A Three-stage model to predict hospital length of Stay and death using Machine learning Algorithms

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Technology
in
Computer Technology

Submitted by

Somnath Ghosh

Registration No. : 160086 of 2021-2022

Examination Roll No: M6TCT24012

Class Roll No: 002110504015

Session: 2021-2024

Under the Supervision of

Prof. Debotosh Bhattacharjee

Department of Computer science and Engineering
Jadavpur University,
188,raja S.C. Mallick Rd,
Kolkata-700032,
West Bengal, India

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Certificate of Recommendation

This is to certify that this is bonafide record of the project entitled **"A Three-stage model to predict hospital length of Stay and death using Machine learning Algorithms"** submitted by **Somnath Ghosh** (University Registration Number 160086 of 2021-2022; Class Roll No. 002110504015; Examination Roll No. M6TCT24012) is hereby approved of a creditable study of a technological subject carried out under my supervision and presented in a manner satisfactory to warrant its acceptance for partial fulfillment of the requirements of the degree of Masters of Technology in Computer Technology. The research results presented in the thesis have not been included in any other submitted for the award of any degree in any other university or institute.

Prof. Debotosh Bhattacharjee

Department of Computer Science &Engineering,
Jadavpur University.Kolkata-32, India
(Supervisor)

Counter signed:

Prof. Nandini Mukherjee,
Head, Department of Computer Science and
Engineering, Jadavpur University,Kolkata-32,India

Prof. Dipak Laha,
DEAN, Faculty of Engineering and Technology,
Jadavpur University,Kolata-32,India

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Certificate of Approval

This is to certify that the thesis entitled **"A Three-stage model to predict hospital length of Stay and death using Machine learning Algorithms**" is a bonafide record of work carried out by **Somnath Ghosh** (Registration Number 160086 of 2021-2022; Class Roll No. 002110504015; Examination Roll No. M6TCT24012) in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Technology** in the Department of Computer Science and Engineering,Jadavpur University during the period of October 2021 to May 2024. It is understood that by this approval the undersigned do not necessarily endorse any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose of which it has been submitted.

_____

Signature of Examiner 1
Date:

_____

Signature of Examiner 2
Date:

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that the thesis entitled **"A Three-stage model to predict hospital length of Stay and death using Machine learning Algorithms"** contains a literature survey and original research work by the undersigned candidate as part of his Degree of **Master of Technology** in **Computer Technology.**

All information has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Somnath Ghosh

Examination Roll No.: M6TCT24012

Registration No.: 160086 of 2021-2022

Thesis Title: A Three-stage model to predict hospital length of Stay and death using Machine learning Algorithms

Signature of the Candidate with date

ACKNOWLEDGEMENT

I owe my deep gratitude to my mentor and thesis guide **Prof. Debotosh Bhattacharjee**, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India,who took a keen interest in my thesis work and guided me all along till the completion of my thesis work by providing all the necessary information for developing a good end-to-end solution.I am extremely thankful to him for his assistance and dedicated involvement in every step throughout the process.His constant suggestions and editorial enhancements have been an everlasting source of motivation and inspiration to me.Working under person with eternal wisdom yet a kind, a simplistic attitude has been an honour.

I would like to thank **Prof. Mahantapas Kundu**, Coordinator of **Centre for Microprocessor Applications for Training, Education and Research (CMATER) Lab**, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for providing me necessary and valuable suggestions.

I would like to thank **Prof. Nandini Mukherjee**, Head of the Department, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for extending his help to complete this thesis.

Most impotantly, none of this could have been happended without the love and support of my family.To my father **Mr. Bimal Kumar Ghosh** , my mother ,**Mrs.Ruma Ghosh** it would be an understatement to say that their unconditional love and encouragement have always helped me in my need.Without their forbearance and whole hearted support, this thesis would not have been able to see the light of day.

Finally, I would like to thank all my professors at Jadavpur university,my classmates, especially **Mr. Amartya Ray** , and the many well wishers who have been a constant source of help and inspiration to me during my thesis work. My thanks and appreciation to all of them for being a part of this wonderful journey and making this possible.

Somnath Ghosh
Examination roll No: M6TCT24012
Dept. of Computer science & Engineering
Jadavpur University
Kolkata,india

Abstract

Globally, healthcare administrators are working to reduce healthcare costs without sacrificing the quality of service provided. The most significant portion of health expenses is related to hospitalization. Length of stay directly impacts the hospitalization cost and resource management. Researchers have been trying to find the relevant factors to accurately identify LOS for the last few years. The early detection of the death of a patient in the hospital is also an essential aspect of hospitalization to manage the hospital resources better to give ultimate care to severe patients. This research developed a strategy to forecast the number of hospitalization days and deaths in a community using the USA Sparcs de-identified data repository. The data is not individually identifiable.

The dataset consists of 2343569 instances and 34 features**.** This study proposes a three-stage model for improved hospital length of stay prediction: admission, post-admission, and discharge stage. These stage-wise models will help the patient and hospital management correctly predict LOS from the beginning of hospitalization to discharge. This study suggests a paradigm for predicting patient length of stay (LOS) using various machine learning (ML), Deep learning (DL), and ensemble techniques. This stage-wise prediction will be performed using a weighted probability averaging ensemble model. The weights of the individual classifiers will be selected based on their performance through a grid search method.

The proposed ensemble model has outperformed the individual base ML models. The proposed ensemble model has improved balanced classification performance across stages, from admission to discharge. For death prediction, we propose a single-stage admission model.This model has a unique seven days death prediction implementation.This prediction will be performed based on Machine learning and deep learning models. The death-to-survived ratio is highly imbalanced in the sparcs dataset. We have evaluated our models using both synthesized data and original data. The performance of the models in the actual data is not so good, and the models performed better in the synthesized environment. We have implemented few imbalanced data models in the original data, and their performance is quite good compared to generic ML models. The death prediction in the post-admission phase can be implemented using a combination of ensemble and synthesized or imbalanced data models.

Contents

Abstract ..... **v**

Chapter 1 ..... **1**

Introduction **1**

1.1 What is Data mining? 1

1.2 Problems faced by the health care systems 2

1.3 Need for data mining in the medical domain 3

1.4 Introduction of structured tabular medical data 3

1.5 Various steps of medical data mining 4

1.6 Importance of features in medical data mining 5

1.7 Impact of synthesized data in medical imaging 5

1.8 Hospitalization phases 6

1.9 The objective of this work 6

Chapter 2 ..... **8**

Literature Review **8**

2.1 Overview 8

2.1.1 Hospital Length of Stay 8

2.1.2 In-hospital Death 9

2.2 Length of stay prediction models 9

2.3 Death prediction models 16

2.4 Summary 19

Chapter 3 ..... **21**

Proposed Methodology of LOS prediction **21**

3.1 Methods 21

3.1.1 Data repository 21

3.1.2 Data Analysis and Visualization 23

3.1.3 Data Pre-processing 28

3.1.3.1 Missing Values handling 28

3.1.3.1.1 Populate with Default value 29

3.1.3.1.2 Removal of features 29

3.1.3.1.3 Removal of missing rows 30

3.1.3.1.4 Populate missing ZIP code 30

3.1.3.2 Feature Engineering 31

3.1.3.3 Feature Selection 31
3.1.3.4 Normalization 32
3.1.3.5 Feature Transformation 32
3.1.3.6 Feature Encoding 32
3.1.4 Statistical relationship between features 33
3.1.5 Proposed Model 35
3.1.5.1 Stage-wise feature selection details 36
3.1.5.2 Machine learning models 37
3.1.5.2.1 Logistic regression 37
3.1.5.2.2 Random Forest Classifier 37
3.1.5.2.3 Adaboost 38
3.1.5.2.4 XGBoost 38
3.1.5.2.5 LightGBM 38
3.1.5.2.6 KNN 38
3.1.5.2.7 Artificial Neural Network 39
3.1.5.2.8 CatBoost 39
3.1.5.2.9 Ensemble Learning 39
3.1.5.2.10 Our Architecture 40
3.1.5.2.10.1 Algorithm 40
3.1.5.2.10.2 Equation of the weighted ensemble model 40
3.1.5.2.10.3 Pseudo code of the algorithm 41
3.1.5.2.10.4 Hyperparameter Tuning of base models 42
3.1.6 Evaluation criteria 42
3.2 Result 42
3.2.1 Admission phase model 43
3.2.2 Post-admission phase model 43
3.2.3 Discharge phase model 44
3.2.4 Feature importance 45
3.3 Discussion 50
Proposed Methodology for Revised LOS Prediction 53
4.1 Problem statement 53
4.2 Solution 53
4.3 Methods 54
4.3.1 Proposed Method 54
4.3.1.1 Naive Bayes 54
4.3.1.2 Tabnet 55

4.3.1.3 CNN 1D 55
4.3.1.4 Decision Tree 55
4.3.1.5 Our Architecture 56
4.3.1.5.1 Algorithm 56
4.3.1.5.2 Pseudo Code of the Model 57
4.3.1.5.3 Hyperparameter Tuning of base models 58
4.4 Evaluation criteria 58
4.5 Result 58
4.5.1 Admission phase model 59
4.5.2 Post-admission phase model 59
4.5.3 Discharge phase model 60
4.5.4 The tuning of the Weighted average Ensemble Model 60
4.5.5 Feature importance 63
4.6 Discussion 71
Chapter 5 73
Proposed Methodology of In-hospital Death Prediction 73
5.1 Methods 73
5.1.1 Data repository 73
5.1.2 Data Analysis and Visualization 75
5.1.3 Data Pre-processing 79
5.1.3.1 Missing Values handling 79
5.1.3.2 Feature Engineering 80
5.1.3.3 Feature Selection 80
5.1.3.4 Feature Normalization and transformation 80
5.1.3.5 Feature Encoding 80
5.1.3.6 Data balancing 81
5.1.4 Statistical relationship between features 82
5.1.5 Proposed Model 85
5.1.5.1 Machine learning models 86
5.1.5.1.1 RusBoost 86
5.1.5.1.2 Easy Ensemble 86
5.1.5.1.3 Balanced random forest Classifier 87
5.1.5.1.4. Ensemble Learning 88
5.1.5.1.5 Our Architecture 88
5.1.5.1.5.1 Algorithm 88

5.1.5.1.5.2 Pseudo code of the Algorithm 91
5.1.5.1.5.3 Hyperparameter tuning of base models 91
5.1.6 Evaluation criteria 92
5.2 Result 92
5.2.1 Experiment strategy 92
5.2.1.1 Pseudo code of iterative feature selection 92
5.2.2 Result Discussion 93
5.2.3 Feature importance 97
5.3 Discussions 100
Conclusion and Future Work 102
6.1 Conclusion and future work 102
References 104

List of Figures

Figure 1:Total charges and costs distribution 26
Figure 2:Plot of total cost versus length of stay 27
Figure 3:Length of stay versus plot of Emergency department, APR medical surgical description, APR risk of mortality, and APR severity of illness death 27
Figure 4:Plot of Age group vs. Length of stays..... 28
Figure 5:Stages-wise length of stay predictive model 37
Figure 6: Schematic diagram of the Los prediction model..... 41
Figure 7: Feature importance of Random Forest for Top 30 features at the discharge stage 46
Figure 8: Feature importance of the XGBoost model for the Top 30 features at the discharge stage.. 47
Figure 9:Feature importance of Xgboost model for Top 30 features at the post-admission stage 47
Figure 10: Feature importance of Random forest model for Top 30 features at the post-admission stage..... 48
Figure 11: Feature importance of the XGBoost model for the Top 30 features at the admission stage48
Figure 12: Feature importance of LightGBM model for Top 30 features at the admission stage..... 49
Figure 13:Feature importance of Catboost model for the Top 30 features at the admission stage..... 49
Figure 14: Performance evaluation of ensemble models in three phases 50
Figure 15: Process flow diagram of revised Length of stay prediction 57
Figure 16: Feature importance of Tabnet model for Top 30 features at the post-admission stage 64
Figure 17: Feature importance of catboost model for Top 30 features at the post admission stage stage 65
Figure 18: Feature importance of the Catboost model for the Top 30 features at the discharge stage. 65
Figure 19:Feature importance of Tabnet model for Top 30 features at the discharge stage 66
Figure 20: Process flow diagram of death prediction..... 90
Figure 21: Feature importance plot for death model using Rusboost model 98
Figure 22: Feature importance plot for death model using RF model..... 98
Figure 23:Feature importance plot for death model using LightGBM model..... 99
Figure 24: Feature importance plot for death model using Balanced RF model..... 99

List of Tables

Table 1: Description of the Sparcs data 21
Table 2: LOS groups..... 23
Table 3: Data distribution of sparcs against its individual levels 23
Table 4: Missing data overview 29
Table 5: Duplicate features 31
Table 6: Encoding mapping details 32
Table 7: Statistical analysis of the features with the target 33
Table 8: Three-stage model feature mapping..... 36
Table 9: Result across all the stages 44
Table 10: Los groups using 5 classes 54
Table 11: Different base model and ensemble model performances 61
Table 12: Tuning of ensemble model to produce optimum results 62
Table 13: Feature importance in admission phase LGBM and CB 66
Table 14: Feature importance in the admission phase for the XGBoost model 67
Table 15: Feature importance in Post admission phase for the Tabnet model and RF model 68
Table 16: Feature importance in Post admission phase for Catboost and XGBoost model..... 68
Table 17: Feature importance in the Discharge phase for RF and Xgboost..... 69
Table 18: Feature importance in the Discharge phase for tabnet and Catboost 70
Table 19: Data distribution of death and survival patients 74
Table 20: Analysis of Sparcs features for death patients 75
Table 21: Features correlation with death..... 83
Table 22: Post-admission stage death prediction model feature mapping 86
Table 23:Death model parameters used 87
Table 24: Performance of the death for all predictors 95
Table 25: Class-level performance of ML models 95
Table 26: Performance class level of base and ensemble models 96
Table 27: Performance of Rusboost with different feature sets 97

Chapter 1

Introduction

1.1 What is Data mining?

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. The word data mining is similar to the concept of mining coal or gold. Through data mining, valuable information about data will be extracted through a few steps known as data mining. Data mining will search the underlying patterns of the data to get the valuable information stored in it. Unfortunately, data mining, appropriately named "knowledge mining from data," is somewhat lengthy. However, in the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. However, in the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. The data we get in data as input will be impure or raw, and that data will be refined to obtain knowledge from it. Data mining has similar terms: knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging[47]. Data mining generally performs two types of work: prediction and descriptive analysis. Predictive analysis is a supervised technique to predict the unknown or future value of the target work, and descriptive analysis is an unsupervised technique in finding patterns describing the data that humans can interpret. The knowledge discovery using data mining is the sequence of the following steps:

1. Data cleaning is essential to remove noisy and inconsistent data. This step cleans the data to make it usable for the next steps.

2. Data integration: Many applications must integrate several data to form the actual data. Some applications use a single data source, so application integration is unnecessary.

3. Data selection: Data relevant to the analysis task are retrieved from the database. Task-specific data is collected for analysis.

4. Data transformation: The data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.

5. Data mining: It is an essential process where intelligent methods are applied to extract data patterns. The refined data collected is used here for predictive or descriptive analysis.

6. Pattern evaluation: The information from the mining step about the interesting data patterns is measured with exciting techniques.

7. Knowledge presentation: The visualization and representation techniques present mined knowledge to users. The user can get data insight from this representation, which helps to select and understand data easily[47].

The Data mining step mentioned in step 5 is sometimes called data mining as this step presents the understanding, but all the required steps are similar. Important data mining cannot be done on inconsistent data, so all steps form the data mining building block. Data mining is being used in many businesses following few are:

1. Sales: Using predictive models, the business companies can estimate product sales on a year, month, or weekly basis.
2. Marketing: Data mining can improve marketing by predicting a product's acceptability areas and factors.
3. Manufacturing: Data mining helps manufacturing by predicting cost, materials, and efficiency.
4. Fraud detection: Data mining helps detect fraud like credit card fraud transaction fraud.
5. Human resources: Data mining helps to predict hire and leave policies.
6. Health care: The application of data mining in healthcare is vast. Data mining is doing excellent work in this field, from diagnosis prediction to bed availability.

1.2 Problems faced by the health care systems

The health care system is the most essential part of a human's life. This is the backbone of human society. After their family, this is the second place that treats them seriously. Obstruction and challenges in the health system may severely impact our lives. Smoothness in health system processing is the central area of focus for the researchers. The health system is classified into outpatient visits and in-patient hospitalization. For both cases, the patient data is collected in two different ways.

Hospital data are taken from patients' information through a structured format and stored in a database for doctors' information to be shared by text in the form of prescriptions. In hospitals, to efficiently manage the patients and run the hospitals, a few things need to be taken care of; those are proper employee management, nurse shifting, doctor’s availability, performance evaluation of the doctor, Emergency department assignment, patient management, hospitalization cost, patient length of stay, beds availability, patient’s conditions evaluation, etc. This management or prediction cannot be run efficiently by any human or group of humans. It needs complex analysis and calculation. Humans cannot perform this activity efficiently. Using the existing data, they can get information about possible outcomes but cannot predict the actual data. The hospitals aim to create a system that can analyze the inputs and predict the activities mentioned above by utilizing hospital data repositories. Researchers are working on these medical grounds to identify the factors contributing towards the successful prediction of the staff mentioned above and trying to develop a predictive model based on the hospital data. This new research area will improve the precision of manual tasks and try to decrease the errors made by manual staff. Sometimes, when complex and massive data is imported and given to the existing manual framework, the error rate will be very high as complex data handling is a burden for humans. Medical insurance claims in the medical sector are also important; manual intervention in claim amount estimation or claim fraud detection is nearly impossible or a rigorous task. The industry felt an automated computer-based technique that would describe and predict healthcare tasks smoothly and accurately without human interaction.

1.3 Need for data mining in the medical domain

The healthcare industry is responsible for producing a large portion of global data. Making sense of that data can give any medical organization, insurance company, and patient a significant advantage. Data mining has been used in many industries to improve productivity, customer experience, and satisfaction and increase product safety and usability—medical data is contained in text form, image form, and tabular data. Image data can tell significant information just by seeing it, but not all information can be extracted similarly from text. However, no conclusion can be drawn just by witnessing the bulk data in tabular form. The relationship between each feature needs advanced computerized analysis and prediction techniques. That can be achieved through data mining. Data mining in healthcare has proven effective in areas such as predictive medicine, customer relationship management, fraud and abuse detection, management, and measuring the effectiveness of certain treatments. Humans cannot do these works because the underlying algorithm is complex and requires large volumes of data. The Role of data mining in healthcare is described below:

1. Identifying high-risk patients: Data mining helps identify individuals at high risk for specific diseases or health conditions. Data mining, for instance, can inform people at risk for several comorbidities and other conditions based on their medical history, family history, way of life, and other information.
2. Personalizing treatment plans: Data mining is utilized to find patients with critical states and help them customize treatment programs to meet each patient's unique needs.
3. Improving patient outcomes: Data mining can be used by healthcare providers to find variables linked to better patient outcomes.
4. Predicting epidemics[45]: Healthcare practitioners can utilize data mining to find trends in data that will help them predict the probability of upcoming outbreaks.
5. Improving Hospital Staffing: Data mining helps predict and analyze staffing to help hospital management.
6. Resource management: Data Mining helps to manage the hospital resources efficiently.
7. Health insurance: Data mining helps predict medical health insurance costs and insurance fraud to help insurance companies.
8. Remote patient monitoring: Data mining helps track patients regularly from a distance and helps detect patients' condition so they can intervene early.
9. Automated disease detection: Data mining helps automated disease detection based on available medical data.

1.4 Introduction of structured tabular medical data

Tabular data is structured data stored in a tabular format comprising rows and columns. Each row represents an instance or observation of the data, while each column represents a specific attribute or feature of that instance. The data contained in Databases, spreadsheets, and comma-separated values (CSV) files are referred to as tabular data. Analyzing tabular data is essential for various tasks, including decision-making, pattern recognition, and prediction.

Businesses extensively utilize tabular data to gain insights into customer behavior and support critical decisions regarding product development, pricing, and marketing strategies. Researchers also rely on tabular data to uncover patterns within scientific datasets and predict future events. With its structured nature, tabular data is a valuable resource for data-driven analysis and informed decision-making across diverse domains[46].
Tabular data contains different data types: numerical, categorical, textual, temporal, and spatial.
Tabular data has several benefits: structured organization in the form of rows and columns, ease of data entry, efficient data retrieval through Excel formula or SQL, ease of analysis, support of data visualization, and integration with database system. It supports machine learning formats. The tabular data format is the same as the machine learning format.
The medical dataset we referred to is the SPARCS USA data repository for 2017. This dataset has 2.3 million rows and 34 columns. The dataset contains demographic information about the patient, hospital-level medical data, and hospitalization costs. This dataset consists of both Numerical and categorical data. The primary diagnosis data might be captured at the outpatient visit or admission stage, and the secondary diagnosis code will be captured post-admission. It has procedures performed on the patient. Patient's severity and risk of mortality are captured in the data. This data contains both the emergency department and general department admissions. Lab tests and patients' private data are not included in this data, and this data is de-identified. The data consists of two columns that can be used as the target column for this predictive task: length of stay and Expired. The Expired feature is derived from the column patient disposition, which describes the patient's disposition upon discharge. The allowable values are Home or self-care, short-term hospital, skilled nursing home, hospice home, and expired, among others. Expired defines the patient who died at a hospital**.** This structured tabular data is input for our research work.

1.5 Various steps of medical data mining

Data mining tasks can be supervised or unsupervised. Our analysis work is based on supervised learning, which consists of pre-processing steps, data mining modeling, results, and evaluation. Supervised learning can be of two types: regression and classification. We will be discussing the classification in the following section. The detailed information about the processes are described below:

1. Pre-processing: The pre-processing includes missing data handling, noise removal, outlier detection, removal, scaling, and transformation.
2. Feature selection: Feature selection is an essential predictive modeling task. The features contributing to the prediction should be prioritized more than those with less predictive power.
3. Feature generation: New features generated from existing features can improve the power of the predictive models. Newly generated features can give significant insights into a patient's condition, which can be an essential factor.
4. Encoding: As a dataset can have mixed data types for categorical data, we need to transform it into a numerical one known as encoding.
5. Train test split: The dataset must be split so that a portion of the data will be utilized for training must be ⅔ of the whole data, and the rest will be utilized for testing. This training testing can be performed using hold out method, K-fold cross-validation, stratified K-fold validation, etc.

6. Modelling: This is the primary data mining step for knowledge extraction. The clean data will be used in this step to perform the prediction. Machine learning and deep learning algorithms like Linear models, Artificial neural networks, Naive Bayes, Decision trees, CNN, Tabnet, and ensemble models Random Forest, AdaBoost, XGBoost, and catboost can make the prediction. Imbalanced data models such as Rusboost, Easy ensemble, and balanced random forest models can be utilized for imbalanced data. Apart from the existing ensemble models, a hybrid ensemble model can be developed utilizing base models to improve the model's predictive power.
7. Validation: Model validation can be performed using the validation metrics accuracy, precision, recall,f1 score, and Auc. These metrics will define the model's performance on the data.
8. Hyperparameter Tuning: The top models will be calibrated based on the metrics to produce optimum results.

1.6 Importance of features in medical data mining

The features in medical data mining have a very significant impact on prediction. The prediction of medical data has some critical areas of research, such as disease diagnosis, risk prediction, and death prediction. Missing important factors or improper feature selection can harm the prediction, which can have serious issues. Missing an essential factor can indirectly decrease the prediction's precision, harming both management and the patients. Features like the mortality factors, lab test data, pre-existing conditions, BP level, and Sugar level are essential for the medical environment. Selecting and removing features in the medical domain requires expertise or a good understanding of business. The knowledge can come from experts' written documents or metadata of the features. For cancer detection, if we predict cancer patients as noncancer patients, that will have a severe impact. Proper factors contributing to prediction must be appropriately identified, and the significance of the features should be correctly defined. This study deals with the prediction of hospital length of stay and death prediction of a patient; for both studies, the dataset is considered to be the Sparcs 2017 dataset. The length of stay is directly related to hospitalization cost and patient care management for extended stays. Death prediction is critical and requires the hospital authority to take extra preventive care. So, wrong predictions of death patients as nondeath patients have a severe impact. So, essential features should be identified through proper business knowledge and metadata from the data.

1.7 Impact of synthesized data in medical imaging

Generally, medical data has some imbalance; some classes have more data, and some have less. For cancer patients, the number of cancer-detected patient count would always be less compared to noncancer patients. It is not always necessary to sample the data if the two classes' importance is not opposite. For critical diseases like cancer, the minority class that is cancer count should have significantly less, so the data mining model always tends to classify it into the majority class. The cancer patient will be classified as a noncancer patient for most of the time. To resolve this issue, we need sampling techniques to come into the picture. To give proper balance to the data, we can perform over and under-sampling to improve the minor class count. SMOTE (synthetic minority oversampling technique) is a popular way to enhance the minority count using the KNN algorithm. Random under sampling can be used to reduce the

majority count. The ensemble method can be also used to improve prediction accuracy for imbalance issues. This ensemble approach will not increase the data distribution, but the algorithm tries to handle the data imbalance by combining multiple weak learners to produce a good model. Always sampling is not a good solution when the classes are not of opposite significance as they are conveying similar information with varying groups.

1.8 Hospitalization phases

Patient visits the hospital for two purposes: one is a for planned check-up that is outpatient visit, and the other one is admission, which is an inpatient visit. The inpatient visit is associated with high expenditure and high resource engagement. The main focus given by the researchers is inpatient visits, so our area of interest is hospitalization cases. Hospitalization inpatient cases can be divided into three phases: admission, post-admission, and discharge. Each phase has its feature size limitation. Limited data will be collected at admission phase, mainly primary diagnosis and patient demographic data. The post-admission phase is critical and can be considered 1-2 days after admission. In this phase, proper diagnosis will be performed, and secondary diagnosis, procedures, and significant medical data will be added. The final phase is the Discharge phase. Several discharge-related features will be added in this phase, like final hospitalization cost and discharge destination. Each phase has an incremental number of added features, and each phase's predictive power differs in its capability to contribute to the prediction.

1.9 The objective of this work

Healthcare systems have started emphasizing efficient resource management and forecasting to assure optimal levels of care, lower associated costs, and enhance patient care. The main goal of hospital managers is to create an effective healthcare plan by distributing the human and physical resources necessary for a successful hospital operation. The most crucial feature is to accurately estimate how long a patient admitted to the hospital will stay. Many hospitalization lengths of prediction analysis have already been performed, primarily based on ICU [4] [26] patients or for specific categories. For ICU patients, some predictions used conventional techniques such as SAPS, APACHE II, and APACHE IV to evaluate the mortality risk. Most of the existing analysis has been limited to a specific hospital, department, or disease study. This work differs from the existing length of stay predictions and mortality prediction models; this prediction will be performed based on the USA SPARCS de-identified 2017 dataset. In this study, we propose a unique three-stage model, admission, post-admission, and discharge stages, to predict the hospital length of stay.

This new generation model does not stick to a particular phase; it explores all possibilities of Los prediction during a hospital stay. Three stages will have increasing features based on the available data in each phase. The feature selection of each stage is performed based on the business knowledge gathered from the metadata of SPARCS[55]. We have utilized various machine learning models such as Logistic regression, Naive Bayes, Decision trees, KNN, ANN, Random Forest, AdaBoost, XGBoost, LightGBM, Catboost, Tabnet, CNN 1D, and ensemble learning for our analysis. A weighted probability average hybrid ensemble model achieved the final prediction by combining the top-performing base ML models across three stages.

The hospital can perform death prediction along with the Length of stay prediction to identify the critical patients.The existing mortality predictive models are limited to disease-specific studies, or the studies are conducted in an ICU setup; few general setup models are present, but our study is generic considering both general and emergency admissions. Most studies use SAPS, APACHE, and SOPA scoring for ICU mortality predictive tools for better interpretation. We have developed a unique single-stage model to predict the death of hospitalization patients in the post- admission stage, so this study will be conducted using post-admission stage data. This death prediction model introduces a '7day' and '7+day' death model to better analyse the death based on the length of stay of the patient. The features significant to the post-admission stage are kept for our analysis of death. Charlson's comorbidities are accounted for in our study to verify the prediction of death. We have utilized various machine learning and deep learning models such as Logistic regression,Naive Bayes, Decision trees, KNN, ANN, Random Forest, AdaBoost, XGBoost, LightGBM,Catboost, Tabnet, CNN 1D, Rusboost, Easy ensemble and Balanced Random forest for our analysis. The final result was verified with a weighted probability ensemble model.

Chapter 2

Literature Review

2.1 Overview

The healthcare system plays a vital role in society. They are treated as the second home after a person's own home. They are always ready to serve you when you are in danger, just as our family is always with you. The health care system serves society by treating patients with an exchange of cost. The cost of the treatment increases with the kind of service provided by the hospital, criticality of the patient, number of hospital stays, facility preferred by the patient, bed taken, and doctor opted; those are important factors of the hospitalization cost. Another critical aspect of hospitalization is survival and death. Death is an unfortunate event for every living species. Hospitalization deals with this unfortunate event of death. The LOSs of death affect both the customer and hospital authority. The patient party's LOSs are indeterministic, but the hospital authority also faces tremendous LOSs in terms of reputation and safety for the patient. Moreover, the number of deaths in a hospital is inversely proportional to the number of admissions in the hospital. The hospital authority should be more careful about the death and predict the length of stay. Proper prediction of these two factors can create goodwill in the hospital, and more admissions can happen to the hospital. In the following section we will, we will introduce the two predicting factors, and the later section will present the past works on these two factors:

2.1.1 Hospital Length of Stay

The length of stay for hospitalization is vital for hospital authorities, patients, and insurance companies. The length of stay(LOS) of hospitalization is defined as the count of days between admission and discharge. The length of stay depends on the patient's condition. Higher lengths of stay will adversely affect the insurance companies and patients, resulting in a higher claim amount and discharge amount. Higher costs will also create a burden on the patient's family. Sometimes, it also has been seen that a smaller stay has higher hospitalization costs than a more extended stay. It depends on the severity of the patient's condition; critical disease will require more expensive treatment. Generally, it has been observed that the longer the LOS, the higher the associated cost. Older people tend to frequent hospital admissions; this results in longer LOS. Comorbidities and pre-existing conditions also severely impact the LOS. Early LOS prediction will help hospital authorities manage resources fruitfully based on the expected LOSs and the patient's condition. They can get an estimation of the bed utilization and the occupancy of the bed. They can also use this prediction for the new admission probability of a single day. The discharge rate of a single day can be calculated based on the LOS prediction.

However, early determination of the hospital's length of stay will severely help the hospital authority manage the hospital service flawlessly. The doctor's commitment to a patient is directly proportional to the LOS; a higher LOS means physicians are more involved with the patient. The hospital medicine supply is highly related to LOS. For minors, LOS's need for daily medicine will be less and longer. LOS needs more stock to be prepared. Longer LOS have higher staff utilization and engagement. Higher length results in higher Costs that affect the patient's party; early prediction of LOS will give them time to gather the required amount. Sometimes, it happens due to longer LOS and associated costs; in the middle of hospitalization,

a patient leaves the hospital and moves to a smaller, less expensive facility. Early detection will help the patient overcome this issue at the very initial stage. The insurance companies will also benefit from the LOS prediction. Insurance companies must pay higher claims than LOSs; slight LOSs have lower claims. Actual LOS prediction helps the insurance company and the patient keep the amount aside.

2.1.2 In-hospital Death

Death is an unusual and unexpected event for any living person. Everyone wants to run far away from death. The hospitals are the places where people go to treat them and try to minimize the risk of death. Several factors are associated with the death of any patient those are criticality of the ill, negligence of the patient party, negligence of the hospital authority, physician's availability, and lack of hospital resources. So, early prediction of the death of a patient in a hospital will reduce the mortality rate of any hospital by taking advance care and remembering the mentioned issues. The hospital authority can take extra preventive action by allocating more resources toward a critical patient; specialist physicians will be assigned to look after the patient, and they can allocate staff to care for the patient. The patient party will also benefit from the prediction by getting good treatment according to the assessed risk. Sometimes, it is demoralizing to get a death prediction, but at the same time, everyone will try to make the prediction a task to make it false. Fewer deaths in a hospital will create good morale in the hospital towards the customer, and it impacts their goodwill; as a result, more admissions will occur. The insurance companies also benefit from this prediction by assigning a scheme for those critical patients so that they can get proper benefits to save their lives. Sometimes, companies pay more premiums for critical patients. The death predictive model will not only predict death but also help prevent death. It has been observed that aged patients have more deaths in hospitals compared to young people. Pre-existing conditions and comorbidities also impact death very much. Proper treatment and care can be taken in the hospital to address these factors.

There have been few works related to the Hospital length of stay and death prediction in the past. Many researchers provide different models to predict the length of stay and death. They utilized different machine-learning techniques to achieve this. The previous studies related to Length of stay and death will be reviewed in the following section:

2.2 Length of stay prediction models

The work related to length of stay prediction will be discussed in this section:

2.2.1 Franck Jaotombo et al. [19] and his team used a French medical database to predict hospital length of stay using machine learning algorithms. This Journal conducts a study on a French hospital database for 2015. They have included all acute care hospitalizations for patients older than 18 years, and LOS is more than one day. They have classified the Length of stay into binary classifications of short and long LOS. They used Tukey's criteria to get the outlier in Length of stay and found 14 days as the cut point. Based on this point, they have divided the binary variable into short stay and prolonged length of stay. The data is imbalanced, with short-stay instances at 90% and PLOS at 10%. Due to data imbalance, they have not considered accuracy and f1-based measurement; instead, they have placed the area under the receiver operating characteristics curve(ROC, AUC). The dataset consists of demographic, clinical data, and hospitalization factors. Five machine learning algorithms have been implemented to predict the LOS and PLOS Logistic regression (LR), Classification trees(CT), Random Forest(RF), Gradient boosting(GB), and Neural network(NN). The Gradient boosting

algorithm has a higher AUC ROC value of 0.8101, and the LR model has a lower value of 0.7947. This study concludes by using the GB model to predict the LOS.

2.2.2 Lingtao Chen et al.[21] and his team predicted hospital stay duration for sepsis patients using six machine-learning techniques. The dataset is collected from a Dutch hospital. They used the outlier detection technique Z score method to identify the outlier and prepared two datasets, one with outliers and the other without outliers. The implementation was done on the Bupa R package and Weka tool. They have compared Lightgbm, XGBoost, Neural network, Random forest, KNN, and linear regression. They opted for regression analysis for the prediction and compared prediction accuracy, RMSE, and MAE for the models. For outlier inclusion, the XGBoost model had a higher prediction accuracy of 73.87 %, Lightgbm had a lower MAE of 3.66, and the Neural network had the lowest RMSE of 5.93. For excluding outliers, KNN had the higher prediction accuracy of 80.50 %, Lightgbm had the lowest MAE of 2.40, and KNN had the lowest RMSE of 3.27.

2.2.3 Nuno Caetano et al. [9] and his team predicted Hospital length of stay using data mining of Portuguese hospitals at the admission stage. The dataset contains 28 variables that seven physicians selected. After removing irrelevant variables with higher missing values and applying the outlier removal technique, they left with 14 variables; they used the hot deck missing value imputation technique to fill in the missing value. The analysis was performed on regression settings to predict the actual length of stay days. They compared six regression models: AP(Naive model), MR(Multiple linear regression), DT(Decision tree), ANN(Artificial neural network), SVM(Support vector machine), and RF(Random Forest). Three classic regression metrics are compared: R2, RMSE, and MAE. The RF model outperformed all other models with an R2 score of 0.813, MAE of 0.224, and RMSE of 0.469. They found that the three most essential variables in the length of stay prediction are episode type, inpatient service, and medical specialty.

2.2.4 Merhan A. Abd-Elrazek et al. [4] and his teammate predicted hospital length of stay in ICU units using general admission features. The study was conducted in the ICU using general admission features. The study has been conducted based on the admission stage only. The features available for prediction are collected at the admission stage; no post-hospitalization features are used in the task. The data is collected from the Suez Canal University Hospital ICU. The patient's LOS has been divided into three classes: class 1 is within two days, class 2 is within seven days, and class 3 is more than seven days. Feature selection techniques Laplacian score, Fsspectrum, MCFS, Greedyfs, and feature similarity were used for feature selection. Nine machine learning algorithms have been used in this study: Neural network, CART, Bagging, Random Forest, SVM, KNN, Naive Bayes, and Fuzzy logic. They have addressed the Bias variance problem that occurred in different ML models. Some models faced bias problems and some variance problems. For Bias reduction, they add a few more features into the set and test the ML algorithms, resulting in bias reduction. For the variance problem, They tried three techniques: increasing the dataset size, Decreasing the number of features, and resampling using bagging. The resampling with bagging was achieved by three experiments that altered the training and testing data distribution. After the bias-variance reduction problem, the Fuzzy logic outperformed other algorithms with 92% accuracy; a CT holds the second position at 90.04%.

2.2.5 . Md. Shahid Ansari et al.[17] and his team predicted prolonged length of stay for dengue patients and identified predictors impacting dengue. The study was conducted at Max Hospital, which is based in India. The features collected are within 24 hours of hospital admission, and regular lab test data are added until discharge. They have divided the length of stay as short stay within five days and extended stay as >5 days. Regression imputation was applied as part of the missing value treatment for nominal or ordinal type columns. Using a feature generation process, they generated 389 features from the existing 40 based on lab test data. Feature selection techniques such as information value, RFE with logistic regression, chi-square test, and L1 feature selection were used to select the optimum features. Logistic regression with elastic net setting and random forest model was used for prediction; the Lr with elastic net has outperformed the Random forest model with AUC 0.75 compared to RF's 0.72. Thus, the LR with elastic net predicted better prolonged LOS in dengue patients. They also identified the top risky features for prolonged LOS with the help of the LR model and RF model's feature importance based on statistical tests, the Kruskal Wallis test, the Chi-square test, and Fisher's exact test.

2.2.6 Joseph Ebinger et al. [8] and his team predicted the hospitalization duration of COVID-19 patients using a machine learning algorithm. The data is collected from a hospital located in California. The study aims to predict prolonged LOS, defined as > 8 days. The study excluded the patients who died within eight days of hospitalization. They created three prediction models: 1. using the data from day 1, 2. Data from the first two days of hospitalization and 3.Data from the first three days of hospitalization. A total of 353 features were used for the prediction. They have used Charlson and Elixhauser scores to generate comorbidity-related features based on the ICD-10 standard code. A total of 42 models were trained and ranked based on their performance of the AUC metric. They have implemented meta-algorithms as ensemble models for the prediction along with base models, and for all three prediction tasks, the ensemble outperformed all other models. For the Day 1 and Day 2 models, the ENET Blender performed best; for the Day 3 model, the AVG Blender performed best in predicting the prolonged length of stay. The precision increased in each stage with new data added, and the highest was achieved on Day 3, with AUC values of 0.819. The predictor of each model is identified using feature importance on a scale of 0-1. The work is implemented using the data robot machine learning platform.

