PREDICTION OF IN-HOSPITAL MORTALITY FOR ICU PATIENTS WITH HEART FAILURE

J. ZHANG[1] , H. LI[1] , N. ASHRAFI[1] , Z. YU[1] , G. PLACENCIA[2] , M. PISHGAR1[1*]

[1]Department of Industrial and Systems Engineering
University of Southern California, California, USA
pishgar@usc.edu

[2]Department of Industrial and Manufacturing Engineering
California State Polytechnic University, Pomona, California, USA
gvplacencia@cpp.edu

ABSTRACT

Heart failure affects millions of people worldwide. It greatly reduces quality of life and is associated with high mortality rates. Despite extensive research, the statistical connection between heart failure and mortality rates for ICU patients remains underexplored, indicating the need for improved prediction models.

This study identified 1,177 patients over 18 years old from the MIMIC-III database using ICD-9 codes. Preprocessing consisted of handling missing data, deleting duplicates, treating skewness, and oversampling to alleviate data imbalances. 18 features were selected within a LightGBM model by checking Variance Inflation Factor (VIF) values, LASSO Regression, and univariate analysis. The final output of the LASSO Logistic Regression model had the highest test AUC-ROC of 0.8766 (95% CI 0.8065 - 0.9429) and accuracy of 0.7291 compared to other baseline models, including Logistic Regression, Random Forest, LightGBM, Support Vector Machine (SVM), and Decision Trees. All models demonstrated good calibration with relatively low Brier scores, highlighting their reliability in predicting in-hospital mortality.

Our models predicted deaths of heart failure ICU patients better than the best results found in both literature and baseline models. These results were based on preprocessing missing values via improved imputation strategies and improved feature selection based on an expanded literature search and improved experiences selecting key features. With the Grid-Search, we had a near-perfect predictive model. These methods greatly increased the predictive accuracy of in-hospital mortality in ICU patients with heart failure.

Keywords: Heart Failure, In-Hospital Mortality, MIMIC-III, Machine Learning

1 INTRODUCTION

Heart failure (HF) affects approximately 6.5 million Americans aged 20 years and over making it a critical field of study. Those with HF experience severe symptoms including difficulty breathing, excessive coughing, and ultimately early death in about a quarter of cases at 1 year. [1] Hospitalization of HF patients often involve a serious infection, e.g. sepsis, in 20% of patients admitted to the ICU with life-threatening conditions. [2] Without further treatment, in-hospital mortality of HF patients will continue to be almost 10% based on [3].

Predictive models forecasting HF patient death in ICUs are critical. The introduction of Electronic Health Records (EHR) has positively affected patients' treatment by utilizing

*Corresponding Author

information generated by the application of the data, improving performance, and increasing efficiency. [4, 5] Machine learning (ML) methods can find patterns and correlations among various features in large, complex data sets. This has improved doctors' ability to diagnose and cure heart failure. [6 - 8] Several studies attempted to develop models to forecast HF patient deaths in ICUs as well with unreliable results. [9 - 10].

Feature selection can choose the most statistically significant attributes, which helps to build better models and to avoid overfitting. Hyperparameters are preset in ML models and can be tuned to fit specific situations and make better predictions. Combining these two methods enables models to make good forecasts that are cost- and resource-sensitive, reliable and actionable in medical fields as evidenced by Gao et al. [11].

Our primary research developed inventive feature selection and data processing techniques to improve predictions. We conducted systematic imputation strategies on a distribution of factors and used univariate analyses based on VIF and Random Forest methods. This resulted in better AUC-ROC than found in the literature. This study conformed with the TRIPOD guidelines, namely, Moons et al. [12] and Amritphale et al. [13]

2 METHOD

2.1 Data Availability

The MIMIC-III (version 1.4) database is an extensive, publicly accessible database that recorded 38,597 adult patients and 49,785 hospital admissions who stayed in ICUs of the Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. This dataset includes information on admissions, patient demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including dates and times).[14] The dataset is also deidentified therefore secondary which does not require approval by an institutional review board nor informed consent. [15])

2.2 Patient Selection

We restricted our study to 1,177 adult patients in the MIMIC-III database diagnosed with heart failure and who were admitted to the ICU. The target group initially consisted of 13,389 patients above 18 years of age selected using relevant International Classification of Diseases-9th Revision (ICD-9) codes. 162 patients without ICU admission were excluded. Another 4,871 patients, who did not have the N-terminal pro b-type natriuretic peptide (NT-proBNP) record were excluded from our study data, as NT-proBNP is a critical marker of heart failure. [16] Lastly, 7,179 were excluded for lack of echocardiography records, a basic tool for heart failure evaluation. [17] The process is detailes in Figure 1.

