Phenotype Exploration of Ischemic Stroke Patients Exhibiting Readmission to ICU within One Year

Group 17
Joseph Suckling, Sarathi Thirumalai Soundararajan,
Nadya Ulibasa, Xuan Wu

Semester 2, 2024

1 Abstract

Objective: This study aims to predict 365-day hospital readmission for ischemic stroke patients using machine learning models and to identify key predictive features to inform clinical decision-making.

Materials and Methods: We used the MIMIC-IV dataset to build and evaluate three ML models: Random Forest, XGBoost, and PCA Clustering. Due to the class imbalance in readmission cases, we employed SMOTE oversampling for Random Forest to improve performance. Hyperparameters for each model were optimized using grid search with cross-validation, and feature importance was analyzed using clustering and permutation methods.

Results: The models achieved moderate performance, with AUC values of 0.64 for Random Forest and 0.78 for XGBoost, indicating limited discriminative ability. While SMOTE improved class balance, the models still struggled to achieve high predictive accuracy. Clustering analysis identified features such as length of stay, troponin T levels, and maximum systolic BP as influential in predicting readmission. Permutation importance for Random Forest highlighted the complex interplay of features affecting patient outcomes.

Discussion: Our findings emphasize the trade-off between cohort identification accuracy and model usability. While ICD codes provided reliable patient classification, they restricted the model's real-time application. Digital phenotyping challenges and the moderate performance of ML models suggest the need for more comprehensive data and advanced methods to improve predictive accuracy.

Conclusion: The study demonstrates the difficulties of developing robust ML models for ischemic stroke readmission prediction. Future work should focus on integrating real-time data and refining digital phenotyping techniques to enhance model performance and clinical relevance.

2 Introduction

2.1 Clinical Problem

Stroke is the second leading cause of death worldwide [1]. Ischemic stroke occurs when a blood clot blocks an artery supplying the brain [2]. Ischemic stroke is also the most common type, accounting for over 87 percent of strokes in the US [3].

Short term readmission of patients after ischemic stroke is common - occurring in approximately 17 percent of patients. Short term readmission is often costly, and patients who are readmitted after ischemic stroke often suffer detrimental health outcomes [3].

The ability to accurately predict a patients' risk of readmission will help clinicians improve patient outcomes and avoid unnecessary hospital expenditure.

2.2 Current Literature

Machine Learning (ML) models have been heavily utilized to predict readmission likelihoods for ICU patients. Research in this field is established and has produced accurate results [4]. However, the prediction of stroke patient readmission is a significantly more nascent field. Prior efforts have focused on statistical models, such as logistic regression, with little effectiveness [5]. While Chen [6] describes ML approaches to stroke readmission prediction as a solution "rarely discussed," several studies have yielded promising results in this area.

Darabi et al. [7] found that clinical variables relating to stroke readmission could be identified through ML models. They included a cohort of 3,184 ischemic stroke patients from the Geisinger Health System records and found the best performance through XGBoost with ROSE-sampling. Lineback et al. [8] looked at both ischemic and haemorrhagic stroke, establishing a cohort of 2,855 from the NM-EDW database. They found that Natural Language Processing (NLP) enhanced ML algorithms led to 'modest' improvements over statistical models. The largest study to date examined 6,558 patients from Xuzhou Medical University Hospital records.

To date, our team cannot find any studies that utilize the MIMIC datasets for ML prediction of stroke readmission.

2.3 Our Approach

We employed a variety of machine learning models, alongside statistical models for comparison, to predict hospital readmission among ischemic stroke patients within one year. Building off the research [9], we also analyzed feature importance within our models to identify key factors influencing readmission risk.

3 Methods

3.1 Dataset

The dataset used is the MIMIC-IV dataset. MIMIC-IV contains de-identified data for over 65,000 ICU patients, including a vast array of relevant clinical data. The data is publicly available and sourced from the electronic health record of the Beth Israel Deaconess Medical Center [10].

3.2 Cohort

Our initial cohort of ischemic stroke patients were selected through ICD-9 codes (those starting with 433 or 434). We discuss the limitation of this later, however, we believe this approach is the most reasonable in the clinical context and is consistent with the literature [8].

The alternative would be a digital phenotyping approach. The reality is that stroke is primarily identified through medical imaging, as well as other specific tests such as blood flow tests [11]. Consequently, the gold standard is to use text processing of radiology reports [12]. We investigated text analysis of radiology reports and the extensive data contained in MIMIC-IV-TEXT. However, we were unable to confidently identify a sufficiently large cohort.

Initially we included all patients from the MIMIC-IV hosp tables. However, because we used features that were predominantly collected in ICU patients, we narrowed our cohort to include only those who had entered the ICU during their hospital admission.

While we initially set out to do 90-day stroke readmission, we found that our cohort was either too small, or too incomplete train our ML models with confidence. In order to balance clinical relevance with the viability of our cohort, we decided to use 365 day readmission. In this way we will have confidence in the validity of our models and hope to capture a broader range of readmission patterns.

Readmission was defined as a readmission to hospital with another ischemic stroke.

3.3 Feature Selection

We have considered 22 initial features from the following 5 aspects:

- Demographics
- ICU Vital Signs
- Infections
- Lab Cardiovascular Tests
- Comorbidities

3.4 Data Processing

3.4.1 Outlier Handling

Outlier handling in healthcare data is crucial to reduce noise and improve model robustness [13]. Outliers were identified based on predefined physiological boundaries specific to medical norms for ischemic stroke patients.

