National College of Ireland

Predictive Analytics for Patient Discharge Using Electronic Health Records

MSc Research Project
MSCDAD_JAN24A_O

Shrey Sanjay Kekade
Student ID: 23194316

School of Computing
National College of Ireland

Supervisor: Dr David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shrey Sanjay Kekade
Student ID:	23194316
Programme:	MSc Data Analytics **Year:** 2024
Module:	MSc Research Project
Supervisor:	Dr David Hamill
Submission Due Date:	12/12/2024
Project Title:	Predictive Analytics for Patient Discharge Using Electronic Health Records
Word Count:	8272 **Page Count** 23

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: Shrey Sanjay Kekade

Date: 11/12/2024

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 Analytics for Patient Discharge Using Electronic Health Records

Shrey Sanjay Kekade
x23194316@stuent.ncirl.ie

Abstract

This study investigates the application of machine learning methods to predict patient discharge outcomes by using EHR data, focusing on two critical challenges in healthcare resource management and patient care continuity. The investigation involved a feature selection and model interpretability to balance predictive accuracy with practical usability in clinical environments. This systematic methodology, consisting of data preprocessing, hyperparameter tuning, and performance assessment using metrics such as F1-score and AUC-ROC, compared a total of five machine learning models: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Neural Network. Of these, the Random Forest model proved to be the most robust, high in accuracy, and easy to interpret using SHAP-based explanations. While the obtained results are promising, the limitation of a single dataset and real-world validation pose a need for improvement and future research. However, this work contributes to predictive analytics in healthcare by providing a replicable framework that integrates advanced machine learning with domain-specific insights. Future work will expand dataset diversity, implement real-time predictive pipelines, and validate models in clinical settings to enhance their utility and scalability.

Keywords: Predictive Analytics, Patient Discharge, Machine Learning, Electronic Health Records (EHR), Healthcare Resource Management.

1. Introduction

The rising demands have pressed the use of predictive analytics to optimise patient care and resource allocation in all hospital settings. Of course, this is very important for the forecast of a patient's status at discharge status which considerably affects both the continuity of care and has important ramifications concerning hospital operations and resource utilisation. With Electronic Health Records (EHR) increasingly available (Ferrão et al., 2021), the ability to apply machine learning to high-dimensional patient data offers a method for improving the accuracy and actionability of discharge predictions. The present study aims to enhance patient discharge-related predictions by advanced machine learning models by focusing on feature selection and interpretability for practical usage in clinical environments.

Even with the broad acceptance of EHRs, issues related to the handling of high-dimensional data, imbalance problems in classes, and interpretability of the models, especially complex models like neural networks, persist. These are considered critical issues because, for real-world applications in healthcare, predictable accuracy and interpretability ensue, in which decisions have to be transparent and explainable to the professional physician (Zapata et al., 2023). The project consequently aims to solve these challenges by developing machine learning models that realize an excellent trade-off between performance and interpretability, contributing to improved outcomes for patients and better management within hospitals.

This inquiry focuses on evaluating the performance of several machine learning algorithms: logistic regression, decision trees, random forests, and neural networks for patients at discharge. Ultimately, the project works to provide healthcare professionals with reliable and interpretable tools to contribute step-by-step to improved decision-making processes regarding patient discharge. This enables resource management and assures better care.

The major research question that guides this project asks the question:

How can machine learning methods, combined with feature selection and explainability techniques, be used to accurately predict patient discharge outcomes using EHR data?

This project will contribute to the development and validation of predictive models that enhance the precision and usability of discharge predictions that improve patient flow and resource use and efficiency in healthcare.

The work is organized into five broad sections, namely Introduction, Literature Review, Methodology, Results, and Conclusion. Sequentially, these sections address the research objectives and various steps undertaken toward the eventual implementation and validation of the developed models in the healthcare environment. Therefore, this topic has many prospects for high contribution to the analytics of health care because it can effectively combine technical and practical aspects.

2. Literature Review

This section reviews the literature in the domain of machine learning for the prediction of discharge status but is limited to papers related to EHR-based predictive modelling, comparative analysis in machine learning techniques and feature selection, model explainability, and fixing challenges connected with AI deployment. The literature confirms that EHR data can drive a sea change in predictive modelling, but there are important gaps, such as more sophisticated methods of feature selection, model explainability, and real-world validation. These insights will thus help form the basis for our research toward building an accurate, interpretable model that is practically applicable to predict patient discharge outcomes, hence contributing toward healthcare management and patient care.

2.1 Introduction to EHR-based Predictive Modelling

EHRs have dramatically changed healthcare data analytics by providing structured data needed for predictive modelling. According to a study by Gao et al. (2022), systematic data mining was applied using EHR data in optimizing hospital resource management and patient discharge planning, demonstrating improved abilities to allocate resources. The integration of machine learning models with EHRs in Singapore hospitals for inpatient discharge predictions is another area that shows data-driven decision-making. According to Ayala Solares et al. (2020), there are various techniques like neural networks and decision trees that can be utilized to arrive at predictive accuracy.

During the COVID-19 pandemic, the role of EHRs in predictive modelling came to the forefront. Zapata et al. (2023) created models using EHRs for predicting possibilities of COVID-19 patients' home discharges, which proved high accuracy in handling pandemic-induced challenges. On the other hand, Lee et al. (2024) explored EHR data in emergency departments for proactive resource allocation, thus proving the versatility of EHRs across healthcare domains.

Problems of data integration and scalability remain. Ferrão et al. (2021) propose a framework for the integration of EHR data into hospital strategies that highlight the need for advanced techniques to handle dataset complexity. A totally different approach is made by Gramaje et al. (2021) who focused on scalable models dealing with large EHR datasets, thus focusing on the dual challenges of data integration and model scalability (Ferrão et al., 2021; Gramaje et al., 2021).

EHRs thus form a very important ingredient of predictive modelling for improved discharge planning and resource management in healthcare. Nevertheless, the issues relating to data integration and lack of scalability raise the need for further research. The next section will be devoted to reviewing machine learning techniques that predict discharge, showing effective algorithms with feature selection and model explainability in place.

2.2 Comparative Analysis of Machine Learning Techniques and Feature Selection

Evaluation of machine learning algorithms is instrumental in predicting discharge outcome. This research, comparing Decision Trees, Bayesian Networks, and Random Forests, has shown that while the latter is very good at handling complex interactions, Bayesian Networks are much more interpretable, thus putting forward a trade-off between accuracy and interpretability. Additional testing revealed that neural networks provide high accuracy but with an extremely high resource cost, while support vector machines represent a much more balanced trade-off in terms of model performance to efficiency, which does reflect the positive correlation of model complexity with its resource requirements. Reviews indicate that logistic regression and decision trees are very much simpler, and improvement might be noted with regard to complex enhancers such as GBM, while tuning for these models is formidable.

In psychiatric hospitals, machine learning approaches do brag of large improvements over the traditional methods, but proper model selection and validation should be done in order to prevent overfitting. Effective feature selection from EHRs including blood pressure and heart rate, according to J. P. Li et al. (2020) and Saleem & Chishti (2019) has been the key to improving predictive accuracy in stroke severity and discharge planning. In this regard, studies on neuro-intensive care outline the role of variables such as Glasgow Coma Scale scores in discharge prediction models. Recursive feature elimination and LASSO regression are some of the methods that have been identified as key drivers for predictive capacity in such models. Class I challenges to the management of high-dimensional EHR data include overfitting issues, which dimensionality reduction techniques like PCA (Bharadiya & Bharadiya, 2023)and t-SNE help reduce. In this view, the combination of feature selection with domain knowledge makes sure to retain clinically meaningful indicators, therefore balancing automated techniques with expert insights (Miao et al., 2022; Wolff et al., 2020)

Therefore, selecting relevant features from EHR data is an important step toward the derivation of accurate predictive models. Two critical aspects in developing robust models for discharge prediction will be addressing high-dimensional data challenges and balancing accuracy with interpretability. The next section will focus on enhancing model explainability and the challenges of deploying AI in healthcare.

