.0	56.8 ± 7.4	54.3 ± 7.0	< 0.001
LVEF, %	57.2 ± 9.0	52.7 ± 11.0	42.3 ± 12.7	41.7 ± 9.8	< 0.001
Length of hospitalization	8.5 ± 3.8	10.3 ± 7.3	15.7 ± 11.4	17.4 ± 13.2	< 0.001

BMI: body mass index; CABG: coronary artery bypass grafting; AF: atrial fibrillation; CVD: cerebrovascular disease; PU: peptic ulcer; GRACE: Global Registry of Acute Coronary Events; TIMI: Thrombolysis in Myocardial Infarction; TC: total cholesterol; LDL-C: low density lipoprotein cholesterol; HDL-C: high density lipoprotein cholesterol; TG: triglyceride; UA: uric acid; ALT: alanine aminotransferase; AST: aspartate aminotransferase; NT-pro BNP: N-terminal pro-B-type natriuretic peptide; hs-CRP: high-sensitivity C-reactive protein; IL-6: interleukin 6; LVEDD: left ventricular end-diastolic diameter; LVEF: left ventricular ejection fraction

Model interpretation

As demonstrated in the SHAP summary plots of the TabNet model for four-class KILLIP classification (Fig. 3A and B), feature contributions were evaluated based on average SHAP values, presented in descending order. SHAP summary plots of other models were presented in Appendix Fig. 3 – Fig. 3. The GRACE and TIMI risk scores, NT-pro BNP, creatinine, and length of hospitalization had a negative impact on predicting "KILLIP 1," indicating that higher values of these features decreased the likelihood of a patient being classified as KILLIP 1. Conversely, LVEF and CCR exhibited a positive effect, increasing the probability of KILLIP 1 classification.

Table 3 Performance of machine learning and deep learning models in predicting KILLIP class (four-class classification). Fivefold cross-validation was performed in all the 1574 patients

	F1	Precision	AUPRC	AUROC
RF	0.786 ± 0.022	0.788 ± 0.009	0.674 ± 0.025	0.797 ± 0.012
XGboost	0.738 ± 0.023	0.761 ± 0.011	0.663 ± 0.027	0.783 ± 0.006
MLP	0.771 ± 0.022	0.775 ± 0.008	0.634 ± 0.028	0.814 ± 0.009
TabNet	0.783 ± 0.024	0.787 ± 0.012	0.684 ± 0.030	0.827 ± 0.005

Note: The values for all evaluation metrics are calculated using weighted averages. Results are presented as mean ± standard deviation across 5 stratified folds (random seed = 42)

Table 4 Performance of machine learning and deep learning models in predicting KILLIP class (binary classification). Fivefold cross-validation was performed in all the 1574 patients

	F1	Precision	AUPRC	AUROC
RF	0.763 ± 0.014	0.758 ± 0.008	0.764 ± 0.018	0.804 ± 0.005
XGboost	0.768 ± 0.018	0.767 ± 0.008	0.782 ± 0.022	0.798 ± 0.004
MLP	0.781 ± 0.016	0.786 ± 0.012	0.773 ± 0.024	0.824 ± 0.006
TabNet	0.774 ± 0.013	0.762 ± 0.009	0.779 ± 0.022	0.831 ± 0.008

Note: The values for all evaluation metrics are calculated using weighted averages. Results are presented as mean ± standard deviation across 5 stratified folds (random seed = 42)

In addition, the SHAP dependence plot provides insight into how individual features influence model predictions. The relationship between actual values and SHAP values for these features is illustrated in Fig. 4, where SHAP values greater than zero correspond to a positive class prediction, indicating a higher KILLIP grade. For example, patients with a GRACE score ≤ 150 or LVEF ≥ 48% had SHAP values above zero, pushing the model's decision toward the "KILLIP 1" class. Similarly, low actual values of NT-pro BNP (≤ 2500) and TIMI scores (≤ 3) also contributed to the prediction of KILLIP 1. For binary KILLIP classification, the corresponding SHAP summary plots for the TabNet model are provided in Appendix Figs. 4 and 5.

Furthermore, local explanations analyzed how specific predictions were made for individual patients using personalized input data. The raw data for one patient is presented in Appendix Table 1. Figure 5A, B, C, and D depict predictions for a patient with KILLIP classifications 1–4, respectively. For instance, Fig. 5A1–A4 shows this patient's probabilities for KILLIP classes 1, 2, 3, and 4, as predicted by the TabNet model: 92.4% (Fig. 5A1), 0.07% (Fig. 5A2), 0.005% (Fig. 5A3), and 0.001% (Fig. 5A4). For this patient, factors such as GRACE risk score, LVEF, and CCR strongly influenced the prediction toward the "KILLIP 1" class. Figure 6. Shown the force SHAP value plot for the test set. The corresponding SHAP local explanations for binary KILLIP classification are shown in Appendix Table 1 and Appendix Figs. 6, 7 and 8.

