Predictive Modeling of Readmission in Patients with Schizophrenia Using Machine Learning Models

MSc Research Project
Data Analytics

Daphne Shekinah Tennison Daniel
Student ID: X23190027

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Daphne Shekinah Tennison Daniel
Student ID:	X23190027
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	12/12/2024
Project Title:	Predictive Modeling of Readmission in Patients with Schizophrenia Using Machine Learning Models
Word Count:	4784
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Daphne Shekinah Tennison Daniel
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modeling of Readmission in Patients with Schizophrenia Using Machine Learning Models

Daphne Shekinah Tennison Daniel
X23190027

Abstract

Schizophrenia is a mental disorder which is chronic, and it affects a person's ability to grasp reality. It not only affects the person who's schizophrenic but also the people surrounding the said person. There are more than 20 million people who are affected by schizophrenia but not many people get the required treatment. This leads to hospitalization and re-hospitalization of patients. There are many rules and regulations set in place for the treatment and the discharge of a patient in different countries. Predicting the risk of a patient's hospital readmission would help not only the patient but also the healthcare professionals who are treating the patient, because learning about how the disorder is affecting the person and treating them with a specific type of care could prove to be helpful. To address this issue, many machine learning models were developed. But there aren't many papers or research done that was specific to Ireland. So, in this project, a dataset that was published by the National Psychiatric Inpatient Reporting System under the Health Research Board from the year 2006 to 2022 was used to predict the readmission of patients with schizophrenia and other related disorders. The proposed model, XGBoost classifier performed the best, even when there was quite a class imbalance. The model achieved an accuracy of 75.38%, recall of 81%, F1-Score of 76% and the precision of the model was 72%. If imputation had been used on the dataset instead of filling up the missing values, the accuracy of the model could've been increased by a significant amount. By using a dataset that does not have many missing values, and combining the model with SMOTE could improve the results drastically.

1 Introduction

Schizophrenia is a complex mental health disorder that affects a person, socially, economically, and personally. It affects the person severely because it affects the person's thought process and their perception of reality. Every year, many people are diagnosed with a new type of schizophrenia with different symptoms. Even though there had been a significant development in the way the disease is managed, there is still a difficulty in providing treatment to people affected globally. According to an article written by Steven Zauderer for Cross River Therapy, only one out of 3 people with schizophrenia recovers fully, even though there are a large number of options for treatment. [1] While nearly 24 million people in the world are affected by schizophrenia, only 33% of them receive proper treatment for it. Schizophrenia symptoms are more observant in the late stages of adolescence or early adulthood, and it can be seen that it affects mostly men.

[1] http://www.crossrivertherapy.com/research/schizophrenia-statistics

Patients diagnosed with schizophrenia have a higher chance of being readmitted to the hospital than patients with other diseases. This is due to the complex and chronic nature of the disorder. Patients with schizophrenia are often not given the right treatment that is meant for them. Since understanding of the workings and symptoms of the disease is changing every day. So, healthcare professionals find it harder to treat them, which in turn leads to patients being readmitted for schizophrenia or any other related mental health disorders.

Machine learning and deep learning models has played an important role for many years in predicting the readmissions of patients with schizophrenia or any kinds of diseases. Though there are many works published on the prediction of readmission of patients with schizophrenia, there were not many studies done on the disorder mainly based in Ireland. That is what is done in this research. Through this research, the question that was planned to be addressed was: What can be done to improve the accuracy of machine learning models in predicting the risk of hospital readmission of patients with schizophrenia? In order to answer the question, three machine learning models were employed on the dataset that helped arrive at a conclusion on which machine learning model would work best. From the results, the further engineering that could be done on the dataset to make the model perform more efficiently was also apparent.

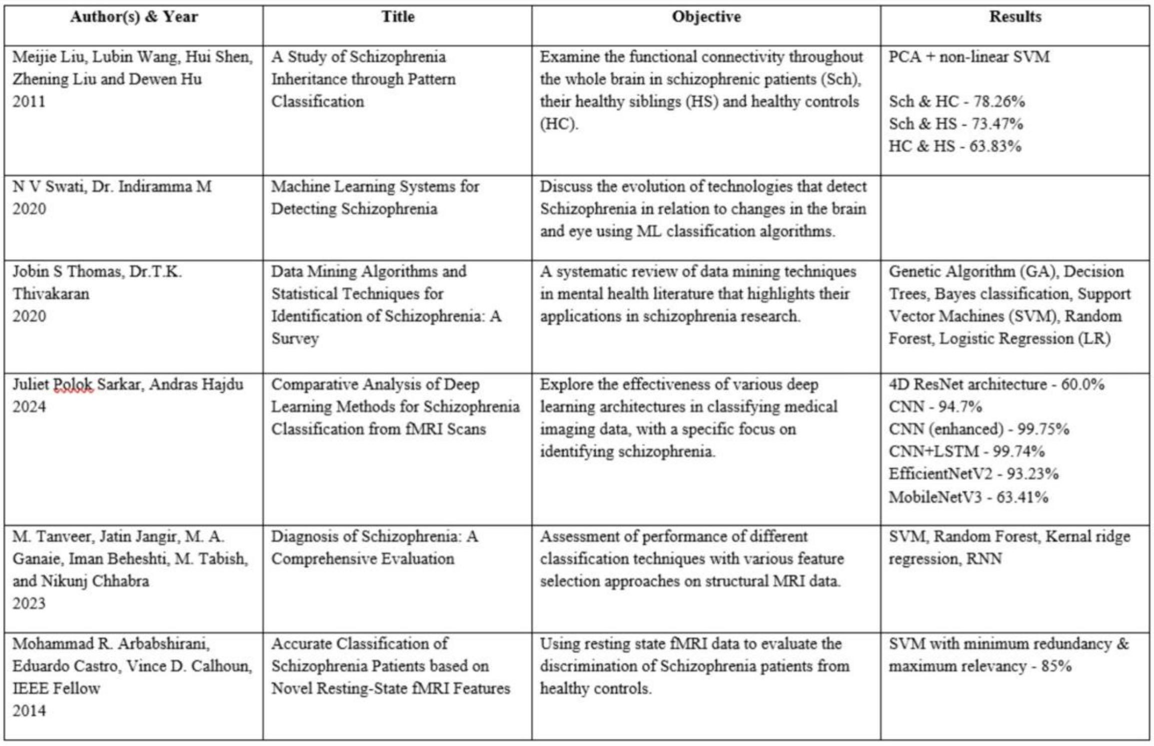
This research paper begins with the detailed review of the existing work done on the topics related to the research, the research methodology followed for the development of the machine learning models and the steps that are involved in the implementation of the machine learning models. Next, the evaluation and results section analyses the results and findings of the models implemented. The last section concludes the research and lays out the plans for future improvements of the research.

