Comparative Analysis of Mortality Rate Prediction in Critical Care Patients: Neural Network vs. Classical Machine Learning Models using US Healthcare Data

Yuta Tsukumo

April 26, 2024

Abstract

This study aims to develop predictive models for mortality rates among critical care patients using neural networks, utilizing large-scale healthcare data from the United States. Additionally, we conduct comparative analysis with classical machine learning techniques. While acknowledging the limitations of generalizing findings to other healthcare systems, we emphasize the significance of accurately predicting mortality rates for informed resource allocation and quality assessment in intensive care settings.

Personal Motivation

Around the world, deep learning is widely applied in medical contexts, such as image diagnosis and interoperate image recognition. However, there is currently limited evidence of deep learning analysis being utilized for national healthcare policy in Japan. By conducting a comparative analysis between deep learning and traditional machine learning methods, I seek to understand their respective strengths and weaknesses. My goal is to contribute to the advancement of data science in healthcare in Japan.

Introduction

Predicting mortality rates for critically ill patients requiring intensive care is essential for making informed decisions regarding allocation of limited medical staff, medical equipment prioritization, assessing quality of treatment facilities, and appropriately classifying severity of illness for clinical research. However, traditional severity assessment standards in critical care, such as Apache scores, are widely used but have been reported to have low calibration for predicting mortality rates. Therefore, creating a model with higher discrimination and calibration using neural network techniques, based on large-scale patient data

from the United States, holds significant importance.

Related Work

In 2022, a study by Jesse et al. at MIT proposed a novel scoring system using The Global Open Source Severity of Illness Score (GOSSIS) data to predict mortality rates for patients requiring intensive care. While the specific code details were not provided in the paper, the study utilized a logistic regression model based on generalized additive mixed models (GAM) to ensure interpretability and explanatory power. In our project, there's potential to achieve higher discrimination and calibration in predicting mortality rates by leveraging neural networks.

Aim

The objective of this study is to create a mortality prediction model for intensive care unit admissions using traditional machine learning methods and deep learning.

Data Sauce

- The GLOBAL OPEN SOURCE SEVERITY OF ILLNESS SCORE (GOSSIS) Consortium https://gossis.mit.edu/
- Kaggle https://www.kaggle.com/datasets/mitishaagarwal/patie

Data Processing

Data includes 91,714 rows (patients) and 85 columns (31.4MB). The columns encompassed various domains, including patient demographics (age, gender, race, etc.), hospitalization details (ICU type, elective surgery, etc.), medical condition (Apache scores, blood pressure, etc.), and comorbidity information (diabetes, immunodeficiency, etc.). The variable 'Unnamed: 83' was removed because it did not contain any values.' After removing 'Unnamed: 83', the missing values accounted for 2.55% of the total data (all cells), while the sample with any missing value accounted for 37.92% of the total sample rows. This result suggests that the use of complete data, in which all samples with missing values are removed, is not desirable in model creation because it results in the loss of a large number of samples. Therefore, the missing values were imputed, since it was deemed acceptable to consider missing values as occurring randomly, the mean of the continuous variable was used to compensate for the missing values. For categorical variables for which the class was initially assigned an object, the class was corrected to be a categorical variable. Before model

creation, rows with missing values for categorical variables were removed, and these data processes resulted in a sample size of 89,488 from the original sample size of 91,714. The dataset was partitioned into training data (70%) and test data (30%).

Exploratory Data Analysis (EDA)

First, the main outcome, in-hospital deaths, accounted for 8% of all patients. Next, for each continuous variable, we created a histogram identified by the binary outcome variable, in-hospital mortality, as well as a bar chart identified by the in-hospital mortality variable for each categorical variable. These distributions showed a trend toward older age in the group of in-hospital deaths and a trend toward fewer in-hospital deaths in the group of elective surgeries. On the other hand, mortality appears to be higher in patients with tubes inserted for respiratory management, with liver failure, uncompromising condition, and with metastatic solid tumors. The graphs did not clearly show any significant differences between the deaths or survivors concerning gender, race, or ICU type.

Evaluation

Due to the imbalance in binary outcomes, the models were evaluated using Receiver Operating Characteristic Area Under the Curve (ROC AUC) and Matthews Correlation Coefficient (MCC) instead of Accuracy. ROC AUC is the most popular metric for evaluating binary outcome models, but it doesn't provide information about precision and negative predictive value. On the other hand, high value of MCC always corresponds to high values for each of the four basic rates: sensitivity, specificity, precision, and negative predictive value.