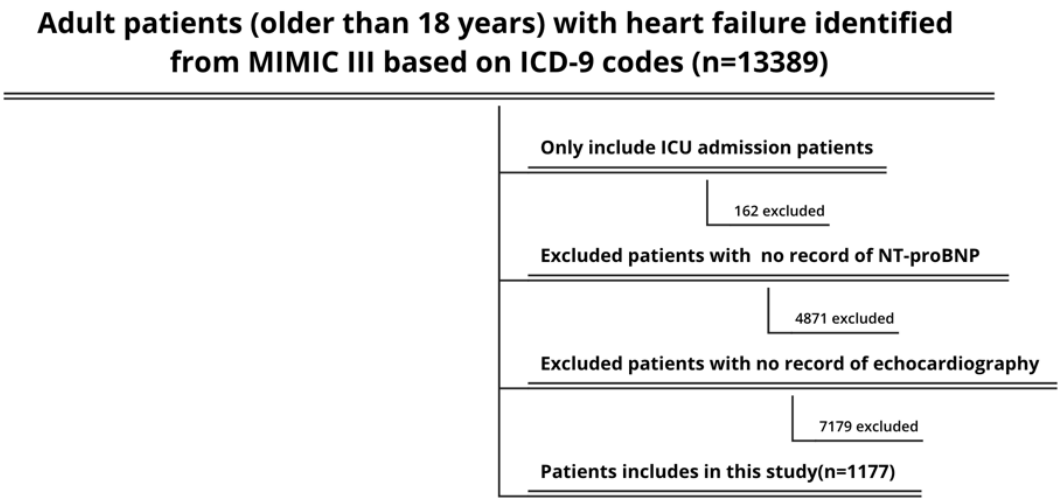

Figure 1: Flow chart illustrating data extraction process for study patient

2.3 Feature Extraction

Using Structured Query Language (SQL) with PostgreSQL (V.9.6), demographic characteristics, vital signs, and laboratory values were extracted from the MIMIC-III dataset. Based on expert opinions, previous studies (Ashrafi et al. [18], Abraham et al. [19], Peterson et al. [20], Jia et al. [21], Lagu et al. [22] and Wang et al. [23]), and clinical relevance, 42 features were extracted from the original MIMIC-III database. Demographic characteristics and vital signs were recorded during the first 24 hours of each admission. Laboratory variables were measured throughout the entire ICU stay. Mean values were analyzed for features with multiple measurements. The primary outcome was in-hospital mortality, defined as 1 if the patient died during their ICU stay or 0 if the patient survived.

2.4 Pre-processing

We began the preprocessing stage by reading and cleaning the raw data obtained from the MIMIC-III database. We excluded missing values, columns with single unique values, and duplicate entries. Rows with null values in the outcome column were also discarded, and variables 'group' and 'ID' were deemed irrelevant.

For imputation, we used median imputation due to skewness and outliers in most features. We addressed outliers by selectively removing extreme data points. We then assessed the distribution of different classes within the outcome variable and identified an imbalance. To address this, we employed oversampling to balance the classes in the training set.

2.5 Feature Selection

The Variance Inflation Factor (VIF) was calculated to prevent multicollinearity in continuous features that could cause high standard errors in the prediction model (Murad et al. [24] and Lafi et al. [25]). Setting a lower VIF limit at 5, as suggested by Zach [26], we deleted variables with multicollinearity to reduce potential variables for modelling to 42.

LASSO regression was used for feature selection for numeric variables because it performaned best eliminating redundant or less informative features as detailed by Muthukrishnan et al. [27]. This process resulted in a small subset of numeric variables with importance scores greater than 0. We applied Random Forest for categorical data, as it best predicted accuracy

for categorical data, and identified 10 categorical features with non-zero feature importance values. [28].

