Optimizing Resource Allocation in Public Healthcare: A Machine Learning Approach for Length-of-Stay Prediction

Paolo Perliti Scorzoni, Anita Giovanetti,
Federico Bolelli, and Costantino Grana

Università degli Studi di Modena e Reggio Emilia
{name.surname}@unimore.it

Abstract. Effective hospital resource management hinges on established metrics such as Length of Stay (LOS) and Prolonged Length of Stay (pLOS). Reducing pLOS is associated with improved patient outcomes and optimized resource utilization (e.g., bed allocation).
This study investigates several Machine Learning (ML) models for both LOS and pLOS prediction. We conducted a retrospective study analyzing data from general inpatients discharged between 2022 and 2023 at a northern Italian hospital. Sixteen regression and twelve classification algorithms were compared in forecasting LOS as either a continuous or multi-class variable (1-3 days, 4-10 days, >10 days). Additionally, the same models were assessed for pLOS prediction (defined as LOS exceeding 8 days). All models were evaluated using two variants of the same dataset: one containing only structured data (e.g., demographics and clinical information), and a second one also containing features extracted from free-text diagnosis. Ensemble models, leveraging the combined strengths of multiple ML algorithms, demonstrated superior accuracy in predicting both LOS and pLOS compared to single-algorithm models, particularly when utilizing both structured and unstructured data extracted from diagnoses.
Integration of ML, particularly ensemble models, has the potential to significantly improve LOS prediction and identify patients at high risk of pLOS. Such insights can empower healthcare professionals and bed managers to optimize patient care and resource allocation, promoting overall healthcare efficiency and sustainability.

Keywords: LOS · pLOS · Machine Learning · Hospital Admissions · Public Healthcare · Sustainability.

1 Introduction

The Italian public healthcare system faces a complex challenge in managing bed availability. The past few decades have witnessed a 30% reduction in hospital bed capacity [2], along with a rise in bed occupancy rates, leading to congestion and extended patient stays. Italy's current provision of 11.6 beds per 100,000

inhabitants falls below the OECD average of 16.9 [1]. This situation is further exacerbated by a significant increase in national health expenditure, rising from approximately 80 billion euros in 2002 to 129 billion euros in 2022 [12, 13], with 20% allocated to inpatient care.

Managing patient flow has become increasingly problematic due to factors such as rising patient volumes, an aging population with higher comorbidity rates, delayed discharges to non-acute care settings, and evolving working practices (as emphasized by the COVID-19 pandemic). Current bed modeling techniques, often relying on midnight census data, lack the granularity needed for optimal space management. A more comprehensive understanding of patient flow dynamics and peak occupancy patterns is essential. In this context, the emergence of Artificial Intelligence (AI) and Machine Learning (ML) offers promising tools to assist bed managers in their daily operations.

2 Background

Length of Stay (LOS), defined as the duration between hospital admission and discharge (i.e., total bed-days occupied by a patient), plays a fundamental role in assessing healthcare service quality. Previous research has demonstrated correlations between LOS and disease severity, readmission rates, and mortality [33, 25]. The reduction of LOS in public healthcare systems benefits both patients and hospitals: early discharge and faster turnover improve inpatient outcomes by preventing complications, reducing risk of adverse events (such as falls, thrombosis, drug reactions, and hospital-acquired infections) [3, 19], and promoting patient autonomy [17]; on the other hand, hospitals experience advantages from optimized treatment strategies, improved resource utilization (e.g., bed allocation), and better control over waiting lists [28]. Furthermore, LOS holds the advantage of uniform measurability, making it comparable even across different healthcare facilities on a global scale [26].

Conversely, Prolonged Length of Stay (pLOS) is associated with functional limitations, cognitive impairment, and a higher burden of comorbidities among patients [5]. Moreover, pLOS often results in cancellations of elective surgeries, increased resource utilization (including raising medical costs, especially in Intensive Care Units, ICUs), and potential delays in admitting critically ill patients. Notably, a small percentage of patients with pLOS can consume up to 50% of available resources [14, 36].

This holds particular relevance in Italy, where demographic shifts like increasing life expectancy (80.6 years for men and 84.8 years for women in 2022 [21]) contribute to an escalation in chronic and degenerative diseases. A timely identification of inpatients with extended stays (often referred to as "bed-blockers") is essential for formulating effective treatment plans. Thus, pLOS serves as a key metric, directly influencing healthcare expenditures and available capacity.

The present study aimed to develop ML-based models for predicting both LOS and pLOS in general patient populations. Several techniques, including regression, support vector machine, KNN, random forest, gradient boosting tree,

neural networks, and ensembles, were compared to identify the most effective models. Additionally, we investigated the most relevant features for accurate prediction.

3 Related Works

Over the past two decades, researchers have employed various statistical techniques to investigate LOS and the influence of covariates such as age, gender, illness severity, diagnosis, and hospital characteristics. More recently, Machine Learning and Deep Learning (DL) have emerged as promising alternatives to these established methodologies in healthcare research [4, 15]. Studies exploring LOS patterns exhibits considerable heterogeneity [35, 42], often focusing on broad patient cohorts [27], specific age ranges [20, 38], explicit discipline areas [24, 32] and medical specialties [9, 34], surgical procedures [8, 40] and oncological surgeries [16, 23], and individual hospital departments. However, only a limited portion of the existing literature addresses the specific context of the Italian public healthcare system.

Trunfio et al. [39] analyzed 2,515 patients undergoing hip-replacement surgery at the University Hospital of Salerno, Italy. Their analysis revealed that Multiple Linear Regression yielded the highest performance in predicting LOS ($R2 = 0.616$), whereas Random Forest and Gradient-Boosted Tree models achieved an accuracy of 71.76% in predicting LOS as a discrete target (less than 7 days, 7-12 days, over 12 days).

