PHConference Proceedings 2022 ISBN: 978-86-87043-83-1

Predicting the Mortality of ICU Patients with Heart Failure: An Improved Stacking Ensemble Model

Te-Nien Chien
National Taipei University of Technology, College of Management
1, Sec. 3, Zhongxiao E. Rd., Taipei, Taiwan
tenienchien@gmail.com

Chengcheng Li
National Taipei University of Technology, College of Management
1, Sec. 3, Zhongxiao E. Rd., Taipei, Taiwan
chengchengli0006@gmail.com

Han-Ling Jiang
University of Manchester, Alliance Manchester Business School
Booth Street West, Manchester, M15 6PB, UK
hanling.jiang@gmail.com

Ta-Te Lee
Taipei Medical University, School of Health Care Administration
172-1, Sec. 2, Keelung Rd., Taipei, Taiwan
b908109088@tmu.edu.tw

ABSTRACT

Cardiovascular diseases have been identified as one of the top three causes of death worldwide, with onset and deaths mostly due to heart failure (HF). In ICU, where patients with HF are at increased risk of death and consume significant medical resources, early and accurate prediction of the time of death for patients at high risk of death would enable them to receive appropriate and timely medical care. The data for this study were obtained from the MIMIC-III database, where we collected vital signs and tests for 6,699 HF patient during the first 24 hours of their first ICU admission. In order to predict the mortality of HF patients in ICUs more precisely, an integrated stacking model is proposed and applied in this paper. In the first stage of dataset classification, the datasets were subjected to first-level classifiers using RF, SVC, KNN, LGBM, Bagging, and Adaboost. Then the fusion of these six classifier decisions was used to construct and optimize the stacked set of second-level classifiers. The results indicate that our model obtained an accuracy of 95.25% and AUROC of 82.55% in predicting the mortality rate of HF patients, which demonstrates the outstanding capability and efficiency of our method. In addition, the results of this study also revealed that platelets, glucose, and blood urea nitrogen were the clinical features that had the greatest impact on model prediction. The results not only improve the understanding of patients' conditions by healthcare professionals but allow for a more optimal use of healthcare resources.

KEYWORDS: heart failure, mortality, intensive care units, machine learning, stacking, electronic health records.

 ISBN: 978-86-87043-83-1

1 INTRODUCTION

Cardiovascular diseases (CVD) have ranked among the top three causes of death worldwide for many years, accounting for an estimated 18.9 million deaths per year, or approximately 31% of global mortality (Virani et al., 2020). The majority of CVD morbidity and mortality are derived from Heart Failure (HF), a common cardiovascular disease in which the heart fails to maintain the body's metabolism. Patients with HF experience a variety of overt symptoms such as shortness of breath, swollen ankles, and physical fatigue, and may also show signs of elevated jugular venous pressure, pulmonary fissures, and peripheral edema caused by cardiac or noncardiac structural abnormalities (L. Chen et al., 2021). As a major cause of cardiovascular morbidity and mortality, HF poses a significant threat to human health and social development (F. Li et al., 2021). In the face of the increasing number of people with HF, despite the rapid advances in medical technology and significant technological advances in diagnosis, assessment, and cardiovascular disease (Incidence, 2017), HF remains a major medical problem worldwide.

According to the World Federation of Societies of Intensive and Critical Care Medicine, an Intensive Care Unit (ICU) is an organized system for the provision of care to critically ill patients that provides intensive and specialized medical and nursing care, an enhanced capacity for monitoring, and multiple modalities of physiologic organ support to sustain life during a period of life-threatening organ system insufficiency(Marshall et al., 2017). As a result, the healthcare system is under a heavy burden, which may affect the criteria for ICU admission, the interventions used, and the duration, all of which may affect the patient's prognosis (Haase et al., 2020). Clinical decision-making in the ICU is time critical and highly dependent on the analysis of physiological data. If there is not enough real-time patient information to make accurate and rapid decisions in a dynamic and rapidly changing environment, it will be challenging for medical professionals to make clinical decisions (W. Chen et al., 2019). Patients admitted to the ICU require close and continuous monitoring to avoid the possibility of rapid deterioration of their health status, therefore, intensive monitoring through ICU equipment generates a large number of medical records and requires efficient and accurate systems to aid in data analysis (El-Rashidy et al., 2020). Using big data for clinical and basic research analysis and applications to improve human well-being and health, such as big data combined with artificial intelligence, can help doctors diagnose and treat diseases and improve the quality of care. Predictive models have been developed over the last few decades as important risk assessment tools and are utilized in a variety of healthcare settings(Kim et al., 2021).

