

# Patient Length of Stay Prediction

Sarah Nassar  
Queen's University  
18sbn@queensu.ca

Wenhan Li  
Queen's University  
19wl36@queensu.ca

Will Wang  
Queen's University  
21wzw1@queensu.ca

Muhammad Maaz Zafar  
Queen's University  
20mmz2@queensu.ca

George Salib  
Queen's University  
george.salib@queensu.ca

Momin Alvi  
Queen's University  
21mna5@queensu.ca

Ryan Zietlow  
Queen's University  
21rjz3@queensu.ca

Meetansh Kharbanda  
Queen's University  
meetansh.kharbanda@queensu.ca

**Abstract**—In the quest to enhance hospital operational planning and efficiency, our research presents a pioneering predictive model for patient length of stay, a critical metric for hospital resource management. We approached the problem through regression analysis, binary classification, and multi-class categorization, employing a suite of machine learning algorithms. For our regression analysis, our neural network model achieved a median squared error of four days. Our classification analysis revealed that a binary classification model distinguishing between short-term (1-10 days) and long-term (11-119 days) stays using a neural network approach outperformed its counterparts by achieving an average F1 score of 72%. Complementing our modelling efforts, we also developed a user interface to allow healthcare professionals to interact with our models. Our project not only advances length of stay prediction methodologies, but also contributes to the broader field of healthcare analytics by demonstrating how machine learning techniques can be implemented to improve hospital operations.

## I. INTRODUCTION

### A. Motivation

In healthcare systems, the optimization of resources and improvement of patient care are objectives of the utmost importance. With the inflating costs of healthcare delivery, efficient resource management has become increasingly essential.

Patient length of stay is a crucial metric, defined as the number of days that an in-patient spends in the hospital during a single admission event. Length of stay serves as a significant indicator of resource consumption within hospital facilities, offering insights into patient flow dynamics and aiding in the evaluation of operational efficiency across care units. Therefore, having estimates for how long patients might stay in the hospital ahead of time can help healthcare staff facilitate more efficient resource allocation, reduce costs, and ultimately enhance the quality of patient care [Stone et al., 2022].

### B. Problem Definition

Accurately predicting a patient's length of stay is important in the hospital context. A tool that predicts patient length of stay can help improve the scheduling and planning process and indicate disease severity or complexity. The purpose of this paper is to investigate the usefulness of machine learning technologies in the context of patient length of stay prediction.

## II. RELATED WORK

A recent review by Stone et al. touches on many techniques and provides an extensive analysis of the problem of length of stay prediction as a whole [Stone et al., 2022]. Operational research-based approaches, such as averaging or compartmental modeling, rely on identifying patterns and characteristics in historical data to make predictions. For example, compartmental modeling splits similar cases into respective compartments. However, these methods lack flexibility and struggle to adapt to new data, limiting relevance. New data introduce new information and patterns that these models cannot account for. Data mining approaches use algorithms such as neural networks and support vector machines to capture complex relationships in data. These models discover patterns that are not immediately apparent. However, they also raise ethical concerns due to their black-box nature making understanding them difficult. Medical professionals and patients must be informed about the decisions affecting the healthcare system, and when these decisions affect the well-being of patients, understanding the reasoning behind them is crucial. This review paper also identifies the shortcomings of length of stay prediction, with the main one being that no unified framework currently exists.

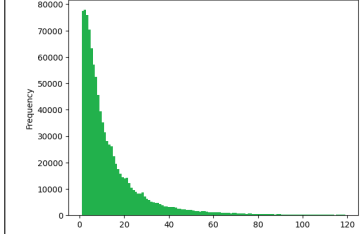
The review paper also examines the state-of-the-art, citing a study that worked on predicting patient length of stay after brain tumor surgery [Muhlestein et al., 2019]. This study presents a robust ensemble model with a root mean squared logarithmic error (RMSLE) of 0.56 on the internal validation, or test, set. The predictive model discussed in the study is tailored for length of stay after brain tumor surgery, which differs from our project goal of training a generic predictor. A targeted approach enhances the model's relevance within the specific scenario, but our project is aimed at exploring the predictive power of machine learning in non-specialized patient scenarios.

## III. METHODOLOGY

### A. Dataset

The dataset we used is from the New York Statewide Planning and Research Cooperative System (SPARCS). We included the publicly available and de-identified hospital

TABLE I  
SOME DATASET FEATURES.