Olivato et al. [31] developed an ML-based system to predict pLOS in COVID-19 patients admitted to the "Spedali Civili" in Brescia. Their model, trained on demographic information and laboratory test results from over 6,000 admissions, attained a ROC-AUC score of 0.76.

Zeleke et al. [41] investigated 12,858 inpatients admitted through the emergency department of an Italian hospital in Bologna. Their Gradient Boosting classifier achieved an accuracy of 75% in predicting pLOS (defined as any stay exceeding 6 days), while Ridge and XGBoost regressors were most effective in forecasting LOS as a continuous outcome, with a prediction error ranging between 6 and 7 days.

In another Italian study by D'Onofrio et al. [11], the application of Random Forest achieved a 77.79% accuracy in predicting LOS for 989 patients undergoing mastectomy surgery at A.O.R.N. "Antonio Cardarelli" in Naples.

Di Matteo et al. [10] implemented an ML-driven system to forecast prolonged LOS (defined as LOS > 7 days) for hip-/knee-arthroplasty patients at "Humanitas Research" Hospital in Milan. Leveraging combined clinical and textual data on 1,517 patients, their model achieved an AUC of 0.789.

While many of these studies focus on specific departments or rely on data partially unavailable at admission (such as lab results), our research addresses this gap by employing various supervised ML algorithms to predict LOS for general inpatients using readily available data extracted from a medico-administrative

platform. We analyzed LOS as a continuous, multi-class, and dichotomous variable, encompassing all medical-surgical departments with the purpose of developing robust and adaptable models for effective generalization. This approach mirrors real-world scenarios where patients may be relocated to alternate wards (often regardless of their primary service) when a department reaches full capacity. Evaluating all medical units collectively provides greater consistency. Additionally, focusing on admission data ensures immediate implementation across diverse hospital settings.

4 Materials and Methods

4.1 Data Selection and Inclusion Criteria

This study was conducted at a general hospital located in Emilia-Romagna, Italy. The "Ospedale di Sassuolo SpA", guided by principles of intensity care, comprises 19 clinical units. We analyzed a dataset of 12,471 hospitalizations from 10,145 unique patients discharged between February 2022 and November 2023. All patients had a minimum stay of 24 hours. Data were extracted from the hospital's EBMS (Electronic Bed Management System), which included information on patient demographics, admission type, clinical features, and hospitalization details. A summary of patient characteristics is provided in Table 1.

To minimize potential biases in model performance, patients undergoing Day Surgery or Day Hospital procedures were excluded due to their predetermined LOS of one day. Additionally, to ensure data integrity, we also excluded patients deceased during hospitalization, inpatients with stays exceeding the 99.95th percentile of the LOS distribution (outliers), and maternity/infancy wards due to their distinct clinical characteristics and potential data collection biases.

We further expanded the initial dataset by integrating historical information regarding each patient's prior hospitalizations, including: the number of previous admissions in the last 12 months (particularly those requiring ICUs or high-intensity care levels), the average and total length of stay during previous hospitalizations, and the average LOS for all patients admitted within the same service as the current hospitalization in the preceding 30 days (in order to capture any trends for a given department).

4.2 Models Development

The research employed two variants of the available dataset. Dataset A contained only structured data, including demographics, clinical information, and admission details. Conversely, dataset B included additional features derived from unstructured free-text diagnoses documented by healthcare practitioners.

Initially, fourteen regression algorithms were implemented to predict LOS as a continuous variable. Performance evaluation metrics included mean absolute error (MAE), root mean squared error (RMSE), R-squared (R2), and adjusted R-squared scores. Additionally, ten classification methods were employed to predict LOS as a multi-class target (1-3 days, 4-10 days, >10 days). Evaluation

Table 1. Patient characteristics.

#	Feature	Total	Type
Patient demographics			
1	Seniority (age divided into 10-years groups)	12471	Categorical
2	Gender (M/F)	12471	Categorical
3	From outer province? (Y/N)	12471	Boolean
4	From outer administrative district? (Y/N)	12471	Boolean
Information from current admission			
5	Admission month (1-12)	12471	Categorical
6	Admission day of week (1-7)	12471	Categorical
7	Admission on weekend (Y/N)	12471	Boolean
8	Admission on working day (Y/N)	12471	Boolean
9	Admission hour of day (0-23)	12471	Categorical
10	Admission on night-time? (Y/N)	12471	Boolean
11	Admission from outer facility? (Y/N)	12471	Boolean
12	Intensity care (L/M/H)	12471	Categorical
13	Admission from ER? (Y/N)	12471	Boolean
14	Short-Stay Observation (SSO)? (Y/N)	12471	Boolean
15	Single room? (Y/N)	12471	Boolean
16	Bed type	12471	Categorical
Clinical information			
17	Terminal patient (End-of-Life)? (Y/N)	12471	Boolean
18	Bedridden patient? (Y/N)	12471	Boolean
19	Multidimensional geriatric assessment requested? (Y/N)	12471	Boolean
20	Integrated Home Care requested? (Y/N)	12471	Boolean
21	Isolation required? (Y/N)	12471	Boolean
22	COVID-19 isolation? (Y/N)	12471	Boolean
23	Contact isolation? (Y/N)	12471	Boolean
24	Structural isolation? (Y/N)	12471	Boolean
25	Other type of isolation? (Y/N)	12471	Boolean
26	Diagnosis (text)	12471	Text
Information from current hospitalization			
27	Hospitalization area	12471	Categorical
28	Hospitalization area type	12471	Categorical
29	Specialty (service)	12471	Categorical
30	Recent transfers count	12471	Numeric
31	Past transfers count	12471	Numeric
32	Movements count	12471	Numeric
33	ICU movements count	12471	Numeric
Information from previous hospitalizations			
34	Patient prev. hospitalizations count (prior 12 mo.)	12471	Numeric
35	Patient prev. ICU hospitalizations count (prior 12 mo.)	12471	Numeric
36	Patient prev. hosp. with high-intensity care count (prior 12 mo.)	12471	Numeric
37	Patient prev. hospitalizations average LOS (prior 12 mo.)	12471	Numeric
38	Patient prev. hospitalizations total LOS (prior 12 mo.)	12471	Numeric
39	Specialty prev. hospitalizations average LOS (prior 1 mo.)	12471	Numeric