2.2.7 Using a neural network, Pei-Fang et al. [18] and his team predicted the length of stay for cardiology patients at the admission stage. This study was conducted in a Christian hospital in Taipei, Taiwan, in a cardiovascular unit for coronary atherosclerosis(CAS), heart failure (HF), and acute myocardial infarction (AMI). Pearson's correlation coefficient identifies the model's feature relationship and risk factors. This study will perform two stages of Length of stay prediction: pre-admission and pre-discharge. They have identified potential features for pre-admission and pre-discharge stages. The pre-admission stage consists of demographic information, primary diagnosis, and comorbidities, and the pre-admission stage includes interventions and payment information. An artificial neural network was used to make the prediction, and the neural network's architecture used for CAS patients differs from that of HF and AMI patients. The number of features for both models is different, so the input of the ANN's pre-admission and pre-discharge models was different. The ANN stage-wise prediction was compared with the linear regression model. The LOS prediction was made with actual

Length of stay data, so regression analysis opted for it. The effectiveness of the models was checked by the coefficient of determination(R2), Mean absolute error(MAE), and Mean relative error(MRE). For CAS patients, the accuracy ranged from 88.07% - 91.53%, MAE 1 to 1.11, and MRE 0.44 to 0.47. They defined the accuracy of LOS as being within one day from actual LOS, and it was observed that the LR model performed better than the ANN model for patients with LOS of 2 days. LR and ANN models cannot predict LOS for more than five days. For AMI and HF Patients, the performance of the pre-admission and pre-discharge model was unsuitable for the AMI and HF patients, so they included two days of deviation in the prediction, and the accuracy increased from 63.69 to 67.47 %. It has been observed that both the LR and ANN models performed better in predicting LOS between 8 and 11 days.in the predischarge model, ANN performed better than the LR model for LOS 3,5,6 and 7 days.ANN model can predict LOS, which is greater than 11 days. LR and ANN models can not correctly predict LOS greater than 18 days. The implemented ANN model has a log sigmoidal activation function in the hidden layer. The number of hidden layer neurons verified is between 10-30, and it was found that 13 is a proper number and used across all the models. The preadmission model was more effective in predicting the LOS than the predischarge model. The LR model performed slightly better than the ANN model due to the high training data used in the LR model compared to the ANN model.

2.2.8 Jan Chrusciel[15] et al. and his team predicted the length of stay using unstructured data; they compared the structured data collected at the hospital with the handwritten unstructured data. They mainly considered emergency department admissions in a French hospital with more than two-day emergency stays. The features collected from the emergency departments are patient demographics, context, and unstructured information. They used the UMLS(Unified Medical Language System) module, which the National Library of Medicine created, to extract the information from the text. The unstructured data extraction was done through the UMLS. They have added a few more features and compared two sets of features: i) Demographic data with structured data and ii) demographic data and unstructured data with a Random forest model for analysis. They have prioritized tree-based models over linear and kernel models to address factors like multicollinearity, overfitting, and identifying complex decision boundaries. The LOS is defined as a short stay within seven days and an extended stay greater than seven days; this is a binary classification prediction. The structured and unstructured data models' predictions were similar, with 75% in unstructured data vs. 74.1% in structured data. They have shown that by using unstructured data, the LOS of hospitalization can be achieved with similar accuracy to that of structured data. This study encourages the use of text data for medical data mining.

2.2.9 In this study conducted at a Spanish hospital in Madrid, Jesus Manuel Puentes Gutiérrez et al. [24] and his team predicted the length of stay across hospital departments. This study utilizes all the departments across the hospital and tries to predict LOS and the factors affecting LOS. A total of 45 features were collected, but 12 were selected using the chi-square test for further processing. The LOS days were split into groups to make the classification prediction. Each department was trained and tested using state-of-the-art machine learning algorithms (KNN, SVM, Naive Bayes, Decision Tree, Random Forest, RNN). They used two unique measurement criteria: 1. Mean error, a percentual deviation of predicted result with real values, and 2. Exact Success percentage that is accuracy. The top list of departments identified as Obstetrics, Ophthalmology, and Endocrinology. Random Forest algorithm obtained the best mean predicted value. This study focuses on the prediction based on individual departments and identifying the factors related to LOS for separate departments.

2.2.10 Chunling Li et al. [27] and his team predicted the length of stay in the intensive care unit on most minor shrinkage and selection operators. The data is collected from Sichuan Provincial People's Hospital with 72 features. The dataset includes Demographic, ICU, Surgical, Drug, and laboratory parameters. The LOS data is skewed to improve the model's performance; the LOS feature is transformed using Box-Cox transformation. The relationship between features has been explored using Apriori Algorithm and Pearson method. After performing all required preprocessing and feature selection, 45 features were left for mining. The LASSO algorithm is applied to predict the LOS based on selected required features. The result obtained using tenfold validation is with RMSE as 0.88 +- 0.13, MAE 0.87+-0.07, R2 0.35+-0.09 with external validation RMSE as 0.74 +- 0.12, MAE 0.82+-0.08, R2 0.33+-0.12. Apart from the whole data prediction, they made the LOS prediction on the two subsets of the Survivor data, which was released from the hospital, and the non-survivor who died there. For survivor data, the LASSO model predicted RMSE as 0.87 +- 0.11, MAE 0.86+-0.07, R2 0.26+-0.06 and for non-survivor data, the LASSO model predicted RMSE as 0.87 +- 0.16, MAE 0.90+-0.08, R2 0.26+-0.10. The LASSO model identified the risk factors associated with the two models, survivor and non-survivor.

2.2.11 Ashwani Kumar et al. [25] and his team predicted Hospital Length of stays for Surgical patients using a two-stage model. They have referred to a dataset from a hospital located in Victoria that has elective surgery patients. The study consists of LOS prediction of surgical patients to identify the group to which a patient belongs according to their LOS. They identified short, medium, and extended stays for their analysis. They selected CART(Classification and regression tree analysis) for their prediction. The CART or decision tree works on a single node at a time to split the data grouping of LOS. It is nonlinear, is not impacted by the curse of dimensionality, is nonparametric, is not impacted by missing values, and is less complex. They have created a two-stage model called Stage I and Stage II based on the available ICU admission status after surgery. Stage I consists of features available before surgery, and stage II consists of features available after surgery, along with stage I features. The study was conducted in the urology, general surgery, orthopedics, and gynecology departments. At first, the Stage I and Stage II models utilized CART regression tree analysis to cluster the LOS based on the continuous LOS value. They measured the relative and cross-validation errors for the four departments concerning the Random forest and KNN regression models. However, the error of the RF and KNN models is lower compared to CART, but for strategic planning, RF and other models use averaging to predict the impacts of the LOS probability distribution, so CART regression analysis is suited more for this modeling. In stage II, using the features from Stage I and the new feature of Stage II ICU admission status, the RF classification model is fitted into the dataset based on variable size of clusters, and the average misclassification rate of the surgery departs is 2.73, 0.90,0.38 and 0.84 % for general, orthopedic, gynecology and urology departments respectively. This study concludes that CART analysis is a helpful tool for clustering patients, and it reduces the relative CV errors up to 58.69% from stage I and overall further up to 9%.Though RF and KNN perform better to reduce errors, their prediction is less informative and averaged across the prediction. However, due to its superior performance, the Stage II model uses an RF classifier to explain the LOS prediction.

2.2.12 Mahsa Rouzabahman et al. [26] and his team predicted the Length of stay and death in the ICU by comparing the clustering method and supervised learning. The study was conducted on the MIMIC II dataset. The Length of stay study will be performed in the ICU, and the death prediction will be performed both in the ICU and after Leaving the ICU. K means clustering

was applied to the training set, varying from 2 to 100 clusters, resulting in 99 scenarios for each scenario in which regression models are fitted. For death, linear and logistic regression was applied, stating that linear regression performed continuous predictions based on values between 0 and 1. The final prediction was performed by rounding off to the nearest integer. The LOS prediction was performed using the Linear regression. Each patient in the test set was searched to fit into the clusters, and in the final assigned cluster, we calculated the accuracy and MAE for Death and LOS prediction. They also calculated the result of Linear and logistic regression for unclustered data. For death prediction, linear regression on clustered data gives higher accuracy than raw data, with a 2.2% improvement and an accuracy of 74.4 %, and for logistic regression, clustered data has an improvement of 0.1 %, having an accuracy of 75.3%. It has been observed that death prediction using three clusters maximizes accuracy while maintaining coverage. For LOS prediction, linear regression with clustering was used, with 3 hours (174 mins) of improvement concerning the unclustered data. The MAE calculated for clustered data is 7300 mins, a 2.5% improvement from the unclustered data. The MAE for LOS is lowest for 7-8 clusters using the clustering technique. They also verified the clustering with regression analysis on a cancer ICU dataset and got 1.8 days of improvement in MAE value concerning unclustered data. It has also been found that the boosted regression technique used in this study predicts LOS when the correlation between the variables is not that significant. Hence, the cluster-boosted regression technique is an essential area of future research for predictive analysis.

2.2.13 Xiangrui Zeng [10] predicted the length of stay for indoor patients using a Light gradient boosting machine. This study refers to the Sparcs 2017 dataset. The study performed LOS prediction using regression analysis. After pre-processing the required data, the study selected features using multiple statistical tests based on different data categories. The statistical research includes Pearson product-moment, spearman coefficient, Kruskal Wallis test, and Spearman correlation coefficient. This study contains six encoding methods in the prediction: Label encoding, count encoding, one hot encoding, target encoding, leave one out encoding, and proposed encoding technique. Out of that, the Proposed encoding technique outperformed all other methods. The proposed encoding techniques consist of one hot and label encoding together. This encoded data is used to train multiple ML regression models: ridge regression, lasso regression, Random forest, XGBoost, and Lightgbm. The light gum model outperformed all other models when the encoding scheme proposed method is applied. The Lightgbm model has an R2 value of 0.96 and an MSE of 2.231. This study assumes that the data present in this dataset is collected during the pre-hospital phase and that the prediction model is a pre-hospital model. This study uses all features present in the dataset for the prediction task.

2.2.14 Md. Mahbubur Rahman et al. [14] and his team predicted Hospital length of stay using a Federated approach using the Sparcs dataset. Federated learning is a privacy-preserving ML technique that uses multiple decentralized clients to perform the training with their data in their machine, and the trained model parameters are sent to the central server where the aggregation happens. The central server can make predictions based on the parameters trained by the clients without seeing their data. The individual clients do not send the data to the central server. Only model parameters are shared across the central server. This study utilized a flower framework for the federated learning task. This study uses Sparcs data based on ten hospitals" data as the ten clients for training. The trained parameters for ten clients are sent to the server for aggregation using the 'FedAvg' strategy. This study predicts LOS using regression analysis. Three Ml models are utilized for this purpose: linear, Lasso, and ridge regression. The

individual performance varies for each hospital, and the accuracies vary. For server-side validation, if we increase the number of clients from 2 to 10, the accuracy on the server side will vary. The Linear best when the number of clients is ten with an R2 value of 0.873, RMSE, and MAE 2.0320 and 1.389, respectively. The federated approach is a groundbreaking ML approach to protect the client's privacy and make predictions in the central server based on the client's trained parameters, keeping the real data private. In this study, no extra features are deleted from the sparks feature set for the prediction.

2.2.15 Raunak Jain et al.[34] and his team predicted the length of stay in hospitals using a machine learning algorithm. This study uses the Sparcs 2016 dataset for the LOS prediction. They used both regression and classification analysis to make this prediction. They have removed the total cost, charges, and patient disposition from the feature set, as these features are collected only at the discharge stage. So, this study was conducted during the admission phase. A distribution-dependent target ending was employed to improve the model's performance. The prediction was performed using a decision tree, random forest, and XGBoost. The XGBoost model has the highest accuracy compared to other models, with an R2 value of 0.44. The top two contributing features are APR DRG description and APR severity of illness description.

2.2.16 Yazan Alnsour et al. [33] and his team predicted patient length of stay using artificial intelligence to assist healthcare professionals in resource planning and scheduling decisions. This study uses the Sparcs dataset for its analysis. They have considered this prediction for the admission stage. This is a single-phase prediction considering the stage as the admission stage. They have removed features only available at the discharge phase, such as total charges, costs, patient disposition, abortion, etc. The features they have considered for admission phases are patient demography, primary and secondary diagnosis, procedures, etc. They utilized both classification and regression analysis to make this prediction. The Regression prediction is performed for short stays of up to 10 days, and the classification task is performed for all the LOS groups. It used XGBoost and Ann for prediction. They used the SMOTE oversampling technique for the classification task, and no sampling technique was implemented for regression. This study uses feature importance to validate the importance of a feature. The XGBoost model is used for this importance calculation; they used permutation importance to verify the importance of a feature. They have used mean decrease in accuracy(MDA) to interpret the significance of a variable. MDA is calculated using actual accuracy and shuffled accuracy(achieved by shuffling the test set randomly). The importance of permutation is proportional to the MDA. For the classification case, the XGBoost outperforms the ANN with an AUC of 0.85 vs 0.82 and an accuracy of 0.64 vs. 0.62. The AUC of XGBoost for category 6 shows the highest value of 0.93 compared to the ANN. For the regression analysis, category 0, LOS less than 6 days XGBoost performed better than ANN with RMSE and MAE as 1.07 days and 0.85 days, respectively. In category two, XGBoost performs better than ANn with RMSE and ANN as 1.34 and 1.12 days. The essential features identified for admission stages are the APR severity illness code, the APR medical surgical description, and the APR MDC code. The study shows that using admission time features, the XGBoost model performs best considering only 16 features.

2.3 Death prediction models

2.3.1 Guilan Kong et al.[35] and his team predicted the in-hospital mortality of sepsis patients in the ICU. This study was conducted for sepsis patients in the ICU using the MIMIC III dataset, an ICU database from the Beth Israel Deaconess Medical Center. The study includes patients aged between 18-90. The sepsis patients who have a SOFA(Sequential organ failure assessment) score greater than two are identified from the dataset. The inclusion of features is restricted to within 24 hours of admission. The study uses physiology, chronic health status, demographics, and other hospitalization-related variables, excluding medication or procedures. This study uses ML algorithms LASSO(Least absolute shrinkage and selection operator)), RF(Random Forest), GBM(Gradient boosting machine), and LR(Logistic regression), established scoring tool SAPS II(Simplified Acute Physiology Score) was employed for this prediction. The death percentage in the dataset is about 17.3%. Though the dataset is imbalanced, no significant error was noticed. The AUROC score has been considered for evaluation to handle the imbalance problem. The GBM model has the best overall performance and a good calibration with a Brier score of 0.104, AUROC 0.845, sensitivity 0.771, and specificity 0.755. The GBM model has poorer interpretability than RF. The LR model performed worse than the GBM model by all measures but performed better than the RF and LASSO models regarding the Brier score and AUROC. SAPS II model shows the poorest prediction performance. This study emphasized clinical application rather than methodology. This study claimed that GBM was compared with RF, LASSO, LR, and SAPS II models for the first time in sepsis mortality prediction.

2.3.2 Using machine learning models, Javed Zarei et al. [36] and his team predicted in-hospital mortality among inpatients with COVID-19. This study used Ahvaz Jundishapure University of Medical Sciences (AJUMS) data. The registry collects demographic data, signs and symptoms, patient outcomes, and PCR and CT results for 38 hospitals. This study created multiple feature sets of sizes 60, 40, 32, and 17 using different feature selection methods, such as the Chi-Square, Fisher exact, and Mann-Whitney U test, based on the importance of the feature utilizing the P value. The data was imbalance death percentage 16.06%, alive 83.94%. The performance of the models was not good with the data, so they used an oversampling technique to increase the data to 36.5 %, which is 49.98 % of deaths. The model prediction was performed based on the four feature sets and three datasets. The machine learning algorithms used in this study are MLP, CHAID(Chi-squared detection of automatic interaction, C5 decision tree, RF(Random Forest), SVM(support vector machine) with RBF kernel, and Bayesian network. The study shows that dataset 3 performs better than all other datasets, and the model selected in this stage is C5. The developed ensemble models lag compared to the C5 model. The performance was also evaluated with an external dataset, and C5 boosting models with feature sets 1 and 2 have better scores. A subpopulation bias analysis was performed on age and sex; it shows that females perform better than male patients, and the performance of the models decreases for older patients. The models with 60 and 40 features performed best, and the C5 model 60 features outperformed all other models. The C5 boosting model with 17 and 32 features has better external validation results. Subpopulation analysis suggests that the C5 model with 32 features has less bias. The essential features for death prediction are intubation, number of comorbidities, age, gender, respiratory distress, blood oxygen saturation level, ICU admission, cough, unconsciousness, positive PCR, and abnormal CT. This study on COVID-19 mortality prediction with sufficient data from 38 hospitals is a benchmark for mortality prediction. Though this study lacks much clinical data, it still produces good results.

2.3.3 Mehmet Kivrak et al. [37] predicted death status for SARS-COV2 patients with deep learning and machine learning methods. The data was refereed from the National Database of Drug Prescription and integrated with clinical chart information for hospitalized subjects. The data includes age, gender, chronic disease(hypertension, diabetes, etc.), COVID severity score, and some enzymes. The study has been performed on deep learning models (CNN) and machine learning models (Random forests, KNN, XGBoost). The importance of the feature was extracted using the Chi-square test and Kruskal Wallis test. The age variable was highly skewed, so Box-Cox transformation was applied. The XGBoost model outperformed all models with an accuracy of 99.7%, and precision, sensitivity, and specificity are 99.7,99.7, and 100 %, respectively. The deep learning model is second on the list, with an accuracy 97.15. The Kappa scores show that XGBoost and DL fit perfectly with Kappa scores of 0.91 and 0.82, respectively. KNN shows a good fit of 0.58, and RF has a value of 0.19, which is not a good fit. The identified top four important variables in DL and XGBoost are Covid-19 severity, Hypertension, COPD, and Gender for the DL model and Covid-19 severity, age, Hypertension, and Gender for the XGBoost model. The Lowest significant value of DL is diabetes and For XGBoost Cancer and COPD. The study has a novel approach to predicting death status using the XGBoost model, and factors affecting death have also been identified.

2.3.4 Dagmar F. Hernandez [38] et al. predicted in-hospital mortality after Transcatheter Aortic Valve replacement using Machine learning. This study used data from the National Inpatient Sample (NIS) database files. The Study includes the patients who are treated after TAVR. The dataset is imbalanced, with 4% dead patients and 96 % alive patients. An unbalanced oversampling technique is applied to make the distribution 50:50. A feature ranking method was implemented using regularized logistic regression with L2 norm. Multiple feature sets were considered 5/10/15/20/30/all based on the feature ranking in descending order. All these feature sets are considered for ML model training and evaluation. The exhauster comorbidity using ICD-9 is considered in the feature set for the prediction. Statistical tests for feature importance were evaluated using chi-square and 2-tailed student's test—four Machine learning algorithms: Logistic Regression, Artificial Neural Network, Naive Bayes, and Random Forest. The AUC score has been identified as the model selection criteria. Logistic regression has the highest AUC value of 0.92; however, other models also have good AUC values: ANN has 0.85, NB has 0.90, and RF has 0.90. The essential variables identified for this prediction are acute kidney injury, cardiogenic shock, fluid and electrolyte disorders, cardiac arrest, sepsis, dyslipidemia, hypertension, coagulopathy, current smoking, vascular complications, etc. This study of NIS TAVR is the second most extensive study on TAVR after STS/ACC TVT, but it outperforms all other studies with good performance using AI. This study is well-equipped for high- and extreme-risk patients but is less practical for low-risk surgical patients. This NIS TAVR scoring system uses AI and performs well for TAVr and any novel structural intervention.

2.3.5 Using machine learning, Jun Ke et al. [39] and his team predicted hospital mortality for patients with acute coronary syndrome. The study was conducted at Fujian Provincial Hospital, having an age > 18 years with a diagnosis of ACS. Exclusion criteria were pregnant women, malignant tumors, end-stage renal disease, severe liver disease and blood disease, and patients with noncardiogenic chest pain. The data includes demographic characteristics, comorbidities, thrombolytic therapy, laboratory test data, and physical examination data. The data has 1.88% death. The machine learning models utilized for this analysis were LR, GBDT, RF, and SVM. They used the Mann-Whitney U test, Chi-squared test, and univariate, multivariate logistic

regression analysis for statistical analysis. The gradient boosting Tree(GBDT) and RF (random forest) models performed best with AUROC 0.918 and 0.913, respectively. The calibration ability of the models was also good. Important features found using the GBDT and RF were NT-proBNP, D-Dimer, Killip, Ctnl, LDH, LVEF, LDH, age, etc. The study shows promising results. According to the existing ACS studies, the factors impacting death in ACS found by the survey are appropriate, and the models' performance is quite good.

2.3.6 Chi-Yung Cheng et al. [40] and his team predicted mortality in patients with sepsis using machine learning utilizing vital sign dynamics. The Chang Gung Medical system provided the data in the Five emergency departments. The study includes patients aged> 17 whose Length of stay is> 3 days, whose Number of admissions is less than three times, and patients with incorrect data. The data is imbalanced, so they used random sampling for the live patient to decrease the patient count. The five vital signs that were used for testing are systolic blood pressure(SBP), Diastolic blood pressure (DBP), heart rate (HR), respiratory rate(RR), body temperature (BP), and patient age and sex were also included. This study contains 6-48 hr lead times where for death, 6-48 hr before death is extracted as death instance and 6-48 before discharge is known as survival instance. The study used four lead times: 6,12,24, and 48. The vital signs are normalized, divided into 255, and transformed into images. Three ML algorithms, CNN, LSTM, and RF, were compared for this study. For the 6hr model, CNN performs best with AUC 0.84; for the 12hr model, CNN performs best with AUC 0.85, 24 hr CNN performs best with AUC 0.82; and for the 48 hr model, CNN is performing better with AUC 0.83. The CNN model performs best for this predictive task compared to LSTM and RF. This study shows that a predictive model for sepsis mortality can be accurately predicted using fewer features and only vital signs.

2.3.7 Mostafa Shanbehzadeh et al.[41] predicted mortality among COVID-19 patients using artificial neural networks. The data referred to by the study belongs to Ayatollah Taleghani Hospital, Iran, with 58 features and four main categories: demographic, clinical manifestations, epidemiological, and hospitalization indicators. The study is performed on patients with age > 18 years. For the data preprocessing step, imputation technique K means clustering was used in missing values. Chi-square and eta coefficient performed the univariate feature selection, and binary Logistic regression was used for multivariate analysis. A total of 13 variables are selected for prediction post-feature selection. This study uses neural networks for the prediction using two types of neural networks: backpropagation (BP) and distributed time delay (DTD). Different combinations of train test splits were evaluated; those are 70-30%, 50-50%, 80-20%, and 60-40% for better model interpretation. MSE (mean squared error) and ROC curve were checked for the model's performance. The 80-20% data split by BP ANN models with ten hidden neurons achieved the best performance. The BP model has a higher AUC of 0.901, whereas DLD has 0.866.

2.3.8 Tatiana Ledneva et al.[42] created a risk-adjusted model for in-hospital / 30-day stroke mortality analysis based on SPARCS 2013 data. The study was performed on the SPARCS 2013 data. This study includes primary diagnoses as strokes in CY2013 from all NYS article 28 hospitals. Three types of strokes are included in this study: Ischemic, subarachnoid hemorrhagic, and intracerebral hemorrhagic. The study includes patients aged> 18, childbirth, pregnancy, and puerperium. This study accounts for deaths occurring within 30 days post-admission or hospice. The discharges within 30 days are not included in the analysis. Of the 184 identified NYS hospitals, 116 had a stroke center, and 68 did not have stroke centers. A multivariate logistic regression model was used for this analysis, and c-statistics metrics were used to assess model performance. The model's performance was judged using the APR-DRG code, and one with an Elixhauser score was used instead of the APR-DRG Code. The APR-

DRG code model performed better at 0.70 vs 0.87 (c-statistics). The logistics model with backward selection was used for good model selection. The age 50-59, risk of mortality, and APR -DRG code were identified as good predictors. The c-statistics of this NYS risk adjustment model perform a good 0.87 compared to the AHRQ model at 0.89. A RAMR (risk-adjusted mortality Rate) was calculated for all hospitals to identify the risk associated with stroke mortality. The RAMR for non-stroke designated hospitals shows lower.

2.3.9 Patrick J. Mccormick et al.[32] and his team validated all patient diagnoses related to group mortality risk and severity of illness modifiers as measures of perioperative risk. The data was taken from the Mount Sinai Hospital for adult patients > 18 years and non-cardiac surgery patients. The data includes discharge disposition, ICD-9 diagnosis and procedure codes, and APR-DRG data. Charlson comorbidity index (CCI) is an essential measure of comorbidity severity for hospitalization cases. The APR-DRG system assigns APR-DRG-CODE, ROM(Mortality risk), and SOI(Severity of illness) based on diagnosis and procedures. The Charlson comorbidity index is calculated based on the diagnosis code. The Charlson scores are divided into two cases: one where it is considered continuous and the second where it is regarded as a class. This study aims to compare the APR-DRG subclasses with the popular Charlson index for the mortality risk prediction of surgical patients. The study was conducted using training and tests split as 50:50 each. The logistic regression model is implemented for the risk prediction. The input has five sets of input variables: Only ROM, Only SOI, ROM, and SOI together, CCI continuous, and CCI groups with in-hospital mortality as the target. Overall mortality data was 1.3 %. The performance metrics used are c-statistics and Brier score. The c-statistics for ROM were 0.974, ROM and SOi were 0.977, SOI only 0.965, and Charlson continuous 0.865. ROM and SOI subclasses for mortality prediction outperform the Charlson index for surgical patients. This study shows that using the traditional Charlson index for mortality prediction is not always good; factors such as APR-DRG ROM perform better in mortality prediction for surgical patients.

2.4 Summary

The abovementioned literature belongs to two case studies: length of stay and death prediction. This study identifies hospital-level factors for hospital length of stay and death prediction and verifies the predictive analysis using different Machine learning models.

1. In this study, LOS prediction will be based on classification. Regression prediction gives an exact value of predicted LOS, but it might not always be accurate. A small error can hugely impact the hospital's authority to manage the resources. The classification prediction will be much more informative with small groups of LOS. Multiclass and binary both have been explored in the LOS classification task. For our model, we have identified seven small groups for LOS; those are 0-5 days, 6-10 days, 11-15 days, 16-30 days, 31-60 days, 61-90 days, and > 90 days. Most studies performed except SPARCS data are limited to a single area, e.g., Department, hospital, and disease. Few studies proposed two/multi-stage models, but they are lacking with diverse data categories. The studies performed on SPARCS data are of single stage only. Previous studies predict LOS on SPARCS data at the admission stage [33] [34] [10] or at the pre-admission stage [10]. The studies identified features captured only at the admission and discharge stages [33]. Some were identified at the admission and admission stage[10]. Our idea is different from these existing studies, and we are introducing a new form of Los prediction using a three-stage model for the SPARCS data utilizing the fact that the data set has certain features that cannot be captured only at pre-admission, admission stage, and discharge stage. A few are Procedure code, mortality risk, illness severity [32], and birth weight. We have identified potential stages for prediction as admission, post-admission, and discharge stages. The prediction will be performed in each stage based on the varying sizes of

features captured at that stage. This stage-wise prediction model will benefit the hospital authority in verifying patient discharge status time by time. Stage-wise predictions will be obtained using a weighted probability average ensemble model and several machine learning models. This stage-wise hybrid ensemble method is unexplored in the length of stay prediction with potential actual small LOS groups. Some studies cut the extreme LOS [19, 21] groups into an average group, but our study also focuses on the extreme cases. Class imbalance is essential for predictive tasks; some studies [33] use the SMOTE technique to encounter this. In this study, no sampling method was incorporated, using the fact that synthetic data might influence the actual prediction; instead, we gave importance to the performances of different evaluation metrics to better understand the precision of the model.

2. In the next phase of the length of stay analysis, we will reduce the size of the stay grouping and make it a 5-class problem with potential groups 0-5 days, 6-15 days, 16-30 days, 31-60 days, and > 60 days. A few deep learning and machine learning models are evaluated: Tabnet, CNN 1D, Decision Tree, and naive Bayes. The ensemble model will consider these new models to build the weighted ensemble model.

3. This study also utilizes the SPARCS 2017 data to predict in-hospital death for the post-admission stage. Most of the studies on death prediction are limited to a single area, e.g., disease-based predictions like sepsis, COVID-19[35-41], etc. In this study, we introduce a generic model to predict inpatient mortality for around 280 primary diagnosis categories. This model will predict death within seven days of hospitalization, death after seven days, and the people discharged from the hospital. This analysis does not contain who died after discharge from the hospital. This study only accounts for the deaths that happen in hospitals. The selection of the post-admission stage is due to the fact that the post-admission phase has several essential factors that can give more predictive power to death status analysis compare to admission phase. The length of stay is a significant factor for hospitalization cases as we have considered the inpatient death prediction at the post admission stage, which is probably 1-2 days post-hospitalization. The length of stay can not be a part of post admission phase it is a discharge feature, so it has been removed from the feature set. We have removed all the features related to the discharge stage from our feature set. Few studies related to Sparcs data used either the Risk of mortality and severity of illness or the Charlson comorbidity index [42,32] to predict death separately. However, our analysis combined both factors into the prediction to build a robust death predictive framework using diverse data. Machine learning and deep learning algorithms are used to predict death and identify its relevant factors. The death in Sparcs data is low, that is, 2%. SMOTE oversampling followed by random under-sampling was performed before prediction to give proper balance. The three classes death prediction, ‘7daydeath’ model vs. ‘7+daydeath’ vs. ‘survived’ model is a new approach in death prediction.This study utilized several auto-data balance machine learning algorithms that work on unbalanced data and verify their performance. The results of those algorithms will be compared with the sampled base models. The top-performed model will be compared with a reduced feature set of sizes 10/20/30/30/40/50/all. Finally, a robust hybrid model using the top-performing model will be used for death prediction.

Chapter 3

Proposed Methodology of LOS prediction

3.1 Methods

The study aims to predict hospitalization length of stay using an incremental three-stage model. Stage-wise, admission to discharge time features will be utilized to indicate the length of stay. The later sections will discuss the dataset, pre-processing, and proposed model.

3.1.1 Data repository

In 1979, the government and the healthcare sector collaborated to create SPARCS, a Comprehensive all-payer data reporting system. The New York State Department of Health contributed to the Hospital Inpatient Discharges (SPARCS De-Identified) 2017 dataset; this is an open dataset with de-identified data; the file is not individually identifiable. It contains essential record-level detail. This repository includes 2.3 Million records with 34 features. They feature patient demographic data, diagnosis data, hospital-level data, and expenses. The patient's lab test-related parameters and private information are excluded and are not part of our prediction. The dataset contains several hospitals in the US based on different regions. The dataset is the repository of 2017 with other diseases found in the US. This dataset includes all patients visiting both emergency and general admissions. The summary of each feature of sparks data is described in Table 1[55].

Table 1: Description of the Sparcs data

Features	Description
Health Service Area	A description of the Health Service Area (HSA) in which the hospital is located. Blank for abortion records. Capital/Adirondack, Central NY, Finger Lakes, Hudson Valley, Long Island, New York City, Southern Tier, Western NY.
Hospital County	A description of the county in which the hospital is located. Blank for abortion records.
Operating Certificate Number	The NYS Department of Health assigns the facility's Operating Certificate Number. Blank for abortion records.
Facility Id	Facility Identifier. Blank for abortion records.
Facility Name	The name of the facility where services were performed was based on the Permanent Facility Identifier (PFI) maintained by the NYSDOH Division of Health Facility Planning. For abortion records, 'Abortion Record – Facility Name Redacted' appears.
Age Group	Age in years at the time of discharge. They are grouped into the following age groups: 0 to 17, 18 to 29, 30 to 49, 50 to 69, and 70 or Older.
Zip Code - 3 digits	The first three digits of the patient's zip code.
Gender	Patient gender: (M) Male, (F) Female, (U) Unknown.
Race	Patient race. Black/African American, Multi, Other Race, Unknown, White. Other Race includes Native Americans and Asian/Pacific Islanders.
Ethnicity	The ethnicity of the patient: Spanish/Hispanic Origin, Not of Spanish/Hispanic Origin, Multi, Unknown.

Length of Stay	The total number of patient days at an acute level and/or other than acute care level (excluding leave of absence days) (Discharge Date - Admission Date) + 1.
Type of Admission	A description of how the patient was admitted to the health care facility: Elective, Emergency, Newborn, Not Available, Trauma, Urgent.
Patient Disposition	The patient's destination or status upon discharge.
Discharge Year	The year of discharge.
CCS Diagnosis Code	AHRQ Clinical Classification Software (CCS) Diagnosis Category Code.
CCS Diagnosis Description	AHRQ Clinical Classification Software (CCS) Diagnosis Category Description
CCS Procedure Code	AHRQ Clinical Classification Software (CCS) ICD-9 Procedure Category Code
CCS Procedure Description	AHRQ Clinical Classification Software (CCS) ICD-9 Procedure Category Description
APR DRG Code	APR-DRG(All Patients Refined Diagnosis Related Groups) Classification Code.
APR DRG Description	APR-DRG Classification Code Description in Calendar Year 2017, Version 34 of the APR- DRG Grouper.
APR MDC Code	All Patient Refined Major Diagnostic Category (APR MDC) Code. APR-DRG Codes 001-006 and 950-956 may group to more than one MDC Code. All other APR DRGs are grouped into one MDC category.
APR MDC Description	All Patient Refined Major Diagnostic Category (APR MDC) Description.
APR Severity of Illness Code	APR-DRG Severity of Illness Code: 1, 2, 3, 4
APR Severity of Illness Description	All Patient Refined Severity of Illness (APR SOI) Description: Minor (1), Moderate (2), Major (3), Extreme (4)
APR Risk of Mortality	All Patient Refined Risk of Mortality (APR ROM) Description: Minor (1), Moderate (2), Major (3), Extreme (4)
APR Medical Surgical Description	APR-DRG specific classification of Medical, Surgical, or Not Applicable.
Payment Typology 1	A description of the type of payment for a hospitalization case.
Payment Typology 2	A description of the type of payment for a hospitalization case.
Payment Typology 3	A description of the type of payment for a hospitalization case.
Birth Weight	The neonate's birth weight is in grams, rounded to 100g.
Abortion Edit Indicator	A flag to indicate if the discharge record contains any indication of abortion ("N" = No; "Y" = Yes)
Emergency Department Indicator	Emergency Department Indicator is set based on the submitted revenue codes. If the record contains an Emergency Department revenue code of 045X, the indicator is set to "Y"; otherwise, it will be "N."
Total Charges	Total charges for the discharge.
Total Costs	Total estimated cost for the discharge.

The target feature for this prediction is Length of Stay. This study discretized the continuous Length of stay into seven groups. Few studies identified [19] Length of stay groups based on median, mean, or outlier values. They identified the groups based on the extreme cut-off points identified by outliers or median values. They have divided the groups based on the cut-off values; those higher than the cut-off values are treated as extended stays, and those below the cut-off are treated as short stays. This study does not rely on dividing groups based on the

frequencies of length of stay groups but rather on practical scenarios. We have also extracted the outlier values based on different methods, such as IQR and percentile (99 percentile), and using Z-Score methods, the identified outlier LOS are 12, 37, and 29. The median value of the LOS is 3.0, and the mean LOS is 5.42. This analysis gives us the length of stay distribution across the data. The potential length of stay groups we have identified is based on our business understanding and practical sense of hospital length of stay. The groups are meant to identify different short expected stays. This grouping might result in data imbalance and impact the classifier's performance for prolonged stays. The benefit of grouping is that essential groups of Los can be validated through this approach. The seven identified potential groups are 0-5 days, 6-10 days, 11-15 days, 16-30 days, 31-60 days, 61-90 days and > 90 days. The data distribution of the different LOS groups is shown in Table 2.

Table 2: LOS groups

Group (In days)	Patient count
0-5 days	1692034
6-10 days	361258
11-15 days	122196
16-30 days	97595
31-60 days	26343
61-90 days	5091
>90 days	3335

3.1.2 Data Analysis and Visualization

The Sparcs data contains patient demographics, hospital diagnosis level, and hospital expense-related data. The data is de-identified, so the patient's personal information and parameters that can identify the patient have been removed. This data consists of record-level basis details. The data does not contain inpatients' comorbidity or any basic vital signs. No lab parameters are also provided in the information. The dataset includes the patient's length of stay, which is the difference between admission and discharge day count. This study will incorporate the prediction of the length of stay. The analysis section will carry out the analysis of the length of stay against a few features. The detailed data analysis concerning its level has been represented in Table 3. The categorical column data distribution is stated in percentage, and the numerical column is expressed in its min, max, median, and mean values.

The study shows that the New York City hospital area has the highest admissions, at 45.78 %; the Manhattan hospital country has the highest admission rate, with 16.9 %.

Table 3: Data distribution of sparcs against its individual levels

Features	Category	Percentage

Health Service Area	New York City	45.78%
	Long Island	14.92%
	Hudson Valley	10.56%
	Capital/Adirond	7.28%
	Others	21.46%
Hospital County	Manhattan	16.92%
	Kings	9.99%
	Queens	8.6%
	Nassau	7.89%
	Others	56.6%
Facility Name	Mount Sinai Hospital	2.23%
	New York Presbyterian Hospital - Columbia Presbyterian Center	2.09%
	North Shore University Hospital	2.06%
	New York Presbyterian Hospital - New York Weill Cornell Center	1.97%
	Long Island Jewish Medical Center	1.87%
		89.78%
Age Group	70 or Older	28.37%
	50 to 69	28.01%
	30 to 49	19.21%
	0 to 17	14.44%
	18 to 29	9.98%
Zip Code - 3 digits	112	12.75%
	104	9.46%
	117	8.18%
	100	6.98%
	113	5.2%
	Other	57.43%
Gender	F	55.34%
	M	44.66%
Race	White	56.2%
	Other Race	25.21%
	Black/African American	17.56%
	Multi-racial	1.03%
Ethnicity	Not Span/Hispanic	80.89%
	Spanish/Hispanic	12.71%
	Unknown	6.02%
	Multi-ethnic	0.38%
Length of Stay	Mean	5.38
	Median	3.0
	Min	1
	Max	120
Type of Admission	Emergency	64.02%
	Elective	18.51%
	Newborn	9.37%
	Urgent	7.46%
	Trauma	0.56%
	Not Available	0.07%
Patient Disposition	Home or Self Care	65.77%
	Home w/ Home Health Services	14.45
	Skilled Nursing Home	9.51%
	Left Against Medical Advice	2.16%
	Expired	2.1%
	Short-term Hospital	1.73%
	Other	4.28%
CCS Diagnosis Description	Liveborn	9.38%
	Septicemia (except in labor)	5.0%
	Hypertension with complications and secondary hypertension	3.15%
	Osteoarthritis	3.02%
	Mood disorders	2.32%
	Others	77.13%

CCS Procedure Description	NO PROC	29.66%
	OT PRCS TO ASSIST DELIV	5.07%
	OTHER THERAPEUTIC PRCS	4.11%
	RESP INTUB/MECH VENTIL	3.16%
	CESAREAN SECTION	3.13%
	Other	54.87%
APR DRG Description	Neonate birthwt >2499g, normal newborn or neonate w other problem	8.12%
	Vaginal delivery	5.92%
	Septicemia & disseminated infections	4.16%
	Cesarean delivery	3.18%
	Heart failure	2.63%
	Other	75.99%
APR MDC Description	Diseases and Disorders of the Circulatory System	12.82%
	Pregnancy, Childbirth and the Puerperium	10.47%
	Newborns and Other Neonates with Conditions Originating in the Perinatal Period	9.59% 9.02%
	Diseases and Disorders of the Musculoskeletal System and Conn Tissue	8.17%
	Diseases and Disorders of the Respiratory System	8.1%
	Diseases and Disorders of the Digestive System	6.44%
	Diseases and Disorders of the Nervous System	5.95%
	Infectious and Parasitic Diseases, Systemic or Unspecified Sites	29.44%
APR Severity of Illness Description	Moderate	38.62%
	Minor	32.59%
	Major	22.89%
	Extreme	5.9%
APR Risk of Mortality	Minor	57.12%
	Moderate	21.68%
	Major	16.14%
	Extreme	5.06%
APR Medical Surgical Description	Medical	74.36%
	Surgical	25.64%
	Not Applicable	0.0%
Payment Typology 1	Medicare	38.7%
	Medicaid	30.05%
	Private Health Insurance	14.22%
	Blue Cross/Blue Shield	11.22%
	Self-Pay	1.91% 3.90%
Birth Weight Ind	Present	90.3%
	Absent	9.7%
Abortion Edit Indicator	N	99.86%
	Y	0.14%
Emergency Department Indicator	Y	58.42%
	N	41.38%
Total Charges	Mean	49723.80
	Median	26574.96
	Min	0.01
	Max	9696644.89
Total Costs	Mean	16647.34
	Median	8924.51
	Min	0.0
	Max	8868612.79
Expired Ind	No	97.9%
	Yes	2.1%

Mount Sinai Hospital has a higher number of admissions. At the same time, middle-aged and old age people have higher hospitalization rates compared to younger ones. For 70 or older, the hospitalization rate is 28.37%, and for 50 to 69, it is 28.01 %. More female patients have been admitted compared to males, with a percentage of 55.34%. 64.02 % of persons opted for an emergency type of admission. Home or self-care disposition is chosen by 65.77 % of people, and the death rate is 2.1%. Live-born, Septicaemia, hypertension-related disease, and osteoarthritis are the most frequent diagnoses present in sparse data. Around 29.66 % of admissions have no procedure performed; the second highest procedure is OT to assist delivery, which is 5.07%. Major diagnoses are Diseases and Disorders of the Circulatory System, Pregnancy, Childbirth, and the Puerperium, Newborns and Other Neonates with Conditions Originating in the Perinatal Period, Diseases and Disorders of the Musculoskeletal System and Conn Tissue, Diseases and Disorders of the Respiratory System. Extreme Severity of Illness has the lowest number of admissions, around 5.9%, and Moderate Severity of Illness has the highest admission percentage' which is 38.62%. The mortality risk in the dataset is less; for patients with minor risk, it is 57.12 %, and for extreme-risk patients, it is 5.06 %. Most of the admissions are medical, with 74.36% and 25.64% being surgical patients. 58.42% of patients are admitted to the emergency department, and 41.38% to general admissions. The population has 2.1% of deaths, and 97.9% survived from hospital.

