Machine Learning-Based Prediction of Hospital Length-of-Stay and Patient Survival

Nikolaos Chaikalis[1], Theodoros Christodoulou[1], Georgia Kalitsi[1], Aikaterini Tzatzimaki[1], Eleni Kaldoudi[1,2], and George Drosatos[2]

[1] School of Medicine, Democritus University of Thrace, 68100 Alexandroupoli, Greece
{nikochai, theochri11, georkali7, aikatzat, kaldoudi}@med.duth.gr
[2] Institute for Language and Speech Processing, Athena Research Center, 67100 Xanthi, Greece
gdrosato@athenarc.gr

Abstract. This study tackles two vital predictive challenges in hospitalised patients with severe illness: (i) Outcome Type (discharge versus death) and (ii) hospital Length-of-Stay. Leveraging a real-world dataset from the RegulaRN database comprising 13,145 patient records with 20 features spanning Request, Admission, Patient, and Clinical categories, we formulate these tasks as classification and regression problems, respectively. We evaluate several state-of-the-art Machine Learning models including eXtreme Gradient Boosting, Histogram Gradient Boosting and Random Forest, alongside Deep Neural Networks such as Multi-Layer Perceptrons. Classification models demonstrate strong performance, achieving F1-score up to 0.8, while regression models exhibit limited predictive accuracy with Mean Absolute Errors around 9–10 days. These results highlight the promise and current limitations of Machine Learning and Deep Learning in enhancing clinical decision support, underscoring the need for richer data and advanced modelling strategies in future research to optimise patient care.

Keywords: Machine Learning, Outcome Type, Length-of-Stay, Classification, Regression.

1 Introduction

The prediction of hospital Length-of-Stay (LoS) and Outcome Type (OT) represents a critical component of healthcare operations management and clinical decision-making across health systems worldwide [1]. Accurate early-stage estimation of these variables can significantly enhance resource allocation, including the distribution of Intensive Care Unit (ICU) and general ward beds, the deployment of healthcare personnel across hospital units, and the availability of essential medical equipment. Moreover, it contributes to maintaining and improving the quality of care delivered throughout various stages of hospitalisation. During public health emergencies, such as the COVID-19 pandemic, the timely and reliable prediction of LoS and OT became even more crucial, serving as a

foundational element for optimising healthcare system efficiency under extreme demand conditions.

The Brazilian Unified Health System (SUS) is one of the largest publicly funded healthcare systems in the world, providing universal and free healthcare coverage to a population of approximately 220 million people. As of 2021, SUS maintained around 430,000 hospital beds, including approximately 45,000 ICU beds [2]. Despite efforts to expand capacity during the COVID-19 pandemic, through the addition of thousands of new ICU beds, these resources proved insufficient to meet the surge in severe cases. Consequently, Brazil ranked among the countries with the highest mortality rates during the pandemic. Beyond bed capacity, several structural and organisational challenges further hindered the response, including the absence of integrated information systems for inter-institutional coordination within SUS, inadequacies in hospital infrastructure, and inefficient resource management.

In most regions of Brazil, healthcare regulation is still managed using SISREG, a national system developed in 2001 to coordinate access to public health services, including medical appointments, diagnostic exams, and specialised procedures. However, the growing demand for healthcare services, particularly ICU beds, during the COVID-19 pandemic exposed the system's critical limitations, such as technical obsolescence, long waiting times, and lack of transparency [3]. In response, certain states developed local alternatives. One such initiative was RegulaRN, a platform developed through a partnership between the State Health Department of Rio Grande do Norte (SESAP-RN) and the Health Innovation Laboratory of the Federal University of Rio Grande do Norte (LAIS/UFRN), in collaboration with FUNCERN, a foundation for supporting the Federal Institute of Rio Grande do Norte [4].

RegulaRN [4] is a digital health regulation platform designed to improve access to and management of public healthcare services. Its key functionalities include clinical and ICU bed regulation, which oversees patient transport and inter-facility transfer logistics; emergency case coordination (SAMU); and real-time hospital management dashboards for monitoring bed occupancy, patient flow, and waiting times. During the COVID-19 crisis, RegulaRN played a pivotal role in improving patient outcomes by reducing delays through the implementation of a clinical prioritisation score, contributing significantly to the mitigation of system overload and mortality.

In subsequent years, the RegulaRN platform was expanded with the addition of several specialised modules, including Regular Ambulatório (for scheduling high-complexity outpatient procedures), Regula Vascular (focused on vascular surgery regulation), and Regula NAE (a telemedicine module that provides expert medical opinions across 12 specialities, aiming to reduce unnecessary patient transfers). Despite these advancements, the platform still lacks the integration of Artificial Intelligence (AI) models capable of predicting critical healthcare variables, such as LoS and OT. Nonetheless, data generated by the RegulaRN system have already been utilised in research studies [4, 5], where Machine Learning (ML) and Deep Learning (DL) techniques were applied to predict the outcomes

of regulated patients, demonstrating the potential of these approaches to support evidence-based decision-making.

Within the scope of the IFMBE Scientific Challenge 2025[3], part of IUPESM WC 2025[4], we apply ML and DL methodologies to a dataset extracted from the RegulaRN platform, aiming to predict both OT and LoS as separate classification and regression tasks. The findings of this study may support the ongoing evolution of the RegulaRN system by informing the development of new AI-driven modules and functionalities that enhance healthcare regulation and resource management. Such improvements could contribute to a more efficient user experience, reduced operational costs, and ultimately, better patient outcomes.

The structure of this paper is organised as follows: Section 2 offers a concise review of the existing literature on LoS and OT prediction. Section 3 details the dataset characteristics and the pre-processing steps undertaken. Section 4 describes the modelling approaches and evaluation methodologies employed. In Section 5, we present the predictive performance results of the proposed models. Finally, Section 6 summarises the key findings and suggests directions for future research.

2 Related Work

The effective selection of ML algorithms is closely tied to the specific characteristics of the dataset, particularly in domains such as medicine where data can vary widely in both type and size. Extensive research has been conducted on predicting hospital LoS and patient OT, employing a variety of methodological approaches.

Mekhaldi et al. [6] focused on predicting LoS using an open-source dataset provided by Microsoft, framing the task as a regression problem. They compared the performance of Random Forest (RF) and Gradient Boosting Models (GBM), applying log transformations, Z-score standardisation and categorical encoding during pre-processing. Their results showed that RF achieved superior performance with a Mean Absolute Error (MAE) of 0.44, compared to 0.55 for GBM, while both models exhibited similar R^2 scores (~ 0.92). In a follow-up study, Mekhaldi et al. [7] expanded their analysis by incorporating Multiple Linear Regression (MLR), Support Vector Machines (SVM), and the SMOTE technique to address data imbalance. In this version, GBM demonstrated the best results, with a MAE of 0.44 and an improved R^2 of 0.94.

Zeleke et al. [8] evaluated nine ML regression models to predict patients' actual LoS in an Italian university hospital, using demographic and clinical admission data. K-means clustering was used for patient stratification, and SHAP values identified the most influential features. Among the models, eXtreme Gradient Boosting (XGBoost) achieved the best performance with a Root Mean Square Percentage Error (RMSPE) of 11 days and a MAE of 7.52 days. Similarly, Jain et al. [9] proposed a comprehensive LoS prediction framework using the New

[3] https://sc-iupesm-2025.dei.uc.pt
[4] https://wc2025.org

York State SPARCS dataset, covering 285 CSS diagnostic groups with separate models for neonatal and non-neonatal patients. They approached the problem both as classification, by binning LoS,and regression (1–120 days), applying RF, CatBoost, Multinomial Logistic Regression (MLR), and Linear Regression. Logistic Regression yielded the highest Brier scores (0.75 for non-newborns, 0.78 for newborns), while CatBoost achieved the best R^2 for non-newborns (0.43), and Linear Regression performed best for newborns ($R^2 = 0.82$).

