ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF STATISTICAL SCIENCES

"PAOLO FORTUNATI"
First Cycle Degree / Bachelor in Statistical Sciences

Comparative Analysis of Machine Learning Algorithms for Clinical Outcome Prediction

Presented by:

Matteo Omizzolo
0001021097

Supervisor:

Prof. Angela Montanari

Session I
2023/2024

Abstract

Predictions made by data-driven machines are increasingly influencing clinical decision-making in healthcare. Various Machine Learning (ML) applications have appeared in recent clinical literature, one of the most relevant being outcome prediction models. This study evaluates the performance of different ML algorithms in predicting in-hospital mortality among Intensive Care Unit (ICU) patients based solely on data collected within the first 24 hours after the first admission to the ICU. The work follows state-of-the-art methodologies in related studies, encompassing data processing, feature selection, inference, and model evaluation. All data are retrieved from Electronic Health Records (EHRs), specifically the Medical Information Mart for Intensive Care (MIMIC-IV) database. The primary focus of this study is on statistical analysis of big data and the exploration of ML applications in clinical settings. Conducted independently of clinical consultation, this research should not be considered from a clinical perspective, but rather as an opportunity for me to further explore different ML applications.

Contents

Chapter 1

Introduction

Recent systematic reviews [1][2] highlight a significant change in clinical decision-making due to the growing acceptance of Machine Learning efficacy in these settings. In particular, outcome prediction models have gained considerable importance in forecasting and improving therapy measures, including essential indications for critical care. The increasing availability of extensive patient data from Electronic Health Records (EHRs) offers the ideal conditions to provide improved prediction of outcomes. Despite its potential for ML applications, modeling EHRs data is complex. The intricate nature of clinical data poses substantial difficulties, including its high dimensionality, noise, heterogeneity, incompleteness, random errors, and systematic biases. In addition, different codes and terminologies can report the same clinical condition, complicating interpretation. These challenges affect the ability of ML algorithms to discern patterns necessary for developing predictive clinical models applicable to real-life situations. Hence, it is crucial to employ effective data processing and feature selection techniques to improve the performance of predictive algorithms.

The clinical outcome this study aims to predict is in-hospital mortality for patients in the Intensive Care Unit (ICU) using data obtained within 24 hours from first admission. The predicted results give scope to compare the performances of different ML algorithms, including K-Nearest Neighbors (KNN), Logistic Regression (LogR), Random Forest (RF), Support Vector Machines (SVM), and Neural Networks (MLP). The algorithms' performances are evaluated using various metrics, including accuracy, precision, recall, f1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

The retrospective study included patients from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, from which demographic data, comorbidity information, vital signs, laboratory test results, clinical scores and treatment details were extracted. The overall analysis involved a comprehensive methodological framework, incorporating data preprocessing, inference analysis, and model evaluation. The goal of the thesis is to compare the performance of the selected ML models to find the most efficient ones at predicting in-hospital mortality for ICU patients.

Chapter 2

Methodology

2.1 Database

All data in this study were obtained from the Medical Information Mart for Intensive Care (MIMIC-IV) database [3], an internationally freely accessible resource approved by the Institutional Review Board of the Massachusetts Institute of Technology. The MIMIC-IV database contains information for more than 40,000 patients admitted to the intensive care unit of Beth Israel Deaconess Medical Center between 2001 and 2019.

2.2 Data Extraction

All data were extracted using Structured Query Language (SQL). The extracted data included:

1. Demographics and Admission Details;
2. Clinical Measurements, including Clinical Scores and Vital Signs;
3. Laboratory Results;
4. Medical History and Treatments, including Diagnoses, Procedures Performed, and Medications Prescribed;

Given the focus of the analysis on predicting in-hospital mortality based on information from the first 24 hours of ICU admission, the maximum, minimum, and mean values were extracted for each quantitative measurement to capture the range and average of the patient's condition during this initial period.

2.2.1 Inclusion Criteria

The inclusion criteria were: 1) first admission to the ICU for each patient; 2) information available within the first 24 hours after admission to the ICU. The exclusion criteria were: 1) patients that died $<$ 24h (excluding these patients helped avoid extreme cases that could bias the models); 2) numerical data with a missing rate of more than 50% were not used for predictions.

2.2.2 Clinical Scores

Robust clinical scoring systems are critical for accurately forecasting in-hospital mortality, allowing for a comprehensive evaluation of ICU patients' conditions using a single, reliable, and validated clinical tool. This study incorporates a variety of clinical scores, each reflecting on a specific aspect of patient health. These are the disease severity scores:

- Glasgow Coma Score (GCS): evaluates neurological impairment, ranging from 3 (deep unconsciousness) to 15 (fully awake);
- Sequential Organ Failure Score (SOFA): quantifies organ dysfunction, ranging from 0 (normal function) to 4 (high degree of dysfunction) for each of the six interested organs;
- Simplified Acute Physiology Score II (SAPS II): estimates the probability of mortality, ranging from 0 to 163 (higher risk of death);
- Oxford Acute Severity of Illness Score (OASIS): a measure of acute illness severity using physiological and hospital data, ranging from 0 to 100 (higher severe illness);
- Mayo End-stage Liver Disease (MELD): designed for liver transplant candidates to predict survival, ranging from 6 (less severe disease) to 40 or more (severe disease).

2.3 Data Preparation

The data preparation process investigated multiple healthcare datasets previously extracted from the MIMIC IV database. First, the target variable was defined. Next, for a comprehensive analysis, the predictors were categorized and divided into numerical, categorical, and dummy variables according to their data types. Diagnoses, procedures, and medical prescriptions were

analyzed separately due to a more complex classification within the database. Proper categorization is essential for implementing specific data-processing techniques for each type, thereby improving the effectiveness of subsequent analyses and models.

2.3.1 Target Variable

To determine and extract in-hospital mortality, the feature ”DOD” (Date of Death) has been taken into account. The initial date format was converted into a binary format, indicating whether a patient died during their hospital stay. If a date was detected, the value 1 was assigned, indicating non-survival. Otherwise, the value 0 was assigned, indicating survival. This transformation is crucial because it is directly fed into the ML models, guaranteeing training on an accurate and meaningful outcome.

2.3.2 Diagnoses, Procedures and Medical Prescription

Diagnoses, procedures, and medical prescriptions are necessary to understand the complex conditions of patients within 24 hours after admission to the ICU. The MIMIC-IV database classifies diagnoses and procedures using the Ninth Version of the International Classification of Diseases (ICD-9) codes and their corresponding ”long descriptions”. The medical prescriptions are classified using ”drug prescribed”, ”dose”, ”dose unit”, and ”route”. Both classifications are integrated with the identification code for each patient’s stay in the ICU (”stay_id”), allowing a standardized aggregation of medical information. Rows with missing values were filtered out, removing incomplete records and guaranteeing only complete cases were considered. For each patient, multiple diagnoses, procedures, and medication prescriptions were identified during the stay in the ICU, posing an ongoing challenge in understanding the overall clinical condition. To address this complication, each ”stay_id” was linked to multiple rows. For the diagnoses and procedures, the ”long descriptions” associated with each ”stay_id” were concatenated. Instead, for the medical prescriptions, the ”drug prescribed”, ”dose”, ”dose unit”, and ”route” were concatenated. This aggregation provided a complete view of each patient’s medical profile during the first 24 hours of admission.