Machine learning (ML) is a branch of artificial intelligence that focuses on training computers to learn from data collected and make improvements based on learned experience, and it is concerned with the problem of constructing computer programs that can automatically improve the accuracy of their output based on experience (Mitchell, 1997). Recently, ML has been increasingly introduced into clinical practice, with applications including preclinical data processing, bedside diagnostic assistance, patient stratification, treatment decision-making, and early warning as part of primary and secondary prevention(Adlung et al., 2021). In recent years, some researchers have applied ML techniques for mortality prediction in HF patients. For example, Negassa et al., developed an ensemble model for the prediction of 30-day mortality in patients with HF after discharge from hospital (Negassa et al., 2021). Luo et al., constructed a risk stratification tool by utilizing the extreme gradient boosting algorithm to correlate clinical features of HF patients with in-hospital mortality rates(Luo et al., 2022). Therefore, accurate prediction of hospital mortality has remained a challenge to date.

 ISBN: 978-86-87043-83-1

Given that each ML method may outperform or have shortcomings in different situations, developing a model that integrates multiple ML methods to obtain better performance has become a new research approach. Bagging involves training several base learners with different bootstrap samples, then consolidating them and voting on the result (Breiman, 1996). Stacking is a powerful ensemble technique that harnesses the predictions of multiple base learners as features to train new meta learners (Wolpert, 1992). Jia (Jia et al., 2021) proposed a stacked approach for ML to efficiently and rapidly construct 3D multi-type rock models using geological and geophysical datasets. In this paper, we deployed the stacking method to perform mortality prediction model construction for HF patients in ICU, which consists of Random Forest (RF), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Light Gradient Boosting Machine (LGBM), Bootstrap aggregating (Bagging), and Adaptive Boosting (AdaBoost).

In this study, we used detailed clinical data from the MIMIC-III database, and focused on the collected data of ICU center patients suffering from heart diseases to predict the mortality of HF patients at 3 days, 30 days, 3 months and 1 year after admission to ICU using ML modeling techniques. This study focuses on proposing a stacking-based model to predict mortality in patients with HF. The base estimators can be adaptively selected and applied to the base layer of the stacking model. In addition, in the selection of key variables, it is expected that the constructed model can also successfully identify the important vital signs indicators of HF patients. We hope to improve the prediction of mortality in patients with heart disease by medical professionals, so that patients, their families and medical professionals are afforded with more informed judgments and make more appropriate prognostic preparations.

2 MATERIALS AND METHODS

2.1 Patient selection and variable selection

The MIMIC-III dataset was used in this study. MIMIC-III integrates comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. The dataset collected de-identified data on 46,520 ICU admissions from 2001 to October 31, 2012 (Johnson et al., 2016). For this study, in order to prevent possible information leakage and ensure a similar experimental setting compared to the related work, we used only the first ICU admission data for each patient. To emphasize the early predictive value, we used data from the first 24 hours of patient admission as input to the predictive model and excluded patients younger than 16 years of age (Gangavarapu et al., 2020; Tang et al., 2021; Yu et al., 2020). A patient cohort was selected based on the following exclusion criteria: The first ICU stay without all subsequent ICU stays, single ICU stay more than 24 hours, patients' first 24 hour of ICU admission, and adult patients (age >= 16). Patients with discharge diagnosis as HF were screened out based on the International Classification of Diseases, 9th Revision codes (ICD-9) 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.xx(W. Q. Guo et al., 2021; Tang et al., 2020). After the first stage of data screening, a total of 7,278 eligible patients were qualified.

The data for the predictive variables selected in this paper were obtained from two tables: admission table and chartevents table in the MIIMIC-III database. As a result of the study, we referred to the predictive variables that have been used by other researchers to predict HF mortality in patients (L. Chen et al., 2021; W. Q. Guo et al., 2021; F. Li et al., 2021) by measuring clinical symptoms 24 hours before admission to the ICU, including Heart Rate, Respiratory Rate, Diastolic Blood Pressure, Systolic Blood Pressure, Temperature,

 ISBN: 978-86-87043-83-1

Oxygen Saturation, Blood Urea Nitrogen, Creatinine, Mean Blood Pressure, Glucose, White Blood Cell, Red Blood Cell, Prothrombin Time, International Normalized Ratio, Platelets, GCS Eye, GCS Motor, GCS Verbal, and Patient's Age and Gender.

In addition, for the treatment of missing values, the data preprocessing method of Guo et al., (C. H. Guo et al., 2020) was adopted in this study to perform a three-stage missing value treatment. Firstly, patients with more than 30% missing values were removed, thus 579 patients were removed. Secondly, variables with more than 40% missing values of the predictor variables were removed. Finally, statistics with missing values greater than 20% for these indicators were removed, and the remaining missing values were interpolated using menus. Finally, a total of 6,699 patients were used in this study.

