ISSN No:-2456-2165

Enhancing Hospital Resource Management: Predicting Patient Length of Stay Using Machine Learning

Anurag Priyadarshi
Engineering Science (Data Science)
University at Buffalo Buffalo, USA

Anshumaan Karna
Computer Science & Engineering IIIT
Naya Raipur, Raipur, India

Abstract:- This project aims to enhance hospital management by predicting patients' length of stay using the MIMIC dataset, ultimately resulting in substantial cost savings and improved resource allocation. In our initial approach, we categorized the target variable, "length of stay" into three classes: short, medium, and long. Employing classification models including Logistic Regression, Random Forests, and Gradient Boosting, we attempted to predict patient outcomes. However, the initial results were unsatisfactory, prompting us to refine our methodology. We expanded the target variable classes to five: very short, short, medium, long, and very long, leading to improved accuracy in predicting short hospital stays. In the second approach, we treated the length of stay as a continuous variable and employed Multiple Linear Regression for modeling. Unfortunately, this ap- proach yielded sub-optimal results compared to the classification techniques. We analyzed the encountered limitations and further propose future steps to enhance the efficiency and accuracy of prediction models, ultimately contributing to more effective hospital resource management.

Keywords:- *Length of Stay, MIMIC III, Classification, Random Forest, Healthcare.*

I. INTRODUCTION

The rising demand for healthcare, especially in developed countries, is driven by an aging population. Policymakers and healthcare organizations aim to align financial incentives with best practices to improve patient outcomes and healthcare af- fordability. Chronic diseases, linked to changing lifestyles and dietary habits, pose a significant challenge, being the leading cause of mortality and disability in the US [1]. Conditions like obstructive pulmonary disease, type 2 diabetes, cancer, and cardiovascular diseases are burdening healthcare systems [2]. Long-term hospital stays have surged over the past decade due to the prevalence of chronic illnesses [2]. In the US, hospitals spend over $377.5 billion annually on patient admissions and stays [3]. Prolonged hospitalizations increase the risk of hospital-acquired conditions [4]. Accurately predicting a patient's hospital stay length is crucial for efficient resource management, cost reduction, and improving patient care [4]. Machine learning and data mining techniques, particularly in intensive care, show promise in optimizing healthcare resource management [5]. This project utilizes the MIMIC database [6], known for its applicability in various healthcare analyses [6].

II. MOTIVATION

The motivation driving this project stems from the immense financial strain imposed on the healthcare system in the United States. In 2020, the nation's healthcare expenditure surpassed an astonishing $4 trillion, with nearly a third of this allocated to hospital charges and services. Understanding the substantial costs associated with patient care, especially concerning the duration of hospital stays, emphasizes the need for efficient resource management.

The outbreak of the COVID-19 pandemic significantly disrupted routine healthcare services [7] [8]. Lock-downs and a healthcare focus on COVID-19 treatment led to the halting of critical vaccination programs against diseases like measles, polio, and meningitis. This interruption in regular healthcare protocols endangered millions of children, underscoring the urgency of efficient patient care and resource allocation.

The pandemic highlighted the importance of categorizing patients based on symptom severity. Hospitals faced over- whelming patient influxes, necessitating a system to prioritize admissions and allocate resources effectively [9]. This under- scored the necessity of a predictive model that could categorize patients into "short", "medium" and "long" stays based on various metrics. Such a model would ensure appropriate care and resource distribution, particularly for the most critical cases.

It's crucial to emphasize that our project does not intend to replace medical professionals, rather it aims to complement their expertise. Recognizing that the pivotal period for a patient's treatment begins upon hospital admission, our model provides crucial insights for medical staff to optimize resource utilization and deliver timely care. By achieving this, we strive to alleviate the strain on healthcare resources and ultimately save both time and lives.

III. OBJECTIVE

Our objective is to construct two distinct models utilizing data extracted from the MIMIC database [6]. These models aim to provide predictive and exploratory insights. The first model is designed to forecast the probability of a categorical outcome, specifically a patient's length of stay. Patients will be categorized into various classes of length of stay based on their individual characteristics and

ISSN No:-2456-2165

circumstances of admission. Subsequently, we will treat the length of stay (target variable) as a continuous variable and develop regression models using diverse regression techniques. Following the implementation of both approaches, we will assess model performance and choose the optimal model based on the results obtained. The practical application of this model in real-life scenarios is a key consideration.