Given the qualitative nature of textual medical data, Natural Language Processing (NLP) was employed to convert the concatenated descriptions into quantitative data. The `Bio_ClinicalBERT` model, a pre-trained model on clinical text, was used to tokenize and encode the textual data

into numerical embeddings. These embeddings can represent the descriptions of a patient's medical condition in a high-dimensional space where semantic similarities between words are translated into spatial proximities. Principal Component Analysis (PCA) was employed to manage the high-dimensionality of the embeddings by transforming the numerical embeddings into a new set of variables (principal components), linear combinations of the original embeddings. Figure 2.1 displays the relationship between the number of principal components and the cumulative explained variance for diagnoses, procedures, and medical prescriptions. The dashed red line and the dotted green line represent the 90% and 95% thresholds of explained variance, respectively, while vertical lines indicate the number of components for each threshold. For the following analysis, the 90% threshold of explained variance was considered to streamline the data while preserving the pattern in a reduced dimension. Hence, a total of 92 components were chosen for procedures, 111 for diagnostics, and 94 for prescriptions.

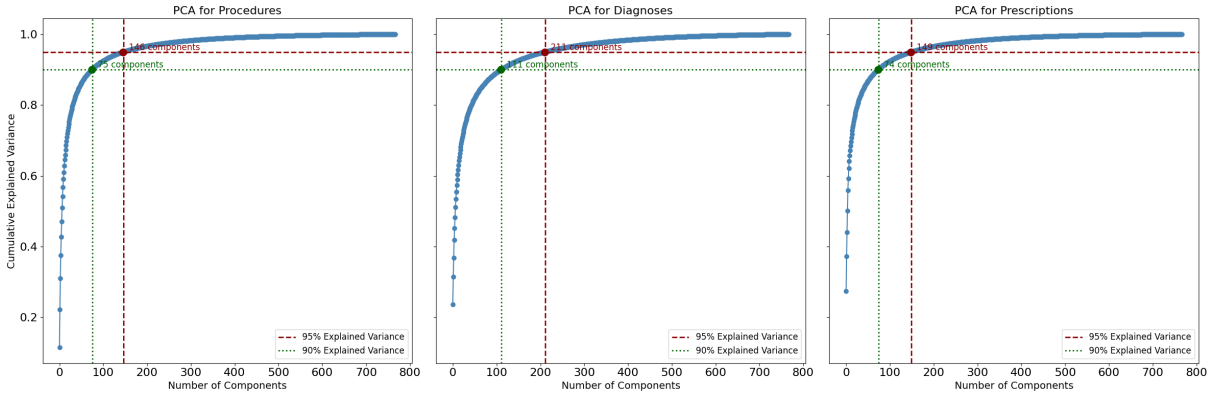

Figure 2.1: Principal Component Analysis (PCA) for Procedures, Diagnoses, and Prescriptions

2.3.3 Categorical Features

In the analysis of categorical features, the different levels were aggregated into broader categories to ensure sufficient information for each level. For example, admission types were grouped into "Emergency-Related", "Urgent", "Elective", and "Surgical", while admission locations were categorized into "Emergency", "Referral", "Transfer", "Direct Admission", "Procedure-Related", and "Unknown". Similarly, ventilation status was grouped into "NonInvasiveVent", "InvasiveVent", and "None", and dialysis types into "None", "Peritoneal", "IHD-Based", and "CRRT-Based". The `OneHotEncoder` was employed to encode categorical data, converting each level into a binary format that can be inputted into machine learning algorithms. Sub-

sequently, encoded variables that showed a significantly high Variance Inflation Factor (VIF) were removed to address potential multicollinearity arising from the ”dummy variable trap”.

2.3.4 Numerical Features

The analysis of numerical attributes involved a multi-step process designed to ensure data quality and consistency. This process included handling outliers, imputing missing values, scaling the data, and checking for multicollinearity.

Outliers were detected using a revised Interquartile Range (IQR) method, which used each feature's 15th and 85th percentiles and identified as outliers those values outside 1.5 times the considered range. The objective of this approach is to specifically address and eliminate data processing errors present in the MIMIC-IV database. The observations were separated into groups of alive and deceased to handle outliers differently, as critical patients often exhibit extreme situations that should not be ignored. After removing the outliers, the number of patients decreased from the initial 27,007 (18,056 alive and 8,951 deceased) to 13,068 (8,749 alive and 4,319 deceased). Features with a significant number of missing values were eliminated by applying a threshold of 40%. This resulted in a reduced dataset of 66 numerical features, from the original 112. To address the remaining missing values, the `KNNImputer` method was employed, which uses the nearest neighbors' values to fill in the missing data. To summarize, the data-cleaning process was thorough, reducing features from 178 to 66 while preserving 13,068 observations.

To ensure that all numerical attributes contributed equally to the ML models' performance, `RobustScaler` was employed. This is a standardization technique that uses the median and interquartile range for scaling, thereby minimizing the influence of the extreme values by preserving the integrity of the scaled data.

The last step consisted of checking for multicollinearity using the correlation matrix (A.1) to avoid redundancy and ensure that each predictor contributes uniquely to the models. Initially, we extracted the maximum, minimum, and mean values for each feature to fully understand the central tendency and range of the data, which is especially crucial for comprehending severe ICU cases. However, this detailed representation introduces significant correlations between

the maximum, minimum, and mean values of the same clinical feature. For instance, heart rate, blood pressure, and temperature show high correlations between their maximum, minimum, and mean values. Additionally, there is a clear correlation between clinical scores and the individual features used to compute them. Clinical scores such as SOFA, SAPS II, and LODS were derived from multiple physiological measurements and laboratory results, which inherently introduces correlation. To address multicollinearity, the mean value of each feature was excluded, as the correlation between maximum and minimum is weaker than that of the mean. As a result, the total number of numerical features was reduced from 66 to 44. Therefore, although incorporating the maximum, minimum, and mean values provides useful insights into patients' conditions, it is imperative to manage the resulting multicollinearity.

2.3.5 Merging of Datasets

The datasets resulting from the data processing were then merged using the "stay_id" as the basis, guaranteeing that only records of the same patient's stay in the ICU were combined. The complete dataset of predictors consisted of 349 features, including all the processed features previously described.

2.4 Feature Selection

A multivariate Cox Proportional Hazards model was employed to identify the critical factors associated with mortality. This model was chosen for feature selection due to its ability to handle time-to-event data, emphasizing the importance of survival times and features with a meaningful impact on mortality.

From the initial 349 features, the model identified 147 features significantly associated with in-hospital mortality. The selected features were then used to train the ML models. All reported p-values were two-tailed and a p-value < 0.05 was considered statistically significant. Table 2.1 presents a concise summary of the Cox model's performance with significant features. The concordance index of 0.83 indicates that the model performs well in distinguishing between various survival durations. The log-likelihood ratio test, with a value of 5393.05 on 147 degrees of freedom, confirms that the included features significantly improve the model fit.

The predictors with a statistically significant impact were further examined and graphed to display their hazard ratios (Figure A.2). Table 2.2 shows the key factors associated with increased mortality. More precisely, the presence of "metastatic solid tumors" (HR = 1.58) was the highest, with a 58% higher risk of in-hospital mortality compared to those without this condition. Additionally, the presence of "age" (HR = 1.55), "severe liver disease" (HR = 1.44), "rrt dialysis" (HR = 1.32), and "malignant cancer" (HR = 1.25) were significantly associated with a higher risk of in-hospital mortality.