The univariate method based on LightGBM was used to order the significance level of all features. A threshold of 0.05 was used to eliminate features that contributed less, based on Bolón-Canedo et al. [29] This reduced the number of variables from 42 to 18 (see Table 1.)

Table 1: Summary of the features selected by category

Category	Feature
Demographic	BMI
Vital signs	Urine_output, Cystolic_blood_pressure, Temperature
Comorbidities	Atrial_fibrillation, Renal_failure, Depression, Hyperlipidemia, Deficiency_anemias, Hypertensive, COPD
Laboratory variables	Leucocyte, RDW, Basophils, Platelets, NT-proBNP, Magnesium_ion, Creatinine
Outcome variable	Mortality

2.6 Modeling

We deliberately used a suite of machine learning models: Logistic Regression, LASSO Logistic Regression, Random Forest, LightGBM, Support Vector Machine, and Decision Tree, as they best analyze complex health data. Grid-Search was used to find the optimum set of hyperparameters for each model. Model performance was assessed by the AUC-ROC values against the test set.

Evaluation criteria included bootstrapped 95% confidence intervals of accuracy and AUC-ROC. Higher AUC-ROCs indicate better discrimination and confidence intervals to estimate model performances. These models facilitated exploring data resources and their elucidation. The Logistic Regression model, with the highest AUC-ROC and narrowest confidence interval, was proposed to predict ICU mortality in heart failure patients along with the other baseline models. The whole process is summarized in Figure 2.

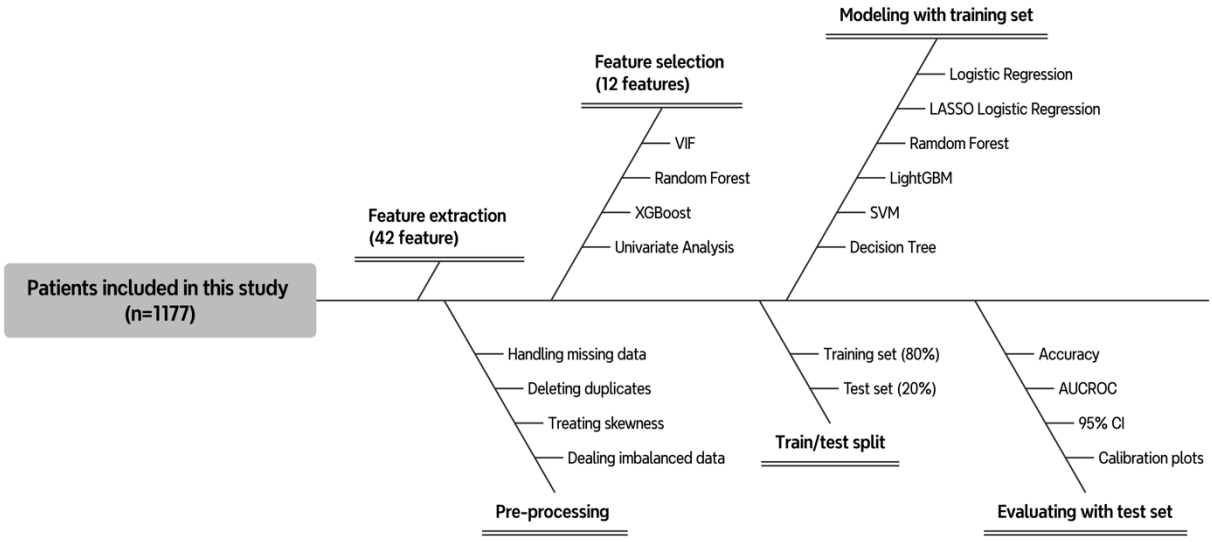

Figure 2: Flow chart illustrating study design

3 RESULT

3.1 Model Evaluation

Table 2 summarizes results of our proposed model and baseline ML models using our evaluation metric. For our Logistic Regression model, the accuracy score is 0.7291 with AUC-ROC values of 0.7710 (95% CI 0.7458 - 0.7927) and 0.8766 (95% CI 0.8065 - 0.9429) for training and test sets respectively. Among baseline models, the LASSO Logistic Regression model performed best, yielding an accuracy score of 0.7291 and AUC-ROC of 0.7712 (95% CI 0.7462 - 0.7927) and 0.8754 (95% CI 0.8038 - 0.9420) in training and test sets (Figure 2).