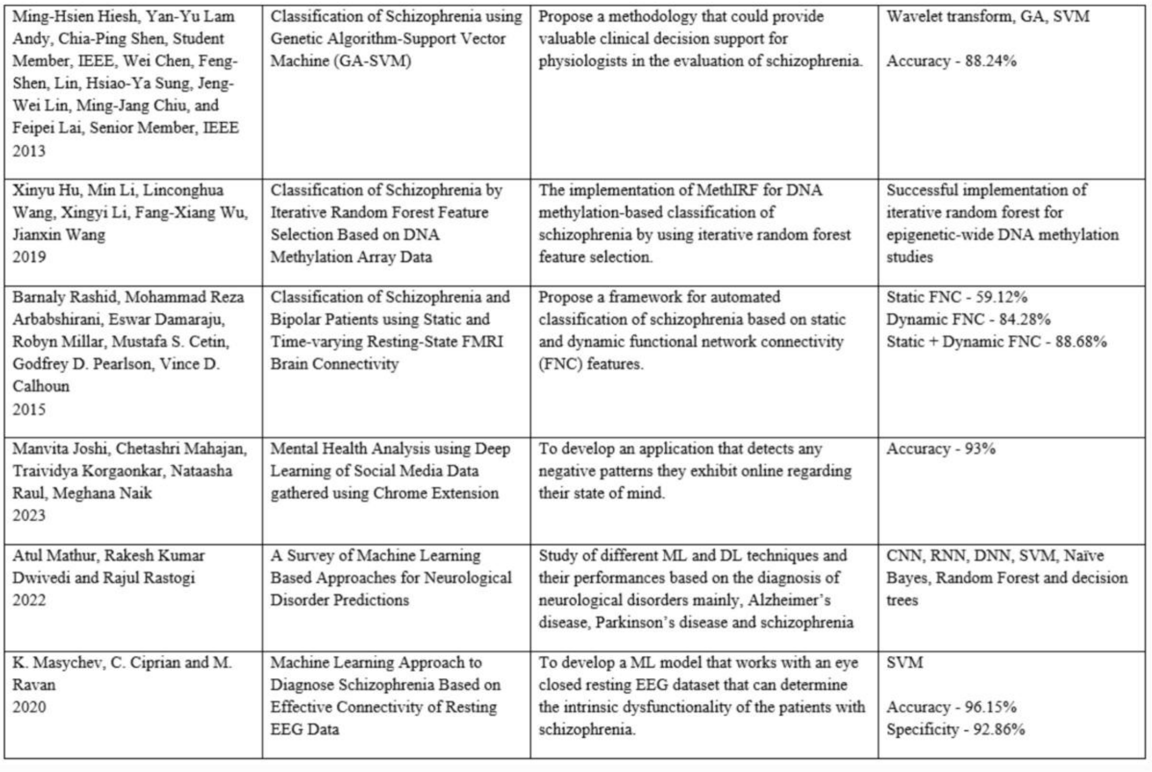
2 Related Work

This section provides an overview of the research that currently exist on the knowledge of using machine learning and deep learning models to predict the readmission of patients diagnosed with schizophrenia. In this review, studies that were published regarding the research topic from the year 2011 to 2024 are considered to gain a comprehensive understanding of the inner workings of the machine learning models and making them more efficient.

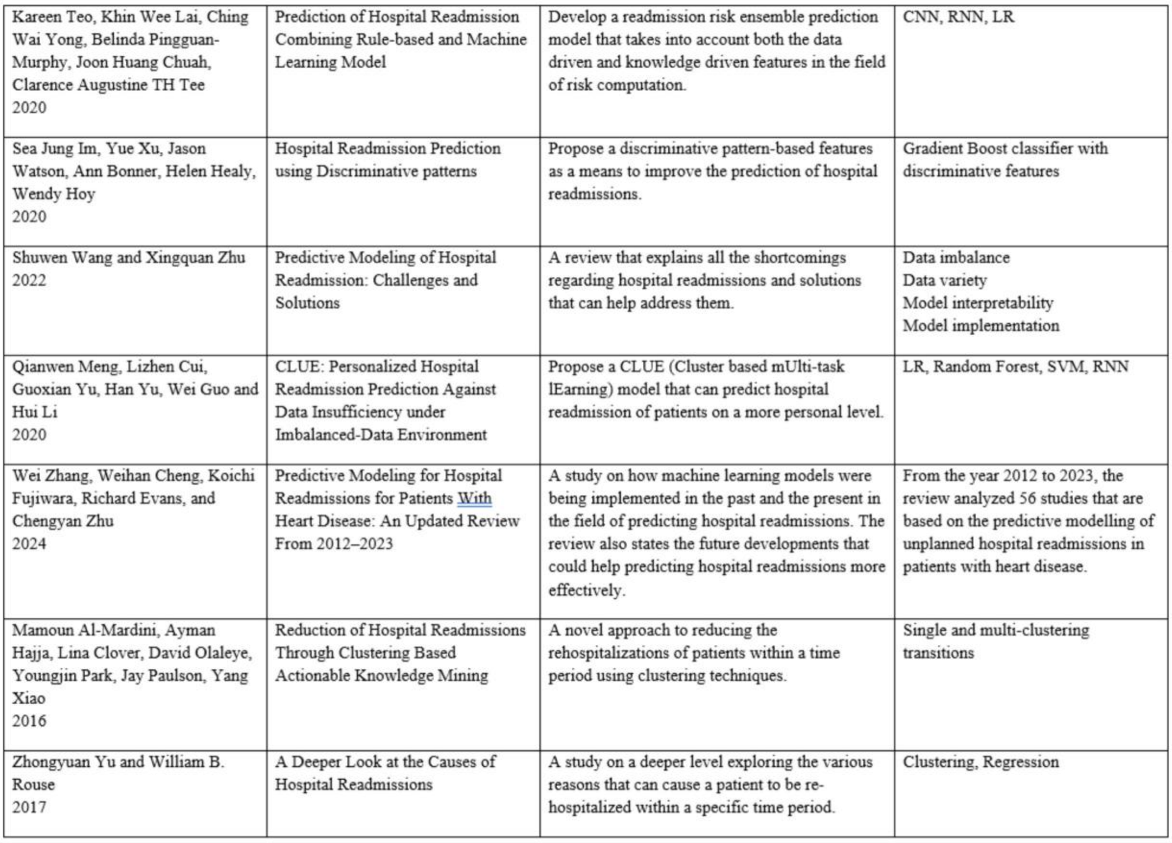
For identifying schizophrenia, THOMAS and THIVAKARAN (2020) did a detailed survey of all the data mining techniques used in the existing mental health literature. The functional connectivity of the brains of patients diagnosed with schizophrenia, their siblings and other healthy subjects were examined by Liu et al. (2011) did a study through pattern classification which had some interesting findings about the disease's inheritance. Swati and Indiramma (2020); Sharma et al. (2023); Sarkar and Hajdu (2024); Mathur et al. (2022) and Hiesh et al. (2013) made elaborate studies on the usage and effectiveness of machine learning and deep learning in the detection of schizophrenia. While different machine learning algorithms were discussed, Tanveer et al. (2023) did a comprehensive study of all the existing classification techniques to diagnose schizophrenia.

Using Functional Magnetic Resonance Imaging (FMRI) scans of patients diagnosed with schizophrenia, various machine learning and deep learning models, with the addition of different features, were trained to detect the disorder. (Sarkar and Hajdu; 2024; Arbabshirani et al.; 2014; Rashid et al.; 2015) From the research, it is apparent that electroencephalogram (EEG) signals are one of the most used way to detect schizophrenia with machine learning. (Masychev et al.; 2020; Qayyum et al.; 2023; Arias and Astudillo; 2023) It is possible to detect schizophrenia using DNA methylation data using Iterative Random Forest. (Hu et al.; 2019) A set of a selected few papers that were found on the topic are presented in the form of a table below for better and easy understanding.