Convenient application for clinical utility

As illustrated in Fig. 7, we have integrated the KILLIP prediction model into a web platform to improve its clinical utility. By entering the actual values for all required features, the application can automatically predict the KILLIP classification for patients after AMI. The web platform is available online at https://prediction-killip-gby.streamlit.app/.

Discussion

In this study, we employed four machine learning and deep learning algorithms to predict the risk of HF after AMI using multidimensional clinical data. These computational methods are well-suited to managing complex and extensive datasets, making them highly effective for developing clinical prediction models. Their ability to

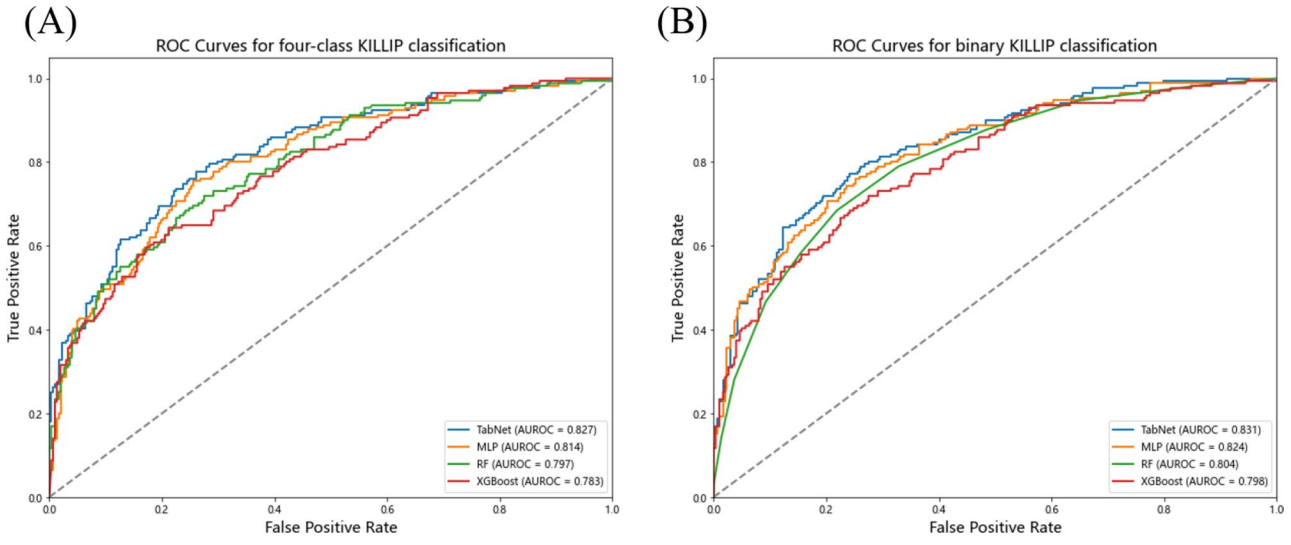

Fig. 2 ROC Curves of machine learning and deep learning models. Fivefold cross-validation was performed in all the 1574 patients. (**A**). ROC Curves for four-class KILLIP classification (**B**). ROC Curves for binary KILLIP classification