Now, we will use visualization tools to visualize the data and try to get more information from it.
Figure 1 shows the distribution of Total charges and Total costs, which has skewed data distribution.

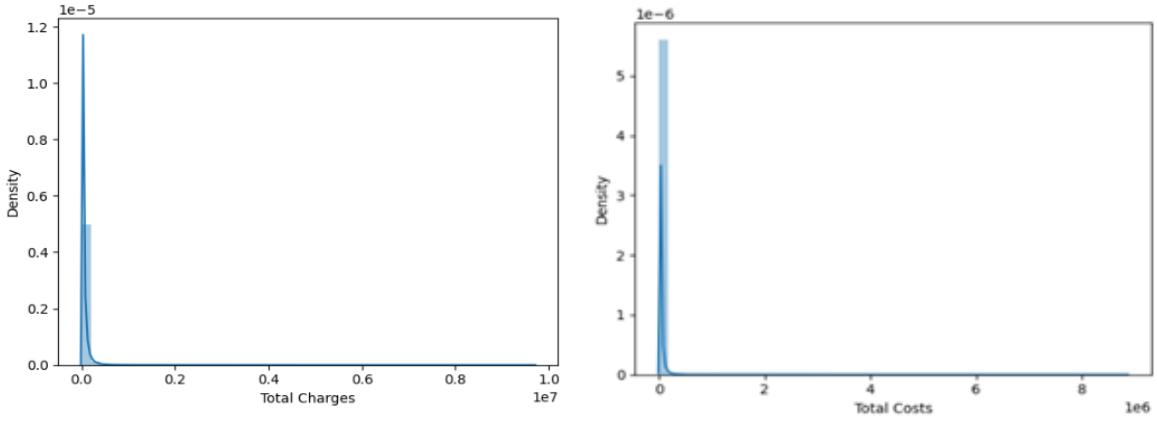


Figure 1:Total charges and costs distribution

Figure 2 shows the plot of the Length of stay against the Total cost; it shows that for short stays (within 20 days), few higher cost estimations are present. The cost is around 120 and 120+, and the estimated cost is high. Most of the estimated cost is within a limit irrespective of Los.

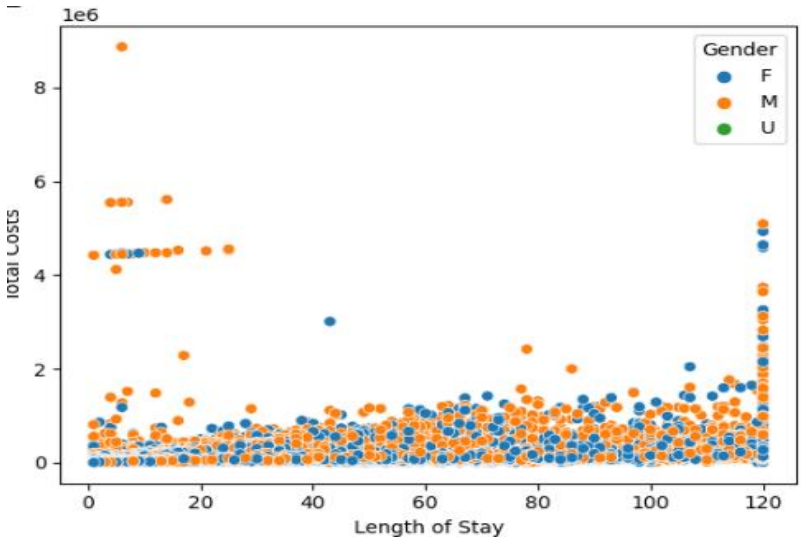


Figure 2:Plot of total cost versus length of stay

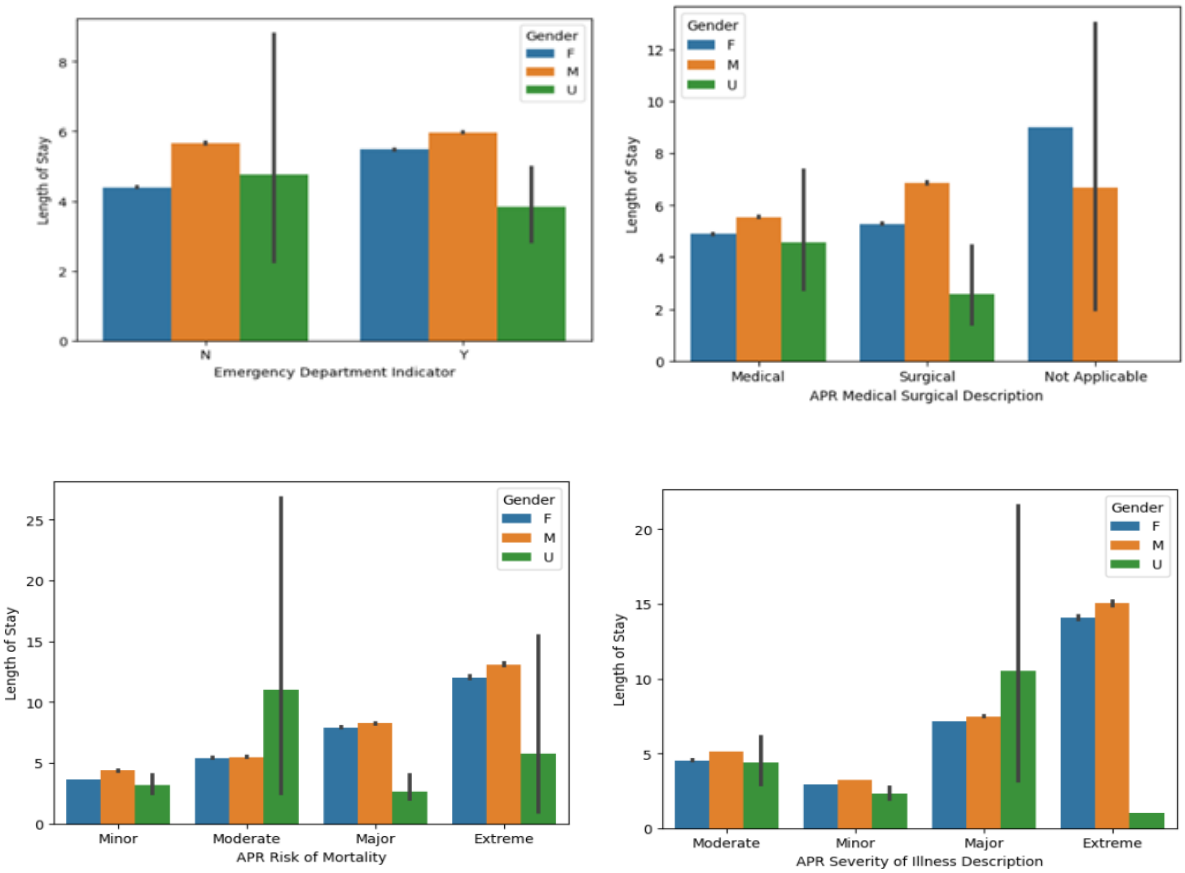


Figure 3:Length of stay versus plot of Emergency department, APR medical surgical description, APR risk of mortality, and APR severity of illness death

The Figure 3 bar plot shows that the length of the emergency and general admission stay have similar Length of stays. Plot b shows that the surgical patient's length of stay is slightly better than the medical patients. Plot c shows that the length of stay of patients with extreme risk is much higher than that of other mortalities. The length of stay generally decreases with a

decrease in the mortality risk. Minor ROM has a length of stay. The plot shows that the extreme severity of illness and patient's length of stay are higher than severities. The length of stay decreases with a decrease in the severity. Minor severity patients have a lower length of stay.

Figure 4 plot shows that the 70 or older age group has a higher stay length than other groups. The length of stay decreases with a decrease in age group. The 0 to 17 age group has less number of hospital stays.

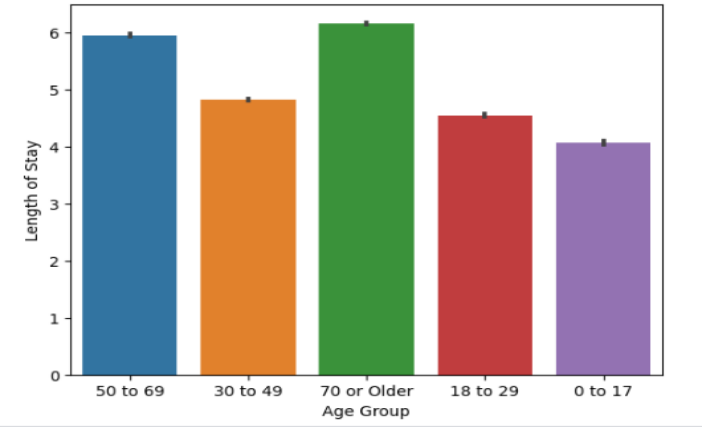


Figure 4:Plot of Age group vs. Length of stays

3.1.3 Data Pre-processing

The raw data we receive is usually incomplete and inconsistent. Those data cannot be directly fed to machine learning models, and it will produce unsatisfactory results; to improve the data quality, we need to clean the data and perform a few processing or transformations to improve the data quality. Data pre-processing generally improves the data quality for ML models. It also removed impure factors present in the data. The raw data is processed to clean data that can be interpretable by the machine. Our data has some inconsistencies: the length of stay has values '120+', which we have modified to make it 120. A few features have junk data. Gender has a value of 'unknown,' APR Medical Surgical Description' has a value of 'Not Applicable,' and birth weight is 'unknown.' We have removed these entries from the dataset. The remaining essential data pre-processing will be discussed in the later sections:

3.1.3.1 Missing Values handling

Missing data handling is the most vital part of the Data mining task. Some algorithms cannot handle missing data, so missing data handling is required. The missing value can degrade the performance of any machine-learning model. We have several columns in our data that have missing values. The detailed information is presented in Table 4:

Table 4: Missing data overview

Feature	Missing value count	Percentage
Hospital Service Area	5155	0.219964
Hospital County	5155	0.219964
Operating Certificate Number	5155	0.219964
Permanent Facility Id	5155	0.219964
Zip Code - 3 digits	39019	1.664939
APR Severity of Illness Description	240	0.010241
APR Risk of Mortality	240	0.010241
Payment Typology 2	878722	37.495034
Payment Typology 3	1737244	74.128135
Birth Weight	2115685	90.276198

Table 4 shows the features having missing value and their count and percentage. The missing values are not MCAR (missing completely at random) or MAR (Missing at random). We have applied a few missing data imputation techniques based on the missing patterns and the information provided in the SPARCS data dictionary [55].Please find the details of the techniques:

3.1.3.1.1 Populate with Default value

For Facility name = 'Redacted for Confidentiality,' some attribute values are intentionally removed for confidentiality. The reason is that those patients had abortions in the hospital since their details have been hidden. We have populated the '999' value in numeric features and 'Not Provided' in the categorical features. Features that are impacted by this are Hospital Service Area, Hospital County, Operating Certificate Number, Permanent Facility ID, and Zip Code - 3 digits.

3.1.3.1.2 Removal of features

A few columns had very high missing percentages. Payment Typology 2 and Payment Typology 3 have 37 % and 74% missing percentages. Filling these higher percentage missing values will impact the model inversely, so we removed these two features from the dataset. Birth weight has missing values more than 90%. Birth weight is an essential feature for newborns. We have kept birth weight, which will significantly impact the analysis of newborns' length of stay and death.

3.1.3.1.3 Removal of missing rows

APR Severity of Illness Description and APR Risk of Mortality have around 240 missing data. As per the SPARCS data dictionary[55], possible APR Severity of Illness Codes are '1','2','3', and '4', which is in number representation of the field APR Severity of Illness Description. APR Risk of Mortality and APR Severity of Illness Description have valid values such as 'Minor,'' Moderate", "Major", "Extreme". For the missing APR Severity of Illness Description and APR Risk of Mortality, the APR Severity of Illness Code is populated as '0', which is insignificant as per SPARCS Metadata as APR Severity of Illness Description and APR Risk of Mortality features cannot be blank so we have removed the 240 rows from the dataset. These missing rows belong to the live-born or pregnancy-related hospital cases.

3.1.3.1.4 Populate missing ZIP code

Zip code has a total of 1.66% of missing rows. The patient's zip code is limited to the first three digits. However, in cases where the population size for the zip code is less than 20,000, if the cell size on patient classification strata is less than 10, or if the record is an abortion, then the zip code is blank. For the redacted facility/abortion cases of the 5155 entries, we have already discussed that it has been updated with the default value as '999'. For the rest of the 33864 entries, we have implemented a Zip code population logic using the features related to Zip code. The features related to the zip code are the hospital service area, hospital county, and facility name. At first, we extracted the unique Hospital Service Area, Hospital County, and Facility Name where the Zip code is missing. Based on the combinations extracted from the Hospital Service Area, Hospital County, and Facility Name, we searched for the available valid zip codes in the dataset. From these available valid Zip Codes, we have randomly selected the Zip codes for our missing value imputation.

3.1.3.2 Feature Engineering

Feature engineering is one of the critical tasks in machine learning modeling. New features extracted from the existing features can increase the predictive power of machine learning models. We have created a few new features from the existing 34 features. Our feature set has a CCS (Clinical Classification code) diagnosis code that gives information about the patient's primary diagnosis. The diagnosis codes are ICD-9 standard codes. Based on this CCS diagnosis code [23], we have generated Comorbidity [23] related features utilizing the Charlson comorbidity index. It will generate 18 more features with different comorbidities and the respective comorbidity score. Patient disposition is the final destination of patients after release from the hospital; we have created one feature, Expired Ind, utilizing patient disposition 'Expired.' It informs that the patient has died in the hospital. The birth weight feature is explicitly for the newborn.

We have created a new feature, birth weight indicator, that utilizes birth weight. This field holds binary values '1' or '0' based on the availability of birth weight. The actual Birth weight field is no longer required, so it has been dropped. We have also generated a few features based on the estimated Total Cost for the features CCS procedure code, Emergency Indicator, and Payment Typology 1. We have calculated the Min, Max, Mean, Median and Standard Deviation for the CCS procedure code, Emergency Indicator, and Payment Typology 1 feature based on the estimated total Costs. The feature relation analysis section will evaluate the importance of the newly generated features. These freshly generated 35 features, along with the remaining features, will be used for prediction.

3.1.3.3 Feature Selection

Feature selection is one of the essential steps in implementing the predictive model. Proper features can increase the model's accuracy, and irrelevant features can significantly degrade the model's performance. In our dataset, five features duplicate features are present . Please refer to Table 5 for the details:

Table 5: Duplicate features

Feature Name	Similar Feature(Will be deleted)
CCS Diagnosis Code	CCS Diagnosis Description
CCS Procedure Code	CCS Procedure Description
APR DRG Code	APR DRG Description
APR MDC Code	APR MDC Description
APR Severity of Illness Code	APR Severity of Illness Description

We will delete one of the features from Table 5. Apart from these five features, we have one feature with a constant value, i.e., discharge year so that we can discard the field directly. We will verify the model's performance with a reduced feature set by selecting more valuable features using techniques such as Mutual Information and the Chi-Square test.

3.1.3.4 Normalization

Normalization is an essential preprocessing step for data mining. Certain distance or weight-based models cannot produce good results without proper scaling. We will apply Min-max normalization to the independent features to normalize them.

3.1.3.5 Feature Transformation

Total estimated cost, total charges, and Length of stay features data distribution is highly skewed. To make them balanced and normally distributed, we used log transformation. Total costs and charges are undergoing this transformation. Length of stay is excluded from transformation as it is a target variable. Transformation helps to build more robust models.

3.1.3.6 Feature Encoding

Feature encoding is a necessary step before data mining. Most models cannot understand non-numeric data, so data needs to be transformed into numeric. Our data has only two numeric independent features. Both ordinal and nominal features are present for categorical features. We have applied a mixed encoding technique employing three different encoding schemes. We have many unique categories for a few categorical features; one hot encoding cannot be used for those features, and it will unnecessarily increase the feature dimension, resulting in a highly complex model and performance degradation. We have opted for Label encoding for many category features (e.g., Procedure code has 200+ unique categories). Please refer to Table 6 for the encoding details.

Table 6: Encoding mapping details

Encoding method	Features used
One hot Encoding	Health Service Area, Race, Ethnicity, Type of Admission, Payment Typology 1, APR Medical Surgical Description, Expired Ind
Ordinal Encoding	APR Risk of Mortality, Age Group, Emergency Department indicator.
label Encoding	Hospital County, Facility Name, Zip Code - 3 digits, Gender, Patient Disposition, Abortion Edit Indicator, Operating Certificate Number, Facility ID, CCS Diagnosis Code, CCS Procedure Code, APR DRG Code, APR MDC Code.

3.1.4 Statistical relationship between features

We have 2343569 instances with 34 features in our dataset. We have performed feature engineering to include more features to give more predictive power to the model. We have included 18 comorbidity-related features as well as 15 features that are related to the total cost. After pre-processing, feature selection, and engineering, 57 features are left with us. We have performed statistical Pearson correlation coefficient tests between target and independent features, and the detailed report is published in Table 7. We have observed that Total cost and Total charge have very high positive correlations of 0.6161 and 0.5767, respectively, with Length of stay. APR Severity of Illness Code, APR Risk of Mortality, and Patient Disposition also have high positive correlations with LOS with coefficients of 0.3872, 0.3392, and 0.2094, respectively. Age Group, Expired_Ind, Emergency Department Indicator, APR MDC Code, and Charlson Index Comorbidity Score also have a significant positive correlation with the LOS. We can see that Birth Weight Ind has a significant negative correlation with the target feature, identified as one of the valuable predictors of LOS. The new cost-related features have a substantial positive and negative correlation with the target feature. Few Comorbidity factors have less impact on the prediction; those are Rheumatologic Disease and Moderate or Severe Liver Disease with p-value < 0.05.

The dataset contains many categorical fields and few numerical fields. The numerical fields are highly skewed. We have also performed a Chi-Square statistical test on the categorical features and a Kendall tau statistical test on the numerical features concerning the detailed analysis of the target feature length of stay, as shown in Table 7. The length of stay is converted into a categorical variable with different levels. The Length of stay's distribution is non-normal. The chi-square test has identified the CCS Procedure Code, CCS Diagnosis Code, APR Risk of Mortality, APR DRG Code, Patient Disposition, APR Severity of Illness Code, Age Group, APR MDC Code, Facility, Birth_Weight_Ind, Charlson Index Comorbidity Score the top features related to the LOS. Kendall Tau shows that total cost and total charge have high significance with length of stay with a correlation coefficient of 0.4842 and total charge of 0.4434. The features related to total cost also have a reasonable correlation, as shown in the table.

Table 7: Statistical analysis of the features with the target

Feature	Pearson Correlation	p-value	Chi-statistics/ Kendal tau	p-value
Operating Certificate Number	0.028272587	0.000000	62123.87178	0.000000
Age Group	0.148518559	0.000000	69652.43807	0.000000
Gender	0.069332564	0.000000	6490.021068	0.000000
Race	-0.036714269	0.000000	2056.058393	0.000000
Patient Disposition	0.209421147	0.000000	165903.1691	0.000000
CCS Diagnosis Code	-0.038629603	0.000000	434217.5321	0.000000
CCS Procedure Code	0.054908829	0.000000	621059.2853	0.000000
APR DRG Code	0.044720187	0.000000	322798.8355	0.000000
APR MDC Code	0.079949673	0.000000	66728.7235	0.000000
APR Severity of Illness Code	0.387247695	0.000000	146951.596	0.000000
APR Risk of Mortality	0.339236665	0.000000	368508.8151	0.000000
APR Medical Surgical Description	0.035295833	0.000000	3273.907162	0.000000
Payment Typology 1	0.040121554	0.000000	4836.371147	0.000000

Emergency Department Indicator	0.087134496	0.000000	17372.59703	0.000000
Birth_Weight_Ind	-0.092799894	0.000000	36211.92656	0.000000
Expired_Ind	0.101501961	0.000000	24167.60626	0.000000
Total Costs_procdre_min	0.095623896	0.000000	0.056993027	0.000000
Total Costs_procdre_avg	0.251958005	0.000000	0.230603946	0.000000
Total Costs_procdre_std	0.193836557	0.000000	0.252045911	0.000000
Total Costs_procdre_med	0.256985534	0.000000	0.185300962	0.000000
Total Costs_Emer_ind_min	0.087134496	0.000000	0.11654761	0.000000
Total Costs_Emer_ind_max	-0.087134496	0.000000	-0.11654761	0.000000
Total Costs_Emer_ind_avg	-0.087134496	0.000000	-0.11654761	0.000000
Total Costs_Emer_ind_std	-0.087134496	0.000000	-0.11654761	0.000000
Total Costs_Emer_ind_med	0.087134496	0.000000	0.11654761	0.000000
Total Costs_pay_typ_max	-0.037436988	0.000000	-0.020211673	1.5E-258
Total Costs_pay_typ_avg	0.125047249	0.000000	0.123978331	0.000000
Total Costs_pay_typ_std	0.075441732	0.000000	0.121966883	0.000000
Total Costs_pay_typ_med	0.127150528	0.000000	0.123504719	0.000000
Total Charges Mod	0.576748042	0.000000	0.443459699	0.000000
Total Costs Mod	0.616172266	0.000000	0.484243024	0.000000
Charlson Index Co-Morbidity Score	0.069836535	0.000000	32089.51034	0.000000
Cerebrovascular Disease	0.045864754	0.000000	5731.798245	0.000000
Dementia	0.037948153	0.000000	3423.720014	0.000000
DHemiplegia or Paraplegia	0.025442307	0.000000	1799.712991	0.000000
Any Malignancy, Including Leukemia and Lymphoma	0.059219741	0.000000	8094.390818	0.000000
Metastatic Solid Tumor	0.03631234	0.000000	4154.705346	0.000000
Zip Code - 3 digits	-0.024583655	1.99E-305	20521.81615	0.000000
AIDS	0.021304985	6.55E-230	1067.118689	1.01E-229
Permanent Facility Id	0.019308616	3.30E-189	51194.47537	0.000000
Abortion Edit Indicator	-0.017376171	1.29E-153	897.6658788	5.34E-190.000000
Ethnicity	-0.016985085	7.22E-147	531.6392619	9.6E-114
Type of Admission	0.016849018	1.48E-144	1414.984181	3.89E-305
Mild Liver Disease	0.015045314	1.16E-115	691.485981	2.43E-148
Diabetes with Chronic Complications	0.014612108	3.29E-109	1476.453779	0.000000
Congestive Heart Failure	0.013189134	2.49E-89	815.9905471	2.64E-175
Myocardial Infarction	-0.013000763	7.50E-87	616.1186648	5.03E-132
Total Costs_procdre_max	0.011240616	2.13E-65	0.075722236	0.000000
Hospital County	0.010070643	7.52E-53	7352.855933	0.000000
Peripheral Vascular Disease	0.008976186	2.38E-42	368.6360756	1.658E-78
Total Costs_pay_typ_min	-0.008652028	1.80E-39	-0.084138491	0.000000
Renal Disease	0.005942956	1.73E-19	603.7694642	2.37E-129
Chronic Pulmonary Disease	-0.004210883	1.58E-10	1297.302501	1.28E-279
Peptic Ulcer Disease	0.004035517	8.71E-10	120.8865469	3.454E-25
Diabetes	-0.003521265	8.79E-08	48.19419097	8.598E-10
Facility Name	-0.003212153	1.06E-06	9498.985962	0.000000
Hospital Service Area	0.002454857	0.0001917	714.780695	2.2E-153
Moderate or Severe Liver Disease	-0.0008112	0.2177673	320.545518	3.999E-68

Rheumatologic Disease	3.72E-05	0.9549814	35.03284211	4.574E-07

The Chi-square statistics about the comorbidity factor Charlson comorbidity factors show that Any Malignancy, Including Leukemia and Lymphoma, Cerebrovascular Disease, Metastatic Solid Tumor, Dementia, DHemiplegia or Paraplegia and Diabetes with Chronic Complications having good significance with the target feature. In contrast, Peptic Ulcer Disease, Diabetes, and Rheumatologic Disease comorbidities show less or no importance with the target variable. The comorbidity factor is an essential predictor of hospitalization case analysis. Our study used the factor for length of stay prediction and other factors in the dataset.

The essential identifiers of the length of stay were discussed using three statistical tests: Pearson correlation, chi-square statistics, and Kendall tau. Using different statistical tests, different factors impacting the length of stay have been identified, and we will verify the actual predictors of the length of stay; when we apply various machine learning models to make the prediction, we will evaluate the essential factor that comes with each machine learning model.

We will not remove any of the features as part of general processing. We will perform a feature selection based on Mutual information and a Chi-Square test to cross-verify the model's performance using a reduced feature set with 55 features.

3.1.5 Proposed Model

Several lengths of stay studies have been proposed for both single and multi-stage, but their work is limited to a specific category, disease, or hospital. The data is usually less diverse, or the sample size needs to be more significant. Most of the studies were concretized on ICU cases. Their study is also limited due to the presence of small data samples. Our referred data has enough samples and knowledge about the length of stay can be gained from that data. This is not limited to only ICU cases. It considers both ICU and general admission data data. This study introduces a new format and way to predict the length of stay. This study presents a three-stage model: admission, post-admission, and final discharge. This three-phase predictive model will give hospital authority and patient parties an incremental estimated hospital stayway. The model will have three predictions: first at the time of admission, next at a post-admission phase, generally after 1-2 days (48 hr.) of hospitalization, and the final at the time of discharge. These multi-stage predictions will be based on the features collected at each stage.

The features will be incremented from the admission phase to the discharge phase, impacting the Los prediction proportionally. This phase-wise prediction depends on the predictive power of the features added at each stage. Different feature sets for the phase-wise prediction model have been identified. Features are determined based on references from the SPARCS data dictionary [55], business understanding, and a few studies [32]. Please refer to Table 8 for the stages and their corresponding features. The patient's demographic and preliminary diagnosis details can be extracted at admission. After admission, some diagnoses and procedures will be performed. Many features related to the diagnosis and condition of the patient will be added, and the same will be utilized for post-admission stage prediction. The final or discharge stage consists of only features available at discharge. The feature's Total cost and Total charges are significant predictors of Length of stay. Total cost is the estimated cost the hospital authority or physician provides after admission. The total cost will be available after admission, so we have included the total cost in the post-admission stage prediction model. The total charge is the actual cost of the hospitalization, which can only be available at the time of discharge.

Table 8: Three-stage model feature mapping

Stage	Features Used for Prediction
Admission	Health Service Area , Hospital County, Operating Certificate Number, Permanent Facility ID, Facility Name, Age Group, Zip Code - 3 digits, Gender, Race, Ethnicity, Ethnicity, CCS Diagnosis Code, Payment Typology 1, Emergency Department Indicator.
Post Admission	CCS Procedure Code, APR DRG Code, APR MDC Code, APR Severity of Illness Code, APR Risk of Mortality, APR Medical Surgical Description, Birth Weight, Abortion Edit Indicator, Total Costs.
Discharge Time	Patient Disposition, Total Charges, Expired Ind
Target	Length of Stay

Discharge disposition is the final destination of a patient after discharge. We have included this feature only in the final discharge stage. The admission stage model prediction will be performed based on the features mentioned in the table, and the subsequent stages will use the features mentioned in the table along with the features of the prior stages, not vice versa. The final discharge stage will use all the features mentioned in the table for prediction.

3.1.5.1 Stage-wise feature selection details

Table 8 shows the different features selected for these three-stage models. The admission stage is when patients opt for hospital admission based on some outpatient or emergency visits. The included features in this stage are related to the patient's demographic hospital details. The patients' primary diagnoses are also included in this phase. The payment typology was also included in the admission phase because the mode of payment opted for is informed at the time of hospitalization. The emergency department is also kept in the admission phase, as general or emergency admissions are known at the time of admissions.

The post-admission phase referred to here is post-hospitalization admission, after 1-2 days of hospitalization. After admissions, the required tests and specialized doctor consultation happened, and based on the doctors' evaluation, a few features were provided that were considered in this phase. This phase added secondary diagnosis and procedures performed after hospitalization. The essential features considered for this phase are significant diagnosis details, procedures performed, diagnosis groups, childbirth weight, abortion, mortality risk, severity of illness, and total costs. The mortality risk and patient seriousness can only be identified when a primary diagnosis[32] has been performed, and patients have stayed in hospitals for more than 24 hr. Child weight and abortion details can be opted for when the surgery has been performed. The surgery information can also be included in this phase as, based on the doctor's diagnosis, the need for surgery is decided that happens post-hospitalization.

The discharge stage is the final stage when the patient will be released from the hospital. The discharge can be of two types: whether the patient has survived or died. The features can be added in this phase based on the discharge-related details, such as the final settled bill patient destination after discharge. The added features of this phase are total charge, discharge destination, and expired Ind. The stage-wise prediction model is shown in Figure 5.

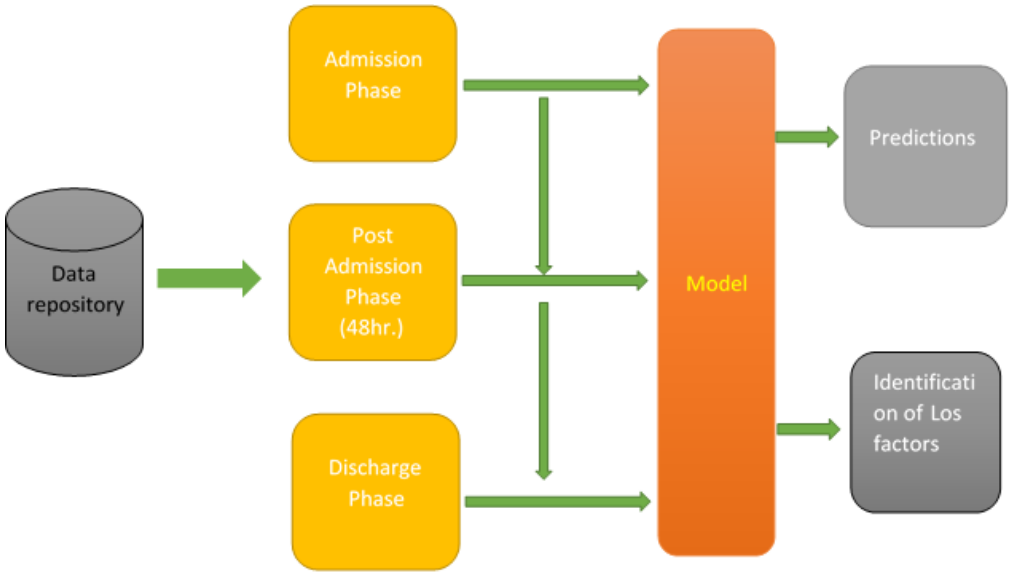


Figure 5:Stages-wise length of stay predictive model

3.1.5.2 Machine learning models

In the proceeding section, we will discuss the different machine learning models utilized in our prediction model and analysis; finally, our proposed new hybrid model will be presented.

3.1.5.2.1 Logistic regression

The logistic model, often known as the logit model, is a statistical model that represents an event's log odds as a linear combination of one or more independent variables. This simple model for classification tasks mainly deals with linear data.

Logistic regression uses Sigmoid Function for Binary Classification and Softmax for multiclass classification. We have multiclass logistic regression and will use default Softmax regression for our prediction. We will use the default setting for logistic regression models, such as penalty as ‘L2’ and iteration as 100.

3.1.5.2.2 Random Forest Classifier

Random Forest is a famous supervised machine learning algorithm. Its idea is ensemble learning, merging several classifiers to solve a challenging issue and enhance the model's functionality. According to its name, "Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Rather than depending on a single decision tree, the random forest forecasts the outcome based on the majority vote of projections from each tree. Because more trees are in the forest, accuracy is higher, and overfitting is avoided. The random forest works on the bootstrap aggregation concept, and it randomly considers a portion of the data to train the individual models to obtain improved accuracy. For our prediction, the parameters used are an estimator of 150 and a max tree depth of 25, irrespective of the stages.

3.1.5.2.3 Adaboost

Combining several weak classifiers, the ensemble modeling technique known as boosting aims to create a robust classifier. It is accomplished by using weak models in series to develop a model. First, a model is made using the training set of data. The second model is then created to fix the previous model's flaws. Models are added this way until the entire training data set is adequately predicted or the maximum number of models have been added. AdaBoost (Adaptive Boosting) is a widespread technique combining multiple weak classifiers to build one robust classifier. We used a number of estimators = 500 for Adaptive boosting ensemble learning irrespective of the prediction stage.

3.1.5.2.4 XGBoost

A distributed gradient boosting library optimized for efficiency and scalability in machine learning model training is called XGBoost. It is an ensemble learning technique that generates a more robust prediction by aggregating the predictions of several weak models. Extreme Gradient Boosting, or XGBoost, is a machine learning algorithm that has gained popularity and widespread usage because it can handle large datasets and achieve state-of-the-art performance in many machine learning tasks, including regression and classification. XGBoost's effective handling of missing values is one of its primary characteristics, enabling it to handle real-world data with missing values without requiring much pre-processing. Furthermore, XGBoost comes with built-in support for parallel processing, which allows training models on big datasets quickly. In this study, the parameters used are an estimator of 1000 and a max tree depth of 8.

3.1.5.2.5 LightGBM

Microsoft created LightGBM, an open-source, distributed, high-performance gradient-boosting framework. Its design prioritizes accuracy, scalability, and efficiency. It is predicated on decision trees that decrease memory utilization and increase model efficiency. It combines several cutting-edge methods, such as Gradient-based One-Side Sampling (GOSS), which maximizes training time and memory consumption by retaining examples with large gradients only during training. Furthermore, LightGBM uses methods based on histograms to generate trees more effectively. These methods increase LightGBM's efficiency and offer an advantage over competing gradient-boosting frameworks. Competing optimizations include leaf-wise tree development and practical data storage formats. We used the Number of estimators = 500 and other configurations as a default across all the stages.

3.1.5.2.6 KNN

KNN is one of the simple models. Because of its simplicity and ease of use, the KNN method is a popular and adaptable machine-learning technique. No presumptions on the distribution of the underlying data are necessary. It is a versatile option for various classification and regression application datasets because it can handle numerical and categorical data. This non-parametric technique bases its predictions on how similar the data points in a particular dataset are. By comparison, KNN is less susceptible to outliers than other algorithms. The KNN algorithm locates the K closest neighbors to a given data point using a metric like Euclidean

distance. The majority vote or the average of the K neighbors is then used to establish the class or value of the data item. Using this method enables the algorithm to anticipate outcomes based on the local structure of the data and adjust to various patterns. We have taken several neighbors =10 across all the stages.

3.1.5.2.7 Artificial Neural Network

Artificial neural networks are based on the architecture and functions of human neurons. Neural nets or neural networks are other names for it. An artificial neural network's first layer, the input layer, transfers data from outside sources to the second layer, the hidden layer. Each neuron in the hidden layer takes in information from the neurons above, calculates the weighted sum, and then relays it to the neurons in the layer below. Because these connections are weighted, the effects of the inputs from the preceding layer are optimized by giving each input a unique weight, which is then modified during training to improve model performance. The Ann structure we used has one input layer, four hidden layers, and the final output layer across all the stages.

3.1.5.2.8 CatBoost

It is an open-source boosting library known as CatBoost, sometimes called Categorical Boosting. It is intended for use with many independent features in regression and classification tasks. A gradient-boosting variation that can handle numerical and categorical features is called a CatBoost. No feature encoding technology such as One-Hot Encoder or Label Encoder is needed to translate category information into numerical features. To lessen overfitting and enhance the dataset's overall performance, an approach known as the symmetric weighted quantile sketch (SWQS) is used, automatically managing the missing values in the dataset. We have used Max Depth =12 and Iterations = 3000 for our prediction across all the stages.

3.1.5.2.9 Ensemble Learning

Through ensemble learning, several machine-learning models are trained to produce improvised predictions that improve the performance of individual machine-learning models. The predictors that have been trained to compute the predictions are essentially referred to as the ensemble. Most of the time, decision trees help solve quantitative problems and facilitate ensemble learning. Ensemble learning computes the final classification based on the ensemble findings generated from decision trees instead of relying solely on a single decision tree's predictive analysis and outcome, i.e., a homogeneous model. Using various machine learning models, including Random Forest Classifier, XGBoost, and CatBoost, among others, one such ensemble learning implementation trains and then displays the combined output of these models to make final judgments. Two different voting classifiers, soft voting, and complex voting classifier, are used to generate the findings. Hard voting is majority voting based on selected classifiers, and soft voting is the maximum average probability of the selected classifiers. The ensemble method for hybrid classifiers not only improves the prediction accuracy but also takes the good from all the classifiers and produces a robust overall improved classifier.
Several soft Ensemble approaches are available: Average probability, Max Rule, Sum rule, and product rule.

3.1.5.2.10 Our Architecture

In our study, we are introducing a novel weighted average probability ensemble-based Los model. This model will take different well-tuned model's predicted probabilities and averages the probabilities based on the weight assigned to each classifier. The classifier's weights will be selected based on the contribution of each model toward generating an optimal ensemble model through a robust grid search method. We are proposing a 3WP ensemble model. The 3WP ensemble model will be selected based on the top three performing base models. The algorithm for the prediction is discussed below:

3.1.5.2.10.1 Algorithm

Input: *Training dataset and testing dataset.*

Procedure: *The steps below will be repeated in three stages: Admission, Post-admission, and discharge.*

1. Train different base models.
2. Test the base models and compare the accuracy, precision, recall, and f1 score.
3. Calibrate the individual base models to obtain optimum precision.
4. Select top models based on descending order of classification metrics considering accuracy, precision, recall, and f1 score.
5. Random Forest, XGBoost, ANN, CatBoost, and LightGBM are the top-performing models selected.
6. Select the top three models based on balanced classification metrics.
7. Obtain an initial weighted average ensemble model using equal weights.
8. Calculate the weights for each model using the grid search method and assign optimum weights to each classifier.
9. Apply weighted probability average ensemble for prediction using optimized weights.
10. Generate probability-weighted voting with a new set of classes.
11. Compare and evaluate the precision of the new ensemble model with the base models.