Table 2: Evaluation results of proposed and baseline models

	Accuracy	AUCROC (training)	95% CI (training)	AUCROC (test)	95% CI (test)	Brier scores
Logistic	0.7291	0.7710	0.7458 - 0.7927	0.8766	0.8065 - 0.9429	0.1103
LASSO	0.7291	0.7712	0.7462 - 0.7927	0.8754	0.8038 - 0.9420	0.1099
RF	0.9064	1.0000	1.0000 - 1.0000	0.7964	0.6990 - 0.8794	0.0952
LightGBM	0.8867	1.0000	1.0000 - 1.0000	0.7240	0.6165 - 0.8272	0.0830
SVM	0.9113	1.0000	1.0000 - 1.0000	0.7090	0.5761 - 0.8308	0.1088
DT	0.7882	0.9896	0.9863 - 0.9930	0.5746	0.4526 - 0.6990	0.1136

Calibration plots (Figure 3) and Brier scores (Table 2) highlighted the reliability of the predicted probabilities. The Logistic Regression model showed good calibration and a Brier score of 0.1103, indicating reliable predictions. The LASSO Logistic Regression and SVM models also had good calibration with Brier scores of 0.1099 and 0.1088, respectively. The Random Forest and LightGBM models had lower Brier scores (0.0952 and 0.0830). It shows these models are well-calibrated. The Decision Tree model, which had the highest Brier score (0.1136) and calibration, demonstrated reasonable predictive capability, albeit not as good as the other models. These results highlighted the importance of both AUC-ROC and calibration in evaluating model performance for clinical decision-making.

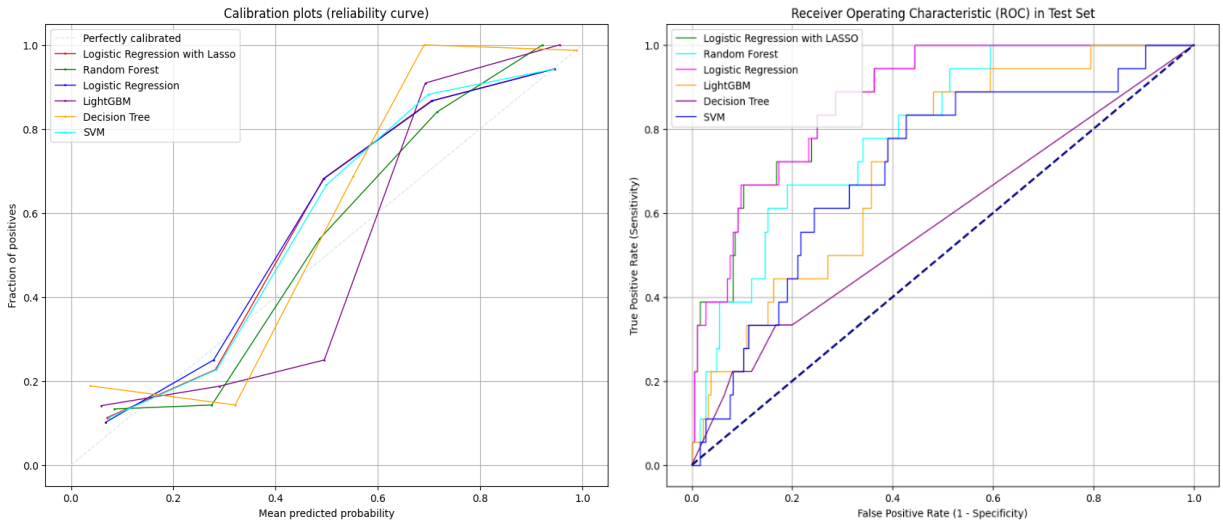

Figure 3: Receiver Operating Characteristic (ROC) Curve and Calibration Plots in Test Set

3.2 Model Comparison

Logistic Regression and LASSO Logistic Regression models showed superior predictive performance. Based on Fein et al. [30], we used a within-subject t-test to determine the difference between the two models using the same dataset. We used 500 bootstrapped AUCROC score for the test. The null hypothesis was set to be "there is no difference between

the Logistic Regression model and LASSO Logistic Regression model," and the alternative hypothesis to be "there exists a difference between these two models.". The small p-value of $1.0822x10^{-43}$ (Table 3) indicated we could reject the null hypothesis with 0.1% statistically significant level, meaning the models were dissimilar. The positive mean difference indicated that the Logistic Regression model had better prediction on average compared with LASSO Logistic Regression. Hence, among all the models we built, the Logistic Regression model was chosen for our proposed model with an AUCROC of 0.8766 (95% CI 0.8065 - 0.9429).