3.4.2 Data Handling

Upon further investigation, some ICU Vital Signs features had discrepancies caused by the presence of multiple units in MIMIC-IV. We filtered out the selected features with fluid and unit pairs that had the most records to create a better-narrowed cohort for our prediction model.

For the machine learning models, we added a 365-day readmission class variable, where ”1” represents a patient readmitted due to ischemic stroke within a year, and ”0” represents a patient who did not experience stroke-related readmission.

In summary, our final features are outlined below in Table 1:

Table 1: Features with Missing Values

Feature	**Missing Values (%)**
Height	44.0
SBP (Maximum)	49.0
WBC (Maximum)	56.0
Lymphocyte Percentage	48.0
RBC (Maximum)	59.0
Lactate Level (Maximum)	40.0
Albumin Level (Minimum)	43.0
Troponin T Level (Maximum)	64.0

3.4.3 Missing Value Imputation and Encoding

Missing data was addressed using a combination of K-nearest neighbors (KNN) imputation for continuous features, mode imputation for binary features, and one-hot encoding for categorical features [14].

3.4.4 Scaling

StandardScaler and RobustScaler were applied to continuous features after imputation to normalize the range of values [15].

3.4.5 Collinearity Handling

High collinearity among features can lead to redundancy and decreased model performance. A correlation matrix was computed to identify pairs of features with correlation coefficients above 0.9. From each highly correlated pair, one feature was removed to minimize redundancy.

3.4.6 Low-Variance Feature Removal

Features with low variance are unlikely to contribute meaningful information for classification or prediction. A variance threshold of 0.01 was applied to remove such features, to enhance computational efficiency in our ML pipelines [16].

3.4.7 Dimensionality Reduction via Principal Component Analysis (PCA)

To further reduce dimensionality and capture the underlying structure of the data, PCA was applied to retain 95% of the variance in the dataset. This step is particularly valuable for ischemic stroke datasets, as it allows us to capture complex interactions between physiological variables without retaining all original features [17].

3.5 Machine Learning Models

3.5.1 Random Forest

The random forest algorithm is an ensemble machine learning algorithm that aggregates results from multiple decision trees using majority voting [18]. It has been applied in related fields such as predicting in-hospital mortality of ICU patients with acute kidney injury [19] and sepsis in patients [20].

When there is class imbalance in the data, synthetic minority oversampling technique (SMOTE) is a useful technique to tackle this issue [21]. Combining SMOTE with with the random forest technique will be used to improve the classification performance.

3.5.2 XGBoost

For our project, XGBoost presents an ideal choice due to its ability to handle missing data efficiently, which aligns with our dataset that includes a substantial proportion of imputed values. Additionally, the algorithm's regularization techniques help reduce overfitting, a common challenge when working with medical data where feature relationships can be intricate and prone to noise. Furthermore, XGBoost's tree-based approach [22] can model complex, non-linear relationships in the data, which is essential in healthcare applications where interactions between patient characteristics, comorbidities, and lab results often exhibit non-linear behaviors.

3.5.3 Clustering

Clustering analysis groups features based on their influence on the model's predictions, revealing patterns and interactions that may be indicative of specific patient phenotypes associated with readmission risk [23]. Clustering can reveal insights about co-occurring clinical conditions or risk factors within the high-risk population.

3.6 Ethics Statement

The data used in this project is derived from the MIMIC-IV dataset, which is publicly available but requires users to complete training in human subjects research and obtain permission for access. We adhered to all data use agreements and ethical guidelines associated with MIMIC to ensure the responsible handling of sensitive patient information. No personally identifiable information is present in the dataset, preserving the anonymity and privacy of individuals. Data generated or processed during this project will not be redistributed, and all analyses have been conducted with care to respect the sensitive nature of medical records.

4 Results

4.1 Exploratory Data Analysis

The initial analysis of the features helped us describe the dataset. The cohort predominantly consisted of elderly patients (60+), with an average weight of around 70–80 kg. Only 10.96% of the patients experienced readmission. While this seems low, it is consistent with what we found in similar literature. Temperature data showed a few febrile cases, suggesting infection, while high systolic blood pressure values in many patients indicated hypertension. Some elevated WBC levels suggested infection, and many lower lymphocyte levels indicated a suppressed immune system. Additionally, diabetes and hypertension comorbidities were prevalent in most patients see (Figure 1).