Values	Count (Frequency)
<b>Age Group (Upon Discharge)</b>	
0-17	117,204 (11.54%)
18-29	110,587 (10.89%)
30-49	205,509 (20.23%)
50-69	314,860 (40.00%)
$\geq 70$	267,572 (26.34%)
<b>Type of Admission</b>	
Emergency	590,606 (58.15%)
Elective	212,104 (20.88%)
Urgent	170,407 (16.78%)
Trauma	24,509 (2.41%)
Newborn	18,106 (1.78%)
<b>Major Diagnostic Category (Top 5)</b>	
Circulatory System	95,464 (9.40%)
Nervous System	94,272 (9.28%)
Digestive System	76,517 (7.53%)
Musculoskeletal System	75,083 (7.39%)
Respiratory System	71,078 (7.00%)
<b>Procedure Type</b>	
Medical	639,806 (62.99%)
Surgical	375,926 (37.01%)
<b>Length of Stay</b>	
1-119 days	

inpatient discharges data file of each year from 2015 to 2019 [csv:2015, 2017], [csv:2016, 2018], [csv:2017, 2019], [csv:2018, 2022], [csv:2019, 2022]. Each year's data file has over 2.3 million records, with each row representing a hospital admission and discharge.

Our pre-processing procedure started out by cleaning the data files. We decided to keep a subset of features that seemed relevant, namely age, gender, race, ethnicity, type of admission, diagnosis description, severity of illness, risk of mortality, and whether the procedure was medical or surgical. We removed rows with null, infrequent, or unavailable values. We also removed duplicate rows and those with a categorical length of stay of "120 +". The details of some of the dataset's features are displayed in Table I.

In addition to regression analysis, we also decided to experiment with various length of stay groupings for binary and multi-class classification strategies. For the binary case, we tried three versions (V1: {1-7 days, 8-119 days}, V2: {1-9 days, 10-119 days}, and V3: {1-10 days, 11-119 days}). For the multi-class case, we tried two versions (V1: {1-4 days, 5-9 days, 10-119} days and V2: {1-6 days, 7-14 days, 15-119 days}).

After cleaning the data and preparing the length of stay target feature, we split the dataset into 80% training and 20% testing stratified by the length of stay. We completed the data preparation step by one-hot encoding the categorical features,

and our final dataset had 1,015,732 rows and 56 columns.

### B. Experiment Setup

We compared several different machine learning algorithms for our project. For regression, we trained linear regression, multi-layer perceptron (MLP) or neural network, decision tree (DT), and random forest (RF) models. For classification, we also trained MLP, DT, and RF models in addition to Bernoulli Naïve Bayes (NB) and logistic regression (LR) models. We kept the default scikit-learn parameters for most models, but increased the maximum number of iterations of LR and MLP to 500. For comparison purposes as well as reproducibility, we used a random seed of zero for the training and testing subset split and LR, MLP, DT, and RF models. In order to avoid negative numbers as regression outputs, we trained the regression models on the natural logarithm (i.e., logarithm with base  $e$ ) of the length of stay and we make sure to exponentiate the model's output for testing and inference. Because the length of stay target variable is measured in days, we also perform rounding after exponentiation.

### C. Evaluation Methods

To measure the performance of our regression models, we used the coefficient of determination or  $R^2$  (1), mean absolute error or Mean AE (2), median absolute error or Median AE (3), root mean squared error or RMSE (4), and root mean squared log error or RMSLE (5). For all these regression metrics except the  $R^2$  value, a smaller number signals better performance. In the equations below,  $n$  is the number of testing data points,  $y_i$  is the actual length of stay for a single hospital admission,  $\hat{y}_i$  is the model's prediction for the length of stay, and  $\bar{y}$  is the average of the actual length of stay over all admissions in the testing subset.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$Mean\ AE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$Median\ AE = median(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|) \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log_e(y_i + 1) - \log_e(\hat{y}_i + 1))^2} \quad (5)$$

To measure the performance of our classification models, we calculated the accuracy (6), F1 score (7), and receiver operating characteristic area under the curve (ROCAUC). For all these classification metrics, a larger number indicates better performance. The F1 score is based on the macro average

TABLE II  
REGRESSION RESULTS.