Statistic	**Value**
Concordance	0.83
Partial AIC	66035.83
log-likelihood ratio test	5393.05 on 147 df
-log2(p) of ll-ratio test	inf

Table 2.1: Summary of Cox Model for Significant Features

Variable	**coef**	**exp(coef)**	**se(coef)**	**z**	**p-value**	**-log2(p)**
metastatic_solid_tumor	0.46	1.58	0.06	7.84	<0.005	47.64
age_at_icu_intime	0.44	1.55	0.03	12.92	<0.005	124.44
severe_liver_disease	0.36	1.44	0.08	4.76	<0.005	19.00
malignant_cancer	0.22	1.25	0.05	4.85	<0.005	11.44
rrt_dialysis_present	0.28	1.32	0.09	3.14	<0.005	9.20

Table 2.2: Features with the Highest Hazard Ratios in the Cox Proportional Hazards Model

2.5 Outcome Prediction

An overview of the results of the data processing and patient selection process is summarized in Figure 2.2. A total of 13068 patients were considered for analysis, with an in-hospital mortality rate of 33.1% (8749 survivors and 4319 non-survivors). The cohort was split into a training set (80%) and a test set (20%), and stratification was used to ensure that the training and test sets accurately represented the overall class distribution.

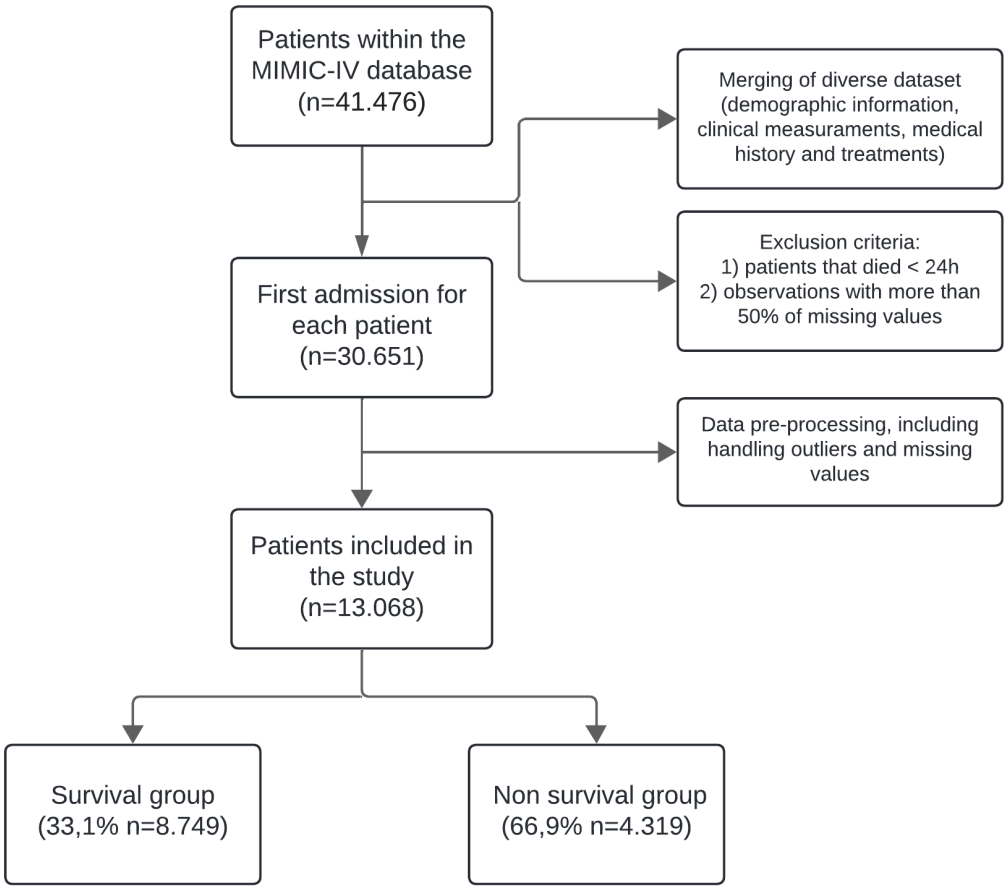

Figure 2.2: Flowchart of patient selection

2.5.1 UMAP and K-Means Application

A clustering analysis was performed on the test set to identify different groups of patients associated with a higher in-hospital mortality rate. For better visualization, the high-dimensional feature space was reduced to two dimensions using the Uniform Manifold Approximation and Projection (UMAP) technique. `UMAP` aims to minimize the disparities between high-dimensional and low-dimensional representations of the data by maintaining the proximities of close points in both spaces. The `K-means` technique was then used to cluster the UMAP-transformed data points. The optimal number of clusters (3) was determined using the elbow method, which plots the Within-Cluster Sum of Squares (WCSS) for each number of clusters, as shown in Figure 2.3.

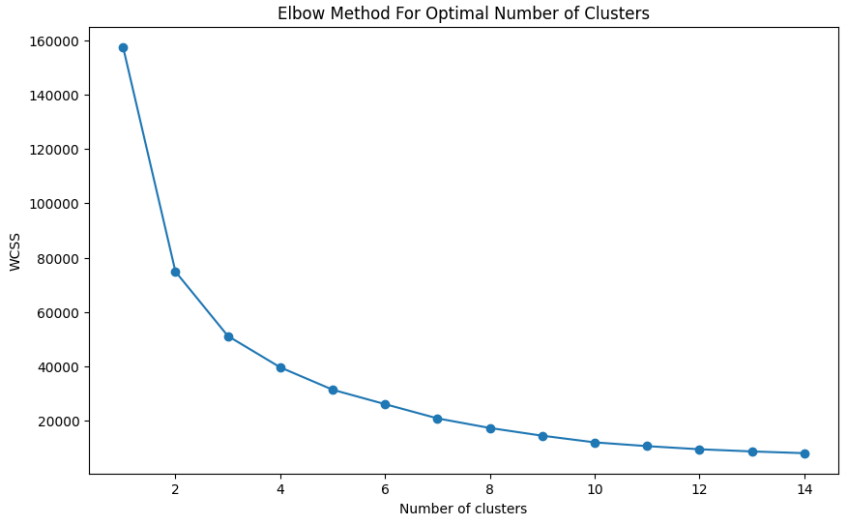

Figure 2.3: Elbow Method For Optimal Number of Clusters

2.5.2 Machine Learning Models

The ML models considered in the study are K-Nearest Neighbors (KNN), Logistic Regression (LogR), Random Forest (RF), Support Vector Machines (SVM), and Neural Networks (MLP). The features used in predicting in-hospital mortality were previously selected through the Cox Proportional Hazards model. `GridSearchCV` conducted hyperparameter tuning for each model, identifying the optimal parameters based on the highest f1-score on the training set during cross-validation. The models were evaluated using performance metrics, such as f1-score, precision, recall, and the Area Under the Receiver Operating Characteristic (ROC) curve (AUC).

Chapter 3

Results

3.1 Inferential Analysis

Table 3.1 shows the patients' baseline characteristics. For a complete table of the population's baseline characteristics, see Table B.1. Inference tests revealed significant disparities between the two groups, survivors and non-survivors. The age was significantly higher in the non-survivor group than in the survivor group (61.10 ± 16.89 vs. 72.13 ± 14.03, $p < 0.001$). Patients in the non-survivor group had higher severity of clinical scores within 24 hours of first admission and a higher incidence of congestive heart failure. For instance, the average SOFA score (3.28 ± 2.23 vs. 4.56 ± 2.92, $p < 0.001$) and SAPS II (29.38 ± 9.95 vs. 38.78 ± 11.13, $p < 0.001$) were significantly lower in non-survivors than in survivors. Furthermore, the incidence of congestive heart failure was significantly higher among non-survivors (23.30% vs. 21.23%, $p = 0.0018$). Myocardial infarction and diabetes with complications were also statistically significant ($p < 0.005$), while the rest of the comorbidities did not present significant differences between the two groups. These findings highlight that older age and higher clinical severity scores are strongly associated with increased in-hospital mortality within the first 24 hours of ICU admission.

