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

Yuta Tsukumo, M.D., MPH

Candidate for a Master of Science in Statistical Practice

Personal Motivation

- Deep learning is applied in medical contexts like image diagnosis and intraoperative image recognition.

- However, so far there are few examples of evidence from deep learning analysis being used for national healthcare policy in Japan.

- By comparing deep learning with traditional machine learning, I aim to grasp their respective strengths and weaknesses, aspiring to contribute to the advancement of the data science in healthcare in Japan.

Study Background

Predicting mortality rates for critically ill patients requiring intensive care is essential for

- ✓ Allocation of limited medical staff / equipment
- ✓ Assessing quality of treatment facilities
- ✓ Appropriately classifying severity for clinical research.

Aim

To create a mortality prediction model for intensive care unit admissions using traditional machine learning methods and deep learning.

Data

✓ Data Source:

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

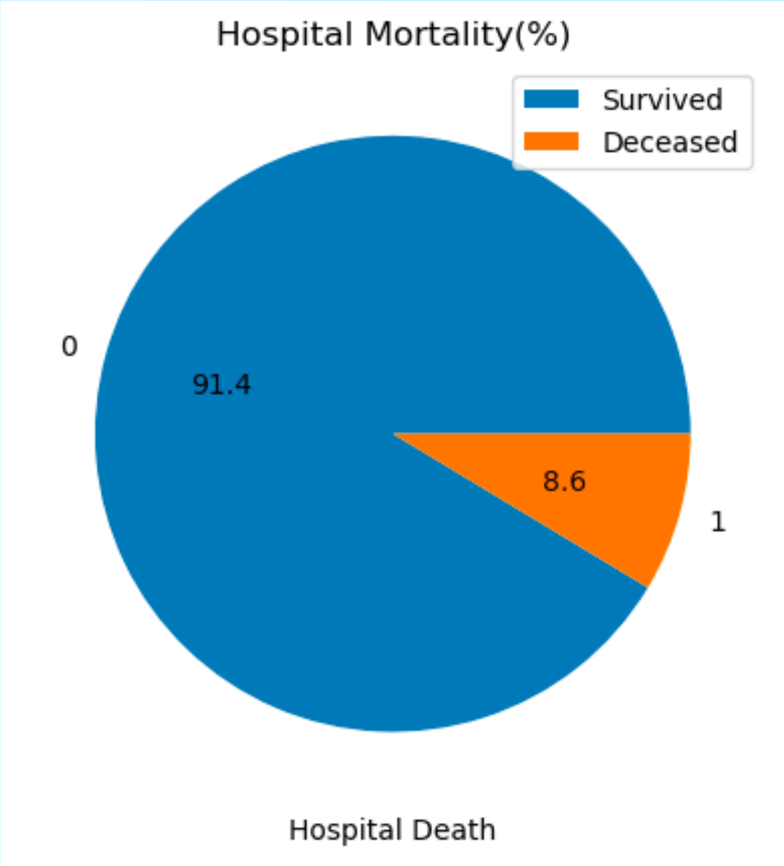
✓ 91,714 rows (patients) and 85 columns (31.4MB)