2.3 Enhancing Model Explainability and Addressing AI Challenges

One of the things that clinical decision-making does is that model explainability is important, and structured EHR data would help tie up with NLP techniques for enhanced model interpretability, making them even more accessible to clinicians. (Kogan et al., 2020). In this regard, strategies aimed at improving model explainability, developing explainable AI using SHAP and LIME, clarify complex outputs of importance toward needing explainable AI in health, where decisions bear heavy consequences (Khodadadi et al., 2023).

Techniques like visual aids and simplifying model architectures help make predictions more understandable, balancing accuracy and transparency (Rahnenführer et al., 2023). Complex models with better performance, such as deep learning networks. A hybrid approach of interpretable models with advanced techniques ensures that acquired benefits are not undermined by muddiness (Chang et al., 2022).

Several challenges, such as those of data privacy, biasing, and scalability, exist that impede the deployment of AI in healthcare. The integration of EHR data in resource management also leads to serious privacy issues that call for very strict data governance policies. There is bias in the model due to imbalanced data sets, which needs to be dealt with if there have to be equitable predictions (Kogan et al., 2020). Oversampling and Synthetic Data Generation techniques enhance model reliability, most needed in the deployment of ethical (Khodadadi et al., 2023). Machine learning approaches are needed in optimising scalability issues with dimensionally high data (Rahnenführer et al., 2023). These studies highlight AI's potential in healthcare but also significant implementation challenges, requiring robust data strategies, scalable models, and fairness in predictions.

Model explainability and addressing AI deployment challenges are key to developing a predictive model that can be trusted. The final section will identify research gaps and contributions that have advanced patient discharge prediction.

2.4 Integration of Cloud Computing in Healthcare Systems

Cloud computing has become increasingly integral to healthcare by enabling efficient data storage, real-time access, and enhanced interoperability across healthcare networks. By leveraging cloud infrastructure, healthcare organizations gain scalable data management solutions that support the storage, and processing demands of electronic health records (EHRs), diagnostic imaging, and other high-volume data. Alassafi (2021) identify the cloud's role in reducing infrastructure costs, promoting resource flexibility, and allowing for the seamless adoption of advanced technologies, such as IoT and Big Data analytics. This integration aids healthcare providers in implementing comprehensive e-health systems, enhancing remote monitoring capabilities and telemedicine initiatives to support patient care continuity.

Further, cloud platforms improve interoperability by enabling unified access to healthcare data across multiple stakeholders, including hospitals, research institutions, and public health organizations. (Mugisha & Paik, 2023) discuss how cloud solutions centralize patient information, making it accessible in real-time to authorized healthcare professionals across various locations. This interoperability fosters collaboration and streamlined patient care, particularly for rural or under-resourced areas. As cloud adoption advances, healthcare systems can leverage real-time analytics and collaborative data sharing to optimize treatment

outcomes and operational efficiency, establishing a more connected, responsive healthcare ecosystem.

2.5 Research Gaps and Contributions

The literature reveals critical gaps in machine learning for healthcare, especially in discharge predictions. There is a need to develop better methods of feature selection to enhance model accuracy using EHR data. Finally, there will be an emerging need to create sophisticated methodologies related to the selection of impactful health indicators. (Sivan & Zukarnain, 2021). Another major gap relates to generalizability across diverse populations and settings. Most of the models developed so far are not flexible, with many realising that the application of these tools in differing scenarios is quite challenging. It's a tough requirement: ensuring the high accuracy of the models over multiple demographics and clinical environments. (Abad et al., 2021).

Real-world validation of most of the developed models is often lacking. Validating predictive models in real clinical settings gives them practical utility and effectiveness, hence more credibility for adoption (Sivan & Zukarnain, 2021). Critical care settings are complex and nuanced; they require robust models. Comprehensive clinical features improve discharge predictions within such settings when incorporated (Abad et al., 2021).

Study	*Focus Area*	*Techniques/Models*	*Key Findings*	*Challenges Addressed*
(Gao et al., 2022)	*Discharge Prediction Techniques*	*Decision Trees, Bayesian Networks, Random Forests*	*Random Forests handle complex interactions well; Bayesian Networks offer better interpretability*	*Accuracy vs. Interpretability*
(Wolff et al., 2020)	*Feature Selection and Stroke Severity Prediction*	*Deep Forest Algorithm*	*Key health indicators like blood pressure and heart rate enhance predictive accuracy*	*Feature Selection, High-dimensional Data*
(Kogan et al., 2020)	*Discharge Length Prediction in Emergency Depts*	*SVM, Neural Networks*	*Neural networks achieve high accuracy; SVMs balance performance and efficiency*	*Model Complexity, Resource Needs, Interpretability*
(Abad et al., 2021)	*AI Deployment in Healthcare*	*Systematic Data Mining Approach*	*Data governance is crucial for privacy; strategies to mitigate model bias are essential*	*Data Privacy, Bias, Scalability*

Table 1: Literature Comparison Table

There are several major gaps in the research, including feature selection, generalizability, validation in real-world settings, and robustness in critical care. Table 1 above provides a compact summary of significant literature. Our study seeks to fill these gaps: as we develop accurate, adaptive, and validated predictive models with the goal of improving patient management and resource allocation in healthcare.

3. Research methodology

The methodology adopted for this study will be based on the KDD process. This structured approach is appropriate for any predictive analytics task, especially in large and complex datasets like EHR (A. Li et al., 2018). The methodology will be broadly divided into five major phases, which include, data understanding, exploratory data analysis, preprocessing, model selection and training, and evaluation metrics, as depicted in Figure 1 below.

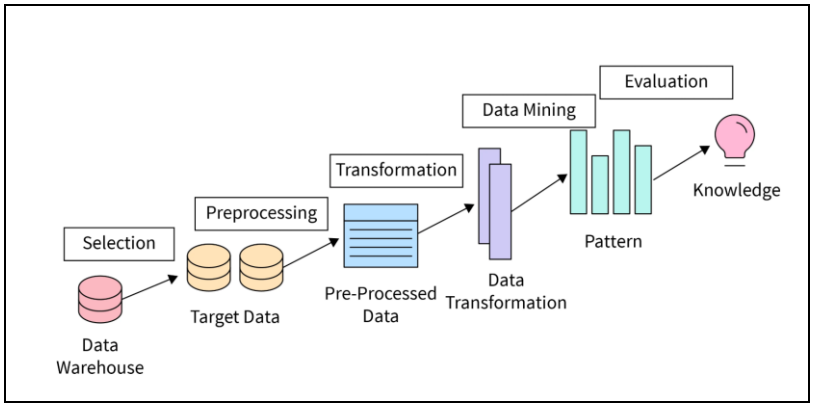

Figure 1: KDD

Data Understanding

First, the raw EHR dataset, named *EHR_data.csv*, was loaded for preliminary analysis from Mendeley Data (https://doi.org/10.17632/7kv3rctx7m.1) (Sadikin & Mujiono, 2020). The structured and unstructured data points included patient characteristics and outcomes in relation to discharge (Miao et al., 2022). In this regard, the sources, formats, and attributes of the data were closely observed to spot missing values and inconsistencies. This initial understanding helped in underlining various features that may strongly impact the prediction of patient discharge.

Exploratory Data Analysis (EDA)

EDA on the data was performed with a view to establishing patterns, relationships, and anomalies in the data. Quantitative features distribution and central tendencies were done by use of descriptive statistics. Box plots allow checking for outliers (Ali et al., 2023), while correlation heatmaps show the level of relationships among various features and thus support the process of choosing features. In exploring categorical data, the use of count plots is performed for class imbalances or predominant categories. These steps unveiled the hidden patterns present the data and prepared it for further pre-processing.

Data Preprocessing

The preprocessing pipeline carried out important steps to clean and prepare the data for modelling. Firstly, it encoded categorical variables into a numerical format to make them compatible with any kind of machine-learning algorithm. It also normalized the distribution by using PowerTransformer for skewness and scaling the features to standardize numerical attributes (A. Li et al., 2018). Principal Component Analysis was carried out to reduce dimensionality as an effort to get rid of multicollinearity and redundancy among the features (Bharadiya & Bharadiya, 2023).