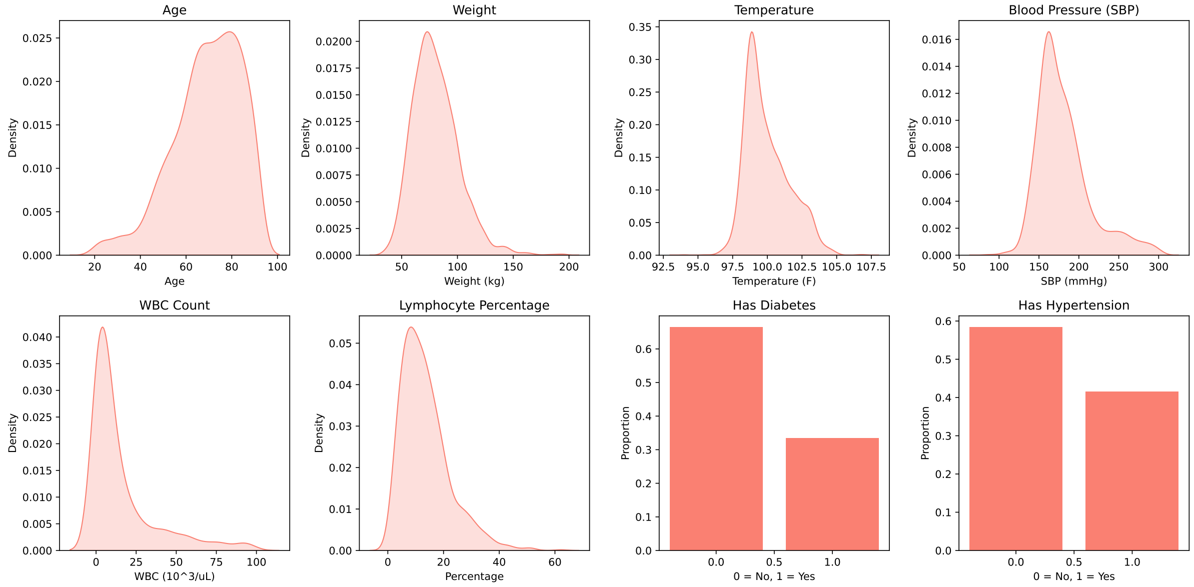

Figure 1: KDE Plot of Features

The gender demographic was 52.8% male and 47.2% female. Among the patients with readmission, 55.7% were male and 44.3% were female, suggesting that males in this cohort were more likely to experience stroke readmission see (Figure 2).

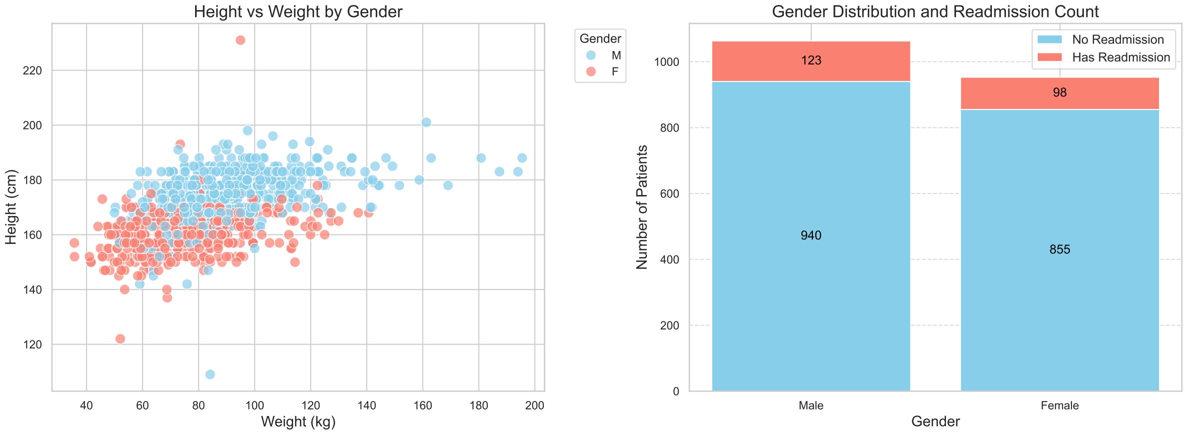

Figure 2: Height, Weight, Gender and Readmission Analysis

This discharge time was used to identify hospital admissions occurring within 30, 120, 180, and 360 days after the previous hospital discharge date, in order to find the optimal duration for our ML model analysis. We selected 365 days, as it had the most admissions (221) see (Figure 3).

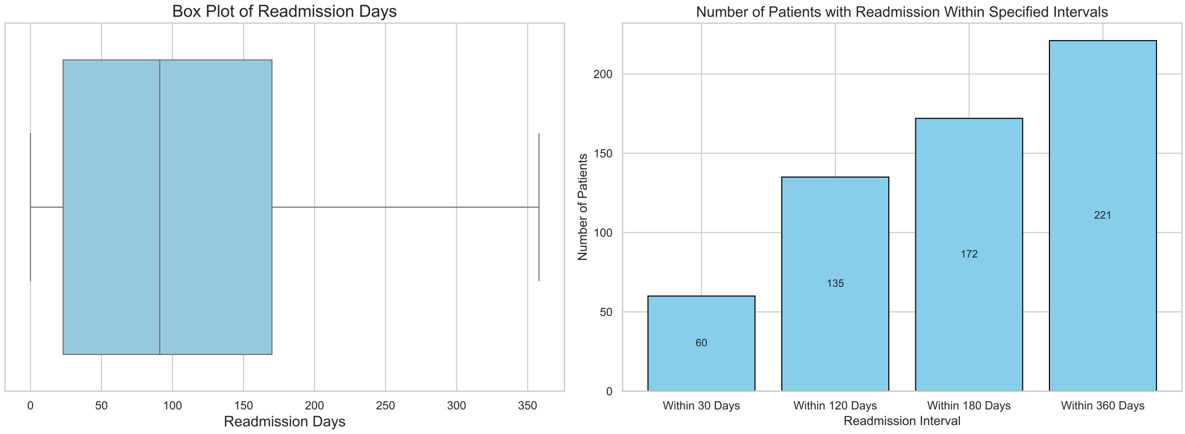

Figure 3: Readmission Days and Patient Counts by Intervals

The (Table 2) represents a comparison of clinical variables between patients with and without hospital readmission.