2.2 Proposed Framework

Mortality rate is a major outcome in acute care: ICU mortality is the highest among hospital units, and early identification of high-risk patients is key to improving outcomes (Harutyunyan et al., 2019). We use ICU admission records of HF patients in the MIMIC-III dataset to predict mortality in HF patients at different time points. We propose an ensemble learning model based on stacking. In the first layer of the stacking model we applied six ML methods, including RF, SVC, KNN, LGBM, Bagging and Adaboost to perform a first-level classifier on the dataset. Then the fusion of these six classifier decisions was used to construct and optimize the stacked set of classifiers. This was followed by subjecting the datasets to second-level classifier using LGBM, and formed the final prediction.

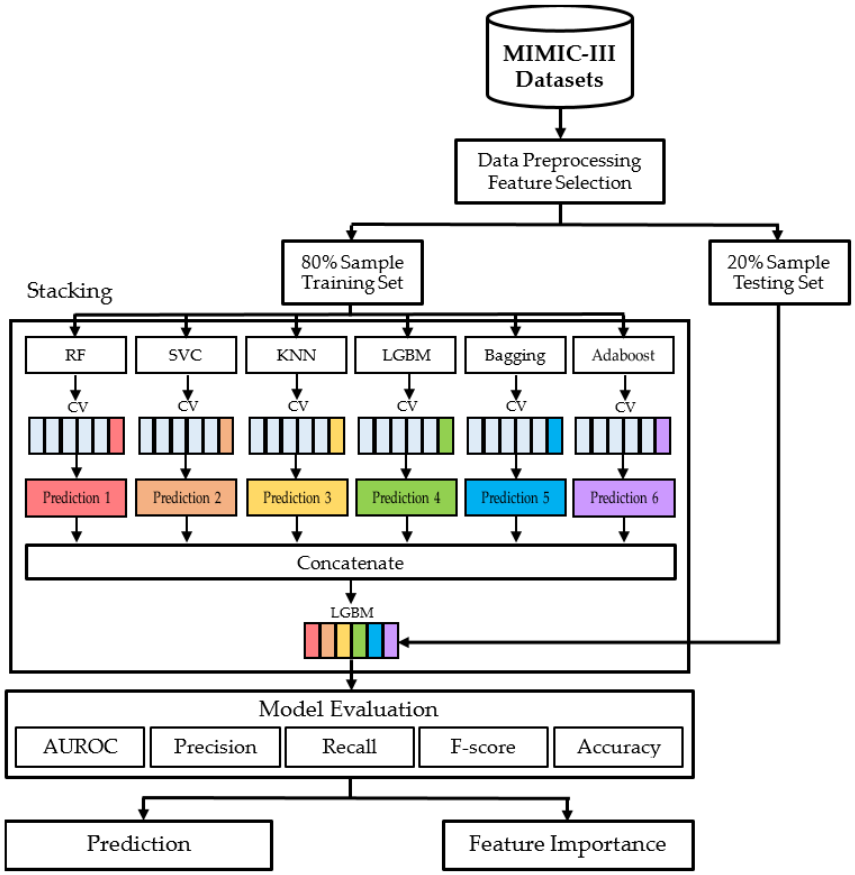

Figure 1. Stacking ensemble based on a cross-validation of all feature subsets.

 ISBN: 978-86-87043-83-1

2.3 Machine Learning

In this study, the most popular and diverse classifiers in related literature were applied in the mortality prediction model for patients with HF in the ICU. Six commonly used ML algorithms include: Random Forest (RF), Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LGBM), Bootstrap aggregating (Bagging), and Adaptive Boosting (AdaBoost).

- Random Forest (RF) is an ensemble supervised ML algorithm. It uses decision trees as the basic classifier. RF generates many classifiers and combines their results by majority voting. The random forest algorithm is well suited to handle datasets with missing values. (K. Li et al., 2021)
- Support Vector Classifier (SVC) performs classification and regression analysis on linear and non-linear data. SVC aims to identify classes by creating decision hyperplanes in a non-linear manner in a higher eigenspace (Nanayakkara et al., 2018).
- The K-Nearest Neighbors (KNN) algorithm is used to predict binary or sequential outputs. The data is divided into clusters and the number of nearest neighbors is specified by declaring the value of "K", a constant. KNN is an algorithm that stores all available instances and classifies new instances based on a similarity measure (K. Li et al., 2021).
- Light Gradient Boosting Machine (LGBM) is an ensemble approach that combines predictions from multiple decision trees to make well-generalized final predictions. LGBM divides the consecutive eigenvalues into K intervals and selects the demarcation points from the K values. This process greatly accelerates the prediction speed and reduces storage space required without degrading the prediction accuracy (Song et al., 2021).
- The Bootstrap aggregating (Bagging) algorithm is an ensemble learning algorithm in the field of machine learning. Bagging algorithm can be combined with other classification and regression algorithms to improve its accuracy, stability, and avoid overfitting by reducing the variance of the results (Ali et al., 2016).
- The self-adaptive nature of the Adaptive Boosting (AdaBoost) method is that the wrong samples of the previous classifier are used to train the next classifier, therefore, the AdaBoost method is sensitive to noisy data and anomalous data. It trains a basic classifier and assigns higher weights to the misclassified samples. After that, it is applied to the next process (Lee et al., 2021).