	Survivor	Non-Survivor	P-value
age_at_icu_intime	61.10 ± 16.89	72.13 ± 14.03	< 0.001
Male, n (%)	7054 (58.00%)	3134 (56.20%)	0.0250
weight	82.45 ± 19.81	75.99 ± 20.34	< 0.001
Clinical Scores			
GCS	14.41 ± 1.08	13.72 ± 2.26	< 0.001
GCS_eyes	3.35 ± 0.69	3.09 ± 0.99	< 0.001
GCS_verbal	3.20 ± 2.26	2.91 ± 2.18	< 0.001
GCS_motor	5.82 ± 0.47	5.34 ± 1.24	< 0.001
SOFA	3.28 ± 2.23	4.56 ± 2.92	< 0.001
OASIS	28.62 ± 7.01	33.09 ± 7.98	< 0.001
SAPSII	29.38 ± 9.95	38.78 ± 11.13	< 0.001
SIRS	2.46 ± 0.96	2.56 ± 0.94	< 0.001
LODS	3.09 ± 1.81	4.59 ± 2.52	< 0.001
MELD	10.11 ± 3.98	14.00 ± 6.81	< 0.001
Co-morbidities			
Myocardial Infarct	1974 (16.23%)	835 (14.97%)	0.0349
Congestive Heart Failure	2582 (21.23%)	1301 (23.33%)	0.0018
Peripheral Vascular Disease	1379 (11.34%)	628 (11.26%)	0.8991
Cerebrovascular Disease	1945 (15.99%)	903 (16.19%)	0.7541
Dementia	369 (3.03%)	180 (3.23%)	0.5194
Chronic Pulmonary Disease	2671 (21.96%)	1266 (22.70%)	0.2804
Rheumatic Disease	357 (2.94%)	150 (2.69%)	0.3879
Peptic Ulcer Disease	382 (3.14%)	182 (3.26%)	0.6998
Mild Liver Disease	1319 (10.85%)	572 (10.26%)	0.2486
Diabetes without Complications	2507 (20.61%)	1174 (21.05%)	0.5177
Diabetes with Complications	877 (7.21%)	448 (8.03%)	0.0571
Paraplegia	573 (4.71%)	247 (4.43%)	0.4275
Renal Disease	2016 (16.58%)	923 (16.55%)	0.9827
Malignant Cancer	1506 (12.38%)	641 (11.49%)	0.0967
Severe Liver Disease	600 (4.93%)	238 (4.27%)	0.0571
Metastatic Solid Tumor	690 (5.67%)	309 (5.54%)	0.7481
AIDS	49 (0.40%)	28 (0.50%)	0.4181

Table 3.1: Baseline characteristics of the study population. Continuous variables are presented as mean and standard deviation. They are compared using a Student's t-test when the data is approximately normally distributed, or the Mann-Whitney U test when it is not normally distributed. The Shapiro-Wilk test is used to assess whether the data are normally distributed. Categorical variables are presented as counts (percentages) and compared using the Chi-Square test or Fisher's exact test, as appropriate.

3.2 Data Visualization

To visually assess the impact of mortality-related critical factors and identify different groups of patients based on clinical condition, various data visualization techniques were employed.

3.2.1 Kaplan Meyer Model

The impact on survival was assessed by performing a Kaplan-Meier survival analysis and the log-rank test. Figure 3.1 displays the probability of survival over time. The overall survival curve, shown on the left, depicts a decreasing likelihood of survival, indicating a high early mortality rate. The survival curves on the right represent the impact of metastatic solid tumors and severe liver disease, previously identified as highly impacting on mortality through the Cox Proportional Hazards model. Both factors, tested with the log-rank, had a p-value < 0.001, confirming that their presence significantly diminishes patients' survival prospects.

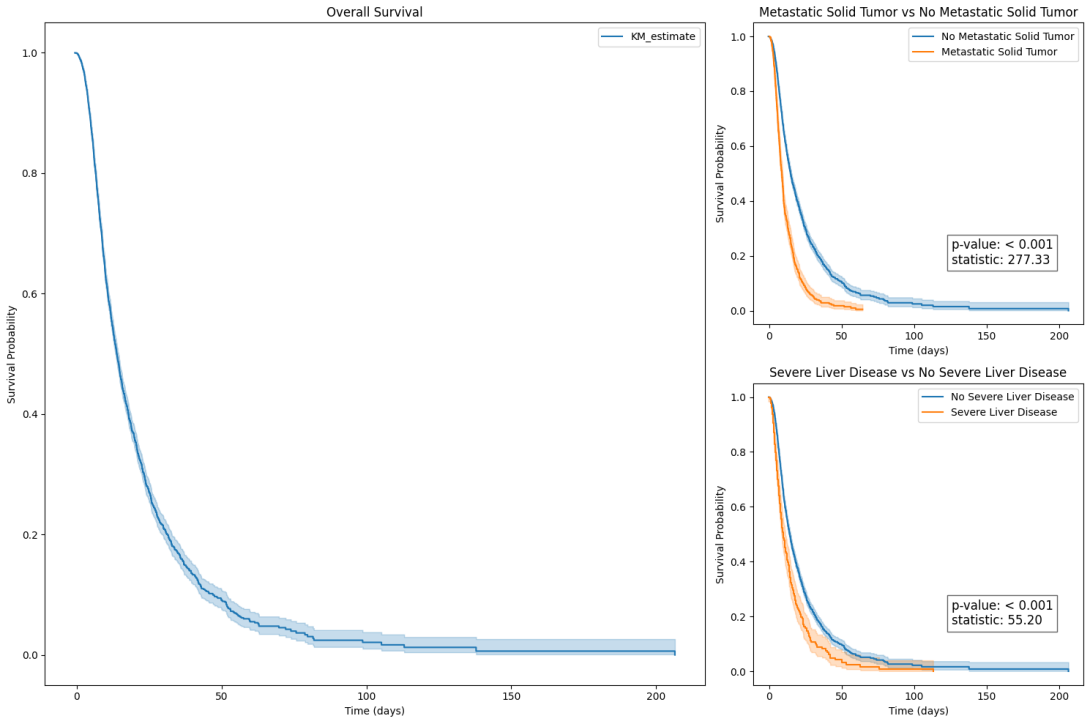

Figure 3.1: Kaplan-Meier Analysis of Survival Probability Over Time

3.2.2 Cluster Analysis

The cluster analysis identified three clusters of patients, which are displayed in Figure 3.2. On the left, the overall survival status of the test cohort is plotted, with blue indicating survivors and

red indicating non-survivors. Cluster 1 exhibits a notably higher concentration of non-survivors, indicating a higher mortality risk. Cluster 2, on the other hand, predominantly consists of survivors, suggesting a lower mortality risk. Cluster 3 shows a mix of survivors and non-survivors, with a slightly higher concentration of survivors compared to Cluster 1.

To further analyze the impact of specific conditions, the clusters of ICU patients based on mortality risk were investigated for the presence of metastatic solid tumors and severe liver disease. Therefore, on the right, the same clusters are visualized with points colored based on the presence of metastatic solid tumors and severe liver disease. Cluster 1 contains several patients with both metastatic solid tumors and severe liver disease, which might correlate with the higher mortality risk observed in this cluster. Cluster 2 predominantly includes patients without the critical factors associated with mortality, supporting its characterization as a lower-risk cluster. Cluster 3, similar to Cluster 1, contains a mix of patients with and without metastatic solid tumors.