The proposed model's architecture is shown in Figure 6.

3.1.5.2.10.2 Equation of the weighted ensemble model

The weighted probability ensemble model is created to choose top classifiers and assign weights to each classifier. The final ensemble model will be obtained by averaging the probability of the classifiers for the weights of each model. The equation for the weighted probability model is defined below:

$$\text{Weighted probability} = \frac{\sum_{i=0}^{n}(Pi*Wi)}{\sum_{i=0}^{n} Wi}$$

N= number of classifiers
Pi= Probability of the classifier
Wi = assigned weight of a classifier

Let's assume three classifiers are chosen for ensemble: classifier 1, classifier 2, and classifier 3. The selected weights are w1, w2, and w3 accordingly. The final weighted probability of the model will be calculated as:
Weighted probability = (Classifer1 probability * w1 + Classifer2 probability * w2 + Classifer1 probability * w3) / (w1+w2+w3).

The final result of the classifier is obtained by fetching the maximum probability of an instance from the weighted probabilities; the location of the maximum probability will be selected as the resultant class.

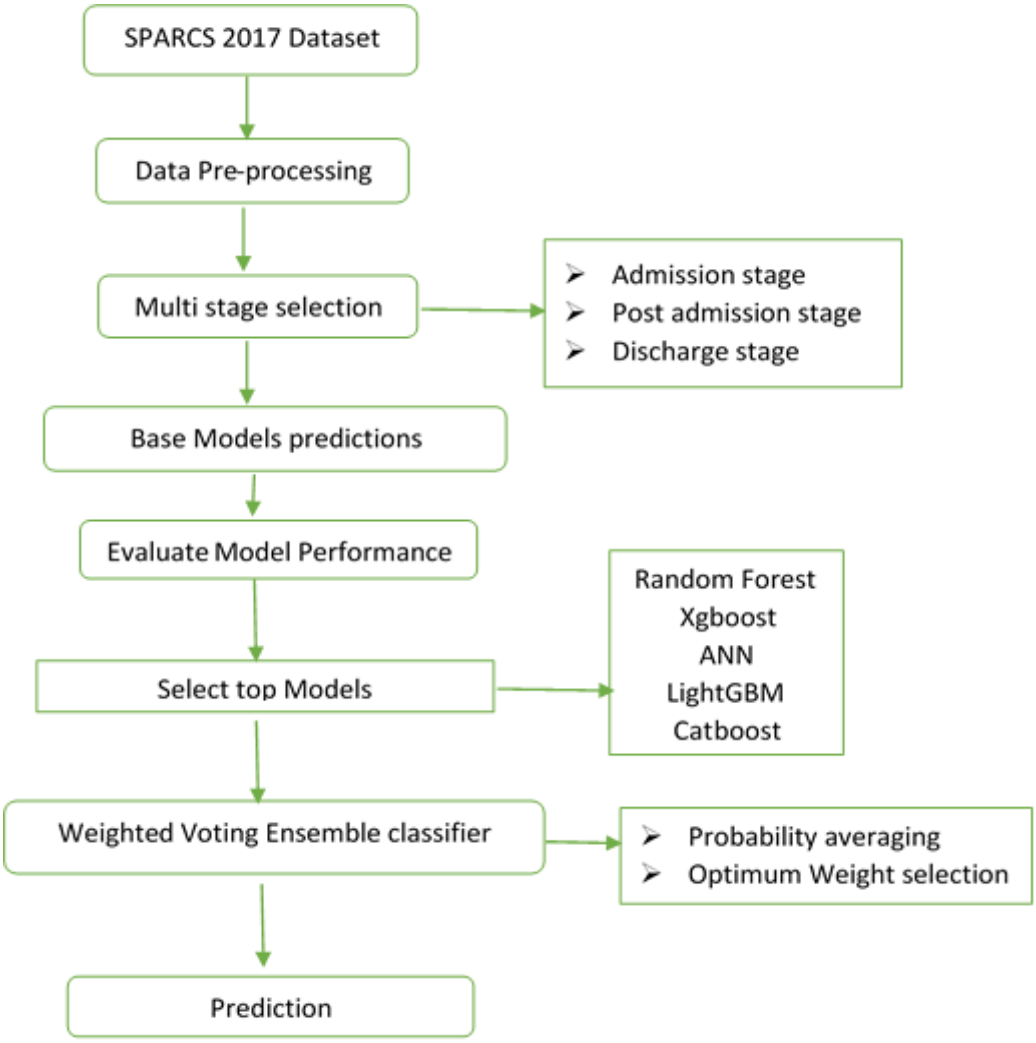


Figure 6: Schematic diagram of the Los prediction model

3.1.5.2.10.3 Pseudo code of the algorithm

1. Initialize a list of `models` containing three different models: 'Randmreg', 'xgb_r', and 'Ann_sklrn.'
2. Predict probabilities for each model on the scaled test data `X_test1_scaled` and store them in a list of `preds`.
3. Convert the list of predictions `preds` into a NumPy array `array_opt.`
4. Calculate the soft ensemble of predictions by averaging the predictions in `array_opt` along the first axis, with equal weights [0.1, 0.1, and 0.1]. Store the result in `soft_ensemble`.
5. Determine the class predictions of the soft ensemble by selecting the maximum value index along the second axis. Store the result in `soft_ensemble_pred`.
6. Calculate the classification report, accuracy, and confusion matrix for the soft ensemble predictions.
7. Initialize an empty DataFrame `df_new`.
8. For each combination of weights (w1, w2, w3) in the ranges [0.1, 0.2... 0.8]:
a. Calculate a weighted ensemble of predictions by averaging the predictions in `array_opt` with the current weights (w1, w2, w3).

b. Determine the class predictions of the weighted ensemble.
c. Calculate the weighted ensemble predictions' accuracy, precision, recall, and F1-score.
d. Append the calculated metrics and the weights to the DataFrame `df_new.`
9. Save `df_new` to a CSV file named 'votingwegtd.csv'.
10. Sort `df_new` by accuracy in descending order.
11. Remove the last `n`(412) rows from `df_new`.
12. Identify the row in `df_new` with the highest precision and store it in `max_acc_row.`
13. Print the weights that achieved the maximum precision.
14. Calculate the soft ensemble using the optimal weights from `max_acc_row.`
15. Determine the class predictions of the soft ensemble.
16. Calculate classification report, accuracy, confusion matrix, precision, recall, F1-score, and AUC-ROC for the soft ensemble predictions.

3.1.5.2.10.4 Hyperparameter Tuning of base models

The weighted ensemble model generated using the base models needs the base models to be highly précised and calibrated. Well-tuned base models give more power to the ensemble model. If the base model's performance is not nearly equal or similar, the ensemble model might not perform well. We have identified three top-performing models based on different stages. Those top models have been tuned to get optimal results for the dataset. The calibrated models are XGBoost, ANN, random forest, LightGBM, and CatBoost. The final tuned parameters are given in the machine learning model section, where we describe the models. The tuned base models help the ensemble model to produce optimal results.

3.1.6 Evaluation criteria

For classification problems, classic evaluation metrics are accuracy, precision, recall, F1-score, and AUC-ROC value. We will be using these metrics to validate our multi-stage model. These metrics define different aspects of the model's performance. We aim to produce maximum efficiency from these metrics by combining the top-performing base models.

3.2 Result

Initially, the USA SPARCS dataset had 2343569 records with 34 features. After data pre-processing, duplicate removal, feature selection, and feature engineering, we have 2308391 instances left in our dataset. The feature dimension increased from 34 to 82. We have used the online data science and ML framework Kaggle for our coding and analysis. Kaggle is an open-source online machine learning and data science research platform. The platform uses Python 3.10 with the latest Scikit Learn, TensorFlow, and Keras ML packages. The models are trained using 80% of the data, and the remaining 20% is used for testing. The comparative study of different stages of the proposed model is illustrated in Table 9. The performance of base models and the weighted average model are also presented in the table. The final optimum weights of the 3WPE are presented in the table. The optimum weight search space is used between 0.1 to 0.8. The number of features utilized for different stages are 55, 79, and 82 for admission, post-admission, and discharge stages, respectively. Two different evaluations are performed: one based on a complete feature set and another with a reduced feature. The reduced feature set evaluation is only applicable during the post-admission phase.

3.2.1 Admission phase model

The accuracy of admission stage models lies between 0.64-0.68. Most models' performance is similar according to accuracy but varies in precision, recall, f1 score, and Auc scores. LightGBM, CatBoost, and XGBoost are performing well not only with respect to accuracy but also with respect to the other parameters, which are performing well compared to other models. For AUC, LightGBM tops the table with a value of 0.7911, and XGBoost is in second position at 0.7853. CatBoost is the winner for precision with a value of 0.34. XGBoost has the highest recall and f1-score, 0.19 and 0.2, respectively. For accuracy, CatBoost is the winner with a value of 0.6821. The performance of random forests and ANN is insignificant at this stage. The 3WPE ensemble classifier combines CatBoost, XGBoost, and LightGBM in the admission phase. The ensemble accuracy is 0.6824, and precision, recall, f1 score, and auc are 0.34, 0.18, 0.18, and 0.8042, respectively, performing decently well with top base models. The weights used for 3WPE are 0.6, 0.1, and 0.1. The class1 (0-5 days) Los group has excellent performance; the recall is 0.98 and precision is 0.71, but the other groups from class2 to class7 performed exceptionally bad. The highest recall is 0.14, and the precision is 0.35. The admission phase model can predict short stays of 0-5 days but cannot predict the length of stay greater than five days.

3.2.2 Post-admission phase model

KNN and AdaBoost models are not performing well in post-admission phase logistic regression. The top-performing models are Random Forest, XGBoost, ANN, and CatBoost. The random forest model's accuracy, AUC, precision, recall, and f1-score are 0.8854, 0.9835, 0.74, 0.68, and 0.71, respectively. For the XGBoost model, accuracy, AUC, precision, recall, and f1-score are 0.8909, .9858, 0.7, 0.66, and 0.68, respectively. For the CatBoost model, accuracy, AUC, precision, recall, and f1-score are 8833, 0.9858, 0.7, 0.64, and 0.67, respectively. For the Artificial neural network, accuracy, AUC, precision, recall, and f1-score are 0.8692, 0.9803, 0.69, 0.62, and 0.64 respectively. In this phase, XGBoost has the highest accuracy and AUC, but the random forest has the highest precision, recall, and f1 score, for the post-admission phase 3WPE ensemble model combines Random Forest, XGBoost, and Artificial Neural network model. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.8921, 0.9863, 0.74, 0.69, and 0.71, respectively. The ensemble model outperformed its base models. The weights used for 3WPE are 0.8, 0.6, and 0.3. The 3WPE ensemble model is also evaluated with a reduced optimal feature set of 55; the results are similar to its original features. In this phase, the class1 (0-5 days) Los group has excellent performance; the recall is 0.96, and the precision is 0.96, but the other groups from class2 to class7 performed poorly compared to class1 but the performance using post-admission features has increased the performance of class2 to class7 a lot compare to the admission phase model. The lowest recall is 0.43 for class 6 (61-90 days), and precision is 0.62 for class 6 (61- 90 days). The post-admission phase model can predict the short and extended stays well compared to the admission phase, but class 3 (11-15 days) and class 6 have low prediction rates.

3.2.3 Discharge phase model

For the discharge model, logistic regression, KNN, and Adaboost models are not performing well. The top-performing models are XGBoost, CatBoost, ANN, and Random Forest. For XGBoost, accuracy, AUC, precision, recall, and f1-score are 0.9053, 0.9891, 0.74, 0.71, and 0.72, respectively. For Random Forest, accuracy, AUC, precision, recall, and f1-score are 0.8995, 0.9868, 0.77, 0.72, and 0.74, respectively; for Artificial Neural networks, accuracy, AUC, precision, recall, and f1-score are 0.8812, 0.9831, 0.71, 0.68 and 0.69 respectively. In the discharge phase, XGBoost has the highest accuracy and AUC, but the random forest has the highest precision, recall, and f1 score. The 3WPE ensemble model combines XGBoost, ANN, and Random forest models for the discharge phase. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.9065, 0.9892, 0.77, 0.73 and 0.75 respectively. The ensemble model outperformed its base models. The weights used for 3WPE are 0.8, 0.5 and 0.2. In this phase, the class1 (0-5 days) Los group has excellent performance; the recall is 0.97, and the precision is 0.96, but the other groups from class2 to class7 performed poorly compared to class1, but the performance using discharge features has increased the performance of class2 to class7 a lot compare to the previous post-admission phase model. The performance of each class increased in this phase. The lowest recall is 0.45, belongs to class 6 (61-90 days), and precision is 0.68 for class 6 (61- 90 days). The discharge phase model can predict the short and extended stays well, similar to the post-admission phase; in the discharge phase, the performance has slightly improved for each class compared to the post-admission phase. Like the post-admission phase, class 3 (11-15 days) and class 6 have low prediction rates.

Table 9: Result across all the stages

Stage	Models	Accuracy	Precision	recall	f1 score	Auc	Weights
	LR	0.6721	0.14	0.14	0.14	0.5237	X
	RF	0.672	0.27	0.18	0.19	0.7687	X
Admission	Adaboost	0.6746	0.2	0.15	0.14	0.6906	X
	XGBoost	0.6769	0.28	0.19	0.2	0.7853	X
	LightGBM	0.6806	0.29	0.18	0.18	0.7911	X
	CatBoost	0.6821	0.34	0.18	0.18	0.803	X
	ANN	0.6762	0.3	0.16	0.15	0.7619	X
	KNN	0.6563	0.22	0.16	0.16	0.6141	X
	3WPE	**0.6824**	**0.34**	**0.18**	**0.18**	**0.8042**	**0.6,0.1,0.1**
	LR	0.7324	0.14	0.14	0.13	0.608	X
	RF	0.8854	0.74	0.68	0.71	0.9835	X
Post	Adaboost	0.7475	0.45	0.48	0.46	0.7485	X
Admission	XGBoost	0.8909	0.7	0.66	0.68	0.9858	X
	LightGBM	0.8493	0.53	0.51	0.51	0.8737	X
	CatBoost	0.8833	0.7	0.64	0.67	0.9837	X

	ANN	0.8692	0.69	0.62	0.64	0.9803	X
	KNN	0.7501	0.37	0.22	0.24	0.738	X
	3WPE	**0.8921**	**0.74**	**0.69**	**0.71**	**0.9863**	**0.8,0.6,0.3**
Discharge	LR	0.732	0.14	0.14	0.13	0.6105	X
	RF	0.8995	0.77	0.72	0.74	0.9868	X
	Adaboost	0.7387	0.48	0.52	0.49	0.7387	X
	XGBoost	0.9053	0.74	0.71	0.72	0.9891	X
	LightGBM	0.8633	0.52	0.5	0.51	0.8483	X
	CatBoost	0.8966	0.74	0.69	0.71	0.9873	X
	ANN	0.8812	0.71	0.68	0.69	0.9831	X
	KNN	0.7519	0.4	0..21	0.24	0.7446	X
	3WPE	**0.9065**	**0.77**	**0.73**	**0.75**	**0.9892**	**0.8,0.5,0.2**
Post admission Reduced features (55)	RF	0.8853	0.74	0.68	0.71	0.9837	X
	XGBoost	0.8908	0.7	0.66	0.68	0.9856	X
	ANN	0.8693	0.69	0.62	0.64	0.9805	X
	CatBoost	0.8833	0.7	0.64	0.67	0.9837	X
	3WPE	**0.892**	**0.74**	**0.69**	**0.71**	**0.9861**	**0.8,0.6,0.3**

3.2.4 Feature importance

This model has three phases: admission, post-admission, and discharge. The features used in each phase have a different significance than the other phase. The models combined across the phases are random forest, XGBoost, ANN, CatBoost, and lightgbm. The artificial neural network model can not provide the importance of features, but the other four models can give importance to the features used for length of stay prediction. Please refer to Figures 7 through 13 for the plots with different feature importance against different stages.

For the admission phase, CCS Diagnostic Code, Payment typology one Medicare, cerebrovascular disease, type of admission newborn, Age group, and Type of emergency are the top 6 features voted by the XGBoost model. At the same time, the LightGBM model voted CCS diagnostic code, facility name, Zip code 3 digits, permanent facility ID, Operating certificate number, Age group, and hospital country. The CatBoost model voted CCS diagnostic code, facility name, Age group, Zip code 3 digits, Gender, permanent facility ID, and Emergency department indicator as top features.

For the post-admission phase, Surgical procedure, total cost, APR severity of illness, Myocardial infarction, type of admission emergency, emergency department indicator, and hospital service are the top features voted by the XGBoost model. At the same time, the random forest model voted total cost, CCS diagnostic code, Apr severity code, facility name, operating certificate number, APR DRG code, permanent facility ID, and surgical procedure as the top features.

For the discharge phase, Surgical procedure, total cost, APR severity of illness, service area New York City, type of admission emergency, Myocardial infarction, service area Hudson Valley, kind of admission emergency, and CCS diagnostic code are the top features voted by

XGBoost model. The random forest model voted total cost, facility name, total charges, operating certificate number, hospital country, CCS diagnostic code, permanent facility ID, APR DRG code, and CCS procedure code as the top features.

The various feature importance plot is shown in the below. The model's performance plot was plotted against a bar plot, and we can see that the model's performance improved across stages from the admission to discharge phases shown in Figure 14.

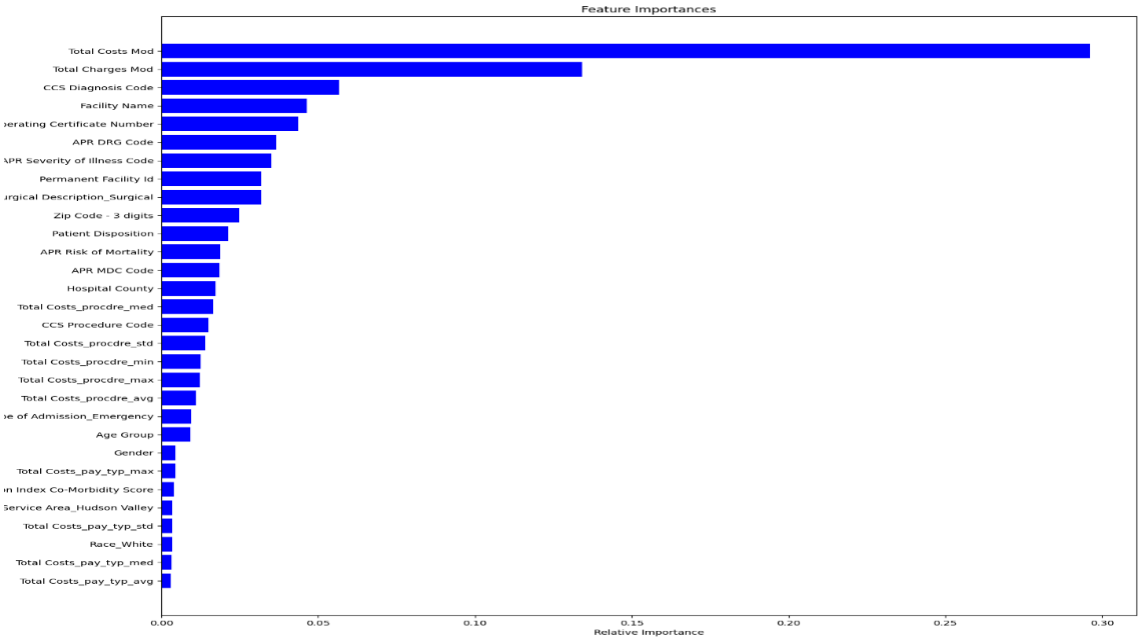


Figure 7: Feature importance of Random Forest for Top 30 features at the discharge stage

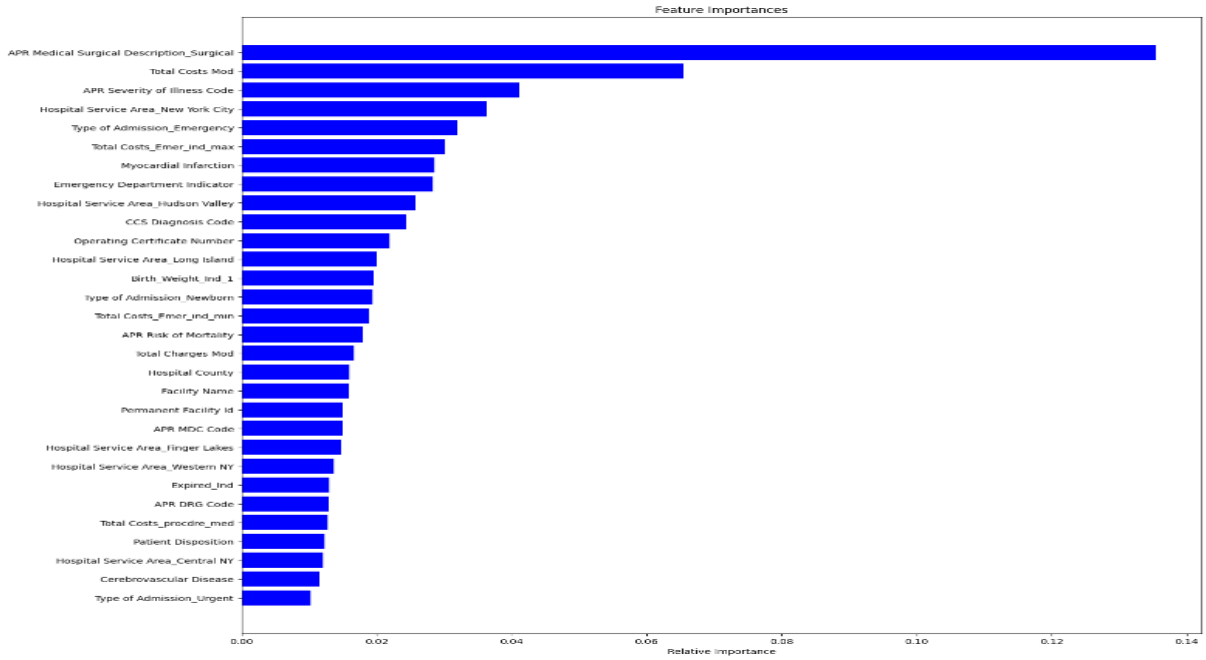


Figure 8: Feature importance of the XGBoost model for the Top 30 features at the discharge stage

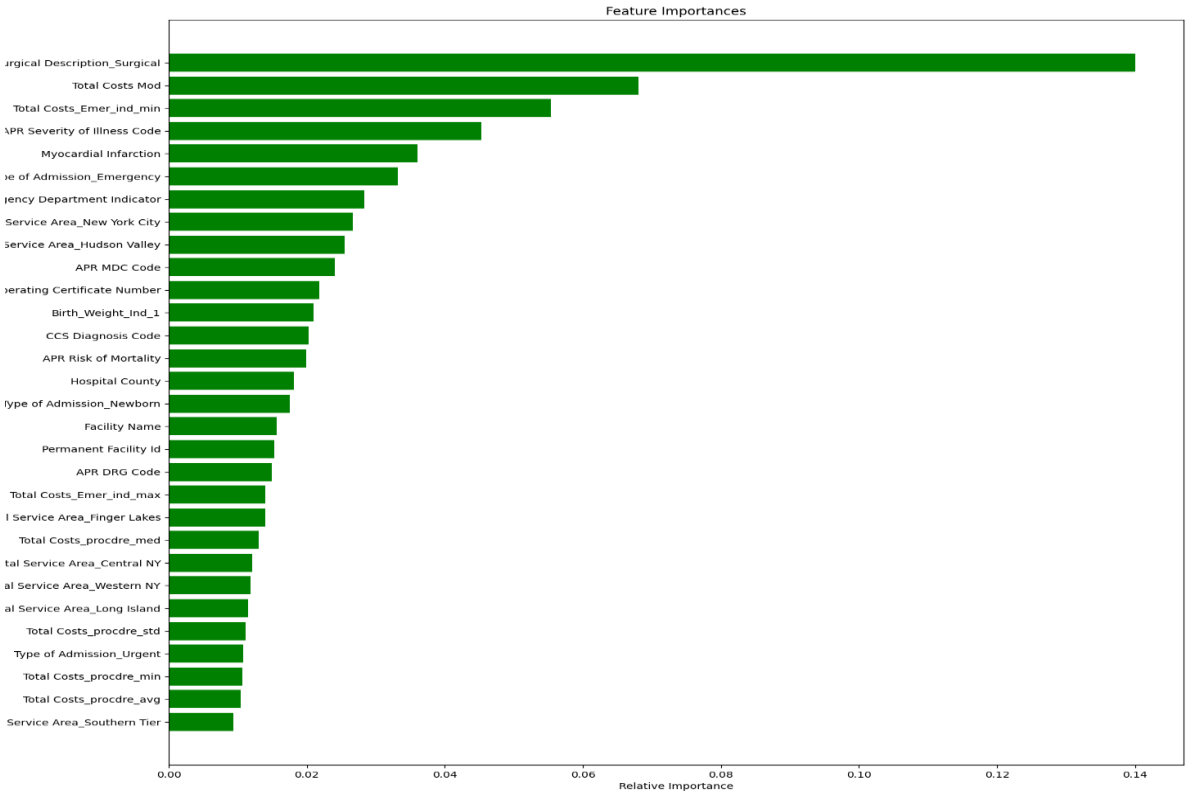


Figure 9:Feature importance of Xgboost model for Top 30 features at the post-admission stage

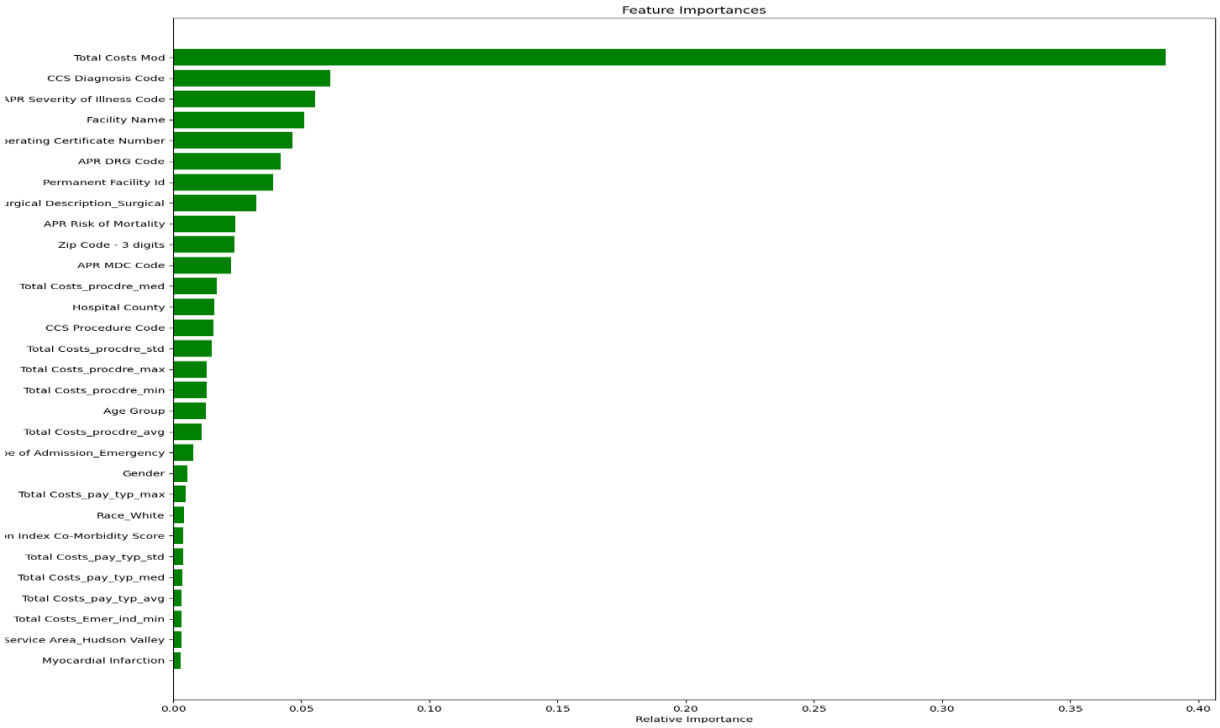


Figure 10: Feature importance of Random forest model for Top 30 features at the post-admission stage

:

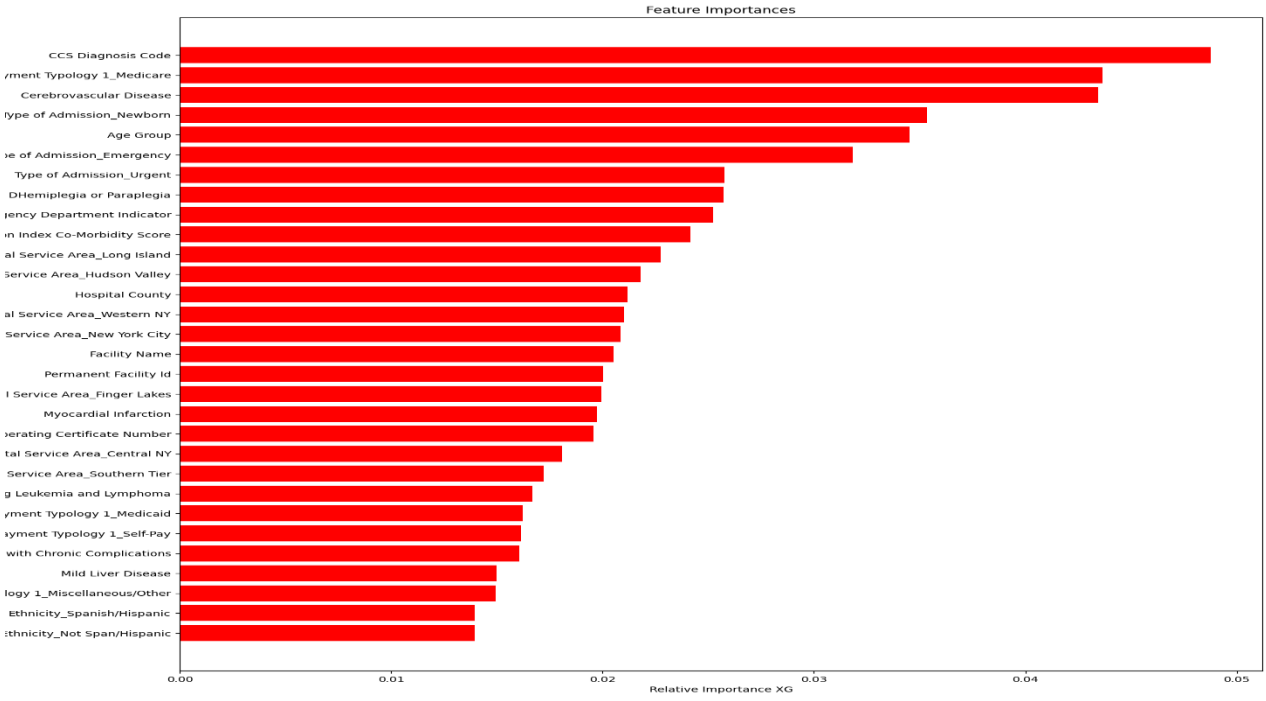


Figure 11: Feature importance of the XGBoost model for the Top 30 features at the admission stage

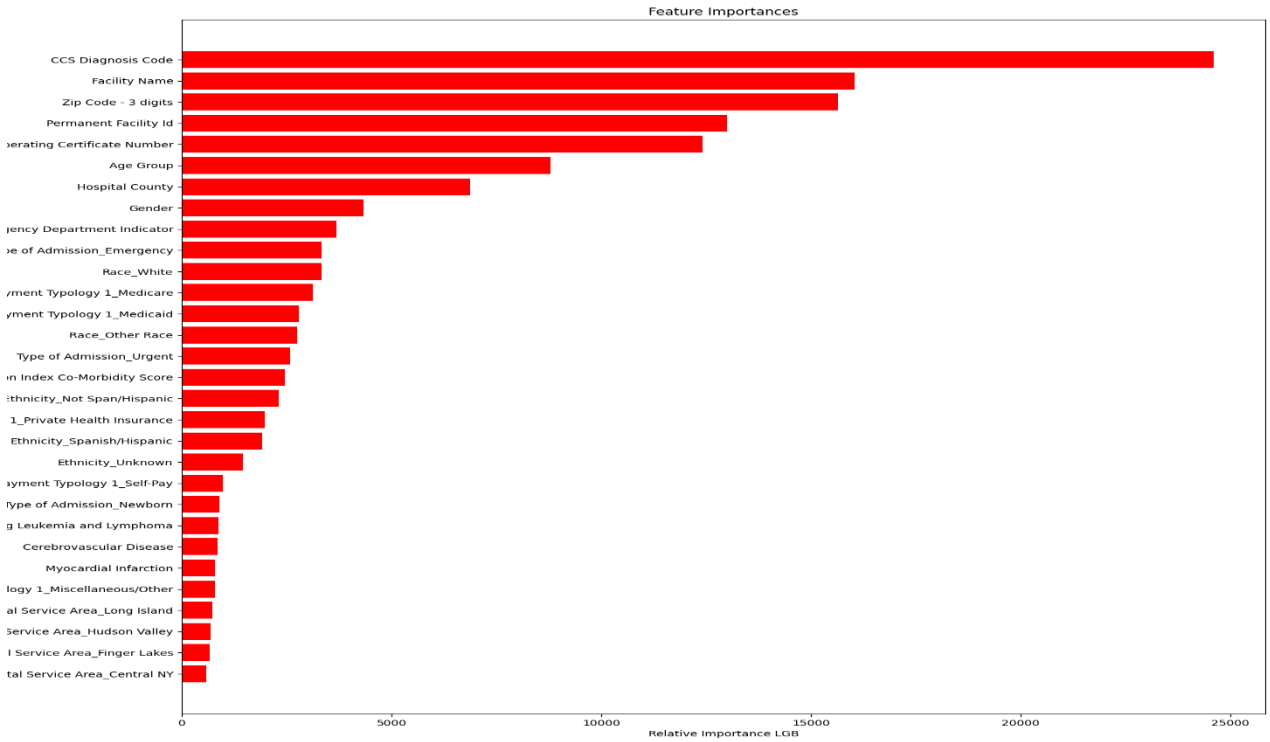


Figure 12: Feature importance of LightGBM model for Top 30 features at the admission stage

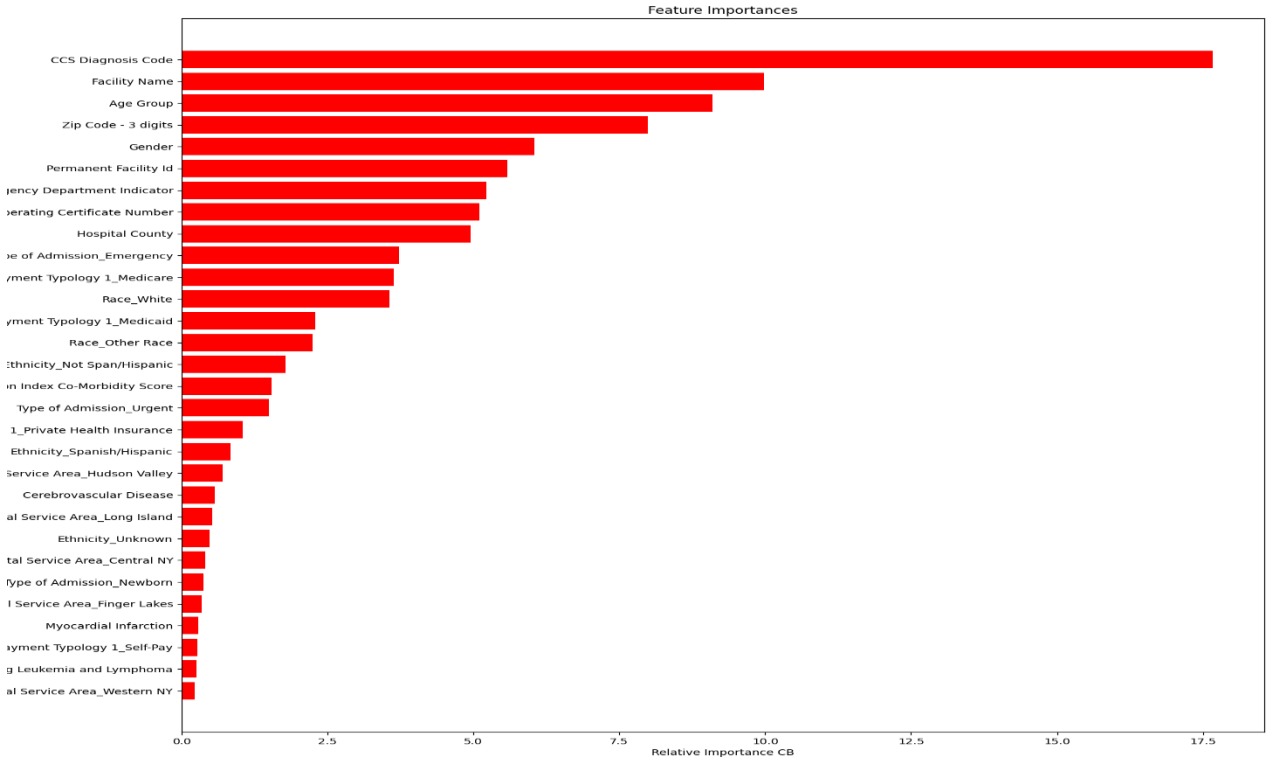


Figure 13:Feature importance of Catboost model for the Top 30 features at the admission stage

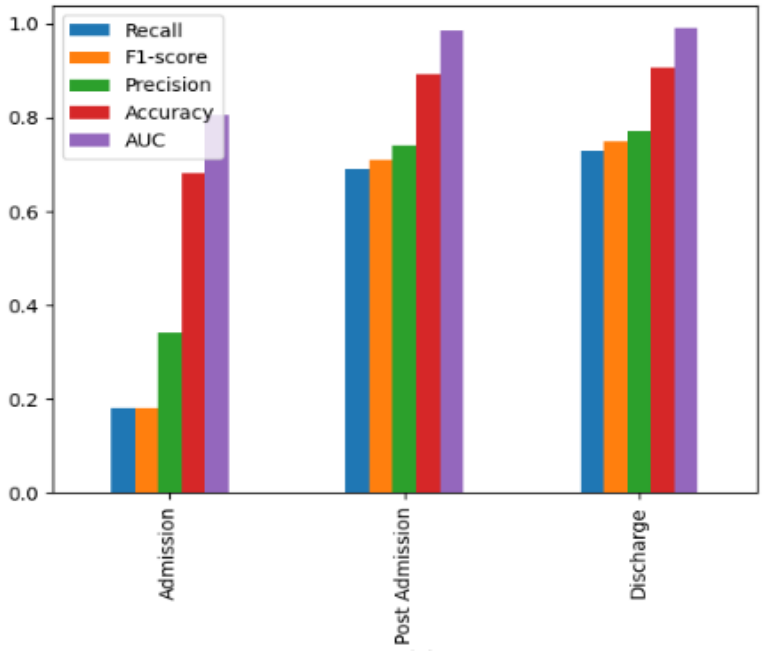


Figure 14: Performance evaluation of ensemble models in three phases

3.3 Discussion

The length of stay prediction is one of the important factors for hospitalization. The length of stay is directly interconnected to the hospital expenses and the patient's well-being. Identifying the factors leading to the correct hospital length of stay prediction is a research topic. This study uses a unique approach to predict hospital length of stay and identify the factors associated with length of stay. This prediction is made using a three-stage model: admission stage, post-admission stage, and discharge stage. The performance of the stage-wise models was improvised from the admission to the discharge phase. The accuracy, AUC, precision, recall, and f1 scores obtained in the admission phase through the 3WPE model are 0.6824, 0.8042, 0.34, 0.18, and 0.18, respectively, an improved version of the base models. The performance in the length of stay in this phase is lacking due to missing essential features. This stage can classify the short stays between 0-5 days, but beyond five days, this admission phase model can not predict the length of stay. The performance of the model beyond five days is the worst. This is due to missing essential factors for length of stay prediction. The accuracy, AUC, precision, recall, and f1 scores obtained in the admission phase through the 3WPE model are 0.8921, 0.9863, 0.74, 0.69, and 0.71, respectively, which is entirely improved from the admission phase. The performance of the ensemble model is quite good in this phase as the features included in this phase have significant importance in length of stay prediction. This post-admission model can predict the Length of stay beyond five days quite well compared to the admission phase. The predictive performance of class 3 (11-15 days) and class 6 (61-90 days) are less compared to other length of stay groups in this phase. The accuracy, AUC, precision, recall, and f1 score obtained in the discharge stage through the 3WPE model are 0.9065, 0.9892, 0.77, 0.73, and 0.75, respectively, which again improved version compared to the admission phase.