Table 2: Summary of Variables by Readmission Status

Variable	**No Readmission**	**Readmission**	**Total**
Demographics			
Age (years)	69.16 ± 15.24	69.72 ± 12.36	2016
Weight (kg)	80.93 ± 21.09	79.25 ± 16.62	2008

Variable	No Readmission	Readmission	Total
Height (cm)	168.90 ± 11.39	168.26 ± 9.79	1126
Gender	M: 940 F: 855	M: 123 F: 98	2016
Length of Stay (days)	5.63 ± 7.45	4.01 ± 4.23	2016
No. of Stays	1.13 ± 0.42	1.19 ± 0.46	2016
ICU Vital Signs			
Systolic Blood Pressure (mmHg)	181.02 ± 33.35	177.89 ± 35.57	1032
Heart Rate (bpm)	79.67 ± 12.37	77.88 ± 11.34	1993
Temperature (°F)	100.04 ± 1.62	99.56 ± 1.22	1992
Respiratory Rate (insp/min)	18.93 ± 3.24	18.53 ± 2.92	2016
Infections			
WBC Count (#/hpf)	15.48 ± 20.71	16.02 ± 23.71	877
RBC Count (#/hpf)	17.08 ± 23.06	10.43 ± 18.50	818
Lymphocyte Percentage	13.43 ± 8.76	14.16 ± 8.01	1056
Lab Cardiovascular Tests			
Lactate Level (mmol/L)	2.97 ± 2.56	2.25 ± 1.26	1216
Albumin Level (g/dL)	3.19 ± 0.68	3.21 ± 0.66	1145
Glucose Extreme Count (mg/dL)	2.54 ± 5.25	2.03 ± 3.71	2016
Sodium Extreme Count (mEq/L)	4.91 ± 9.30	3.45 ± 5.92	2016
Troponin T Level (ng/L)	0.82 ± 2.37	0.91 ± 2.12	722
Comorbidities			
Hypertension	1052 (58.61%)	126 (57.01%)	2016
Diabetes	578 (32.20%)	97 (43.89%)	2016
Atrial Fibrillation	695 (38.72%)	74 (33.48%)	2016
Charlson Comorbidity Index	0.09 ± 0.29	0.14 ± 0.35	2016

4.2 Feature Selection

The final features after data processing is shown as the following PCA result, which shows a summary of each principal component, the feature with the strongest effect, and its loading value (Table 3):

Table 3: Influential Feature for Principal Components

Principal Component	Most Influential Feature	Loading Value
PC1	Length of Stay	0.406903
PC2	Height	0.578569
PC3	Troponin T Level	0.497462
PC4	WBC Count	0.612487
PC5	RBC Count	0.470055
PC6	Systolic Blood Pressure	0.520356
PC7	Troponin T Level	0.553311
PC8	Troponin T Level	0.429682
PC9	Respiratory Rate	0.617803
PC10	Lymphocyte Percentage	0.486586

Principal Component	**Most Influential Feature**	**Loading Value**
PC11	No. of Stays	0.404918
PC12	Glucose Extreme Count	0.487040
PC13	Albumin Level	0.681930
PC14	Age	0.470885
PC15	Sodium Extreme Count	0.611774
PC16	Weight	0.643377
PC17	Length of Stay	0.764128

4.3 Random Forest

We implemented a Random Forest (RF) algorithm on the selected features to train the model. To optimize the model's performance, we use GridSearchCV which tunes hyperparameters and uses cross-validation to evaluate their performance. In addition, we set a balanced class weight in the model to give more weight to the minority class.

In our experiment, we compared the baseline RF model with the RF model enhanced by SMOTE oversampling. The results as shown in the table 4 displays a difference in the model's performance, demonstrating the effectiveness of handling class imbalance through synthetic data generation.

Table 4: Random Forest Performance Metrics

Method	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**ROC AUC**
Random Forest	0.908	0.598	1.0	0.139	0.570
Random Forest + SMOTE	0.816	0.610	0.269	0.419	0.641

The plain RF model achieved high overall accuracy, primarily due to classifying most cases as the dominant class 0 (no readmission). SMOTE mitigated this issue by providing better balance between classes, resulting in higher recall for class 1 (readmission). However, the performance for class 1 remained limited and the AUC for both models was still relatively low (0.570 and 0.641), indicating poor discriminative ability.