Method

Classical predictive models such as logistic regression, random forest, gradient boosting, and the generalized additive model (GAM) have been considered. Additionally, three neural network models were fitted with varying numbers of hidden layers (1, 2, or 5). The optimizer used was Adam, and the loss function employed was binary cross-entropy loss. The weight was kept constant across all models; the ratio of 12 (death) to 1 (survival). It is noted that hyperparameter tuning was not conducted for the learning rate and weight decay, given the use of the Adam optimizer. Hyperparameter tuning, performed using 5-fold cross-validation, focused on optimizing the dimensions of each hidden layer, dropout proportion, batch size, and number of epochs to enhance model performance. The models were trained on 70% of the data and evaluated for model performance on the remaining 30% test dataset.

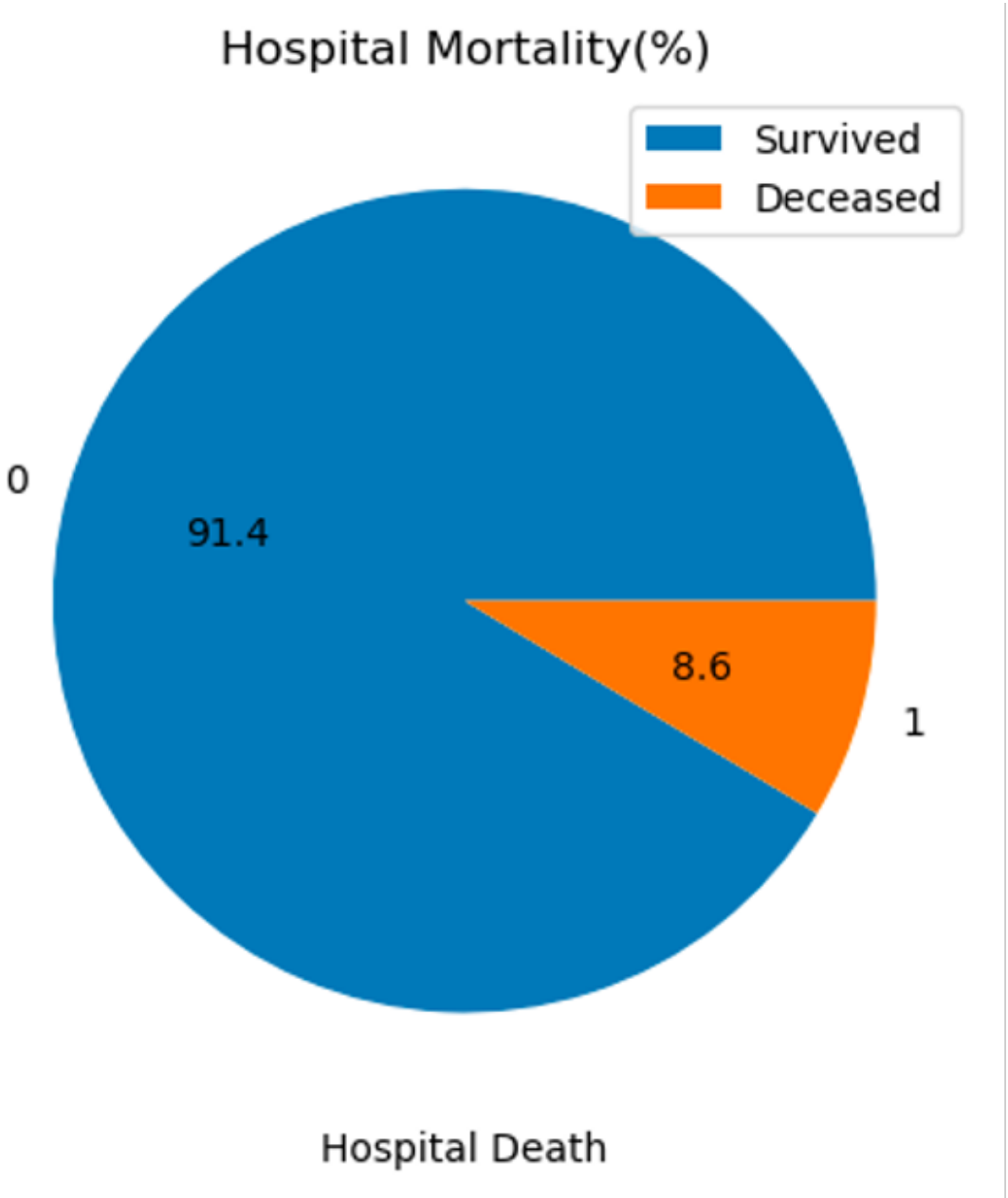

Figure 1: Hospital Mortality

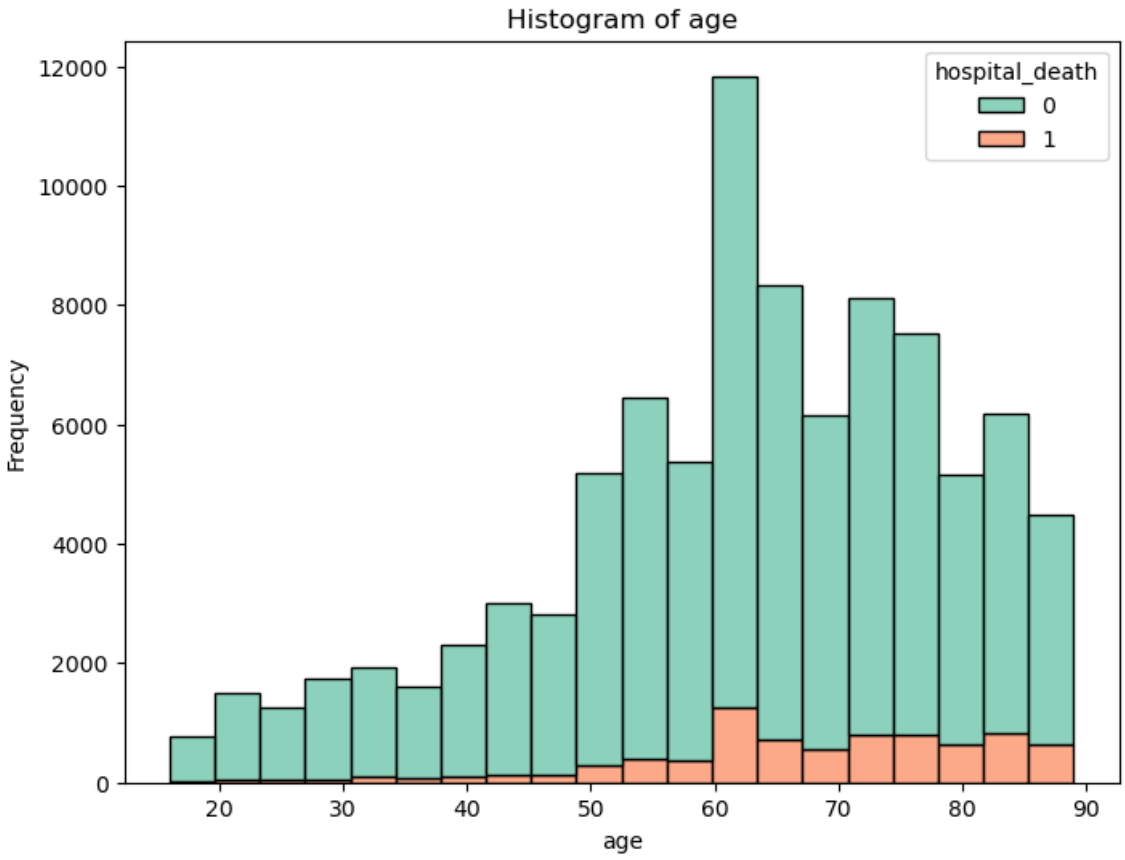

Figure 2: Mortality Distribution by Age

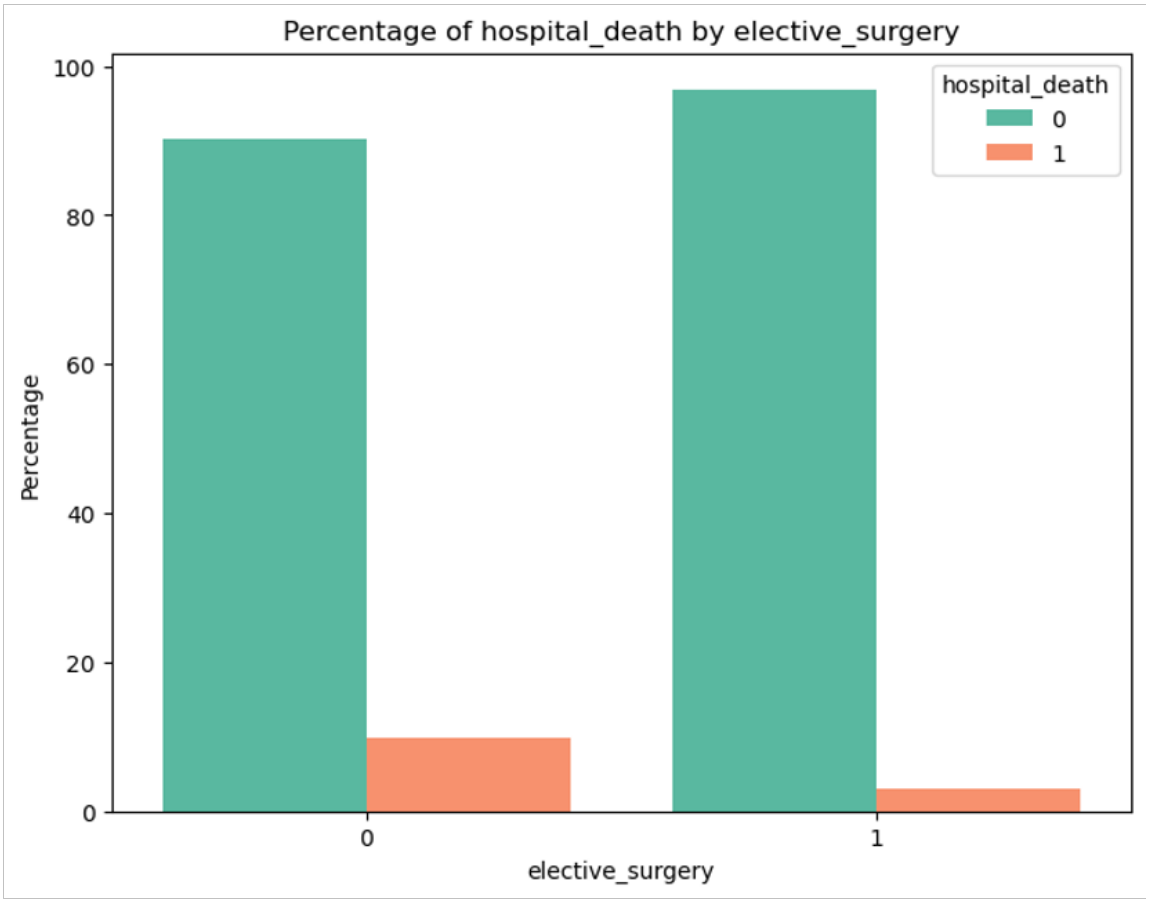

Figure 3: Percentage Comparison: Hospital Deaths with vs. without Elective Surgery