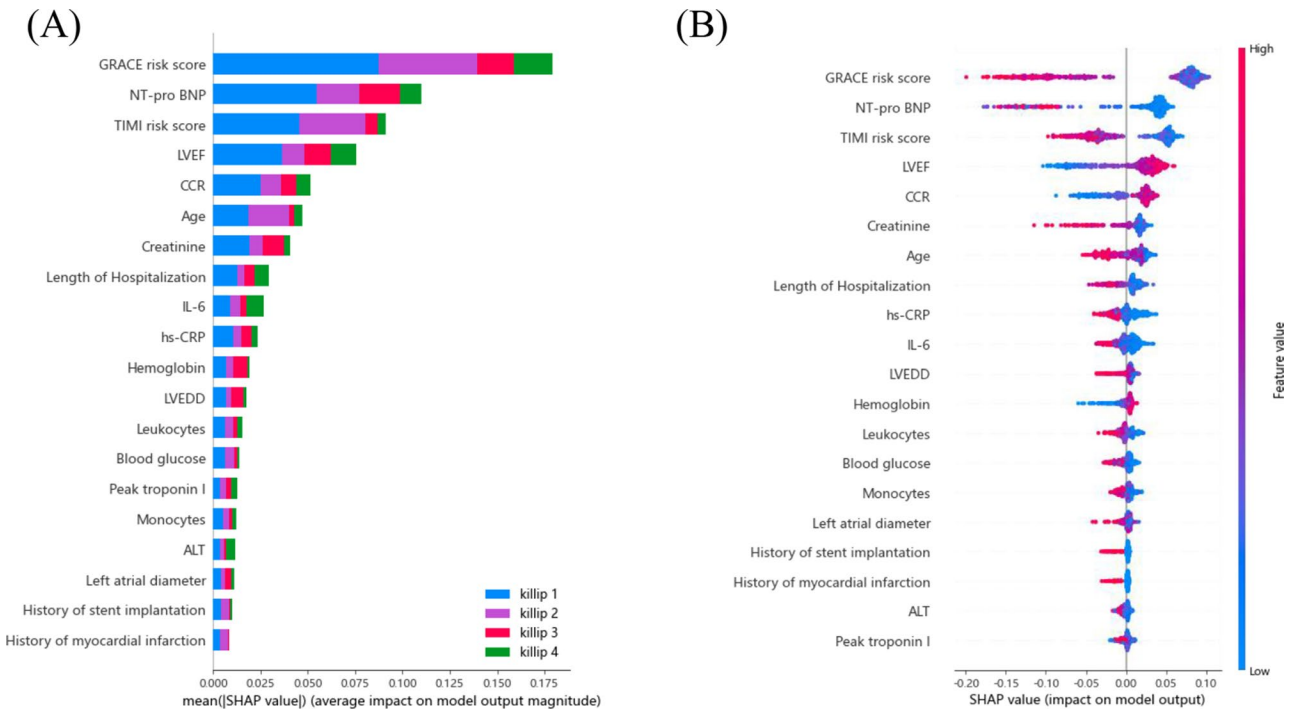

Fig. 3 Feature importance by the SHAP method for the Tabnet model. **(A)** SHAP summary bar plot derived from 1574 patients. **(B)** SHAP summary dot plot for KILLIP 1 classification (1005 patients). The colors of the dots represent the actual feature values for each patient, with red indicating higher values and blue indicating lower values. Dots are stacked vertically to represent density

handle diverse data types and identify intricate relationships between variables allows for improved accuracy in clinical risk predictions. By integrating easily accessible multidimensional clinical data with advanced machine learning and deep learning algorithms, we have enhanced the potential of clinical prediction tools in identifying patients at risk for post-AMI HF.

Indeed, several established scoring systems such as the MAGGIC risk score, Framingham criteria, and ADHERE risk tree have been developed to predict the onset or outcomes of heart failure. However, these tools are typically designed for use in broader heart failure populations (including both de novo HF and chronic HF), rather than specifically for acute heart failure prediction in the context of acute myocardial infarction. In contrast, our study aimed to predict the severity of heart failure during hospitalization following AMI, using the Killip classification as the outcome metric—a standard prognostic tool in AMI settings. While we did incorporate well-validated cardiovascular scores such as the GRACE and TIMI scores as input features, both of which have shown predictive value in post-AMI prognosis, including heart failure risk, we acknowledge the value of referencing other HF-specific risk tools for broader contextualization.

Among the models tested, the TabNet model achieved the highest AUROC value for KILLIP classification. TabNet effectively combines the strengths of deep learning and tree-based models, employing a sequential attention mechanism to select crucial features at each decision step. Prior research has demonstrated the TabNet method's excellent predictive value in medical contexts.

In medical data, the issue of class imbalance is common. To address this, we utilized the SMOTE over-sampling method to generate synthetic samples similar to the original ones, thereby increasing data diversity and enhancing the model's performance and generalizability. This approach better reflects real-world medical scenarios and provides more reliable support for research and clinical decision-making.

While machine learning and deep learning models are often regarded as "black boxes", their lack of interpretability can be a challenge in clinical settings. To improve transparency, we applied the SHAP method, which offers both global and local explanations of model predictions. SHAP helps elucidates the model's overall functionality and details how specific predictions are made for individual patients. By highlighting key clinical variables contributing to the risk of HF, SHAP visualizations can assist practitioners in identifying key factors early.