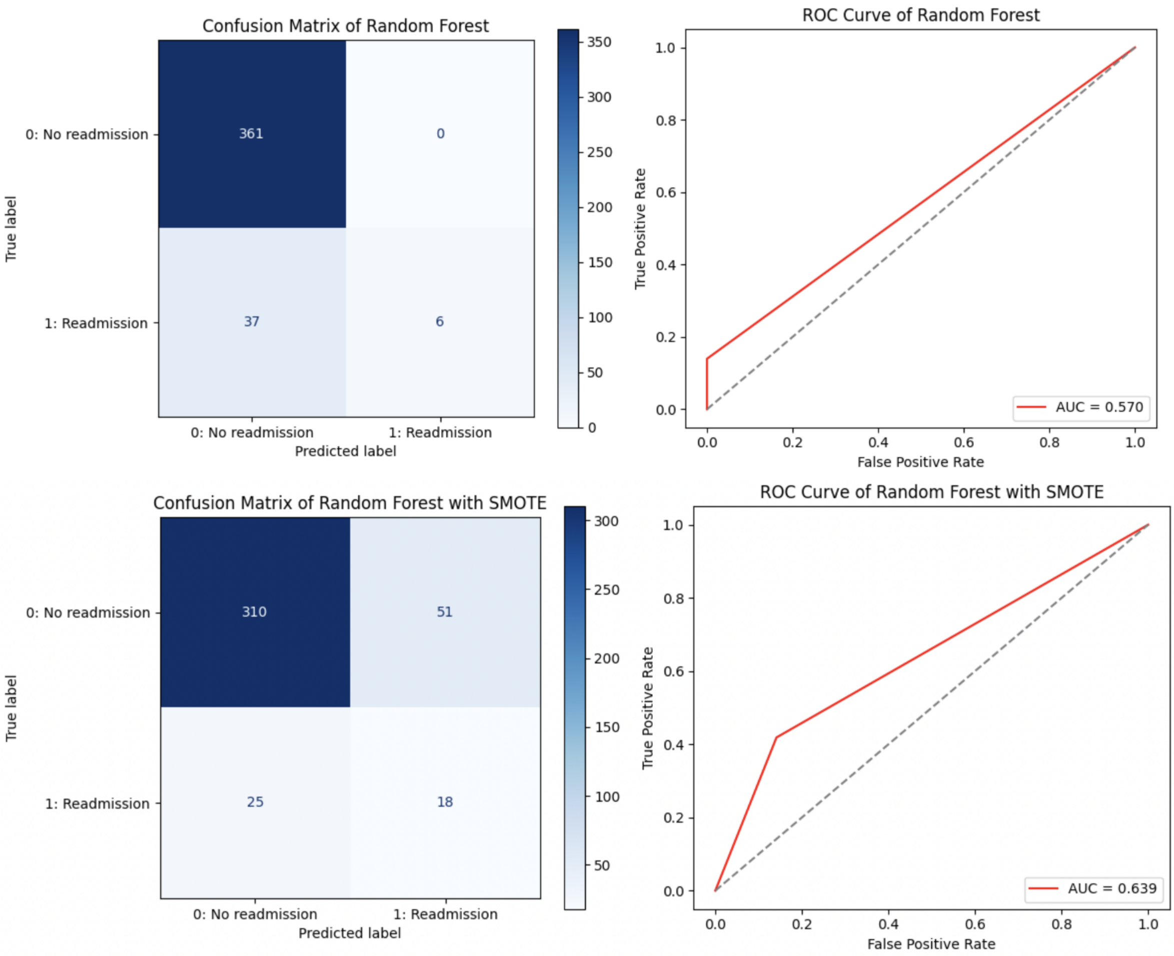

Figure 4: Random Forest Performance Comparison

4.4 XGBoost

We implemented XGBoost to predict stroke readmission. Initially we applied a default XGBoost model to establish our baseline, then iterated with hyperparameter tuning using RandomizedSearchSV to improve performance. We utilized a custom recall function that aimed to reduce false negatives - where we fail to identify patients who were actually readmitted. This was done because false negatives pose a major clinical concern.

Table 5: Comparison of XGBoost Performance Metrics

Method	Accuracy	F1 Score	Precision	Recall	ROC AUC
XGBoost Baseline	0.93	0.50	0.88	0.35	0.76
With Hyperparameter Tuning	0.91	0.56	0.78	0.44	0.78

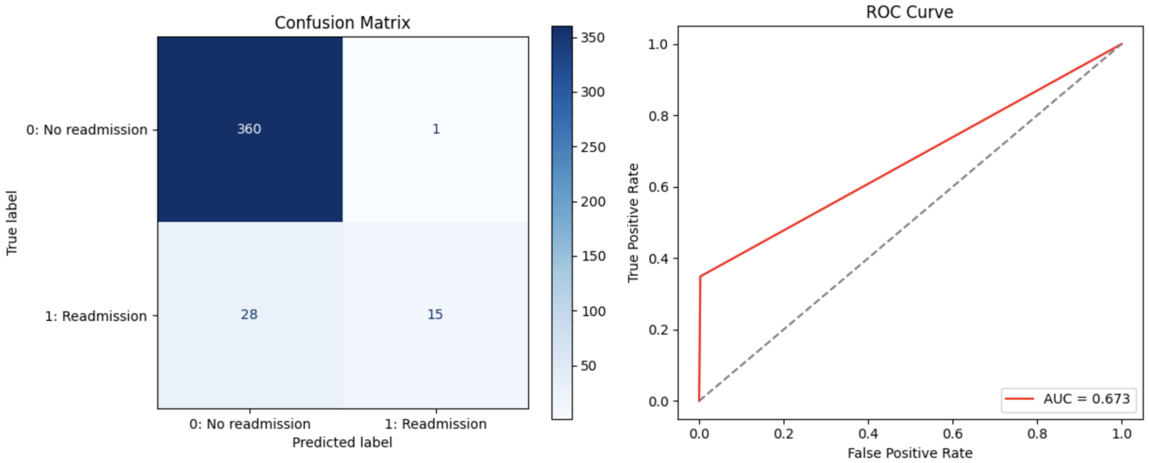

Figure 5: Hypertuned XGBoost Performance

Feature importance for the models trained through XGBoost are shown below.