Medeiros et al. [10] extended LoS prediction to a specialised patient population by focusing on paediatric admissions to a public hospital in Brazil. Utilising data from a paediatric ward, the authors applied appropriate data pre-processing and evaluated the performance of several regression models, including MLR, RF, Support Vector Regression (SVR), Ridge Regression, and Partial Least Squares. Following hyperparameter tuning via Grid Search, RF achieved the best results, with an R^2 of 0.6567 and a MAE of 3.51 days. SVR also showed promising performance, yielding an R^2 of 0.5735 and a MAE of 3.86 days.

Another key area of medical research involves predicting in-hospital patient mortality, where various data types and methodological approaches are employed. Shamout et al. [11] provide a comprehensive review of ML models for predicting in-hospital patient outcomes using Electronic Health Record (EHR) data. They cover key pre-processing steps such as feature engineering, standardisation, encoding, and normalisation techniques like Z-score and min-max scaling. The review surveys a variety of models, from traditional methods like Decision Trees and Logistic Regression to advanced approaches including SVM and DL. It also discusses common evaluation metrics and highlights directions for future research in clinical OT prediction.

Barreto et al. [4, 5] investigated both ML and DL methods using the ReguläRN COVID-19 bed-regulation database, the same source used in our study, though their focus was solely on predicting patient OT, not LoS. In their initial work, they applied feature selection, pre-processing, and SMOTE for class imbalance, followed by hyperparameter tuning on models such as Decision Trees, RF, and Deep Neural Netwroks (DNNs) with various optimisers. The best performance came from a Multi-Layer Perceptron optimised with Stochastic Gradient Descent, achieving 84.01% accuracy, 79.57% precision, and an F1-score of 81.00%. The RMSprop optimiser improved recall and specificity (84.67%) and ROC-AUC (91.6%). In a follow-up study using updated data (2021–2024), the authors expanded their analysis with additional ensemble models, XGBoost, AdaBoost, and GBM. XGBoost delivered the highest accuracy and recall (87.77%), while RF and GBM achieved top precision (87.85%) and F1-score (87.56%), respectively.

In recent years, many studies have adopted an integrated approach to simultaneously predict both LoS and patient mortality, an objective aligned with our work. Iwase et al. [12] explored this dual-task prediction in ICU patients using EHR data, categorising LoS into three groups: short (<1 week), medium (1–2 weeks), and long (>2 weeks). For mortality prediction, they employed RF, XGBoost, Neural Networks, and Logistic Regression, while LoS prediction was

based on RF and Logistic Regression using APACHE II and SOFA scores. RF yielded the best AUC for mortality (0.945) and strong results for short (0.961) and long (0.830) stays. Similarly, Alghatani et al. [13] applied six ML models, Logistic Regression, Linear Discriminant Analysis (LDA), RF, kNN, SVM, and XGBoost, as both binary classifiers and regressors for ICU mortality and LoS prediction using the MIMIC-III dataset. They extracted demographic and vital sign features (e.g., heart rate, blood pressure, glucose, SPO2), summarising each patient's data using statistical descriptors (mean, standard deviation, Q1–Q4). RF achieved the highest accuracy for mortality (89%) and binary LoS classification (threshold: median of 2.64 days), while SVR provided the best regression performance with a MAE of 2.81 days.

3 Data Description and Pre-Processing

The data-driven foundation of this study begins with an in-depth exploration and preparation of the dataset used for predictive modelling. We first describe the structure and origin of the dataset, including the feature groups and target variables. This is followed by a detailed statistical characterisation, highlighting key distributions, imbalances, and outliers that may influence model performance. We then conduct a correlation analysis—both linear and non-linear—to identify the strength of relationships between input features and the prediction targets: OT and LoS. Finally, we outline the pre-processing steps applied to enhance data quality and feature representation, including encoding, feature extraction, thresholding, clinical categorisation, handling of missing values, and scaling. All analyses and transformations are performed on the training set, as the test data remains unseen during model development.

3.1 Dataset and features

The dataset, collected from October 2021 to January 2024 under Project Regula SESAP-RN/FUNCERN, originates from the Technological Innovation in Health group at the Federal University of Rio Grande do Norte, in collaboration with the state's Secretary of Public Health. A subset of 13,415 records was extracted, split into 8,049 training, 2,683 validation, and 2,413 test samples. Each record includes two targets and twenty input features, grouped into four categories: Request, Admission, Patient Data, and Clinical Information. The variables in each category are as follows:

- **Request**: Request Date, Type (Adult/Paediatric), Bed Type (Ward/ ICU).
- **Admission**: Admission date, Bed Type (Ward/ ICU), Health Unit.
- **Patient Data**: Gender (Male/Female), Age (Number), Patient's Federal Unit.
- **Clinical Information**: ICD code (ICD-10), Blood pressure (mmHg), Glasgow Coma Scale (3-15), Hematocrit (%), Hemoglobin (g/dL), Leukocytes (unitscells/mm^3), Lymphocytes (%), Urea (mg/dL), Creatinine (mg/dL), Platelets ($10^3/\mu L$), Diuresis (mL/day).

The two output variables are the Clinical outcome (Death or Survival) and the Length of Hospital Stay (in days).

3.2 Dataset Statistics

To guide model training effectively, we began by exploring the dataset through frequency tables for categorical variables and descriptive statistics for numerical features, as shown in the tables below.

From Table 1, we observe a nearly equal distribution of male and female patients, while paediatric cases are rare and could introduce bias in model predictions. The dataset is imbalanced regarding outcomes, with 75.71% of patients surviving. Request and admission bed types align closely, suggesting most patients received the bed type requested. As shown in Table 2, 72.69% of admissions occurred at just two health units: 1ª SÃO JOSÉ DE MIPIBU and 2ª MOSSORÓ. Table 3 highlights clinical statistics, Patient Age, and LoS. Clinical features show extreme outliers, likely due to noise or measurement errors. LoS ranges widely (0–607 days) with high variability, complicating precise prediction. Notably, patient age is also broadly distributed, reflecting the dataset's demographic diversity and adding valuable heterogeneity for model learning.

Table 1. Frequency distribution of binary categorical variables.

Variable	**0**	**1**
Patient Gender (0=Female, 1=Male)	3898	4150
Request Type (0=Paediatric, 1=Adult)	258	7790
Request Bed Type (0=Ward, 1=ICU)	4909	3139
Admission Bed Type (0=Ward, 1=ICU)	4891	3157
Outcome Type (0=Survival, 1=Death)	6094	1955

Given the distinct nature of categorical variables, we assessed their individual impact on the two outcome variables. Specifically, we analysed the frequency, mortality and LoS for each ICD code, as severe diseases are expected to increase mortality and hospital stay duration. The heatmap in Figure 1 highlights the top 20 deadliest diseases in the training set by absolute death count. Pneumonia (J18.9) was the leading cause of death, followed by urinary tract infection (N39.0), stroke, acute myocardial infarction, and other respiratory illnesses—many likely linked to COVID-19. The overall mortality rate was 24.28%, with average LoS of 15.73 days for deceased patients and 15.23 days for survivors. These findings underscore the importance of prompt diagnosis and treatment of infectious and cardiovascular diseases, which pose the greatest risks. Notably, similar LoS between survivors and non-survivors suggests that longer hospitalisation alone does not strongly predict mortality.