Model	$R^2$	Mean AE	Median AE	RMSLE	RMSLE
Linear	0.12	8.68	<b>4.00</b>	15.27	0.80
MLP	<b>0.23</b>	<b>8.11</b>	<b>4.00</b>	<b>14.31</b>	<b>0.75</b>
DT	0.20	8.46	5.00	14.58	0.79
RF	0.22	8.33	<b>4.00</b>	14.39	0.78

of the values from each individual class. For the multi-class configuration, the ROCAUC is based on the one-vs-rest strategy. In the equations below,  $TP$  is the number of true positive predictions,  $TN$  is the number of true negative predictions,  $FP$  is the number of false positive predictions, and  $FN$  is the number of false negative predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (7)$$

#### IV. RESULTS AND DISCUSSION

Table II shows the performance of the four regression algorithms. The neural network proved to be the algorithm that was best able to capture the relationship between the one-hot encoded binary features and the length of stay target feature. Our generic predictor achieved an RMSLE of 0.75, which is comparable to the RMSLE of 0.56 that was achieved by the state-of-the-art in specifically predicting patient length of stay after brain tumour surgery [Muhlestein et al., 2019].

Table III presents the evaluation metric scores of the five classification algorithms when trained on two different multi-class length of stay groupings and five binary length of stay groupings. Across both classification types, the neural network consistently out-performed the other algorithms. Both multi-class configurations performed similarly, but the second version of the length of stay grouping (1-6 days, 7-14 days, 15-119 days) has a slight advantage. The binary classification models perform significantly better than their multi-class counterparts. All three binary configurations performed similarly, but the third version of the length of stay grouping (1-10 days, 11-119 days) is in the lead.

Although the performance of our models is promising, there are factors that may contribute to the current level of predictive error. Hospitals are dynamic environments and their operating efficiency can fluctuate based on numerous circumstances, such as the time of day or season. This can pose a challenge for consistent model accuracy since the dataset we used does not include detailed timing information. Additionally, there exist variations in efficiency levels among different hospitals, so a length of stay prediction may be correct for some hospitals but incorrect for others. Preliminary experiments indicate the potential for improved performance when including hospital county in the input feature space. For example, the MLP regression model trained with the county information and tested on a different and larger testing subset improves the

TABLE III  
CLASSIFICATION RESULTS.

Grouping	Model	Accuracy	F1	ROCAUC
<b>Binary</b>				
1-7 days 8-119 days	NB	0.66	0.66	0.72
	LR	0.68	0.68	0.75
	MLP	0.72	<b>0.72</b>	0.79
	DT	0.69	0.69	0.74
	RF	0.70	0.69	0.75
1-9 days 10-119 days	NB	0.67	0.66	0.73
	LR	0.69	0.68	0.75
	MLP	0.72	<b>0.72</b>	<b>0.80</b>
	DT	0.70	0.70	0.76
	RF	0.70	0.70	0.77
1-10 days 11-119 days	NB	0.67	0.66	0.73
	LR	0.69	0.67	0.75
	MLP	<b>0.73</b>	<b>0.72</b>	<b>0.80</b>
	DT	0.71	0.70	0.76
	RF	0.71	0.70	0.77
<b>Multi-Class</b>				
1-4 days 5-9 days 10-119 days	NB	0.52	0.43	0.67
	LR	0.55	0.41	0.69
	MLP	<b>0.58</b>	0.49	0.73
	DT	0.53	0.46	0.67
	RF	0.54	0.46	0.68
1-6 days 7-14 days 15-119 days	NB	0.51	0.46	0.68
	LR	0.53	0.45	0.70
	MLP	0.57	<b>0.51</b>	<b>0.74</b>
	DT	0.53	0.49	0.69
	RF	0.53	0.50	0.70

mean absolute error from 8.11 to 5.34 and the median absolute error from four days to three days. However, we decided to keep the original feature set for our results in order to enable the possibility of external validation with data from other, non-New York, regions.

Our results also suggest that the features considered may not be sufficient to fully represent a patient's hospital stay scenario, especially since our analysis does not directly take into account the occurrence of complications and misdiagnoses. These observations indicate a need for a more detailed approach to feature selection and model training to improve prediction performance.

After saving the best three models from each type of analysis, we developed a user interface with Jupyter Widgets to allow users to enter a patient's information and view the prediction outputs. A screenshot of the user interface can be seen in Fig. 1.

##### A. Ethical Considerations

As a healthcare project, patient health and safety should be the top priority. Inaccurately predicting a patient's length of stay can impact the allocation of resources and the amount of time and care dedicated, which indicates the need for thorough testing. Further, the dataset used for this project comes from a specific population (New York), so this constitutes data bias and the model cannot be expected to work for all populations and demographics without extensive external validation.

##### B. Replication Package

Our code and link to the user interface can be found at <https://github.com/S-N-2019/PatientLengthOfStayPrediction>.

