Predicting In-Hospital Mortality with Machine Learning: Insights from the MIMIC III Database

Selina Chen, Sabrina Weng
University of Pennsylvania, Philadelphia, United States of America

I. Introduction

The prevalence of Heart Failure (HF) among adults in the United States ranges from approximately 1.9% to 2.6% across the general population, with incidence rates increasing among older adults, expected to reach 8.5% among those aged 65 to 70 years [1]. According to the Centers for Disease Control and Prevention (CDC), around 6.2 million U.S. adults currently live with heart failure, and in 2018, HF was cited on 379,800 death certificates, representing 13.4% of all deaths [2]. Heart failure represents a significant public health challenge due to its high morbidity and mortality rates.

Despite technological advancements that enhance data collection capabilities, the complex relationships between demographic characteristics, vital signs, and mortality in HF patients remain insufficiently understood. To address this gap, this project utilizes the comprehensive MIMIC-III database to identify predictors of in-hospital mortality, thereby aiding healthcare professionals in early risk assessment and intervention strategies.

Risk factors for heart failure, as noted by the CDC [2], include coronary artery disease, diabetes, high blood pressure, obesity, and valvular heart disease. These conditions exacerbate the risk of developing heart failure. This project is predicated on the understanding that a patient's demographic details and vital signs are critical health status indicators. Existing literature supports the notion that attributes such as age, gender, and vital sign abnormalities significantly influence HF patient outcomes during hospitalization. Our goal is to empirically test specific research hypotheses to validate this understanding and generate insights that could inform more tailored and effective treatment plans for HF patients in the ICU.

The importance of this project is highlighted by its potential to transform the management of patients in Intensive Care Units (ICUs). By developing a dependable machine-learning prediction model, healthcare professionals can gain early access to mortality risk indicators. This advancement enables the implementation of timely and personalized interventions. Moreover, the project is designed to reveal complex relationships between demographic traits, vital signs, and mortality rates—areas that have not been extensively explored in the past.

Leveraging the rich dataset available, this initiative will also investigate the correlations between specific subsets of features and mortality outcomes. The goal is to identify which features most significantly influence mortality, thereby providing deeper insights that can enhance patient care strategies and outcomes.

Consequently, we hypothesize that age, gender, and hypertension are significant predictors of in-hospital mortality, aiming to validate these factors through our research efforts.

II. Material and Methods

2.1 Dataset

Our dataset is derived from the publicly accessible MIMIC-III (Medical Information Mart for Intensive Care III) database. MIMIC-III is a comprehensive database that contains de-identified health-related data for over forty thousand patients who were admitted to critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012 [3]. The database encompasses a wide array of information including demographics, bedside vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and both in-hospital and post-discharge mortality data. Further details on the MIMIC-III database and how to access it are available on the Physionet website (mimic.physionet.org).

The dataset consists of 25 distinct tables such as ADMISSIONS, CALLOUT, CAREGIVERS, PRESCRIPTIONS, SERVICES, TRANSFERS, etc. These tables provide extensive patient data, enabling a wide range of machine learning analyses and predictive modeling.

For our project, we utilized PostgreSQL version 9.6 on Kaggle to compile a dataset comprising the following tables: ADMISSIONS, PATIENTS, ICUSTAYS, D_ICD DIAGNOSIS, DIAGNOSIS_ICD, LABEVENTS, D_LABIEVENTS, CHARTEVENTS, D_ITEMS, NOTEEVENTS, and OUTPUTEVENTS. This curated dataset focuses specifically on ICU patients diagnosed with heart failure, aiming to analyze and predict mortality outcomes based on the gathered patient information.

Data Characteristics:

- **Target Variable**: The "outcome" column is binary, where 0 denotes survival and 1 denotes death.
- **Demographic Variables:** Age is represented as a numerical variable, while gender is categorical with 1 indicating male and 2 indicating female.
- **Vital Signs:** Recorded within the first 24 hours of admission, these are numerical and include heart rate, systolic blood pressure, respiratory rate, diastolic blood pressure, temperature, SPO2, and urine output.
- **Comorbidities:** Identified using ICD-9 Codes, these variables are either binary or numerical and include conditions such as hypertension, atrial fibrillation, deficiency anemias, depression, diabetes, renal failure, CHD without MI, BMI, COPD, and hyperlipemia.