Treatment of outliers was done using capping so that the effect of very extreme values is curtailed. Eventually, VIFs were checked to make sure that the problem of multicollinearity had disappeared. The cleaned data was now moved into a new CSV file called cleaned_data.csv. The cleaned data was subjected to updated visualization and descriptive statistics for validation to make sure that the quality has really improved.

Model Selection and Training

The following techniques were tested for various predictive solutions: Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network. In this case, these models were selected due to the wide range of methodologies, from simple linear ones up to ensemble and deep-learning-based techniques. Each model uses simulations of hyperparameter tuning using GridSearchCV (Belete & Huchaiah, 2022). The dataset was randomly divided into training and testing subsets, with 80% of the data used for training and 20% kept for the test; this therefore allowed validation of the generalizability of the developed models.

Evaluation Metrics

The performances for these models were measured in a variety of ways to create the most comprehensive view of performance: the percentage of samples classified correctly, a balanced measure between precision and recall using the F1-score, class distinguishability through AUC-ROC among the rest, and the relationship of sensitivity to specificity via precision-recall curves. Collectively, the aforementioned metrics informed the most suitable and robust model to real-world deployment.

4. Design Specifications

4.1 System Architecture

The system architecture as illustrated in Figure 2 for this study consists of two main pipelines: preprocessing and modeling. The preprocessing pipeline involves loading the raw dataset (EHR_data.csv), performing Exploratory Data Analysis (EDA), and applying preprocessing steps such as categorical variable encoding, feature scaling, and dimensionality reduction using PCA (Bharadiya & Bharadiya, 2023). These steps ensure the data is clean, normalized, and optimized for modelling. The processed dataset is then stored as cleaned_data.csv for further use.

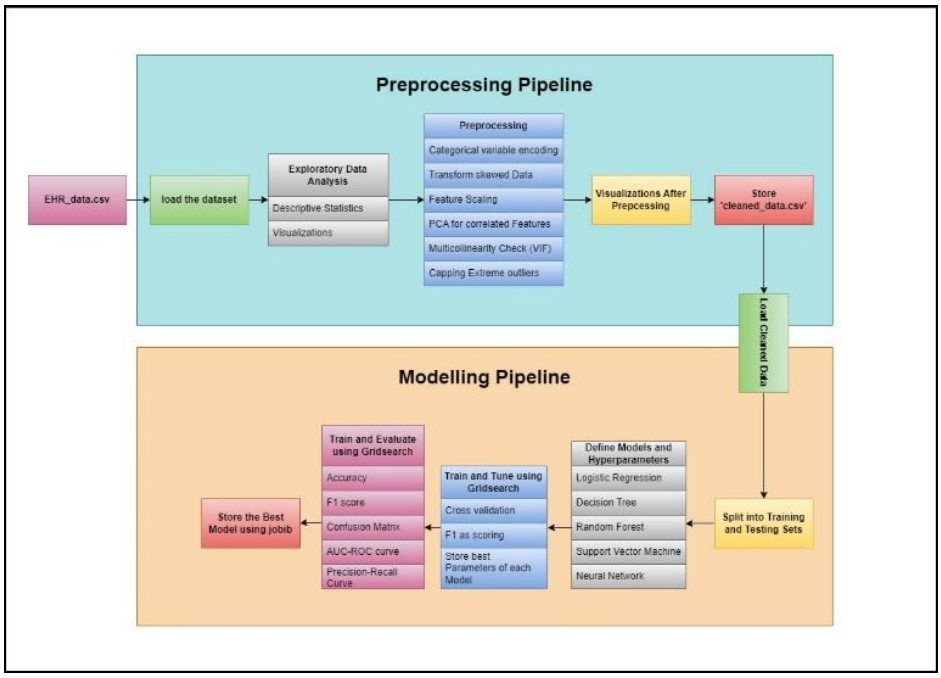

Figure 2: Architecture diagram

The modelling pipeline involves splitting the cleaned dataset into training and testing subsets, defining machine-learning models with tailored hyperparameters, and optimising these parameters using GridSearchCV. The best-performing model is evaluated using predefined metrics and saved with joblib for deployment.

4.2 Model Specifications

The combination of these five models provides a comprehensive evaluation of predictive techniques, from interpretable linear models to advanced ensemble and deep learning approaches. By leveraging the strengths of each model, this study aims to develop a predictive framework that balances accuracy, interpretability, and practical applicability in healthcare settings. The justification for the models can be seen in Table 2. The diversity in model selection ensures that the findings contribute to both the theoretical insights and practical advancements in EHR-based predictive analytics.

Model	*Justification*	*Key Literature*
Logistic Regression	*Simple, interpretable model effective in healthcare applications where transparency is critical.*	*(Mugisha & Paik, 2023)*
Decision Tree	*Captures non-linear relationships and provides hierarchical decision insights, making it accessible to healthcare professionals.*	*(J. P. Li et al., 2020)*
Random Forest	*Robust ensemble method capable of handling complex interactions and minimizing overfitting, suitable for high-dimensional EHR data.*	*(Miao et al., 2022)*
Support Vector Machine	*Balances performance and computational efficiency; kernel methods handle non-linear patterns effectively.*	*(Wolff et al., 2020)*
Neural Network	*High accuracy in capturing intricate data patterns, though resource-intensive;*	*(Kogan et al.,*

	evaluates trade-offs in advanced modeling techniques.	*2020)*

Table 2: Model Selection Justification

Each model was systematically fine-tuned to optimize performance, balancing accuracy, precision, recall, and other metrics, ensuring the final deployed model was both reliable and interpretable.

5. Implementation

5.1 Hardware, Tools, and Libraries Used

For these experiments, the used environment is one with an AMD Ryzen 5 4600H with Radeon Graphics processor, 8 GB of RAM (7.42 GB usable), and 217.82 GB storage running Windows 11 Home Single Language (Version 23H2). The main development flow used Python 3.10.13 for compatibility and efficiency, with work iterated in *JupyterLab* for ease of execution. Additionally, joblib was utilized for saving and reloading trained machine learning model, enabling efficient maintenance and management of work outputs.

The *pandas* and *numpy* libraries were important in the chain of preprocessing for data manipulation and numerical computations. *Matplotlib* and *seaborn* were other libraries used for data visualization, mainly for exploratory data analysis. Feature scaling was done using *StandardScaler*, and normalization of skewed data distributions was done using *PowerTransformer (A. Li et al., 2018)*. For dimensionally reducing the multicollinear features, PCA from *scikit-learn* was used. Finally, VIF from the *statsmodels* library was used to check for multicollinearity among features.

The machine learning pipeline consisted of the following algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Multi-Layer Perceptron from the *scikit-learn* library. Hyperparameter tuning was performed in a structured fashion using *GridSearchCV*, and model evaluation consisted of accuracy, F1-score, and AUC-ROC. Advanced visualizations included ROC and precision-recall curves and were created with the help of *matplotlib* and *seaborn*.

This combination of hardware, tools, and libraries ensures a robust and efficient framework for the building, evaluation, and deployment of predictive models.

5.2 The Preprocessing Pipeline

The preprocessing pipeline began by importing the libraries to ensure that the environment was equipped with tools for data manipulation, statistical checks, visualization, and transformation. After loading the initial dataset, an exploratory data analysis phase was conducted to gain insight into the structure and quality of the data. This initial EDA first computed the descriptive statistics to show basic distributions and central tendencies (Fig. 3), followed by a series of visualizations. Distribution plots, for example, showed how some features change with regard to the target variable, while box plots were an effective way of determining the presence and amount of outliers. Correlation heatmaps made the relationships between features quite clear and helped guide decisions about subsequent dimensionality reduction steps. Besides, count plots made the understanding of categorical variables much easier by pointing out class imbalances or dominant categories.