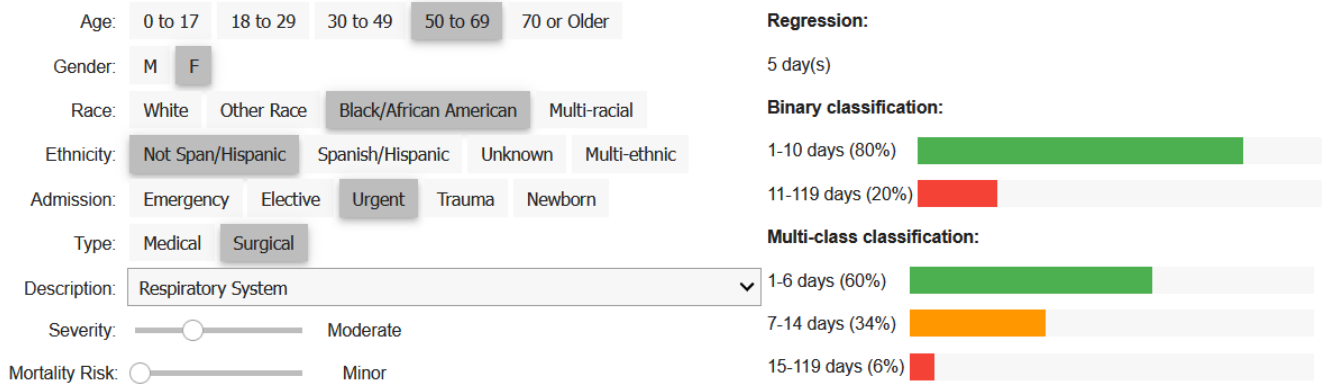


Fig. 1. User interface built with Jupyter Widgets.

## V. CONCLUSION

Overall, the purpose of this project was to explore machine learning techniques for the task of patient length of stay prediction, which is an important metric for hospital efficiency and planning. We used a large dataset composed of admission records from New York over several years and tried framing the problem as regression, binary classification, and multi-class classification. Our best regression model was a neural network with a  $R^2$  value of 0.23 and a median squared error of four days and our best classification model was also a neural network for predicting a short-term (1-10 days) or long-term (11-119 days) stay with an average F1 score of 72%. Finally, we built a user interface to allow users to interact with our models and obtain length of stay predictions given a patient's information.

## VI. FUTURE WORK

Our current performance results are promising, but there is always room for improvement. Our next steps are to look into more years of data to train our models, try different combinations for the length of stay groupings, and invest more time into data understanding and feature engineering. To make more precise estimates, we can also try other machine learning algorithms and more specialized neural network architectures. We would also like to meet with professionals in the healthcare industry to investigate our models' usefulness in medical settings and potentially write a further research paper. Moreover, we can conduct external validation by testing our models on data from other hospitals, such as Canadian ones, to assess their versatility and reliability across diverse healthcare settings. Furthermore, we envision integrating our models into electronic health record (EHR) systems to automate the provision of length of stay predictions without necessitating direct user input for each patient. This integration would streamline workflow processes for healthcare professionals and facilitate more efficient resource allocation within hospital settings.

## VII. LIMITATIONS

Data preparation is an essential part in machine learning, and as such, data scientists have to identify appropriate features and understand their internal relationships. Since most members of our team do not have a background in healthcare, the features we chose to include are not guaranteed to be the best combination of information hinting at a patient's expected length of stay.

## VIII. ACKNOWLEDGEMENTS

We would like to thank two medical students, Aryan Shah and Stephanie Shaw, who were part of the QMIND + QMED collaboration and contributed valuable insights on the task of patient length of stay prediction. We would also like to Aryan Shah for his feedback on this paper.

## REFERENCES

- [csv:2015, 2017] csv:2015 (2017). Hospital inpatient discharges (sparses de-identified): 2015.
- [csv:2016, 2018] csv:2016 (2018). Hospital inpatient discharges (sparses de-identified): 2016.
- [csv:2017, 2019] csv:2017 (2019). Hospital inpatient discharges (sparses de-identified): 2017.
- [csv:2018, 2022] csv:2018 (2022). Hospital inpatient discharges (sparses de-identified): 2018.
- [csv:2019, 2022] csv:2019 (2022). Hospital inpatient discharges (sparses de-identified): 2019.
- [Muhlestein et al., 2019] Muhlestein, W. E., Akagi, D. S., Davies, J. M., and Chambliss, L. B. (2019). Predicting inpatient length of stay after brain tumor surgery: developing machine learning ensembles to improve predictive performance. *Neurosurgery*, 85(3):384.
- [Stone et al., 2022] Stone, K., Zwiggelaar, R., Jones, P., and Mac Parthaláin, N. (2022). A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4):e0000017.