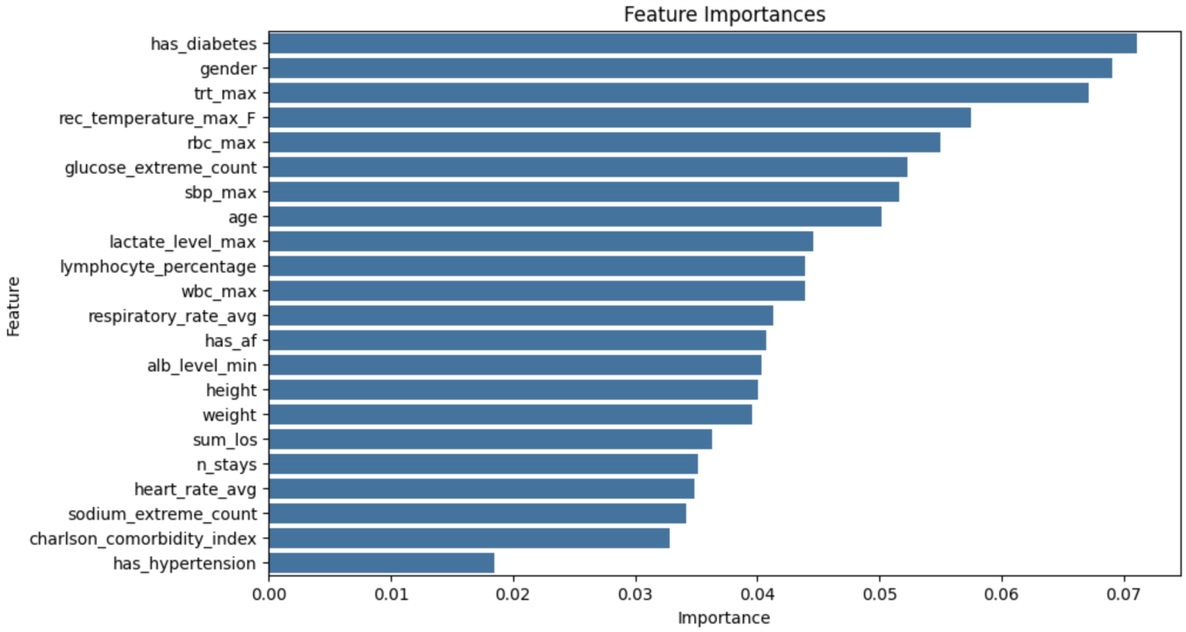

Figure 6: Hypertuned XGBoost Feature Importances

4.5 ML Model Results

The performance of our machine learning models, including Random Forest and XGBoost, was found to be inadequate, with AUC scores of 0.64 and 0.78. This relatively low discriminative ability suggests that the models struggled to effectively distinguish between readmission and non-readmission cases among ischemic stroke patients. Several factors may have contributed to this outcome, including the inherent class imbalance in the dataset and potential limitations in feature representation.

Table 6: Comparison of ML Models

Method	Accuracy	F1 Score	ROC AUC
Random Forest	0.82	0.61	0.64
XGBoost	0.93	0.51	0.78

4.6 Feature Importance Analysis

4.6.1 Clustering Analysis

The clustering analysis provided insights into feature groupings and their collective impact on model predictions. We can derive from the cluster results that the highest influences come from features such as Length of Stays, Troponin T Level, and Systolic Blood Pressure. In the meantime, features like RBC Count, Lymphocyte Percentage also contribute to the prediction of readmission for ischemic stroke patients.