- **Laboratory Variables:** Measured throughout the ICU stay, these include hematocrit, RBC count, PT, INR, MCH, MCHC, MCV, RDW, lymphocytes, bicarbonate, leucocytes, NT-proBNP, pH, glucose, blood potassium, blood sodium, blood calcium, anion gap, chloride, lactic acid, neutrophils, basophils, creatine kinase, creatinine, urea nitrogen, platelets, magnesium ion, PCO2, and ejection fraction (EF).

This comprehensive dataset serves as the foundation for our analytical and predictive modeling efforts, aiming to enhance the management and treatment outcomes of heart failure patients in ICU settings.

2.2 EDA

Our initial phase involved a thorough exploratory data analysis to deepen our understanding of the dataset. We streamlined the dataset by removing unnecessary columns and assessed the structure by examining the shape of the DataFrame. We also evaluated the completeness of the data, identifying the number of NaN values in each column, calculating the percentage of missing values, and noting the data type of each column.

To ensure data integrity, we implemented a function called *drop_na* to eliminate duplicate values within the columns. We further enhanced our analysis by creating a correlation matrix, which provided preliminary insights into the relationships between each variable and our target variable, "outcome."

Originally, we hypothesized that age and gender would significantly influence mortality rates. However, the initial correlation coefficients suggested that these demographic factors did not have a strong correlation with the outcome. Instead, we discovered that certain laboratory values, such as Anion gap, lactic acid, and leukocyte counts, showed the strongest correlations with the outcome.

Subsequently, we employed the *SimpleImputer* to fill in missing values using the mean for numerical data and utilized `LabelEncoder` for encoding categorical data. We conducted a final check to ensure no remaining missing values across the dataset. To further our analysis, we generated additional visualizations to explore the distribution and frequency of each variable, enhancing our understanding of the dataset's characteristics.

2.3 Model Training and Prediction

We initiated our model training process by partitioning our dataset into training and testing sets, utilizing an 80:20 split and setting a random state of 42 to ensure reproducibility.

To identify the most effective model for our dataset, we developed a function named *find_best_model*. This function was employed to train and evaluate eight different machine learning models that we studied during the semester. These models include Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Naive Bayes, AdaBoost, and XGBoost. Our primary objective was to determine which model yields the highest performance based on various evaluation metrics.

We supplemented our approach by reviewing relevant literature to understand the appropriate applications for each model. Logistic Regression is particularly suited for binary classification tasks, making it a strong candidate for our project which involves predicting outcomes from a mix of feature types. Decision Trees offer a versatile solution for classification and regression problems and are highly interpretable, capable of processing both numerical and categorical data. This could be particularly advantageous given the diverse nature of our dataset.

Random Forests are beneficial for mitigating overfitting in Decision Trees, enhancing model accuracy, which could be crucial if our data exhibits overfitting. SVMs are tailored for classification tasks and could be integral for classifying complex patterns in our project data. KNN is adaptable and requires no prior training to make predictions, offering immediate applicability for both classification and regression tasks.

Although Naive Bayes is typically used for text classification, its principles could still be explored in our context to assess its efficacy with non-textual data. AdaBoost is known for its robustness in handling large datasets and complex attribute interactions, which, despite our dataset's size not being large, could still illuminate complex feature relationships. Lastly, XGBoost provides a sophisticated and efficient approach to both regression and classification challenges.

By systematically evaluating these models, we aim to glean deeper insights into which models are most suitable in terms of accuracy, training efficiency, and compatibility with the specific characteristics of our data.

III. Results

As an initial run on all eight machine learning models we discussed previously, Table 1 shows the scores that each model has.

Model	**Score**
Logistic Regression	0.860
Random Forest	0.860
AdaBoost	0.847
SVM	0.839

XGBoost	0.835
Naive Bayes	0.831
KNN	0.826
Decision Tree	0.771

Table 1. Scores showcasing the eight machine learning model performance

From our analysis of the comparative table, we observed that the logistic regression and random forest models exhibit the strongest performance among the eight models evaluated, with each scoring above 0.70. Although the differences in their scores are not substantial, these models stood out as the most effective.

To further assess the capabilities of each model, we will conduct a detailed evaluation using precision, recall, and F1 scores.

3.1 Logistic Regression

For logistic regression, Table 2 shows the distribution of the scores for each metric correspondingly. The model has an overall accuracy of 86.02%. This means that the model accurately predicts whether patients have heart disease or not 86.02% of the time. This is a general indicator of the model's overall effectiveness but does not reveal how well it performs in each category (those with and without heart disease).

	precision	**recall**	**f1-score**	**support**
0.0	0.87	0.97	0.92	198
1.0	0.67	0.26	0.38	38
accuracy	-	-	0.86	236
macro avg	0.77	0.62	0.65	236
weighted avg	0.84	0.86	0.83	236

Table 2. Precision, recall, and f1-score for the logistic regression model

Precision indicates the reliability of the model's positive predictions.