Analysis of categorical variables rather than ICD, revealed a mortality rate of 42% in ICU patients compared to 13% in Wards. Mortality rates were similar

Table 2. Frequencies of Admission Health Units and Patient Federal Units.

Admission Health Unit	**Frequency**
1ª SÃO JOSÉ DE MIPIBU	3408
2ª MOSSORÓ	2443
3ª JOÃO CAMARA	669
4ª CAICÓ	551
5ª SANTA CRUZ	540
6ª PAU DOS FERROS	200
7ª METROPOLITANA	165
8ª AÇU	73
Patient Federal Unit	**Frequency**
RN (Rio Grande do Norte)	7879
Rest	170

Table 3. Summary statistics for numerical variables.

Variable	**Mean**	**Median**	**Std**	**Min**	**Max**	**IQR**
patientAge	60.40	64.00	22.01	1	111	31.00
systolicPressure	123.40	120.00	80.85	5	6464	25.00
diastolicPressure	73.93	75.00	16.23	0	600	16.00
glasgowScale	13.06	15.00	3.69	3	15	2.00
hematocrit	0.35	0.35	0.08	0.01	0.97	0.10
hemoglobin	15.22	12.00	15.63	0.10	99	3.40
leucocytes	14.12	10.70	37.03	0	991	7.53
lymphocytes	0.17	0.15	0.12	0.01	0.98	0.14
urea	56.45	39.00	55.41	0	615	43.00
creatinine	2.35	1.04	4.67	0	94	1.11
platelets	256.55	238.00	126.59	1	999	140.00
diuresis	1131.76	900.00	1432.68	1	9999	500.00
lengthOfStay	15.36	9.00	21.77	0	607	13.00

across genders (around 25%). Mortality by federal unit tended to be higher for non-RN cases, likely due to low occurrence numbers and severe illness (>20%). Average LoS showed little variation by Bed Type or Gender (16.2 days in Wards, 14 in ICU; 14.8 days for Females, 15.9 for Males). LoS and mortality rates increased in federal units with fewer cases, mirroring trends seen by OT. Patient age was categorised into four groups (<18, 18–39, 40–59, 60+), confirming expected increases in mortality with age. Notably, the 18–39 group exhibited the longest average LoS, influenced by outliers with stays exceeding 200 days. Further details, including OT and LoS by admission unit, are presented in the Figures 2 and 3.

In Figure 4, we have the percentage of death per the admission health unit of Rio Grande do Norte. Although 7th METROPOLITANA has the most cases, however it does not have the higher mortality (26%). 6a and 5a have higher

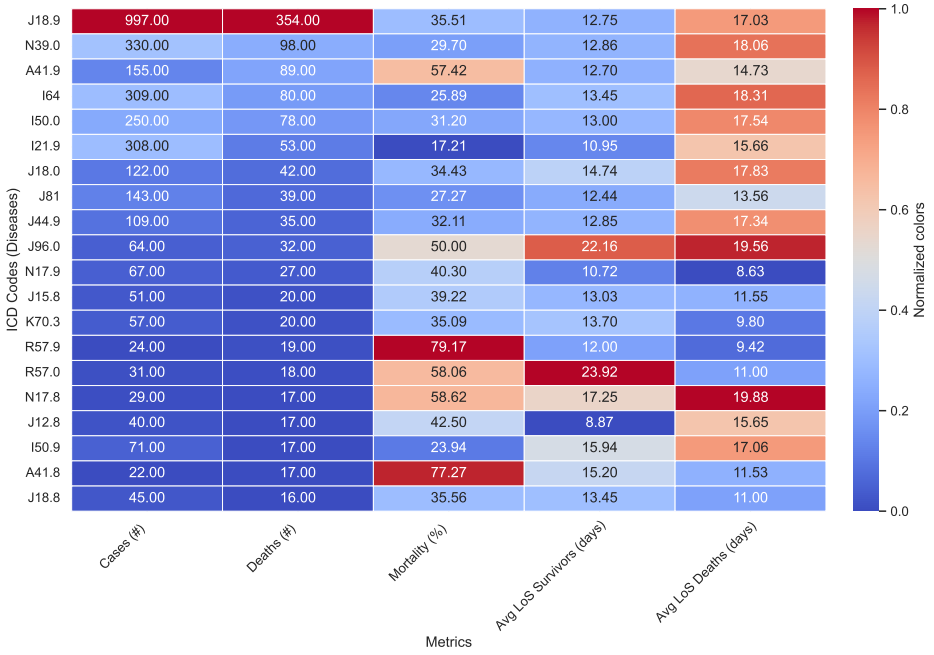

Fig. 1. Top-20 deadliest diseases by ICD code in the training dataset.

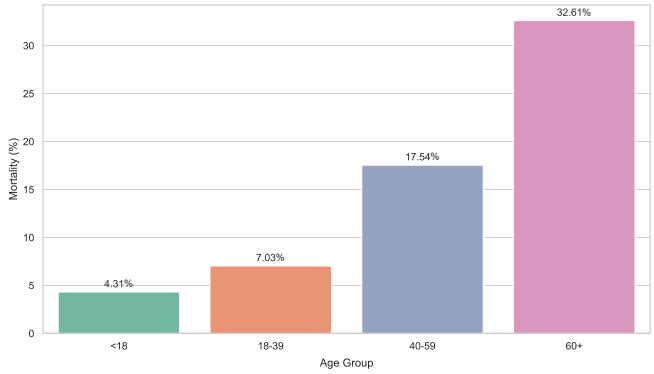

Fig. 2. Mortality percentage per age group.

mortality, while they have the fewer cases as well. That also makes sense, since the general mortality rate is 24%, so more cases, means more survivors and not more dead patients.

In Figure 5, we observed that the hospital with the highest number of cases also had the longest average LoS. This may be attributed to a larger number of

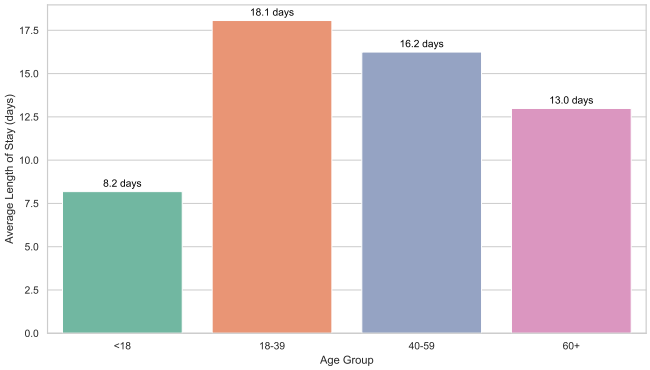

Fig. 3. Average LoS per age group.

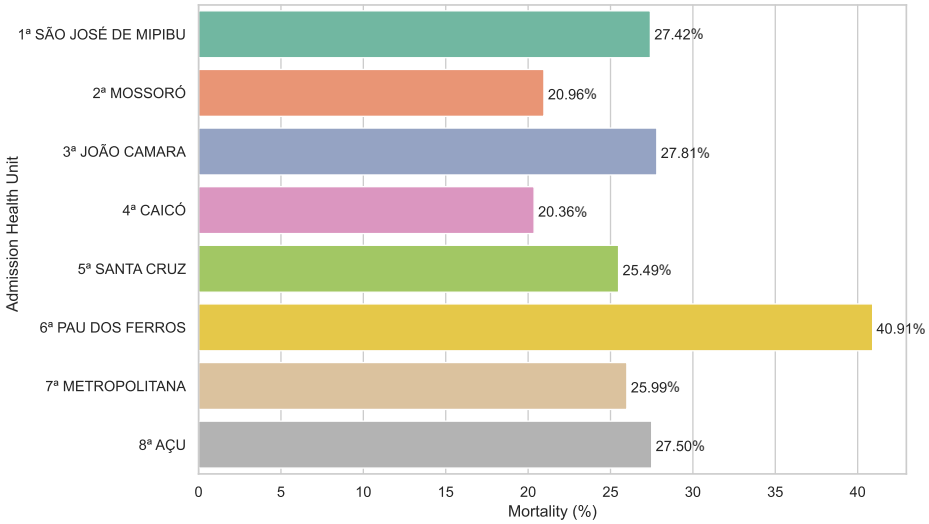

Fig. 4. Mortality percentage by Admission Health Unit in the training dataset.

patients who survived but required extended hospitalisation during their recovery process.