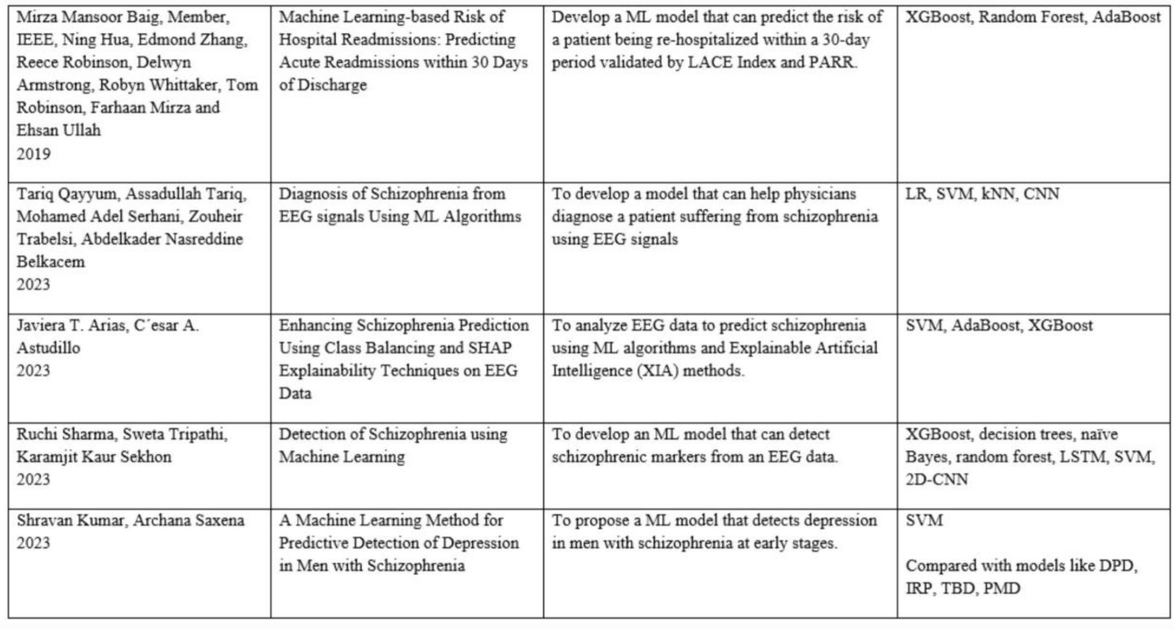




Author(s) & Year	Title	Objective	Results
Meijie Liu, Lubin Wang, Hui Shen, Zhening Liu and Dewen Hu 2011	A Study of Schizophrenia Inheritance through Pattern Classification	Examine the functional connectivity throughout the whole brain in schizophrenic patients (Sch), their healthy siblings (HS) and healthy controls (HC).	PCA + non-linear SVM Sch & HC - 78.26% Sch & HS - 73.47% HC & HS - 63.83%
N V Swati, Dr. Indiramma M 2020	Machine Learning Systems for Detecting Schizophrenia	Discuss the evolution of technologies that detect Schizophrenia in relation to changes in the brain and eye using ML classification algorithms.	
Jobin S Thomas, Dr.T.K. Thivakaran 2020	Data Mining Algorithms and Statistical Techniques for Identification of Schizophrenia: A Survey	A systematic review of data mining techniques in mental health literature that highlights their applications in schizophrenia research.	Genetic Algorithm (GA), Decision Trees, Bayes classification, Support Vector Machines (SVM), Random Forest, Logistic Regression (LR)
Juliet Polok Sarkar, Andras Hajdu 2024	Comparative Analysis of Deep Learning Methods for Schizophrenia Classification from fMRI Scans	Explore the effectiveness of various deep learning architectures in classifying medical imaging data, with a specific focus on identifying schizophrenia.	4D ResNet architecture - 60.0% CNN - 94.7% CNN (enhanced) - 99.75% CNN+LSTM - 99.74% EfficientNetV2 - 93.23% MobileNetV3 - 63.41%
M. Tanveer, Jatin Jangir, M. A. Ganaie, Iman Beheshti, M. Tabish, and Nikunj Chhabra 2023	Diagnosis of Schizophrenia: A Comprehensive Evaluation	Assessment of performance of different classification techniques with various feature selection approaches on structural MRI data.	SVM, Random Forest, Kernal ridge regression, RNN
Mohammad R. Arbabshirani, Eduardo Castro, Vince D. Calhoun, IEEE Fellow 2014	Accurate Classification of Schizophrenia Patients based on Novel Resting-State fMRI Features	Using resting state fMRI data to evaluate the discrimination of Schizophrenia patients from healthy controls.	SVM with minimum redundancy & maximum relevancy - 85%
Ming-Hsien Hiesh, Yan-Yu Lam Andy, Chia-Ping Shen, Student Member, IEEE, Wei Chen, Feng-Shen, Lin, Hsiao-Ya Sung, Jeng-Wei Lin, Ming-Jang Chiu, and Feipei Lai, Senior Member, IEEE 2013	Classification of Schizophrenia using Genetic Algorithm-Support Vector Machine (GA-SVM)	Propose a methodology that could provide valuable clinical decision support for physiologists in the evaluation of schizophrenia.	Wavelet transform, GA, SVM Accuracy - 88.24%
Xinyu Hu, Min Li, Linconghua Wang, Xingyi Li, Fang-Xiang Wu, Jianxin Wang 2019	Classification of Schizophrenia by Iterative Random Forest Feature Selection Based on DNA Methylation Array Data	The implementation of MethIRF for DNA methylation-based classification of schizophrenia by using iterative random forest feature selection.	Successful implementation of iterative random forest for epigenetic-wide DNA methylation studies
Barnaly Rashid, Mohammad Reza Arbabshirani, Eswar Damaraju, Robyn Millar, Mustafa S. Cetin, Godfrey D. Pearlson, Vince D. Calhoun 2015	Classification of Schizophrenia and Bipolar Patients using Static and Time-varying Resting-State FMRI Brain Connectivity	Propose a framework for automated classification of schizophrenia based on static and dynamic functional network connectivity (FNC) features.	Static FNC - 59.12% Dynamic FNC - 84.28% Static + Dynamic FNC - 88.68%
Manvita Joshi, Chetashri Mahajan, Traividya Korgaonkar, Nataasha Raul, Meghana Naik 2023	Mental Health Analysis using Deep Learning of Social Media Data gathered using Chrome Extension	To develop an application that detects any negative patterns they exhibit online regarding their state of mind.	Accuracy - 93%
Atul Mathur, Rakesh Kumar Dwivedi and Rajul Rastogi 2022	A Survey of Machine Learning Based Approaches for Neurological Disorder Predictions	Study of different ML and DL techniques and their performances based on the diagnosis of neurological disorders mainly, Alzheimer's disease, Parkinson's disease and schizophrenia	CNN, RNN, DNN, SVM, Naïve Bayes, Random Forest and decision trees
K. Masychev, C. Ciprian and M. Ravan 2020	Machine Learning Approach to Diagnose Schizophrenia Based on Effective Connectivity of Resting EEG Data	To develop a ML model that works with an eye closed resting EEG dataset that can determine the intrinsic dysfunctionality of the patients with schizophrenia.	SVM Accuracy - 96.15% Specificity - 92.86%