2.4 Stacking Ensemble Technique

Stacking is an ensemble method for connecting multiple different types of classification models through a meta-classifier (Verma & Pal, 2020). Generally, base learners are called first-level learners and combinators are called second-level learners or meta learners. The basic principle of stacking is as follows. First, the first-level learner is trained using the initial training dataset. Then, the output of the first-level learner is used as the input feature of the meta learner. Finally, a new dataset is formed using the corresponding original labels as new labels to train the meta learner (Cui et al., 2021). As heterogeneous ensembles have better generalization performance and prediction accuracy, this paper proposes a stacked ensemble classifier that can be divided into two stages. First, we use RF, SVC, KNN, LGBM, Bagging and Adaboost as the base classifiers in the first stage. The individual six classification models are trained using the complete training set; then, the probabilistic outputs obtained in the first

 ISBN: 978-86-87043-83-1

stage are fed into the meta-classifier in the second stage, and then the meta-classifier is fitted based on the output meta-features of each classification model using the chosen ensemble techniques. The meta classifier can be trained on the predicted category labels or probabilities from the ensemble technique.

2.5 Synthetic Minority Oversampling Technique (SMOTE)

In ML, the problem is imbalanced when the class distribution is highly skewed. Imbalanced classification problems usually occur in many applications and pose obstacles to traditional learning algorithms (Raghuwanshi & Shukla, 2021). SMOTE is a powerful solution to the classification imbalance problem and has delivered robust results in various domains. The SMOTE algorithm adds synthetic data to a small number of classes to form a balanced dataset (Raghuwanshi & Shukla, 2021). The core of the method is to perform random undersampling and oversampling for larger samples and smaller samples, respectively. In the MIMIC-III used in this study, only a few patients died during their ICU admission. Therefore, the SMOTE method, which uses synthetic minority sampling techniques to preprocess highly imbalanced data sets, was used in this study.

2.6 Evaluation criteria

The performance of the full classification is measured using different evaluation parameters, which consist of binary values (positive and negative). Two general evaluation measures, precision and recall, were used to evaluate the sentiment of tweets based on positive and negative polarity, including Accuracy and F-score for micro averaging purposes. Four functional accuracy measures were taken into account based on the outcomes of the confusion matrix named true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The evaluation parameters used to measure the performance of our proposed system are listed below: we applied Precision, Recall, F-score, and Accuracy, which are widely used in the research field, to evaluate the results of our study, defined as follows:

$$Precision = PPV = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{2}$$

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

In addition, this study adopts the area under the receiver operating characteristic (AUROC) to measure the predictive performance of the model. The area under the receiver operating curve is primarily used to measure the classification threshold performance of classifiers. ROC is a curve consisting of points generated by the true positive rate (TPR) and false positive rate (FPR) of model. TPR signify the probabilities that models can correctly locate positive samples. Such probabilities are commonly referred to as recall rates and represent revenue. By contrast, FPR signify the probabilities that models incorrectly locate positive samples and represent losses. AUROC values range from 0 to 1, where the larger the value, the more superior the result (Song et al., 2021)

 ISBN: 978-86-87043-83-1

3 RESULTS

3.1 Baseline Characteristics

The study was from the MIMIC-III database and ultimately used the ICU admission records of 6,699 HF patients. Table 1 provides demographic information on HF patients. The mean age of the patients in this study was 70.3 years, 55% of whom were male, and the mean number of days hospitalized and the mean number of days in the ICU was 5.8 days.

Table 1. Features involved in the model.

	Overall	Alive at ICU	Dead at ICU
General			
Number	6,699 (100%)	5,754 (85.89%)	945 (14.11%)
Age (Q1-Q3)	70.31±13.04	69.88±13.03	72.92±12.74
Gender (male)	3,694 (55.14%)	3,185 (55.35%)	509 (53.86%)
Outcomes			
Hospital LOS (days) [Q1-Q3]	13.04 [5.99-16.00]	12.78 [6.06-15.77]	14.60 [5.27-19.09]
ICU LOS (days) [Q1-Q3]	5.79 [1.93-6.23]	5.40 [1.88-5.77]	8.17 [2.38-10.38]
Care Unit Type			
CCU	1,852 (27.65%)	1,633 (28.38%)	219 (23.17%)
CSRU	1,319 (19.69)	1,240 (21.55%)	79 (8.36%)
MICU	2,537 (37.87%)	2,049 (35.61%)	488 (51.64%)
SICU	635 (9.48%)	526 (9.14%)	109 (11.34%)
TSICU	356 (5.31%)	306 (5.32%)	50 (5.29%)