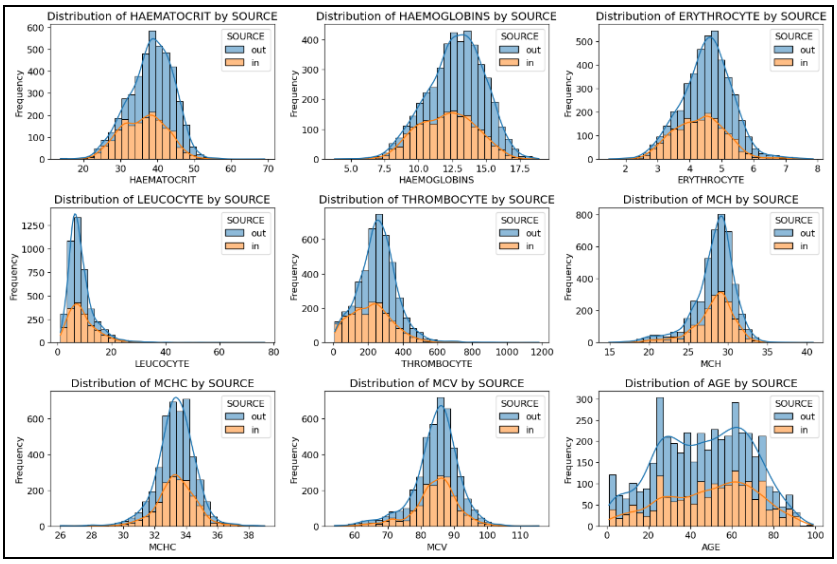

Figure 3: Distribution of EHR Features

After the picture of the raw data had crystallized, the focus shifted to a structured sequence of tasks regarding preprocessing. Categorical variables were encoded to enable non-numerical attributes to be transformed into a form suitable for machine learning models. The *PowerTransformer* handled the skew of the data distribution, making it symmetric and thus improving the performance and stability of downstream models (Fig. 4).

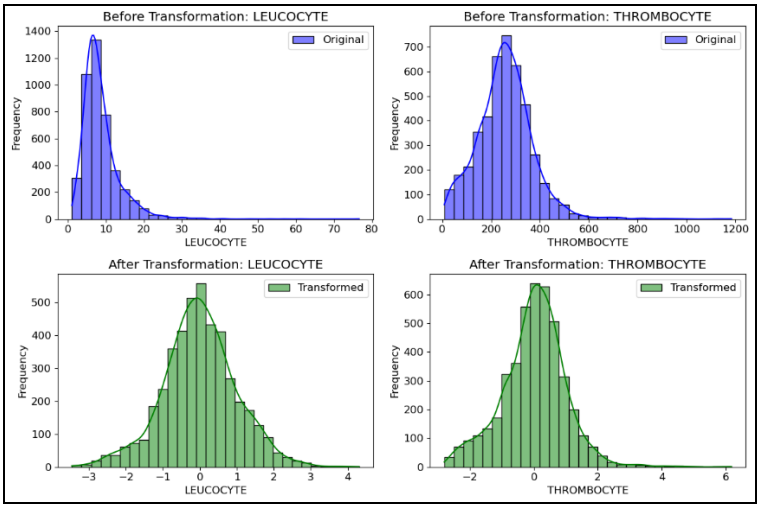

Figure 4: Effect of PowerTransformer on Skewed Features

Feature scaling, mostly through standardization, prevented variables with different scales from having a disproportionate effect on the learning algorithms. Besides, it made use of PCA for decreasing the dimensionality and addressing the problem of multicollinearity to let the model focus on those principal components explaining the data with most variance using only a few non-redundant features. The selection of this feature set was refined further by performing checks based on the Variance Inflation Factor for problems regarding multicollinearity. In a case where outliers could not be removed, capping methods were used to contain them without completely discarding data that could potentially be useful.

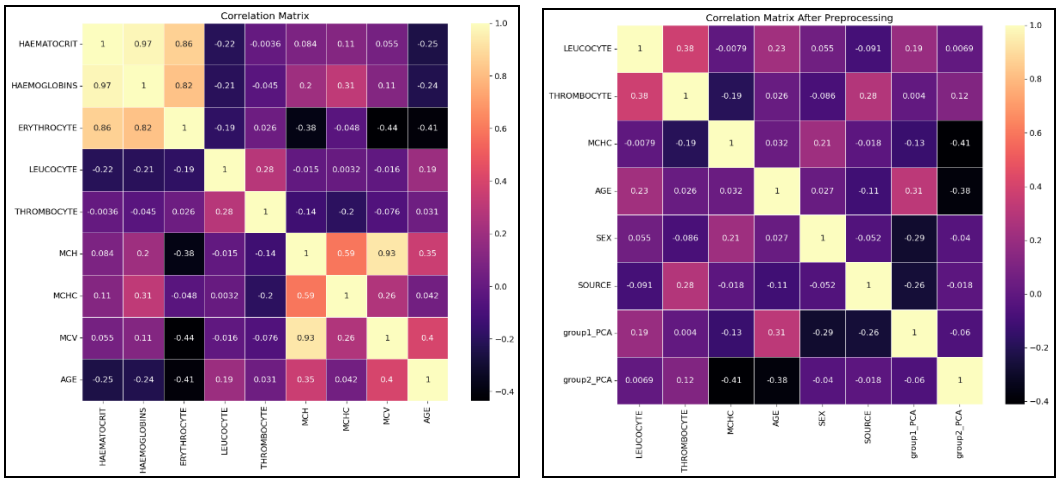

Figure 5: Correlation Matrix Before and After Preprocessing

Once these main preprocessing steps were carried out, visualizations were undertaken a second time. This post-preprocessing EDA allowed us to verify that scaling, normalization, PCA, and capping outliers improved the general quality of the dataset. The newly generated distribution plots had more normalized feature distributions; box plots depicted less influence from outliers (Fig. 6), and correlation heatmaps (Fig. 5) confirmed the non-redundant manageability of data.

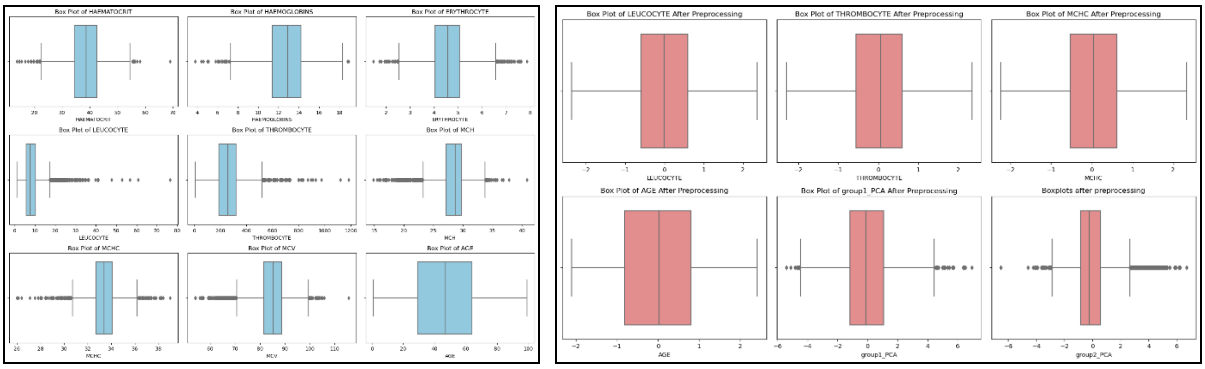

Figure 6: Box plot of Features Before and After Preprocessing

Finally, the now cleaned and pre-processed dataset-without problematic outliers, skewness, and too much dimensionality-was written to a new CSV file. This output-now clean and well-structured data would form the very basis of the subsequent modelling pipeline, ensuring that any machine learning models trained or validated would be done with data of the highest integrity.

5.3 The Modeling Pipeline

After getting a cleaned and well-structured dataset from the pre-processing phase, the modelling pipeline commenced by importing this refined data. The cleaned dataset, now free of problematic outliers, in a state of reduced dimensionality, and balanced scales, therefore presented a stable bedrock on which predictive models could be built. Upon loading the data, the target variable, *SOURCE*, was separated from the features to clearly distinguish the input variables from the outcomes that the models would seek to predict.