Features selected for this phase have a significant impact on Los prediction. The behavior of the discharge stage model is quite similar to that of the post-admission phase; the performance of class 3 and class 6 improved in this phase, but compared to other classes, these two classes

are still struggling with prediction. The 3WPE model has outperformed its base ML models in each phase. The admission phase has limited features related to demographics and diagnosis. The performance improved in the post-admission phase due to the availability of essential features like estimated cost, procedures performed, severity score, etc. The discharge phase performed best in this three-stage predictive model. Features like actual hospitalization cost, patient disposition, and expired Ind has been included in this phase.

This study utilizes the Charlson comorbidity score [23] and seventeen comorbidity factors for length of stay prediction. The statistical analysis shows that the Charlson comorbidity factors significantly impact the length of stay prediction. Many studies use different tools for ICU death predictors, such as SAPS APACHE, in their analysis, but we are not depending on these classical tools as these tools are designed for ICU, and our data consists of both ICU and general admissions. Considering the comorbidity factor, we were more interested in building our framework by utilizing the machine learning approach to predict the length of stay. Machine learning is more robust and performs better than these tools as these tools are based on a particular data format. This study uses comorbidity factors along with basic patient level data to verify their impact in length of stay prediction. The study shows Comorbidities have significant impact but all factors are not equally contributed to the prediction.

Compared to other studies [33], they used sparcs data to predict the Length of stay only during the admission phase. They removed discharge-type features from their dataset and kept only admission features. They have divided the model based on admission and discharge only. In earlier studies on sparcs [34], they excluded total cost, total charge, and patient disposition from the feature set, assuming that those are discharge-related features, though total cost is the estimated cost for hospitalization, which will be provided in the post-admission phase. Some studies use sparcs data to predict admission time/pre-admission time, considering all the features [10]. Our model differs from this existing approach, and we have introduced a three-phase model to give proper importance to features collected in each stage. Many studies used sampling techniques [33] to resolve the class imbalance problem. We have not applied any sampling technique in our study; rather, we have given importance to the performance of all classification metrics. The synthetic data might break the purity of the data, including a similar group of interest; no opposite group is present in our study. We have observed that short stays have better accuracy than Long stays in admission stages, but extended stay prediction improved across stages. This study using a weighted average ensemble model across three stages showed promising results, competing with the existing length of stay models[33,34,10].

However, this study has some limitations. Firstly, SPARCS does not share the patients' tab test-related parameters. Those parameters might have a good impact on accurately predicting the stay length. It also lacks features related to the number of available beds in the hospital, the number of outpatient visits performed by the patient in that year, and whether patients have any existing comorbidities like high blood pressure or diabetes re-admission. This information might have helped to predict the length of stay more accurately. Though the dataset has some limitations, we can still manage good performance.

The good part of the dataset is that it is not limited to any specific region or hospital; it has diverse categories of patients and includes many hospital data. The analysis is based on the whole data, and we have not performed any outlier removal techniques. In many studies, they applied outlier removal, and the accuracy increased[19][21]. One of the problems with the data is class imbalance; the last three Los groups, 31-60,61-90 and above 90+, have a small number

of samples. Many studies use sampling techniques to counter this issue. We have not applied any sampling technique to the data; instead, we have given importance to the precision, recall, and f1 score metrics, as those can tell the absolute truth of the prediction. Synthetic data might break its purity. A short stay has better accuracy compare to extended stay prediction.

This study shows an impactful model for length of stay prediction, capable of three distinct prediction stages from admission to discharge. This is an all-in-one model for length-of-stay prediction with varying sizes of features. This study utilizes multiple machine learning models to create a robust weighted average probability model. Hospital authorities and medical insurance companies can use this new model to balance their resources better.

Chapter 4

Proposed Methodology for Revised LOS Prediction

4.1 Problem statement

The multistage prediction performed in the previous section has some drawbacks; as the data is imbalanced, the performance of the minority classes is not that significant compared to the majority groups. The prediction performed in class 1, the Los group of 0-5 days, has the highest precision and recall across stages. The performance of Class 2 to class 7 is not as good as compare to class 1, but class 3 (11-15 days) and class 6 (61-90 days) group's performance is low across all groups. The precision-recall for the admission stage for class 3 is 0.66, 0.55 for the discharge stage, 0.70, and 0.61, respectively. The precision-recall for a post-admission stage for class 6 is 0.62, 0.43 for the discharge stage, 0.68, 0.45, respectively. The performance metrics for class 3 and class 6 are lower compare all other los groups. For the admission stage, the performance of classes other than class 1 is worse; all other groups are performing poorly. The performance of class 3 and class 6 impacted the model's overall performance in the post admission and discharge phases. In this study, we aim to improve the overall performance of the LOS prediction across all the length of stay class groups by addressing the issues in class 3 and class 6. The following section describes the methods and the processes.

4.2 Solution

In this section, we will discuss the solution and methods implemented to improve the model's performance, addressing the above mentioned issues. The methodology does not implement any sampling techniques to boost the minority classes. The groups of the length of stays we are dealing with in this study are not meant to describe opposite groups. An example of an opposite grouping is disease prediction, which refers to whether a patient has cancer. Our study deals with similar information throughout the groups with different periods. We have analyzed the confusion matrix of the various stages of the developed model and found that class 3's misprediction is higher in class 2, and class 6's misprediction is higher in class 5. The 3WPE model's misclassification of class 3 is mainly happening in class 2, and class 6 is happening in class 5. In this study, to address the issue, we have clubbed classes 2 and class3 together and 5 and class6 together to form new groups. In this new approach, we have decreased the classes from 7 to 5, and the latest sets of LOS classes are 0-5, 6-15, 16-30, 31-60, and > 60. In this methodology, the existing machine learning approaches that were used in the previous study has been used; we have included a few more models from both machine and deep learning models. Those are Naïve Bayes, Decision trees, CNN 1D and Tabnet. The CNN models are mainly used for image processing tasks; they use 2D convolution in those tasks. Our data is structured tabular data, so 2D convolution cannot be used; instead, we used the CNN 1D model, which is a dimension convolution model. Apart from the CNN model, we have used a tabular deep learning model called Tabnet for loss prediction. Nowadays, deep learning models are gaining interest in tabular data [56-58]. A few other studies use deep learning models for

prediction using tabular data. Los' prediction using deep learning is not evaluated as such. This approach utilizes deep learning and machine learning to predict Los. This prediction also uses the weighted average ensemble model discussed in the previous chapter but with few new deep and machine learning models. This combination of machine learning and deep learning is an unexplored area for length-of-stay prediction.

4.3 Methods

The data repository and pre-processing are not altered in this chapter, as all the processing is the same with respect to Chapter 3. For reference, please refer to Chapter 3 methods for the details. Only the Los class distribution has been changed in this study. The new set of class mappings is given in the below table, table 10:

Table 10: Los groups using 5 classes

Group (In days)	Patient count
0-5 days	1692034
6-15 days	483453
16-30 days	97595
31-60 days	26343
>60 days	8426

4.3.1 Proposed Method

This study also utilizes the proposed three stages of modeling: admission, post-admission, and discharge. The feature subset used in this study is the same as the previous study in Chapter 3. We are not giving the features list again in this section; please visit Chapter 3, proposed model section, for reference. The model that will be utilized is the same novel weighted probability model suggested by the previous chapter, chapter 3. The only difference is that we have used new models in this prediction: Naïve Bayes, a decision tree, and two deep learning models, CNN 1D and Tabnet.

The introduction regarding the new models is given below:

4.3.1.1 Naive Bayes

A collection of supervised learning algorithms known as naive Bayes methods is based on applying the Bayes theorem and the naive assumption of conditional independence between each pair of features given the value of the class variable. Naive Bayes classifiers appear to have oversimplified assumptions, but they have proven highly effective in many real-world scenarios, most notably spam filtering and document classification. A minimal amount of training data is needed to estimate the required parameters. Comparing naive Bayes learners and classifiers against more advanced techniques reveals how fast they can be. We have used default Sklearn settings for the algorithm across all the stages.

4.3.1.2 Tabnet

In 2019, researchers at Google Cloud presented TabNet. TabNet aims to efficiently apply deep neural networks to tabular data, which is still processed data from various applications, including marketing, banking, healthcare, retail, and finance. Deep learning in the tabular dataset is motivated by its notable performance gain on massive datasets when applied to other domains, such as image, language, and speech data, compared to previous machine learning approaches. Thus, it should function with tabular data. Another explanation might be that, in contrast to deep neural networks, tree-based algorithms are less effective in learning how to minimize error by applying methods like gradient descent. TabNet offers an interpretable and high-performing deep learning architecture for tabular data. It chooses which characteristic to select to produce high interpretability and practical training using a technique known as the sequential attention mechanism. In our model, we have used the number of steps =5, width of decision prediction layer and width of attention embedding = 64, Number of independent Gated Linear Units, and Number of shared Gated Linear Units =2 across all the stages.

4.3.1.3 CNN 1D

One kind of Deep Learning neural network design frequently utilized in computer vision is the convolutional neural network (CNN). The branch of artificial intelligence, "computer vision," allows computers to comprehend and analyze images and other visual input. The expanded form of artificial neural networks, convolutional neural networks (CNNs), is primarily used to extract features from grid-like matrix datasets—for instance, visual datasets with many data patterns, such as pictures or movies. 2D CNN is a popular CNN model mainly used for image classification tasks. Apart from 2D CNN, 1D and 3D CNN is also available. 3D CNN uses high-dimension image data, while 1D CNN is predominantly used for tabular time series data. As our data is in tabular form, we have chosen the CNN 1D model. The application of CNN 1D on Length of stay prediction is new to this domain. We have used 3 CNN 1D layers with max pooling and three dense layers for our stage-wise predictive model.

4.3.1.4 Decision Tree

A decision tree is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing outcomes or predictions. Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value.

The process of creating a decision tree involves:

1. Selecting the Best Attribute: Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.
2. Splitting the Dataset: The dataset is divided into subsets based on the selected attribute.
3. Repeating the Process: The process recursively for each subset, creating a new internal node or leaf node until a stopping criterion is met (e.g., all instances in a node belong to the same class or a predefined depth is reached) [51].

We have kept the max depth for the decision tree at 25 throughout the stages.

4.3.1.5 Our Architecture

Our study uses the weighted average probability ensemble base developed in the previous chapter. This model will take different well-tuned model's predicted probabilities and averages the probabilities based on the weight assigned to each classifier. The classifier's weights will be selected based on the contribution of each model towards generating an optimal ensemble model through a robust grid search method. We are proposing a 3WP and 4WP ensemble model. The 3WP and 4WP ensemble models will be selected based on the top three or four performing base models. In this study, apart from the 3WP ensemble model, we have also introduced another 4WP model for our analysis. The algorithm for the prediction is discussed below:

4.3.1.5.1 Algorithm

Input: *Training dataset and testing dataset.*

Procedure: *The steps below will be repeated in three stages: Admission, Post-admission, and discharge.*

1. Train different base models.
2. Test the base models and compare the accuracy, precision, recall, and f1 score.
3. Calibrate the individual base models to obtain optimum precision.
4. Select top models based on descending order of classification metrics considering accuracy, precision, recall, and f1 score.
5. Random Forest, XGBoost, Tabnet, CatBoost, and LightGBM are the top-performing models selected.
6. Select the top three or four models based on balanced classification metrics.
7. Obtain an initial weighted average ensemble model using equal weights.
8. Calculate the weights for each model using the grid search method and assign optimum weights to each classifier.
9. Apply weighted probability average ensemble for prediction using optimized weights.
10. Generate probability-weighted voting with a new set of classes.
11. Compare and evaluate the precision of the new ensemble model with the base models.

The architecture for the model is the same, but it only differs with the inclusion of the new models. Please refer to Figure 15 for the schematic diagram of the process flow.

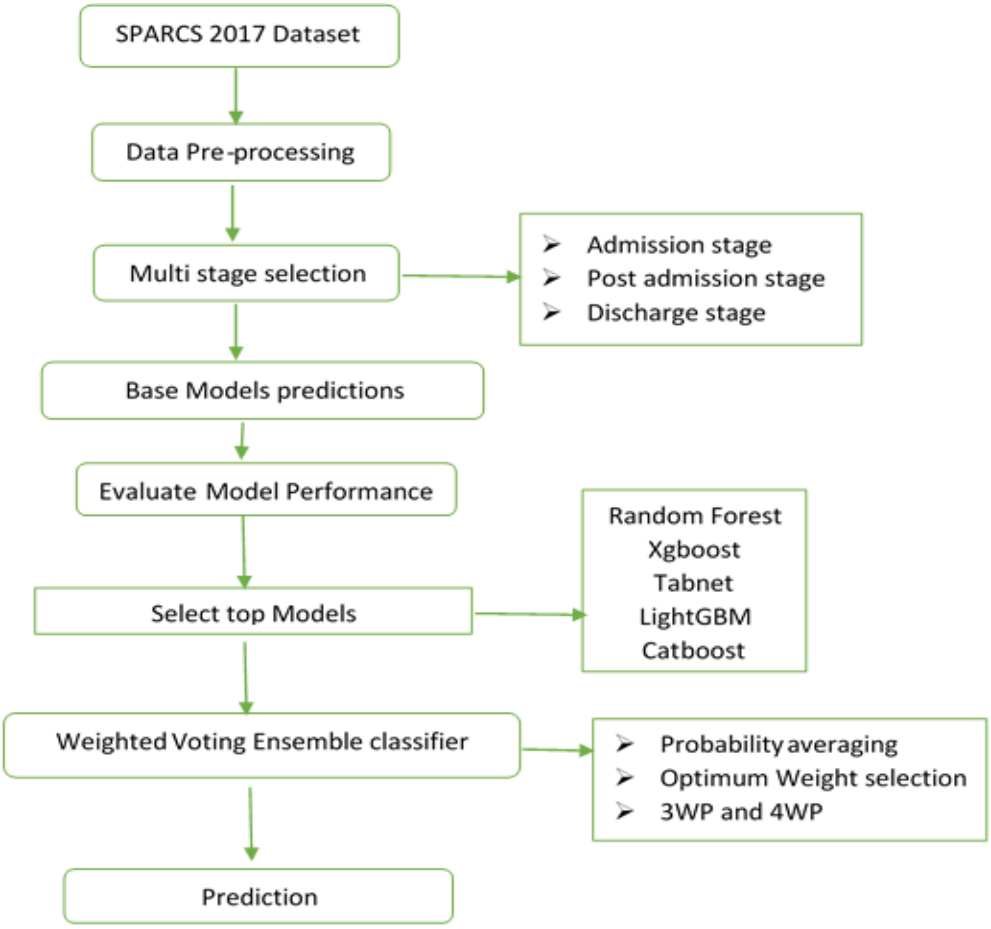


Figure 15: Process flow diagram of revised Length of stay prediction

4.3.1.5.2 Pseudo Code of the Model

1. Prompt the user to enter the number of classifiers for combination and store it in `num_of_classifier.`

2. If `num_of_classifier` is equal to 4:
   a. Define a list of `models` containing tuples of models and their abbreviations: Random Forest (Randmreg), XGBoost (xgb_r), TabNet (tabnet), and CatBoost (cb_r).
   Else:
   b. Define a list of `models` containing tuples of models and their abbreviations: Random Forest (Randmreg), XGBoost (xgb_r), and TabNet (tabnet).
3. Predict probabilities for each model on the scaled test data `X_test1_scaled` and store them in a list of `preds`.
4. Convert the list of predictions `preds` into a NumPy array `array_opt`.
5. Calculate the soft ensemble of predictions by averaging the predictions in `array_opt` along the first axis, with equal weights [0.1, 0.1, 0.1, 0.1] if `num_of_classifier` is 4, else [0.1, 0.1, 0.1].
6. Determine the class predictions of the soft ensemble by selecting the maximum value index along the second axis. Store the result in `soft_ensemble_pred`.
7. Calculate the classification report, accuracy, and confusion matrix for the soft ensemble predictions.
8. Initialize an empty DataFrame `df_new`.
9. Set variables `k` to 0 and `a` to 250.
10. Iterate through different combinations of weights (w1, w2, w3, w4) in the ranges [0.1, 0.2, ..., 0.8]:
    a. Calculate a weighted ensemble of predictions by averaging the predictions in `array_opt` with the current weights (w1, w2, w3, w4).
    b. Determine the class predictions of the weighted ensemble.

c. Calculate accuracy, balanced accuracy, precision, recall, and F1-score for the weighted ensemble predictions.
d. Increment `k`.
e. If `k` equals `a`, print the current count.
f. Append the calculated metrics along with the weights to the DataFrame `df_new`.
11. Sort `df_new` by accuracy in descending order.
12. If `num_of_classifier` is equal to 4, set `n` to 3900, else set it to 412.
13. Remove the last `n` rows from `df_new`.
14. Identify the row in `df_new` with the highest precision and store it in `max_acc_row`.
15. Calculate the soft ensemble using the optimal weights from `max_acc_row`.
16. Determine the class predictions of the soft ensemble.
17. Calculate classification report, accuracy, confusion matrix, precision, recall, F1-score, balanced accuracy, and AUC-ROC for the soft ensemble predictions.
18. Print the classification report, accuracy, confusion matrix, precision, recall, F1-score, and balanced accuracy.

4.3.1.5.3 Hyperparameter Tuning of base models

The weighted ensemble model generated using the base models needs the base models to be highly précised and calibrated. Well-tuned base models give more power to the ensemble model. If the base model's performance is not nearly equal or similar, the ensemble model might not perform well. We have identified three top-performing models based on different stages. Those top models have been tuned to get optimal results for the dataset. The calibrated models are XGBoost, Tabnet, random forest, LightGBM, and CatBoost. The final tuned parameters are given in the proposed model section, where we describe the models. Apart from the tablet, CNN, decision tree, and naïve Bayes model for all other tuned parameters, please refer to the previous chapter. The tuned base models help the ensemble model to produce optimal results.

4.4 Evaluation criteria

For classification problems, classic evaluation metrics are accuracy, precision, recall, F1-score, and AUC-ROC value. We will be using these metrics to validate our multi-stage model. These metrics define different aspects of the model's performance. We aim to produce maximum efficiency from these metrics by combining the top-performing base models.

4.5 Result

Initially, the USA SPARCS dataset had 2343569 records with 34 features. After data pre-processing, duplicate removal, feature selection, and feature engineering, we have 2308391 instances left in our dataset. The feature dimension increased from 34 to 82. We have used the online data science and ML framework Kaggle for our coding and analysis. Kaggle is an open-source online machine learning and data science research platform. The platform uses Python 3.10 with the latest Scikit Learn, TensorFlow, and Keras ML packages. The models are trained using 80% of the data, and the remaining 20% is used for testing. The comparative study of different stages of the proposed model is illustrated in Table 11. The performance of base models and the weighted average model are also presented in the table. The final optimum weights of the 3WPE and 4WPE models are presented in the table. The optimum weight search space is used between 0.1 to 0.8. The number of features utilized for different stages are 55, 79, and 82 for admission, post-admission, and discharge stages, respectively. Two other

evaluations are performed: one is based on a complete feature set and another with a reduced feature. The reduced feature set evaluation is only applicable during the post-admission phase.

4.5.1 Admission phase model

The accuracy of admission stage models lies between 0.67-0.69. Most of the models' performance is similar in accuracy but varies in precision, recall, f1 score, and Auc scores. In this study, LightGBM, CatBoost, and XGBoost are performing well not only with respect to accuracy but also with respect to other parameters, as compared to other models. For AUC, LightGBM tops the table with a value of 0.8068, and XGBoost is in second place at 0.7922. LightGBM is the winner for precision with a value of 0.44. XGBoost has the highest recall and f1-score, 0.27 and 0.28, respectively. For accuracy, LightGBM is the winner with a value of 0.6996. The performance of random forests and ANN is insignificant at this stage. The new models CNN 1D and Tabnet are performing on average. The 3WPE ensemble classifier combines CatBoost, XGBoost, and LightGBM in the admission phase. The ensemble accuracy is 0.6995, and precision, recall, f1 score, and auc are 0.44, 0.26, 0.27, and 0.8087, respectively, which perform decently similarly concerning top base model LightGBM, but the AUC value is improved in 3WPE model. The weights used for 3WPE are 0.7, 0.1, and 0.4. The class1 (0-5 days) Los group performed excellently; the recall is 0.94 and precision is 0.74, but the other groups from class2 to class7 performed exceptionally severely. The highest recall is 0.25, and the precision is 0.47. The admission phase model can predict the short stays of 0-5 days; the prediction for class 2 6-15 days improved slightly but not up to the mark with a recall of 0.25 and precision of 0.47. Other groups' performance was similar to the previously developed model. The 4WPE model is not implemented in this stage as the peformnace in this stage was not that significant.

4.5.2 Post-admission phase model

In post-admission phase logistic regression, KNN, Naïve Bayes, and adaboost models are not performing well. Top performing models are Random Forest, XGBoost, Tabnet, CatBoost, and CNN 1D. The random forest model's accuracy, AUC, precision, recall, and f1-score are 0.9114, 0.9862, 0.81, 0.77, and 0.79 respectively. For the XGBoost model, accuracy, AUC, precision, recall, and f1-score are 0.9160, .9882, 0.79, 0.76, and 0.78 respectively. For the CatBoost model, accuracy, AUC, precision, recall, and f1-score are 0.9139, 0.9877, 0.8, 0.75, and 0.77 respectively. For Tabnet, accuracy, AUC, precision, recall, and f1-score are 0.9084, 0.9863, 0.81, 0.77, and 0.79, respectively. For the CNN 1D model, accuracy, AUC, precision, recall, and f1-score are 0.9015, 0.9846, 0.79, 0.75 and 0.77 respectively. In this phase, XGBoost has the highest accuracy and AUC, but random forest and tabnet have the highest precision, recall, and f1 score. The 3WPE ensemble model combines Random Forest, XGBoost, and Tabnet for the admission phase. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.9178, 0.9888, 0.82, 0.78, and 0.80, respectively. The 3WPE ensemble model outperformed its base models. The weights used for 3WPE are 0.8, 0.7, and 0.6. We have also combined the top four models in this phase to create 4WPE: Random Forest, XGBoost, Tabnet, and CatBoost. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.9179, 0.9889, 0.82, 0.78, and 0.80, respectively. The weights used for 4WPE are 0.5, 0.6, 0.5 and 0.2. The performance of the 3WPE and 4WPE are similar. The 4WPE needs a higher optimal weight search time. However, the 3WPE is the preferred one for the modeling. The 3WPE ensemble model is also evaluated with a reduced optimal feature set of 55; the results are similar to its original features. In this phase, the class1 (0-5 days) Los group performs excellently; the recall is 0.96, and the precision is 0.96. The performance of the other groups

from class 2 to class 5 is lower compared to class 1 but is much improved compared to the previous one. This stage's minimum precision and recall are 0.74 and 0.66 for class 4 (31-60 days). The post-admission stage model's performance is far better than the admission stage. This model can better predict both short and long stays.

4.5.3 Discharge phase model

For the discharge model, logistic regression, KNN, Naïve Bayes, and Adaboost models are not performing well. The top-performing models are XGBoost, CatBoost, Tabnet, Random Forest, and CNN 1D. For XGBoost, accuracy, AUC, precision, recall, and f1-score are 0.9293, 0.9914, 0.82, 0.8, and 0.81, respectively. For Random Forest, accuracy, AUC, precision, recall, and f1-score are 0.9235, 0.9893, 0.84, 0.8, and 0.82, respectively. For Tabnet, accuracy, AUC, precision, recall, and f1-score are 0.9232, 0.9897, 0.83, 0.81, and 0.82, respectively. For the CatBoost model, accuracy, AUC, precision, recall, and f1-score are 0.9258, 0.9906, 0.83, 0.78, and 0.8, respectively. For the CNN 1D model, accuracy, AUC, precision, recall, and f1-score are 0.9145, 0.988, 0.83, 0.75, and 0.78 respectively. In the discharge phase, XGBoost has the highest accuracy and AUC, but random forest and tabnet have the highest precision, recall, and f1 score. The 3WPE ensemble model combines XGBoost, Tabnet, and Random forest models for the discharge phase. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.9320, 0.992, 0.85, 0.82, and 0.84 respectively. The ensemble model outperformed its base models. The weights used for 3WPE are 0.7, 0.7 and 0.6. We have also combined the top four models in this phase to create 4WPE: Random Forest, XGBoost, Tabnet, and CatBoost. The combined model's accuracy, AUC, precision, recall, and f1-score are 0.9320, 0.992, 0.85, 0.82 and 0.83 respectively. The weights used for 4WPE are 0.7, 0.8, 0.8 and 0.2. The performance of the 3WPE and 4WPE are similar. The 4WPE needs a higher optimal weight search time. However, the 3WPE is preferred for modeling in this phase. The class1 (0-5 days) Los group performs excellently; the recall is 0.97, and the precision is 0.97. The performance of the other groups from class 2 to class 5 is lower than class 1, but it is much improved compared to the previous one. This stage's minimum precision and recall are 0.78 and 0.72 for class 4 (31-60 days). The discharge stage model performs far better than the post-admission stage. This model can better predict both short and long stays. The discharge phase model has a better prediction of the length of stay.

4.5.4 The tuning of the Weighted average Ensemble Model

Properly tuning weighted ensemble models can produce better results, and different aspects of the models' performance parameters can be evaluated. In this approach, we have tried to change the weights of the weighted ensemble model to achieve the highest accuracy, precision, recall, and f1-score across all the stages. The tuning aims to achieve higher single metrics performance by compromising other metrics. The calculated weights and results of such tuning are shown in Table 12. The **boldness** signifies that particular metrics have been given top priority. For the admission phase, we used the 3WPE, so tuning was performed only on this 3WPE model. We tuned the model to achieve the highest accuracy, precision, recall, and f1-score values. The highest individual metrics have been presented in **bold.** For the admission stage, the performance of the balanced 3WPE accuracy is 0.6995, and precision, recall, and f1 scores are 0.44, 0.26, and 0.27. The following table shows the variation of each metric to achieve its higher value and its corresponding other metrics values are also displayed. The highest

accuracy, precision, recall, and f1-score achieved by this are 69.98, 0.44, 0.27, and 0.28, respectively, but this can not be achieved altogether.

The post-admission phase has implemented two weighted ensemble models: 3WPE and 4WPE. The highest individual metrics performance through adjusting weights has been published in the table. For highest accuracy model the weights are given in the table as well as for highest recall model the weights according to it are provided in table. Using both the 4WPE and 3WPE tuning has been performed. The balanced model we got through ensembling can be compared through the table. Our model is selected based on giving equal importance to all the metrics, but only individual metrics are given priority in this table.

For example, the 3WPE model's highest accuracy, precision, recall, and f1-score achieved by this tuning is 0.9180, 0.83, 0.78, and .80, but this can not be achieved. Our balanced 3WPE model solves this problem and tries to get a balanced result to reach near the highest for all the metrics. The tuned ensemble models are also provided in the 3WPE and 4WPE models table for the discharge stage. In the ensemble model, we have produced a balanced metrics performance compared to the highest metrics results achieved in the table. If any requirement arises to prioritize single metrics, the table will be referred to address the model.

Table 11: Different base model and ensemble model performances

Stage	Models	Accuracy	Precision	recall	f1 score	Auc	Weights
Admission	LR	0.6718	0.42	0.2	0.17	0.659	X
	RF	0.6721	0.33	0.26	0.27	0.7474	X
	Adaboost	0.6849	0.33	0.22	0.21	0.7073	X
	XGBoost	0.693	0.38	0.27	0.28	0.7922	X
	LightGBM	0.6996	0.44	0.26	0.27	0.8068	X
	CatBoost	0.698	0.41	0.26	0.27	0.8038	X
	ANN	0.6878	0.4	0.24	0.23	0.7671	X
	KNN	0.6639	0.32	0.24	0.24	0.6414	X
	NB	0.6714	0.31	0.2	0.17	0.6319	X
	DT	0.6741	0.32	0.26	0.27	0.6781	X
	CNN	0.6873	0.43	0.24	0.23	0.7572	X
	Tabnet	0.6862	0.34	0.23	0.22	0.7606	X
	3WPE	**0.6995**	**0.44**	**0.26**	**0.27**	**0.8087**	**0.7,0.1,0.4**
Post Admission	LR	0.8206	0.46	0.34	0.36	0.9235	X
	RF	0.9114	0.81	0.77	0.79	0.9862	X
	Adaboost	0.8135	0.63	0.61	0.61	0.8156	X
	XGBoost	0.916	0.79	0.76	0.78	0.9882	X
	LightGBM	0.8997	0.72	0.7	0.71	0.9717	X
	CatBoost	0.9139	0.8	0.75	0.77	0.9877	X
	ANN	0.8964	0.76	0.76	0.76	0.9826	X
	KNN	0.7741	0.52	0.31	0.34	0.7767	X
	NB	0.6611	0.37	0.25	0.24	0.7194	X

	DT	0.8884	0.74	0.72	0.73	0.8941	X
	CNN	0.9015	0.79	0.75	0.77	0.9846	X
	Tabnet	0.9084	0.81	0.77	0.79	0.9863	X
	3WPE	**0.9178**	**0.82**	**0.78**	**0.8**	**0.9888**	**0.8,0.7,0.6**
	4WPE	**0.9179**	**0.82**	**0.78**	**0.8**	**0.9889**	**0.5,0.6,0.5,0.2**
Discharge	LR	0.8299	0.48	0.38	0.4	0.93	X
	RF	0.9235	0.84	0.8	0.82	0.9893	X
	Adaboost	0.8168	0.65	0.62	0.63	0.8247	X
	XGBoost	0.9293	0.82	0.8	0.81	0.9914	X
	LightGBM	0.9054	0.74	0.72	0.73	0.9715	X
	CatBoost	0.9258	0.83	0.78	0.8	0.9906	X
	ANN	0.9052	0.8	0.76	0.78	0.985	X
	KNN	0.778	0.53	0.31	0.34	0.7846	X
	NB	0.6644	0.39	0.25	0.24	0.7234	X
	DT	0.8973	0.77	0.74	0.75	0.8956	X
	CNN	0.9145	0.83	0.75	0.78	0.988	X
	Tabnet	0.9232	0.83	0.81	0/82	0.9897	X
	3WPE	**0.932**	**0.85**	**0.82**	**0.84**	**0.992**	**0.7,0.7,0.6**
	4WPE	**0.932**	**0.85**	**0.82**	**0.83**	**0.992**	**0.7,0.8,0.8,0.2**
Post admission Reduced features -55	RF	0.9113	0.81	0.77	0.79	0.9863	X
	XGBoost	0.9161	0.79	0.76	0.78	0.9883	X
	Tabnet	0.9085	0.81	0.77	0.79	0.9864	X
	CatBoost	0.914	0.8	0.75	0.77	0.9878	X
	3WPE	**0.9177**	**0.82**	**0.78**	**0.8**	**0.9887**	**0.8,0.7,0.6**

Table 12: Tuning of ensemble model to produce optimum results

Stage	Models	Accuracy	Precision	recall	f1 score	Weights
Admission	3WPE	**69.98**	0.4367	0.2590	0.2674	0.8,0.1,0.1
	3WPE	69.96	**0.4376**	0.2583	0.2663	0.8,0.1,0.4
	3WPE	69.55	0.3875	**0.2651**	0.2761	0.1,0.8,0.1
	3WPE	69.55	0.3875	0.2651	**0.2761**	0.1,0.8,0.1
Post Admission	3WPE	**91.80**	0.8141	0.7786	0.7955	0.5,0.7,0.3
	3WPE	91.56	**0.8286**	0.7822	0.8040	0.8,0.1,0.6
	3WPE	91.70	0.8274	**0.7834**	0.8042	0.7,0.4,0.7
	3WPE	91.61	0.8285	0.7827	**0.8043**	0.7,0.2,0.7
	4WPE	**91.81**	0.8145	0.7758	0.7942	0.5,0.7,0.3,0.3
	4WPE	91.62	**0.8288**	0.7813	0.8037	0.7,0.2,0.8,0.1
	4WPE	91.65	0.8269	**0.7824**	0.8034	0.3,0.3,0.7,0.1
	4WPE	91.68	0.8285	0.7822	**0.8040**	0.8,0.3,0.8,0.1

	3WPE	**93.22**	0.8431	0.8157	0.8288	0.1,0.3,0.2
	3WPE	93.12	**0.8546**	0.8213	0.8372	0.5,0.2,0.4
	3WPE	93.15	0.8539	**0.8232**	0.8379	0.7,0.5,0.8
	3WPE	93.15	0.8539	0.8232	**0.8379**	0.7,0.5,0.8
Discharge	4WPE	**93.22**	0.8440	0.8159	0.8294	0.3,0.8,0.6,0.1
	4WPE	93.08	**0.8546**	0.8189	0.8359	0.4,0.1,0.3,0.1
	4WPE	93.15	0.8536	**0.8221**	0.8372	0.7,0.5,0.8,0.1
	4WPE	93.15	0.8536	0.8221	**0.8372**	0.7,0.5,0.8,0.1

4.5.5 Feature importance

This model has three phases: admission, post-admission, and discharge. The features used in each phase have a different significance than the other phase. The models combined across the phases are random forest, XGBoost, Tabnet, CatBoost, and lightgbm. As we are not changing anything in the processing or modeling, the feature importance of the prior ensemble can be referred to for common models. Please refer to Figures 16 through 19 for the plots with different feature importance against different stages.

For the admission phase, CCS Diagnostic Code, Payment typology one Medicare, cerebrovascular disease, type of admission newborn, Age group, and Type of emergency are the top 6 features voted by the XGBoost model. At the same time, the lightGBM model voted CCS diagnostic code, facility name, Zip code 3 digits, permanent facility ID, Operating certificate number, Age group, and hospital country. The CatBoost model voted CCS diagnostic code, facility name, Age group, Zip code 3 digits, Gender, permanent facility ID, and Emergency department indicator as top features.

For the post-admission phase, Surgical procedure, total cost, APR severity of illness, Myocardial infarction, type of admission emergency, emergency department indicator, and hospital service in New York are the top features voted by the XGBoost model. In comparison, the random forest model voted total cost, CCS diagnostic code, Apr severity code, facility name, operating certificate number, APR DRG code, permanent facility ID, and surgical procedure as the top features. The tabnet model votes total costs, race white, admission not available, ethnicity not span / Hispanic, Moderate or Severe Liver Disease, birth weight ind, abortion ind, service area Hudson Valley as the top features. The CatBoost model votes total cost, facility name, CCS diagnostic code, hospital country, operating certificate number, permanent facility ID, and CCS procedure code as the top performers.

For the discharge phase, Surgical procedure, total cost, APR severity of illness, service area New York City, type of admission emergency, Myocardial infarction, service area Hudson Valley, type of admission emergency, and CCS diagnostic code are the top features voted by XGBoost model. The random forest model voted total cost, facility name, total charges, operating certificate number, hospital country, CCS diagnostic code, permanent facility ID, APR DRG code, and CCS procedure code as the top features. The tabnet model votes total costs, Leukemia and lymphoma, renal disease, hospital service area Long Island, abortion ind, total charges, and patient disposition as the top performers. The CatBoost model votes for total

costs, facility name, total charges, operating certificate number, hospital country, CCS diagnostic code, permanent facility ID, APR DRG code, and CCS procedure code as the top performers.

The various feature importance plot is shown in the below. In this part, we have only demonstrated the importance of the plot for the CatBoost and Tabnet models feature. The detailed feature importance value obtained through different models has been presented from Table 13 to Table 18. The contribution of essential features towards the length of stay prediction has been successfully identified.