Our SHAP analysis revealed that higher GRACE risk score, TIMI risk score, age, and elevated levels of NT-pro BNP, creatinine, hs-CRP, and IL-6 were associated with an increased risk of HF after AMI. Conversely, higher CCR and LVEF were linked to a decreased risk. Although GRACE and TIMI scores were originally designed to predict mortality and recurrent myocardial infarction in

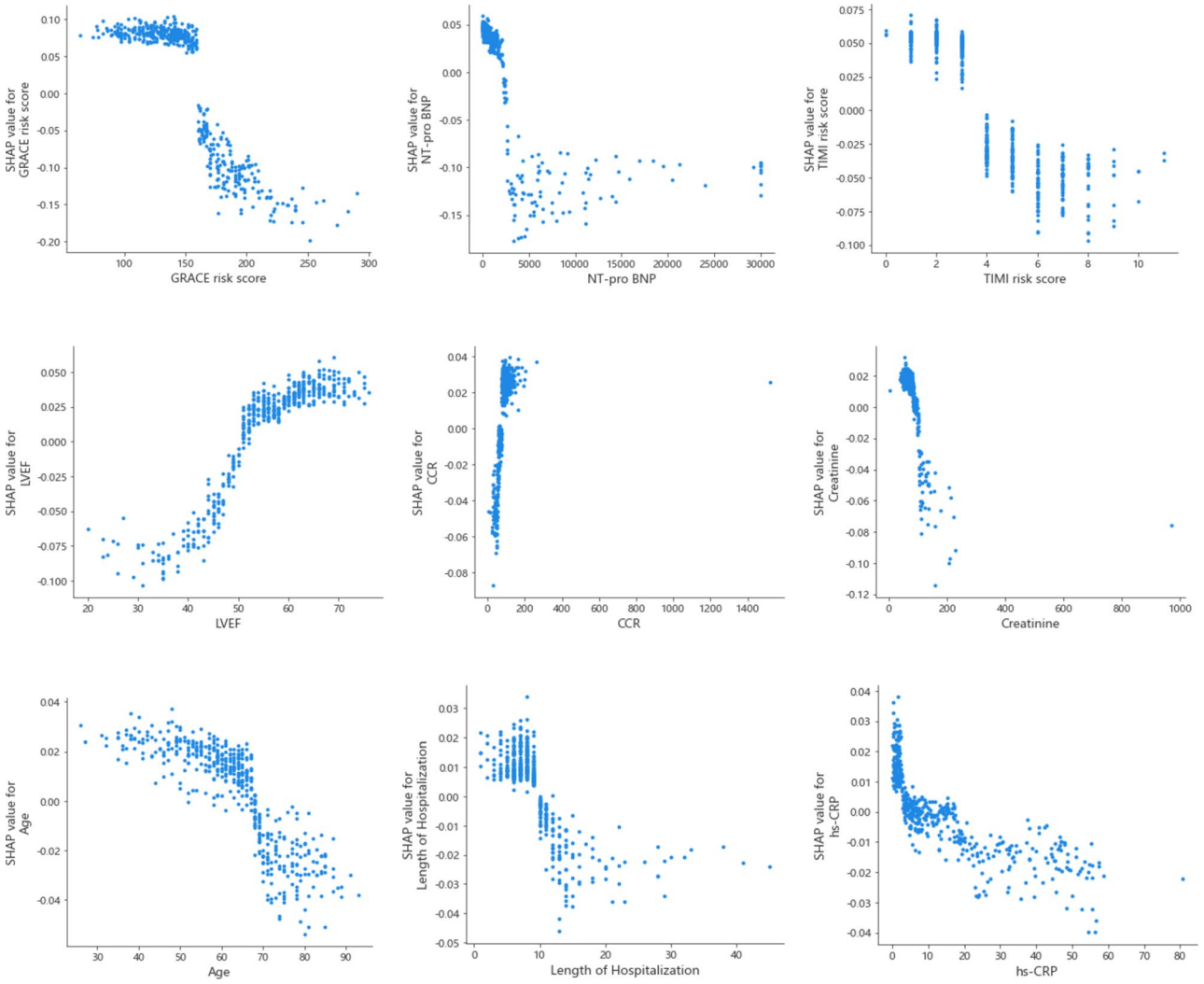

Fig. 4 Global model explanation by the SHAP method for the TabNet model. SHAP dependence plot for KILLIP 1 classification (1005 patients). Each dot represents a patient and shows how a single feature affects the model's output. SHAP values greater than zero push the decision toward the "KILLIP 1" class