Having prepared the data, it would make a lot of sense to split the data into training and testing subsets. It ensures that models will be trained on one portion of the data while being evaluated on an unseen one by assigning about 80% of the data to training and about 20% to testing. This structure allowed less overfitting, besides generalizing performances of the final model better.

Having established the train-test framework, the pipeline now needed to define a set of models and their respective hyperparameters. Accordingly, a list of well-known machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and a Neural Network (MLPClassifier) were chosen for selection to span low to high complexities and different approaches. Each model type was further complemented with a predefined grid of hyperparameters for which the optimal configuration would be systematically searched. This is attested by the fact that for Logistic Regression, the tuning of regularisation strengths was completed; for Decision Trees, different tree depths and sample splits were tried; and for Neural Networks, changes in the size of hidden layers and different activation functions were tried. Random Forests were tuned by changing the number of trees and maximum depth, and for SVMs, iterative testing with different kernel values and penalty values was done to find the optimal combination.

Training and evaluation in this pipeline formed the critical core. Using *GridSearchCV*, each model was put into an extensive hyperparameter search process, where several configurations for each model were tried with their respective scores using the F1 metric. This is a convenient metric because it balances between precision and recall. After identifying the best hyperparameters for each model, that model's best version was retrained and further evaluated on the test set. Accuracy, precision, recall, F1-score, and AUC-ROC are some performance metrics that can be used to measure the holistic view concerning strengths and weaknesses that the model portrays. Consequently, visualization through confusion matrices, ROC curves, and precision-recall curves intuitively showed findings for a clear understanding of how each model was performing in handling true and false predictions, including trade-offs between sensitivity and specificity.

The last part was to compare all the models side by side. It was easy to compare which candidate model was better after each other through the tabular transformation of the evaluation results and also to plot comparative line graphs for the main metrics of interest. the best model was prepared for deployment, validated through the testing phase and supported by diagnostic visualisations that established its reliability. Conclusive performance characteristics confirmed the best hyperparameters chosen in the modelling pipeline.

6. Evaluation

6.1 Experiment: Logistic Regression

First was Logistic Regression, a linear model often used as a baseline in classification tasks. After systematic hyperparameter tuning using GridSearchCV, the model chose a regularization strength C of 0.1 and used the 'lbfgs' solver for optimization. This setting provided a good balance between generalization and computational efficiency, ensuring that the model neither underfit nor overfit the data.

After selecting the best parameters, Logistic Regression was applied to the test set to see its performance. From the resultant metrics, it could be observed that this model achieved an accuracy of about 0.71, hence correctly classifying about 71% of the samples. The precision

and recall were about 0.71 and 0.85, respectively, which postulated that the model was good in identifying the positive class but sometimes misclassified certain negatives as positives. Thus, this confirmed a reasonable balance between precision and recall with an F1-score of about 0.77, which therefore showed moderate discriminatory power across the two classes with an AUC-ROC of approximately 0.77.

Visualizations from Figure 7 provided deeper insights: the confusion matrix brought out the balance between true positives and false positives, while the ROC curve emphasized the ability of the model to do well in distinguishing between classes at various thresholds.

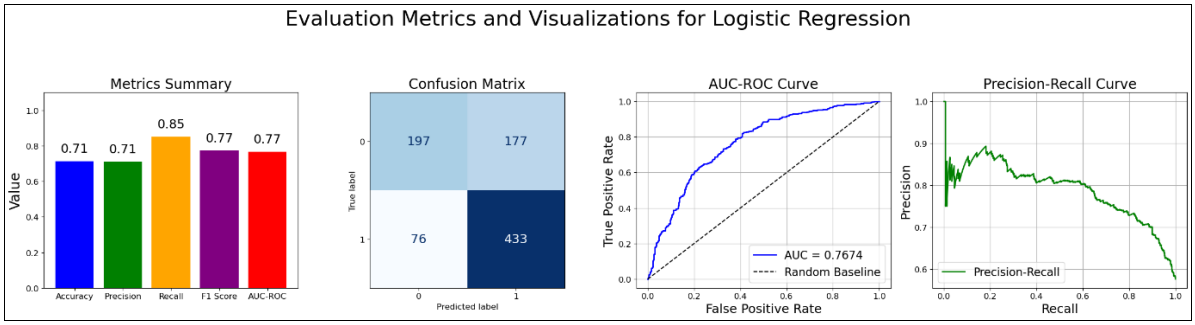

Figure 7: Evaluation Metrics and Visualizations for Logistic Regression

The precision-recall curve shows how the changes of the decision threshold affected the model sensitivity to the positive class. To wrap up, this first experiment with Logistic Regression set a solid baseline from which, although improvement was possible, the linear approach of the model had already caught most of the key patterns in the data.

6.2 Experiment: Decision Tree

The second experiment was carried out using a nonlinear model with inherent interpretability and understandability, namely the Decision Tree classifier. A number of hyperparameters were tuned concerning the criterion, depth, and splitting rules. 'Gini' was used as a criterion, while max_depth was set to 10. It was combined with a lenient constraint on the splitting: min_samples_leaf=1 and min_samples_split=2. In this way, it allowed the tree to grow up to a reasonable complexity without excessive overfitting. On the held-out set, Decision Tree achieves an accuracy of around 0.73, compared with the best Logistic Regression performance of less than that.

With a precision of about 0.73 and recall of about 0.80, it says that the tree was performing quite well in recognizing the positive examples and was less prone to mistakenly labelling a negative example.

This outputs a balanced F1-score of about 0.77, while the AUC-ROC, at about 0.77, confirmed that this tree was indeed yielding a moderate yet distinct discrimination between classes. This finding was reinforced by visualization tools (Fig. 8). The confusion matrix depicted that the hierarchical splits of the Decision Tree resulted in fewer misclassifications as compared to the linear model. The ROC curve's true positive rate for a given false positive shows more stability in improvement, and the precision-recall curve represents a model that can maintain reasonable levels of precision on a wide range of recall values. In summary, the Decision Tree performed worst of all but still showed a better adaptation to the data's inherent complexities than the logistic regression linear model.

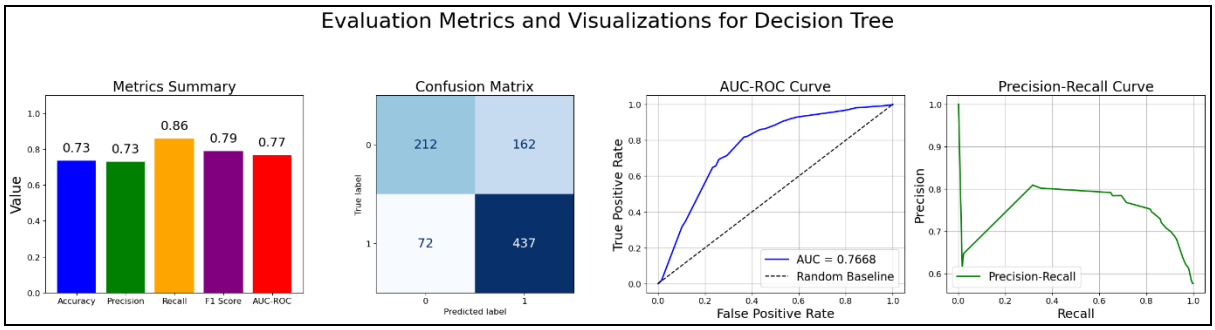

Figure 8: Evaluation Metrics and Visualizations for Decision Tree

6.3 Experiment: Random Forest

The third experiment scaled up the complexity using a Random Forest classifier. By aggregating multiple decision trees trained on bootstrapped samples of data and subsets of features, the Random Forest tamed the variance that might affect a single decision tree. Hyperparameter tuning explored different numbers of trees (n_estimators), maximum depths, and splitting criteria. The selected hyperparameters were n_estimators=500, max_depth=10, min_samples_leaf=1, min_samples_split=2, and notably bootstrap=False for increased diversity. That was the balance between model complexity and computation, considering a large forest, yet keeping trees focused and diverse.