Kareen Teo, Khin Wee Lai, Ching Wai Yong, Belinda Pingguan-Murphy, Joon Huang Chuah, Clarence Augustine TH Tee 2020	Prediction of Hospital Readmission Combining Rule-based and Machine Learning Model	Develop a readmission risk ensemble prediction model that takes into account both the data driven and knowledge driven features in the field of risk computation.	CNN, RNN, LR
Sea Jung Im, Yue Xu, Jason Watson, Ann Bonner, Helen Healy, Wendy Hoy 2020	Hospital Readmission Prediction using Discriminative patterns	Propose a discriminative pattern-based features as a means to improve the prediction of hospital readmissions.	Gradient Boost classifier with discriminative features
Shuwen Wang and Xingquan Zhu 2022	Predictive Modeling of Hospital Readmission: Challenges and Solutions	A review that explains all the shortcomings regarding hospital readmissions and solutions that can help address them.	Data imbalance Data variety Model interpretability Model implementation
Qianwen Meng, Lizhen Cui, Guoxian Yu, Han Yu, Wei Guo and Hui Li 2020	CLUE: Personalized Hospital Readmission Prediction Against Data Insufficiency under Imbalanced-Data Environment	Propose a CLUE (Cluster based mUlti-task lEarning) model that can predict hospital readmission of patients on a more personal level.	LR, Random Forest, SVM, RNN
Wei Zhang, Weihan Cheng, Koichi Fujiwara, Richard Evans, and Chengyan Zhu 2024	Predictive Modeling for Hospital Readmissions for Patients With Heart Disease: An Updated Review From 2012–2023	A study on how machine learning models were being implemented in the past and the present in the field of predicting hospital readmissions. The review also states the future developments that could help predicting hospital readmissions more effectively.	From the year 2012 to 2023, the review analyzed 56 studies that are based on the predictive modelling of unplanned hospital readmissions in patients with heart disease.
Mamoun Al-Mardini, Ayman Hajja, Lina Clover, David Olaleye, Youngjin Park, Jay Paulson, Yang Xiao 2016	Reduction of Hospital Readmissions Through Clustering Based Actionable Knowledge Mining	A novel approach to reducing the rehospitalizations of patients within a time period using clustering techniques.	Single and multi-clustering transitions
Zhongyuan Yu and William B. Rouse 2017	A Deeper Look at the Causes of Hospital Readmissions	A study on a deeper level exploring the various reasons that can cause a patient to be re-hospitalized within a specific time period.	Clustering, Regression



Mirza Mansoor Baig, Member, IEEE, Ning Hua, Edmond Zhang, Reece Robinson, Delwyn Armstrong, Robyn Whittaker, Tom Robinson, Farhaan Mirza and Ehsan Ullah 2019	Machine Learning-based Risk of Hospital Readmissions: Predicting Acute Readmissions within 30 Days of Discharge	Develop a ML model that can predict the risk of a patient being re-hospitalized within a 30-day period validated by LACE Index and PARR.	XGBoost, Random Forest, AdaBoost
Tariq Qayyum, Assadullah Tariq, Mohamed Adel Serhani, Zouheir Trabelsi, Abdelkader Nasreddine Belkacem 2023	Diagnosis of Schizophrenia from EEG signals Using ML Algorithms	To develop a model that can help physicians diagnose a patient suffering from schizophrenia using EEG signals	LR, SVM, kNN, CNN
Javiera T. Arias, C´esar A. Astudillo 2023	Enhancing Schizophrenia Prediction Using Class Balancing and SHAP Explainability Techniques on EEG Data	To analyze EEG data to predict schizophrenia using ML algorithms and Explainable Artificial Intelligence (XIA) methods.	SVM, AdaBoost, XGBoost
Ruchi Sharma, Sweta Tripathi, Karamjit Kaur Sekhon 2023	Detection of Schizophrenia using Machine Learning	To develop an ML model that can detect schizophrenic markers from an EEG data.	XGBoost, decision trees, naïve Bayes, random forest, LSTM, SVM, 2D-CNN
Shravan Kumar, Archana Saxena 2023	A Machine Learning Method for Predictive Detection of Depression in Men with Schizophrenia	To propose a ML model that detects depression in men with schizophrenia at early stages.	SVM Compared with models like DPD, IRP, TBD, PMD

Joshi et al. (2023) developed a mobile application that detects any kinds of negative patterns exhibited by individuals on social media and it achieved 93% accuracy. In general, it can be seen that schizophrenia is diagnosed in men more than women. So, Kumar and Saxena (2023) created a machine learning model that detects depression in men in the early stages using Support Vector Machine (SVM). The re-admissions of men in the hospital for schizophrenia is almost always higher than that of women.

Now, on the note of re-admissions, these are some of the literature available on the usage of machine learning for the prediction of hospital readmission of patients with schizophrenia. The work done by Wang and Zhu (2022) explains in detail all the negative effects of re-admissions and solutions that can help overcome them. Likewise, Yu and Rouse (2017) did a study that explores the causes of re-hospitalization of patients within a specific time-frame on a deeper level. To predict the readmission of patients within a 30 day period, Baig et al. (2019) developed an ML model using XGBoost, Random Forest and AdaBoost. This study greatly contributed to the models chosen for this research.

An ensemble prediction model was developed by Teo et al. (2020) to predict the risk of readmission of patients. Likewise, Im et al. (2020) developed a machine learning model with discriminative features to predict hospital readmissions. Zhang et al. (2024) did an extensive study of 56 existing papers based on unplanned readmission of patients with heart disease. Last but not the least, Al-Mardini et al. (2016) proposed a novel approach to reduce the risk of readmission of patients with single and multi-clustering techniques.

Though, the abundance of current research existing on the topic is not less, the absence of research on the prediction of readmission of patients with schizophrenia based specifically in Ireland is the key motivator for this project. Building this section helped in choosing the best machine learning models that could be implemented in the prediction of hospital admissions and the prediction of readmission of patients diagnosed with schizophrenia and other related mental disorders. The research methodology was formed with the help of the knowledge gained from the literature review done.

3 Methodology

This section details the dataset that was chosen for this research and the steps undertaken to implement the chosen machine learning algorithms in a technical manner. The trends and patterns found in the dataset is also presented in graphs. The pictorial representation of the methodology followed in the research is given in Figure 1.

3.1 Dataset

The dataset chosen for this research is taken from the open data platform of the Central Statistics Office of India. The dataset is published by the National Psychiatric Inpatient Reporting System (NPIRS) under the Health Research Board. [2]It is a statistics dataset which contains the number of hospital admissions of patients for various reasons including schizophrenia and other schizophrenia related diseases from the year 2006 to 2022. The hospital admission statistics in the dataset include the diseases that are classified using the ICD-10 Diagnostic Group. The sex of the patients was also given, this helped in making better connections within the dataset. The dataset was downloaded from the website on October 31, 2024, before the changes were made to represent ICD-11 Diagnostic Group with more modern terminology till the year 2023. The website was last updated on November 7, 2024.