3.3 Correlation Analysis

An essential step before applying any predictive model is analysing the correlation between target variables and input features. To this end, we computed both linear (Pearson) and non-linear (Phik) correlations for OT and LoS against all input features. The Pearson correlation matrix (Figure 6) shows that the outcome variable has a moderate positive linear relationship with request and Admission Bed Type ($r = 0.33$ and $r = 0.32$), Patient Age ($r = 0.28$) and Urea

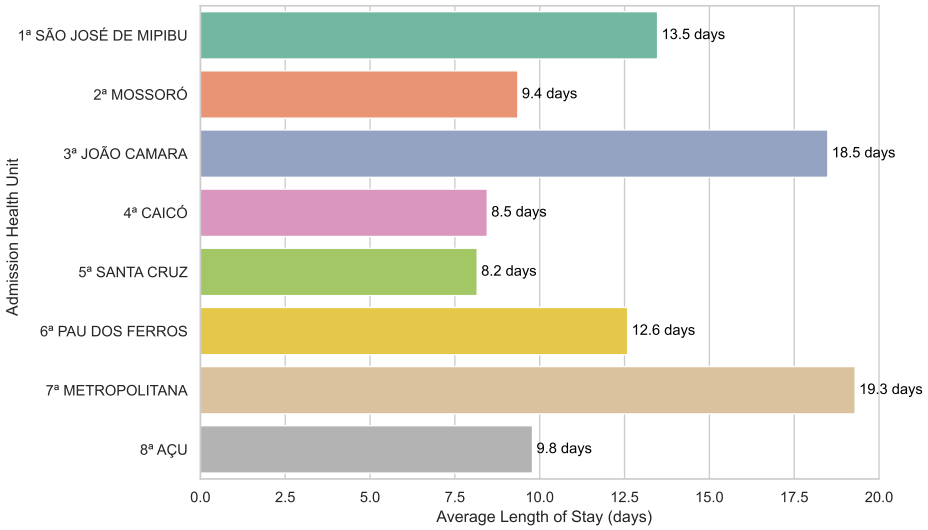

Fig. 5. LoS based on Admission Health Unit in train dataset.

levels ($r = 0.26$). Moderate negative correlations are observed with Lymphocytes ($r = -0.24$) and a strong negative correlation with the Glasgow Scale ($r = -0.57$). These results indicate that certain clinical and administrative variables influence patient outcomes, at least linearly. In contrast, LoS exhibits generally weak linear correlations with all features, the strongest being with days between Request and Admission ($r = 0.09$), underscoring the complexity of modelling this variable.

To capture both linear and non-linear associations, particularly involving categorical variables, we also calculated Phik correlations (Figure 7). Unlike Pearson's correlation, Phik is bounded between 0 and 1 and effectively handles categorical data without assuming linearity, reflecting the overall strength of association regardless of relationship type. For OT, Phik reveals strong associations with the Glasgow Scale ($\phi = 0.73$), Request Bed Type ($\phi = 0.50$) and Admission Bed Type ($\phi = 0.48$). Regarding LoS, Phik indicates slightly stronger but still weak associations, with the strongest links to ICD code ($\phi = 0.18$), and Admission Health Unit ($\phi = 0.09$). These findings suggest that while some non-linear or categorical associations with LoS exist, their overall strength remains low, highlighting the diverse patient characteristics in the training sample and the inherent challenge in accurately predicting Length of Hospital Stay.

3.4 Pre-Processing

As noted earlier, clinical features exhibit significant skewness and extreme outliers, and all categorical variables require appropriate numeric encoding before

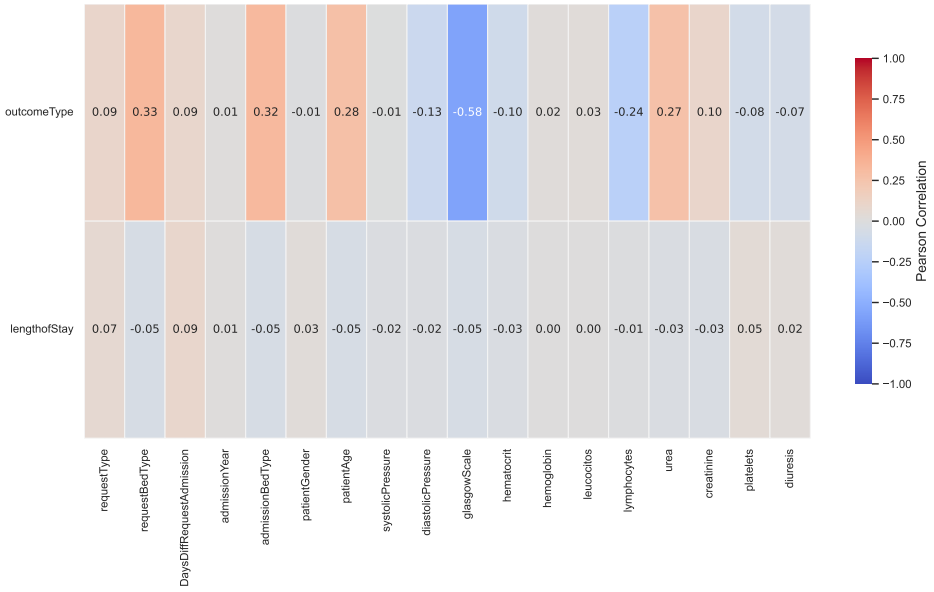

Fig. 6. Pearson correlation matrix.

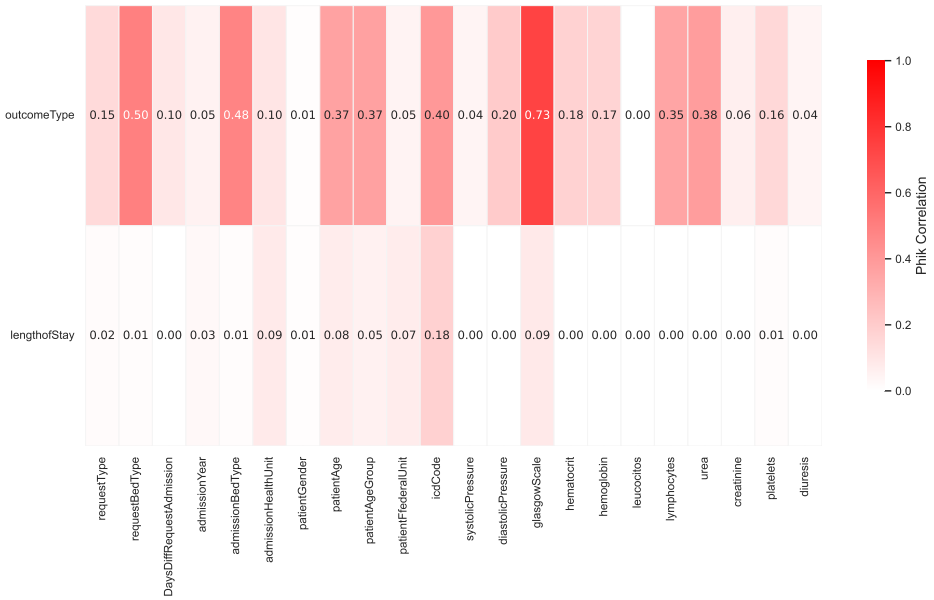

Fig. 7. Phik correlation matrix.

model training. This section outlines the pre-processing steps applied to improve data quality and ensure clean, consistent inputs.