- Class 0 (no heart disease): a precision of 0.87 means that when the model predicts a patient does not have heart disease, it is correct 87% of the time
- Class 1 (has heart disease): a precision of 0.67 suggests that when the model predicts heart disease, it is correct about 67% of the time. This is particularly important in a

medical setting where false positives can lead to unnecessary stress and medical procedures for patients.

Recall reflects the model's ability to detect all relevant cases in each category.

- Class 0: a recall of 0.97 indicates the model is very good at identifying patients who do not have heart disease, catching 97% of those cases.
- Class 1: a recall of 0.26 is low, meaning the model identifies only 26% of all patients who actually have heart disease. This is a critical point because it suggests many patients with heart disease might not be identified, potentially missing crucial early interventions.

F1-score is essential in medical applications where both the accuracy and the completeness of case identification are crucial.

- Class 0: the high F1-score (0.92) for patients without heart disease means the model is efficient at identifying these cases
- Class 1: the low F1-score (0.38) for detecting heart disease indicates a poor performance, with significant implications for clinical practice as it suggests a need for improvement to avoid missing cases of heart disease.

We also plotted the Receiver Operating Characteristic (ROC) curve for this scenario, in which our area under the curve was 0.77. The ROC curve is a graphical representation used to show the diagnostic ability of a binary classifier system as its discrimination threshold is verified. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

An AUC value of 0.77 suggests that the model has good discrimination ability. It is significantly better than random guessing but it is not excellent. In the clinical context, specifically for heart disease prediction, an AUC of 0.77 suggests that the model is relatively effective in identifying patients at risk of heart disease but might misclassify a significant number of cases.

3.2 Random Forest

Similarly, we looked at the metrics for the random forest model. The values obtained are listed in Table 3 below.

	precision	**recall**	**f1-score**	**support**
0.0	0.86	1.00	0.92	198
1.0	1.00	0.13	0.23	38
accuracy	-	-	0.86	236
macro avg	0.93	0.57	0.58	236
weighted avg	0.88	0.86	0.81	236

Table 2. Precision, recall, and f1-score for the logistic regression model

The overall accuracy for the random forest model is also 86.02%. As a summary for the precision, recall, and f1-score for each class:

- Class 0 (no heart disease):
    - a precision of 0.86 means that when the model predicts that a patient does not have the condition, it is correct 86% of the time
    - a recall of 1.00 means that the model identifies 100% of the actual cases of patients without the condition
    - a high f1-score of 0.92 shows an excellent balance between precision and recall for patients without the condition
- Class 1 (with heart disease):
    - a precision of 1.00 means when the model predicts that a patient has the condition, it is correct every time
    - a recall of 0.13 means that the model only correctly identifies 13% of the actual cases of patients with the condition
    - a f1-score of 0.23 shows a poor balance between precision and recall for this class

The ROC curve exhibits a similar pattern as the logistic regression model, with an AUC value of 0.79.

As we see, the model needs to be adjusted to improve its sensitivity (recall) for class 1 (with heart disease) without sacrificing the precision that it currently exhibits. This might involve re-evaluating the features used for prediction, adjusting the decision threshold, or employing different modeling techniques.

3.3 Neural Network

In order to improve our model, we constructed a basic neural network using the Keras library featuring the following architecture:

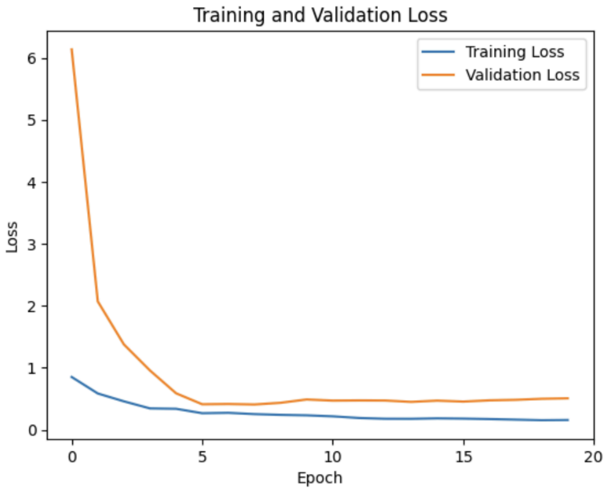
1. BatchNormalization: This layer normalizes the activations from the previous layer for each batch, enhancing the stability and speed of the training process by mitigating internal covariate shifts.