3.2 Dataset Cleaning

The dataset has 4,284 rows and 7 columns, which includes the one numeric column called VALUE. The other columns found in the dataset are year, type of admission, sex, dia-

[2] Dataset: http://data.cso.ie/table/HRA05

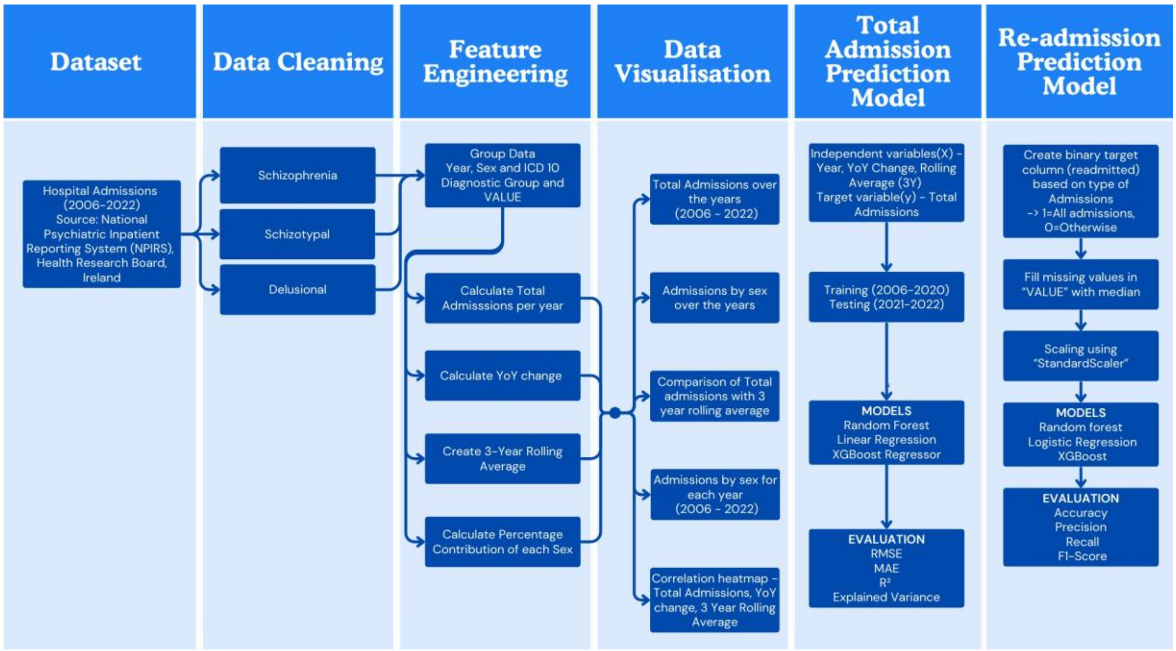

Dataset
Data Cleaning
Feature Engineering
Data Visualisation
Total Admission Prediction Model
Re-admission Prediction Model
Hospital Admissions (2006-2022) Source: National Psychiatric Inpatient Reporting System (NPIRS), Health Research Board, Ireland
Schizophrenia
Schizotypal
Delusional
Group Data Year, Sex and ICD 10 Diagnostic Group and VALUE
Calculate Total Admisssions per year
Calculate YoY change
Create 3-Year Rolling Average
Calculate Percentage Contribution of each Sex
Total Admissions over the years (2006 - 2022)
Admissions by sex over the years
Comparison of Total admissions with 3 year rolling average
Admissions by sex for each year (2006 - 2022)
Correlation heatmap - Total Admissions, YoY change, 3 Year Rolling Average
Independent variables(X) - Year, YoY Change, Rolling Average (3Y) Target variable(y) - Total Admissions
Training (2006-2020) Testing (2021-2022)
MODELS Random Forest Linear Regression XGBoost Regressor
EVALUATION RMSE MAE R² Explained Variance
Create binary target column (readmitted) based on type of Admissions -> 1=All admissions, 0=Otherwise
Fill missing values in "VALUE" with median
Scaling using "StandardScaler"
MODELS Random forest Logistic Regression XGBoost
EVALUATION Accuracy Precision Recall F1-Score

Figure 1: Methodology Diagram

gnostic group, units and admission values. However, there are around 1500 missing entries in the VALUE column. This required data preprocessing. The dataset was extracted for rows containing keywords such as ”Schizophrenia,” ”schizotypal,” and ”delusional.” This filtered the dataset to only the data relevant for this research, which helps on focussing the analysis on schizophrenia related diseases.

As the next step, the missing values in the VALUE column were replaced with 0, which meant that it is taken that there were no hospital admissions in that particular year. While this might not be an ideal approach, this simplifies the analysis process without the risk of introducing outliers. This cleaned data provided a structured dataset for visualisation and predictive modelling. The cleaned dataset also ensures that only relevant and meaningful data was used in the process. Even though feature engineering was done on this dataset, there was still an imbalance in the data. To select a ML model, that would be able to handle it, a study done using a CLUE model proposed Random Forest and Logistic Regression. (Meng et al.; 2020)

3.3 Exploratory Data Analysis and Data Visualizations

To understand the trends and patterns of the dataset, exploratory data analysis was used, and multiple visualizations were drawn up to represent the trends found in the hospital admissions from the year 2006 to 2022 for various reasons across Ireland. To understand the trends of yearly admissions, a line plot was used to display the number of admissions over the years 2006 to 2022 for schizophrenia related issues.

From Figure 2, it is apparent that the hospital admissions of patients with schizophrenia related issues have not been consistent. It can be seen that the admissions peaked in the year 2009 and there’s a significant decline in the year 2010. The lowest admission rate is recorded in the year 2017. The changes in Year-over-Year (YoY) is also calculated in order to highlight the shift in the admission trends over the years.

To compare the male and female admissions over the years, a line plot that segments the differences in admission rates by gender is plotted. The graph provides valuable insights into the admission trends in each gender both together and separately.

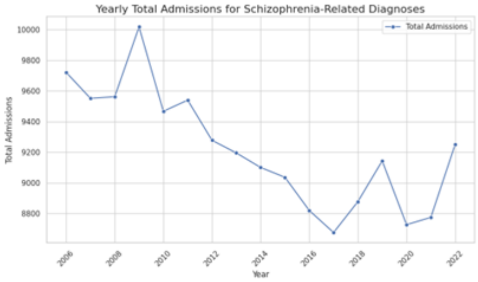


Figure 2: Total Admissions in Each Year

It can be seen that the admission rates of men are consistently higher than that of women. It can also be seen that the admission rates of both men and women exhibit similar patterns in increase and decrease over the years, while the total admission rates show minor fluctuations over the years. The shaded area in Figure 3 representing both sexes highlights regular fluctuations, that could possibly suggest clinical patterns in admission rates.

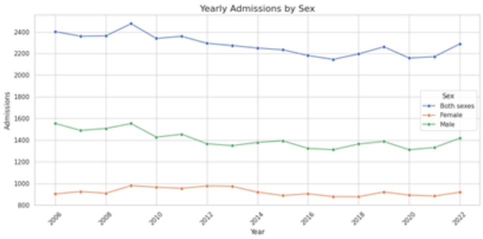


Figure 3: Total Admissions Each Year by Gender

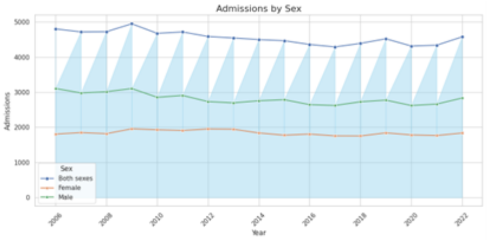


Figure 4: Admissions in Each Year by Gender

Since there are some fluctuations in the admission rates, a three-year rolling average is calculated to smoothen out the fluctuations. This helps in the detection of long term trends and also helps filter out the noise in the admissions data. The rolling average is plotted alongside total admissions, and it is seen that the line plot of rolling average is smoother than the total admissions plot.