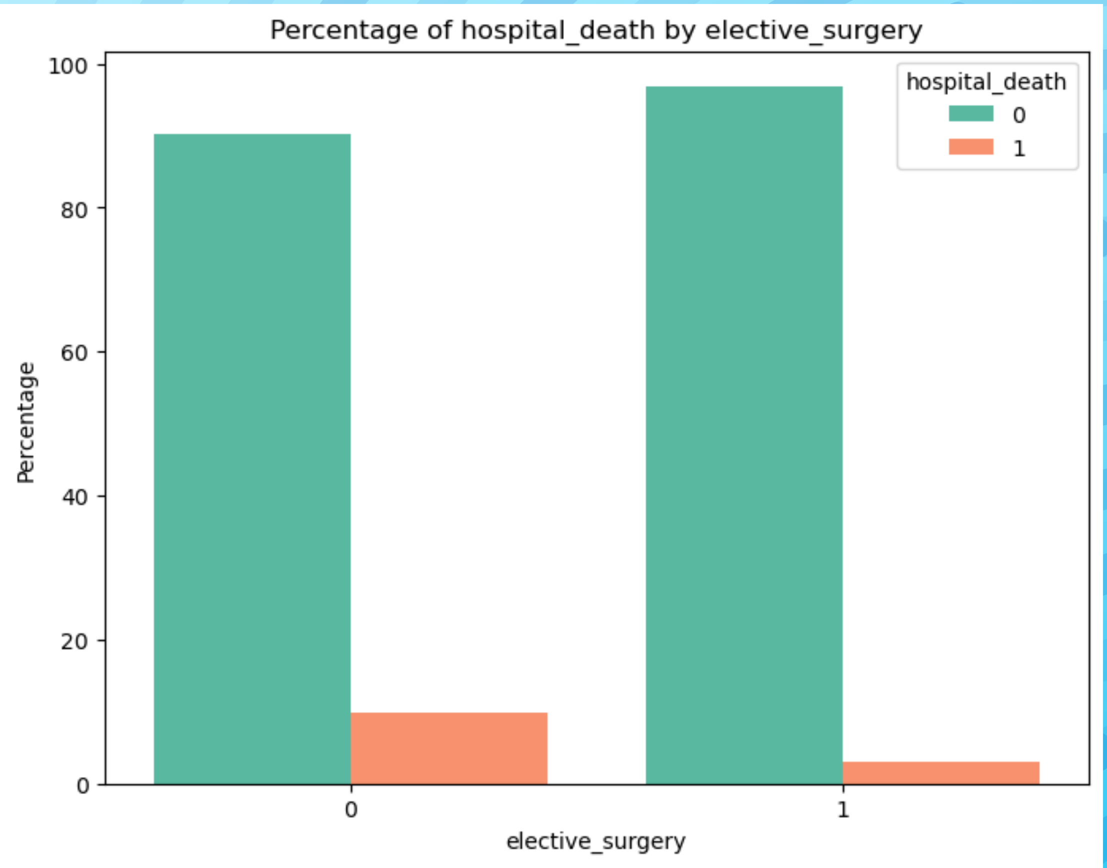
✓ Variables

- Patient demography (Age, Gender, Race, BMI, etc.)
- Hospitalization (ICU type, Elective surgery, etc.)
- Medical condition (Apache scores, blood pressure, Heart Rate, Respiratory Rate, etc.)
- Comorbidity (Diabetes, Immunodeficiency, etc.)

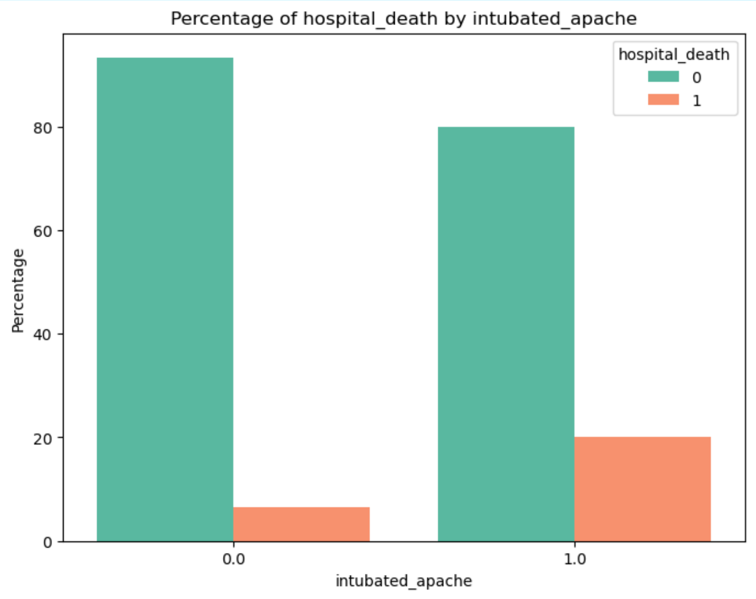
✓ Data Partitioning
Training data 70%, Test data 30%