Table 3. T-test results using 500 bootstrapped AUCROC between Logistic Regression and LASSO Logistic Regression models

	P-Value	T-Statistic	Mean Difference
Logistic vs LASSO	1.0822x10-43	15.3184	0.0013

3.3 SHapley Additive ExPlanations (SHAP) Analysis

SHAP analysis helped identify which features most influenced the ML model prediction, as suggested by Hamilton et al. [31]. Importance values were computed and presented by descending order (Figure 4). Leukocyte seemed to be the most crucial feature in mortality prediction of HF patients in ICU. The majority of points in red for Leucocyte, RDW, Creatinine, Magnesium_ion, NTproBNP, and temperature were positioned on the right side of the zero-center line. This suggests fatal outcomes for patients with higher values of these features. In contrast, parameters like Urinary output, Platelets, Basophils, BMI, and systolic blood pressure had more red points on the left side. This indicated ICU-HF patients exhibiting low values of these parameters had higher possibility of death.

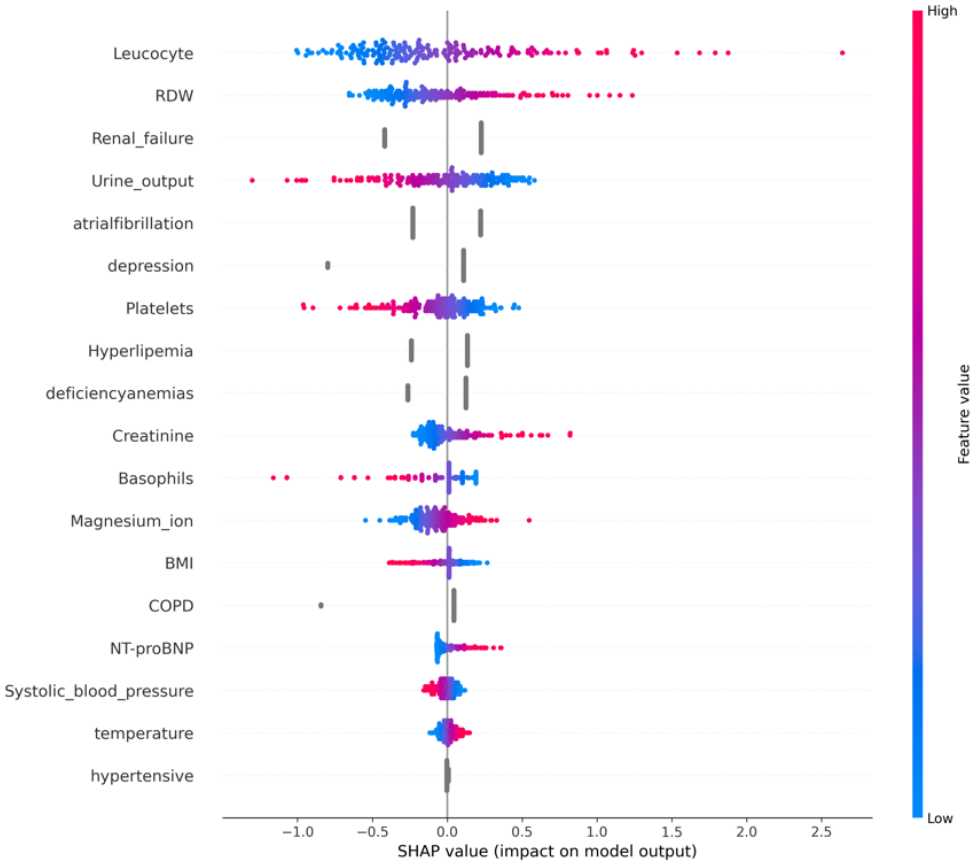

Figure 4: SHAP values for each variable using the proposed Logistic Regression model

4 DISCUSSION

Many studies have tried to predict in-hospital mortality among ICU patients with heart failure using the MIMIC-III database. The majority of these studies though were limited by feature selection and imbalanced datasets. For example, Chiu et al. [32] utilized an ensemble algorithm to generalize a combination of models but had difficulties dealing with feature selection and managing imbalanced data.