metrics for classification encompassed accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUROC). The same ten classifiers were also used to predict pLOS, defined as any hospitalization exceeding eight days, corresponding to the 75th percentile.

To extract features from diagnoses, a text cleaning process was implemented, involving the removal of stop words and irrelevant or non-domain-specific terms. A pre-trained BERT base model was then employed to tokenize the text and generate embeddings. Subsequently, a data preprocessing pipeline was applied to normalize all numerical variables using a StandardScaler and to one-hot encode categorical features. This step ensured all features were on a comparable scale and that categorical features were numerically represented. Furthermore, principal component analysis (PCA) was applied to embeddings from dataset B

to reduce the dimensionality of the data to 100 components, thereby mitigating computational complexity.

Each dataset was randomly partitioned into a training set comprising 80% of the admissions (9,976) and a holdout/validation set encompassing the remaining 20% (2,495 admissions), using stratified sampling. A five-fold cross-validation approach was employed for each task to compare algorithms and isolate the top performers. Hyperparameter tuning was then conducted for each selected model. Finally, performance was assessed on the independent test set.

Moreover, Voting and Stacking ensemble methods were added in the final evaluation. These methods aggregate predictions from multiple base models (using voting and stacking aggregation techniques, respectively) to improve overall accuracy and reduce model bias.

Classification Models. According to the Italian Ministry of Health [18], the national average LOS for acute care in 2020 was 7.5 days. In Emilia-Romagna, the region in which the hospital subject of this study is located, the average LOS for acute care in 2020 was 7.6 days, closely aligned with national data. Guided by these benchmarks and the requirements of the hospital under investigation, it was decided to split LOS into three distinct groups:

- Group 1: LOS $\leq$ 3 days (5,830 hospitalizations)
- Group 2: 4 $\leq$ LOS $\leq$ 10 days (4,445 hospitalizations)
- Group 3: LOS $>$ 10 days (2,196 hospitalizations)

Furthermore, the chosen thresholds ensured a roughly even distribution of observations across groups. As a result, no data-balancing techniques were employed.

Binary Classification Models. Consistent with national trends, this study adopted an 8-day threshold to classify hospitalizations as either "short" or "prolonged". Instances of LOS falling within the range of 1 to 8 days accounted for 9,578 hospitalizations, while prolonged LOS (>8 days) represented 2,893 admissions (23.2% of all hospitalizations). Elderly age groups (70-79 years, 80-89 years, and 90-99 years) experienced longer LOS compared to younger cohorts, corresponding to 28%, 37%, and 12% of all prolonged stays, respectively. The majority of cases were observed in general medicine wards (39%) and long-term care (16%). Additionally, 67.3% of bed-blockers required medium-intensity care, while 26.3% and 6.4% required low- and high-intensity care, respectively.

The prevalence of nearly three times more instances of class 0 ("short" LOS) compared to class 1 ("prolonged" LOS) poses a potential challenge, as the model might overfit to the majority class (class 0) and struggle to accurately identify the minority class (class 1). To address this imbalance, we employed techniques like SMOTE and ADASYN to generate synthetic minority class instances, aiming to enhance model performance. However, these methods did not yield significant performance improvements on our dataset. This suggests that ensemble methods, which combine predictions from multiple learners, might be more effective for handling class imbalance, particularly for metrics beyond accuracy.

To maintain methodological coherence, the pLOS prediction task leveraged the same battery of classifiers and evaluation metrics employed for the multiclass LOS task.

Table 2. Results of tuned models on datasets A and B (regression task).

	Dataset A				Dataset B			
Model	**MAE ↓**	**RMSE ↓**	**R2 ↑**	**Ad. R2 ↑**	**MAE ↓**	**RMSE ↓**	**R2 ↑**	**Ad. R2 ↑**
Stacking Regressor	**2.806**	**4.614**	**0.635**	**0.633**	2.705	4.622	0.633	0.632
Voting Regressor	2.824	4.617	0.634	0.633	2.722	4.537	0.647	0.645
XGB Regressor	2.844	4.634	0.632	0.630	2.776	4.585	0.639	0.638
CatBoost Regressor	2.831	4.639	0.631	0.629	**2.726**	**4.520**	**0.649**	**0.648**
Linear Regression	2.976	4.692	0.622	0.621	2.911	4.622	0.633	0.632
Ridge	2.963	4.706	0.620	0.618	2.862	4.620	0.634	0.632
GB Regressor	2.946	4.723	0.617	0.616	2.923	4.739	0.615	0.613
LGBM Regressor	2.903	4.770	0.609	0.608	2.756	4.616	0.634	0.633
Elastic-net	2.995	4.776	0.609	0.607	2.887	4.679	0.624	0.623
SVR	2.883	4.784	0.607	0.606	2.792	4.694	0.622	0.620
Lasso	3.051	4.867	0.594	0.592	3.004	4.811	0.603	0.601
RF Regressor	2.968	4.903	0.588	0.586	2.843	4.787	0.607	0.605
KNN Regressor	2.989	5.012	0.569	0.567	2.868	4.957	0.578	0.577
AdaBoost Regressor	3.601	5.275	0.522	0.521	3.576	5.285	0.521	0.519
MLP Regressor	3.298	5.587	0.464	0.462	3.090	5.350	0.509	0.507
DT Regressor	3.882	6.617	0.249	0.246	3.757	6.604	0.252	0.249