acute coronary syndrome patients, they can also serve as indirect indicators of HF risk, as shown in previous studies [20, 21]. HF is a common complication following AMI, especially in cases with significant myocardial damage or additional cardiovascular risk factors. Age, a well-established risk factor of HF, is likely related to age-related changes in cardiac and vascular function [22, 23]. NT-pro BNP, a biomarker reflecting cardiac workload and function, is essential for diagnosing and prognosticating HF [24]. Additionally, elevated levels of inflammatory markers such as IL-6 and hs-CRP, have been linked to increased HF risk after AMI [25, 26]. Impaired kidney function, as indicated by elevated creatinine levels and decreased CCR, is closely related to HF development after AMI [27]. Finally, a decreased LVEF, which indicates impaired heart pumping function, is strongly associated with the development of HF and poor prognosis post-AMI. In summary, monitoring these factors is essential for early identification of high-risk patients and timely intervention to improve outcomes. However, it is critical to acknowledge that SHAP is a post hoc interpretability method. Its values reflect feature importance within the model's decision-making framework rather than genuine causal relationships or clinical pathophysiological mechanisms [28, 29].

KILLIP grading is traditionally based on changes in blood pressure and lung auscultation during AMI hospitalization. However, our study incorporated multidimensional clinical indicators into a predictive model, providing a more accurate assessment of HF risk than the KILLIP grading alone. The development of predictive models like ours enhances the comprehensive clinical thinking of cardiovascular physicians. To further facilitate clinical use, we are making the predictive models accessible through a web platform, aiming to promote widespread clinical application and adoption.