Our study mitigated such issues using an innovative method of missing value imputation by equal distribution for one variable after a systematic imputation based on the median value of each variable. This minimized undesired effects of skewed data and outliers. Also, this study employed thorough feature selection techniques of VIF, LASSO, and Random Forest, and univariate analyses within the LightGBM framework. This method provided the most relevant features to improve the predictive ability of the model. Furthermore, the use of Grid-Search for hyperparameter fine-tuning guaranteed that each hyperparameter was set at its best level. The adjustments led to the impressive performance of the Logistic Regression model. Applying SHAP helped us to understand the decision-making process of the model, thereby promoting understanding of the clinical issues. The proposed Logistic Regression model achieved an AUC-ROC of 0.8766, a 9.18% improvement over the best AUC-ROC of 0.8029 reported by Li et al., who used XGBoost and LASSO regression models. [3]

5 CONCLUSION

This research focused on developing an ML model to predict the mortality of ICU patients with HF using data from the MIMIC-III database. We compared five baseline models to our proposed Logistic Regression model. The Logistic Regression model demonstrated superior performance over baseline models and the best existing models, achieving a higher AUC-ROC and a narrower 95% confidence interval. All models showed strong calibration and low Brier scores to validate their robustness and accuracy in predicting patient survival outcomes.

These enhancements were attributed to a rigorous feature selection process that reduced the initial set of features to 18 key variables. Comprehensive hyperparameter tuning using Grid-Search optimization ensured the best possible performance of the Logistic Regression model. SHAP analysis confirmed the clinical relevance of selected features, such as leucocyte count and RDW, further validating the model's robustness.

Our framework offers valuable support to medical professionals by helping them identify ICU HF patients at high mortality risk. The predictive model evaluates the risk of death using various biomarkers, enabling clinicians to implement preventive actions effectively. This capability is particularly advantageous in critical care settings, where timely and accurate predictions can significantly impact patient outcomes.

6 REFERENCES

[1] **Emmons-Bell, S., Johnson, C., Roth, G.** 2022. Prevalence, incidence and survival of heart failure: a systematic review, Heart, 108(17), pp. heartjnl-2021-320131.

[2] **Dlugacz YD, Stier L, Lustbader D, Jacobs MC, Hussain E, Greenwood A.** Expanding a Performance Improvement Initiative in Critical Care from Hospital to System. The Joint Commission Journal on Quality Improvement. 2002 Aug;28(8):419–34.

[3] **Li J, Liu S, Hu Y, Zhu L, Mao Y, Liu J.** Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. Journal of Medical Internet Research. 2022 Aug 9;24(8):e38082.

[4] **Miotto R, Li L, Kidd BA, Dudley JT.** Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6(1):26094.

[5] **Lin YW, Zhou Y, Faghri FAO, Shaw MJ, Campbell RH.** Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS ONE. 2019;14(7):e0218942.

[6] **Rahmani AM, Yousefpoor E, Yousefpoor MS, Mehmood Z, Haider A, Hosseinzadeh M, et al.** Machine Learning (ML) in Medicine: Review, Applications, and Challenges. Mathematics. 2021 Nov 21;9(22):2970.

[7] **Sidey-Gibbons JAM, Sidey-Gibbons CJ.** Machine learning in medicine: a practical introduction. BMC Medical Research Methodology [Internet]. 2019;19(1):64.

[8] **Brnabic A, Hess LM.** Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Medical Informatics and Decision Making. 2021 Feb 15;21(1).

[9] **Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z.** 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 [Internet]. 2021 Jul 23;11(7):e044779.

[10] **Chiu CC, Wu CM, Chien TN, Kao LJ, Li C, Jiang HL.** Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure. Journal of Clinical Medicine [Internet]. 2022 Jan 1 [cited 2024 May 14];11(21):6460.

[11] **Gao, J., Lu, Y., Ashrafi, N., Domingo, I., Alaei, K., & Pishgar, M.** (2024). Predicting sepsis mortality using machine learning methods. *medRxiv*.