5 Results

5.1 Regression Models

Among the sixteen regressors evaluated for predicting LOS as a continuous variable on dataset A (Table 2), the ensemble StackingRegressor achieved the strongest performance (MAE 2.81, R2 score 0.635), followed by VotingRegressor and XGBRegressor. When considering dataset B, which incorporated unstructured data, CatBoostRegressor emerged as the superior model (MAE 2.73, R2 score 0.649), followed closely by VotingRegressor and XGBRegressor.

The integration of embedded representations derived from free-text diagnoses resulted in a measurable, albeit slight, performance enhancement across all models. This improvement can be ascribed, at least partially, to the ability of embeddings to encapsulate the semantic meaning of diagnoses, a task that is challenging to accomplish solely through conventional structured features. In addition, embeddings offer the advantage of modeling the relationships among diverse diagnoses, which becomes particularly valuable in the presence of comorbidities. Importantly, including the admitting diagnosis does not introduce bias or confound the study endpoint (e.g., data leakage) as this information is inherently available at the time of hospitalization.

5.2 Classification Models

Among the classifiers employed to forecast LOS on dataset A, VotingClassifierSoft exhibited superior performance (accuracy 73.55%, F1-score 73.25%, AUROC 87.94%). The StackingClassifier and CatBoostClassifier trailed closely behind. As for dataset B, VotingClassifierSoft again emerged as the top performer (accuracy 76.27%, F1-score 75.96%, AUROC 89.60%), followed by CatBoostClassifier and StackingClassifier. Consistent with the findings from the regression analysis, the employment of embeddings derived from diagnoses yielded a modest improvement in performance (Table 3).

Table 3. Results of tuned models on datasets A and B (multi-class classification task).

	Dataset A				Dataset B			
Model	**Acc. ↑**	**F1 ↑**	**A.ROC ↑**	**A.PRC ↑**	**Acc. ↑**	**F1 ↑**	**A.ROC ↑**	**A.PRC ↑**
Voting Soft	**0.735**	**0.732**	**0.879**	**0.777**	**0.763**	**0.760**	**0.896**	**0.794**
Stacking	0.732	0.730	0.879	0.774	0.763	0.761	0.893	0.789
CatBoost	0.733	0.729	0.878	0.776	0.758	0.754	0.894	0.788
XGB	0.728	0.724	0.875	0.775	0.762	0.760	0.892	0.789
RF	0.730	0.728	0.873	0.767	0.750	0.745	0.886	0.778
GB	0.729	0.726	0.872	0.764	0.760	0.757	0.890	0.783
LGBM	0.724	0.722	0.870	0.760	0.748	0.745	0.887	0.778
Log. Regression	0.730	0.727	0.864	0.750	0.740	0.737	0.883	0.773
KNN	0.702	0.697	0.847	0.726	0.718	0.710	0.862	0.733
MLP	0.678	0.679	0.827	0.707	0.719	0.717	0.852	0.729
AdaBoost	0.710	0.707	0.787	0.651	0.738	0.737	0.817	0.677
DT	0.654	0.652	0.724	0.519	0.655	0.655	0.720	0.513

The feature importance analysis conducted using CatBoost on dataset A (Figure 1) revealed several key factors affecting length of stay. Ranked in descending order of importance, the most significant features include: the overall average length of stay for same-service hospitalizations within the previous 30 days (*ba_specialty_prev_hosp_avg_los*); the number of recent transfers across wards (*ba_recent_transfers_count*), possibly indicative of increasing medical complexity (i.e. patients needing specialized units); the surgery hospitalization area (*b_hosp_area_G*); the number of bed movements during hospitalization (*ba_movements_count*), including those within the same ward; low-intensity care level (*ba_intenscare_L*), suggesting the idea that stays associated with critical conditions tend to be shorter due to a focus on stabilizing the patient's condition; age in range 80-89 years (*p_age_range_80-89*), implying a higher risk of prolonged stays for elderly inpatients, possibly due to age-related vulnerabilities or comorbidities; the average LOS for same-patient hospitalizations in the prior year (*ba_patient_prev_hosp_avg_los*); and the need for a multidimensional geriatric assessment (*cs_is_uvm_req*), typically associated with frail individuals and elderly.

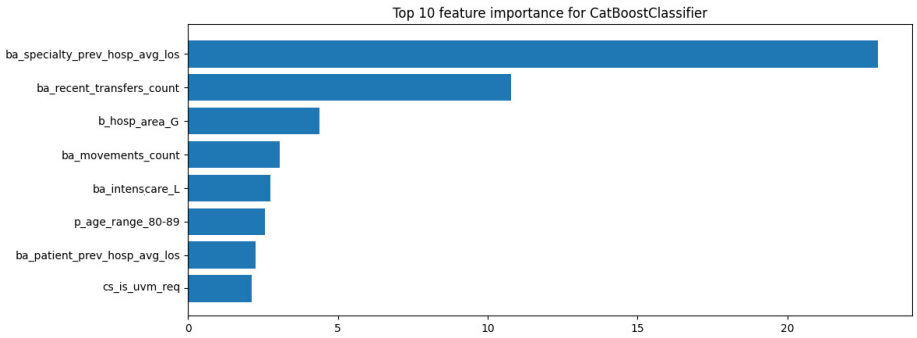

Fig. 1. Feature Importance for CatBoostClassifier on dataset A (multi-class classification task).