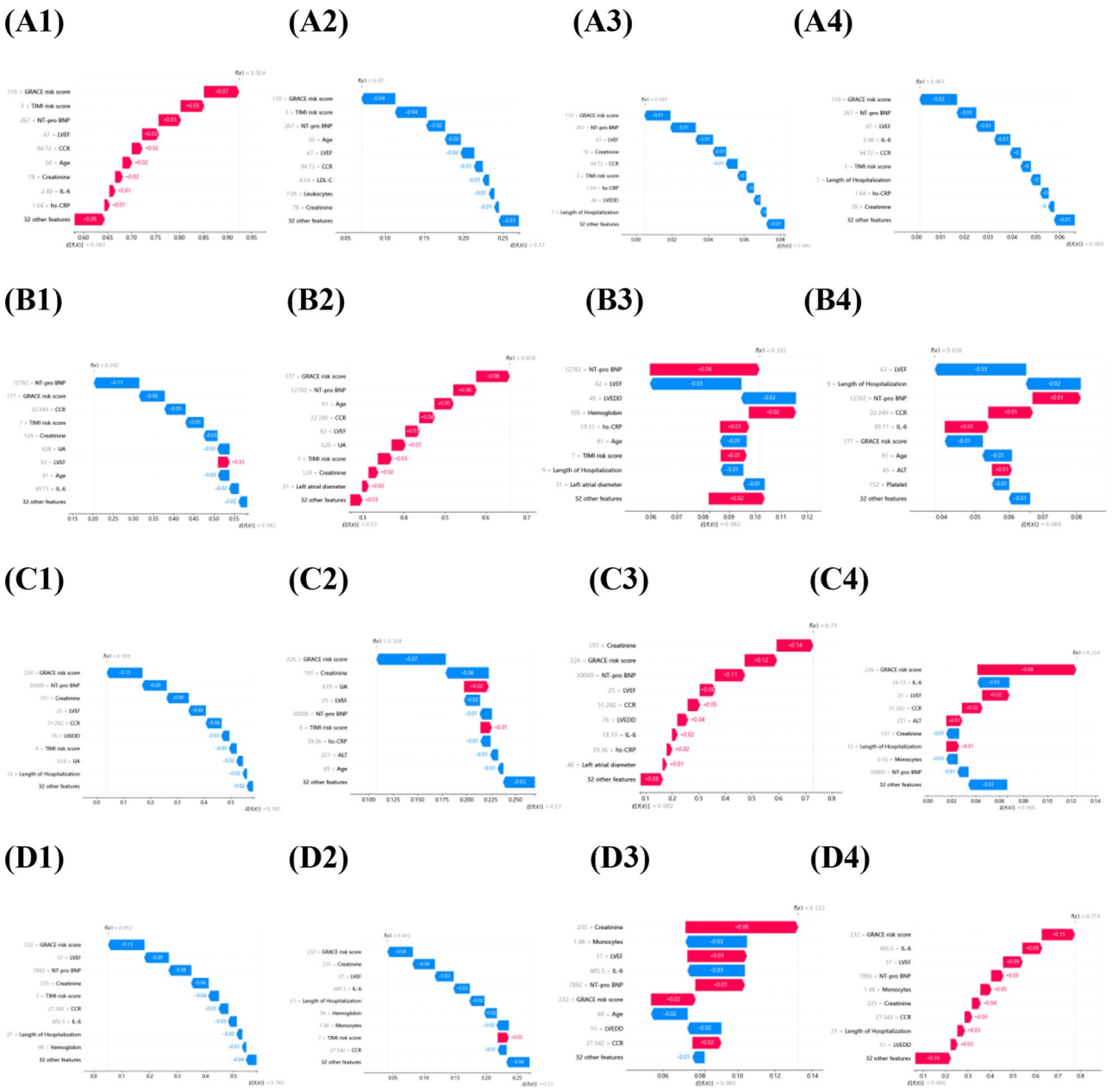

Fig. 5 Local model explanation by the SHAP method for the TabNet model. (**A1**–**D1**, **A2**-**D2**, **A3**-**D3**, **A4**-**D4**) represent prediction result plots for randomly selected patients from each KILLIP class 1 through 4. The raw data for each patient is presented in Appendix Table 1

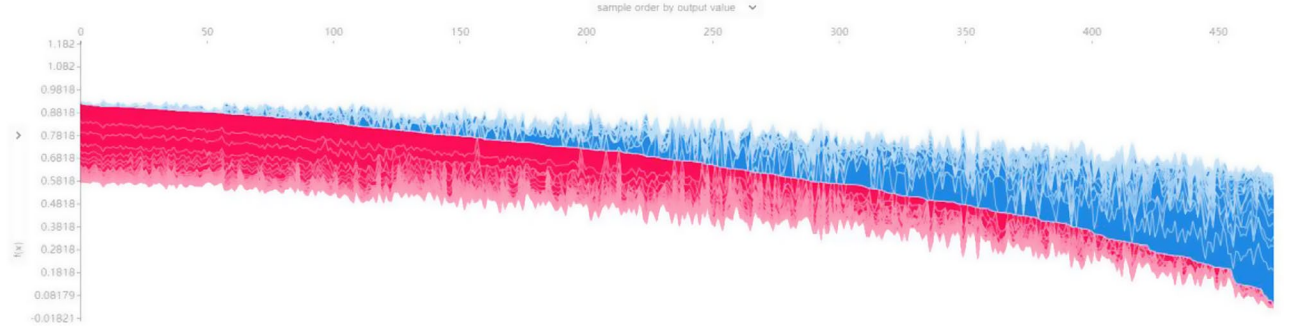

Fig. 6 Force SHAP value plot for the test set (315 patients). Each patient is represented along the x-axis, while the contributions of features are shown on the y-axis. A larger red area for an individual patient indicates a higher probability of the prediction being classified as "KILLIP 1."

Fig. 7 The web platform for KILLIP classification prediction model

There are several limitations to this study. First, the data were derived from a single-center dataset, and external validation with multi-center data is lacking. Second, the reliance on SHAP for interpretability introduces methodological constraints. As a post hoc explanation tool, SHAP values do not establish causality and may prioritize variables that are statistically predictive within the model rather than clinically actionable targets. This could lead to over-reliance on model-derived associations without rigorous biological validation. Third, the data modalities included in the study were somewhat limited. Specifically, our analysis focused primarily on structured clinical variables (e.g., laboratory biomarkers, risk scores, and demographic features), while omitting unstructured data modalities such as imaging data, longitudinal follow-up records, and genomic or proteomic biomarkers. This may restrict the model's ability to capture subtle pathophysiological interactions that could further refine HF risk stratification. Future research will focus on expanding the dataset by collecting more comprehensive clinical data from AMI patients across multiple institutions to refine and improve the accuracy of the prediction model.

Conclusion

By harnessing the power of artificial intelligence, we have developed a KILLIP classification prediction model to assess the risk of HF after AMI. This model enhances risk stratification, optimizes treatment strategies, guides early clinical interventions, reduces the incidence of post-AMI heart failure, and improves patient outcomes, demonstrating significant clinical utility. Its clinical utility is further demonstrated through its integration into a user-friendly web platform, accessible to both remote and local healthcare settings. The platform's visual design framework ensures that the predictive tool is both practical and actionable across a range of clinical environments.