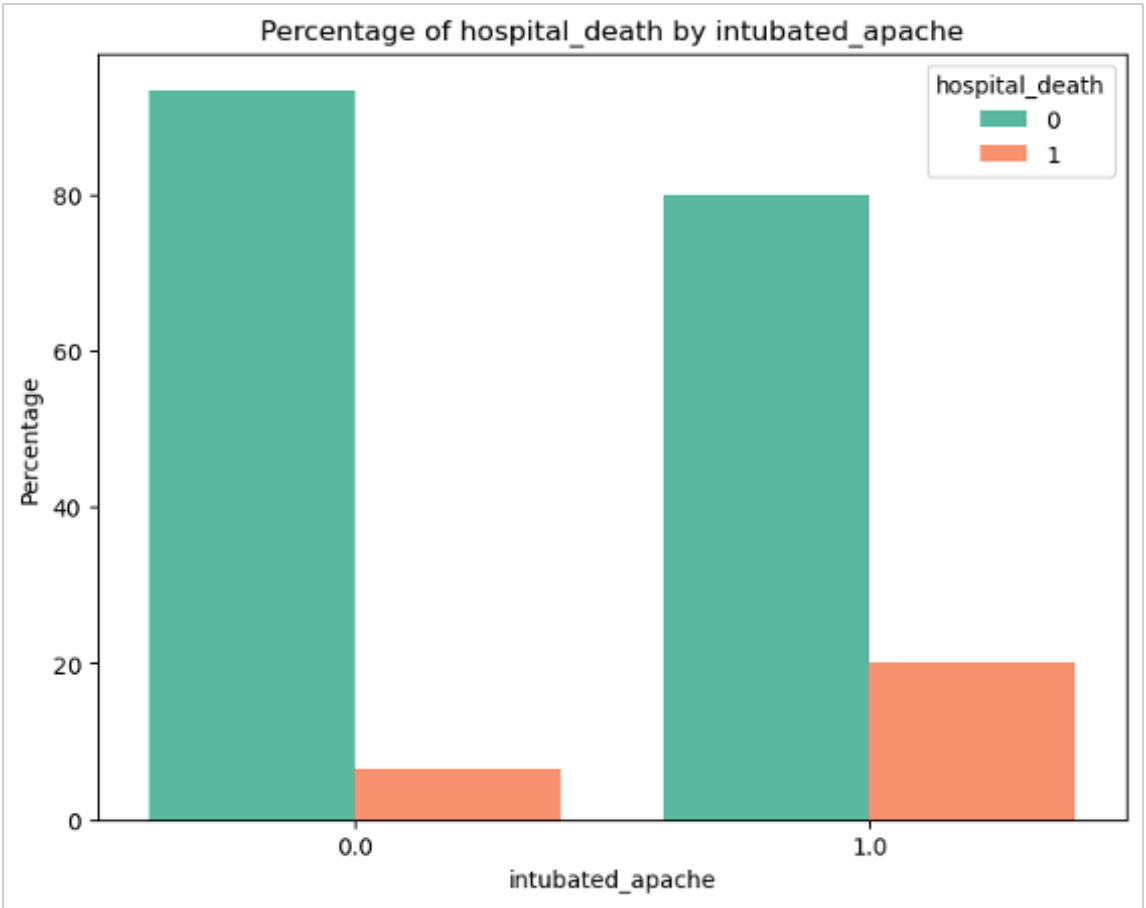

Figure 4: Percentage Comparison: Hospital Deaths with vs. without Intubation

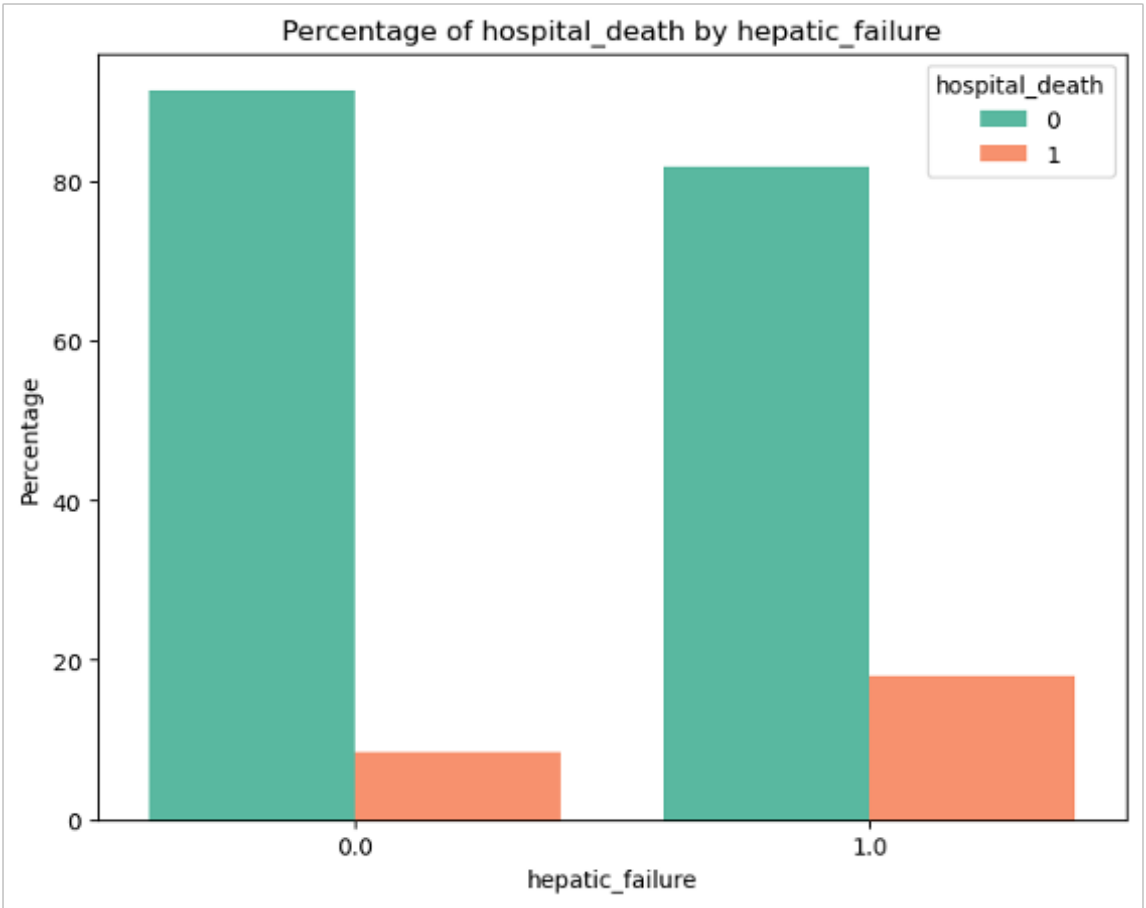

Figure 5: Percentage Comparison: Hospital Deaths with vs. without Hepatic Failure