2. Dense Layers (128 neurons): These fully connected layers consist of 128 neurons each, utilizing the ReLU (Rectified Linear Unit) activation function. The ReLU function is favored for introducing non-linearity into the network, which is crucial for complex pattern recognition in data.

3. Dropout (20%): A dropout layer with a rate of 20% is included to combat overfitting. This technique randomly sets a specified fraction of input units to zero during the training phase, which helps make the model robust by preventing it from relying too heavily on any single or small group of neurons.

4. Dense Layer (2 neurons): Another fully connected layer follows, containing 2 neurons and employing a softmax activation function. This function is typically used in multi-class classification tasks to generate a probability distribution across various classes.

The network is compiled using categorical cross-entropy as the loss function, the Adam optimizer for efficient iterative optimization, and accuracy as the metric for performance evaluation. This setup is aimed at efficient processing and learning from the dataset to yield reliable predictive outputs.



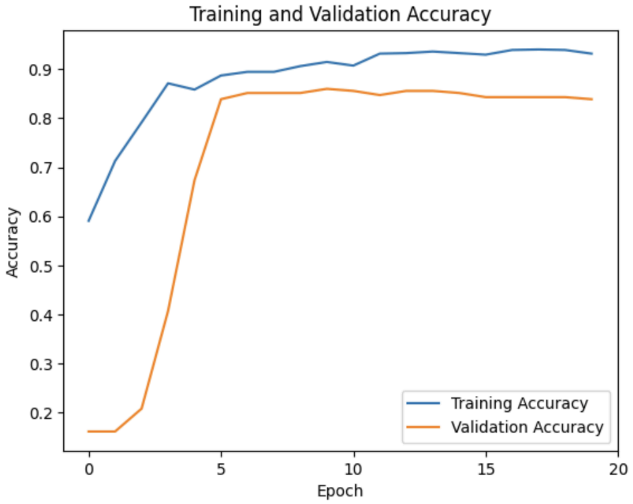


Figure 1. Training and Validation Loss of the NN model **Figure 2.** Training and Validation Accuracy of the NN model

Eventually, we were able to achieve a testing accuracy of 92.37%.

IV. Discussion

5.1 Limitations

Despite the promising results achieved by our models, this study faces several limitations. First, the reliance on data solely from the MIMIC-III database, which includes patients from a single institution, may limit the generalizability of our findings to other populations or healthcare settings. Additionally, the retrospective nature of the study restricts our ability to infer causality from the associations observed. Moreover, while the dataset is rich and comprehensive, the potential for missing or inaccurately recorded data remains, which could impact the reliability of our predictive models.

5.2 Future Recommendations

To enhance the robustness and applicability of our findings, future research should consider several strategies. Expanding the dataset to include multiple centers across different geographic regions could help validate and potentially enhance the generalizability of the model. Employing advanced machine learning techniques such as deep learning might also uncover more complex patterns and interactions that are not readily apparent with the models used in this study. Additionally, integrating real-time data acquisition into the model could facilitate the development of dynamic prediction tools that adjust their predictions as new data becomes available during a patient's ICU stay.

Conclusion

This study has demonstrated the potential of machine learning models to predict in-hospital mortality among heart failure patients using data from the MIMIC-III database. The logistic regression and random forest models showed the most promising performance, highlighting key predictors of mortality such as anion gap, lactic acid levels, and leukocyte counts. These insights could be instrumental in early risk stratification and the tailoring of treatment strategies in the ICU, ultimately aiming to improve patient outcomes. While the models currently exhibit limitations, notably in their sensitivity to detecting patients at the highest risk, the groundwork laid by this research provides a strong foundation for further exploration and refinement of machine learning applications in critical care.

References

[1] *American Heart Association Abstracts to Manuscripts—Barriers to Publication - Journal of Cardiac Failure*, www.onlinejcf.com/article/S1071-9164(17)30492-X/fulltext. Accessed 22 Apr. 2024.

[2] “Heart Failure.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 5 Jan. 2023, www.cdc.gov/heartdisease/heart_failure.htm.

[3] Shahane, Saurabh. “In Hospital Mortality Prediction.” *Kaggle*, 3 Sept. 2021, www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction/data.