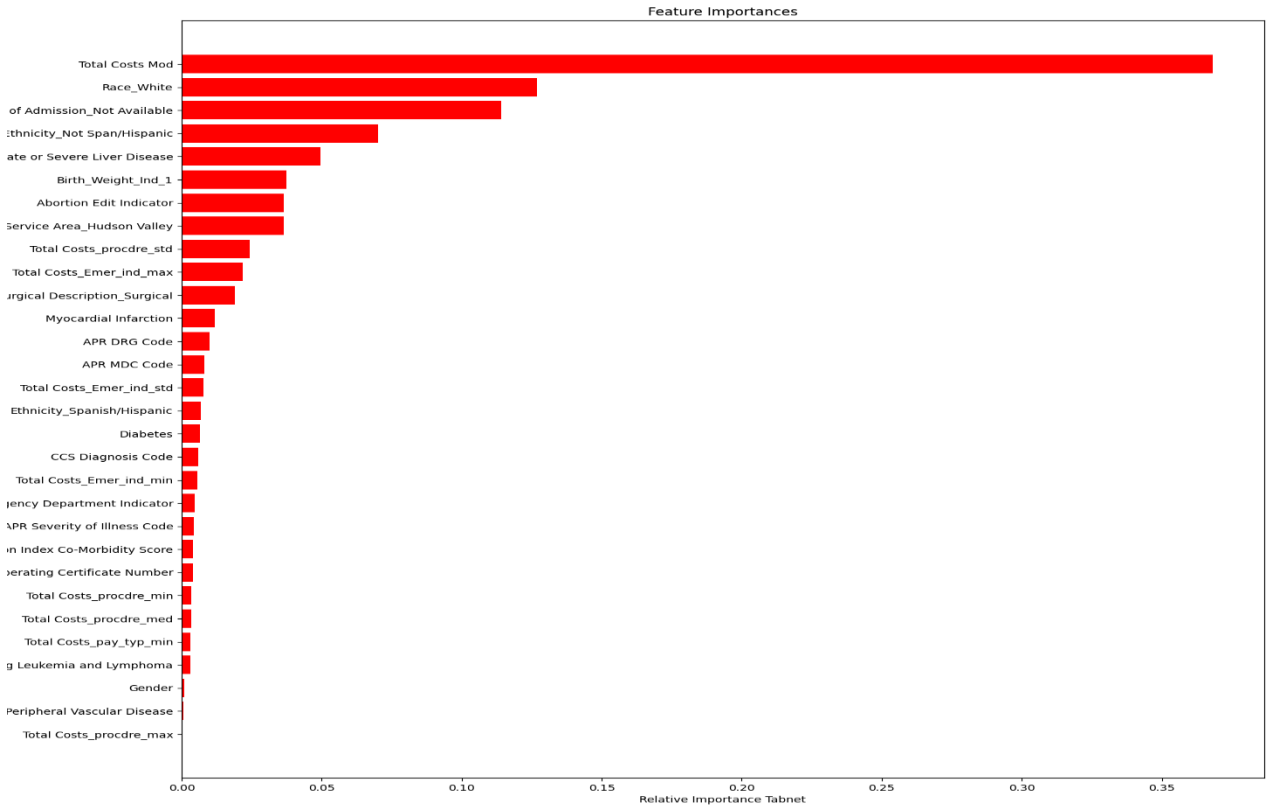


Figure 16: Feature importance of Tabnet model for Top 30 features at the post-admission stage

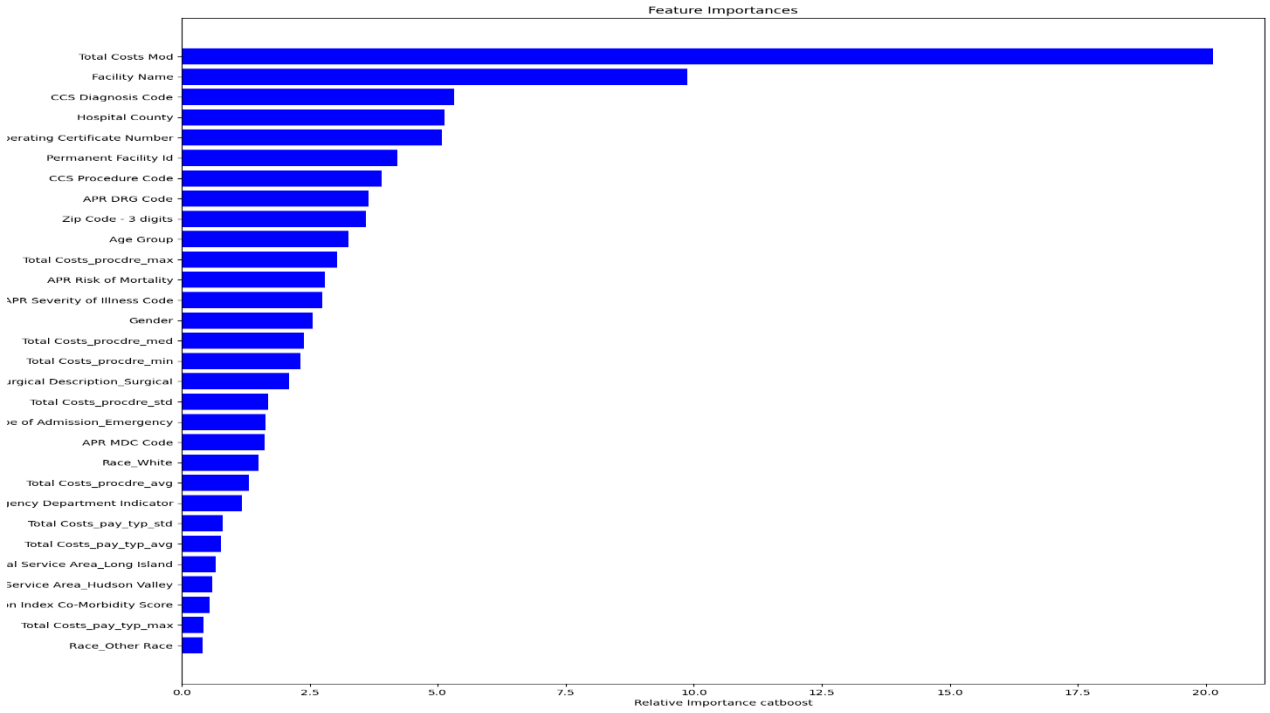


Figure 17: Feature importance of catboost model for Top 30 features at the post admission stage stage

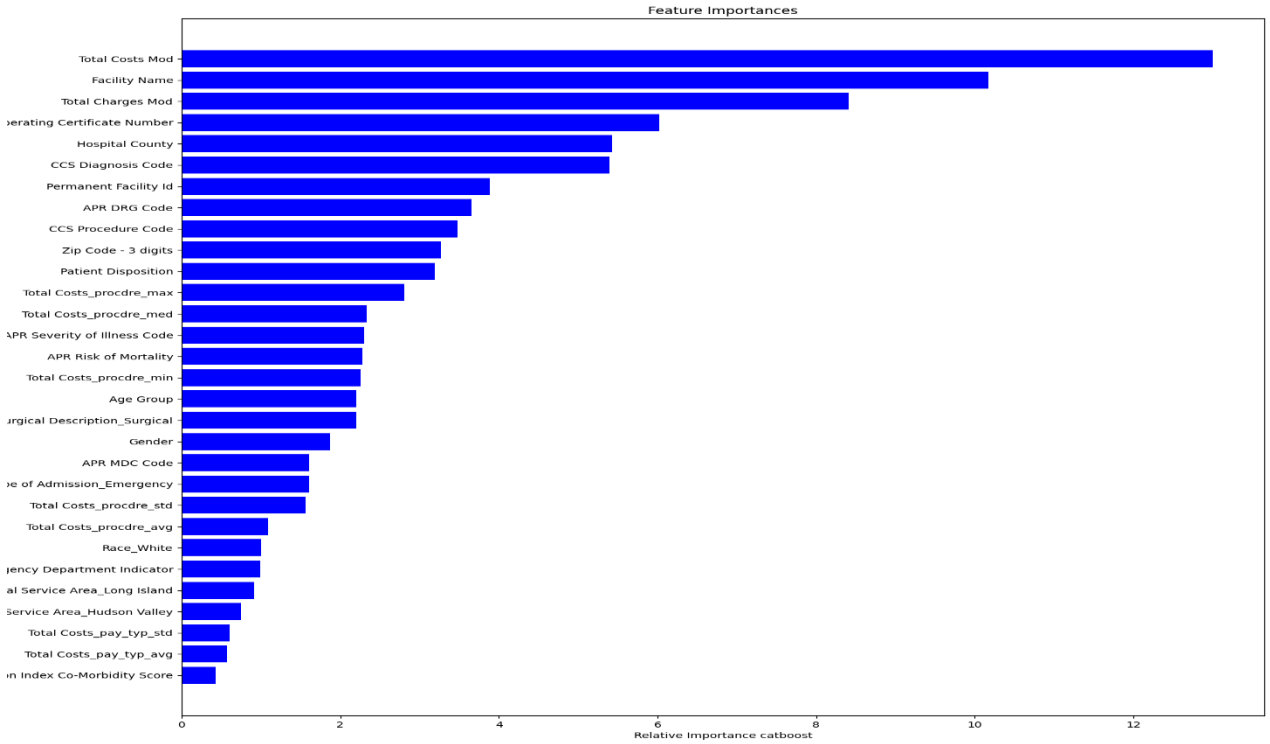


Figure 18: Feature importance of the Catboost model for the Top 30 features at the discharge stage

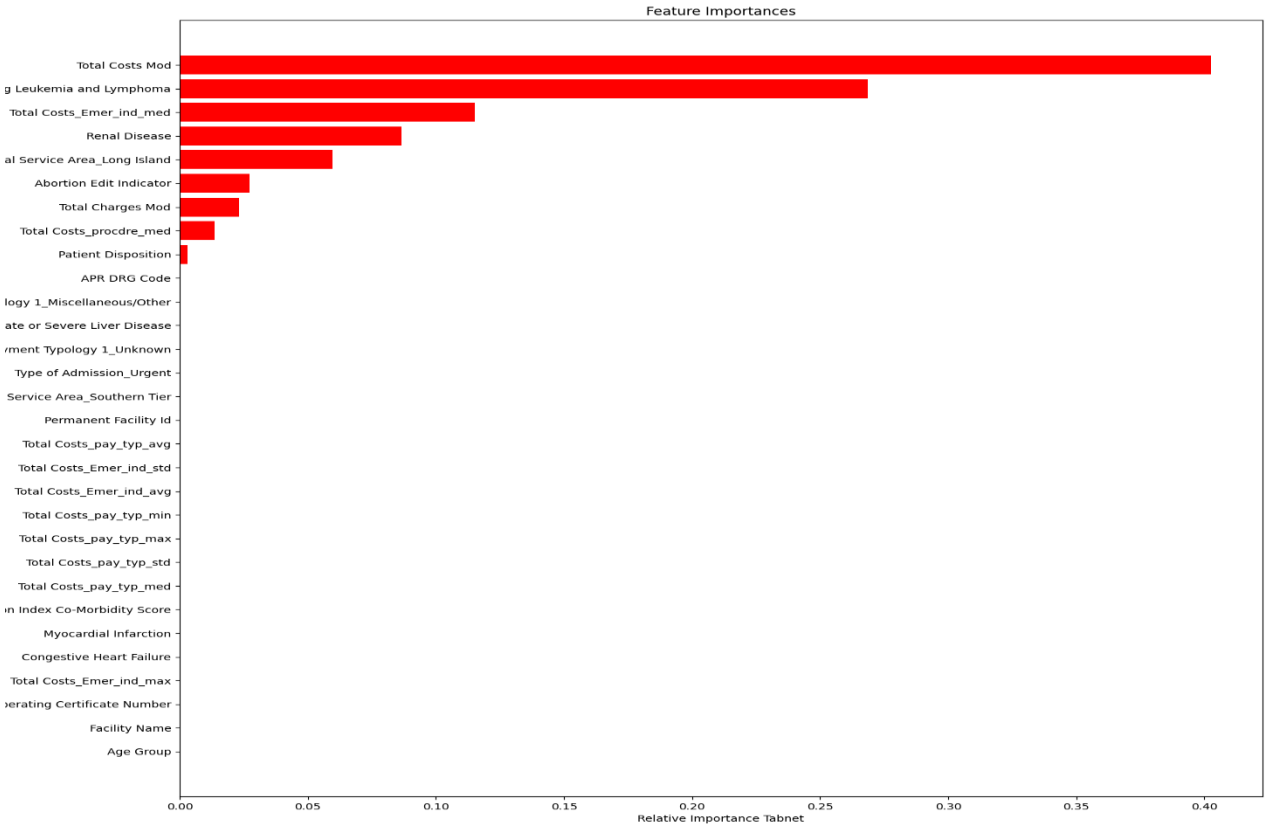


Figure 19:Feature importance of Tabnet model for Top 30 features at the discharge stage

Table 13: Feature importance in admission phase LGBM and CB

Feature	LGM importance	Feature	CB importance
CCS Diagnosis Code	24598	CCS Diagnosis Code	17.67000309
Facility Name	16046	Facility Name	9.976749463
Zip Code - 3 digits	15639	Age Group	9.102938936
Permanent Facility Id	12988	Zip Code - 3 digits	7.983373119
Operating Certificate Number	12406	Gender	6.039496667
Age Group	8797	Permanent Facility Id	5.577477756
Hospital County	6868	Emergency Department Indicator	5.224945586
Gender	4336	Operating Certificate Number	5.107589537
Emergency Department Indicator	3688	Hospital County	4.95734027
Type of Admission_Emergency	3341	Type of Admission_Emergency	3.726950613
Race_White	3326	Payment Typology 1_Medicare	3.631046948
Payment Typology 1_Medicare	3117	Race_White	3.560966014
Payment Typology 1_Medicaid	2792	Payment Typology 1_Medicaid	2.287727832
Race_Other Race	2749	Race_Other Race	2.237466082
Type of Admission_Urgent	2575	Ethnicity_Not Span/Hispanic	1.784015912
Charlson Index Co-Morbidity Score	2463	Charlson Index Co-Morbidity Score	1.548102457
Ethnicity_Not Span/Hispanic	2306	Type of Admission_Urgent	1.499929903

Payment Typology 1_Private Health Insurance	1975	Payment Typology 1_Private Health Insurance	1.045321614
Ethnicity_Spanish/Hispanic	1915	Ethnicity_Spanish/Hispanic	0.83415747
Ethnicity_Unknown	1448	Hospital Service Area_Hudson Valley	0.700407217
Payment Typology 1_Self-Pay	977	Cerebrovascular Disease	0.571338938
Type of Admission_Newborn	905	Hospital Service Area_Long Island	0.516405592
Any Malignancy, Including Leukemia and Lymphoma	882	Ethnicity_Unknown	0.476848697
Cerebrovascular Disease	851	Hospital Service Area_Central NY	0.397189856
Myocardial Infarction	800	Type of Admission_Newborn	0.380492594
Payment Typology 1_Miscellaneous/Other	799	Hospital Service Area_Finger Lakes	0.34428287
Hospital Service Area_Long Island	737	Myocardial Infarction	0.279100203
Hospital Service Area_Hudson Valley	697	Payment Typology 1_Self-Pay	0.275276694
Hospital Service Area_Finger Lakes	667	Any Malignancy, Including Leukemia and Lymphoma	0.253804757

Table 14: Feature importance in the admission phase for the XGBoost model

Feature	XG importance
CCS Diagnosis Code	0.048768092
Payment Typology 1_Medicare	0.04365532
Cerebrovascular Disease	0.043428816
Type of Admission_Newborn	0.035341095
Age Group	0.034541454
Type of Admission_Emergency	0.031819865
Type of Admission_Urgent	0.025774188
DHemiplegia or Paraplegia	0.025715923
Emergency Department Indicator	0.025222303
Charlson Index Co-Morbidity Score	0.024144884
Hospital Service Area_Long Island	0.022739628
Hospital Service Area_Hudson Valley	0.021789126
Hospital County	0.021178069
Hospital Service Area_Western NY	0.021020668
Hospital Service Area_New York City	0.02084331
Facility Name	0.020533856
Permanent Facility Id	0.020017838
Hospital Service Area_Finger Lakes	0.019962842
Myocardial Infarction	0.019728705
Operating Certificate Number	0.019568942
Hospital Service Area_Central NY	0.018082261
Hospital Service Area_Southern Tier	0.01720745
Any Malignancy, Including Leukemia and Lymphoma	0.016670361
Payment Typology 1_Medicaid	0.016214702
Payment Typology 1_Self-Pay	0.016159065
Diabetes with Chronic Complications	0.016072856
Mild Liver Disease	0.014975576
Payment Typology 1_Miscellaneous/Other	0.014953505

Ethnicity_Spanish/Hispanic	0.01396388

Table 15: Feature importance in Post admission phase for the Tabnet model and RF model

Features	Tabnet Importance	Feature	RF importance
Total Costs Mod	0.368191085	Total Costs Mod	0.387302783
Race_White	0.126963011	CCS Diagnosis Code	0.061371614
Type of Admission_Not Available	0.114023558	APR Severity of Illness Code	0.055268728
Ethnicity_Not Span/Hispanic	0.070265176	Facility Name	0.051220522
Moderate or Severe Liver Disease	0.049439234	Operating Certificate Number	0.046466542
Birth_Weight_Ind_1	0.037376988	APR DRG Code	0.041889855
Abortion Edit Indicator	0.036461888	Permanent Facility Id	0.038865188
Hospital Service Area_Hudson Valley	0.036352409	APR Medical Surgical Description_Surgical	0.03259545
Total Costs_procdre_std	0.024393923	APR Risk of Mortality	0.024177757
Total Costs_Emer_ind_max	0.021767673	Zip Code - 3 digits	0.023933838
APR Medical Surgical Description_Surgical	0.018959584	APR MDC Code	0.022602953
Myocardial Infarction	0.011865412	Total Costs_procdre_med	0.017072088
APR DRG Code	0.010129964	Hospital County	0.015968866
APR MDC Code	0.00823176	CCS Procedure Code	0.01587119
Total Costs_Emer_ind_std	0.007668617	Total Costs_procdre_std	0.014952603
Ethnicity_Spanish/Hispanic	0.006824362	Total Costs_procdre_max	0.013168074
Diabetes	0.006700745	Total Costs_procdre_min	0.013066788
CCS Diagnosis Code	0.006070518	Age Group	0.012814843
Total Costs_Emer_ind_min	0.005682549	Total Costs_procdre_avg	0.011187331
Emergency Department Indicator	0.004829449	Type of Admission_Emergency	0.007914012
APR Severity of Illness Code	0.004218281	Gender	0.00573762
Charlson Index Co-Morbidity Score	0.004184136	Total Costs_pay_typ_max	0.004825228
Operating Certificate Number	0.004178459	Race_White	0.004115308
Total Costs_procdre_min	0.003548284	Charlson Index Co-Morbidity Score	0.003968384
Total Costs_procdre_med	0.003278657	Total Costs_pay_typ_std	0.003909751
Total Costs_pay_typ_min	0.003222435	Total Costs_pay_typ_med	0.003490667
Any Malignancy, Including Leukemia and Lymphoma	0.003046731	Total Costs_pay_typ_avg	0.003414292
Gender	0.000898365	Total Costs_Emer_ind_min	0.003378774
Peripheral Vascular Disease	0.000687878	Hospital Service Area_Hudson Valley	0.003229956

Table 16: Feature importance in Post admission phase for Catboost and XGBoost model

Feature	CB importance	Feature	XG importance
Total Costs Mod	20.14031341	APR Medical Surgical Description_Surgical	0.14002115

Facility Name	9.880467237	Total Costs Mod	0.06805395
CCS Diagnosis Code	5.317387816	Total Costs_Emer_ind_min	0.055375516
Hospital County	5.136422906	APR Severity of Illness Code	0.045339834
Operating Certificate Number	5.087312786	Myocardial Infarction	0.03604824
Permanent Facility Id	4.208871425	Type of Admission_Emergency	0.03321989
CCS Procedure Code	3.90187058	Emergency Department Indicator	0.028291674
APR DRG Code	3.655207297	Hospital Service Area_New York City	0.026733482
Zip Code - 3 digits	3.603341842	Hospital Service Area_Hudson Valley	0.025443025
Age Group	3.256683953	APR MDC Code	0.02403143
Total Costs_procdre_max	3.038210761	Operating Certificate Number	0.021839606
APR Risk of Mortality	2.789156339	Birth_Weight_Ind_1	0.021009175
APR Severity of Illness Code	2.751510454	CCS Diagnosis Code	0.02022989
Gender	2.560914978	APR Risk of Mortality	0.019948872
Total Costs_procdre_med	2.379124158	Hospital County	0.018117072
Total Costs_procdre_min	2.325376074	Type of Admission_Newborn	0.017505586
APR Medical Surgical Description_Surgical	2.100989509	Facility Name	0.015643863
Total Costs_procdre_std	1.696666735	Permanent Facility Id	0.015304687
Type of Admission_Emergency	1.629017498	APR DRG Code	0.014918247
APR MDC Code	1.622482558	Total Costs_Emer_ind_max	0.013994264
Race_White	1.49555923	Hospital Service Area_Finger Lakes	0.013976391
Total Costs_procdre_avg	1.315468078	Total Costs_procdre_med	0.013037579
Emergency Department Indicator	1.183989795	Hospital Service Area_Central NY	0.012037359
Total Costs_pay_typ_std	0.796187949	Hospital Service Area_Western NY	0.01181156
Total Costs_pay_typ_avg	0.764690173	Hospital Service Area_Long Island	0.011540774
Hospital Service Area_Long Island	0.663735871	Total Costs_procdre_std	0.011165927
Hospital Service Area_Hudson Valley	0.600997263	Type of Admission_Urgent	0.010818395
Charlson Index Co-Morbidity Score	0.553670817	Total Costs_procdre_min	0.010723388
Total Costs_pay_typ_max	0.421564699	Total Costs_procdre_avg	0.010389408

Table 17: Feature importance in the Discharge phase for RF and Xgboost

Features	RF Importance	Features	XG Importance
Total Costs Mod	0.323758584	APR Medical Surgical Description_Surgical	0.14797086
Total Charges Mod	0.128589523	Total Costs Mod	0.06614875
CCS Diagnosis Code	0.057534799	Total Costs_Emer_ind_max	0.06567397
APR Severity of Illness Code	0.043086297	APR Severity of Illness Code	0.04017929
Facility Name	0.042308394	Hospital Service Area_New York City	0.029262332
Operating Certificate Number	0.041758118	Type of Admission_Emergency	0.029049894
APR DRG Code	0.034090543	Myocardial Infarction	0.028838575
APR Medical Surgical Description_Surgical	0.030954	Hospital Service Area_Hudson Valley	0.027013205
Permanent Facility Id	0.029611608	Emergency Department Indicator	0.026935887
Zip Code - 3 digits	0.022708124	CCS Diagnosis Code	0.022358947
APR MDC Code	0.021474216	Hospital Service Area_Long Island	0.021442367

Patient Disposition	0.021395404	Birth_Weight_Ind_1	0.020747086
APR Risk of Mortality	0.02087106	APR Risk of Mortality	0.020605087
Hospital County	0.015758532	Type of Admission_Newborn	0.018606942
Total Costs_procdre_med	0.014918165	Operating Certificate Number	0.01842364
CCS Procedure Code	0.013469721	Total Costs_Emer_ind_min	0.01836998
Total Costs_procdre_std	0.01278133	Total Charges Mod	0.01562042
Total Costs_procdre_min	0.011581905	APR MDC Code	0.015569372
Total Costs_procdre_max	0.010823199	Permanent Facility Id	0.015348256
Total Costs_procdre_avg	0.009620224	Hospital Service Area_Finger Lakes	0.015113214
Type of Admission_Emergency	0.008124797	Hospital County	0.014821176
Age Group	0.007870494	Facility Name	0.014662114
Total Costs_pay_typ_max	0.003792934	Expired_Ind	0.013611542
Gender	0.003726256	Hospital Service Area_Western NY	0.012841681
Total Costs_Emer_ind_avg	0.003550345	Patient Disposition	0.012231928
Hospital Service Area_Hudson Valley	0.003479279	APR DRG Code	0.011818585
Charlson Index Co-Morbidity Score	0.003213358	Total Costs_procdre_med	0.010830347
Total Costs_Emer_ind_med	0.002956067	Hospital Service Area_Central NY	0.010365255
Total Costs_pay_typ_std	0.002855534	Cerebrovascular Disease	0.009501406
Total Costs_Emer_ind_max	0.002790549	Type of Admission_Urgent	0.009498738

Table 18: Feature importance in the Discharge phase for tabnet and Catboost

Feature	Tabnet importance	Feature	CB importance
Total Costs Mod	0.402565503	Total Costs Mod	13.00028935
Any Malignancy, Including Leukemia and Lymphoma	0.268687012	Facility Name	10.17000623
Total Costs_Emer_ind_med	0.115310738	Total Charges Mod	8.407059381
Renal Disease	0.086595401	Operating Certificate Number	6.021144902
Hospital Service Area_Long Island	0.059558811	Hospital County	5.423035584
Abortion Edit Indicator	0.027116158	CCS Diagnosis Code	5.396714389
Total Charges Mod	0.02327459	Permanent Facility Id	3.887345918
Total Costs_procdre_med	0.013661181	APR DRG Code	3.652444248
Patient Disposition	0.003127742	CCS Procedure Code	3.476477329
APR DRG Code	6.27062E-05	Zip Code - 3 digits	3.274394571
Payment Typology 1_Miscellaneous/Other	2.9191E-05	Patient Disposition	3.191681461
Moderate or Severe Liver Disease	9.2607E-06	Total Costs_procdre_max	2.805707624
Payment Typology 1_Unknown	1.59322E-06	Total Costs_procdre_med	2.331233027
Type of Admission_Urgent	7.97431E-08	APR Severity of Illness Code	2.297197792
Hospital Service Area_Southern Tier	2.18767E-08	APR Risk of Mortality	2.277901921
Permanent Facility Id	1.02658E-08	Total Costs_procdre_min	2.255078969
Total Costs_pay_typ_avg	1.32757E-10	Age Group	2.206246827
		APR Medical Surgical Description_Surgical	2.20372942

Gender	1.871988972
APR MDC Code	1.603915764
Type of Admission_Emergency	1.603544087
Total Costs_procdre_std	1.566586051
Total Costs_procdre_avg	1.090432911
Race_White	1.001694167
Emergency Department Indicator	0.994308838
Hospital Service Area_Long Island	0.909187326
Hospital Service Area_Hudson Valley	0.74926884
Total Costs_pay_typ_std	0.609468066
Total Costs_pay_typ_avg	0.56908991

4.6 Discussion

The length of stay prediction is one of the important factors for hospitalization. The length of stay is directly interconnected to the hospital expenses and the patient's well-being. Identifying the factors leading to correct hospital length of stay prediction is a research topic. This study uses a unique approach to predict hospital length of stay and identify the factors associated with length of stay. This prediction is made using a three-stage model: admission stage, post-admission stage, and discharge stage. The performance of the stage-wise models was improvised from the admission to the discharge phase. The accuracy, AUC, precision, recall, and f1 scores obtained in the admission phase through the 3WPE model are 0.6995, 0.8087, 0.44, 0.26, and 0.27, respectively, an improved version of the base models. The performance in the length of stay in this phase is lacking due to missing essential features. This stage can classify the short stays between 0-5 days, but beyond five days, this admission phase model can not predict the length of stay. The performance of the model beyond five days is the worst. This is due to missing essential factors for length of stay prediction. This model's performance improved compared to the prior methodology, but for the admission phase, the improvement is not that significant. The accuracy, AUC, precision, recall, and f1 score obtained in the admission phase through the 3WPE model are 0.9178, 0.9888, 0.82, 0.78, and 0.8, respectively, which is quite improved from the admission phase. The performance of the ensemble model is quite good in this phase as the features included in this phase have significant importance in length of stay prediction. This post-admission model can predict the Length of stay beyond 5 days quite well compared to the admission phase. The issue of class 3 and class 6 misclassification improved. The accuracy, AUC, precision, recall, and f1 scores obtained in the discharge stage through the 3WPE model are 0.9320, 0.992, 0.85, 0.82, and 0.84, respectively, which again improved the version compared to the post-admission phase.

Features selected for this phase have a significant impact on Los prediction. The behavior of the discharge stage model is quite similar to that of the post-admission phase. The issue of class 3 and class 6 improved, and we got an improved classifier that can classify both short and long stays well. The 3WPE model has outperformed its base ML models in each phase. The admission phase has limited features related to demographics and diagnosis. The performance improved in the post-admission phase due to the availability of essential features like estimated cost, procedures performed, severity score, etc. The discharge phase performed best in this

three-stage predictive model. Features like actual hospitalization cost, patient disposition, and expired Ind has been included in this phase.

This methodology clubbed misclassified different length of stay groups into one group, which helps improve the classifier's performance. Deep learning models like Tabnet and CNN 1D significantly impact prediction, improving the model's overall performance. The tabnet model has very good accuracy as well, and other factors like precision, recall, and f1-score of the tabnet are higher among all the base models' performance matches with the random forest model, though its accuracy is lower than the XGBoost, overall it performed pretty well. The Tabnet model is utilized in this methodology to combine with other models to generate the weighted ensemble model, improving overall performance. The CNN 1D also gives good accuracy, precision, recall, and f1-score.

In the future, we can utilize CNN 1D more in medical data mining work with robust modeling. Identification of proper length of stay groups based on the confusion matrix from the previous ensemble model helps to improve the performance of this model. Combining machine learning and deep learning model outperforms the ensemble model compared to using only machine learning models. This study shows an impactful model for length of stay prediction, capable of three distinct prediction stages from admission to discharge. This is an all-in-one model for length-of-stay prediction with varying sizes of features. This study utilizes multiple machine learning models to create a robust weighted average probability model. Hospital authorities and medical insurance companies can use this new model to balance and manage resources better. This new model is a unique generic model for classifying length of stay, and its performance is good compared to that of the existing models[33,34,10].

Chapter 5

Proposed Methodology of In-hospital Death Prediction

5.1 Methods

The study aims to predict the hospital mortality for the post-admission phase. The post-admission phase is when a patient is admitted into the hospital, and doctors do a thorough check up. Several tests are performed to better analyze the condition of the patient. In this stage, the patient's condition has been appropriately diagnosed, and decisions about the patient's condition and treatment have been made. Predicting mortality in this phase will help the hospital authority, and doctors take extra care of the patients with a higher chance of death. The processes to achieve the mortality prediction consist of a few steps: an introduction to the dataset, data pre-processing, and the proposed model, which will be discussed in the following sections.

5.1.1 Data repository

In 1979, the government and the healthcare sector collaborated to create SPARCS, a Comprehensive all-payer data reporting system. The New York State Department of Health contributed to the Hospital Inpatient Discharges (SPARCS De-Identified) 2017 dataset[55], an open dataset with de-identified data; the file is not individually identifiable. It contains essential record-level detail. This repository includes 2.3 Million records with 34 features. They feature patient demographic data, diagnosis data, hospital-level data, and expenses. The patient's lab test-related parameters and private information are excluded and are not part of our prediction. The dataset contains several hospitals in the US based on different regions. The dataset is the repository of 2017 with different diseases reported in the US. This dataset includes all patients visited both in emergency and general admissions. The summary of each feature of Sparcs data is described in Table 1 of Chapter 3. This analysis also uses the exact data for length of stay prediction. This is an extension of the work done in chapters 3 and 4; this study predicts the death that occurred in the hospital with respect to it’s severity based on Los.

The target feature for this prediction is death. Death happens in a hospital any time between the day of admission and the number of days. This dataset consists of a maximum length of stay of 120. Death can vary between 1 to 120 days. This study predicts the death of inpatients considering the patients' length of stay. A patient can die immediately after hospitalization or after a prolonged stay. Based on this understanding of the relationship between death and length of stay, we have divided the death prediction into two parts. The first one is the death that happened within seven days of hospitalization, and the second one is the death that occurred after seven days of hospitalization. This seven-day death model can help identify potential death groups based on survival days. If the patients died within a seven7-day period, then the criticality of the patient will be considered as extreme compared to the 7+ group. Accordingly, the care should be taken for the critical patients. Based on the criticality, the 7+ death group

will get less priority than the 7-day group. The least priority will be given to the surviving group; the resource arrangement and assignment by the hospital will follow this model. These seven-day and post-seven-day in-hospital mortality prediction models will benefit the hospital authority in predicting death for the survival days, that is, the length of stay. This is a new approach to in-hospital death prediction. This study does not remove any entry from the dataset. All records have been considered for the prediction. The prediction includes all age groups, diseases, and hospitals for the mortality prediction. Most studies utilized the adult group for death prediction (age>18 years) and a specific disease. This study refers a data which has diverse category that predicts death and tries to discover the factors contributing to death prediction. The prediction is a multiclass problem with identified groups being '7daydeath', 7+ daydeath', and 'survival'**.** The dataset has 2.1 % of deaths and 97.9% of survivors.
This study used Charlson's comorbidity factors to predict in-hospital death. The CCS diagnosis code is used to generate the comorbidity factors; the CCS Diagnosis code present in sparcs data is in ICD-9 standard form, and we have converted diagnosis codes into Charlson factors based on ICD-9 conversion logic[23]. These 17 Charlson comorbidities factors will be utilized along with the features present in the dataset for prediction. Most studies based on Sparcs data used either the Charlson comorbidity factor alone or mortality factors such as APR DRG Code, Risk of mortality, and Severity of illness group[32,42] alone to predict death. However, this study combines both factors to predict death, resulting in a robust length of stay-specific death model. This study has around 49000(2%) death samples and 22300000 survivors, which creates a severe data imbalance. To improve the balance, we used SMOTE to make 10% of the overall death percentage, followed by random under-sampling. We have used the chi-square feature selection technique SelctKbest to reduce the feature set with features size of 10/20/30/40/50/all and tested the features with the best-performing model to verify the prediction performance based on small feature sets. The data distribution and the groups of death and survival have been stated in Table 19.

Table 19: Data distribution of death and survival patients

Group (In days)	Patient count
7daysdeath	28776
7+daysdeath	20230
Survived	2258975

5.1.2 Data Analysis and Visualization

The Sparcs data contains patient demographics, hospital diagnosis level, and hospital expense-related data. The data is de-identified, so the patient's personal information and parameters to identify patients have been removed. This data consists of essential details from the record level. The data does not contain inpatients' comorbidity or any basic vital signs. No lab parameters are also provided in the information. The dataset comprises Patient disposition, which is the final disposition after hospital; the valid values are Home or Self Care, Home w/ Home Health Services, Skilled Nursing Home, Expired, etc. The Expired disposition is utilized to identify the death that happened in the hospital. The expired Ind is the target variable here. The target with three classes, 7-day death, 7+ day death, and survived, has been created based on combining the 'Expired' disposition and length of stay. The study focuses on hospital death prediction for 7 and 7+ days. The detailed data analysis of dead patients with respect to their level has been represented in Table 20. The categorical column data distribution is stated in percentage, and the numerical column is expressed in its min, max, median, and mean values.

The study shows that the New York City hospital area has the highest number of deaths, at 41.9 %, and The Manhattan hospital has the highest death rate, at 12.18 %.

Table 20: Analysis of Sparcs features for death patients

Features	Category	Percentage
Health Service Area	New York City	41.9%
	Long Island	15.05%
	Hudson Valley	10.36%
	Western NY	8.72%
	Capital/Android	8.14%
	Others	15.83%
Hospital County	Manhattan	12.18%
	Kings	9.98%
	Bronx	9.96%
	Nassau	8.27%
	Queens	7.13%
	Others	52.48%
Facility Name	Calvary Hospital Inc	4.55%
	Strong Memorial Hospital	2.4%
	New York Presbyterian Hospital - Columbia Presbyterian Center	2.24%
	North Shore University Hospital	2.23%
	Albany Medical Center Hospital	2.08%
	Others	86.5%
Age Group	70 or Older	64.15%
	50 to 69	27.73%
	30 to 49	5.1%
	0 to 17	1.86%
	18 to 29	1.09%

Zip Code - 3 digits	112 117 104 100 113 others	12.19% 7.64% 7.16% 5.98% 4.93% 62.1%
Gender	M F	50.6% 49.4%
Race	White Other Race Black/African American Multi-racial	64.59% 19.1% 15.34% 0.91%
Ethnicity	Not Span/Hispanic Spanish/Hispanic Unknown Multi-ethnic	86.74% 7.38% 5.62% 0.23%
Length of Stay	Mean Median Min Max	10.58 6.0 1 120
Type of Admission	Emergency Elective Urgent Newborn Trauma Not Available	81.42% 8.81% 7.68% 1.18% 0.88% 0.03%
CCS Diagnosis Description	Septicemia (except in labor) Acute cerebrovascular disease Respiratory failure; insufficiency; arrest (adult) Hypertension with complications and secondary hypertension Acute myocardial infarction Others	29.04% 6.25% 5.34% 4.15% 3.52% 51.7%
CCS Procedure Description	RESP INTUB/MECH VENTIL NO PROC OT VASC CATH; NOT HEART OTHER THERAPEUTIC PRCS BLOOD TRANSFUSION Others	29.52% 21.63% 5.16% 4.72% 3.97% 35%
APR DRG Description	Septicemia and disseminated infections Infectious parasitic diseases including HIV w O.R. procedure Heart failure Respiratory Failure Moderately extensive procedure unrelated to principal diagnosis	22.52% 5.73% 4.2% 3.63% 3.24% 60.68%

APR MDC Description	Infectious and Parasitic Diseases, Systemic or Unspecified Sites Diseases and Disorders of the Circulatory System Diseases and Disorders of the Respiratory System Diseases and Disorders of the Nervous System Diseases and Disorders of the Digestive System Diseases and Disorders of the Hepatobiliary System and Pancreas Diseases and Disorders of the Kidney and Urinary Tract	29.5% 15.45% 15.42% 10.56% 6.35% 4.1% 3.58% 15.04%
APR Severity of Illness Description	Extreme Major Moderate Minor	59.4% 33.28% 6.41% 0.91%
APR Risk of Mortality	Extreme Major Moderate Minor	60.88% 32.37% 5.84% 0.91%
APR Medical Surgical Description	Medical Surgical	75.72% 24.28%
Payment Typology 1	Medicare Medicaid Private Health Insurance Blue Cross/Blue Shield Miscellaneous/Other Self-Pay	70.63% 11.44% 7.74% 4.54% 2.87% 1.38%
Birth Weight Ind	0 1	98.68% 1.32%
Abortion Edit Indicator	N Y	99.99% 0.01%
Emergency Department Indicator	Y N	76.91% 23.09%
Total Costs	Mean Median Min Max	42398.90 17997.49 0.21 5093860.44
Expired Ind	7 Group 7+ Group	58.72% 41.28%
Charlson Index Co-Morbidity Score	0 1 2 6 3	67.32% 15.27% 11.99% 3.13% 2.28%
Myocardial Infarction	Yes	3.83%
Congestive Heart Failure	Yes	1.08%
Peripheral Vascular Disease	Yes	1.50%
Cerebrovascular Disease	Yes	6.54%

Dementia	Yes	0.42%
Chronic Pulmonary Disease	Yes	1.57%
Rheumatologic Disease	Yes	0.21%
Peptic Ulcer Disease	Yes	0.07%
Mild Liver Disease	Yes	0.93%
Diabetes	Yes	0.00%
Diabetes with Chronic Complications	Yes	0.62%
DHemiplegia or Paraplegia	Yes	0.00%
Renal Disease	Yes	1.99%
Any Malignancy, Including Leukemia and Lymphoma	Yes	9.37%
Moderate or Severe Liver Disease	Yes	2.28%
Metastatic Solid Tumor	Yes	2.86%
AIDS	Yes	0.27%

Calvary Hospital Inc. hospital's death rate is the highest. The data rate is highest in the 70 or Older age group; the percentage is 64.15%. The death rate is lowest in the 18 to 29 age group. Older and adult have a higher death rate compared to young people. The 0 to 17 age group has a higher death rate compared to the 18 to 29 age group. The 0 to 17 age group mainly belongs to the newborn and children. The death rate is higher for male patients, with a percentage of 50.6%. For the Admission type, as Emergency has more deaths, the percentage is 81.42%. Septicemia (except in labor), Acute cerebrovascular disease, Respiratory failure, insufficiency, arrest (adult), Hypertension with complications and secondary hypertension, and Acute myocardial infarction are the primary diagnoses having higher death rates. Septicemia (except in labor) has the highest death rate, with a percentage of 29.04%.

The procedures that have higher death are RESP INTUB/MECH VENTIL, OT VASC CATH, NOT HEART, OTHER THERAPEUTIC PRCS, and BLOOD TRANSFUSION. The patient without having any procedure also has a higher death rate. Primary diagnoses where death rate is higher are Infectious and Parasitic Diseases, Systemic or Unspecified Sites, Diseases and Disorders of the Circulatory System, Diseases and Disorders of the Respiratory System, Diseases and Disorders of the Nervous System, Diseases and Disorders of the Digestive System, Diseases and Disorders of the Hepatobiliary System and Pancreas, Diseases and Disorders of the Kidney and Urinary Tract. Patient with Extreme Severity of illness has higher deaths, around 59.4%, and the death rate decreased as the severity decreased. The patient with an Extreme Risk of mortality has a higher death rate.

The percentage is 60.88%. It got decreased with the severity. A minor risk of mortality results in fewer deaths, around 0.91%. Nonsurgical patients have a higher death rate, which is around 75.72%. The Emergency Department has a higher death rate of 76.91%. The death percentage for Length of stay less than 7 days is 58.72%, and length of stay greater than 7 days has fewer deaths, around 41.28%. The Charlson comorbidity factor with a score '0' has a higher death rate of 67.32%. The Charlson comorbidities, Any Malignancy, Including Leukemia and Lymphoma, has a higher death rate of 9.32%. Second on the list is Cerebrovascular Disease, with a rate of 6.54%, and a third is Myocardial Infarction, with a rate of 3.83%. In contrast, Diabetes and Hemiplegia or Paraplegia have a death rate of 0%.