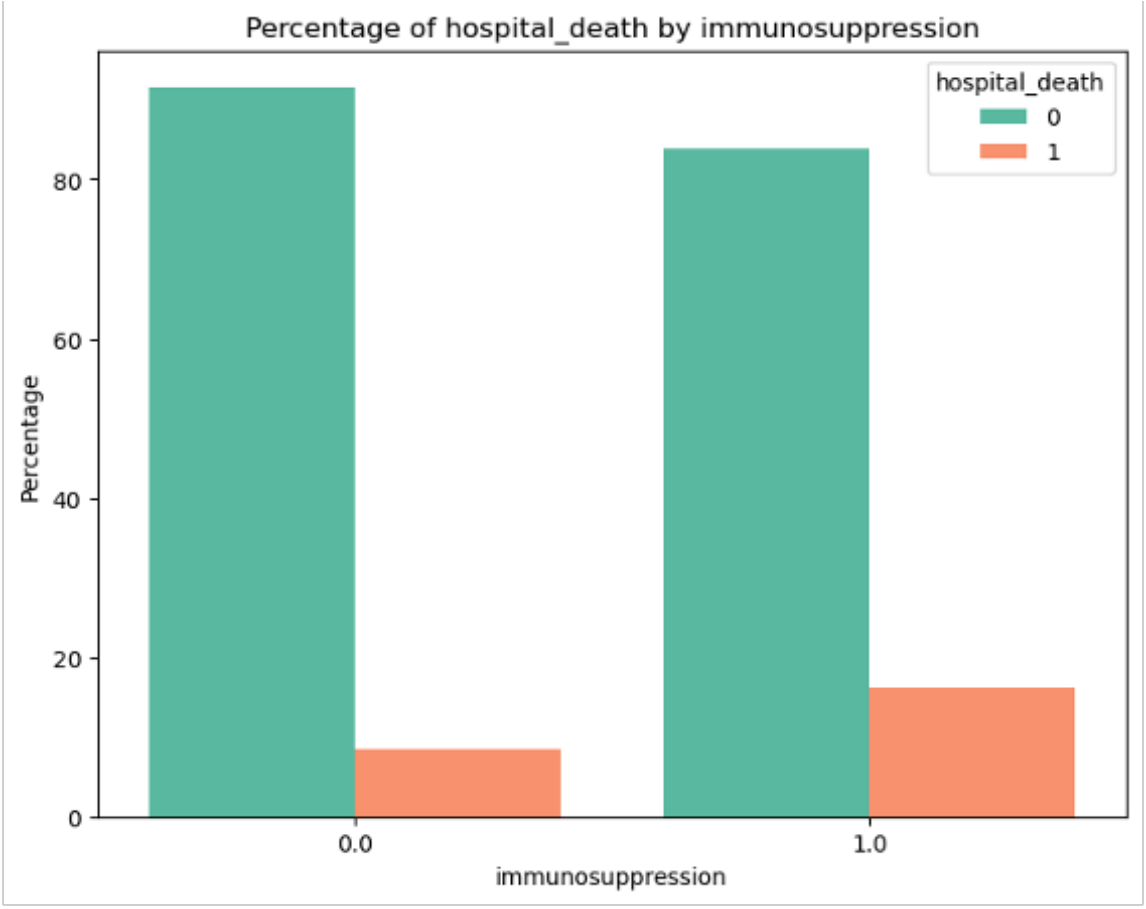

Figure 6: Percentage Comparison: Hospital Deaths with vs. without Immunosuppression

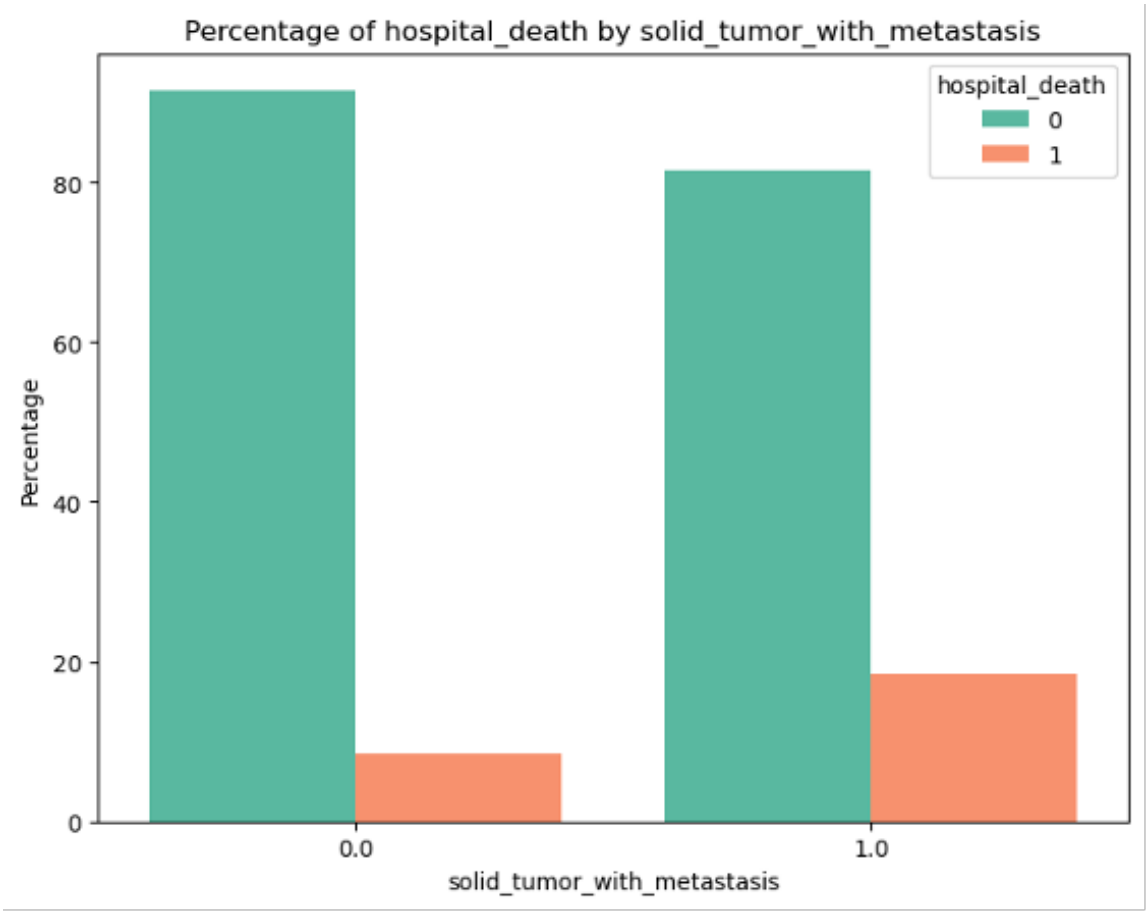

Figure 7: Percentage Comparison: Hospital Deaths with vs. without Solid Tumor Metastasis

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

(worst and minimum value -1; best and maximum value +1)

Figure 8: Matthews Correlation Coefficient (MCC)

Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* 2023 Feb 17;16(1):4. DOI: 10.1186/s13040-023-00322-4. PMID: 36800973; PMCID: PMC9938573.