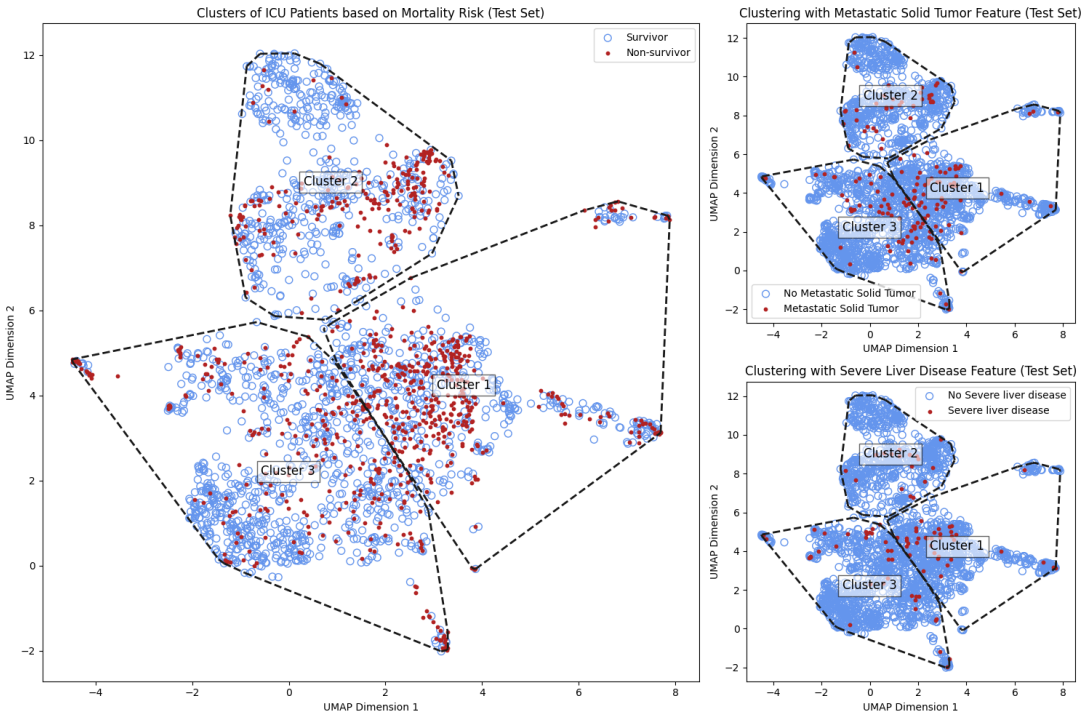

Figure 3.2: Clustering of ICU Patients Based on Mortality Risk

3.3 Comparison of ML Models Performances

This section reports the performance of the ML models employed on the test cohort to predict ICU mortality. The models assessed include K-Nearest Neighbors, Logistic Regression, Random Forest, Support Vector Machines, and Multi-Layer Perceptron. The results for each model are detailed below, highlighting their respective strengths and weaknesses in predicting patient mortality. It is important to clarify that survivors are indicated with the value 0, while non-survivors are indicated with the value 1.

3.3.1 K-Nearest Neighbors

For the K-Nearest Neighbors classifier, the best combination found was {`leaf_size: 1; n_neighbors: 32; p: 1`} with an f1-score of 0.572 on the training set during cross-validation (Table 3.2). This model employs the `Manhattan distance` metric, which calculates the absolute differences in coordinates between data points, taking into account `32 nearest neighbors`. Setting `leaf_size=1` allows the algorithm to improve search speed quickly by switching to brute-force search. Upon testing, the model demonstrated strong performance, with an accuracy of 0.77. The precision for the survivor Class was 0.79, slightly higher compared to 0.71 for the non-survivor Class. However, the recall was significantly better for the survivor Class (0.90) than for the non-survivor Class (0.51), indicating that the model misses a considerable number of deceased patients. Overall, the f1-score indicates superior performance for the survivor class (0.84 vs. 0.59). The confusion matrix reveals that the classifier accurately identified 1576 out of 1750 instances for Class 0 and 434 out of 864 for Class 1.

	Precision	**Recall**	**F1-Score**	**Support**
0 (survived)	0.79	0.90	0.84	1750
1 (not survived)	0.71	0.51	0.59	864
Accuracy			0.77	2614
Macro avg	0.75	0.70	0.72	2614
Weighted avg	0.76	0.77	0.76	2614

Table 3.2: K-Nearest Neighbors Classifier Performance Metrics

3.3.2 Logistic Regression

For the Logistic Regression classifier, the best combination found was {`C: 0.7; class_weight: balanced; max_iter: 90; solver: lbfgs`} with an f1-score of 0.734 on the training set during cross-validation (Table 3.3). The `lbfgs solver` was selected for regularization; in particular, `C = 0.7` indicates a moderate balance between well-fitting training data and generalizing to new data. The model automatically adjusts weights to account for class imbalances, resulting in improved performance for the minority class. Upon testing, the model achieved an overall accuracy of 0.81. The precision, recall, and f1-scores for Class 0 were 0.89, 0.83, and 0.86, respectively, and for Class 1, they were 0.69, 0.79, and 0.74, respectively. These results confirm the outstanding performance of the LogR model in distinguishing between the two groups. The confusion matrix shows that the classifier correctly identified 1444 out of 1750 instances for Class 0 and 681 out of 864 for Class 1.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.89	0.83	0.86	1750
1	0.69	0.79	0.74	864
Accuracy			0.81	2614
Macro avg	0.79	0.81	0.80	2614
Weighted avg	0.82	0.81	0.82	2614

Table 3.3: Logistic Regression Classifier Performance Metrics

3.3.3 Random Forest

For the Random Forest classifier, the best combination found was {`min samples leaf: 33; n estimators = 800`} with an f1-score of 0.635 on the training set during cross-validation (Table 3.4). The model uses `800 trees`, ensuring a minimum of `33 samples` to prevent overfitting. Upon testing, the RF model achieved an overall accuracy of 0.80. The precision, recall, and f1-scores for Class 0 were 0.80, 0.94, and 0.86, respectively, and for Class 1, they were 0.81, 0.53, and 0.64, respectively. The RF model showed a comprehensively good performance; however, the low recall for Class 1 suggests that various deceased patients were misclassified. The confusion matrix shows that the classifier correctly identified 1641 out of 1750 instances for Class 0 and 458 out of 864 for Class 1.

	Precision	Recall	F1-Score	Support
0	0.80	0.94	0.86	1750
1	0.81	0.53	0.64	864
Accuracy			0.80	2614
Macro avg	0.80	0.73	0.75	2614
Weighted avg	0.80	0.80	0.79	2614

Table 3.4: Random Forest Classifier Performance Metrics

3.3.4 Support Vector Machine

For the Support Vector Machine classifier, the best combination found was {`C: 0.630; kernel = linear`} with an f1-score of 0.715 on the training set during cross-validation (Table 3.5). A `linear` decision boundary was used; `C: 0.630` indicates moderate regularization, balancing the fit between the training data and generalization to new data. Upon testing, the SVM model provides good overall performance with an accuracy of 0.82. The precision, recall, and f1-scores for Class 0 were 0.84, 0.91, and 0.87, respectively, and for Class 1, they were 0.77, 0.66, and 0.71, respectively. The confusion matrix shows that the classifier correctly identified 1585 out of 1750 instances for Class 0 and 567 out of 864 for Class 1.