As observed, the Random Forest showed a clear gain in performance, reaching an accuracy of about 0.76. The precision was at about 0.75, with recall as high as about 0.87. This shows its strong capability to catch true positives with fewer classification errors. Improved F1-score to about 0.80 reflects a good balance between precision and recall. Moreover, this model enhances the AUC-ROC metric, reaching as high as 0.81, and showing a good discriminative power. Compared to all the experiments before, Random Forest stood out as the best so far, as it leveraged the strength of ensemble methods in picking up complex patterns and interactions among the data (Fig. 9).

These improvements were even clearer in the visualization charts. The confusion matrix showed fewer classification errors, thus reinforcing the model's deep understanding of both classes. The ROC curve placed the Random Forest comfortably above the baseline, maintaining high true positive rates even as the false positives increased. Similarly, the precision-recall curve was consistently stronger, proving that the model sustained high precision across a range of recall levels.

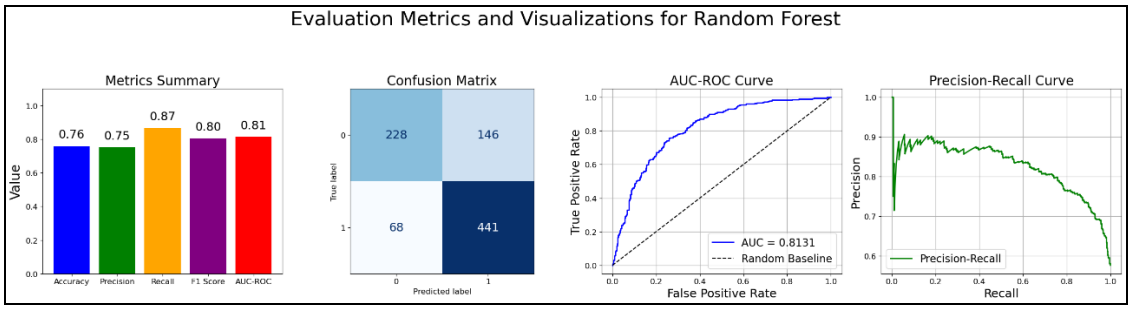

Figure 9: Evaluation Metrics and Visualizations for Random Forest

6.4 Experiment: Support Vector Machine

The fourth experiment introduces the SVM, a model known to find complex decision boundaries with high-dimensional data. Great care was taken to tune the kernel and regularisation parameters of the SVM toward optimal performance. The best configuration from GridSearchCV had a penalty parameter C of 1, a kernel type of 'rbf', a gamma parameter set to 'auto', and a polynomial degree of 2. These parameters balanced the model's flexibility with its capacity to generalise so that it didn't overfit or struggle to find meaningful patterns.

On test data, the evaluation of the SVM yields an accuracy score of approximately 0.74, where most three-quarters predictions have turned right into the desired actual results of this classification model. Much more detailed was the precision and recall, 0.73 for the number of true positive predictions against the predicted positives, and just less than 0.86 for recall for relative sensitivity to class-positive events or situations. The resulting F1-score of about 0.79 confirmed a balanced trade-off between these metrics, and an AUC-ROC score of about 0.81 reflected robust discriminative capability. This placed the performance of the SVM squarely between the Decision Tree and Random Forest results in previous experiments.

This was further supported by the visualizations (Fig. 10): the confusion matrix showed fewer errors in classification when compared to simpler models, and the ROC curve, while falling, remained above the baseline, confirming the performance of the SVM as stable and reliable across a spectrum of decision thresholds. The precision-recall curve showed how the SVM could maintain reasonable precision levels as recall increased, a good testament to the model complexity and tuneable decision boundary. Although not outperforming the performance of the ensemble-based Random Forest, the SVM showed that careful tuning of kernel-based methods could yield a performance close to the more sophisticated ensemble methods.

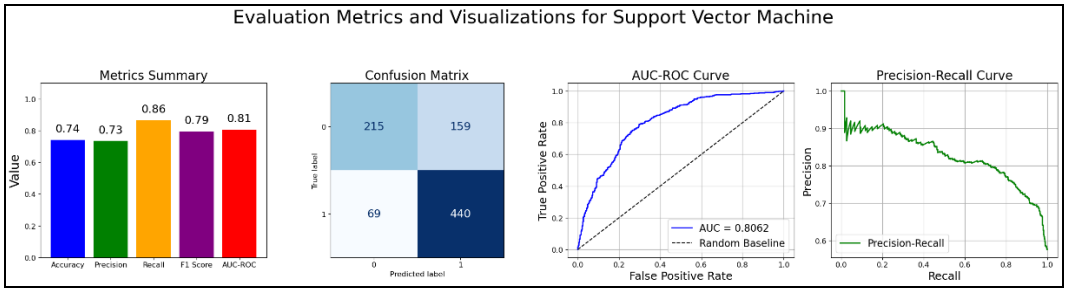

Figure 10: Evaluation Metrics and Visualizations for Support Vector Machine

6.5 Experiment: Neural Network

The fifth experiment was given to a Neural Network MLPClassifier for the implementation of deep learning to capture complex and nonlinear relationships in the data. Perform hyperparameter tuning through hidden layer sizes, activation functions, learning rate, and alpha concerning overfitting, making sure model capacity is enough. The best configuration obtained was a hidden layer structure with 150, the *'relu'* as an activation function for better propagation of gradients, a learning rate set to *'constant'*, and an alpha valued at 0.0001 for soft regularization.

This neural network reached an accuracy of about 0.76, outperforming both Logistic Regression and Decision Tree performances, while closely matching the Random Forest and SVM. Precision near 0.77 and recall at about 0.83 showed that the model was strong in identifying positive instances without inflating false positives too much. With an F1-score of

approximately 0.80, the Neural Network balanced precision and recall effectively. Although a bit surprising, considering that some simple techniques are known to perform rather well on this data, its AUC-ROC of approximately 0.80 assured that this more complex and parameter-heavy approach was indeed showing legitimate discriminatory power rather than simply overfitting the training data.

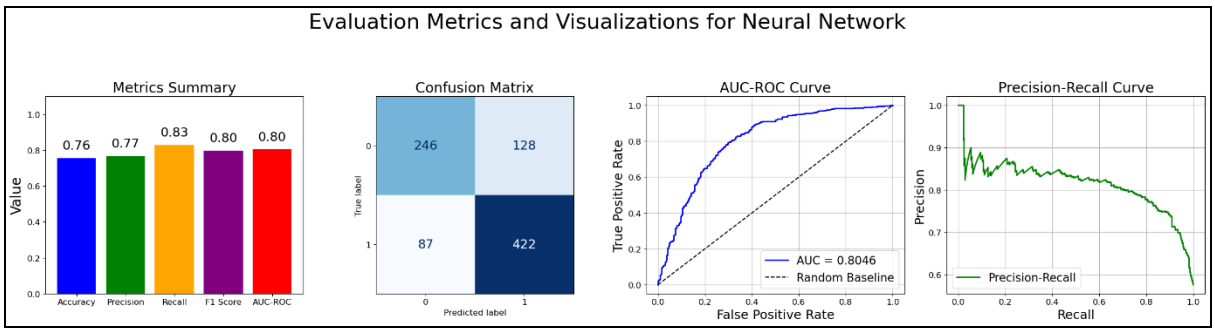

Figure 11: Evaluation Metrics and Visualizations for Neural Network

Associated plots from Figure 11 confirmed these results: from the confusion matrix, it was possible to see that the classification improved, while from the ROC curve, the increases of the true positive rate steadily increased by varying the threshold. Precision-recall stayed put, demonstrating the stability of maintaining a reasonable level for the Neural Network in its precision, even while moving along with different thresholds on the recall. Although it did not beat the best performance from the Random Forest, this Neural Network showed the potential of such advanced modelling techniques, given a sufficient amount of data and tuning effort.

6.6 Experiment: Model Comparison

The sixth and final step in the evaluation was the drawing together of all previous models tested, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Neural Network, into a comprehensive comparison analysis. Aggregation into Table 3 of key metrics like accuracy, precision, recall, F1-score, AUC-ROC, and showing all of the results on a comparative plot allow us to take an integrated overview and identify which model is actually superior while trying to balance everything about prediction performance.