Through predictive analysis and by considering patient allocation with respect to healthcare resource utilization, our objective is to derive insights that can better inform healthcare systems. These insights will enable proactive allocation of crit- ical healthcare resources and will ultimately enhance overall patient care outcomes.

IV. DATA DESCRIPTION

This study employs data from the publicly available MIMIC III (Medical Information Mart for Intensive Care) clinical database [6] as seen in Fig. 1). MIMIC datasets are widely uti- lized in diverse research domains including clinical medicine, epidemiology, and physiology [10] [11] [12]. These datasets are openly accessible to researchers globally, subject to a data use agreement, providing detailed yet deidentified information about a large cohort of ICU patients. The MIMIC III dataset contains comprehensive clinical data from patients treated at the Beth Israel Deaconess Medical Center in Boston, Mas- sachusetts, covering the years from 2001 to 2012.

Comprising a total of 26 interlinked tables, the MIMIC III dataset uses unique identifiers like Patient ID to establish con- nections between tables. These tables offer extensive insights into patient admissions, initial conditions upon admission, demographic profiles, caregiver details, prescribed medica- tions, and various other aspects. Crucially, it also includes information about the length of hospital stays for patients, a key focus of our current study.

For the purposes of this academic study, we have chosen to work with an aggregated version of the MIMIC III dataset, conveniently accessible on Kaggle. This aggregated version condenses the data into a single file, featuring 28 variables and a total of 59,000 entries.

V. APPROACH

A. *Initial Exploratory Data Analysis*

During the Initial Exploratory Data Analysis (EDA), it was discovered that three categorical variables had missing values or entries: “admission diagnosis”, “religion”, and “marital status”. To address these missing values, we assigned “UN- KNOWN CATEGORY” to the respective entries.

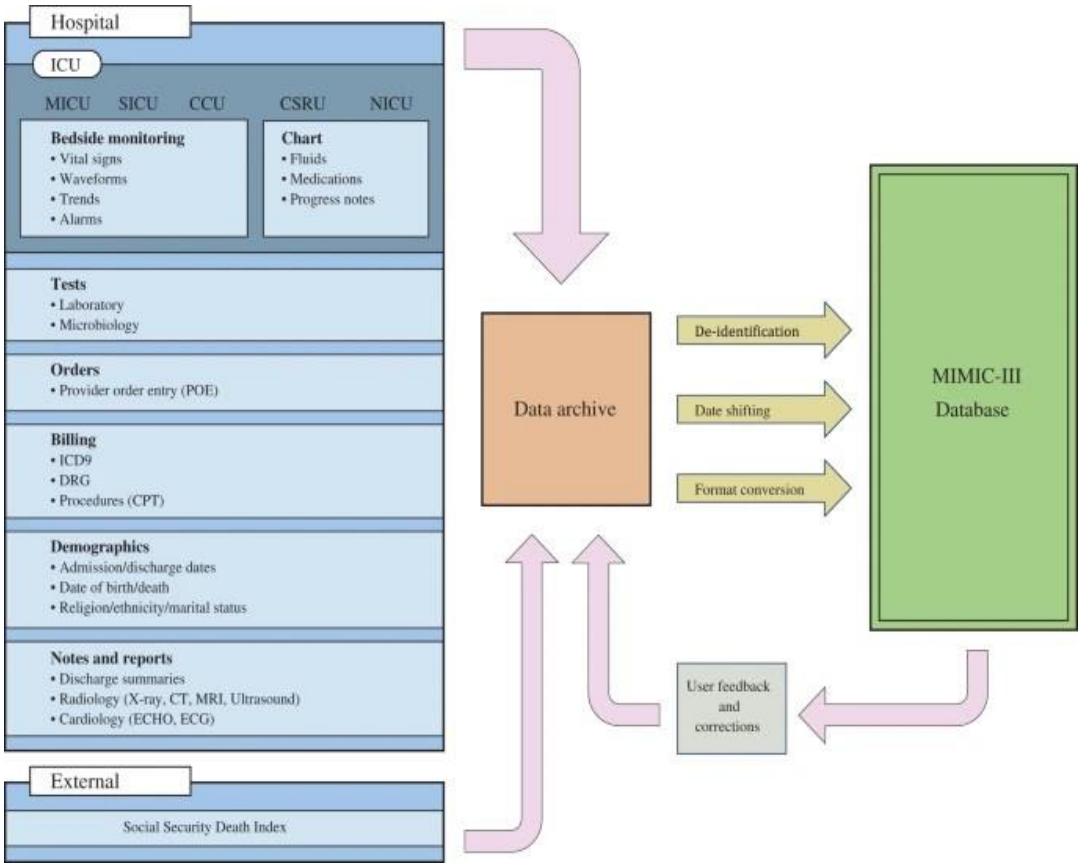


Fig. 1: Overview of MIMIC III clinical database.