Variable Encoding: Categorical variables were encoded based on their characteristics. Binary encoding was applied to features with two categories (e.g., Patient Gender, Request Type, Request and Admission Bed Type). For high-cardinality features (e.g., ICD code, Patient Federal Unit), Frequency encoding was used. Categorical variables with a small number of categories (e.g., Admission Health Units) were One-Hot encoded. The Glasgow Coma Scale was grouped into severity levels: mild (3–8), moderate (9–12), and high (13–15).

Feature Extraction: Features were derived from admission and request dates, including the interval between request and admission, admission weekday, and admission month. Blood pressure was separated into systolic and diastolic components. To facilitate age-based analysis, patient age was categorised into four clinically relevant groups: 0 – Children (<18 years), 1 – Young adults (18–39 years), 2 – Middle-aged adults (40–59 years), and 3 – Elderly (60+ years). For classification and most regression models (e.g., Lasso, RF, XGBoost), composite features were constructed by combining related clinical measurements for each patient. Three composite features were defined: hematocrit and hemoglobin formed an "oxygenation" feature [14], leukocytes and lymphocytes an "immune" feature [15] and urea and creatinine a "renal" feature [16]. For each patient, the values of the variables within a group were averaged to produce a single composite feature. This process was performed separately within each combination of age group and gender, allowing the aggregated values to reflect both the physiological relationships among variables and the demographic context in which they were measured.

Threshold Cleaning: We applied thresholding to clinical laboratory and vital sign features (e.g., Blood Pressure, Hematocrit, Hemoglobin) by defining clinically informed lower and upper bounds for each one of them (Table 4). Any values falling outside these ranges were capped at the nearest boundary. For example, abnormally high or low diuresis measurements were constrained within medically acceptable limits.

Clinical Categorisation: The physiological characteristics, including Hemoglobin and Platelets, were grouped based on the age group and gender of the patients. Clinical thresholds were adjusted for each subgroup (paediatric males, adult females), and characteristics were categorised as low, normal, or high accordingly.

Missing values: In numeric columns we fill the missing values (NaN) with the mean value of each column.

Scaling: For Regression models, both log and Yeo-Johnson transformations were applied to the LoS variable, with the Yeo-Johnson transformation providing a better fit due to its ability to handle zero and positive values, enhancing prediction accuracy. Numeric variables in models such as Lasso, Light Gradient Boosting (LigthGBM), RF, Ridge, XGBoost, and MLP were scaled using the

Table 4. Clinical thresholds for each numerical variable.

Variable	Category	Threshold
Systolic BP (mmHg) [17]	Hypotension	< 90
	Normal	90–119
	Elevated	120–129
	Hypertension Stage 1	130–139
	Hypertension Stage 2	140–179
	Hypertensive crisis	≥ 180
Diastolic BP (mmHg) [17]	Hypotension	< 60
	Normal	60–79
	Hypertension Stage 1	80–89
	Hypertension Stage 2	90–119
	Hypertensive crisis	≥ 120
Hematocrit (%) [14]	Critical low	< 30
	Low	30–40 (Male); 30–36 (Female)
	Normal	40–54 (Male); 36–48 (Female)
	High	> 54 (Male); > 48 (Female)
Hemoglobin (g/dL) [14]	Critical low	< 10
	Low	10–14 (Male); 10–12 (Female)
	Normal	14–18 (Male); 12–16 (Female)
	High	> 18 (Male); > 16 (Female)
Creatinine (mg/dL) [14]	Low	< 0.6 (Male); < 0.5 (Female)
	Normal	0.6–1.2 (Male); 0.5–1.1 (Female)
	High	1.2–2.0 (Male); 1.1–2.0 (Female)
	Critical high	≥ 2.0
Platelets ($\times 10^3/\mu$L) [18]	Any, Pediatric (< 18), Low	< 165
	Any, Pediatric (< 18), Normal	165–473
	Any, Pediatric (< 18), High	> 473
	Female, Adult (18–39), Low	< 136
	Female, Adult (18–39), Normal	136–436
	Female, Adult (18–39), High	> 436
	Female, Adult (40–59), Low	< 136
	Female, Adult (40–59), Normal	136–436
	Female, Adult (40–59), High	> 436
	Male, Adult (18–39), Low	< 120
	Male, Adult (18–39), Normal	120–369
	Male, Adult (18–39), High	> 369
	Male, Adult (40–59), Low	< 120
	Male, Adult (40–59), Normal	120–369
	Male, Adult (40–59), High	> 369
	Female, Adult (> 60), Low	< 119
	Female, Adult (> 60), Normal	119–396
	Female, Adult (> 60), High	> 396
	Male, Adult (> 60), Low	< 112
	Male, Adult (> 60), Normal	112–361
	Male, Adult (> 60), High	> 361
Leucocytes (Unitscells/mm^3) [19]	Low	< 4.5
	Normal	4.5–11.0
	High	11.0–20.0
	Very high	> 20.0
Urea (mg/dL) [14]	Low	< 5
	Normal	5–20
	High	20–40
	Very high	> 40
Lymphocytes (%) [20]	Very low	< 10
	Low	10–20
	Normal	20–40
	High	> 40
Diuresis (mL/day) [21]	Anuria	< 100
	Oliguria	100–800
	Normal	800–2500
	Polyuria	> 2500

RobustScaler, which is resilient to outliers by leveraging the median and interquartile range. Binary variables and certain categorical features—including LoS, Patient Age, and days between request and admission, were excluded from scaling. For classification models, specific pre-processing steps included One-Hot Encoding of blood measurements and the Glasgow Coma Scale in XGBoost. Additionally, the MLP classification model categorised LoS into clinically meaningful bins to improve classification performance.

4 Methods and Evaluation

This section outlines the modelling approach used to predict two clinical outcomes: patient discharge type (survival or death) and hospital LoS. We first describe the ML and DL models employed for both classification and regression tasks. Next, we detail the hyperparameter optimisation process applied to each model, following best practices established in prior studies [4, 5]. Finally, we present the evaluation metrics used to assess model performance, including both standard and domain-specific measures.

4.1 Machine Learning Models

For both regression and classification tasks, we employed a range of traditional ML models – including Histogram Gradient Boosting (HistGBM) [22], Logistic Regression [23], RF [24], TabNet [25], XGBoost [26], CatBoost [27], Lasso/Ridge regression [28], and LightGBM [29] – chosen for their robustness with mixed feature types and strong empirical performance. Additionally, we implemented Neural Network models, specifically MLP, capable of capturing complex non-linear relationships through deep architectures and modern optimisation techniques. Each model was selected to represent diverse learning paradigms and to establish a comprehensive baseline for performance comparison. All models were subjected to systematic hyperparameter tuning using Python's Optuna tool [30].