5.1.3 Data Pre-processing

The raw data we receive is usually incomplete and inconsistent. Those data cannot be directly fed to machine learning models, and it will produce unsatisfactory results; to improve the data quality, we need to clean the data and perform a few processing or transformations to improve the data quality. Data pre-processing generally improves the data quality for ML models. It also removed impure factors present in the data. The raw data is processed to get clean data that can be interpretable by the machine. Our data has some inconsistencies: the length of stay has values ‘120+’, which we have modified to make it 120. Some features include junk data like Gender has the value ‘unknown’, APR Medical Surgical Description' has a value ‘Not Applicable,' and birth weight has the value ‘unknown’; we have removed these entries from the dataset. The essential data pre-processing will be discussed in the later sections:

5.1.3.1 Missing Values handling

This study also refers to the Sparcs dataset as the length of stay prediction. The missing value handling step is the same as the length of stay prediction. We have not done something new in

this study. For details of disappeared values imputation, please refer to Chapter 3, missing value handling description.

5.1.3.2 Feature Engineering

Feature engineering is one of the essential tasks in machine learning modeling. New features extracted from the existing features can increase the predictive power of machine learning models. We have created a few new features from the existing 34 features. Our feature set has a CCS (Clinical Classification code) diagnosis code that gives information about the patient's primary diagnosis. The diagnosis codes in sparcs data are ICD-9 standard codes. Based on this CCS diagnosis code [23], we have generated Comorbidity [23] related features utilizing the Charlson comorbidity index. It generates 18 more features with different comorbidities and the respective comorbidity score.

Patient disposition is the final destination of patients after release from the hospital. We have created one feature, Expired Ind, utilizing patient disposition 'Expired.' It informs that the patient has died in the hospital. This Expired Ind will be merged along with the length of stay to generate a new feature called day expired, which will have three values: '7Yes', '7+Yes', and 'No'. '7Yes' denotes that the death happened within seven days of hospitalization, death happened in the hospital after 7 days of hospitalization, and the patient survived in the hospital, respectively. This day's expiration date is the target variable for our death prediction. Expired Ind is no longer required as the day expired conveys similar information with length of stay as a factor. The birth weight feature is explicitly for the newborn. We have created a new feature, birth weight indicator, that utilizes birth weight. This field holds binary values '1' or '0' based on the availability of birth weight. The actual Birth weight field is no longer required, so it has been dropped.

5.1.3.3 Feature Selection

This step is unaltered compared to the feature selection steps used in the length of stay prediction; for reference, please visit Chapter 3, feature selection step. The selected top-performing model will be evaluated based on a subset of features. The feature subset will be chosen based on the Chi-square feature selection method utilizing the Select K best function. Features 10/20/30/40/50/all will be selected using chi-square feature selection based on decreasing order of importance. The subsets will be trained and tested with the top-performing model to verify the model's performance with the reduced feature dimension.

5.1.3.4 Feature Normalization and transformation

This step is unaltered compared to the feature selection steps used in the length of stay prediction. For reference, please visit Chapter 3, which features the normalization and transformation steps.

5.1.3.5 Feature Encoding

This step is unaltered compared to the feature selection steps used in the length of stay prediction. For reference, please visit Chapter 3, feature encoding step.

5.1.3.6 Data balancing

Data balancing is an essential step for classification. Imbalanced data can heavily affect the performance of the classifier. The minority class's performance will be worse. The classifier will not be able to classify the minor class; The major classes will get highly classified, whereas the minor will have the least accuracy. Our dataset is highly imbalanced; it has 2.1% death samples, and 97.9% survived samples. The performance of the classifier using original data is terrible. It classifies the primary class('Survived') well but cannot correctly classify minor classes(Death instances). The technique of improving the classifier's performance by addressing the class imbalance problem is called sampling. Sampling can be of two types: undersampling and oversampling. Undersampling reduces the primary class to make them equal to the minority class, and over-sampling increases the count of minor classes. We performed both SMOTE and random undersampling in this approach to improve the classifier performance. Many studies use SMOTE, a random under-sampling technique, to address the class imbalance problem.

SMOTE stands for Synthetic Minority Over-sampling Technique. It is an oversampling technique used to balance the class distribution of a dataset by creating synthetic minority class samples. SMOTE is a data augmentation technique that generates new synthetic samples by interpolating between existing minority-class samples.SMOTE creates synthetic samples by joining the nearest neighbors in the feature space. The basic idea behind SMOTE is to create new minority class samples by taking small steps from one of the minority class samples to one of its k nearest neighbors in the feature space, where k is a parameter of the Algorithm. The Algorithm creates a new sample by randomly selecting one of the k nearest neighbors and then adding a small perturbation to the feature vector between the original sample and the selected neighbor. This creates new synthetic data that is similar to the minority class samples in the feature space but is not an exact copy of any of the existing samples.
The SMOTE algorithm works in the following way:

1. Select a minority class sample from the original dataset.
2. Find its k nearest minority class neighbors in the feature space.
3. Randomly select one of the k nearest neighbors.
4. Generate a new synthetic sample by interpolating between the selected minority class sample and the randomly selected neighbor.
5. Repeat steps 1-4 until the desired number of synthetic samples is generated.

The k parameter in SMOTE determines the nearest neighbors to consider when generating synthetic samples. Typically, the value of k is set to 5, but it can be adjusted depending on the size of the dataset and the degree of imbalance.
K could be a hyperparameter to optimize using cross-validation. However, SMOTE relies on the k nearest neighbors algorithm, which does not scale well, and therefore, optimizing this value could be quite time- and resource-consuming.[50].

Random sampling, the most straightforward under-sampling technique, involves randomly selecting examples from the majority class and deleting them from the training dataset. This is referred to as random undersampling. Although simple and effective, a limitation of this technique is that examples are removed without any concern for how useful or necessary they might be in determining the decision boundary between the classes. This means it is possible, or even likely, that valuable information will be deleted. The major drawback of random undersampling is that this method can discard potentially useful data that could be important for the induction process. Data removal is a critical decision; hence, many proposals for undersampling use heuristics to overcome the limitations of the non-heuristics decisions.[53].

In this prediction, we have made the minority distribution 10% and random under-sampling as 'major.'The 10% oversampling was performed using the SMOTE technique, and random undersampling was performed to reduce the majority of samples. Several other percentages of oversampling were performed using the distribution of 20% and 50%, but the performance of the classifier towards the minority, that is, death, was not good. So, we have opted for 10% SMOTE with random under sampling for improved accuracy. This will be used throughout our analysis. The model shows a drift in performance when SMOTE has more than 10% sample. We have not decreased the oversampling below 10%. Training the model with less than 10% minor samples will be unrealistic. Without the under-sampling technique, the model accuracy in identifying minor classes was terrible. The random under-sampling increases the accuracy of minor classes. So, we have implemented SMOTE and random under-sampling together for our analysis. Another set of analyses will also be carried out using the original data. The performance of the base classifiers using original data is not good. We have used a few classifiers, which work well on unbalanced data. These classifiers sample the data first and then perform the training. The sampling technique was integrated into those model's Algorithms themselves, so there was no need for explicit sampling for those models. Those models can be treated as the alternative to the Sampling techniques. A few models are Rusboost, Easy Ensemble, and balanced random forest Classifiers.

5.1.4 Statistical relationship between features

We have 2343569 instances with 34 features in our dataset. We have performed feature engineering to include more features to give more predictive power to the model. We have included 18 Charlson comorbidity-related features. After pre-processing, feature selection, and engineering, 41 features are left with us. We have performed statistical Pearson correlation coefficient tests between target and independent features, and the detailed report is published in Table 21. We have observed that CCS Diagnosis Code and Birth_Weight_Ind have very high positive correlations with death, 0.1252 and 0.036, respectively.

Most features negatively correlate with the target variable death for death prediction. APR Risk of Mortality, APR Severity of Illness Code, Age Group, Total Costs, Emergency Department Indicator, and Charlson Index Comorbidity Score have high negative correlations with death. The coefficients are -0.2791, -0.2295, -0.107, -0.05, -0.05, -0.049 respectively**.** Comorbidity features Any Malignancy, Including Leukemia and Lymphoma, Cerebrovascular Disease, and Metastatic Solid Tumors, which have a high correlation with death, whereas Diabetes, Peptic Ulcer Disease, and Dementia have a shallow impact on death prediction. The Diabetes has p values < 0.05, which is insignificant for death analysis as the dataset contains many categorical and few numerical fields. In this analysis, Total cost is the only numeric field considered; the distribution of total cost is highly skewed, as shown in the figure in Chapter 3 data analysis. The chi-square test is ideal for categorical variables, so we have also performed a Chi-Square statistical test on the categorical features to re-verify their importance and a Kendall Tau statistical test on the numerical features. The detailed analysis is shown in Table 21.

The chi-square test has identified CCS Diagnosis Code, APR Risk of Mortality, CCS Procedure Code, APR Severity of Illness Code, Permanent Facility ID, Operating Certificate Number, Age Group, Charlson Index Comorbidity Score, APR DRG Code as the prime factors related to the death prediction. The statistics showed that the Risk of mortality and severity of illness have high priority compared to the Charlson comorbidity score, but Charlson comorbidity factors significantly impact the death analysis. The comorbidity factors of Any Malignancy, Including Leukemia and Lymphoma, Metastatic Solid Tumor, and Cerebrovascular Disease, have a higher correlation with death. Diabetes and peptic Ulcer Disease have the lowest impact on death. Diabetes has a p-value of less than 0.05. Both the statistics identified that Diabetes has the lowest impact on death analysis. Kendall Tau shows that total cost negatively correlates with death, with a coefficient of -0.064. All these factors will be used for death prediction, and a reduced feature set selected using Chi-square statistics will be utilized for death prediction. Unlike length of stay analysis, the Total estimated cost has less impact on death, whereas medical details have a higher effect on death prediction.

Table 21: Features correlation with death

Feature	Pearson Correlation	p-value	Chi-statistics/ Kendall tau	p-value
CCS Diagnosis Code	0.125266769	0	1477403.171	0
Birth_Weight_Ind	0.036114181	0	3117.984496	0
Operating Certificate Number	0.017739875	4.99E-160	26480.79757	0
Permanent Facility Id	0.016773431	2.80E-143	27740.50628	0
APR Medical Surgical Description	0.015244376	1.08E-118	1888.804769	0
Ethnicity	0.014855543	8.20E-113	132.6070763	1.60228E-29
Diabetes with Chronic Complications	0.01272174	3.04E-83	377.2897196	1.18189E-82
Facility Name	0.009315227	1.78E-45	9127.037712	0
APR DRG Code	0.00857038	9.23E-39	9391.309792	0
Rheumatologic Disease	0.00648567	6.58E-23	97.79965149	5.79529E-22
Hospital County	0.00577332	1.76E-18	2131.726522	0

APR MDC Code	0.005753029	2.31E-18	337.1380711	6.18583E-74
Hospital Service Area	0.005635592	1.11E-17	152.3658782	8.20673E-34
Abortion Edit Indicator	0.005092386	1.02E-14	65.94941122	4.77823E-15
Chronic Pulmonary Disease	0.004716248	7.74E-13	50.8362437	9.14218E-12
DHemiplegia or Paraplegia	0.002851552	1.47E-05	20.4046968	3.70831E-05
Peptic Ulcer Disease	0.001712023	0.009291357	13.81200658	0.001001754
Diabetes	0.001021457	0.12067689	3.197935146	0.202105069
Dementia	-0.00187949	0.004295787	22.91551768	1.05672E-05
Renal Disease	-0.004994664	3.23E-14	56.67525623	4.93315E-13
AIDS	-0.005066767	1.38E-14	171.5786177	5.52302E-38
Type of Admission	-0.005564292	2.81E-17	89.3631112	3.93593E-20
Moderate or Severe Liver Disease	-0.008370454	4.71E-37	168.8315039	2.18125E-37
Peripheral Vascular Disease	-0.010648665	7.05E-59	263.303176	6.67485E-58
Zip Code - 3 digits	-0.011275166	8.66E-66	5701.077256	0
Myocardial Infarction	-0.011492496	2.81E-68	376.2052292	2.0327E-82
Congestive Heart Failure	-0.011718284	6.50E-71	317.3364298	1.23387E-69
Mild Liver Disease	-0.012254	2.27E-77	524.898129	1.0467E-114
Gender	-0.014884212	3.06E-113	432.4050417	1.27186E-94
Race	-0.019936269	1.42E-201	567.4430085	6.0439E-124
CCS Procedure Code	-0.029739202	0	163020.0321	0
Payment Typology 1	-0.030439638	0	1806.967623	0
Metastatic Solid Tumor	-0.031452435	0	3569.691659	0
Cerebrovascular Disease	-0.038458755	0	3329.063341	0
Any Malignancy, Including Leukemia and Lymphoma	-0.042930853	0	7602.535827	0
Charlson Index Co-Morbidity Score	-0.049972159	0	17270.82268	0
Emergency Department Indicator	-0.05123473	0	2626.939663	0
Total Costs Mod	-0.053401903	0	-0.064454316	0
Age Group	-0.107053182	0	21098.31152	0
APR Severity of Illness Code	-0.229524194	0	54087.55429	0
APR Risk of Mortality	-0.279126701	0	240364.1034	0

The crucial identifiers of death were discussed using three statistical tests: Pearson correlation, chi-square statistics, and Kendall tau.Using different statistical tests, different factors impacting hospital death have been identified, and we will verify the actual predictors of death when we apply various machine learning models to make the prediction. We will evaluate the critical factor that comes with each machine-learning model.

We will not remove any of the features as part of general processing. We will perform a feature selection based on the Chi-Square test to cross-verify the model's performance using a reduced feature set with 10/20/30/40/50/features.

5.1.5 Proposed Model

Several in-hospital death studies have been proposed, but their work is limited to the ICU or the general ward. Most studies have performed analyses based on specific diseases like Sepsis, COVID-19 19, etc. Some studies use standard ICU mortality score tools like SAPS and APACHE for verification and prediction. Some studies used these tools to compare the result with their developed model and found that these models' output is not up to compare to their developed model. Some studies use deep learning models and machine learning techniques to predict death and identify the factors associated with death. This study consists of both ICU and general admissions, and it does not use any of the tools SAPS and APACHE. This study diverts from single disease analysis by considering around 200+ diagnoses/diseases. This is a generic model for the death prediction . The died patient's length of stay shows a median value of 6 days and an average of 10.58 days. This model proposes a death prediction model considering the length of stay as a parameter. This method proposes a '7-day death' vs. ' 7+day death ' vs. 'Survival' model. The '7-day death' model is when death happens within seven days from admission, and the '7+day death' model is that when death happens after 7 days. ‘Survival’ is defined as those who got released from a hospital after cure. The 7-day death model was selected by considering the median value of length of stay is 6.

The studies presented for death have less diverse samples and they gave more emphasis on particular diseases, considering different lab test parameters. Those studies are mainly disease-specific. They might not work well for generic data where multiple diseases are being considered. Our study does not include any lab parameters, so this study is of a generic model utilizing only primary record-level data. This is a significant study where critical data that are tough to gather are not considered for modeling. Patients' sensitive data should be hidden; this model does not use this data. This death prediction study follows a different architecture than the length of stay study discussed in previous chapters. The length of stay of a patient is associated with the budget. However, death is related to the well-being of the patient. The length of stay study can be utilized to reduce the budget, whereas death prediction can be used to save the patient's life.

Due to that, the death prediction will be performed only in the post-admission phase, unlike length of stay, which has three-phase predictions. The post-admission phase is generally considered after 1-2 days (48 hr.) of hospitalization. This post-admission stage prediction will be performed based on the features collected after hospitalization. The post-admission phase features used for this prediction will be similar to that length of stay prediction model but only differing in that this study excludes the length of stay feature from the analysis. Because in the post-admission phase, the length of stay will not be known to the hospital authority as it is a discharge time feature.

Please refer to Table 22 for hospital or post-admission phase prediction features. The feature selection for the model is similar to the length of stay prediction, where the discharge stage

features have been removed from the feature set. The detailed feature selection based on different phases has already been discussed in Chapter 3 of Length of Stay Prediction. The post-admission phase death prediction will be performed by combining the admission and post-admission phase features mentioned in the table. The discharge phase feature will not be used for our prediction.

Table 22: Post-admission stage death prediction model feature mapping

Stage	Features Used for Prediction
Admission	Health Service Area, Hospital County, Operating Certificate Number, Permanent Facility ID, Facility Name, Age Group, Zip Code - 3 digits, Gender, Race, Ethnicity, Ethnicity, CCS Diagnosis Code, Payment Typology 1, Emergency Department Indicator.
Post Admission	CCS Procedure Code, APR DRG Code, APR MDC Code, APR Severity of Illness Code, APR Risk of Mortality, APR Medical Surgical Description, Birth Weight, Abortion Edit Indicator, Total Costs.
Discharge Time	Patient Disposition, Total Charges, Expired Ind, Length of stay
Target	Dayexpired

5.1.5.1 Machine learning models

In this analysis, several machine learning and deep learning models have been utilized for the prediction; more or less, we have used the same machine learning models used in length-of-stay prediction. In this analysis, we have used a few new machine learning models, such as the imbalance machine learning model. The details about those models have been discussed in the following section.

5.1.5.1.1 RusBoost

Rusboost is a hybrid sampling/boosting algorithm, called RUSBoost, for learning from skewed training data. This Algorithm provides a more straightforward and faster alternative to available boosting algorithms. The Rusboost combines the random undersampling, synthetic minority oversampling technique, and AdaBoost in a single module. RUSBoost is an attractive alternative for improving the classification performance of learners built using imbalanced data. RUSBoost combines data sampling and boosting, providing a simple and efficient method for improving classification performance when training data is imbalanced.

RUSBoost is computationally less expensive than other sampled-based boosting methods and has significantly shorter model training times. This combination of simplicity, speed, and performance makes RUSBoost an excellent technique for learning from imbalanced data[49]. For our analysis, the rusboost was implemented using the Imblearn Python package. The hyperparameters used are Max depth at 50 and the number of estimators at 500.

5.1.5.1.2 Easy Ensemble

This is another algorithm based on sampling/boosting algorithms utilized for imbalanced data. A bag of balanced boosted learners is also known as EasyEnsemble.This Algorithm is known as EasyEnsemble.The classifier is an ensemble of AdaBoost learners trained on different balanced bootstrap samples. The balancing is achieved by random under-sampling [48]. This Algorithm is also similar to the rusboost Algorithm, which works on the imbalanced dataset.

The setting used for this ensemble is sampling strategy as 'majority.' The Imblearn learn python package is used to implement this algorithm.

5.1.5.1.3 Balanced random forest Classifier

This is an implementation of a random forest algorithm using sampling known as Balanced Random Forest. The balanced Random Forest is used in the case of imbalanced data. It works like average RF but balances the prevalence class by undersampling for each bootstrapping iteration.

The Balanced Random Forest (BRF) algorithm is shown below [54]:

1. For each iteration in a random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacements, from the majority class.

2. Induce a classification tree from the data to maximum size without pruning. The tree is induced with the CART algorithm, with the following modification: Instead of searching through all variables for the optimal split at each node, only search through a set of significant randomly selected variables.

3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

We have kept the default settings for a balanced random forest classifier for our analysis.

Other models utilized for the prediction are logistic regression, Random Forest, AdaBoost, LightGBM, Catboost, Xgboost, Decision tree, ANN, CNN 1D, and Tablet. Details about these models have been already discussed in the previous chapter. The following section will provide the parameters used for these models in Table 23.

Table 23:Death model parameters used

Model	Parameters
XGBoost	Estimator=500
LightGBM	Estimator=500 feature fraction=0.6
Catboost	Estimator=2000 Max depth=8
Random Forest	Estimator =150 Max depth =20
Adaboost	Estimator=500
ANN	4 Hidden layers
Logistic regression	Default settings
CNN	4 Convolution layers, 1 global max pooling 1 dense layer

Tabnet	n_d,n_a=64,n_steps=5, clip value=2,n_shared=2,n_independent=2 gamma=1.5
DT	max_depth=20

5.1.5.1.4. Ensemble Learning

In the previous chapters, we discussed the ensemble models. Some in-built ensemble models are bagging, boosting models, stacking, and voting classifiers. Bagging and boosting models are made up of similar base learners known as homogeneous models, whereas stacking and blending are examples of heterogeneous models created using different base classifiers. The voting classifiers are also examples of heterogeneous ensemble models produced by combining different base models. In the previous chapter, we generated a hybrid heterogeneous ensemble model utilizing several base models. That model gave us good results compared to its base models.

5.1.5.1.5 Our Architecture

Our study proposes a weighted average probability ensemble-based death prediction model. This is unique; it combines the same machine learning model, but each is a different well-tuned model. The predicted probabilities of individual models are taken, and the probabilities are averaged. Classification is performed based on the weight assigned to each classifier. The classifier's weights will be selected based on the contribution of each model towards generating an optimal ensemble model through a robust grid search method. We are proposing a 3WPB (weighted probability balanced model) ensemble model. The base model for the 3WPB ensemble model is selected based on the performance of the base models, which have the best performance across all models. This ensemble model combines both base ML models and new imbalanced models. The Algorithm for the prediction is discussed below:

5.1.5.1.5.1 Algorithm

Input: *Training dataset and testing dataset.*

Procedure: *Repeat these steps to obtain the final model*

1. Train different base models.
2. Test the base models and compare the accuracy, precision, recall, and f1 score.
3. Calibrate the individual base models to obtain optimum precision.
4. Select the top model based on the descending order of classification metrics considering accuracy, recall, and specificity.
5. Random Forest, LightGBM, Rusboost, and Balanced RF are the top-performing models selected.
6. The selected top model will be tuned to obtain maximum precision.
7. Obtain an initial weighted average ensemble model using equal weights.

8. Calculate the weights for each model using the grid search method and assign optimum weights to each classifier which is having high recall.
9. Apply weighted probability average ensemble for prediction using optimized weights.
10. Generate probability-weighted voting with a new set of classes.
11. Compare and evaluate the precision of the new ensemble model with the base model.

Please refer to Figure 20 for the death prediction process flow diagram.

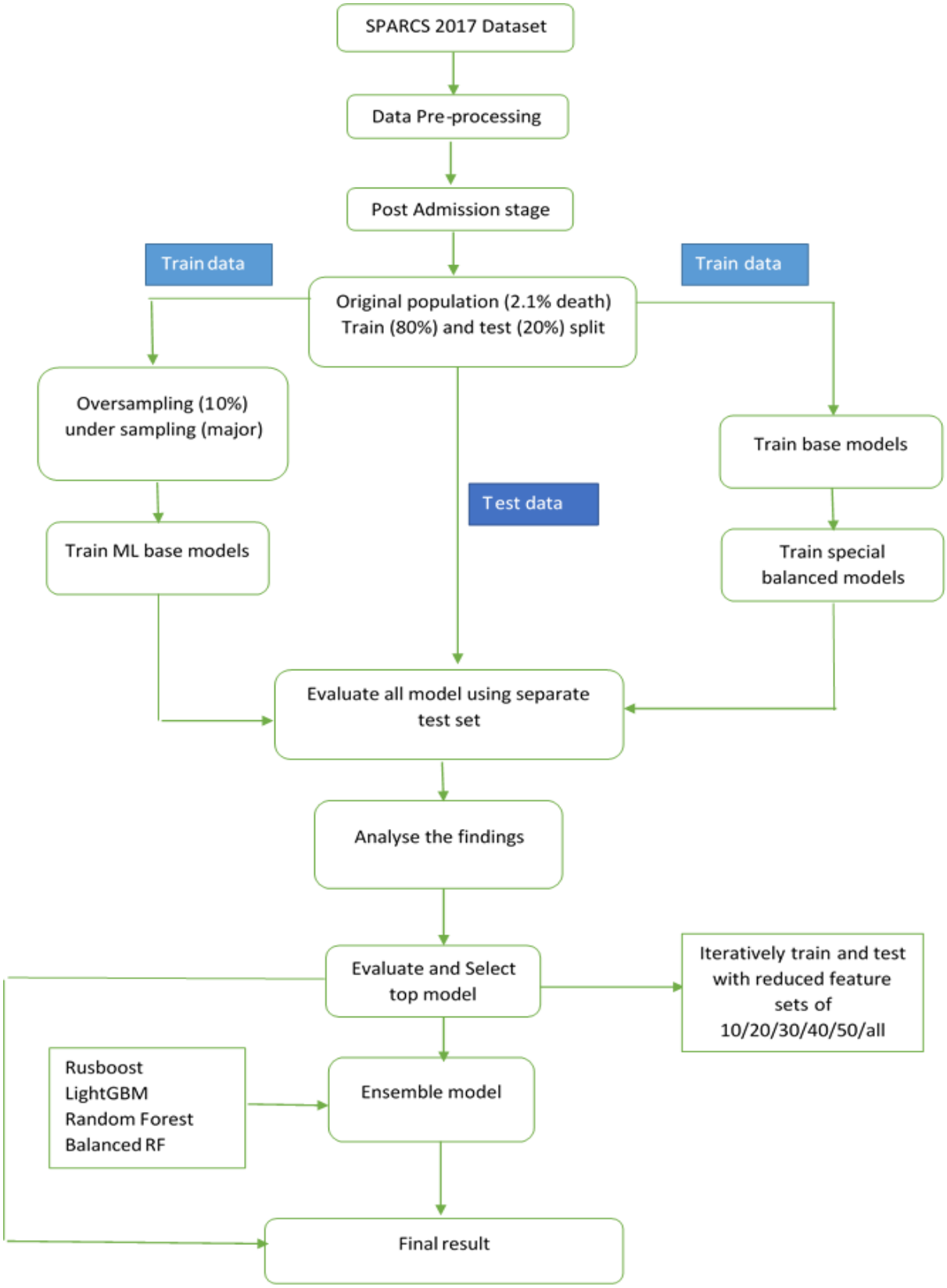


Figure 20: Process flow diagram of death prediction

5.1.5.1.5.2 Pseudo code of the Algorithm

1. Define models as a list of tuples with model objects and their names:
models = [(Rust, 'rust'), (lghtgbm, 'lghtgbm'), (Randmreg, 'RF'), (Rb_bal, 'BalRF')]
2. Initialize preds as an empty list
3. For each model and its type in models:
a. If the model type is 'CNN,' use the predict method and add the result to preds
b. Otherwise, use the predict_proba method and add the result to preds
4. Convert preds to a numpy array and assign it to array_opt
5. Calculate the soft ensemble by averaging array_opt along axis 0 with weights [0.1, 0.1, 0.1]
soft_ensemble = np.average(array_opt, axis=0, weights=[0.1, 0.1, 0.1])
6. Calculate soft_ensemble_pred as the argmax of soft_ensemble along axis 1
7. Calculate f1 as the classification report of y_test11 and soft_ensemble_pred
8. Calculate acc as the accuracy score of y_test11 and soft_ensemble_pred
9. Calculate conf as the confusion matrix of y_test11 and soft_ensemble_pred
10. Initialize df_new as an empty data frame
11. For w1 from 1 to 8:
a. For w2 from 1 to 8:
i. For w3 from 1 to 8:
1. Calculate weights as [w1/10, w2/10, w3/10]
2. Calculate wted_preds2 as the average of array_opt along axis 0 with weights
3. Calculate wted_ensemble_pred2 as the argmax of wted_preds2 along axis 1
4. Calculate weighted_accuracy as the accuracy score of y_test11 and wted_ensemble_pred2
5. Calculate f1score as the f1 score of y_test11 and wted_ensemble_pred2 with average='macro'
6. Calculate recall as the recall score of y_test11 and wted_ensemble_pred2 with average='macro'
7. Calculate price as the precision score of y_test11 and wted_ensemble_pred2 with average='macro'
8. Append a new row with weights and performance metrics to df_new
12. Save df_new to '/kaggle/working/votingwegtd.csv'
13. Find the row in df_new with the maximum recall and assign it to max_acc_row
14. Calculate the soft ensemble by averaging array_opt along axis 0 with weights from max_acc_row
soft_ensemble = np.average(array_opt, axis=0, weights=[max_acc_row[0], max_acc_row[1], max_acc_row[2]])
15. Calculate soft_ensemble_pred as the argmax of soft_ensemble along axis 1
16. Calculate f1 as the classification report of y_test11 and soft_ensemble_pred
17. Calculate acc as the accuracy score of y_test11 and soft_ensemble_pred
18. Calculate conf as the confusion matrix of y_test11 and soft_ensemble_pred
19. Calculate f1score as the f1 score of y_test11 and soft_ensemble_pred with average='macro'
20. Calculate recall as the recall score of y_test11 and soft_ensemble_pred with average='macro'
21. Calculate prec as the precision score of y_test11 and soft_ensemble_pred with average='macro'
22. Calculate aucroc as the roc_auc score of y_test11 and soft_ensemble with multi_class='ovr'
23. Calculate spcty as the specificity score of y_test11 and soft_ensemble_pred with average=None
24. Calculate spcty as the specificity score of y_test11 and soft_ensemble_pred with average='macro'

5.1.5.1.5.3 Hyperparameter tuning of base models

The weighted ensemble model generated using balanced and base models needs the base models to be highly precise and calibrated. Well-tuned base models give more power to the ensemble model. If the base model's performance is not nearly equal or similar, the ensemble model might not perform well. We have identified four top-performing models based on evaluation metrics. The top models are Random Forest, LightGBM, Rusboost, and Balanced Random Forest.

5.1.6 Evaluation criteria

For classification problems, classic evaluation metrics are accuracy, precision, recall, F1-score, specificity, and AUC-ROC value. We will be using these metrics to validate our in-hospital mortality prediction model. These metrics define different aspects of the model's performance. We aim to produce maximum efficiency from these metrics by combining base models. As this is a death prediction model of a hospital, the correct identification of hospital death will be given more weight than the model identifying nondeath as death. Recall and specificity will be prioritized for death prediction to better validate the model's performance.

5.2 Result

Initially, the USA SPARCS dataset had 2343569 records with 34 features. After data pre-processing, duplicate removal, feature selection, and feature engineering, we have 2307981 instances left in our dataset. The feature dimension increased from 34 to 64. We have utilized the online data science and ML platform Kaggle for our coding and analysis. Kaggle is an open-source online machine learning and data science research platform. The platform uses Python 3.10 with the latest Scikit learn, Tensorflow, learn learn, PyTorch, and Keras ML packages. The models are trained using 80% of the data, and the remaining 20% is used for testing. The in-hospital death prediction results are illustrated in Tables 24 through 26. The performance of the base and ensemble models are also presented in the table. The final optimum weights of the 3WPB are presented in the table. The optimum weight search space is used between 0.1 to 0.8.

5.2.1 Experiment strategy

The experiments on death prediction have been carried out in two phases. One is with original unsampled data, and the other is with sampled data. The ensemble model will take top performers considering both sets of samples. The first experiment uses the sampled training data to verify the performance of all base learners. The original data is also used to train and test base models.

A few balanced ML models are also used along with these base models. Finally, Chi-square feature selection will be used to select a reduced feature set based on the ranking of the features to verify the model's performance literately. The feature size will be 10/20/30/40/50/all. The pseudo-code of iterative feature selection and model evaluation logic is given below.

5.2.1.1 Pseudo code of iterative feature selection

1. Initialize feat to 10
2. While the feat is greater than 9:
  a. Print "no of features" and the value of feat
  b. Initialize chi2_selector using SelectKBest with chi2 and k set to feat
  c. Fit and transform X_train2_scaled with y_train21 using chi2_selector, and assign the result to X_train1_bst
  d. Get the column indices of the selected features using chi2_selector and assign them to cols_idxs
  e. Select the columns from X_train21 at indices cols_idxs and assign to features_df_new
  f. Transform X_test2_scaled using chi2_selector and assign the result to X_test1_bst
  g. Decrement feat by 10

3. Import RUSBoostClassifier from learn.ensemble
4. Initialize Rust as an instance of RUSBoostClassifier with:
a. DecisionTreeClassifier (max_depth=50) as the estimator
b. n_estimators set to 500
5. Fit Rust with X_train1_bst and y_train21
6. Predict using Rust on X_test1_bst and assign the result to predict_Rust
7. Calculate f1 as the classification report of y_test21 and predict_Rust
8. Calculate acc as the accuracy score of y_test21 and predict_Rust
9. Calculate conf as the confusion matrix of y_test21 and predict_Rust
10. Predict probabilities using Rust on X_test1_bst and assign the result to y_pred_prob
11. Calculate across as the roc_auc_score of y_test21 and y_pred_prob with multi_class set to 'over'
12. Calculate spicy as the specificity score of y_test21 and predict_Rust with average=None
13. Calculate spcty as the specificity score of y_test21 and predict_Rust with average='macro'

5.2.2 Result Discussion

In the first phase, the training and testing were done using the original dataset. The classifiers' performance was not good in the original sample because of the high data imbalance. The Xgboost model has performed best in original data compared with other learners; it has accuracy, precision, recall, f1 score, specificity, and AUC of 0.9807, 0.68, 0.53, 0.58,0.7717, and 0.967, respectively. Refer to Table 24 for the detailed result. Though the model has high accuracy, the precision, recall, and specificity scores are very low; specifically, the recall is unsuitable for identifying dead patients. This model has been discarded from the analysis. The original data is also used to train and test the three auto-balanced ML models, Rusboost, easy ensemble, and balanced random forest; the accuracy of the models is between 0.80 to 0.84, but the precision of the models is low for the death classes '7day' and '7+day' precision is in the range of 0.07-0.11 for all the models, whereas the models have decent recall ranging from 0.8-0.85. The positive classes (death) also have a good recall greater than 0.77. The average overall specificity of the models is within the range of 0.91- 0.93.

In the subsequent analysis phase, the sampled data was used for model training; testing was done on a separate set before sampling. The smote technique and random under-sampling are combined for sampling. The minority samples were increased by 10%, followed by random under-sampling. In this sampling setting, the performance of the classifiers are good. The linear regression, Adaboost, Decision tree, and CNN perform slightly bad than the other learners. RF, LightGBM, Catboost, and XGBoost are the top-performing models. RF model has accuracy, precision, recall, f1 score, specificity, and AUC values as 0.9089, 0.43, 0.82, 0.49, 0.9217, and 0.9630, respectively. LightGBM model has accuracy, precision, recall, f1 score, specificity, and AUC values as 0.9303, 0.46, 0.82, 0.53, 0.9214, and 0.9704, respectively. The Catboost model has accuracy, precision, recall, f1 score, specificity, and AUC values as 0.9298, 0.46, 0.82, 0.52, 0.9208, and 0.9698, respectively. Performances of Catboost and LightGBM are similar. Xgboost model has accuracy, precision, recall, f1 score, specificity, and AUC values as 0.9408, 0.47, 0.79, 0.54, 0.9086, and 0.9686 respectively. The performance of the four models is more or less equal. We have combined two models from this manually sampled data and two from the auto-balanced model. Combining two different model types can convey a mixed prediction, making the model less error-prone. The new hybrid ensemble model averages the prediction, which keeps the prediction from both models and produces a robust and diverse death classifier. The error made by different predictors can be ruled out by the hybrid model, which combines three/four models to create a balanced weighted ensemble

model known as the 3WPB/4WPB model. This model's performance is suitable compared to its base models. The base models are based on manual and auto sampling; averaging the results bias about individual models will be rectified. The performance of the 3WPB and 4WPB models are accuracy, precision, recall, f1 score, specificity, and AUC as 0.8827, 0.42, 0.85, 0.46, 0.9364, 0.9682 for 3WPB and .0.8916, 0.43, 0.85, 0.47, 0.9373, 0.9686 for 4WPB respectively. This new proposed model has overall good balanced accuracy, recall, and specificity performance compared to other models. Any model between 3WPB and 4WPB can be selected, but for complexity and time constraints, 3WPB will be best.

The detailed results of the models are presented in two tables, table 25 and table 26. This table shows the class-level performance of the individual models. It has been observed that the surviving class has performed best across all the models. For all the models, the precision of the '7day' and '7+day' models is consistently lower than that of the recall of the models except for the 'survived' class, which has better precision. For most models, the difference between precision and recall is very high; Xgboost,LightGBM and Catboost have slightly better precision for the two death classes. The precision for the 'survived' class varies from 0.99 to 1. The '7day' class has a precision value of 0.07-0.23, and '7+day' has 0.08-0.19 for these two classes XGBoost tops the table. The recall value is quite good compared to the precision value of the models, whereas the 'survived' class has a high recall of around 0.8 to 0.94, the '7day' has a value in the range of 0.69-0.83 RUSBoost model has a higher recall, and '7+day' has recall in the range of 0.65 to 0.88 again rusboost is the winner for this class. The models can classify death and survival quite well as the true positive rate is reasonable, though it is making a mistake by wrongly classifying nondeath as death as precision for the '7day' and '7+day' models is low. The specificity of the models for the 'survived' class is in the range of 0.77-0.95; again, rusboost has high specificity for the 'survived' class; the LightGBM, Catboost, and RF model also have reasonable specificity for 'survived' class in the range of 0.82-0.85. The '7day' class varies in the range of 0.88-0.97, with easy ensemble having the lowest and Xgboost having the highest in this group. For the '7+day' class, this varies in the range of 0.91-0.97; the result is the same as the '7day' model. The specificity of the model is quite good. Again, to produce a balanced model that not only makes good results but tries to minimize the error of the individual model, we prepared a 3WPB and 4WPB model. The performance of these two models is quite good. The precision for the '7day' model is between 0.15 -0.17, the '7+day' model is 0.11, and survived model 1. The recall value lies between 0.76-0.77 for the '7day' model, 0.9-0.91 for the '7+day' model, and 0.88-0.89 for the 'survived class, which is quite good and balanced. The specificity lies between the 'survived' class is 0.91-0.92, the '7day' model is 0.94-0.95, and the '7+day' is 0.93. This ensemble model averages and reduces the bias created by individual models.

Using the single Rusboost model, a different feature subset was tested; the result is shown in Table 27. The Rusboost model with all the features performed best. The performance of the Rusboost model with up to 40 features was satisfactory, but after that it has declined abruptly.