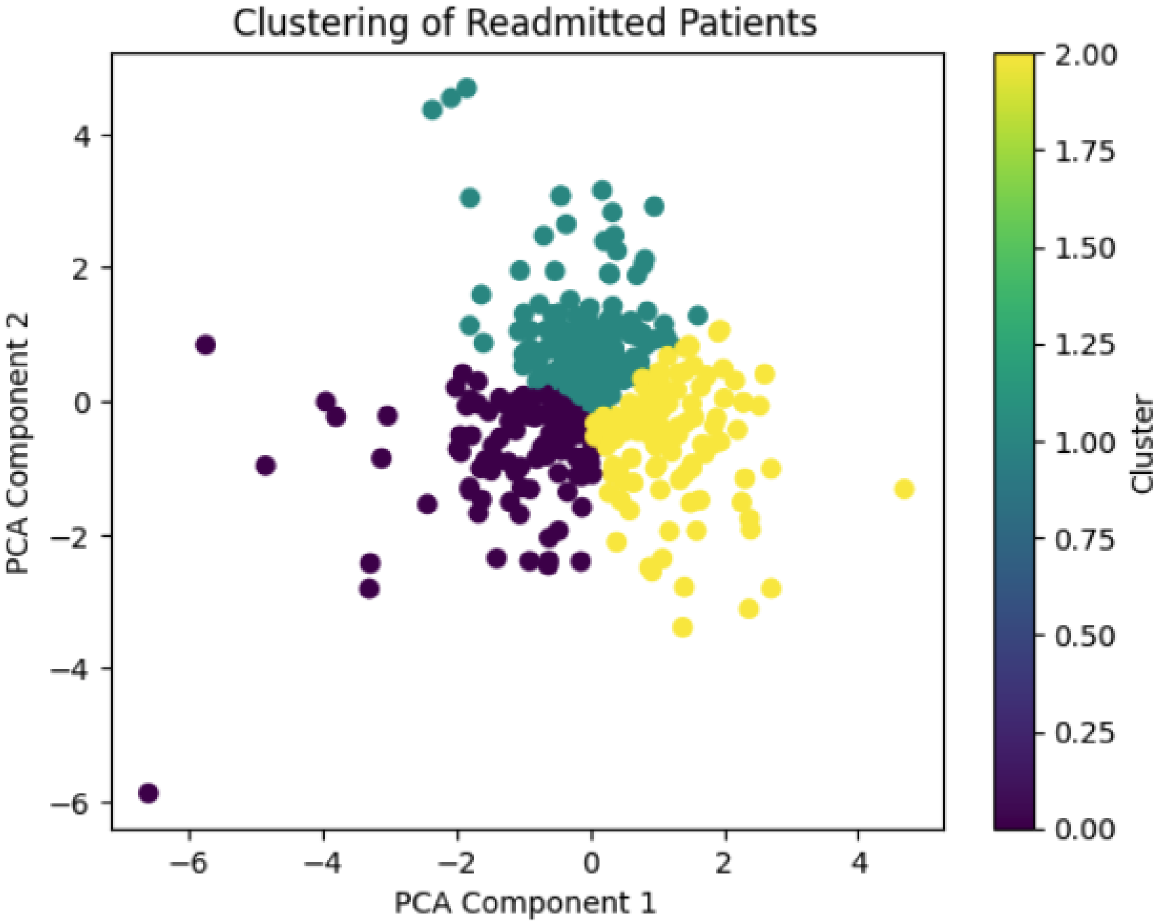

Figure 7: Clustering with PCA

Features in **Cluster 1** showed significant drops in model accuracy when shuffled, underscoring their critical role in predictions.

4.6.2 Permutation Importance

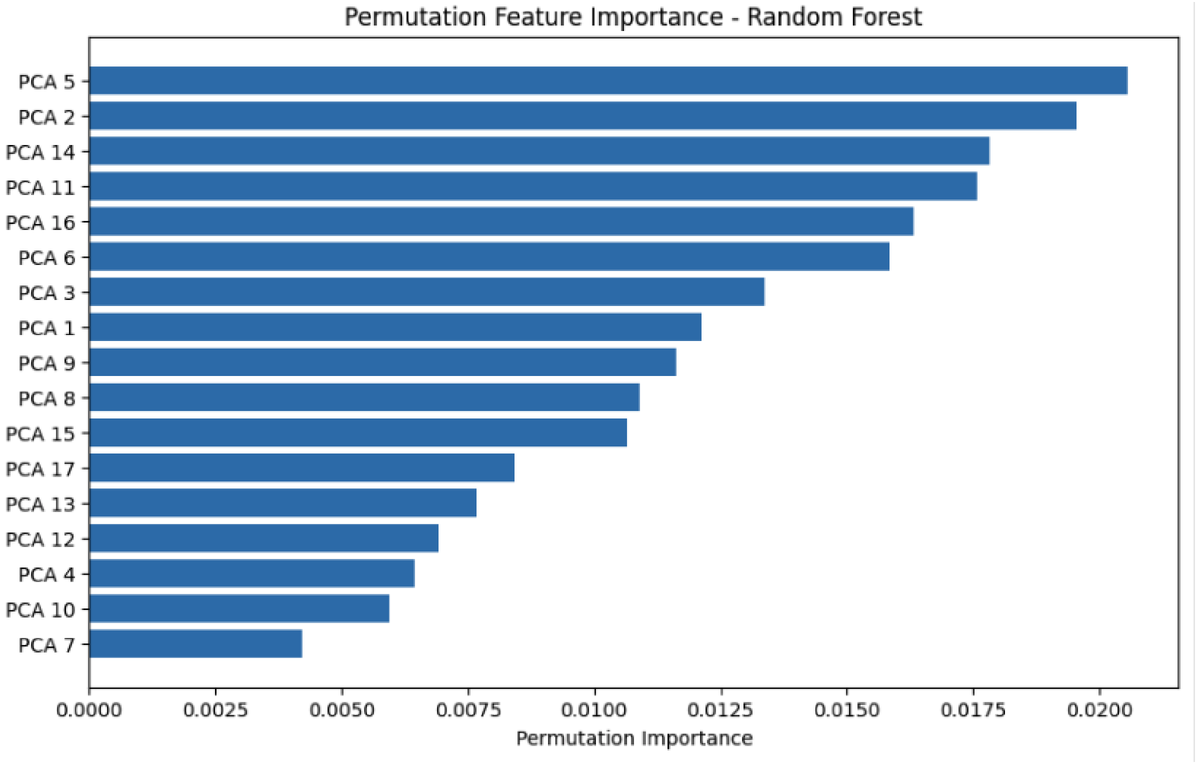

Figure 8: Permutation Feature Importance

5 Discussion

The performance comparison between XGBoost and Random Forest models highlights a trade-off between different evaluation metrics. The XGBoost model achieved a higher accuracy (0.93) and substantially improved ROC AUC score (0.78), compared to that of the Random Forest model, 0.82 and 0.64 respectively. This implies that XGBoost was better at handling class imbalance, using boosting to capture complex patterns effectively and focus on difficult-to-predict minority classes, as our stroke readmission rates meant was only 10 percent of the total cohort.