4.2 Hyperparameter Optimisation

In Table 5, we depict all the combinations of hyperparameters that were feeded into the optimiser in order to find the best parameters of each model, both for Regression and Classification. For the tree models like XGBoost, RF, LightGBM, HistGBM and CatBoost a variety of parameters is tried with regards to each algorithm like number of estimators (number of trees in the ensemble), max depth of tree, max iterations (number of trees sequentially added to the prediction), learning rate (the individual contribution of each tree), sub-sample (% of train data rows), col-sample/max features (% of train data columns), min samples split (minimum samples required to split a node), min samples leaf (minimum samples required to be at an end node), gamma (min loss reduction to split a node), l2 leaf regularisation/bootstrap (large weights are penalised, prevents overfitting). Lasso and Ridge regression parameters are alpha (weights

Table 5. Selection of hyperparameters and tested ranges for each model.

Model	Hyperparameter	Range Tested
MLP	Hidden layers	[1–6]
	Neurons per layer	[8–256]
	Dropout rate	[0.0–0.5]
	Activation function	[tanh, relu, sigmoid]
	Learning rate	[0.00001–0.01]
	Batch size	[16–128]
	Epochs	[10–100]
Lasso Regression	Alpha	[0.00001–1]
	Selection	[cyclic, random]
	Max iterations	[1000–5000]
	Tolerance	[0.000001–0.001]
LightGBM	Number of estimators	[100–300]
	Max depth	[3–15]
	Learning rate	[0.01–0.3]
	Number of leaves	[15–150]
XGBoost	Number of estimators	[50–1500]
	Max depth	[3–20]
	Subsample	[0.7–1]
	Colsample by tree	[0.7–1]
	Gamma	[0–5]
	Max features	[sqrt, log2]
Ridge Regression	Alpha	[0.00001–10]
	Solver	[auto, svd, cholesky, lsqr]
	Max iterations	[1000–5000]
	Tolerance	[0.000001–0.001]
TabNet	Decision layer size	[8–64]
	Attention embedding size	[8–64]
	Sequential steps	[3–10]
	Gamma	[1–2]
	Lambda sparse	[0.00001–0.01]
	Batch size	[64–512]
CatBoost	Max iterations	[100–5000]
	Learning rate	[0.01–0.2]
	Max depth	[3–12]
	L2 leaf regularization	[1–5]
RF	Number of estimators	[100–2000]
	Max depth	[3–30]
	Min samples split	[2–10]
	Min samples leaf	[1–4]
	Bootstrap	[True, False]
HistGBM	Max iterations	[100–300]
	Learning rate	[0.01–0.2]
	Max depth	[3–7]
	Min samples leaf	[20–100]

to the input feature, prevents overfitting), max iterations/tolerance (no closed form solution, needs these 2 convergence criteria), selection (method to update coefficients in Lasso) and solver (optimisation algorithm in Ridge regression). For the MLP we tested for the number of hidden layers and neurons per layer, dropout rate (% of neurons to be deactivated, prevents overfitting), activation functions for the hidden layers, learning rate (step size in gradient descend for the optimisation of weights), batch size (number of samples in one pass before weights update) and number of epochs (one full pass of the training dataset).

For the MLP configuration, Optuna initially suggested a relatively large architecture with six hidden layers of 256 neurons each, a dropout rate of 0.3, ReLU activation, 100 training epochs, and a batch size of 112. However, in practice, significantly smaller networks—comprising only 2–3 layers with 16–32 neurons each—yielded superior performance on the validation set, indicating that overly complex architectures may lead to overfitting in this context.

For the tree-based classifiers, optimal hyperparameters typically included a maximum depth between 8 and 10, 200–300 estimators, and a learning rate around 0.05. The best configurations also featured a sub-sample ratio of 0.74, column sampling by tree of 0.94, and a gamma value of 2. Regarding sample control, the minimum samples for split and leaf were 10 and 3, respectively, while bootstrap sampling was consistently set to False.

For linear models such as Ridge and Lasso regression, the optimal alpha was approximately 0.01, with the SVD solver performing best. The maximum number of iterations ranged from 2200 to 2600, and the optimal tolerance was 0.0001. Notably, the hyperparameter configurations for the tree-based models were largely consistent across both the classification and regression tasks, highlighting the robustness of these settings.

4.3 Evaluation Metrics

The primary objective of this study is to assess the performance of our predictive models across two target variables. To this end, we employ a set of well-established evaluation metrics tailored to each task. For the classification task, we report Accuracy and F1-Score, the latter of which incorporates both precision and recall to provide a balanced measure of model performance. For regression, we utilise the MAE and Root Mean Squared Error (RMSE) to quantify prediction errors. Additionally, we utilise two custom, domain-specific metrics: the Discharge Type Score (DTscore) and the LoS Type Score (LSscore), which offer further insight into the overall effectiveness of our methods within the specific clinical context.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{6}$$

$$\text{DTscore} = 10 \cdot (1 - F_1) \tag{7}$$

$$\text{LSscore} = \min \left(10,\ \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \right) \tag{8}$$

For hyperparameter optimisation in the classification task, the F1-score was employed as the objective metric, with class weighting applied to address class imbalance. The F1-score represents the harmonic mean of precision (the proportion of true positives among predicted positives) and recall (the proportion of true positives among actual positives). It is bounded between 0 and 1, with higher values indicating better model performance. While Accuracy measures the overall proportion of correct predictions, it is often misleading in imbalanced datasets due to its bias toward the majority class. In the regression task, the MAE was chosen as the optimisation criterion because it is less sensitive to outliers compared to the RMSE, thereby providing a more robust evaluation of prediction errors.

4.4 Code Availability

To support reproducibility and further experimentation, all source code, configuration files, and instructions used in this study are publicly available at: `https://github.com/MSc-BMI-DUTH-ATHENA/WC2025`.

5 Results

Although the hyperparameter optimisation process described in the previous section provides valuable insights into configuring the algorithms to minimise prediction errors, additional fine-tuning and manual experimentation around the optimal parameter ranges are necessary to finalise the models. In the following tables, we present the evaluation scores for the metrics introduced earlier, showcasing the best-performing runs for both classification and regression tasks on the training and validation datasets. These results reflect a combination of Optuna's automated suggestions and targeted manual adjustments to enhance model performance.

5.1 Model Performance on Training and Validation Sets

In the classification task (Table 6), most models demonstrated strong performance, with F1-scores ranging from 0.6523 to 0.7941 on the validation set. The HistGBM model achieved the highest F1-score of 0.7941, outperforming both XGBoost (0.7783) and RF (0.7749). In contrast, the MLP model yielded the lowest F1-score of 0.6523. These findings indicate that while MLPs are powerful for capturing complex, non-linear patterns in large datasets, their inherent complexity may limit their ability to generalise effectively in smaller datasets.

Table 6. Classification models performance on training and validation sets.

Model	Training Set			Validation Set		
	Accuracy	F1-score	DTscore	Accuracy	F1-score	DTscore
HistGBM	0.8231	0.8100	1.8998	0.8069	0.7941	2.0586
XGBoost	0.8101	0.7939	2.0613	0.7943	0.7783	2.2170
RF	0.8177	0.7904	2.0963	0.8017	0.7749	2.2512
TabNet	0.7889	0.7847	2.1532	0.7734	0.7693	2.3070
Logistic Regression	0.7695	0.7833	2.1669	0.7544	0.7679	2.3205
MLP	0.7721	0.6653	3.3467	0.7570	0.6523	3.4771

Table 7. Regression models performance on training and validation sets.