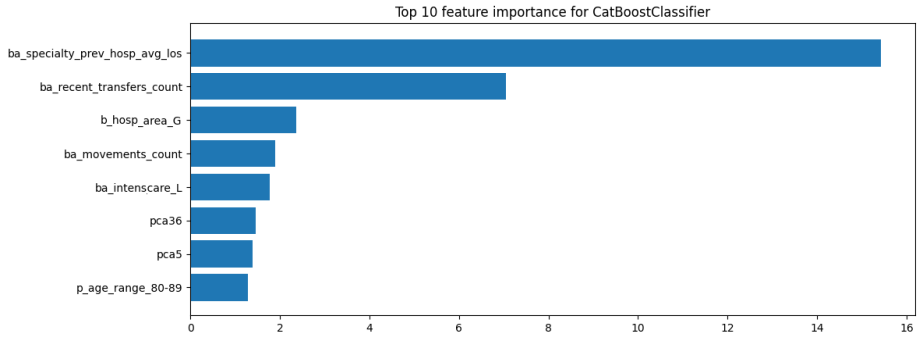

Fig. 2. Feature Importance for CatBoostClassifier on dataset B (multi-class classification task).

Interestingly, upon applying the same analysis to dataset B, PCA components —derived from diagnosis text embeddings— started to emerge as significant features (Figure 2). Although PCA offers a valuable tool for mitigating the curse of dimensionality, it can also lead to a less interpretable model. This limitation arises from the transformation of the original features into a new set of linearly combined variables, making it challenging to directly map the transformed representation back to the original interpretable elements.

5.3 Binary Classification Models

In line with the multi-class classification task, AUROC and AUPRC (Area Under the Precision-Recall Curve) served as the primary evaluation metrics for the binary classification models. Ranging from 0 to 1, AUROC effectively captures the trade-off between true and false positives across all possible thresholds. Conversely, AUPRC prioritizes the identification of positive samples, making it particularly advantageous in scenarios involving imbalanced datasets.

LogisticRegression provided the most accurate predictions for prolonged length of stay in dataset A (accuracy 86.61%, F1-score 64.62%, AUROC 90.54%), followed by VotingClassifierSoft and CatBoostClassifier. On the other hand, when analyzing dataset B, VotingClassifierSoft (accuracy 86.53%, F1-score 64.48%, AUROC 91.67%), CatBoostClassifier, and StackingClassifier demonstrated superior predictive capabilities for pLOS (Table 4).

Additionally, we assessed the alignment between predicted and actual pLOS risks using calibration curves for the three best-performing models (Figure 3).

In binary classification with probabilistic outputs, calibration curves offer a visual assessment of predicted probabilities compared to true class frequencies. An ideal model would exhibit a diagonal calibration curve, signifying perfect concordance between predicted and observed probabilities. This facilitates critical evaluation of model reliability, enabling selection of models with trustworthy estimates for informed decision-making.

Table 4. Results of tuned models on datasets A and B (binary classification task).

	Dataset A				Dataset B			
Model	**Acc. ↑**	**F1 ↑**	**A.ROC ↑**	**A.PRC ↑**	**Acc. ↑**	**F1 ↑**	**A.ROC ↑**	**A.PRC ↑**
Log. Regression	**0.866**	**0.646**	**0.905**	**0.794**	0.866	0.654	0.911	0.798
Voting Soft	0.863	0.636	0.905	0.789	**0.865**	**0.645**	**0.917**	**0.808**
CatBoost	0.867	0.654	0.903	0.785	0.867	0.654	0.912	0.799
Stacking	0.867	0.654	0.903	0.785	0.867	0.654	0.912	0.799
GB	0.861	0.636	0.901	0.782	0.863	0.650	0.908	0.793
LGBM	0.861	0.642	0.898	0.777	0.862	0.642	0.907	0.791
AdaBoost	0.862	0.642	0.896	0.778	0.853	0.634	0.888	0.769
RF	0.863	0.635	0.893	0.777	0.857	0.584	0.904	0.788
XGB	0.842	0.606	0.884	0.751	0.862	0.655	0.901	0.785
KNN	0.848	0.545	0.884	0.743	0.862	0.516	0.874	0.726
MLP	0.827	0.618	0.867	0.733	0.862	0.625	0.874	0.745
DT	0.799	0.569	0.719	0.423	0.862	0.558	0.713	0.412