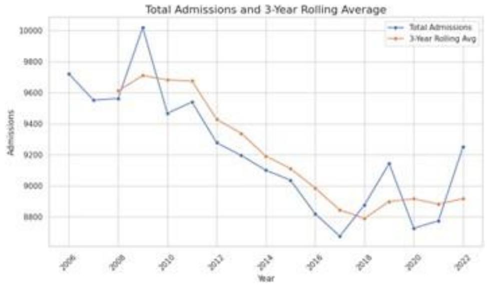


Figure 5: Total Admissions and 3-Year Rolling Average

The percentage of admission rates of both men and women contributing to the total admission rates were calculated for every year. This graph shows the categorized hospital admissions by gender from each year from 2006 to 2022. Each bar in the graph represents the admissions data that indicates the inter-quartile range, and the whiskers show the variability in the admissions rate which captures the full range of admissions data except the outliers. While fluctuations can be seen in the range and median values over the years, the distribution within the categories remain relatively consistent.

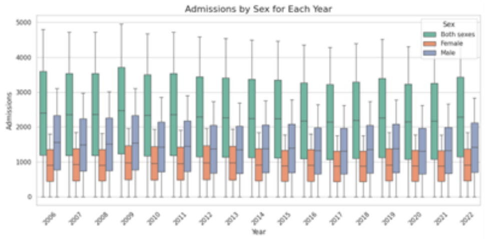


Figure 6: Admissions by Sex in Each Year

To visualise the correlations between variables namely total admissions, year, YoY changes and the rolling average, a heatmap is generated. This heatmap highlights the relationship between the variables. Total admissions and rolling average exhibit strong correlation and a weak correlation is seen between admissions and Year-over-Year changes.

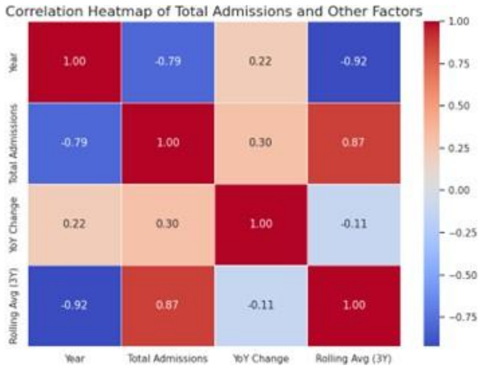


Figure 7: Correlation Heatmap

3.4 Feature Engineering

In order to enhance the dataset's predictive efficiency, the following features were added: 3-year Rolling Average, Year-over-Year Changes and Gender Contribution. The 3-year rolling average helped smoothen out the year-to-year fluctuations and also capture some of the underlying trends. The rolling average is the sum of any three years' number of admissions divided by three. The YoY changes is the difference in the number of admissions between the consecutive years. This greatly helps in highlighting the sharp increases or decreases in the admission trends. The percentage contribution of each gender to the total admissions helped analyse the relative dynamics between men and women patients with schizophrenia. This could be considered a valuable information for understanding gender-based inconsistencies in the trends of hospitalization of patients with schizophrenia.

The models for the prediction of readmission were chosen in such a way that there can be a comprehensive comparison of both linear and ensemble-based machine learning techniques for binary classification of the dataset.

4 Design Specification and Implementation

This section of the report gives the detailed explanation of the implementation of the research that has been proposed and how the question was answered using distinctive features of the machine learning model. This project is mainly focussed on analysing the hospital admission trends related to the diagnoses of schizophrenia in Ireland. After analysing the admission trends, predictive models for hospital readmissions and total admissions were developed. The proposed methodology is being implemented in steps involving data cleaning, visualising the trends and patterns, feature engineering and modelling predictive analysis to arrive at meaningful insights and build models to predict readmission. The code was run in Google Colab. The implementation process is explained in detail below.

4.1 Total Admissions Prediction Model

For the estimation of the total admissions of patients who are diagnosed with schizophrenia related diseases, predictive modelling is used on the historical data recorded in Ireland. To make the dataset fit to implement the predictive models, the dataset is split into two. The first set is from the year 2006 to 2020 and the second set is the rest of the data from the year 2021 to 2022. The first set is considered as the training set, so that the models are trained on the admission data from the earlier years. The models were made to associate with the relationships between the features like the 3-year rolling average and the year of admission, and total admissions. The second set is taken as the testing set, so the model can evaluate on the recent and undiscovered data. This helped create a simulation on how the model performs on future data that is practically unforeseen. For the prediction of total admissions, three regression models were trained and implemented namely, Linear Regression, Random Forest Regression and XGBoost Regressor.

4.1.1 Linear Regression

This model helps to create a linear relationship between the predictor (x_train) and the target variable (y_train). The coefficients of each of the predictors is estimated by this model. By doing so, the residual sum of squares between the predicted values and the observed values is minimized. With the use of LinearRegression from the sklearn library, this model is implemented. This model required very minimal parameter tuning. This makes this model an ideal baseline for the comparison of other models. This model is ideal for this particular data because, the dataset has numerical features that shows linear trends.

4.1.2 Random Forest

Random Forest is an ensemble learning method that makes more than one decision tress in every step and makes the best decision in each step to reduce over-fitting. This also improves the model's ability to generalize. The model was initialized with 100 estimators. In every step, the tree is trained on a specific subset of the data, which makes it sensitive to minute variations and noise. The parameter, random_state guarantees reproducibility by fixing the random seed. While the previous model was suitable for linear relationships, random forest is better suited for non-linear interactions between the features selected, which makes it viable for various distributions of data.

4.1.3 XGBoost

This model, like the Random Forest is an ensemble model based on decision trees, which makes predictions by reducing errors in every step. To prevent over-fitting, XGBoost combines the strength of sequential improvement from gradient boosting and regularization. From xgboost library, XGBRegressor is used to implement the model. For regression tasks, objective="reg:squarederror" parameter is used and to limit the number of boosting rounds to a 100, n_estimators=100 is used. XGBoost is chosen for its ability to successfully handle datasets that have complex relationships between the features. XGBoost can also efficiently handle both linear and non-linear patterns in the dataset.

4.2 Readmissions Prediction Model

Based on the admission related features that were collected in the previous steps, three machine learning models were implemented on the modified dataset to predict the risk of a patient being readmitted to the hospital for schizophrenia related diseases. The models that were implemented on the dataset are Logistic Regression, Random Forest and XGBoost. To prepare the dataset to predict the readmission of patients, the target variable needed to be created by transforming the Type of Admission column. The name of the target variable was Readmitted. The target variable was assigned a binary label. 1 for the type of admission being "All admissions" and 0 for all the other entries. The independent variables are Year which apprehends the temporality of the data, Sex which contains the gender of the patients being admitted, ICD 10 Diagnostic Group which is a categorical column that represents the diagnostic groups, out of which schizophrenia related diseases were selected and VALUE which is the numerical column of the dataset containing the number of patients being admitted in a particular year for a particular type of disease.