Model	Training Set			Validation Set		
	MAE	RMSE	LSscore	MAE	RMSE	LSscore
RF	8.9299	16.1600	8.9299	9.2177	16.6850	9.2177
HistGBM	8.9843	16.2789	8.9843	9.2915	16.8380	9.2915
CatBoost	9.0216	16.2489	9.0216	9.3265	16.8010	9.3265
Ridge Regression	9.0403	16.3585	9.0403	9.3265	16.8890	9.3265
LightGBM	9.2344	18.9511	9.2344	9.3101	19.1100	9.3101
XGBoost	9.0842	16.2833	9.0842	9.3869	16.8230	9.3869
TabNet	9.3049	16.3984	9.3049	9.5949	16.9040	9.5949
Lasso Regression	9.6392	17.1754	9.6392	9.8040	17.4720	9.8040
MLP	9.2136	16.3350	9.2136	10.2721	18.2120	10.0000

In the regression task (Table 7), model performance was generally limited, reflecting the weak correlation between features and the target variable, as discussed in Section 3. The difficulty of predicting the exact LoS, a highly skewed variable with most patients hospitalised for 20–30 days and few exceeding 100 days, further challenged model accuracy. Although our setup excludes the simplification of binning, the achieved MAE of 9–10 days may still be acceptable given the full range of LoS (0–607 days). On the validation set, all models performed comparably, with tree-based approaches showing slightly better results. RF yielded the lowest MAE (9.2177), followed closely by HistGBM (9.2915) and

CatBoost (9.3265). In contrast, the MLP model showed the weakest performance (MAE = 10.2721), in alignment with its limited classification performance. This suggests that the added complexity of MLPs may lead to overfitting in regression, especially when the underlying patterns are weak or nonlinearities are insufficiently informative for generalisation.

5.2 Cross-Validation Results and Model Robustness

To ensure the robustness and generalisability of our models, we performed 10-fold cross-validation on the combined training and validation sets using the optimal hyperparameters identified through Optuna. This methodology removes the need for further manual tuning and enables a comprehensive assessment of model stability across diverse, unseen data splits. Tables 8 and 9 present the average validation performance across folds for each classification and regression model, respectively, based on the three key evaluation metrics described previously.

Table 8. 10-fold cross-validation average scores for classification.

Model	Accuracy	F1-score	DTscore
HistGBM	0.8131	0.8032	1.9675
TabNet	0.8067	0.7941	2.0588
XGBoost	0.7995	0.7906	2.0936
RF	0.8042	0.7828	2.1716
Logistic Regression	0.7618	0.7429	2.5710
MLP	0.7571	0.6524	3.4759

Table 9. 10-fold cross-validation average scores for regression.

Model	MAE	RMSE	LSscore
HistGBM	9.5009	17.1489	9.3991
RF	9.6024	17.4681	9.5837
CatBoost	9.6101	17.5991	9.5623
TabNet	9.7179	17.5417	9.6670
Ridge Regression	9.8185	17.8283	9.7859
XGBoost	9.7672	17.5080	9.7297
MLP	9.8550	17.8010	9.7839
LightGBM	9.9103	20.5662	9.7036
Lasso Regression	10.169	18.3714	9.8914

The evaluation scores remained consistent across folds, demonstrating low sensitivity to data partitioning and underscoring the robustness and generalisability of the models. While regression performance was generally modest, the

classification task showed strong average F1-scores across all models, reflecting the effectiveness of the optimised hyperparameters. Due to the minimal differences in scores, 10-fold cross-validation slightly altered the relative ranking of models compared to single optimal runs. HistGBM achieved the highest average F1-score (0.8032), reaffirming its superior performance, followed closely by TabNet (0.7941) and XGBoost (0.7906). In contrast, MLP consistently underperformed relative to the other models, as previously discussed, with an F1-score of 0.6524. In the regression task, most models exhibited comparable results, with MLP (MAE = 9.855), LightGBM (9.9103), and Lasso Regression (10.169) showing slightly higher errors. HistGBM (9.5009) and RF (9.6024) achieved the lowest MAE, consistent with prior evaluations. We further explored a residual learning approach by fine-tuning XGBoost on the prediction errors of a base RF regressor. Although this hybrid strategy improved training performance, it consistently led to higher validation MAE, often exceeding 10, indicating overfitting and limited generalisation to unseen data.

5.3 Additional Exploratory Experiments

In addition to our primary modelling pipeline, we explored several strategies to enhance performance; however, these alternatives did not produce consistent improvements and were ultimately excluded.

To address class imbalance in the classification task, we applied CTGAN to synthetically generate additional samples for the minority "death" class. Initial tests with 500 synthetic samples slightly degraded both the F1-score and MAE. Scaling the generation to 8000 synthetic cases creating a balanced 50/50 distribution, similarly failed to improve performance, indicating that CTGAN-based augmentation was not effective in our context.

In the regression setting, we experimented with dynamic quantile selection in CatBoost. Specifically, we trained a meta-regressor to predict the expected error for each α value in the Quantile Loss Function (ranging from 0.01 to 0.99), allowing the model to select the best quantile for each sample during inference. Despite its theoretical appeal, this approach did not outperform a single CatBoost model trained with a fixed α and was therefore not retained.

6 Conclusions

Predicting patient OT and LoS at an early stage remains a complex yet crucial task for healthcare stakeholders, including clinicians, administrators, and patients. The COVID-19 pandemic highlighted significant challenges faced by hospitals worldwide, such as those in Rio Grande do Norte, Brazil, where surges in demand strained resources. Consequently, advanced technological solutions and enhancements to existing systems are essential to improve healthcare efficiency and resource management. The IFMBE Scientific Challenge 2025 presented specific obstacles – heterogeneous patient populations, skewed LoS distributions with extreme outliers, imbalanced data, and weak feature-target correlations –

that limited the predictive accuracy of ML and DL models for precise LoS estimation. In contrast, OT prediction proved more tractable, likely due to stronger associations with clinical features such as ICD-coded diagnoses, facilitating identification of patients at higher risk of adverse outcomes.

Future research should focus on expanding the dataset significantly by incorporating a larger number of patient records and additional clinical features to uncover deeper and potentially novel associations with the prediction targets. Moreover, conducting cross-sectional studies across multiple countries would enhance the generalisability of findings related to patient OT and LoS. As highlighted in the literature review, alternative modelling approaches, such as categorising LoS into clinically relevant groups (e.g., short, medium, and prolonged stays), may better support resource allocation and hospital management decisions. Furthermore, integrating ML and Artificial Intelligence models into existing hospital information systems, such as RegulaRN, combined with real-time patient vital sign monitoring [31, 32], could substantially improve the accuracy and timeliness of critical predictions, thereby enhancing both patient care and operational efficiency.

In conclusion, advancing predictive modelling through richer data, broader validation, and real-time integration holds considerable promise for transforming hospital management and patient outcomes in increasingly complex healthcare environments.

Acknowledgement. This work was carried out in the context of the Inter-Institutional Master's Program "Biomedical Informatics" with the support of the School of Medicine, Democritus University of Thrace and the Athena Research Center in Greece.