In line with our problem statement of predicting a patient's length of stay, which is initially a continuous variable, we categorized it into three major groups to facilitate analysis:

- Short Stays: 0-5 days
- Medium Stays: 6-10 days
- Longer Stays: 10 days and above

Converting the continuous variable into a categorical one allows for clearer insights. By consolidating various categor- ical variables into fewer classes, we can derive meaningful observations. For instance, when plotting box plots for dif- ferent age groups and their respective lengths of stay (as illustrated in Fig. 2), it was observed that younger individuals tend to have shorter hospital stays, while older individuals are predominantly categorized under longer and medium stays.

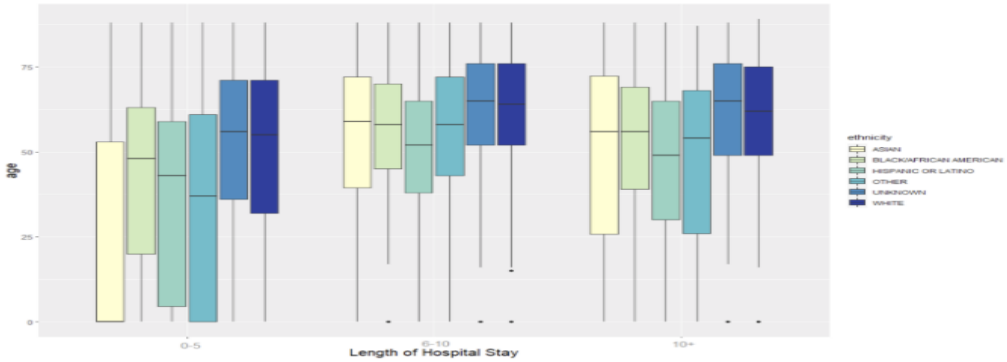


Fig. 2: Patient length of stay (LOS) by age, across various patient ethnicity

B. Data Preprocessing

The data pre-processing comprises of the following steps:

- *Missing Value Imputation:* In the dataset, there are no null values for the continuous variables, whereas there were null values for few categorical variables "admit diagnosis", "religion", "marital status". The null entries for the given categorical variables were replaced by "OTHER", "NOT SPECIFIED", "UNKNOWN" respectively.
- *Dropping the Non-Predictive Variables:* We dropped 3 non predictive variables in the dataset, "patient id", "ad- mit procedure", "LOS group".
- *Re coding the other categorical variables into fewer categories:* Categorical variables with various subcategories were reduced to fewer categories. Admit Location variable had different subcategories like "Info not available", "Transfer from skilled nurse, Transfer within this facility, Transfer from Hospital","HMO Referral /SICK". All the subcategories like "Info not available" were assigned as "Unknown" while "TRANSFER" was assigned to various transfer locations as subcategories. The Table I shows the re coded category for different categorical variables.
- *Categorizing our target variable:* Initially our target variable was a continuous variable, but as per our problem statement we converted it into different segments /factors (sub categories).
- *RFE technique for Feature selection:* There were 28 variables in our dataset (including the target variable). After the data pre processing part we have 24 variables, so to finalize the variables for our modelling part we implemented the Recursive Feature Elimination technique [13]. We used "treebagfuncs" that is Xgboost [14] function, and the num- ber of folds were 10 and method was "repeated cv". After implementing the RFE technique by the above criteria we got 16 important features out of which 11 variables were used for the modelling as seen in the Fig. 3 . The 11 variables used for modelling are age, gender, admit type, admit diagnosis, num transfers, num notes, num diagnosis, num callouts, num procedures, num drugs, LOS days.