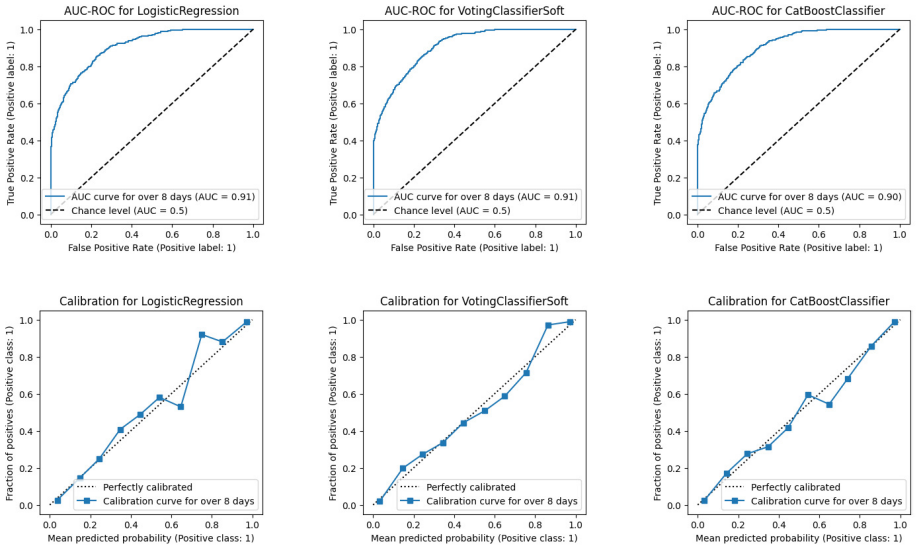

Fig. 3. AUROC and calibration plot for LogisticRegression, VotingClassifierSoft and CatBoostClassifier on dataset A (binary classification task).

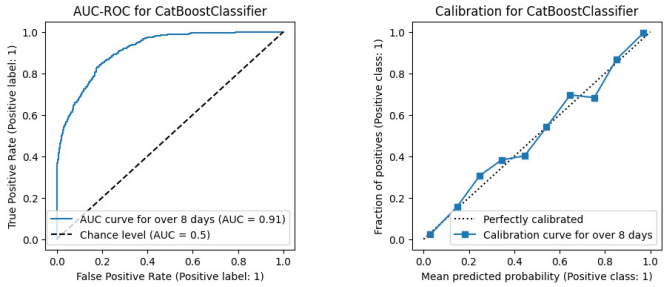

Fig. 4. AUROC and calibration plot for CatBoostClassifier on dataset B.

In dataset A, both VotingClassifierSoft and CatBoostClassifier proved well-calibrated, with points on their calibration plots clustering closely around the ideal diagonal line. However, VotingClassifierSoft revealed a propensity to overestimate the probabilities of pLOS in high-risk patients, while CatBoost exhibited a slight underestimation bias. When considering dataset B, CatBoost showed a reduced tendency to underestimate the likelihood of pLOS in high-risk inpatients (Figure 4).

6 Discussion

While prior research has demonstrated success in applying machine learning to predict LOS for distinct patient cohorts, such heart failure treatments [37], stroke interventions [6], and cesarean sections [29], this area remains relatively unexplored for the broader, heterogeneous inpatient population. Fragmented, problem-specific models necessitate intensive maintenance, hindering scalability and sustainability, especially in resource-constrained public health systems where patient focus is primary. The present work addresses this gap by targeting a wider spectrum of diagnoses and conditions. By avoiding the limitations inherent in single-ward studies (e.g., ICU [15, 24]), which may be constrained by internal dynamics, our methodology achieved encouraging generalizability for real-world implementation.

Our findings support existing evidence that ensemble models consistently outperform individual algorithms in specific tasks [30]. Through the aggregation of multiple base learners, ensemble models effectively mitigate biases and improve overall predictive capability. Specifically, tree-based ensemble models allow for converting complex models into transparent decision rules. By employing tools such as SHAP Values, Partial Dependence Plots, and Individual Conditional Expectation Plots, practitioners can gain a deeper understanding of the model's decision-making process, which is pivotal for the acceptance and the adoption of ML-driven systems in clinical settings.

The comparative analysis of different variations of the same data source proved the instrumental role of integrating text extracted from diagnoses in capturing subtle nuances not represented in structured data alone. This is particularly relevant to leveraging the expertise of medical staff and the outcomes derived from physical examinations. Our results align with prior investigations [7, 22], which have demonstrated that text-derived features may enhance predictive performance.

A further analysis identified factors significantly associated with prolonged stays, including the average LOS for previous admissions in the same service and the number of transfers within the current hospitalization. Additionally, when incorporating text from diagnosis, the need for a multidimensional geriatric assessment emerged as another key indicator for identifying patients at high risk of pLOS. These findings underscore the potential of machine learning to identify clinically relevant predictors of LOS and pLOS, potentially informing patient management and resource allocation strategies.

Lastly, our models leverage readily available data from institutional Electronic Health Records (EHRs), collected within the first 24 hours of hospitalization, enabling bed managers to promptly utilize them as decision-making aids upon patient admission.

Despite its strengths, our study is not without limitations. First, its retrospective design using historical data, may introduce potential selection bias. Second, a trade-off exists between model generalizability and potential accuracy gains due to the exclusion of vital signs and laboratory results in favor of readily available data. While incorporating this information could improve accuracy, its delayed availability and inconsistent collection across departments (e.g., long-term care, rehabilitation) would also limit the model's transferability in healthcare settings with less comprehensive EHR data gathering. Additionally, although clinical parameters are essential for patient monitoring, their direct influence on long-term outcomes like prolonged LOS (measured in days) might be less pronounced. Our study intentionally opted for features available at admission to enable early prediction and resource allocation, prioritizing real-world applicability. Nonetheless, future efforts will focus on incorporating more granular clinical data as they become available during a patient's stay, to evaluate their impact on model performance. Finally, the monocentric nature of the study, relying on data from a single hospital, restricts external validation. In future work, we plan to consider datasets from seven additional Italian hospitals, enabling a more compelling assessment of generalizability across different environments. However, we posit that striving for absolute generalizability may not be the most effective strategy. Given documented influences of structural, organizational, and administrative factors on LOS, a tailored approach might prove more valuable, acknowledging healthcare system diversity while providing an adaptable framework for LOS prediction within individual hospitals.