[12] **Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Collins GS.** Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Internal Med. 2015;162(1):W1–73

[13] **Amritphale A, Chatterjee R, Chatterjee S, Amritphale N, Rahnavard A, Awan GM, Fonarow GC.** Predictors of 30-day unplanned readmission after carotid artery stenting using artifcial intelligence. Adv Ther. 2021;38(6):2954–72.

[14] **A. E. W. Johnson et al.,** "Mimic-III, a freely accessible Critical Care Database," Scientific Data, vol. 3, no. 1, 2016. doi:10.1038/sdata.2016.35

[15] **Johnson A, Pollard T, Mark R.** MIMIC-III Clinical Database [Internet]. Physionet.org. 2016. Available from: https://physionet.org/content/mimiciii/1.4/

[16] **Johnson A, Pollard T, Shen L, Lehman LW, Feng M, Ghassemi M, et al.** OPEN SUBJECT CATEGORIES Background & Summary. 2016; Available from: https://lcp.mit.edu/pdf/JohnsonSD2016.pdf

[17] **Krishnamoorthy VK, Sengupta PP, Gentile F, Khandheria BK.** History of echocardiography and its future applications in medicine. Critical Care Medicine. 2007 Aug;35(Suppl):S309–13.

[18] **Ashrafi, N., Liu, Y., Xu, X., Wang, Y., Zhao, Z., & Pishgar, M.** (2024). Deep learning model utilization for mortality prediction in mechanically ventilated ICU patients. *medRxiv*.

[19] **Abraham WT, Fonarow GC, Albert NM, et al.** Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the organized program to initiate

lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF). J Am Coll Cardiol 2008;52:347–56. 10.1016/j.jacc.2008.04.028

[20] **Peterson PN, Rumsfeld JS, Liang L, et al.** A validated risk score for in-hospital mortality in patients with heart failure from the American heart association get with the guidelines program. Circ Cardiovasc Qual Outcomes 2010;3:25–32.10.1161/CIRCOUTCOMES.109. 854877

[21] **Jia Q, Wang Y-R, He P, et al.** Prediction model of in-hospital mortality in elderly patients with acute heart failure based on retrospective study. J Geriatr Cardiol 2017;14:669–78. 10.11909/j.issn.1671-5411.2017.11.002

[22] **Lagu T, Pekow PS, Stefan MS, et al.** Derivation and validation of an in-hospital mortality prediction model suitable for profiling Hospital performance in heart failure. J Am Heart Assoc 2018;7. 10.1161/JAHA.116.005256.

[23] **Wang N, Gallagher R, Sze D, et al.** Predictors of frequent readmissions in patients with heart failure. Heart Lung Circ 2019;28:277–83. 10.1016/j.hlc.2017.10.024

[24] **Murad MH, Wang Z, Chu H, Lin L.** When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. BMJ. 2019 Jan 22;k4817.

[25] **Lafi, S.; Kaneene, J.** An explanation of the use of principal-components analysis to detect and correct for multicollinearity. Prev. Vet. Med. 1992, 13, 261–275.

[26] **Zach.** A Guide to Multicollinearity & VIF in Regression [Internet]. Statology. 2019.

[27] **Muthukrishnan R, Rohini R.** LASSO: A feature selection technique in predictive modeling for machine learning [Internet]. IEEE Xplore. 2016. p. 18–20.

[28] **Speiser JL, Miller ME, Tooze J, Ip E.** A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert Systems with Applications. 2019 Nov;134(1):93–101.

[29] **Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A.** A review of feature selection methods on synthetic data. Knowledge and Information Systems. 2012 Mar 30;34(3):483–519.

[30] **Fein EC, Gilmour J, Machin T, Hendry L.** Section 3.4: Paired T-test Assumptions, Interpretation, and Write Up. usqpressbookspub [Internet]. 2022 Jun 16

[31] **Hamilton RI, Papadopoulos PN.** Using SHAP Values and Machine Learning to Understand Trends in the Transient Stability Limit [Internet]. arXiv.org. 2023.

[32] **Chiu, C. C., Wu, C. M., Chien, T. N., Kao, L. J., Li, C., & Jiang, H. L. (2022).** Applying an improved stacking ensemble model to predict the mortality of ICU patients with heart failure. Journal of Clinical Medicine, 11(21), 6460.