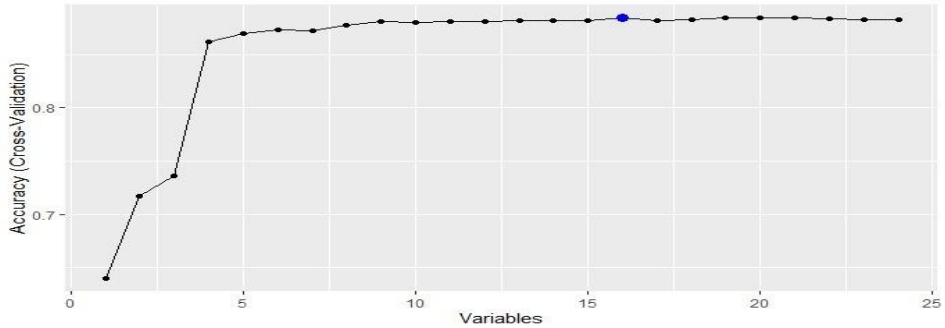


Fig. 3: Recursive Feature Elimination technique

ISSN No:-2456-2165

- *Further Exploratory Data Analysis:*

✓ Length of stay v/s Age: As per the Fig. 4 for the very short stays, the median age is 50 that means persons with median age of 50 years had a very short stay at the hospital. Whereas the people with a median age above 65 had the longest stay at the hospital

✓ Age v/s Length of Stay (For both Males and Females): As per Fig. 5, females with a median age of 50 had a very short stay at the hospital. Whereas both the Males and Females with a median age of above 65 had the longest stay at the hospital.

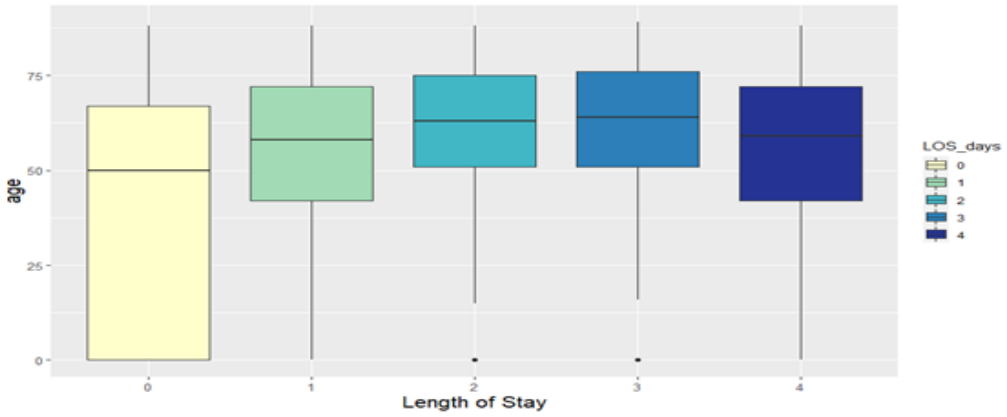


Fig. 4: Patient length of stay (LOS) v/s Age.

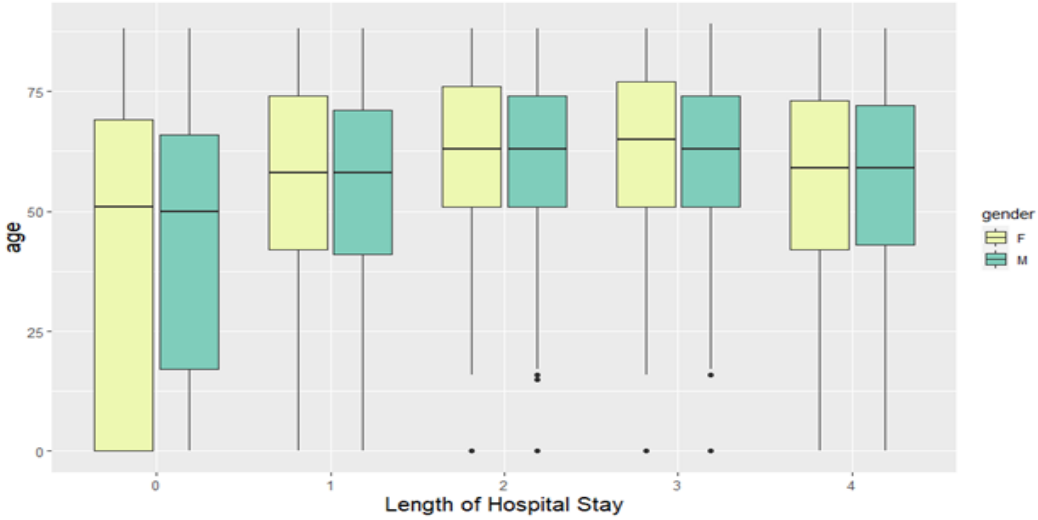


Fig. 5: Age v/s Length of Stay(LOS)