Abbreviations

HF	Heart failure
AMI	Acute myocardial infarction
AI	Artificial intelligence
SHAP	Shapley Additive Explanation
RF	Random Forest
MLP	Multi-Layer Perceptron

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12872-025-04818-1.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

C.G. and B.G. drafted the manuscript, performed the machine learning models, and interpreted the data; X.H. and T.Z. conducted data curation, and preliminary analysis; T.T. contributed to analytical methodology development; H.L. and J.X. conceived and designed the study, supervised the research, and critically revised the manuscript. All authors have read and approved the final submitted manuscript.

Funding

This work was supported by the Beijing Natural Science Foundation (No. L246059, Z240021, 7242264), the National Natural Science Foundation of China (No.82100265) and the R&D Program of Beijing Municipal Education Commission (No. KM202310025020).

Data availability

The data and materials can be obtained from the authors upon reasonable request.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Xuanwu Hospital, Capital Medical University with the approval document number (2022–129) and was processed according to the principles of the Declaration of Helsinki. All enrolled patients signed informed consent forms.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 February 2025 / Accepted: 2 May 2025
Published online: 12 May 2025

References

1. Samsky MD, Morrow DA, Proudfoot AG, Hochman JS, Thiele H, Rao SV. Cardiogenic shock after acute myocardial infarction. JAMA. 2021;1840. https://doi.org/10.1056/NEJM199110173251601.
2. Hernandez AF, Udell JA, Jones WS, Anker SD, Petrie MC, Harrington J, Matheus M, Seide S, Zwiener I, Amir O, Bahit MC, Bauersachs J, Bayes-Genis A, Chen Y, Chopra VK, Figtree A, Ge G, Goodman JG, Gotcheva S, Goto N, Gasior S, Jamal T, Januzzi W, Jeong JL, Lopatin MH, Lopes Y, Merkely RD, Parikh B, Parkhomenko PB, Ponikowski A, Rossello P, Schou X, Simic M, Steg D, Szachniewicz PG, van der Meer J, Vinereanu P, Zieroth D, Brueckmann S, Sumin M, Bhatt M, Butler DL. Effect of empagliflozin on heart failure outcomes after acute myocardial infarction: insights from the EMPACT-MI trial. Circulation. 2024;149(21):1627–38. https://doi.org/10.1161/CIRCULATIONAHA.124.069217.
3. Hori Y, Sakakura K, Jinnouchi H, Taniguchi Y, Tsukui T, Hatori M, Kasahara T, Watanabe Y, Yamamoto K, Seguchi M, Fujita H. Determinants of serious in-hospital complications in patients with Killip class 1/2 ST-segment elevation myocardial infarction who underwent primary percutaneous coronary intervention. Heart Vessels. 2024 Mar;18. https://doi.org/10.1007/s00380-024-02382-w.
4. Deo RC. Machine learning in medicine. Circulation. 2015;1920–30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
5. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14. https://doi.org/10.1167/tvst.9.2.14.
6. Gao Z, Liu X, Kang Y, et al. Improving the prognostic evaluation precision of hospital outcomes for heart failure using admission notes and clinical tabular data: multimodal deep learning model. J Med Internet Res. 2024;26:e54363. https://doi.org/10.2196/54363.
7. Bat-Erdene BI, Zheng H, Son SH, Lee JY. Deep learning-based prediction of heart failure rehospitalization during 6, 12, 24-month follow-ups in patients with acute myocardial infarction. Health Inf J. 2022;28(2):14604582221101529. https://doi.org/10.1177/14604582221101529.
8. Li Y, Hu Y, Jiang F, Chen H, Xue Y, Yu Y. Combining WGCNA and machine learning to identify mechanisms and biomarkers of ischemic heart failure development after acute myocardial infarction. Heliyon. 2024;10(5):e27165. https://doi.org/S2405-8440(24)03196-7.