MICU Denotes Medical ICU; SICU Denotes Surgical ICU; CCU Denotes Coronary Care Unit; CSRU Denotes Cardiac Surgery Recovery Unit; TSICU Denotes Trauma Surgical ICU

3.2 Mortality prediction results of different models

In this study, a 10-fold cross-validated training and testing was used, with 80% of the data used for training the model and 20% for testing the model, followed by extensive statistical analysis to evaluate performance. Table 2 lists the six different ML methods used in the first phase of this study, and the addition of the Stacking technique in the second phase involved seven techniques to generate the AUROC of HF patient mortality prediction tasks over four different time periods. This study determined the highest AUROC for predicting 3 days mortality in patients with HF, i.e., death within 3 days could be accurately predicted within 24 hours of patient admission. Figure 2 indicates that the data collected from ICU admissions can be used to predict mortality within 3 days, which is a better prediction outcome; Figure 2 also clearly shows that all four models have high AUROC after adding stacking technique in the second stage, which means they can distinguish mortality from non-mortality cases well.

Table 2. The AUROCs of different classifiers.

	RF	SVC	KNN	LGBM	Bagging	Adaboost	Stacking
3 Days	0.7598	0.7249	0.7490	0.7868	0.7534	0.7230	**0.8255**
30 Days	0.7472	0.7179	0.7476	0.7724	0.7442	0.7168	**0.8052**
3 Months	0.7433	0.7002	0.7313	0.7596	0.7338	0.7005	**0.7830**
1 Year	0.6998	0.6671	0.6958	0.7269	0.7014	0.6706	**0.7532**

 ISBN: 978-86-87043-83-1

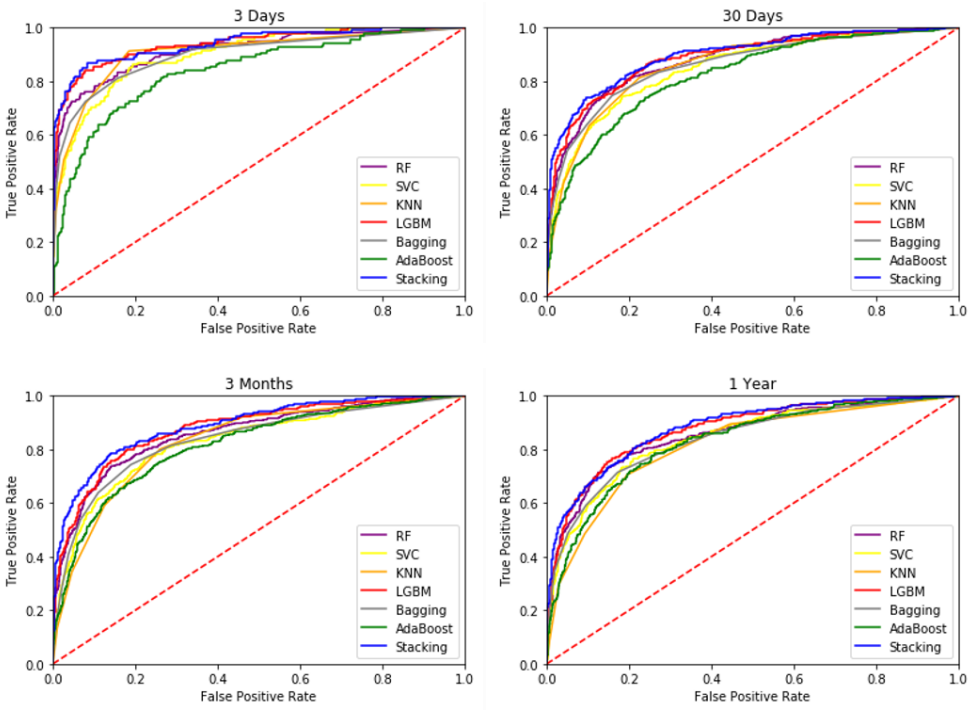

Figure 2 The ROC curves of all base models.

In addition to the above AUROC, this study also compares Precision, Recall, F-score, and Accuracy, as shown in Table 3, when evaluating the performance of different attribute sets used in the classification algorithm. The best precision and recall was shown in a run where the 3 days mortality was used to select the best subset of attributes. In addition, it can be observed from the addition of Stacking technique in the second stage results in a higher accuracy, with the highest value of 0.9525 for predicting mortality within 3 days. It also demonstrated that this study could achieve very good results in predicting mortality in HF patients using data from 24 hours before the patients were admitted to ICU.

Table 3. Diagnostic Precision, Sensitivity, F-Score, and Accuracy of different classifiers.