7 Conclusions

This study provides robust evidence that ensemble-based prediction models outperform traditional techniques in forecasting LOS and identifying general inpatients at high risk of prolonged LOS across diverse services and wards. Our core methodology's reliance on readily available EHR data, coupled with algorithms that do not necessitate resource-intensive procedures or specialized hardware, promotes its potential integration into various healthcare settings and workflows, serving as a second-opinion tool to support both medical staff and healthcare management in their daily tasks.

A notable achievement was the preliminary embedding of our inference model into the EBMS employed at Ospedale di Sassuolo. By leveraging a combination of RESTful APIs and HL7 messaging, this integration significantly enhanced the user experience for bed managers involved in the experimental phase. Beside providing a comprehensive visual representation of bed occupancy and availability across departments, the augmented version of the system now delivers predictive insights into expected LOS for each patient, particularly those at high risk

of prolonged hospitalization. Early identification of such patients allows healthcare providers to proactively implement targeted interventions, including closer monitoring and timely discharge planning, potentially mitigating the incidence of protracted stays. This preemptive approach can contribute to smoother patient flow, increased bed availability, a lower rate of rescheduled interventions, improved patient satisfaction, and, ultimately, reduced overall healthcare expenditures. These results emphasize the priority for hospitals to embrace adaptation and innovation to meet the demands of an aging population with chronic conditions, while containing costs and optimizing resources to ensure the sustainability of public healthcare for future generations.

References

1. Health at a glance 2023 oecd indicators (2023), https://doi.org/10.1787/7a7afb35-en [Accessed: (2024/01/08)]
2. Oecd indicators on healthcare resources: Hospital beds by sector between 2001 and 2021 (2024), https://stats.oecd.org/ [Accessed: (2024/01/08)]
3. Ackroyd-Stolarz, S., Guernsey, J.R., Mackinnon, N.J., Kovacs, G.: The association between a prolonged stay in the emergency department and adverse events in older patients admitted to hospital: a retrospective cohort study. BMJ Quality & Safety **2011**(20), 564–569 (2011)
4. Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., Levin, S.: Real-time prediction of inpatient length of stay for discharge prioritization. Journal of the American Medical Informatics Association **23(e1), e2-e10** (2016)
5. Bo, M., Fonte, G., Pivaro, F., Bonetto, M., Comi, C., Giorgis, V., Marchese, L., Isaia, G., Maggiani, G., Furno, E., Falcone, Y., Isaia, G.C.: Prevalence of and factors associated with prolonged length of stay in older hospitalized medical patients. Geriatrics & gerontology international **16**(3), 314–321 (2016)
6. Chen, R., Zhang, S., Li, J., et al.: A study on predicting the length of hospital stay for chinese patients with ischemic stroke based on the xgboost algorithm. BMC Medical Informatics and Decision Making **23**(1), 1–10 (2023)
7. Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A., Sanchez, S.: The prediction of hospital length of stay using unstructured data. BMC Medical Informatics and Decision Making **21**(1), 351 (2021)
8. Chuang, M.T., Hu, Y.H., Lo, C.L.: Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. International Transactions in Operational Research **25**(1), 75–90 (2018)
9. Daghistani, T.A., Elshawi, R., Sakr, S., Ahmed, A.M., Al-Thwayee, A., Al-Mallah, M.H.: Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. International journal of cardiology **288**, 140–147 (2019)
10. Di Matteo, V., Tommasini, T., Morandini, P., Savevski, V., Grappiolo, G., Loppini, M.: Machine learning prediction model to predict length of stay in patients undergoing hip or knee arthroplasties: Results from a high volume single center multivariate analysis. Pre-prints **2023110915** (2023)
11. D'Onofrio, G., D'Amore, A., Onofaro, F., Caputi, E., Napoli, A., Triassi, M., Marino, M.R.: Prediction of hospital length stay for patients undergoing mastectomy. Stud Health Technol Inform. **29**(305), 261–264 (June 2023), pMID: 37387012