Before the models were trained to predict hospital readmissions, all the missing values in the numerical column, VALUE were replaced with the median values of the column. The categorical columns like Sex and ICD 10 Diagnostic Group, the missing values were entered as “Unknown”. Then OneHotEncoder was used to transform all the categorical features into their numerical form, which makes the dataset ready for machine learning models. StandardScaler from scikit-learn was used for feature scaling. By doing this, it is ensured that the mean and standard deviation of all the numerical values is 0 and 1 respectively. This step is crucial for algorithms like Logistic Regression and XGBoost, since they are sensitive to feature magnitudes. To achieve consistency, the same scaler function was used on the training data. Then it was applied to both the datasets, training and testing.

4.2.1 Logistic Regression

Logistic Regression is a linear machine learning model which is widely used for binary classification. It creates a relationship of linear nature between the target and the independent variables. To implement this model, solver=’liblinear’ was used, this is learnt to be best suited for building simple models. The maximum number of iterations was set as 1000, so the model has sufficient iterations to arrive at the solution. The regularization parameter is set to 1.0, this helps prevent overfitting. To train the model, the scaled training data (x_train_scaled) and the target data (y_train) were used. By minimizing the log-loss function, this model helps optimise the input features’ coefficients. The model that is trained was used to predict the binary class (y_pred) on the scaled test data (x_test_scaled). This model serves as the baseline model in this prediction analysis.

4.2.2 Random Forest

Random Forest is an ensemble learning method that creates multiple decision trees in every step and uses the average of the outputs to make predictions. This model is a robust machine learning model, and it also helps reduce overfitting. Just like the random forest that’s modelled for the total admissions, the total number of trees in the forest is specified as a 100. This step helps balance the computational cost and the prediction accuracy. The random state is declared as 42. This controls the randomness of the data in feature selection and bootstrapping. Just like the logistic regression model, this model is also trained on the same training data, with the same scaler. The predictions are generated under the name rf_y_pred for the scaled dataset just like the previous model. This model made apparent the non-linear relationships between the features without needing to tune the parameter at an extensive amount.

4.2.3 XGBoost

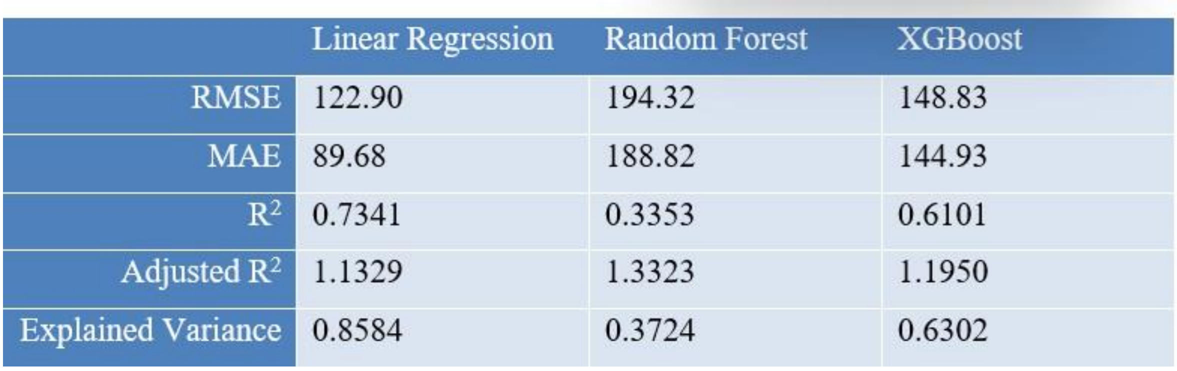
This is a gradient boosting algorithm that sequentially build decision trees and corrects the errors that were made by the previous trees. This model also adds regularization techniques to prevent overfitting of the data. In the implementation of the model, it was assigned to perform binary classification with logistic loss. The maximum number of iterations were set as 100 like the previous models. This is a flexible model, which makes it ideal for handling both linear and non-linear relationships in the data effectively.

The dataset was pre-processed consistently for all the models and the features were scaled in a similar way, so the differences that are seen in the performance of the models is attributed to the models' characteristics rather than the data preparation. The baseline for the prediction is Logistic Regression, the non-linear relationships were captured by Random Forest and XGBoost was used for its ability of sequential boosting to achieve high accuracy.

5 Evaluation

For evaluating the regression models that were used for the prediction of total admissions, four important evaluation metrics were used and they are: Root Mean Squared Error (RMSE) to measure the mean of the errors, Mean Absolute Error (MAE) to calculate the mean of the difference between the prediction and actual values, Coefficient of Determination (R^2) to show how the model evaluates the variance in the target variable and Explained Variance Score (EVS) shows the models' ability to evaluate the variability of the data.

Out of the three models used, linear regression performed the best with an RMSE and MAE of 122.90 and 89.68, respectively. This suggests that the model has the ability to predict total admissions most accurately. The R^2 of the model is the closest to one, 0.7341 and the EVS is 0.8584. RMSE and MAE of XGBoost is 148.83 and 144.93, respectively. The model explained around 61% of the variance of the data. Though the performance of the model is not better than linear regression, the model's ability to handle complex patterns might prove useful in datasets of large volumes.



	Linear Regression	Random Forest	XGBoost
RMSE	122.90	194.32	148.83
MAE	89.68	188.82	144.93
R^2	0.7341	0.3353	0.6101
Adjusted R^2	1.1329	1.3323	1.1950
Explained Variance	0.8584	0.3724	0.6302

Figure 8: Results of Total Admissions Model

The RMSE of random forest is the highest when compared to the other three models, which proves that the model performed poorly with this dataset. The R^2 suggests that the model only captured 33% of the variance in the data. The mediocre performance of the model could be attributed to the reason that the size of the dataset was small, and the features were limited. In conclusion, linear regression outperformed the other two models. This could be because of the linearity of the data. XGBoost also worked better than the random forest model. Radom forest found it hard to perform on this dataset amidst its robustness possibly because of overfitting.

The evaluation metrics used on the classification models for the prediction of hospital readmission of patients with schizophrenia are Accuracy, Precision, Recall, F1-Score. The

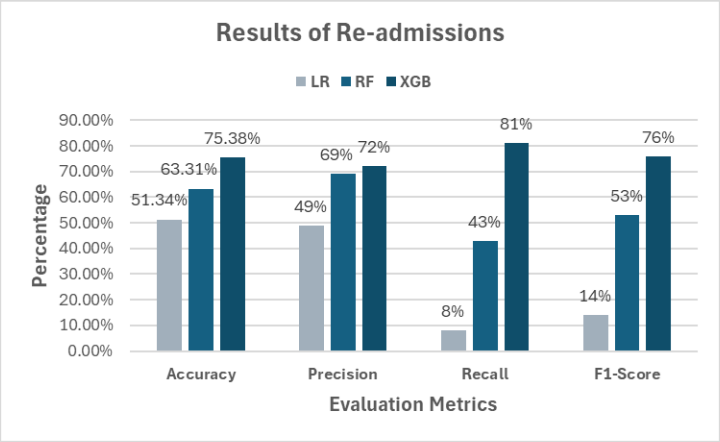


Figure 9: Results of Re-admissions Model

confusion matrix of all the models were also printed as part of the evaluation. Of all the readmission models, XGBoost performed the best when compared to other classification models. It produced the highest accuracy of 75.4% with a good precision and recall of 72% and 81%, respectively for the prediction of readmissions. The F1-Score of both the classes combined, non-readmissions and readmissions, was 75%. This suggests that XGBoost could be used in other larger datasets to predict readmissions of patients.