	Precision	Recall	F1-Score	Support
0	0.84	0.91	0.87	1750
1	0.77	0.66	0.71	864
Accuracy			0.82	2614
Macro avg	0.81	0.78	0.79	2614
Weighted avg	0.82	0.82	0.82	2614

Table 3.5: Support Vector Machine Classifier Performance Metrics

3.3.5 Multi-Layer Perceptron

The best parameters for the Multi-Layer Perceptron classifier were {`alpha: 4.75; hidden layer sizes = 42; max iter: 2000`}, which gave an f1-score of 0.702 on the training set during cross-validation (Table 3.6). The hidden layer has `42 neurons`; `alpha = 4.75` indicates strong regularization to prevent overfitting with sufficient iterations for convergence. Upon testing, the MLP model performed well overall, with an accuracy of 0.82. The precision, recall, and f1-scores for Class 0 were 0.84, 0.90, and 0.87, respectively, and for Class 1, they were 0.77, 0.65, and 0.70, respectively. The confusion matrix shows that the classifier correctly identified 1579 out of 1750 instances for Class 0 and 563 out of 864 for Class 1.

	Precision	Recall	F1-Score	Support
0	0.84	0.90	0.87	1750
1	0.77	0.65	0.70	864
Accuracy			0.82	2614
Macro avg	0.80	0.78	0.79	2614
Weighted avg	0.82	0.82	0.82	2614

Table 3.6: Multi-Layer Perceptron Classifier Performance Metrics

3.3.6 ROC Curves

ROC curves were plotted to comprehensively assess the various ML models' performances and visualize their trade-offs between true positive rate and false positive rate. Figure 3.3 displays the curves with the Area Under the Curve (AUC) values annotated in the legend. LogR, SVM, and RF models demonstrated remarkable accuracy in correctly predicting classes, as seen by their AUC of 0.89. The MLP model also demonstrated commendable performance with an AUC of 0.88. However, the KNN model exhibited a relatively lower capacity for distinguishing between classes, with an AUC of 0.82.

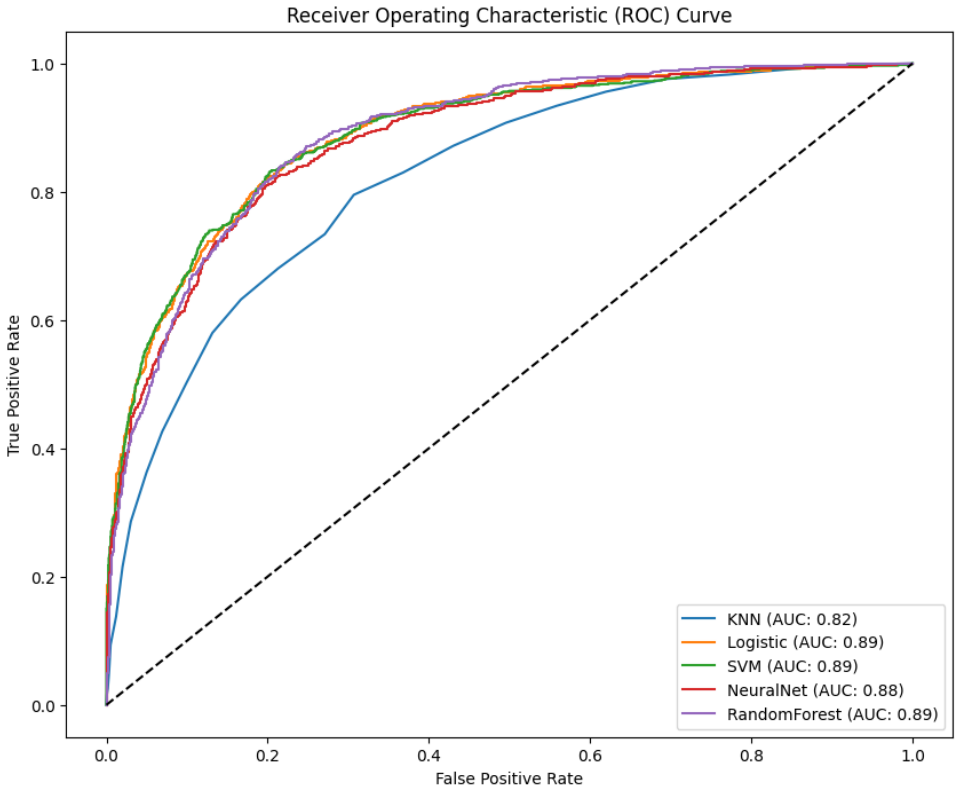

Figure 3.3: Receiver Operating Characteristic (ROC) Curve

Chapter 4

Conclusion

The study developed a comprehensive statistical framework to improve prediction accuracy through extensive data processing and detailed feature selection. The focus was on comparing the performance of various ML models in predicting in-hospital mortality given the clinical condition of the initial 24 hours from admission to the ICU.

The feature selection conducted with the Cox Proportional Hazards model offered a crucial temporal perspective, identifying the critical factors necessary for an accurate prediction. Moreover, the Kaplan-Meier and Cluster Analysis visually represented ICU patients' survival and mortality patterns. The survival curves revealed that the presence of metastatic solid tumors or severe liver disease significantly worsens survival prospects, as confirmed by the log-rank test. The clustering distinguished three unique patient groups according to mortality risk, and further investigated for the presence of metastatic solid tumor and severe liver disease. Cluster 1, predominantly consisting of non-survivors, was associated with a high mortality risk. Cluster 2, predominantly consisting of survivors, was identified as a lower-risk group. Cluster 3 showed a mix of survivors and non-survivors, similar to Cluster 1, but with a slightly better survival rate.

In predicting ICU mortality the models with the best performance were Logistic Regression, Support Vector Machine, and Random Forest, with AUC values of 0.89. These models also displayed high precision, recall, and F1 scores across survivor and non-survivor classes, making them reliable in distinguishing between the two groups. In particular, LogR and SVM showed a strong balance in their precision and recall, ensuring accurate identification for both survivors and non-survivors. RF also demonstrated strong performance, especially in precision,

but its recall for non-survivors was lower, suggesting the classification of numerous deceased patients as alive. The Neural Network also performed accurately, with an AUC of 0.88 and balanced performance metrics. It demonstrated high precision and recall for survivors, as well as reasonable performance for non-survivors, making it a robust choice for mortality prediction. On the other hand, the K-Nearest Neighbors model exhibited the lowest performance among the models evaluated, with an AUC of 0.82. The model particularly struggled with recall for non-survivors, indicating a significant number of misclassifications in deceased patients. This lack of performance for deceased patients makes this model less reliable for predicting ICU mortality compared to the other models.

Overall, the Logistic Regression, Support Vector Machine, and Random Forest models emerged as the most effective for predicting in-hospital mortality, delivering a satisfactory ability to correctly classify patients given the information gathered within the first 24 hours from the first admission to the ICU.

4.1 Limitations and Practical Implications

The study had several limitations, as highlighted in previous similar studies [4]. First, the study used a retrospective design. Therefore, it was difficult to establish causality and distinguish between treatment-induced and patient-induced causes for some changes, potentially leading to bias. Second, the high number of missing values in many relevant features resulted in the exclusion of various patients, potentially impacting the study's generalizability. However, though the extent and the scope of this study were limited, it nonetheless gives way to further improvements and developments.

Bibliography

[1] Duncan Shillan et al. “Use of machine learning to analyse routinely collected intensive care unit data: a systematic review”. In: *Critical Care* 23.1 (2019), p. 284. ISSN: 1364-8535. DOI: 10.1186/s13054-019-2564-9. URL: https://doi.org/10.1186/s13054-019-2564-9.

[2] Fadi Shamout, Tianpei Zhu, and David A. Clifton. “Machine Learning for Clinical Outcome Prediction”. In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 116–126. DOI: 10.1109/RBME.2020.3007816. URL: https://doi.org/10.1109/RBME.2020.3007816.