9. Li X, Shang C, Xu C, Wang Y, Xu J, Zhou Q. Development and comparison of machine learning-based models for predicting heart failure after acute myocardial infarction. BMC Med Inf Decis Mak. 2023;23(1):165. https://doi.org/10.1186/s12911-023-02240-1.
10. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. https://doi.org/10.1136/bmj.m689.
11. Popat A, Yadav S, Patel SK, Baddevolu S, Adusumilli S, Rao Dasari N, Sundarasetty M, Anand S, Sankar J, Jagtap YG. Artificial intelligence in the early prediction of cardiogenic shock in acute heart failure or myocardial infarction patients: A systematic review and Meta-Analysis. Cureus. 2023;15(12):e50395. https://doi.org/10.7759/cureus.50395.
12. Mohammad M, Olesen K, Koul S, et al. Development and validation of an artificial neural network algorithm to predict mortality and admission to hospital for heart failure after myocardial infarction: a nationwide population-based study. Lancet Digit Health. 2022;4:e37–45. https://doi.org/10.1016/S2589-7500(21)00228-4.
13. Arik SÖ. and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 8. 2021. https://doi.org/10.1609/aaai.v35i8.16826
14. Warren SMC, Walter Pitts. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5:115–33. https://doi.org/10.1007/BF02478259.
15. Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386. https://doi.org/10.1037/h0042519.
16. David E, Rumelhart GE, Hinton, Ronald J, Williams. Learning representations by back propagating errors. Nature. 1986;323(6088):533–6. https://doi.org/10.1038/323533a0.
17. Breiman L. Random forests. Machine learning 45 (2001): 5–32. https://doi.org/10.1023/A:1010933404324
18. Chen T. and Carlos Guestrin. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. https://doi.org/10.1145/2939672.2939785
19. Lundberg SM, Su-In, Lee. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30 (2017).
20. Guo C, Han X, Zhang T, Zhang H, Li X, Zhou X, Feng S, Tao T, Yin C, Xia J. Lipidomic analyses reveal potential biomarkers for predicting death and heart failure after acute myocardial infarction. Clin Chim Acta. 2024;562:119892. https://doi.org/10.1016/j.cca.2024.119892.
21. Zhang T, Han X, Zhang H, Li X, Zhou X, Feng S, Guo C, Song F, Tao T, Yin C, Xia J. Identification of molecular markers for predicting the severity of heart failure after AMI: an Olink precision proteomic study. Clin Chim Acta. 2024;555:117825. https://doi.org/10.1016/j.cca.2024.117825.
22. Peikert A, Martinez FA, Vaduganathan M, Claggett BL, Kulac IJ, Desai AS, Jhund PS, de Boer RA, DeMets D, Hernandez AF, Inzucchi SE, Kosiborod MN, Lam CSP, Shah SJ, Katova T, Merkely B, Vardeny O, Wilderäng U, Lindholm D, Petersson M, Langkilde AM, McMurray JJV, Solomon SD. Efficacy and safety of Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction according to age: the DELIVER trial. Circ Heart Fail. 2022;15(10):e010080. https://doi.org/10.1161/CIRCHEARTFAILURE.122.010080.
23. Redfield MM, Borlaug BA. Heart failure with preserved ejection fraction: A review. JAMA. 2023;329(10):827–38. https://doi.org/10.1001/jama.2023.2020.
24. Luo H, Xiang C, Zeng L, Li S, Mei X, Xiong L, Liu Y, Wen C, Cui Y, Du L, Zhou Y, Wang K, Li L, Liu Z, Wu Q, Pu J, Yue R. SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation. Sci Rep. 2024;14(1):17728. https://doi.org/10.1038/s41598-024-67844-7.
25. Alogna A, Koepp KE, Sabbah M, Espindola Netto JM, Jensen MD, Kirkland JL, Lam CSP, Obokata M, Petrie MC, Ridker PM, Sorimachi H, Tchkonia T, Voors A, Redfield MM, Borlaug BA. Interleukin-6 in patients with heart failure and preserved ejection fraction. JACC Heart Fail. 2023;11(11):1549–61. https://doi.org/10.1016/j.jchf.2023.06.031.
26. Gui XY, Rabkin SW, C-Reactive Protein. Interleukin-6, Trimethylamine-N-Oxide, Syndecan-1, nitric oxide, and tumor necrosis factor Receptor-1 in heart failure with preserved versus reduced ejection fraction: a Meta-Analysis. Curr Heart Fail Rep. 2023;20(1):1–11. https://doi.org/10.1007/s11897-022-00584-9.
27. Bart BA, Goldsmith SR, Lee KL, Givertz MM, O'Connor CM, Bull DA, Redfield MM, Deswal A, Rouleau JL, LeWinter MM, Ofili EO, Stevenson LW, Semigran MJ, Felker GM, Chen HH, Hernandez AF, Anstrom KJ, McNulty SE, Velazquez EJ, Ibarra JC, Mascette AM, Braunwald E. Heart failure clinical research network. Ultrafiltration in decompensated heart failure with cardiorenal syndrome. N Engl J Med. 2012;367(24):2296–304. https://doi.org/10.1056/NEJMoa1210357.
28. Slack D et al. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. https://doi.org/10.1145/3375627.3375830
29. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.