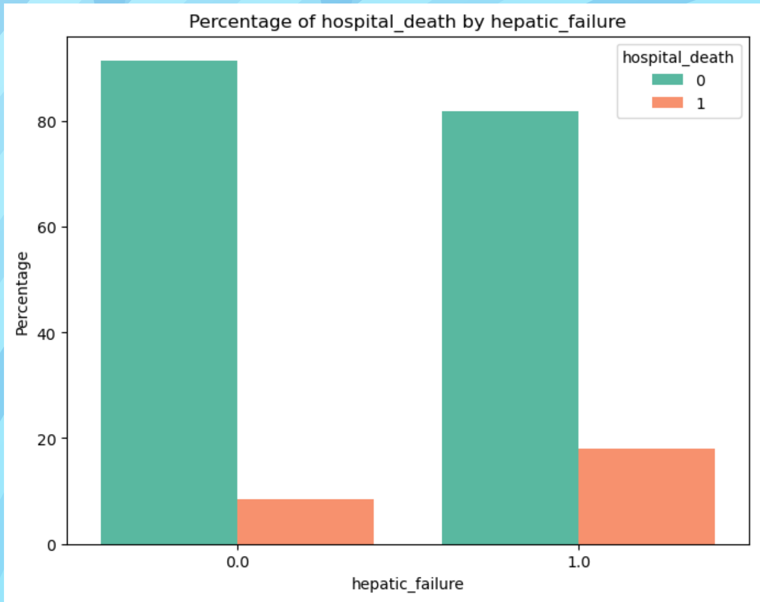
EDA (excerpts 1)

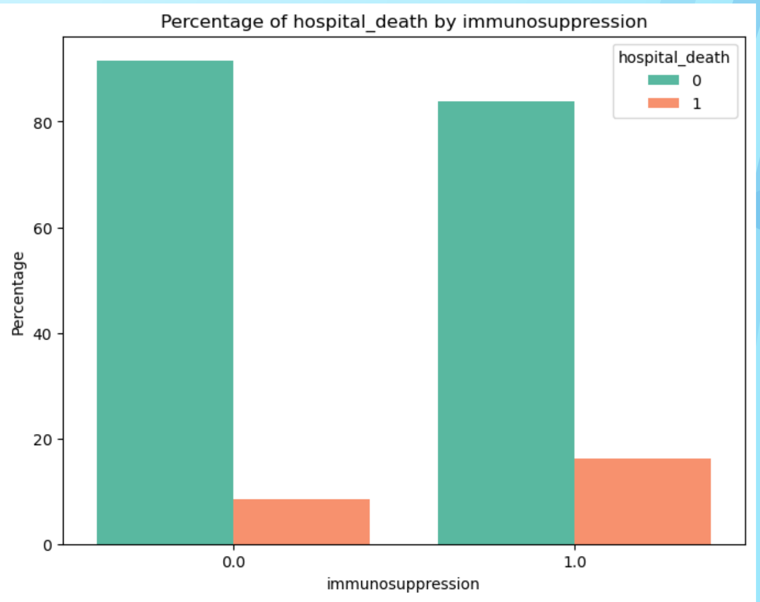


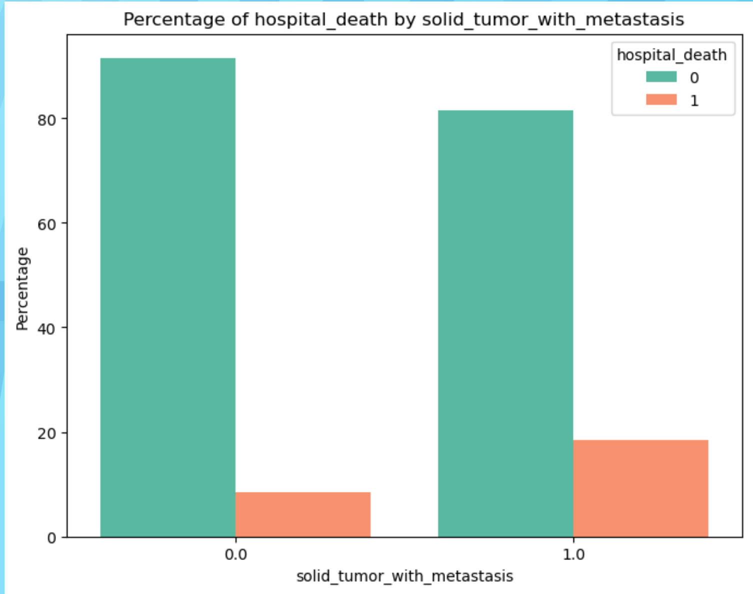


EDA (excerpts 2)









Methods

- 29 variables from 85 columns were selected as inputs based on EDA results and medical knowledge.