	Method	Precision	Recall	F-score	Accuracy
3 Days	RF	0.9167	0.3429	0.4989	0.9343
	SVC	0.8991	0.1920	0.3105	0.9208
	KNN	0.6635	0.5198	0.5824	0.9288
	LGBM	0.8763	0.5821	0.6989	0.9425
	Bagging	0.8176	0.3959	0.5322	0.9343
	AdaBoost	0.5937	0.3082	0.4044	0.9139
	Stacking	**0.8030**	**0.6682**	**0.7286**	**0.9525**
30 Days	RF	0.7857	0.5455	0.6435	0.8457
	SVC	0.7389	0.4962	0.5934	0.8262
	KNN	0.6399	0.6145	0.6264	0.8126
	LGBM	0.7704	0.6069	0.6789	0.8533
	Bagging	0.7526	0.5508	0.6355	0.8387
	AdaBoost	0.6827	0.5170	0.5871	0.8142
	Stacking	**0.7831**	**0.6747**	**0.7247**	**0.8690**
3 Months	RF	0.7878	0.5372	0.6385	0.8429
	SVC	0.7577	0.4505	0.5647	0.8206
	KNN	0.6507	0.5692	0.6072	0.8095
	LGBM	0.7712	0.5792	0.6613	0.8466
	Bagging	0.7476	0.5304	0.6200	0.8320

 ISBN: 978-86-87043-83-1

	AdaBoost	0.6844	0.4780	0.5625	0.8079
	Stacking	**0.7720**	**0.6311**	**0.6939**	**0.8564**
1 Year	RF	0.7600	0.4412	0.5577	0.8397
	SVC	0.7490	0.3722	0.4961	0.8267
	KNN	0.6252	0.4767	0.5408	0.8144
	LGBM	0.7405	0.5069	0.6017	0.8460
	Bagging	0.7114	0.4583	0.5569	0.8332
	AdaBoost	0.6558	0.4046	0.5001	0.8147
	Stacking	**0.7428**	**0.5651**	**0.6414**	**0.8551**

3.3 Interpretation of variable importance

Feature importance is the main contribution of each feature to improve the predictive power of the whole model. It provides an intuitive view of the importance of features to see which features have a greater impact on the final model, but it is not possible to determine how the features relate to the final prediction. We can observe that the clinical characteristics of Platelets, Glucose, Blood Urea Nitrogen, Age, Heart Rate, Systolic Blood Pressure, and Diastolic Blood Pressure are important factors influencing the prediction of HF, and previous studies of HF Similar results have been found in previous studies of HF (Barnett et al., 2019; Wallner et al., 2018). However, the effects of GCS eye, GCS motor, GCS verbal, Red Blood Cell, International Normalized Ratio and Gender were not as significant.

In our study, we collect some most important clinical features that contributed most to the model at four different prediction times. The contributions are ranked from most important to least important. We can find that, in addition to considering an HF patient's Platelets, Glucose, and Blood Urea Nitrogen, Systolic Blood Pressure and Diastolic Blood Pressure are also important variables in predicting 3 days, 30 days, and 3 months mortality of HF patients in the ICU. Heart Rate is an important variable to be considered when predicting 1 year mortality in patients with HF in ICU.

4 DISCUSSION

The feasibility of the ML technique for mortality prediction in HF patients has been previously demonstrated. Luo et al., constructed a risk stratification tool using the extreme gradient boosting algorithm to correlate patient clinical characteristics with in-hospital mortality, and this new ML model outperformed traditional risk prediction methods with an AUC of 0.831(Luo et al., 2022). Negassa et al., developed an ensemble model for 30-day post-discharge mortality and used discrimination, range of prediction, Brier index and explained variance as metrics to evaluate model performance, the discrimination achieved by the ensemble model was higher 0.83 (Negassa et al., 2021). The aim of this study is to predict the mortality of ICU patients by ML model from structural vital signs data collected during the ICU stay of HF patients. Through in-depth analysis of the experimental results, the following conclusions were drawn from this study. First, the data collected during the first 24 hours after the admission of HF patients to the ICU were used for modeling analysis, and the ML stacking method was effective and rapid in constructing predictive models. The results indicated that our proposed stacking method has good performance in predicting mortality in HF patients with Accuracy (95.25%) and AUROC (82.55%). Second, this study also found that Platelets, Glucose, and Blood Urea Nitrogen were the clinical characteristics that had the greatest impact on model prediction and were important indicators to be considered in the selection of important variables, which is in line with previous studies on HF (Barnett et al., 2019).

 ISBN: 978-86-87043-83-1

There are some limitations to this study. First, in this study, in order to consider the convenience and completeness of data collection, only ICU databases with the easiest access to complete dynamic patient information were considered. The MIMIC-III data used in this study were obtained from the Beth Israel Deaconess Single Medical Center in Boston, Massachusetts, USA. However, the broader application of our ML model requires further validation in different statistical populations. More comprehensive results and validation may be obtained for patients in other types of medical institutions such as large hospitals or small clinics, or in medical institutions of other scales. Second, the data used in this study were limited to the data collected during the first ICU stay, excluding the records and reports of patients readmitted to the hospital. The collection of data from multiple ICU hospitalizations may provide a comprehensive assessment of time series issues, or may provide more different levels of analysis to patients, healthcare professionals, and patients' families for future evaluation and prognosis.