C. *Modelling Approach*

- *Approach 1:* We built a classification model, to predict the output variable LOS(Length of Stay)/. Various classifica- tion model, like Logistic Regression [15] [16], Random Forest [17], and Boosting [18] were implemented on the unseen dataset.
- *Approach 2:* As the target variable is continuous we can use the regression techniques as well , like multiple linear regression. Multiple linear regression [19], uses several explanatory variables to predict the outcome of a response variable. Hence, the patient's length of stay can be modelled as a linear function of multiple variables.

VI. RESULTS

Initially, the target variable (LOS) had 3 classes: short (0- 5), medium (5-10), and long (>10 days). The performance of all three models Logistic Regression, Random Forests, and Boosting, is presented below. With three target classes, all models exhibited subpar performance with an accuracy below 50%. The Gradient Boosting model slightly outperformed the other two models with an accuracy of 45.71%, an error rate of 54.28%, and a 95% confidence interval in the range of 0.4523.

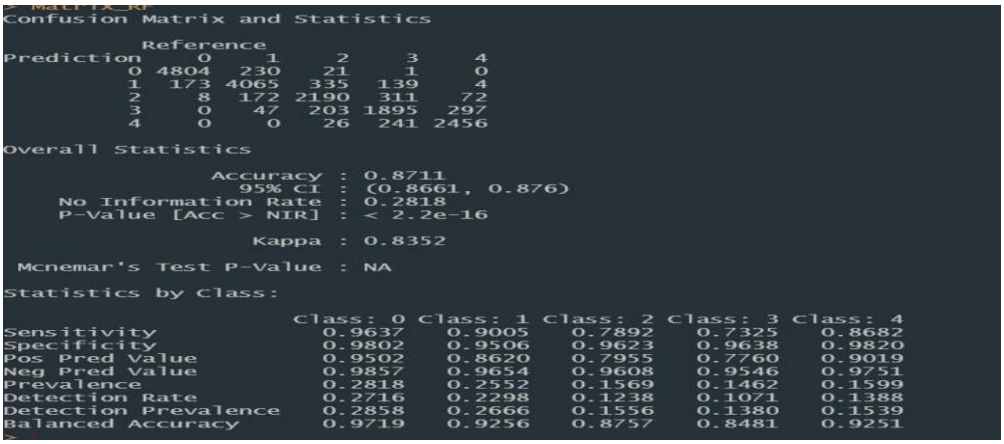
Table 1: Recoding Categorical Variables

Variable	Original Category	Recoded Category
ADMIT	INFO NOT AVAILABLE TRSF WITHIN FACILITY, TRANSFER	UNKNOWN TRANSFER
RELIGION	UNOBTAINABLE CATHOLIC, JEWISH, PROTESTANT QUAKER, NOT SPECIFIED	NOT SPECIFIED OTHER
ETHNICITY	ASIAN CHINESE, ASIAN-THAI, ASIAN-INDIAN, ASIAN- VIETNAMESE, ASIAN-JAPNESE WHITE-RUSSIAN, WHITE-OTHER, WHITE-OTHER EUROPEN	ASIAN WHITE HISPANIC-
MARITAL STATUS	LIFE PARTNER UNKNOWN DIVORCED,SEPER	MARRIED UNKNO WN
ADMIT DIAGNOSIS	CORONARY ARTERY DISEASE\\CORONARY ARTERY BYPASS GRAFT / SDA UPPER GI BLEED, LOWER GI BLEED, UPPER GASTRO INSTESTINAL BLEED	CORONARY ARTERY DISEASE GASTROINTESTINAL BLEED

to 0.4620. The confusion matrix metrics, such as precision and recall, also displayed significantly poor results, with none of them surpassing the 70% threshold in any of the three models. To further enhance these results, we expanded the target variable into two additional classes, resulting in a total of 5 classes. In this scenario, the performance improved signifi- cantly for all three models. As seen in the table below, Random Forests outperformed the other two models with an accuracy of 87.11%, an error rate of 18.37%, and a 95% confidence interval in the range of 0.8661 to 0.8760 as seen in the Fig. 6. clearly indicates that our Random Forest model is successfully predicting the hospital length of stay for the short days class. Finally, we considered the target variable as continuous and applied a linear regression model for prediction. However, the outcomes from this model, as depicted below, were notably inferior compared to the classification models with 3 classes. The Multiple R-squared value was computed to be 0.0971, while the adjusted R-squared was 0.09702.