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*

- Missing values of continuous/ integers variables were imputed by means. Rows with missing values for categorical variables were excluded in the analysis.

- Because of the imbalance in binary outcomes, ROC AUC and Matthews Correlation Coefficient (MCC) were used instead of Accuracy to compare models.

Matthew Correlation Coefficient (MCC)

- ✓ ROC AUC is the most popular metric for evaluating binary outcome models, but it doesn't provide information about precision and negative predictive value.

- ✓ <u>High value of MCC always corresponds to high values for each of the four *basic rates*</u>: sensitivity, specificity, precision, and negative predictive value

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

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

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.

Neural Network

Fitted with Three Model (Hidden layer = 1, 2 or 5)

Input layer: 62 Nodes
Optimizer: Adam
Loss Function: Binary Cross-Entropy Loss

Fixed Hyperparameters
- Weight = 12 (Death) : 1 (Survive)
- Learning Rate = 0.001
- Weight Decay = 0.001

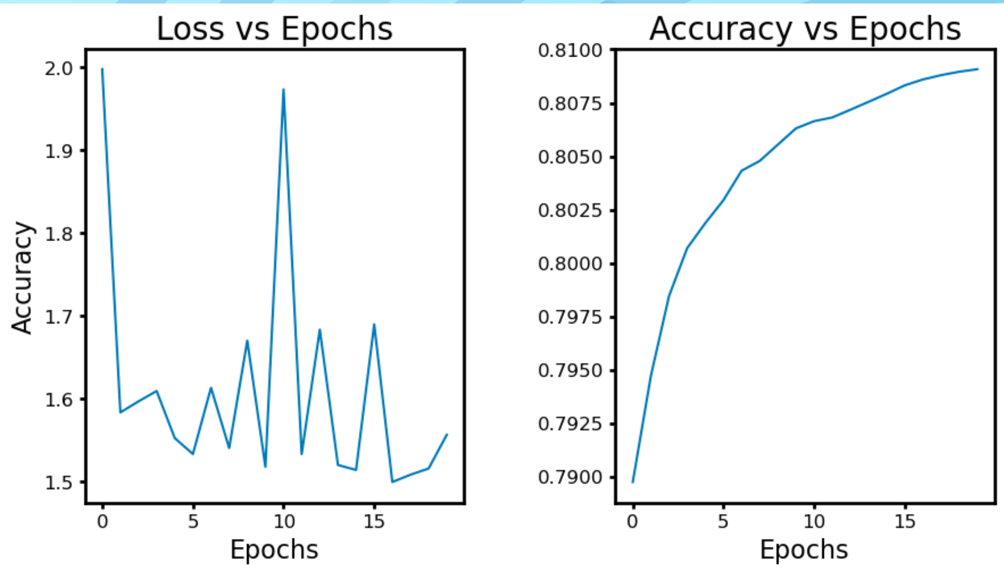
Hyperparameter Tuning (5- fold Cross Validation)
- Dimensionality of Each Hidden Layer
- Dropout Proportion
- Batch Size
- Epochs