12. of Economy, I.M., Finance: Monitoraggio della spesa sanitaria, rapporto n. 7 (2020), https://www.rgs.mef.gov.it/VERSIONE-I/attivita_istituzionali/monitoraggio/spesa_sanitaria/2020/ [Accessed: (2024/01/08)]
13. of Economy, I.M., Finance: Monitoraggio della spesa sanitaria, rapporto n. 10 (2023), https://www.rgs.mef.gov.it/VERSIONE-I/attivita_istituzionali/monitoraggio/spesa_sanitaria/ [Accessed: (2024/01/08)]
14. Evans, J., Kobewka, D., Thavorn, K., D'Egidio, G., Rosenberg, E., Kyeremanteng, K.: The impact of reducing intensive care unit length of stay on hospital costs: evidence from a tertiary care hospital in canada. Canadian Journal of Anesthesia **65**(6), 627–635 (2018)
15. Gholipour, C., Rahim, F., Fakhree, A., Ziapour, B.: Using an artificial neural networks (anns) model for prediction of intensive care unit (icu) outcome and length of stay at hospital in traumatic patients. Journal of clinical and diagnostic research: JCDR **9**, 4 (2015)
16. Gohil, R., Rishi, M., Tan, B.H.: Pre-operative serum albumin and neutrophillymphocyte ratio are associated with prolonged hospital stay following colorectal cancer surgery. British journal of medicine and medical research **4**(1), 481 (2014)
17. Hauck, K., Zhao, X.: How dangerous is a day in hospital? A model of adverse events and length of stay for medical inpatients. Medical care pp. 1068–1075 (2011), http://www.jstor.org/stable/23053852
18. of Health, I.M.: Rapporto annuale sull'attività di ricovero ospedaliero (2020), https://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?id=3277, last accessed 2024/01/08
19. Heit, J.A., Silverstein, M.D., Mohr, D.N., Petterson, T.M., O'Fallon, W.M., Melton, L.J.: Risk factors for deep vein thrombosis and pulmonary embolism: a population-based case-control study. Archives of internal medicine **160**(6), 809–815 (2000)
20. Hesselink, G., Van Den Bogaert, M., Akkermans, R.P., Schoon, Y.: Risk factors for prolonged length of stay of older patients in an academic emergency department: a retrospective cohort study. Emergency medicine international **2019** (2019)
21. Istat: Annual italian statistics (2022), http://dati.istat.it/, last accessed 2024/01/08
22. Jiang, L.Y., Liu, X.C., Nejatian, N., et al.: Health system-scale language models are all-purpose prediction engines. Nature pp. 1–6 (2023)
23. Jo, Y.Y., Han, J., Park, H.W., Jung, H., Lee, J.D., Jung, J., Cha, H.S., Sohn, D.K., Hwangbo, Y.: Prediction of prolonged length of hospital stay after cancer surgery using machine learning on electronic health records: retrospective cross-sectional study. JMIR medical informatics **9**, 2 (2021)
24. Ma, X., Si, Y., Wang, Z., Wang, Y.: Length of stay prediction for icu patients using individualized single classification algorithm. In: Computer methods and progr in biomedicine, 186, 105224. AMS (2020)
25. Marfil-Garza, B.A., Belaunzarán-Zamudio, P.F., Gulias-Herrero, A., Zuñiga, A.C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., Sifuentes-Osornio, J.: Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary healthcare center in mexico. PloS one **13**, 11 (2018)
26. Marshall, A., Vasilakis, C., El-Darzi, E.: Length of stay-based patient flow models: recent developments and future directions. Health care management science **8**, 213–220 (2005)
27. 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: World

conference on information systems and technologies. pp. 202–211. Springer International Publishing, Cham (April 2020)
28. Molloy, I.B., Martin, B.I., Moschetti, W.E., Jevsevar, D.S.: Effects of the length of stay on the cost of total knee and total hip arthroplasty from 2002 to 2013. the journal of bone and joint surgery. American **99**, 5 (2017)
29. Montella, E., Marini, M.R., Majolo, M., Raiola, E., Russo, G., Longo, G., Lombardi, A., Borrelli, A., Triassi, M.: Regression and classification methods for predicting the length of hospital stay after cesarean section: a bicentric study. In: Proceedings of the 6th International Conference on Medical and Health Informatics. p. 135–140. Association for Computing Machinery, New York, NY, USA (2022)
30. Muhlestein, W.E., Akagi, D.S., Davies, J.M., Chambless, L.B.: Predicting inpatient length of stay after brain tumor surgery: developing machine learning ensembles to improve predictive performance. Neurosurgery **85**(3), 384 (2019)
31. Olivato, M., Rossetti, N., Gerevini, A.E., Chiari, M., Putelli, L., Serina, I.: Machine learning models for predicting short-long length of stay of covid-19 patients. Procedia Computer Science **207**, 1232–1241 (2022)
32. Rocheteau, E., Liò, P., Hyland, S.: Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: Proceedings of the conference on health. pp. 58–68. inference, and learning (April 2021)
33. Rojas-García, A., Turner, S., Pizzo, E., Hudson, E., Thomas, J., Raine, R.: Impact and experiences of delayed discharge: A mixed-studies systematic review. Health Expectations **21**(1), 41–56 (2018)
34. Rotar, E.P., Beller, J.P., Smolkin, M.E., Chancellor, W.Z., Ailawadi, G., Yarboro, L.T., Hulse, M., Ratcliffe, S.J., Teman, N.R.: Prediction of prolonged intensive care unit length of stay following cardiac surgery. Seminars in Thoracic and Cardiovascular Surgery **34**(1), 172–179 (March 2022)
35. Stone, K., Zwiggelaar, R., Jones, P., Mac Parthaláin, N.: A systematic review of the prediction of hospital length of stay: Towards a unified framework. PLOS Digital Health **1**, 4 (2022)
36. Stricker, K., Rothen, H.U., Takala, J.: Resource use in the icu: short-vs. long-term patients. Acta Anaesthesiologica Scandinavica **47**(5), 508–515 (2003)
37. Sud, M., Yu, B., Wijeysundera, H.C., et al.: Associations between short or long length of stay and 30-day readmission and mortality in hospitalized patients with heart failure. JACC: Heart Failure **5**(8), 578–588 (2017)
38. Thompson, B., Elish, K.O., Steele, R.: Machine learning-based prediction of prolonged length of stay in newborns. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1454–1459. IEEE (2018)
39. Trunfio, T.A., Borrelli, A., Improta, G.: Is it possible to predict the length of stay of patients undergoing hip-replacement surgery? International Journal of Environmental Research and Public Health **19**(10), 6219 (2022)
40. Trunfio, T.A., Scala, A., Giglio, C., Rossi, G., Borrelli, A., Romano, M., Improta, G.: Multiple regression model to analyze the total los for patients undergoing laparoscopic appendectomy. BMC Medical Informatics and Decision Making **22**(1), 1–8 (2022)
41. Zeleke, A.J., Palumbo, P., Tubertini, P., Miglio, R., Chiari, L.: Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a gradient boosting algorithm analysis. Front. Artif. Intell. **6** (2023), article 1179226
42. Zolbanin, H.M., Davazdahemami, B., Delen, D., Zadeh, A.H.: Data analytics for the sustainable use of resources in hospitals: predicting the length of stay for patients with chronic diseases. Information & Management **59**(5), 103282 (2022)