Since the classification models outperformed the regression model, we opted to use classification as our final model.

VII. DISCUSSION



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4
         0 4804  230   21    1    0
         1  173 4065  335  139    4
         2    8  172 2190  311   72
         3    0   47  203 1895  297
         4    0    0   26  241 2456

Overall Statistics

               Accuracy : 0.8711
                 95% CI : (0.8661, 0.876)
    No Information Rate : 0.2818
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8352

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9637   0.9005   0.7892   0.7325   0.8682
Specificity            0.9802   0.9506   0.9623   0.9638   0.9820
Pos Pred Value         0.9502   0.8620   0.7955   0.7760   0.9019
Neg Pred Value         0.9857   0.9654   0.9608   0.9546   0.9751
Prevalence             0.2818   0.2552   0.1569   0.1462   0.1599
Detection Rate         0.2716   0.2298   0.1238   0.1071   0.1388
Detection Prevalence   0.2858   0.2666   0.1556   0.1380   0.1539
Balanced Accuracy      0.9719   0.9256   0.8757   0.8481   0.9251
```

Fig. 6: Random Forest Model Summary.

Other performance metrics, such as precision and recall, also saw significant improvements for all three models. In the case of Random Forest, precision and recall rates were over 75% for all classes and even reached as high as 95-96% for the short and very short days classes. It was observed that precision was highest for the short days class at 95% and lowest for the long days class at 77%. Even the sensitivity (recall) was high for the short and very short-day classes. This

VIII. LIMITATION

Applying predictive analysis to healthcare settings is chal- lenging due to the numerous unforeseen parameters that can affect model results. Despite our best efforts to leverage the data for generalizing the model and quantifying the cost of different lengths of stay, unforeseen parameters, particularly in emergency cases where admission diagnosis may not accu- rately identify patients' underlying

ISSN No:-2456-2165

conditions pose significant challenges. Given the complex and sensitive nature of this domain, it's challenging to generalize the models extensively. It's important to note that the data used for this analy- sis comprises admissions to Beth Israel Deaconess Medical Center in Boston, Massachusetts, one of the highly regarded hospitals in the United States. Consequently, the patient data we observed were predominantly from specific ethnic groups, primarily covered by private insurance or Medicare, with a median patient age of 59 years. Notably, a majority of the patients were classified as emergency room admissions, and the mortality rate in these records was significantly higher than the US average mortality rate for emergency room admissions. This dataset's patient profile indicates that this sample of patient records does not accurately represent the wider demographic composition of patients across the United States. Therefore, caution must be exercised in generalizing the results of our analysis to the broader US population. Additionally, it's essential to consider that this dataset includes patient records from 2001 to 2012, so recent advancements in hospital infrastructure, diagnostic capacity, and resource capacities may not be reflected in this data.

A significant limitation of classification models using healthcare data is the lack of comprehensive measures. In prioritizing feature validity over model accuracy, we ended up with a total of 11 variables, of which nine are categorical (admit type, admit diagnosis, num transfers, num notes, num callouts, num procedures, num drugs, LOS day), one is binary (gender male), and one is continuous (age). While the MIMIC III database [6] offers more detailed information on patient-caretaker interactions, we chose to utilize only the average quantity of these measures rather than the results of tests, procedures, and patient notes. To enhance the performance of classification models, we can incorporate additional critical metrics such as pulse rate, heart rate, oximeter readings, temperature, nerve and pain reflexes, as well as other general indicators like pre-existing medical conditions at the time of admission.

IX. CONCLUSION