References

[1] Almeida, G., Brito Correia, F., Borges, A.R., Bernardino, J.: Hospital length-of-stay prediction using machine learning algorithms—A literature review. Applied Sciences **14**(22) (2024). DOI 10.3390/app142210523

[2] Santos, P.P.G.V.D., Oliveira, R.A.D.D., Albuquerque, M.V.D.: Inequalities in the provision of hospital care in the Covid-19 pandemic in Brazil: An integrative review. Saúde em Debate **46**(spe1), 322–337 (2022). DOI 10.1590/0103-11042022e122i

[3] Gomes, S., Silva, A.L.N.D., Segatto, C.I., Santos, A.: The coordination role of Rio Grande do Norte state government in response to COVID-19: Innovation in times of crisis? Saúde e Sociedade **31**(4) (2022). DOI 10.1590/s0104-12902022210523en

[4] Barreto, T.D.O., Veras, N.V.R., Cardoso, P.H., Fernandes, F.R.D.S., Medeiros, L.P.D.S., Bezerra, M.V., Andrade, F.M.Q.D., Pinheiro, C.D.O., Sánchez-Gendriz, I., Silva, G.J.P.C., Rodrigues, L.F., Morais, A.H.F.D., Dos Santos, J.P.Q., Paiva, J.C., Andrade, I.G.M.D., Valentim, R.A.D.M.: Artificial intelligence applied to analyzes during the pandemic: COVID-19

beds occupancy in the state of Rio Grande do Norte, Brazil. Frontiers in Artificial Intelligence **6** (2023). DOI 10.3389/frai.2023.1290022

[5] Barreto, T.D.O., Farias, F.L.D.O., Veras, N.V.R., Cardoso, P.H., Silva, G.J.P.C., Pinheiro, C.D.O., Medina, M.V.B., Fernandes, F.R.D.S., Barbalho, I.M.P., Cortez, L.R., Santos, J.P.Q.D., Morais, A.H.F.D., Souza, G.F.D., Machado, G.M., Lucena, M.J.N.R., Valentim, R.A.D.M.: Artificial intelligence applied to bed regulation in Rio Grande do Norte: Data analysis and application of machine learning on the "RegulaRN Leitos Gerais" platform. PLOS ONE **19**(12), e0315,379 (2024). DOI 10.1371/journal.pone.0315379

[6] Mekhaldi, R.N., Caulier, P., Chaabane, S., Chraibi, A., Piechowiak, S.: A comparative study of machine learning models for predicting length of stay in hospitals. Journal of Information Science and Engineering **37**(5) (2021). DOI 10.6688/JISE.202109_37(5).0003

[7] Mekhaldi, R.N., Caulier, P., Chaabane, S., Chraibi, A., Piechowiak, S.: Using machine learning models to predict the length of stay in a hospital setting. In: Advances in Intelligent Systems and Computing, pp. 202–211. Springer International Publishing, Cham (2020). DOI 10.1007/978-3-030-45688-7_21

[8] Zeleke, A.J., Palumbo, P., Tubertini, P., Miglio, R., Chiari, L.: Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department. Informatics in Medicine Unlocked **47**, 101,499 (2024). DOI 10.1016/j.imu.2024.101499

[9] Jain, R., Singh, M., Rao, A.R., Garg, R.: Predicting hospital length of stay using machine learning on a large open health dataset. BMC Health Services Research **24**(1) (2024). DOI 10.1186/s12913-024-11238-y

[10] Boff Medeiros, N., Fogliatto, F.S., Karla Rocha, M., Tortorella, G.L.: Predicting the length-of-stay of pediatric patients using machine learning algorithms. International Journal of Production Research **63**(2), 483–496 (2025). DOI 10.1080/00207543.2023.2235029

[11] Shamout, F., Zhu, T., Clifton, D.A.: Machine learning for clinical outcome prediction. IEEE Reviews in Biomedical Engineering **14**, 116–126 (2021). DOI 10.1109/rbme.2020.3007816

[12] Iwase, S., Nakada, T.a., Shimada, T., Oami, T., Shimazui, T., Takahashi, N., Yamabe, J., Yamao, Y., Kawakami, E.: Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports **12**(1) (2022). DOI 10.1038/s41598-022-17091-5

[13] Alghatani, K., Ammar, N., Rezgui, A., Shaban-Nejad, A.: Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. JMIR Medical Informatics **9**(5), e21,347 (2021). DOI 10.2196/21347

[14] Clark, V.L., Kruse, J.A.: Clinical methods: The history, physical, and laboratory examinations. JAMA **264**(21), 2808–2809 (1990). DOI 10.1001/jama.1990.03450210108045

[15] Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.: Adaptive immunity to infection. In: Immunobiology: The Immune System in Health and Disease, 5 edn., chap. 10. Garland Science, New York (2001)

[16] Yu, A.S.L., Chertow, G.M., Luyckx, V.A., Marsden, P.A., Skorecki, K., Taal, M.W.: Brenner and Rector's The Kidney, 2-Volume Set, 11 edn. Elsevier (2019)

[17] Reboussin, D.M., Allen, N.B., Griswold, M.E., Guallar, E., Hong, Y., Lackland, D.T., Miller, E.P.R., Polonsky, T., Thompson-Paul, A.M., Vupputuri, S.: Systematic review for the 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. Journal of the American College of Cardiology **71**(19), 2176–2198 (2018). DOI 10.1016/j.jacc.2017.11.004

[18] Zaninetti, C., Biino, G., Noris, P., Melazzini, F., Civaschi, E., Balduini, C.L.: Personalized reference intervals for platelet count reduce the number of subjects with unexplained thrombocytopenia. Haematologica **100**(9), e338–e340 (2015). DOI 10.3324/haematol.2015.127597

[19] Williams-Reid, H., Johannesson, A., Buis, A.: Wound management, healing, and early prosthetic rehabilitation: Part 3 – A scoping review of chemical biomarkers. Canadian Prosthetics & Orthotics Journal **8**(1) (2025). DOI 10.33137/cpoj.v8i1.43717

[20] Pagana, K.D., Pagana, T.J., Pagana, T.N.: Mosby's Diagnostic and Laboratory Test Reference, 17 edn. Elsevier (2024)

[21] Cohen, R., Fernie, G., Roshan Fekr, A.: Fluid intake monitoring systems for the elderly: A review of the literature. Nutrients **13**(6) (2021). DOI 10.3390/nu13062092

[22] Guryanov, A.: Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In: Lecture Notes in Computer Science, pp. 39–50. Springer International Publishing, Cham (2019). DOI 10.1007/978-3-030-37334-4_4

[23] Shipe, M.E., Deppen, S.A., Farjah, F., Grogan, E.L.: Developing prediction models for clinical use using logistic regression: An overview. Journal of Thoracic Disease **11**(S4), S574–S584 (2019). DOI 10.21037/jtd.2019.01.25

[24] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). DOI 10.1023/a:1010933404324

[25] Wang, H., Ding, J., Wang, S., Li, L., Song, J., Bai, D.: Enhancing predictive accuracy for urinary tract infections post-pediatric pyeloplasty with explainable AI: An ensemble TabNet approach. Scientific Reports **15**(1) (2025). DOI 10.1038/s41598-024-82282-1

[26] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, San Francisco California USA (2016). DOI 10.1145/2939672.2939785

[27] Hancock, J.T., Khoshgoftaar, T.M.: CatBoost for big data: An interdisciplinary review. Journal of Big Data **7**(1) (2020). DOI 10.1186/s40537-020-00369-8

[28] Tibshirani, R.: Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society Series B: Statistical Methodology **73**(3), 273–282 (2011). DOI 10.1111/j.1467-9868.2011.00771.x

[29] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)

[30] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, p. 2623–2631. Association for Computing Machinery, New York, NY, USA (2019). DOI 10.1145/3292500.3330701

[31] Ting, L.P.Y., Chen, H.P., Liu, A.S., Yeh, C.Y., Chen, P.L., Chuang, K.T.: Early detection of patient deterioration from real-time wearable monitoring system (2025). DOI 10.48550/ARXIV.2505.01305. Version Number: 2

[32] Gabriel, P., Rehani, P., Troy, T., Wyatt, T., Choma, M., Singh, N.: Continuous patient monitoring with AI: Real-time analysis of video in hospital care settings. Frontiers in Imaging **4** (2025). DOI 10.3389/fimag.2025.1547166