Model	*best_params*	*accuracy*	*precision*	*recall*	*f1_score*	*AUC_ROC*
Logistic Regression	*{'C': 0.1, 'solver': 'lbfgs'}*	*0.713476784*	*0.709836066*	*0.850687623*	*0.773905273*	*0.767447969*
Decision Tree	*{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}*	*0.734994337*	*0.729549249*	*0.858546169*	*0.788808664*	*0.766833363*
Random Forest	*{'bootstrap': False, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}*	*0.757644394*	*0.751277683*	*0.866404715*	*0.804744526*	*0.813081117*
Support	*{'C': 1, 'degree': 2, 'gamma':*	*0.741789354*	*0.734557596*	*0.864440079*	*0.794223827*	*0.806175998*

Vector Machine	*'auto', 'kernel': 'rbf'}*					
Neural Network	*{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (150,), 'learning_rate': 'constant'}*	*0.756511891*	*0.767272727*	*0.829076621*	*0.796978281*	*0.804639484*

Table 3: Evaluation Metrics for Tuned Models

Each of the models had some advantages. Logistic Regression is understandable, very efficient, and an easy-to-understand baseline that did a reasonable job, though definitely there is further space to consider sophisticated relationships. The Decision Tree introduced nonlinearity and slightly improved metrics, showing a better understanding of subtle features than the linear approach did. The Random Forest leveraged the power of an ensemble even further in accuracy, precision, recall, and especially in AUC-ROC, a belief that ensemble methods were really capable of handling the intricacies in this dataset robustly. Whereas the Support Vector Machine and the Neural Network continued to push the boundaries, with the SVM closely nudging Random Forest's high standards and the Neural Network providing a nuanced trade-off between complexity and accuracy.

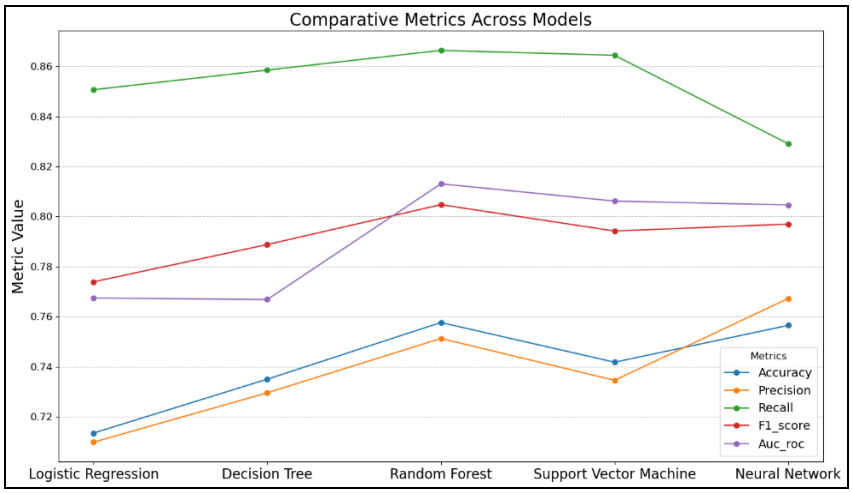

Figure 12: Comparative Metrics Across Models

The plotted metrics in Figure 12 tell an interesting story. While accuracy and precision gradually rose from the simpler to the more complex models, recall is especially vital in healthcare contexts where missing positive cases can be quite costly, improved markedly with ensemble methods and kernel-based approaches. The F1-score, the harmonic mean of precision and recall, rose steadily to higher values as models became more advanced, indicating that better overall quality of classification was achieved by more sophisticated or ensemble-based methods. Meanwhile, the AUC-ROC, which reflects the model's discriminative ability across thresholds, peaked for the Random Forest and remained high for both SVM and Neural Network, emphasising the fact that these models were proficient in distinguishing between classes even under varying decision boundaries.

The final comparison yielded one clear winner: while SVM and Neural Network were competitive, both confirming the potential of nonlinear and deep architectures, Random Forest consistently provided the best metrics. The balanced improvement of this model in all the key performance indicators, together with its non-declining AUC-ROC, best qualified it

for real-world deployment. The relative interpretability of the Random Forest using feature importance analysis and SHAP explanations added confidence that the stakeholders would be in a position to understand-and thus trust-the model's decisions in a clinical setup.

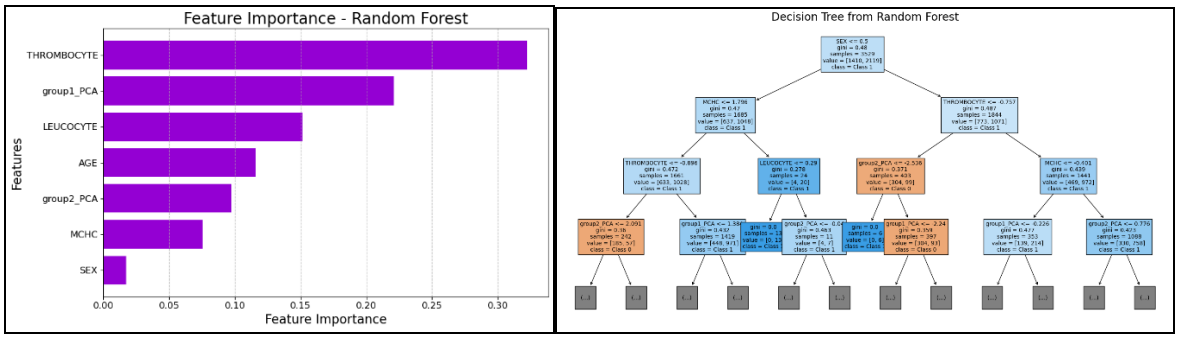

Figure 13: Feature Importance and a Decision Tree from Random Forest

This was clearly reflected in the visualisations, where the Random Forest model outperformed the rest. From Figure 13, the decision tree from the Random Forest ensemble provides a very clear hierarchical representation of important decision paths that the model follows in making predictions, showcasing its capabilities for complex relationships. A feature importance plot further points to features such as *THROMBOCYTE*, *group1_PCA*, and *LEUCOCYTE* that contributed to the critical building up of the model's predictive accuracy (Fig. 13). The combination of interpretability and high performance underlines the robustness of Random Forest as the best choice to predict the outcome in the case of patient discharge. These insights have been pivotal in understanding the key drivers of the model's predictions, providing valued inputs for clinical decision-making.

In a nutshell, the comparative exercise highlighted not only the incremental value of exploring multiple algorithms and tuning their hyperparameters but also guided the selection of a final model. Random Forest came out to be the best among them, balancing robust accuracy, strong discriminative capability, and interpretability. The evaluation chapter thus reached its logical conclusion by placing individual model performances in a broader comparative landscape, laying a foundation for deploying the chosen model in practice.

6.7 Discussion

The study employed a systematic approach to predicting patients' discharge using machine learning models from EHR data. While strong predictive performance was attained with the results, there is a dire need for critical assessment in methodology and identification of areas which may require an upgrade. One limitation of the design used in the study is that the analysis relied on one dataset; hence, the generalisation of the model to other clinical settings could be compromised. Incorporation of data from so many different healthcare facilities could increase robustness and applicability. Besides, even though the preprocessing steps were followed very carefully, outliers and removing the skewness in some variates may have affected those model performances. A deeper investigation involving alternative scaling methods or robust estimators may alleviate these questions.

The results of the current study compare pretty well with the literature, mainly about the high performances of Random Forest and Neural Networks in performing predictive tasks on healthcare data. However, the deviations in certain metrics-like precisions across Logistic Regression-predict that it might depend a lot on the data-specific pattern, underlining the

feature engineering-domain-specific tuning aspect that still has to be investigated in further detail.

Despite these limitations, the experiment contributed significantly to leveraging EHR data for predictive analytics. The combination of preprocessing strategies, systematic model evaluation, and robust visualization techniques provides a replicable framework for similar studies. This study showed that machine learning can be used to support clinical decision-making, though with due consideration for limitations in any real-world application.