The model that least accurately predicted readmissions is Logistic Regression, with an accuracy of 51.3%. There was a noticeable difference in the model's ability to predict readmissions and non-readmissions. The recall for Class 0, which is non-readmissions was 92% while the recall for readmissions is only 8%. It can be concluded that logistic regression was not best suited for this dataset because of the imbalance in class distribution. Though Random Forest performed better than logistic regression with an accuracy of 63.1%, it was not an ideal prediction model for this dataset. The F1-Score of the model for predicting the majority class, that is the non-readmissions class is 70%. However, the F1-Score of the minority class, the readmissions is 53% with a lot of room for improvement.

6 Conclusion and Future Work

The dataset that was chosen for this research had quite a few missing values which showed inconsistent patterns in the trends regarding the total admissions. But from the data available it could be seen that the number of men admitted to the hospital for schizophrenia and other schizophrenia related diseases were higher than the number of women admitted for the same purposes. Since the patterns found in the total admissions were fairly linear, linear regression performed the best when it came to predicting total admissions of patients. For the prediction of readmission, classification models were used, because those models tend to handle datasets with imbalance in classes. Due to the lack of features in the dataset, the prediction of readmission proved to a difficult task and XGBoost outperformed the other classification models.

Using a balanced dataset could help draw up more conclusions based on the data recorded, which in turn will help determine the readmission risks more effectively. With the incorporation of socio-demographic factors and the characteristic of the hospitals, the analysis of readmission of patients can be improved prominently. To handle the missing values, simple filling methods were used in this project. If class imbalance is an issue, techniques like resampling or SMOTE (Synthetic Minority Over-sampling Technique) could be to used handle it. Instead of filling methods, the usage of imputation methods can prove to be more efficient. To implement non-linear regression or classification tasks, neural networks can be used, so that even the complex patterns and trends found in the dataset can be explored. Hyperparameter tuning can be used on random forest and XGBoost to optimize the performance of the models.

Despite the dataset chosen being an unbalanced one, with the help of data preprocessing and feature engineering, the prediction of readmission of patients with schizophrenia based on the information collected from Ireland was moderately successful. The XGBoost that was used performed fairly well with an accuracy of 75.4%. The visualisations that were drawn up from the dataset provided valuable insights into the hospital admissions of patients with schizophrenia and other related mental health disorders in Ireland. Even though there were challenges like missing data was encountered, the results show that with the addition of more information and further feature engineering, the accuracy of machine learning models can be improved to predict the risk of hospital readmission in patients with schizophrenia.

References

Al-Mardini, M., Hajja, A., Clover, L., Olaleye, D., Park, Y., Paulson, J. and Xiao, Y. (2016). Reduction of hospital readmissions through clustering based actionable knowledge mining, *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 444–448.

Arbabshirani, M. R., Castro, E. and Calhoun, V. D. (2014). Accurate classification of schizophrenia patients based on novel resting-state fmri features, *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6691–6694.

Arias, J. T. and Astudillo, C. A. (2023). Enhancing schizophrenia prediction using class balancing and shap explainability techniques on eeg data, *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pp. 1–5.

Baig, M. M., Hua, N., Zhang, E., Robinson, R., Armstrong, D., Whittaker, R., Robinson, T., Mirza, F. and Ullah, E. (2019). Machine learning-based risk of hospital readmissions: Predicting acute readmissions within 30 days of discharge, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2178–2181.

Hiesh, M.-H., Lam Andy, Y.-Y., Shen, C.-P., Chen, W., Lin, F.-S., Sung, H.-Y., Lin, J.-W., Chiu, M.-J. and Lai, F. (2013). Classification of schizophrenia using genetic algorithm-support vector machine (ga-svm), *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6047–6050.

Hu, X., Li, M., Wang, L., Li, X., Wu, F.-X. and Wang, J. (2019). Classification of schizophrenia by iterative random forest feature selection based on dna methylation array data, *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 807–811.

Im, S. J., Xu, Y., Watson, J., Bonner, A., Healy, H. and Hoy, W. (2020). Hospital readmission prediction using discriminative patterns, *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 50–57.

Joshi, M., Mahajan, C., Korgaonkar, T., Raul, N. and Naik, M. (2023). Mental health analysis using deep learning of social media data gathered using chrome extension, *2023 4th International Conference for Emerging Technology (INCET)*, pp. 1–8.

Kumar, S. and Saxena, A. (2023). A machine learning method for predictive detection of depression in men with schizophrenia, *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)*, pp. 1–6.

Liu, M., Wang, L., Shen, H., Liu, Z. and Hu, D. (2011). A study of schizophrenia inheritance through pattern classification, *2011 2nd International Conference on Intelligent Control and Information Processing*, Vol. 1, pp. 152–156.

Masychev, K., Ciprian, C. and Ravan, M. (2020). Machine learning approach to diagnose schizophrenia based on effective connectivity of resting eeg data, *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6.

Mathur, A., Dwivedi, R. K. and Rastogi, R. (2022). A survey of machine learning based approaches for neurological disorder predictions, *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, pp. 586–590.

Meng, Q., Cui, L., Yu, G., Yu, H., Guo, W. and Li, H. (2020). Clue: Personalized hospital readmission prediction against data insufficiency under imbalanced-data environment, *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 469–472.

Qayyum, T., Tariq, A., Serhani, M. A., Trabelsi, Z. and Belkacem, A. N. (2023). Diagnosis of schizophrenia from eeg signals using ml algorithms, *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2564–2570.

Rashid, B., Arbabshirani, M. R., Damaraju, E., Millar, R., Cetin, M. S., Pearlson, G. D. and Calhoun, V. D. (2015). Classification of schizophrenia and bipolar patients using static and time-varying resting-state fmri brain connectivity, *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 251–254.

Sarkar, J. P. and Hajdu, A. (2024). Comparative analysis of deep learning methods for schizophrenia classification from fmri scans, *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 69–74.

Sharma, R., Tripathi, S. and Sekhon, K. K. (2023). Detection of schizophrenia using machine learning, *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 505–512.

Swati, N. and Indiramma, M. (2020). Machine learning systems for detecting schizophrenia, *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 877–880.

Tanveer, M., Jangir, J., Ganaie, M. A., Beheshti, I., Tabish, M. and Chhabra, N. (2023). Diagnosis of schizophrenia: A comprehensive evaluation, *IEEE Journal of Biomedical and Health Informatics* **27**(3): 1185–1192.

Teo, K., Wee Lai, K., Wai Yong, C., Pingguan-Murphy, B., Huang Chuah, J. and Tee, C. A. T. (2020). Prediction of hospital readmission combining rule-based and machine learning model, *2020 International Computer Symposium (ICS)*, pp. 352–355.

THOMAS, J. S. and THIVAKARAN, T. (2020). Data mining algorithms and statistical techniques for identification of schizophrenia: A survey, *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 246–251.

Wang, S. and Zhu, X. (2022). Predictive modeling of hospital readmission: Challenges and solutions, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19**(5): 2975–2995.

Yu, Z. and Rouse, W. B. (2017). A deeper look at the causes of hospital readmissions, *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 919–923.

Zhang, W., Cheng, W., Fujiwara, K., Evans, R. and Zhu, C. (2024). Predictive modeling for hospital readmissions for patients with heart disease: An updated review from 2012–2023, *IEEE Journal of Biomedical and Health Informatics* **28**(4): 2259–2269.