The primary aim of this project is to comprehend the diverse range of patient data available and generate a model to understand and visualize the underlying patterns. We intend to utilize these models to predict a patient's hospital length of stay, thereby reducing the burden of monotonous hospitalization efforts. Leveraging retrospective data from the MIMIC III dataset [6], we successfully developed classifications to comprehend potential underlying factors that can influence a patient's length of stay. After comparing the performance of different predictive classification models, we concluded that given the current limitations of the dataset, the present models can effectively predict a patient's length of stay within the short length range.

This project has provided us with valuable insights into the complexity of analytical tools used in healthcare analytic and the implications these models may have in practice. It has also shed light on the challenges one can face when attempting to apply purely data-driven approaches in the medical domain. Predictive modeling holds significant potential for contributing to advancements in hospital resource management, diagnostic analysis, and even early detection of diseases and disabilities.

To operationalize these academic predictive models, careful consideration should be given to evaluating the pros and cons they may bring in practical circumstances. The model should be designed to analyze and interpret circumstances in a manner that benefits all the involved communities, including patients and medical staff.

X. FUTURE WORK

In terms of future work, we plan to collaborate with multiple doctors to incorporate their insights from initial diagnoses into our model. This collaborative effort aims to enhance the current model's accuracy and enable a perspective that integrates medical expertise rather than relying solely on a data-based approach.

REFERENCES

[1]. P. E. Petersen and H. Ogawa, "The global burden of periodontal dis- ease: towards integration with chronic disease prevention and control," Periodontology 2000, vol. 60, no. 1, pp. 15–39, 2012.
[2]. B. Friedman, H. J. Jiang, A. Elixhauser, and A. Segal, "Hospital inpatient costs for adults with multiple chronic conditions," Medical Care Research and Review, vol. 63, no. 3, pp. 327–346, 2006.
[3]. H. Catalyst, "Interview with stanford cio, carolyn byerly: Launching a clinical data warehouse in months," Health Catalyst, 2016.
[4]. M. Hassan, H. P. Tuckman, R. H. Patrick, D. S. Kountz, and J. L. Kohn, "Hospital length of stay and probability of acquiring infection," International Journal of pharmaceutical and healthcare marketing, vol. 4, no. 4, pp. 324–338, 2010.
[5]. T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," Journal of global health, vol. 8, no. 2, 2018.
[6]. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," Scientific data, vol. 3, no. 1, pp. 1–9, 2016.
[7]. A. S. Fauci, H. C. Lane, and R. R. Redfield, "Covid-19—navigating the uncharted," 2020.
[8]. T. P. Velavan and C. G. Meyer, "The covid-19 epidemic," Tropical medicine & international health, vol. 25, no. 3, p. 278, 2020.

[9]. L. Yang, S. Liu, J. Liu, Z. Zhang, X. Wan, B. Huang, Y. Chen, and Y. Zhang, "Covid-19: immunopathogenesis and immune therapeutics,"
[10]. Signal transduction and targeted therapy, vol. 5, no. 1, pp. 1–8, 2020. [10] Z. Zhang, D. Yang, and Y. Zhang, "Disease diagnosis based on multi- view contrastive learning for electronic medical records.," IAENG International Journal of Applied Mathematics, vol. 53, no. 3, 2023.
[11]. R. Long, D. Yang, and Y. Liu, "Diseasenet: A novel disease diagnosis deep framework via fusing medical record summarization," IAENG International Journal of Computer Science, vol. 49, no. 3, 2022.
[12]. Z. Lin, D. Yang, H. Jiang, and H. Yin, "Learning patient similarity via heterogeneous medical knowledge graph embedding," IAENG Interna- tional Journal of Computer Science, vol. 48, no. 4, 2021.
[13]. X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in Sixth International Conference on Machine Learning and Applications (ICMLA 2007), pp. 429–435, IEEE, 2007.
[14]. T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., "Xgboost: extreme gradient boosting," R package version 0.4-2, vol. 1, no. 4, pp. 1–4, 2015.
[15]. R. E. Wright, "Logistic regression.," 1995.
[16]. D. Bo¨hning, "Multinomial logistic regression algorithm," Annals of the institute of Statistical Mathematics, vol. 44, no. 1, pp. 197–200, 1992.
[17]. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," Journal of chemical information and computer sciences, vol. 43, no. 6, pp. 1947–1958, 2003.
[18]. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189–1232, 2001.
[19]. P. Burton, L. Gurrin, and P. Sly, "Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling," Statistics in medicine, vol. 17, no. 11, pp. 1261–1291, 1998.