Model Creation

Based on the EDA results and medical knowledge, 29 variables were selected from a total of 85 columns as follows:

- Age
- BMI
- Days Before ICU Admission
- GCS scores
- Heart Rate
- Arterial Pressure
- Respiratory Rate
- Elective Surgery
- Ethnicity
- Gender
- Source of ICU Admission
- Types of ICU
- Post-Operative
- Atrial Fibrillation
- Intubated
- Ventilated

- AIDS
- Cirrhosis
- Diabetes
- Hepatic Failure
- Immunosuppression
- Leukemia
- Lymphoma
- Solid Tumor with Metastasis

Result

The results are shown in the following table (Figure 9). For Shallow Networks and two Deep Networks, the number of Nodes in each hidden layer, as well as the Dropout probability, Batch Size, and Epochs tuned by cross-validation, are also shown in the table.
The results showed that Gradient Boosting, GAM had the highest ROC AUC of 0.856 and Shallow Network had the highest MCC of 0.341.

Model	ROC AUC	MCC	(Accuracy)
Logistic Regression	0.836	0.275	0.919
Random Forests	0.849	0.291	0.922
Gradient Boosting	**0.856**	0.288	0.922
GAM (Generalized Additive Model)	**0.856**	0.328	0.922
Support Vector Machine	0.708	0.203	0.9182
Shallow Networks (Hidden Layer = 1) Dimensionality: Layer 1 = 16), Dropout Prob = 0.3, Batch Size = 32, Epochs =20	0.830	**0.341**	0.808
Deep Networks (Hidden Layer = 2) Dimensionality: Layer 1 = 32, Layer 2 = 16 Dropout Prob = 0.3, Batch Size = 32, Epochs =20	0.819	0.339	0.816
Deep Networks (Hidden Layer = 5) Dimensionality: Layer 1 = 64, Layer 2 = 64, Layer 3 = 32, Layer 4 = 16, Layer 5 = 8) Dropout Prob = 0.3, Batch Size = 16, Epochs =5	0.789	0.294	0.737

Figure 9: Model Performance Comparison

Discussion / Conclusion

Among neural networks, models with shallower hidden layers demonstrated superior performance. While deep learning is increasingly utilized in medical applications, especially in diagnostic imaging, its superiority over classical machine learning models may not always be evident in simpler analyses like mortality prediction.
However, the hyperparameters in this study were not exhaustively analyzed due to time and computational constraints. Better performance with deep learning may be achievable through more sophisticated tuning.
From a technical perspective, our initial attempt to fit a model using the Apache score included in the data did not yield good model fit. However, when a model was created that included some components of the Apache score, such as vital signs, as variables instead of using Apache score itself, the model performance improved. This enhancement is attributed to the increased flexibility of the model with the addition of more variables.
Furthermore, in terms of optimization, we compared SGD (Stochastic Gradient Descent) and Adam. As reported in various studies, Adam required less time for model convergence and exhibited better overall model performance compared to SGD.
Although the model's performance in this study was evaluated as favorable based on AUC, it did not perform well according to MCC. This discrepancy may be attributed to the unbalanced data.

Github repo link

https://github.com/2023DS598/Project

References

[1] Raffa, J. D., Johnson, A. E. W., O'Brien, Z., Pollard, T. J., Mark, R. G., Celi, L. A., Pilcher, D., & Badawi, O. (2022). The Global Open Source Severity of Illness Score (GOSSIS). Critical care medicine, 50(7), 1040–1050. https://doi.org/10.1097/CCM.0000000000005518

[2] Nassar, A. P., Jr, Mocelin, A. O., Nunes, A. L., Giannini, F. P., Brauer, L., Andrade, F. M., & Dias, C. A. (2012). Caution when using prognostic models: a prospective comparison of 3 recent prognostic models. Journal of critical care, 27(4), 423.e1–423.e4237. https://doi.org/10.1016/j.jcrc.2011.08.016

[3] Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min, 16(1), 4. https://doi.org/10.1186/s13040-023-00322-4