Result

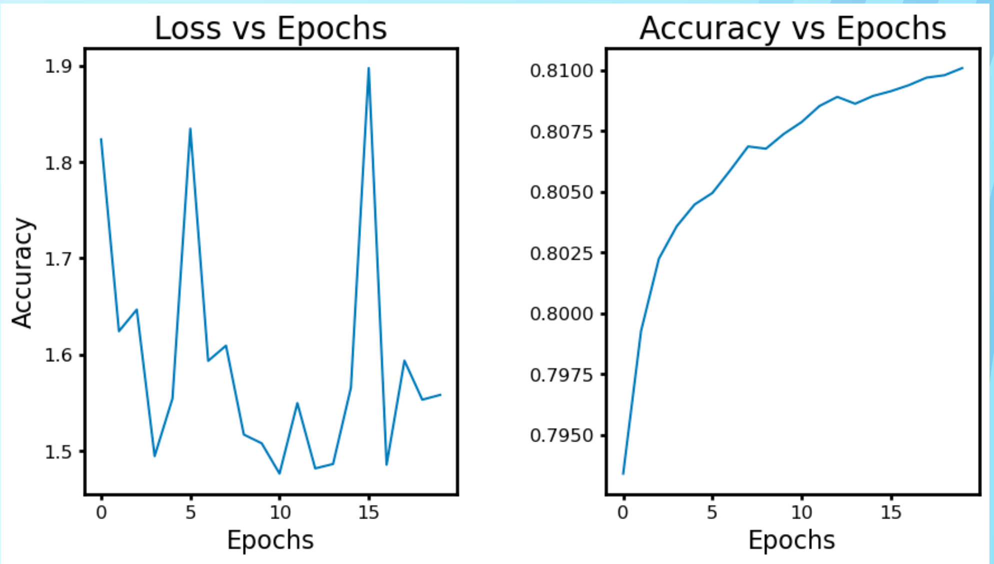
Model	ROC AUC	MCC	(Accuracy)
Logistic Regression	0.836	0.275	0.919
Random Forests	0.849	0.291	0.922
Gradient Boosting	0.856	0.288	0.922
GAM (Generalized Additive Model)	**0.856**	0.328	0.922
Support Vector Machine	0.708	0.203	0.9182
Shallow Networks (Hidden Layer = 1) Dimensionality: Layer 1 = 16), Dropout 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

Learning Curve

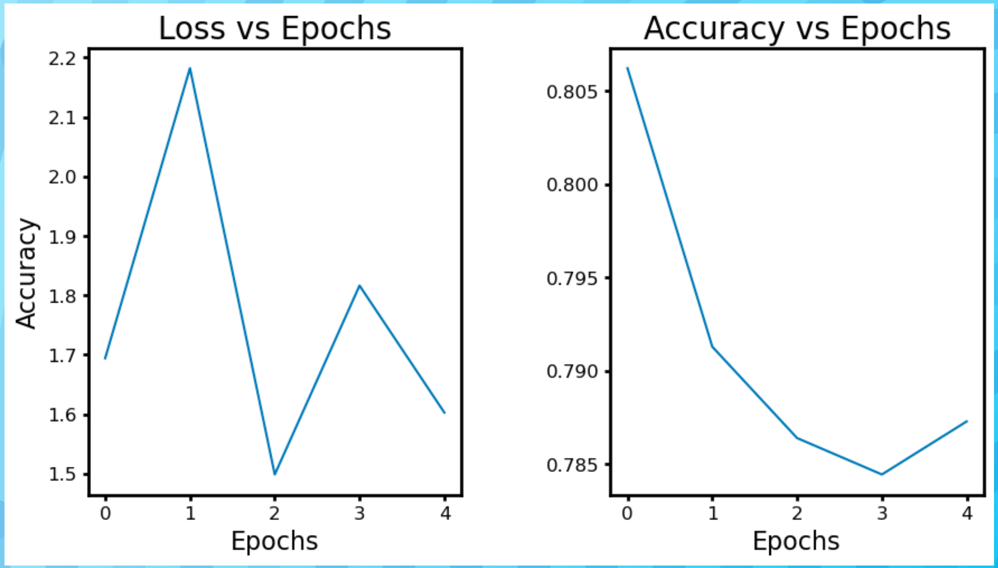
Shallow Networks



Deep Networks (Layers = 2)



Deep Networks (Layers = 5)



Summary (Results)

1. The results of the comparison between classical machine learning models and deep learning showed that the Generalized Additive Model (GAM) and Shallow Networks performed the best.

2. Among neural networks, models with shallower hidden layers demonstrated superior performance.

Discussion

1. 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.

2. 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.

3. The model's performance in this study was considered favorable based on AUC; however, it did not perform well according to MCC, possibly due to the unbalanced data.

Questions?