5 CONCLUSIONS

In our study, we used detailed clinical data from the MIMIC-III database to predict mortality at 3 days, 30 days, 3 months, and 1 year after admission to the ICU for patients with HF, using ML modeling techniques, based on data collected from patients suffering from cardiovascular diseases and being treated in an ICU. We developed mortality prediction models for HF patients admitted to ICU that can be used in ICU monitoring techniques. Our model involves a small number of routinely collected variables that can be easily used in the clinical setting. Despite the recent development of various severity scores and ML models for early mortality prediction, such predictions remain challenging. Compared to existing solutions, this study provides a complement to current clinical decision-making methods by proposing a new stacked ensemble approach to predict mortality in patients with HF in the ICU. The base estimators can be adaptively selected and applied to the base layer of the stacking model. Our stacking model is significantly better than the traditional ML approach in mortality prediction, and can successfully screen out the important clinical features of HF patients. It can empower health care professionals to better predict mortality in HF patients, and provide patients, their families and medical professionals with more information to determine the status of patients and make more appropriate prognosis. For follow-up studies, this study suggests the following recommendations for future studies:

The MIMIC-III data is relatively rich and complete, and this study only modeled and predicted the mortality of patients; subsequent studies can be conducted to evaluate the readmission, length of stay, medication use, and complications of patients with reference to the framework of this study. This type of study can be made more objective and complete if it can be extended to conduct more comprehensive evaluation and analysis. The evolution of variables over time can be collected from patient EHR data in an attempt to obtain better predictive effects. In terms of research methods, future research can attempt different ML methods as well as deep learning methods that have recently been applied to solve time-series data more effectively. For example, long short-term memory or recurrent neural network(Ping et al., 2020) are common deep learning models.

REFERENCES

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. *Med, 2*(6), 642-665. doi:10.1016/j.medj.2021.04.006

Ali, S., Majid, A., Javed, S. G., & Sattar, M. (2016). Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data. *Computers in Biology and Medicine, 73*, 38-46. doi:10.1016/j.compbiomed.2016.04.002

Barnett, O., Horiuchi, Y., Wettersten, N., Murray, P., & Maisel, A. (2019). Blood urea nitrogen and biomarker trajectories in acute heart failure. *European Journal of Heart Failure, 21*, 257-257. Retrieved from <Go to ISI>://WOS:000468990703024

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Chen, L., Yu, H. P., Huang, Y. P., & Jin, H. Y. (2021). ECG Signal-Enabled Automatic Diagnosis Technology of Heart Failure. *Journal of Healthcare Engineering, 2021*, 8. doi:10.1155/2021/5802722

Chen, W., Long, G., Yao, L., & Sheng, Q. Z. (2019). AMRNN: attended multi-task recurrent neural networks for dynamic illness severity prediction. *World Wide Web, 23*(5), 2753-2770. doi:10.1007/s11280-019-00720-x

Cui, S. Z., Qiu, H. X., Wang, S. T., & Wang, Y. Z. (2021). Two-stage stacking heterogeneous ensemble learning method for gasoline octane number loss prediction. *Applied Soft Computing, 113*. doi:10.1016/j.asoc.2021.107989

El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S., & El-Bakry, H. M. (2020). Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model. *IEEE Access, 8*, 133541-133564. doi:10.1109/access.2020.3010556

Gangavarapu, T., Jayasimha, A., Krishnan, G. S., & Kamath, S. S. (2020). Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems, 190*. doi:10.1016/j.knosys.2019.105321

Guo, C. H., Lu, M. L., & Chen, J. F. (2020). An evaluation of time series summary statistics as features for clinical prediction tasks. *Bmc Medical Informatics and Decision Making, 20*(1). doi:10.1186/s12911-020-1063-x

Guo, W. Q., Peng, C. N., Liu, Q., Zhao, L. Y., Guo, W. Y., Chen, X. H., & Li, L. (2021). Association between base excess and mortality in patients with congestive heart failure. *Esc Heart Failure, 8*(1), 250-258. doi:10.1002/ehf2.12939

Haase, N., Plovsing, R., Christensen, S., Poulsen, L. M., Brochner, A. C., Rasmussen, B. S., . . . Perner, A. (2020). Characteristics, interventions, and longer term outcomes of COVID-19 ICU patients in Denmark-A nationwide, observational study. *Acta Anaesthesiologica Scandinavica*. doi:10.1111/aas.13701

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data, 6*. doi:10.1038/s41597-019-0103-9

Incidence, G. B. D. D. I. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016 (vol 390, pg 1211, 2017). *Lancet, 390*(10106), E38-E38. Retrieved from <Go to ISI>://WOS:000413823200006