7. Conclusion and Future Work

7.1 Conclusion

This study has developed a machine learning model to predict a patient's discharge outcome using EHR data. In this work, five models have been developed and tested: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Neural Network. Random Forest proved the most robust, providing the best balance between precision, recall, and overall accuracy. Results confirm the feasibility of the application of machine learning to EHR data for predictive analytics but also point out the main challenges that relate to the problem of skewness of the data and poor generalization. These limitations open paths toward future research efforts, for instance, ensembling using various datasets or the realisation of real-time pipelines.

The described inquiry already provides a successful answer to the research question, since it shows that ML techniques combined with feature selection and interpretability methods can predict patient discharge outcomes using EHR data with high accuracy. Out of the five models explored, Random Forest became the most powerful with balanced accuracy and AUC-ROC at 0.81, easily interpretable through feature importance analysis. Dimensionality reduction techniques such as PCA and VIF highlighted **THROMBOCYTE levels** and PCA-derived components as the most important predictors in explaining variances of predictions. While promising, these also allude to the limitations of dataset diversity and real-world validation that need to be addressed in future enhancements for clinical applicability.

The results of this work will be of great help in understanding and developing actionable insights into healthcare analytics, besides providing a stepping stone for further research. In fact, the proposed framework with enhanced scalability and validation might be the starting point for developing sophisticated clinical support systems, possibly aiding the medical fraternity with timely and accurate decisions.

7.2 Future Work

Future work on this project should be directed toward the expansion of the dataset through the integration of EHR data from multiple healthcare institutions to enhance generalizability across diverse clinical environments. Real-time predictive pipelines will be developed to suit dynamic healthcare scenarios, enhancing timeliness and relevance in predictions.

Advanced explainable AI methods, such as SHAP and LIME, can be integrated for clear, interpretable insights of model decisions for clinicians and are therefore trust-inducing in the

system. Final deployment and validation at real clinical settings are crucial with regards to the practical utility, reliability, and scalability of the developed models for wider diffusion.

Acknowledgement

I sincerely thank Dr David Hamill, my supervisor, for his invaluable guidance, weekly meetings, and consistent support throughout this project. His feedback and direction for each section of the report were crucial to its completion and success.

REFERENCES

Abad, Z. S. H., Maslove, D. M., & Lee, J. (2021). Predicting Discharge Destination of Critically Ill Patients Using Machine Learning. *IEEE Journal of Biomedical and Health Informatics*, *25*(3), 827–837. https://doi.org/10.1109/JBHI.2020.2995836

Alassafi, M. O. (2021). Success indicators for an efficient utilization of cloud computing in healthcare organizations: Saudi healthcare as case study. *Computer Methods and Programs in Biomedicine*, *212*. https://doi.org/10.1016/j.cmpb.2021.106466

Ali, M., Zhu, P., Huolin, M., Pan, H., Abbas, K., Ashraf, U., Ullah, J., Jiang, R., & Zhang, H. (2023). A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization. *Natural Resources Research*, *32*(3), 1047–1066. https://doi.org/10.1007/s11053-023-10184-6

Ayala Solares, J. R., Diletta Raimondi, F. E., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Pinho Gomes, A. C., Payberah, A. H., Zottoli, M., Nazarzadeh, M., Conrad, N., Rahimi, K., & Salimi-Khorshidi, G. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. In *Journal of Biomedical Informatics* (Vol. 101). Academic Press Inc. https://doi.org/10.1016/j.jbi.2019.103337

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*(9), 875–886. https://doi.org/10.1080/1206212X.2021.1974663

Bharadiya, J., & Bharadiya, J. P. (2023). A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning. *Article in International Journal of Innovative Research in Science Engineering and Technology*, *8*(5). https://doi.org/10.5281/zenodo.8002436

Chang, Y. H., Shih, H. M., Wu, J. E., Huang, F. W., Chen, W. K., Chen, D. M., Chung, Y. T., & Wang, C. C. N. (2022). Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department. *BMC Emergency Medicine*, *22*(1). https://doi.org/10.1186/s12873-022-00632-6

Ferrão, J. C., Oliveira, M. D., Gartner, D., Janela, F., & Martins, H. M. G. (2021). Leveraging electronic health record data to inform hospital resource management: A systematic data mining approach. *Health Care Management Science*, *24*(4), 716–741. https://doi.org/10.1007/s10729-021-09554-4

Gao, R., Cheng, W. X., Suganthan, P. N., & Yuen, K. F. (2022). Inpatient Discharges Forecasting for Singapore Hospitals by Machine Learning. *IEEE Journal of Biomedical and Health Informatics*, *26*(10), 4966–4975. https://doi.org/10.1109/JBHI.2022.3172956

Gramaje, A., Thabtah, F., Abdelhamid, N., & Ray, S. K. (2021). Patient Discharge Classification Using Machine Learning Techniques. *Annals of Data Science*, *8*(4), 755–767. https://doi.org/10.1007/s40745-019-00223-6

Khodadadi, A., Ghanbari Bousejin, N., Molaei, S., Kumar Chauhan, V., Zhu, T., & Clifton, D. A. (2023). Improving Diagnostics with Deep Forest Applied to Electronic Health Records. *Sensors*, *23*(14). https://doi.org/10.3390/s23146571

Kogan, E., Twyman, K., Heap, J., Milentijevic, D., Lin, J. H., & Alberts, M. (2020). Assessing stroke severity using electronic health record data: A machine learning approach. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-019-1010-x

Lee, S.-Y., Chinnam, R. B., Dalkiran, E., Krupp, S., & Nauss, M. (n.d.). *Prediction of emergency department patient disposition decision for proactive resource allocation for admission*. https://doi.org/10.1007/s10729-019-09496-y/Published

Li, A., Yang, X., Dong, H., Xie, Z., & Yang, C. (2018). Machine learning-based sensor data modeling methods for power transformer PHM. *Sensors (Switzerland)*, *18*(12). https://doi.org/10.3390/s18124430

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access*, *8*, 107562–107582. https://doi.org/10.1109/ACCESS.2020.3001149

Miao, Z., Sealey, M. D., Sathyanarayanan, S., Delen, D., Zhu, L., & Shepherd, S. (2022). *A Data Preparation Framework for Cleaning Electronic Health Records and Assessing Cleaning Outcomes for Secondary Analysis 2 3 4*.

Mugisha, C., & Paik, I. (2023). Bridging the Gap between Medical Tabular Data and NLP Predictive Models: A Fuzzy-Logic-Based Textualization Approach. *Electronics (Switzerland)*, *12*(8). https://doi.org/10.3390/electronics12081848

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A. L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., & McShane, L. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. In *BMC Medicine* (Vol. 21, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12916-023-02858-y

Sadikin, Mujiono (2020), "EHR Dataset for Patient Treatment Classification", Mendeley Data, V1, doi: 10.17632/7kv3rctx7m.1 https://doi.org/10.17632/7kv3rctx7m.1

Saleem, T. J., & Chishti, M. A. (2019). Exploring the Applications of Machine Learning in Healthcare. *International Journal of Sensors, Wireless Communications and Control*, *10*(4), 458–472. https://doi.org/10.2174/2210327910666191220103417

Sivan, R., & Zukarnain, Z. A. (2021). Security and privacy in cloud-based e-health system. In *Symmetry* (Vol. 13, Issue 5). MDPI AG. https://doi.org/10.3390/sym13050742

Wolff, J., Gary, A., Jung, D., Normann, C., Kaier, K., Binder, H., Domschke, K., Klimke, A., & Franz, M. (2020). Predicting patient outcomes in psychiatric hospitals with routine data: A machine learning approach. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-1042-2

Zapata, R. D., Huang, S., Morris, E., Wang, C., Harle, C., Magoc, T., Mardini, M., Loftus, T., & Modave, F. (2023). Machine learning-based prediction models for home discharge in patients with COVID-19: Development and evaluation using electronic health records. *PLoS ONE*, *18*(10 October). https://doi.org/10.1371/journal.pone.0292888