[3] A. Johnson et al. *MIMIC-IV (version 2.2)*. https://physionet.org/content/mimiciv/2.2/. 2023.

[4] Xiao-Yan Ding, Zhi-Zhong Chen, and Han Chen. “Visualizing ICP “Dose” of neurological critical care patients”. In: *Intensive Care Medicine* 50.5 (May 2024), pp. 781–783. ISSN: 1432-1238. DOI: 10.1007/s00134-024-07424-5. URL: https://doi.org/10.1007/s00134-024-07424-5.

Appendix A

Methodology

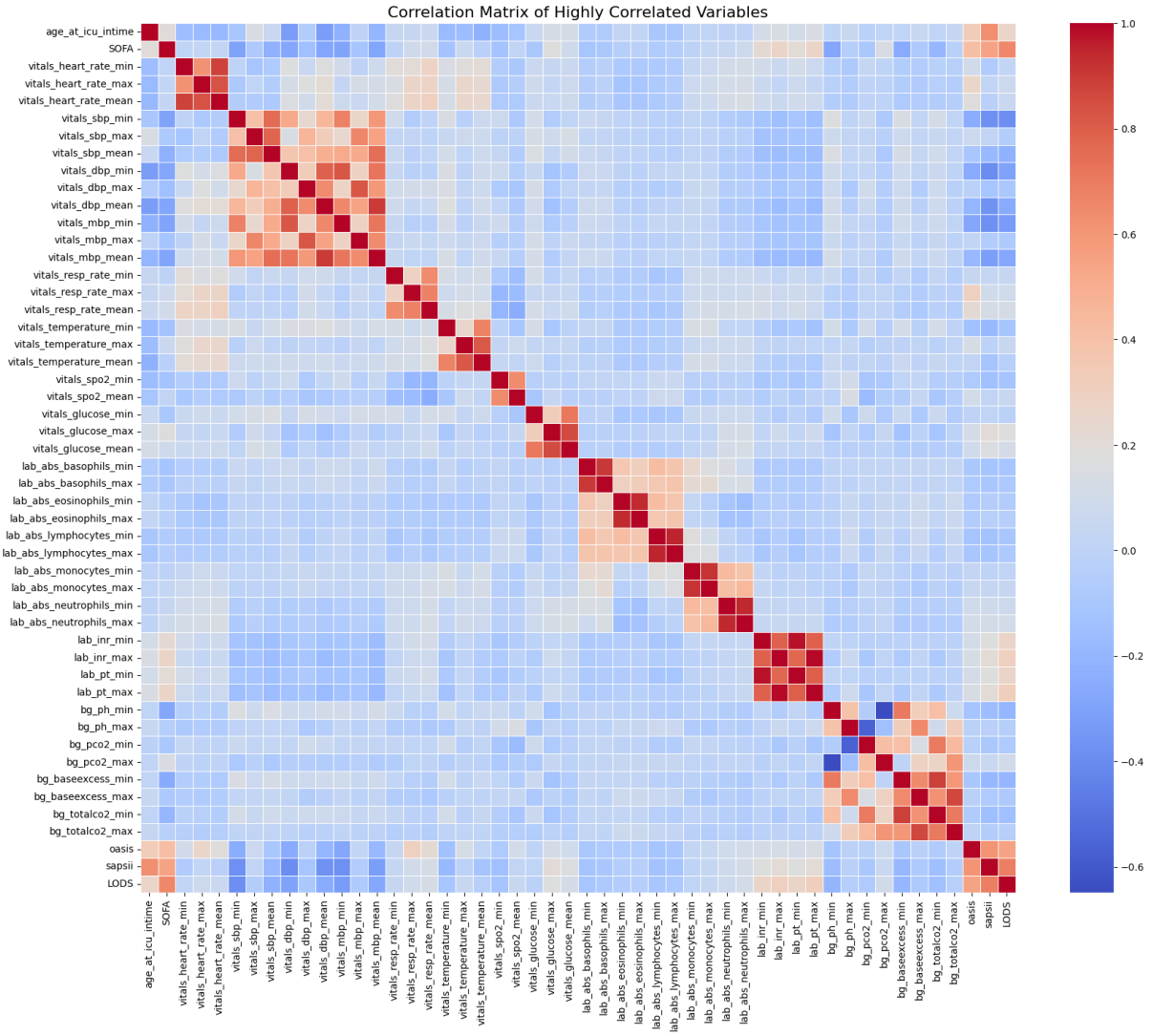

Figure A.1: Correlation Matrix of Highly Correlated Features

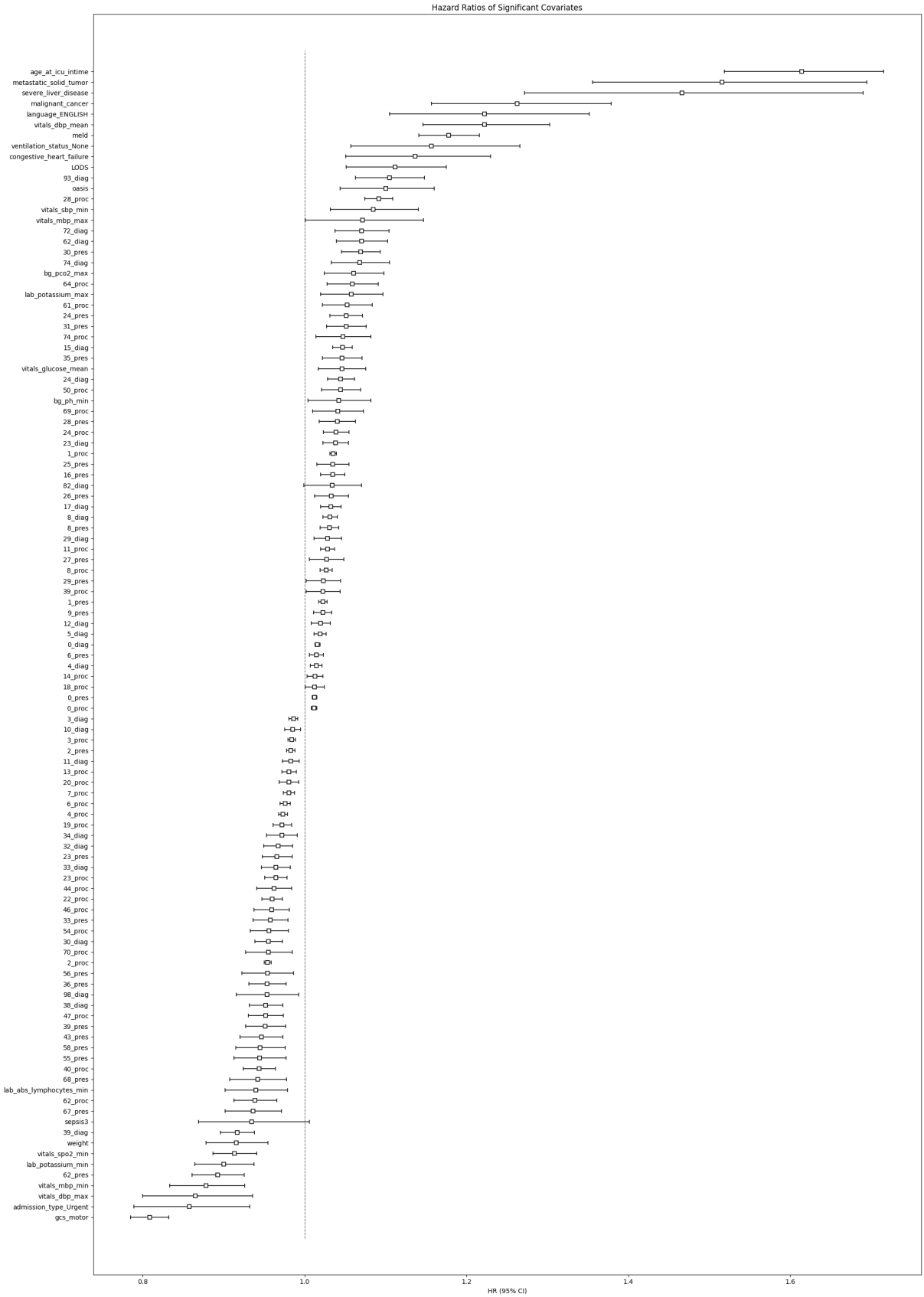

Figure A.2: Hazard Ratios of Significant Predictors (p-value < 0.01)