Jia, R., Lv, Y., Wang, G., Carranza, E., Chen, Y., Wei, C., & Zhang, Z. (2021). A stacking methodology of machine learning for 3D geological modeling with geological-geophysical datasets, Laochang Sn camp, Gejiu (China). *Computers & Geosciences, 151*, 104754.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., . . . Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Sci Data, 3*, 160035. doi:10.1038/sdata.2016.35

Kim, J. Y., Yee, J., Park, T. I., Shin, S. Y., Ha, M. H., & Gwak, H. S. (2021). Risk Scoring System of Mortality and Prediction Model of Hospital Stay for Critically Ill Patients Receiving Parenteral Nutrition. *Healthcare, 9*(7). doi:10.3390/healthcare9070853

Lee, Y. W., Choi, J. W., & Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Computers in Biology and Medicine, 129*. doi:10.1016/j.compbiomed.2020.104151

Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., & Lian, Z. (2021). 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, 11*(7), e044779.

Li, K., Shi, Q. W., Liu, S. R., Xie, Y. L., & Liu, J. L. (2021). Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. *Medicine, 100*(19), 5. doi:10.1097/md.0000000000025813

Luo, C. D., Zhu, Y., Zhu, Z., Li, R. X., Chen, G. Q., & Wang, Z. (2022). A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. *Journal of Translational Medicine, 20*(1). doi:10.1186/s12967-022-03340-8

Marshall, J. C., Bosco, L., Adhikari, N. K., Connolly, B., Diaz, J. V., Dorman, T., . . . Pelosi, P. (2017). What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. *Journal of Critical Care, 37*, 270-276.

Mitchell, T. (1997). Machine learning.

Nanayakkara, S., Fogarty, S., Tremeer, M., Ross, K., Richards, B., Bergmeir, C., . . . Kaye, D. M. (2018). Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *Plos Medicine, 15*(11), 16. doi:10.1371/journal.pmed.1002709

Negassa, A., Ahmed, S., Zolty, R., & Patel, S. R. (2021). Prediction Model Using Machine Learning for Mortality in Patients with Heart Failure. *American Journal of Cardiology, 153*, 86-93. doi:10.1016/j.amjcard.2021.05.044

Ping, Y., Chen, C., Wu, L., Wang, Y., & Shu, M. (2020). *Automatic detection of atrial fibrillation based on CNN-LSTM and shortcut connection.* Paper presented at the Healthcare.

Raghuwanshi, B. S., & Shukla, S. (2021). Classifying imbalanced data using SMOTE based class-specific kernelized ELM. *International Journal of Machine Learning and Cybernetics, 12*(5), 1255-1280. doi:10.1007/s13042-020-01232-1

Song, J. Z., Liu, G. X., Jiang, J. Q., Zhang, P., & Liang, Y. C. (2021). Prediction of Protein-ATP Binding Residues Based on Ensemble of Deep Convolutional Neural Networks and LightGBM Algorithm. *International Journal of Molecular Sciences, 22*(2), 21. doi:10.3390/ijms22020939

Tang, Y. Y., Lin, W. C., Zha, L. H., Zeng, X. F., Zeng, X. M., Li, G. J., . . . Yu, Z. X. (2020). Serum Anion Gap Is Associated with All-Cause Mortality among Critically Ill Patients with Congestive Heart Failure. *Disease Markers, 2020*, 10. doi:10.1155/2020/8833637

Tang, Y. Y., Zeng, X. F., Feng, Y. L., Chen, Q., Liu, Z. H., Luo, H., . . . Yu, Z. X. (2021). Association of Systemic Immune-Inflammation Index With Short-Term Mortality of Congestive Heart Failure: A Retrospective Cohort Study. *Frontiers in Cardiovascular Medicine, 8*, 15. doi:10.3389/fcvm.2021.753133

Verma, A. K., & Pal, S. (2020). Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method. *Applied Biochemistry and Biotechnology, 191*(2), 637-656. doi:10.1007/s12010-019-03222-8

 ISBN: 978-86-87043-83-1

Virani, S. S., Alonso, A., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., . . . Amer Heart Assoc, C. (2020). Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation, 141*(9), E139-E596. doi:10.1161/cir.0000000000000757

Wallner, M., Eaton, D. M., von Lewinski, D., & Sourij, H. (2018). Revisiting the Diabetes-Heart Failure Connection. *Current Diabetes Reports, 18*(12). doi:10.1007/s11892-018-1116-z

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259.

Yu, R. X., Zheng, Y. L., Zhang, R. K., Jiang, Y. Q., & Poon, C. C. Y. (2020). Using a Multi-Task Recurrent Neural Network With Attention Mechanisms to Predict Hospital Mortality of Patients. *Ieee Journal of Biomedical and Health Informatics, 24*(2), 486-492. doi:10.1109/jbhi.2019.2916667