Table 24: Performance of the death for all predictors

Sampling	Models	Accuracy	Precision	recall	f1 score	Specificity	Auc	Weights
With Sampling (10% SMOTE + RNDM)	LR	0.8953	0.42	0.76	0.46	0.8961	0.9447	X
	RF	0.9089	0.43	0.82	0.49	0.9217	0.963	X
	Adaboost	0.8882	0.42	0.78	0.46	0.8991	0.8865	X
	XGBoost	0.9408	0.47	0.79	0.54	0.9086	0.9686	X
	LightGBM	0.9308	0.46	0.82	0.53	0.9214	0.9704	X
	Catboost	0.9298	0.46	0.82	0.52	0.9208	0.9698	X
	Tabnet	0.9108	0.43	0.79	0.48	0.9102	0.9569	X
	ANN	0.9127	0.43	0.79	0.49	0.9129	0.9591	X
	CNN	0.92	0.44	0.77	0.49	0.9002	0.9526	X
	DT	0.8915	0.41	0.76	0.45	0.8972	0.8218	X
	4WPH	0.8916	0.43	0.85	0.47	0.9373	0.9682	0.1,0.2,0.1,0.1
	3WPH	0.8827	0.42	0.85	0.46	0.9364	0.9682	0.4,0.5,0.3
Without Sampling	Rusboost	0.8427	0.4	0.85	0.43	0.9318	0.9607	X
	Easy Ensemble	0.8024	0.39	0.8	0.39	0.9132	0.915	X
	Bal RF	0.8429	0.4	0.85	0.42	0.9322	0.9622	X
	XGBoost	0.9807	0.68	0.53	0.58	0.7717	0.967	X

Table 25: Class-level performance of ML models

Model	Class	Precision	recall	f1 score	Specificity
LR	7daydeath	0.12	0.73	0.21	0.9372
LR	7+daydeath	0.13	0.65	0.22	0.9613
LR	Survived	1	0.9	0.94	0.7897
RF	7daydeath	0.16	0.78	0.26	0.9494
RF	7+daydeath	0.15	0.77	0.25	0.9616
RF	Survived	1	0.91	0.95	0.8542
Adaboost	7daydeath	0.11	0.74	0.2	0.9286
Adaboost	7+daydeath	0.14	0.71	0.24	0.9624
Adaboost	Survived	1	0.89	0.94	0.8062
LightGBM	7daydeath	0.2	0.77	0.32	0.9633
LightGBM	7+daydeath	0.18	0.75	0.29	0.9703
LightGBM	Survived	1	0.93	0.96	0.8306
Catboost	7daydeath	0.2	0.77	0.32	0.9632
Catboost	7+daydeath	0.18	0.75	0.29	0.9694
Catboost	Survived	1	0.93	0.96	0.8299

ANN	7daydeath	0.15	0.74	0.25	0.95
ANN	7+daydeath	0.16	0.72	0.26	0.9654
ANN	Survived	1	0.92	0.95	0.8232
CNN	7daydeath	0.17	0.69	0.27	0.9577
CNN	7+daydeath	0.15	0.69	0.25	0.966
CNN	Survived	0.99	0.92	0.96	0.777
Tabnet	7daydeath	0.15	0.73	0.24	0.9365
Tabnet	7+daydeath	0.15	0.71	0.25	0.958
Tabnet	Survived	1	0.91	0.95	0.797
DT	7daydeath	0.12	0.71	0.21	0.9365
DT	7+daydeath	0.12	0.67	0.21	0.958
DT	Survived	1	0.94	0.94	0.797

Table 26: Performance class level of base and ensemble models

Model	Class	Precision	recall	f1 score	Specificity
xgboost	7daydeath	0.23	0.73	0.35	0.9707
XGBoost	7+daydeath	0.19	0.7	0.35	0.974
XGBoost	Survived	1	0.93	0.97	0.781
Rusboost	7daydeath	0.11	0.83	0.19	0.9149
Rusboost	7+daydeath	0.1	0.88	0.17	0.927
Rusboost	Survived	1	0.84	0.91	0.9535
Bal RF	7daydeath	0.1	0.82	0.18	0.9128
Bal RF	7+daydeath	0.1	0.88	0.18	0.929
Bal RF	Survived	1	0.84	0.91	0.9529
Easy Ensemble	7daydeath	0.07	0.77	0.14	0.8836
Easy Ensemble	7+daydeath	0.08	0.84	0.15	0.9179
Easy Ensemble	Survived	1	0.8	0.89	0.9381
3WPH (proposed)	7daydeath	0.15	0.77	0.26	0.9483
3WPH (proposed)	7+daydeath	0.11	0.91	0.19	0.9348
3WPH (proposed)	Survived	1	0.88	0.94	0.926
4WPH (proposed)	7daydeath	0.17	0.76	0.28	0.9544
4WPH (proposed)	7+daydeath	0.11	0.9	0.2	0.9376
4WPH (proposed)	Survived	1	0.89	0.89	0.9197

Table 27: Performance of Rusboost with different feature sets

Feature set	Model	Accuracy	Precision	recall	f1 score	Specificity	Auc
All	Rusboost	0.8427394	0.4	0.85	0.43	0.9318	0.9607
50	Rusboost	0.8423885	0.4	0.85	0.43	0.9319	0.9605
40	Rusboost	0.8391086	0.4	0.85	0.42	0.9308	0.9583
30	Rusboost	0.8357745	0.4	0.83	0.42	0.92889	0.9531
20	Rusboost	0.8271024	0.38	0.68	0.38	0.925	0.9122
10	Rusboost	0.802867	0.37	0.65	0.37	0.9154	0.902

5.2.3 Feature importance

This new model of in-hospital death prediction using sparcs data has been developed, and the top performing models identified as Random Forest, LightGBM, Rusboost, and Balanced RF model. The top 30 features of these models have been plotted and shown in figure 21 to figure 24.

The top features identified by the random forest model are APR Risk of Mortality, Total Costs Mod, APR Severity of Illness Code, Facility Name, CCS Procedure Code, CCS Diagnosis Code, APR DRG Code, Operating Certificate Number, Zip Code - 3 digits, Permanent Facility Id, Payment Typology 1_Medicare, Hospital County, Emergency Department Indicator, Age Group, Gender, APR MDC Code. The features having lower importance are Diabetes, Abortion Edit Indicator, etc.

The top features identified by the LightGBM model are APR DRG Code, CCS Procedure Code, Facility Name, Total Costs Mod, CCS Diagnosis Code, Operating Certificate Number, Permanent Facility ID, Zip Code - 3 digits, Hospital County, APR MDC Code, APR Risk of Mortality, APR Severity of Illness Code, Age Group, Charlson Index Co-Morbidity Score, etc. The lower importance features are AIDS, Diabetes, and Abortion Edit Indicators, etc.

The top features identified by the Rusboost model are Total Costs Mod, APR Risk of Mortality, CCS Procedure Code, Facility Name, CCS Diagnosis Code, APR DRG Code, Operating Certificate Number, Zip Code - 3 digits, Permanent Facility ID, Hospital County, Age Group, APR Severity of Illness Code, APR MDC Code, Charlson Index Co-Morbidity Score, etc. The features having lower importance are Diabetes, Hospital Service Area Not Provided, etc.

The top features identified by the Balanced RF model are Total Costs Mod, APR Risk of Mortality, APR Severity of Illness Code, CCS Diagnosis Code, CCS Procedure Code, APR DRG Code, Age Group, Facility Name, Permanent Facility ID, Operating Certificate Number, Zip Code - 3 digits, APR MDC Code, Hospital County, Payment Typology 1_Medicare, APR Medical Surgical Description Surgical, Gender, Charlson Index Co-Morbidity Score etc. The features having lower importance are Diabetes, Abortion Edit Indicator etc.

The top death features are Total Costs Mod, APR Risk of Mortality, APR Severity of Illness Code, CCS Diagnosis Code, CCS Procedure Code, APR DRG Code, Age Group, and APR MDC Code. In contrast, the Charlson comorbidity factor is less important than these features.

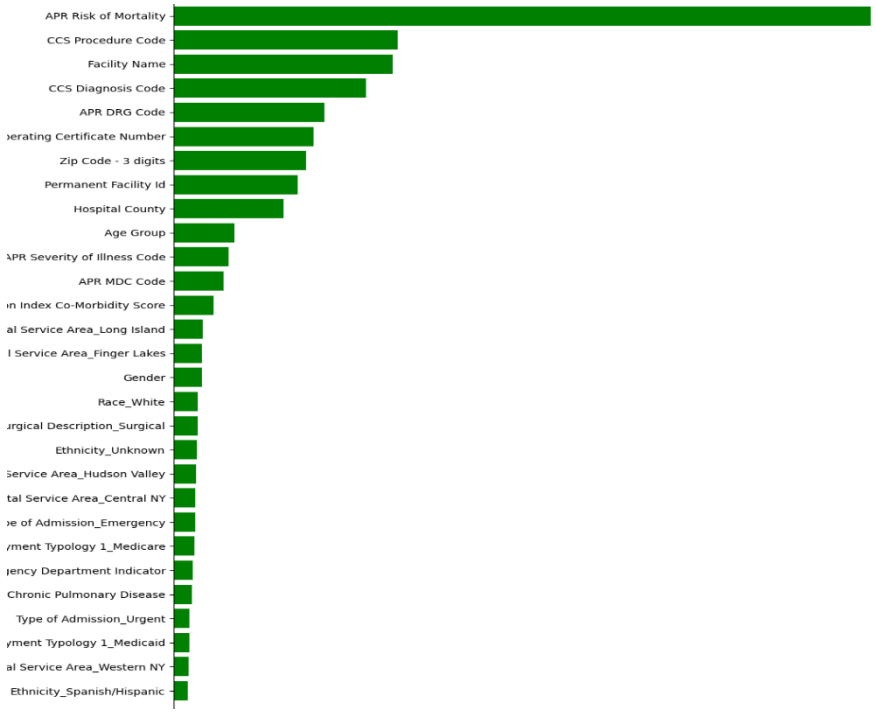


Figure 21: Feature importance plot for death model using Rusboost model

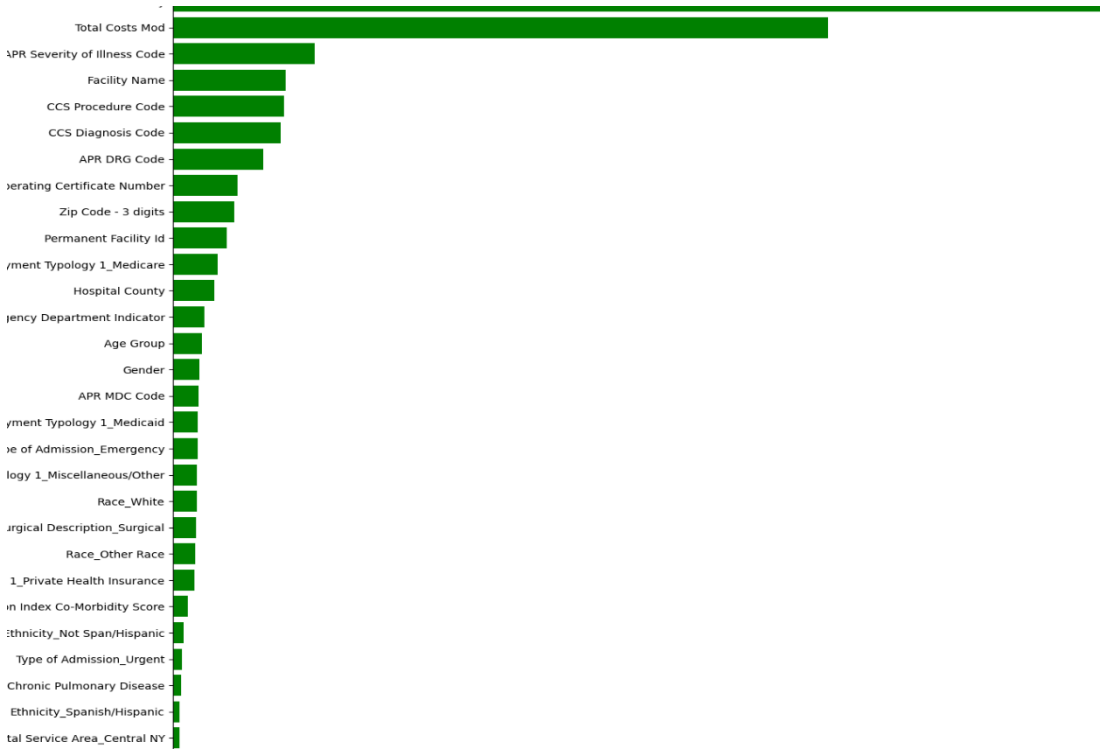


Figure 22: Feature importance plot for death model using RF model

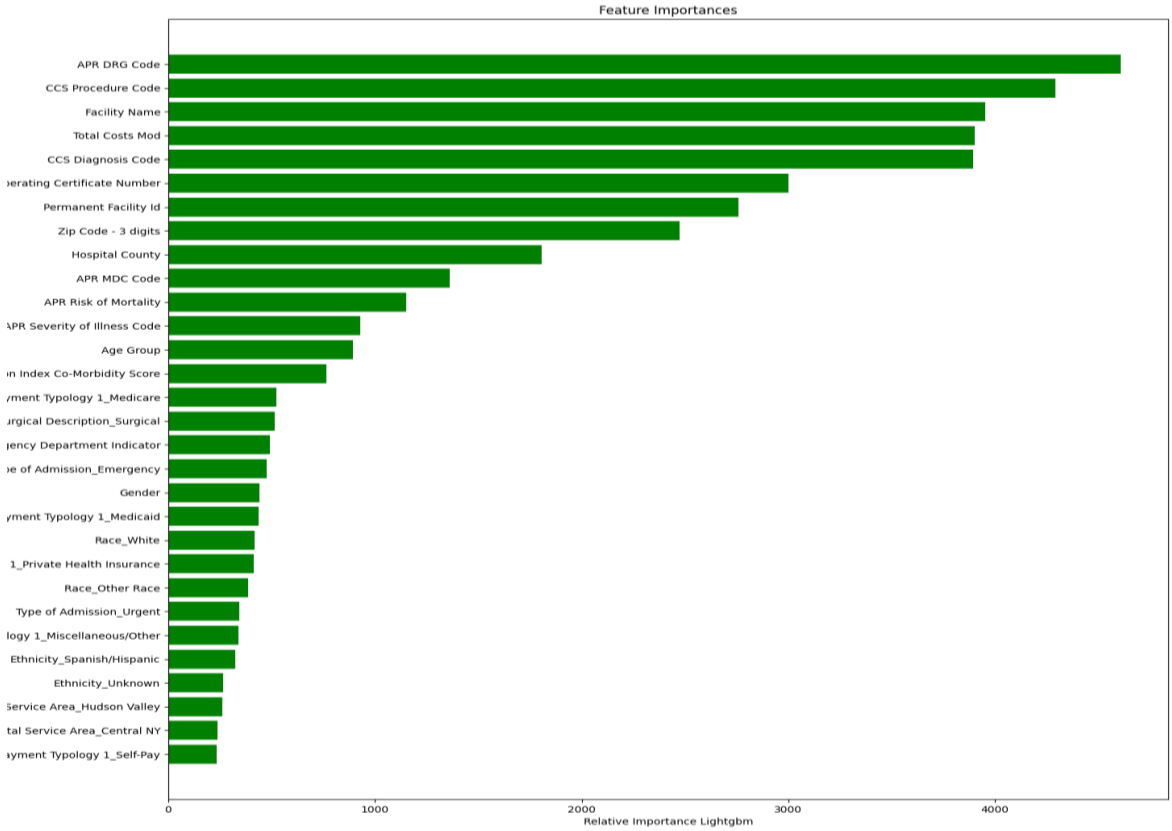


Figure 23:Feature importance plot for death model using LightGBM model

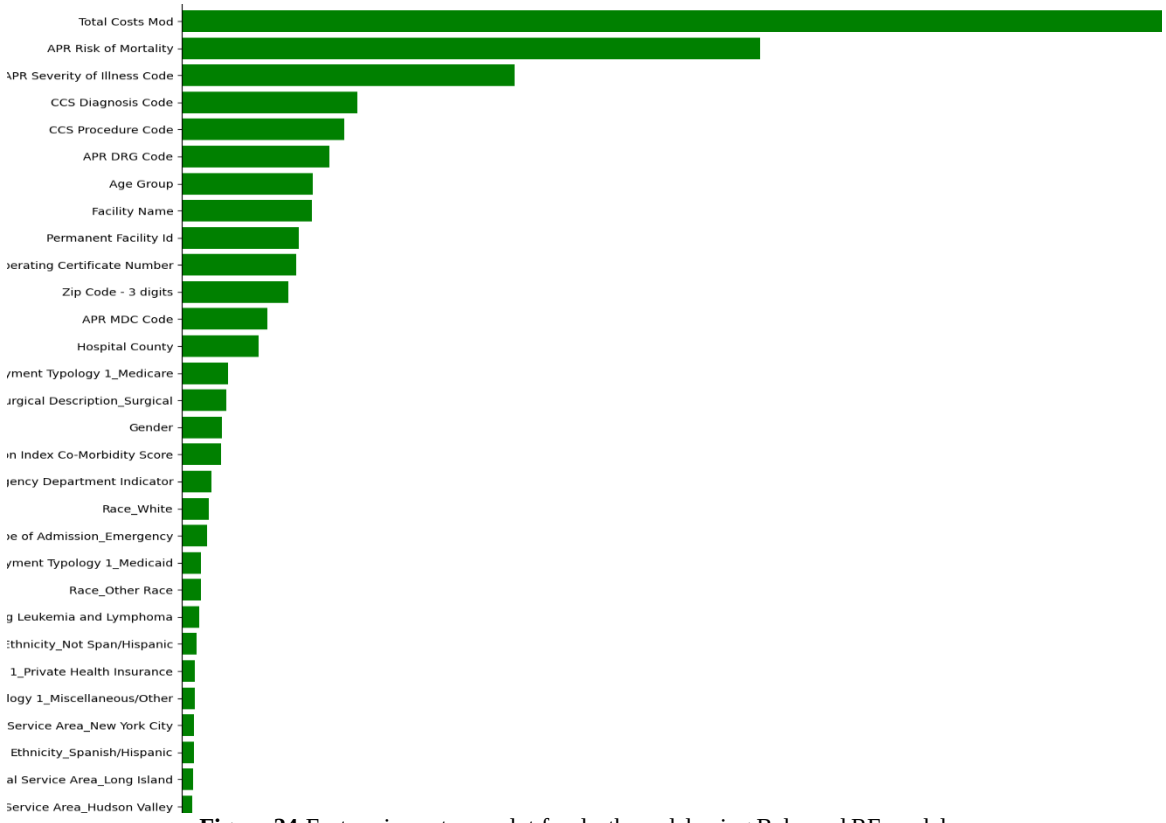


Figure 24: Feature importance plot for death model using Balanced RF model

5.3 Discussions

Death prediction is vital for patients admitted to the hospital. One of the most critical tasks of the hospital, apart from disease diagnosis and treatment, is the analysis of the severity of the patient and the risk of mortality of the patient. The death prediction tool can help the hospital and physician identify severe patients, and care can be taken accordingly. For inpatients, successful death prediction can benefit the hospital authority by allowing them to take optimum patient care to reduce the risk of death. This death prediction model introduces a '7day death' and '7+day death' prediction model. This new model will inform the hospital authority whether the patient will die within 7 days or after 7 days post admission. This new model will help hospital administrators identify critical patients so that treatment plans can be prioritized accordingly. This can be achieved by utilizing several machine learning models; the final model will be an ensemble model for death prediction.

This study utilizes the Charlson comorbidity score [23] and seventeen comorbidity factors for death prediction; the statistical analysis showed that the Charlson comorbidity factors significantly impact death prediction. The Severity of illness , Risk of mortality , Diagnosis code , Age group, total cost have more impact on death prediction compare to Charlson factors. Many studies use different tools for ICU death prediction, such as SAPS and APACHE. In their analysis, this study does not use this tool as our data has both ICU and general admissions. This study focuses on building a machine-learning model based on a death prediction tool. Studies show that machine learning is more robust and performs better than these tools. This study uses both comorbidity factors and the features present in the dataset to verify their impact on death prediction.

Compared to other studies, Mehmet Kivrak et al. [37] predicted the death of covid 19 patients with the XGBoost model with an accuracy of 99.7%, precision, sensitivity, and specificity of 99.7,99.7 and 100 %, respectively, the dataset they have referred has less sample compare to us and their focus is only on a single disease, we have predicted using 3WPB and our model has accuracy, precision, recall, f1 score, specificity, Auc score as 0.8827, 0.42, 0.85, 0.46, 0.9364, 0.9682. Our model has low precision, mainly for the '7day' and '7+day' models; this study aims to identify death correctly. The mistake of saying death to non-death is a severe mistake, but saying non-death to death is less severe. The precision of the two death classes is mainly misclassifying non-death to death, which is not that harmful. Whereas the model has good recall, the death-to-non-death error is less.

The actability of the model is good. Other studies mainly considered death prediction using two classes: death and survived, but our study breaks death into two classes: '7daydeath', '7+daydeath', and the final one, ' survived.' This model is more robust and effective than the previous models as several diseases are considered for this analysis. Some studies predict [42] death after release from the hospital also, but this study mainly focuses on death in the hospital. The release from the hospital is only triggered when the patient is cured or died. For cured patients, the death prediction at home or in any other disposition is of different research; it is not part of this analysis. In this study, the length of stay feature has been removed as the length of study should be unknown for the post-admission phase. Most studies on death prediction models mainly focus on particular diseases, but this study is generic; it considers around 200+ primary diseases. The balanced models used in this study can detect death classes better compared to base ML models.The rusboost has better recall for '7daydeath' , ensemble has for '7+daydeath' combinedly both the model can be used for death prediction.

However, this study has some limitations. Firstly, SPARCS does not share the patients' tab test-related parameters. Those parameters might have created a good impact on accurately predicting the patients' death. It also lacks features related to the number of available beds in the hospital, number of outpatient visits, number of ICU admissions, and readmissions performed by the patient in that year. Whether patients have any existing comorbidities like high blood pressure or diabetes, this information might have helped to predict the death more accurately. Though the dataset has some limitations, we can still manage acceptable performance. The good part of the dataset is that it is not limited to any specific region or hospital; it has diverse categories of patients and includes many hospital data. The analysis is based on the whole data, and we have not performed any outlier removal techniques. The '7day death' model is selected based on patients' median length of stay is 6. The 7-day model makes a good death prediction after having few drawbacks. Though the rusboost model identifies the death class well, the 3WPB model has overall good acceptance as the metrics are more balanced. The performance of the reduced feature set using the top model rusboost shows that up to 40 features, the rusboost model produces good performance; after 40 features, the performance degrades the 10 features that have the lowest values among all.

This study shows an impactful model for in-hospital death prediction, capable of death prediction within 7 days; after 7 days of hospitalization, this creates a unique model for death prediction and finding out the factors related to death. This study utilizes multiple machine learning models and creates an ensemble model. Hospital authorities and medical insurance companies can use this new model to better balance and manage their resources for optimum care.

Chapter 6

Conclusion and Future Work

6.1 Conclusion and future work

Forecasting how long a patient will stay in the hospital is a potentially valuable strategy for healthcare resource management systems. The length of stay prediction helps both the patient and hospital management understand how long the hospitalization will take. The loss is associated with a medical budget for the patient, so if any patient can get an estimation early, it will help them accumulate the required budget. The hospital management will also benefit from allocating and reserving the assets for a patient's care. This will benefit the insurance companies by allowing them to get an estimation of the payout.

Death prediction is also a vital prediction for in-hospital patients. The timely death prediction will reduce the risk of death and be a parameter to improve hospital activity. A proper death prediction model can result in good hospital management and increase their reputation. Accurately, death prediction can save patients' lives by assigning proper specialized doctors and medical facilities. The 7-day death model will help the hospital authority identify critically-risk patients.

We have proposed a three-stage model to predict hospital length of stay and to direct the issue of early prediction of Los. Our model comprises three stages: admission, post-admission, and discharge. Studies on the SPARCS dataset have considered only single-phase models [33] [34] [10]. We have introduced a three-stage model to predict the length of stay to give flexibility to the patient and hospital party to get an estimation from admission to discharge stage. The accuracy improved from admission to discharge phase by adding new features.

The first prediction will be performed at admission with quite a few features. The performance at this stage is less significant than less information availability. In the post-admission stage, prediction will be performed with increased features accumulated after hospitalization. This model is more robust and accurate than the admission model. The discharge model will perform the final prediction. This final stage model can be used as a reference point for the hospital authority and insurance companies to verify how well their models contribute and will be utilized for analysis purposes. These three-phase predictions have been achieved by an ensemble model called 3WPE. This three-phase model identifies the factors contributing to each stage, and this model is new compared to the existing models. The performance of the three-phase model is high enough compared to the existing state-of-the-art models, including the predictions that happen in sparcs data.

For death prediction, we have proposed a '7daydeath', '7+daydeath', and survived model for in-hospital patients. To direct the issue of early death prediction, this model is ideal as it does not just predict death or survival; it also gives a suspected length of stay for patient death. This prediction will be performed only on inpatient patients, considering the phase post-admission,

unlike the length of stay prediction performed at three stages. This study finds the Charlson comorbidity factors along with the features available in the dataset for death prediction, unlike other studies on sparcs, which consider only Charlson factors alone or the sparcs features alone to predict death. This prediction will utilize many hospitals and diseases to create a robust model. The studies that carried out the death prediction are mainly given priority to specific diseases or admissions related to ICU only. This study uses a generic model for the prediction. This 7day model is a unique approach for death prediction in the post-admission phase using only primary patient-level data. The study shows a competitive result. An ensemble model achieved the final prediction.

From the future scope perspective, there is a scope for improvement of the model's performance by adding other models into our analysis and building a more improved ensemble model. The main problem lies in the calibration of the base models, as the data size is enormous, and it takes a significant amount of CPU time and cost to train and predict models. In the future, we will try to work on this prediction to add more value.

References

1. Y. Xie, G. Schreier, D. C. W. Chang, S. Neubauer, S. J. Redmond, and N. H. Lovell, "*Predicting number of hospitalization days based on health insurance claims data using bagged regression trees*," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2014, pp. 3314-3317. doi: 10.1109/EMBC.2014.6944181.
2. Y. Xie, G. Schreier, M. Hoy, Y. Liu, S. Neubauer, D. C. W. Chang, S. J. Redmond, and N. H. Lovell, "*Analyzing health insurance claims on different timescales to predict days in hospital*," Journal of Biomedical Informatics, vol. 60, pp. 276-286, 2016. doi: 10.1016/j.jbi.2016.01.002.
3. R. Jabeen and S. N. Fatima, "*Predicting days in hospital using health insurance claims*," International Journal of Scientific Engineering and Technology Research (IJSETR), vol.5, no. 29, pp. 6081-6086, Sep. 2016. [Online]. Available: http://ijsetr.com/uploads/5/3/2/9/53299371/ijsetr_vol_5_issue_29.pdf.
4. Merhan A. Abd-Elrazek, A. A. Eltahawi, M. H. Abd Elaziz, and M. N. Abd-Elwhab, "*Predicting length of stay in hospitals intensive care unit using general admission features*," Ain Shams Engineering Journal, vol. 12, no. 2, pp. 1561-1570, 2021. doi: 10.1016/j.asej.2021.02.018.
5. S. Li, Y. Zhang, Q. Zhang, Y. Chen, S. Liu, Q. Wang, et al., "*Predicting hospital length of stay with neural networks based on electronic health records*," Journal of Biomedical Informatics, vol. 113, p. 103631, 2021. doi: 10.1016/j.jbi.2020.103631.
6. G. Srimannarayana and P. Venkata Rajesh, "*A study on days of hospitalization of insured with the claims data*," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 3, pp. 5230-5238, 2021. [Online]. Available: https://turcomat.org/article/download.
7. C. Wang et al., "*Prediction of days in hospital for children using random forest*," ResearchGate, Oct. 2017. [Online]. Available: https://www.researchgate.net/.
8. Joseph Ebinger, M. Wells, D. Ouyang, T. Davis, N. Kaufman, S. Cheng, and S. Chugh, "*A machine learning algorithm predicts duration of hospitalization in COVID-19 patients*," Intelligence-Based Medicine, vol. 5, p. 100035, 2021. doi: 10.1016/j.ibmed.2021.100035.
9. Nuno Caetano, R. M. S. Laureano, and P. Cortez, "*A data-driven approach to predict hospital length of stay: A Portuguese case study*," In Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS-2014), pp. 407-414, DOI: 10.5220/0004892204070414.
10. Xiangrui Zeng, "*Length of stay prediction model of indoor patients based on Light Gradient Boosting Machine*," Hindawi, Computational Intelligence and Neuroscience, vol. 2022, pp. 9517029, Aug. 2022. doi: 10.1155/2022/9517029.
11. Dimitris Bertsimas and J. Pauphilet, "*Predicting inpatient flow at a major hospital using interpretable analytics*," May 2020. doi: 10.1101/2020.05.12.20098848.
12. Miguel Ortiz-Barrios et al., "*Predicting 15-day unplanned readmissions in hospitalization departments: An application of logistic regression*," Ingeniare. Revista chilena de ingeniería, vol. 29, no. 2, pp. 378-398, 2021.
13. Lakshmi Prayaga, K. Devulapalli, and C. Prayaga, "*Predicting Inpatient Status for the Next 30/60/90 Days With Machine Learning*," International Journal of Big Data and Analytics in Healthcare, vol. 6, no. 2, pp. 1-18, Jan. 2021, doi: 10.4018/IJBDAH.20210701.oa9.
14. Md. Mahbubur Rahman, D. Kundu, S. A. Suha, U. R. Siddiqi, and S. K. Dey, "*Hospital patients' length of stay prediction: A federated learning approach*," Journal of King Saud

University - Computer and Information Sciences, vol. 34, no. 10A, pp. 7874-7884, Nov. 2022. doi: 10.1016/j.jksuci.2022.07.006.
15. Jan Chrusciel, F. Girardon, L. Roquette, D. Laplanche, A. Duclos, and S. Sanchez, "*The prediction of hospital length of stay using unstructured data*," Journal of Medical Internet Research, vol. 23, no. 12, p. e01722, 2021. doi: 10.1186/s12911-021-01722-4.
16. Brennan Klein et al., "*Forecasting hospital-level COVID-19 admissions using real-time mobility data*," Communications in Medicine, vol. 3, no. 1, p. 25, Feb. 14, 2023, doi: 10.1038/s43856-023-00253-5.
17. Md Shahid Ansari et al., "*Identification of predictors and model for predicting prolonged length of stay in dengue patients*," *Health Care Management Science*, vol. 24, no. 4, pp. 786-798, Dec. 2021, doi: 10.1007/s10729-021-09571-3.
18. Pei-Fang Jennifer Tsai et al., "*Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network*," *Journal of Healthcare Engineering*, vol. 2016, p. 7035463 2016. doi: 10.1155/2016/7035463.
19. Franck Jaotombo et al., "*Machine-learning prediction for hospital length of stay using a French medico-administrative database*," *Journal of Market Access & Health Policy*, vol. 11, no. 1, p. 2149318, Nov. 26, 2022. doi: 10.1080/20016689.2022.2149318.
20. Scott Levin, S. Barnes, and M. Toerper, "*Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay*," *BMJ Innovations*, vol. 7, no. 2, Dec. 2020, Art. no. bmjinnov-2020-000420, doi: 10.1136/bmjinnov-2020-000420.
21. Lingtao Chen and H. B. Klasky, "*Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients with Sepsis: A Comparative Study*," in *IEEE* SoutheastCon 2022, New Jersey, United States of America, Apr. 2022, pp. 302-309, DOI:10.1109/SoutheastCon48659.2022.9764052.
22. Marta Wilk, D. W. R. Marsh, S. De Freitas, and J. Prowle, "*Predicting Length of Stay in Hospital Using Electronic Records Available at the Time of Admission*," *Stud Health Technol Inform*, vol. 270, pp. 191-195, 2020, doi: 10.3233/SHTI200186.
23. Deyo, R. A., Cherkin, D. C., & Ciol, M. A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *Journal of Clinical Epidemiology,45*(6),613-619.DOI:10.1016/0895-4356(92)90133-8.[Online].
Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540553/table/t1-amia_2012_symp_0809/.
24. Jesus Manuel Puentes Gutiérrez, M.-Á. Sicilia, S. Sánchez-Alonso, and E. García-Barriocanal, "*Predicting Length of Stay Across Hospital Departments*," *IEEE Access*, vol. 9, pp. 38946-38958, 2021. DOI: 10.1109/ACCESS.2021.3066562.
25. Ashwani Kumar and H. Anjomshoa, "*A Two-Stage Model to Predict Surgical Patients' Lengths of Stay From an Electronic Patient Database*," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 848-856, Mar. 2019. DOI: 10.1109/JBHI.2018.2819646.
26. Mahsa Rouzbahman, A. Jovicic, and M. Chignell, "*Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU*?," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 851-858, May 2017. DOI: 10.1109/JBHI.2016.2525731.
27. Chunling Li, L. Chen, J. Feng, D. Wu, Z. Wang, J. Liu, and W. Xu, "*Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator*, "IEEEAccess, vol.7, pp.108981-108987,2019.DOI: 10.1109/ACCESS.2019.2934166.
28. G. Thippa Reddy et al., "*An Ensemble-based Machine Learning model for Diabetic Retinopathy Classification*," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Feb. 2020, doi: 10.1109/ic-ETITE47903.2020.235.

29. Artittayapron Rojarathand W. Songpan, "*Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems*," Applied Intelligence, vol. 51, no. 8, pp. 3692-3707, Jul. 2021. DOI: 10.1007/s10489-020-02106-3.
30. Isha Gandhi and M. Pandey, "Hybrid Ensemble of classifiers using voting," in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Oct. 2015, DOI: 10.1109/ICGCIoT.2015.7380496.
31. Ramya Akula, N. Nguyen, and I. Garibay, "*Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes*," presented at the 2019 SoutheastCon, Huntsville, AL, USA,11-14April2019.DOI: 10.1109/SoutheastCon42311.2019.9020358.
32. Patrick J. McCormick, H. Lin, S. G. Deiner, and M. A. Levin, "*Validation of the All Patient Refined Diagnosis Related Group (APR-DRG) Risk of Mortality and Severity of Illness Modifiers as a Measure of Perioperative Risk*," J. Med. Syst., vol. 42, no. 5, p. 81, Mar. 2018, doi: 10.1007/s10916-018-0936-3.
33. Yazan Alnsour, M. Johnson, A. Albizri, and A. Harfouch, "*Predicting Patient Length of Stay Using Artificial Intelligence to Assist Healthcare Professionals in Resource Planning and Scheduling Decisions*," J. Global Inf. Manage., vol. 31, no. 1, 2023. [Online]. Available: https://orcid.org/0000-0001-7198-7197. DOI: 10.4018/JGIM.323059.
34. Raunak Jain, M. Singh, A. R. Rao, and R. Garg, "*Machine Learning Models To Predict Length Of Stay In Hospitals*," presented at the 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Jun. 11-14, 2022, doi: 10.1109/ICHI54592.2022.00105.
35. Guilan Kong, K. Lin, and Y. Hu, "*Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU*," BMC Medical Informatics and Decision Making, vol. 20, no. 1, p. 251, Oct. 2020, doi: 10.1186/s12911-020-01271-2.
36. Javad Zarei, A. Jamshidnezhad, M. H. Shoushtari, A. M. Hadianfard, M. Cheraghi, and A. Sheikhtaheri, "*Machine Learning Models to Predict In-Hospital Mortality among Inpatients with COVID-19: Underestimation and Overestimation Bias Analysis in Subgroup Populations*," Journal of Healthcare Engineering, vol. 2022, Art. no. 1644910, Jun. 2022, doi: 10.1155/2022/1644910.
37. Mehmet Kivrak, E. Guldogan, and C. Colak, "*Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods*," Computer Methods and Programs in Biomedicine, vol. 202, Art. no. 105951, Feb. 2021, doi: 10.1016/j.cmpb.2021.105951.
38. Dagmar F. Hernandez-Suarez, Y. Kim, P. Villablanca, T. Gupta, J. Wiley, B. G. Nieves-Rodriguez, J. Rodriguez-Maldonado, R. F. Maldonado, I. da Luz Sant'Ana, C. Sanina, P. Cox-Alomar, H. Ramakrishna, A. Lopez-Candales, W. W. O'Neill, D. S. Pinto, A. Latib, and A. Roche-Lima, "*Machine Learning Prediction Models for In-Hospital Mortality After Transcatheter Aortic Valve Replacement*," JACC: Cardiovascular Interventions, vol. 12, no. 16, pp. 1508-1518, Jul. 2019, doi: 10.1016/j.jcin.2019.06.013.
39. Jun Ke, Y. Chen, X. Wang, Z. Wu, Q. Zhang, Y. Lian, and F. Chen, "*Machine learning-based in-hospital mortality prediction models for patients with acute coronary syndrome*," American Journal of Emergency Medicine, vol. 51, pp. 205-212, Dec. 2021, doi: 10.1016/j.ajem.2021.12.070.
40. Chi-Yung Cheng, C.-T. Kung, F.-C. Chen, I.-M. Chiu, C.-H. R. Lin, C.-C. Chu, C. F. Kung, and C.-M. Su, "*Machine learning models for predicting in-hospital mortality in patients with sepsis: Analysis of vital sign dynamics*," Frontiers in Medicine, vol. 9, pp. 964667, Nov. 2022, doi: 10.3389/fmed.2022.964667.
41. Mostafa Shanbehzadeh, R. Nopour, and H. Kazemi-Arpanahi, "*Design of an artificial neural network to predict mortality among COVID-19 patients*," Informatics in Medicine Unlocked, vol. 29, p. 100983, Jun. 2022, doi: 10.1016/j.imu.2022.100983.

42. Tatiana Ledneva, "*In-Hospital/30-Day Stroke Mortality Risk Adjusted Rate Development, SPARCS, 2013*," New York State Department of Health, 2016. [Online]. Available: https://www.health.ny.gov/statistics/sparcs/docs/in_hospital_30_day_stroke_mortality_risk_adjusted_rate.pdf.
43. Yan Zhang, W. Xu, P. Yang, and A. Zhang, "*Machine learning for predicting sepsis-related death: a systematic review and meta-analysis*," BMC Med. Inform. Decis. Mak., vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12911-023-02383-1.
44. C. Sunil Kumar, A. Govardhan, and B. Sunil Srinivas, "*Data Mining Issues and Challenges in Healthcare Domain*," Int. J. Eng. Res. Technol. (IJERT), vol. 3, no. 1, pp. 1-6, Jan. 2014, ISSN: 2278-0181.
45. "Role of Data Mining in Healthcare," Online Manipal, https://www.onlinemanipal.com/blogs/role-of-data-mining-in-healthcare:
46. "Tabular Data in the Age of AI," Kaggle, https://www.kaggle.com/code/kobbiemanrique/tabular-data-in-the-age-of-ai.
47. J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques, 3rd Edition," Morgan Kaufmann, 2011. Available: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.
48. Xu-Ying Liu, J. Wu, and Z.-H. Zhou, "*Exploratory Undersampling for Class-Imbalance Learning*," IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539-550, Apr. 2009, doi: 10.1109/TSMCB.2008.2007853.
49. C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "*RUSBoost: A Hybrid Approach to Alleviating Class Imbalance*," IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 40, no. 1, pp. 185-197, Jan. 2010, doi: 10.1109/TSMCA.2009.2029559.
50. "Overcoming Class Imbalance with SMOTE," Train in Data Blog, https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote.
51. "Decision Tree," GeeksforGeeks, https://www.geeksforgeeks.org/decision-tree.
52. S. Mishra, "Handling Imbalanced Data: SMOTE vs. Random Undersampling," Int. Res. J. Eng. Technol. (IRJET), vol. 4, no. 8, pp. 317, Aug. 2017, e-ISSN: 2395-0056, p-ISSN: 2395-0072.
53. "Undersampling Algorithms for Imbalanced Classification," Machine Learning Mastery, https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/.
54. "imblearn.ensemble.BalancedRandomForestClassifier,"imbalanced-learn, https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedRandomForestClassifier.html
55. "Hospital Inpatient Discharges - SPARCS De-Identified," New York State Department of Health, Health Data NY, https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-z7e7
56. S. A. Fayaz, M. Zaman, S. Kaul, and M. A. Butt, "*Is Deep Learning on Tabular Data Enough? An Assessment*," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 13, no. 4, p. 466, 2022. [Online]. Available: http://www.ijacsa.thesai.org.
57. R. Shwartz-Ziv and A. Armon, "*Tabular data: Deep learning is not all you need*," Information Fusion, vol. 81, pp. 84-90, May 2022. doi: 10.1016/j.inffus.2021.11.011.
58. R. H. Alvi et al., "*Deep learning approach on tabular data to predict early-onset neonatal sepsis*," J. Inf. Telecommun., vol. 5, no. 3, pp. 1-21, Dec. 2020, doi: 10.1080/24751839.2020.1843121.

59. S. Kumari, D. Kumar, and M. Mittal, "*An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier*," Int. J. Cogn. Comput. Eng., vol. 2, pp. 40-46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.