Appendix B

Results

	Survivor	Non-Survivor	P-value
age_at_icu_intime	61.10 ± 16.89	72.13 ± 14.03	< 0.001
Male, n (%)	7054 (58.00%)	3134 (56.20%)	0.0250
weight	82.45 ± 19.81	75.99 ± 20.34	< 0.001
Vitals			
vitals_heart_rate_min	67.74 ± 13.05	69.83 ± 14.95	< 0.001
vitals_heart_rate_max	99.47 ± 17.14	102.25 ± 19.92	< 0.001
vitals_sbp_min	93.28 ± 14.64	92.44 ± 17.28	0.0019
vitals_sbp_max	146.93 ± 19.88	150.30 ± 23.14	< 0.001
vitals_sbp_mean	118.48 ± 14.15	119.72 ± 16.99	0.0035
vitals_dbp_min	47.67 ± 10.12	45.25 ± 11.09	< 0.001
vitals_dbp_max	85.31 ± 16.85	87.74 ± 18.82	< 0.001
vitals_dbp_mean	63.12 ± 10.18	61.91 ± 10.74	< 0.001
vitals_mbp_min	61.11 ± 10.88	58.48 ± 12.36	< 0.001
vitals_mbp_max	101.82 ± 15.83	103.14 ± 18.11	0.0002
vitals_resp_rate_min	11.42 ± 3.08	12.53 ± 3.51	< 0.001
vitals_resp_rate_max	26.36 ± 5.47	27.66 ± 5.94	< 0.001
vitals_resp_rate_mean	17.88 ± 2.96	19.22 ± 3.64	< 0.001
vitals_temperature_min	36.32 ± 0.56	36.28 ± 0.58	< 0.001
vitals_temperature_max	37.39 ± 0.60	37.37 ± 0.68	< 0.001
vitals_spo2_min	93.31 ± 2.94	92.28 ± 4.46	< 0.001
vitals_spo2_max	99.76 ± 0.56	99.71 ± 0.60	< 0.001
vitals_spo2_mean	97.45 ± 1.48	97.20 ± 1.75	< 0.001

Continued on next page

Table B.1: Baseline characteristics of the study population (continued)

	Survivor	Non-Survivor	P-value
vitals_glucose_mean	128.85 ± 28.43	138.96 ± 41.66	< 0.001
urine_output	1935.18 ± 1012.68	1533.74 ± 994.10	< 0.001
Laboratory Values			
lab_potassium_min	3.94 ± 0.48	3.92 ± 0.58	0.0092
lab_potassium_max	4.40 ± 0.55	4.56 ± 0.77	< 0.001
lab_abs_basophils_min	0.03 ± 0.02	0.03 ± 0.02	< 0.001
lab_abs_basophils_max	0.03 ± 0.02	0.03 ± 0.02	< 0.001
lab_abs_eosinophils_max	0.10 ± 0.08	0.09 ± 0.08	< 0.001
lab_abs_lymphocytes_min	1.51 ± 0.72	1.16 ± 0.62	< 0.001
lab_abs_monocytes_min	0.51 ± 0.27	0.52 ± 0.31	0.7772
lab_abs_neutrophils_min	8.78 ± 3.76	9.15 ± 4.76	0.1344
lab_inr_min	1.16 ± 0.15	1.27 ± 0.31	< 0.001
lab_ptt_min	27.83 ± 4.32	29.44 ± 5.91	< 0.001
lab_ptt_max	31.84 ± 8.45	35.13 ± 14.36	< 0.001
Blood Gases			
bg_ph_min	7.33 ± 0.06	7.33 ± 0.08	< 0.001
bg_ph_max	7.42 ± 0.05	7.42 ± 0.06	< 0.001
bg_po2_min	108.08 ± 48.34	93.92 ± 47.46	< 0.001
bg_pco2_min	36.56 ± 5.46	37.20 ± 7.06	< 0.001
bg_pco2_max	46.74 ± 7.39	46.48 ± 9.74	< 0.001
Clinical Scores			
GCS	14.41 ± 1.08	13.72 ± 2.26	< 0.001
gcs_eyes	3.35 ± 0.69	3.09 ± 0.99	< 0.001
gcs_verbal	3.20 ± 2.26	2.91 ± 2.18	< 0.001
gcs_motor	5.82 ± 0.47	5.34 ± 1.24	< 0.001
SOFA	3.28 ± 2.23	4.56 ± 2.92	< 0.001
OASIS	28.62 ± 7.01	33.09 ± 7.98	< 0.001
SAPSII	29.38 ± 9.95	38.78 ± 11.13	< 0.001
SIRS	2.46 ± 0.96	2.56 ± 0.94	< 0.001
LODS	3.09 ± 1.81	4.59 ± 2.52	< 0.001
MELD	10.11 ± 3.98	14.00 ± 6.81	< 0.001
Co-morbidities			
Myocardial Infarct	1974 (16.23%)	835 (14.97%)	0.0349
Congestive Heart Failure	2582 (21.23%)	1301 (23.33%)	0.0018
Peripheral Vascular Disease	1379 (11.34%)	628 (11.26%)	0.8991

Continued on next page

Table B.1: Baseline characteristics of the study population (continued)

	Survivor	Non-Survivor	P-value
Cerebrovascular Disease	1945 (15.99%)	903 (16.19%)	0.7541
Dementia	369 (3.03%)	180 (3.23%)	0.5194
Chronic Pulmonary Disease	2671 (21.96%)	1266 (22.70%)	0.2804
Rheumatic Disease	357 (2.94%)	150 (2.69%)	0.3879
Peptic Ulcer Disease	382 (3.14%)	182 (3.26%)	0.6998
Mild Liver Disease	1319 (10.85%)	572 (10.26%)	0.2486
Diabetes without Complications	2507 (20.61%)	1174 (21.05%)	0.5177
Diabetes with Complications	877 (7.21%)	448 (8.03%)	0.0571
Paraplegia	573 (4.71%)	247 (4.43%)	0.4275
Renal Disease	2016 (16.58%)	923 (16.55%)	0.9827
Malignant Cancer	1506 (12.38%)	641 (11.49%)	0.0967
Severe Liver Disease	600 (4.93%)	238 (4.27%)	0.0571
Metastatic Solid Tumor	690 (5.67%)	309 (5.54%)	0.7481
AIDS	49 (0.40%)	28 (0.50%)	0.4181

Table B.1: Baseline characteristics of the study population. Continuous variables are presented as mean and standard deviation. They are compared using a Student's t-test when the data is approximately normally distributed, or the Mann-Whitney U test when it is not normally distributed. The Shapiro-Wilk test is used to assess whether the data are normally distributed. Categorical variables are presented as counts (percentages) and compared using the Chi-Square test or Fisher's exact test, as appropriate.