One important metric in our clinical context is false negatives - where our model predicts no readmission but the patient is in fact readmitted. Due to the clinical severity of a repeated stroke, minimizing false negatives is crucial for a dependable model. Consequently, our XGBoost model was trained using hyperparameter tuning that was aimed at increasing recall. The recall of our model improved from 0.35 before tuning to 0.44 afterwards. Similarly, recall for our baseline Random Forest improved significantly after applying SMOTE. We have learned that improving baseline models through SMOTE and custom hyperparameter tuning is effective in a clinical context to improve recall.

In terms of the most important metric AUC, Random Forest and XGBoost demonstrated only moderate performance, with AUCs of 0.64 and 0.78, respectively. These values suggest that the models are far from being optimal for reliably predicting stroke readmission.

When looking at the feature importance from XGBoost, it is clear that no obvious pattern is present. Seemingly unrelated features such as gender, diabetes, and peak troponin levels are most significant. Clustering analysis provided further insights into feature importance, identifying variables such as length of stay, troponin T levels, and maximum systolic BP as key predictors. Additionally, Random Forest's permutation importance highlighted the impact of feature interactions, underscoring the complexity of factors influencing readmission risk.

Thus, our research was not conclusively able to determine metrics that could be used clinically in a clinicians assessment of a stroke patient's likelihood for readmission.

5.1 Limitations

One limitation of our approach is relying on ICD codes to gather our initial cohort, as these codes are typically assigned at or after patient discharge. This dependency makes our model less suitable for immediate clinical use. As described in our methods, digital phenotyping proved challenging due to the lack of accurate biomarkers and the difficulty of accurately analysing radiology notes to diagnose ischemic stroke. Using digital phenotyping would lead to a smaller cohort with low confidence in the identification of ischemic stroke patients. Thus, we opted for ICD codes to maintain model accuracy.

Although our model isn't directly applicable in real-time clinical settings, it remains valuable for identifying clinically significant features and conditions related to stroke readmission. We prioritized cohort confidence over immediate usability, using ICD codes and NIHSS instead of digital phenotyping.

Another limitation was extending the the readmission window to one year to create a sufficiently large cohort. While 30 days is the most clinically relevant time frame, we prioritized a complete and rich cohort, believing that any findings would still have clinical relevance, especially for identifying predictive features.

5.2 Learnings

Our analysis highlighted the trade-offs between accurate cohort identification and clinical usability. While using ICD codes provided confidence in correctly identifying ischemic stroke patients, it restricted the model's real-time applicability, as these codes are assigned at or after discharge. Additionally, our exploration of digital phenotyping revealed the complexities of using unstructured data like radiology notes and the lack of precise biomarkers, making it challenging to establish an accurate, smaller cohort.

5.3 Future Work

Future work could explore real-time data integration, such as lab results or continuous monitoring metrics, to enhance the model's clinical utility. Additionally, refining digital phenotyping techniques and leveraging advanced natural language processing for radiology notes may improve early identification and prediction of at-risk patients.

6 Conclusion

In this study we explored ML models to predict readmission for ischemic stroke patients using the MIMIC-IV dataset. We employed Random Forest with SMOTE, and XGBoost with randomized hyperparameter tuning. Our AUC scores were 0.64 and 0.78 respectively. Due to the clinical severity of false negatives in our clinical context, we investigated approaches to maximize recall. While we were able to improve recall through applying SMOTE and custom hyperparameter tuning, our results are not reliable enough to be considered for a clinical setting. Finally, our attempts to identify features that might be clinically relevant in stroke readmission outcomes were unsuccessful. Our EDA involved clustering and PCA, however was unsuccessful in establishing predictive features. These results suggest that accurately predicting readmission in patient populations remains challenging.

Our results have highlighted the complexity involved in modeling clinical data. In addition, our struggles to establish a cohort that is large, complete, and relevant clinically proved deeply challenging. Prior literature on different datasets suggested ML accuracy for stroke readmission obtained only moderate results. Our findings were in line with this.

7 Contribution

Section	Joseph S	Sarathi T S	Nadya U	Xuan W
Conceptualization	✓	✓	✓	✓
Visualisation		✓		✓
Query	✓	✓	✓	✓
Dataset Analysis		✓		
Model Coding	✓		✓	
Results Analysis	✓		✓	✓
Phenotyping	✓	✓	✓	✓

Table 7: Author Contributions by Section

References

[1] D. Kuriakose and Z. Xiao, "Pathophysiology and treatment of stroke: Present status and future perspectives," *International Journal of Molecular Sciences*, vol. 21, no. 20, 2020. [Online]. Available: https://www.mdpi.com/1422-0067/21/20/7609

[2] C. Hui, P. Tadi, M. Z. K. Suheb, and L. Patti, "Ischemic stroke," in *StatPearls [Internet]*. StatPearls Publishing, 2024.

[3] L. W. Zhou, M. G. Lansberg, and A. de Havenon, “Rates and reasons for hospital readmission after acute ischemic stroke in a us population-based cohort,” *Plos one*, vol. 18, no. 8, p. e0289640, 2023.

[4] J. C. Rojas, K. A. Carey, D. P. Edelson, L. R. Venable, M. D. Howell, and M. M. Churpek, “Predicting intensive care unit readmission with machine learning using electronic health record data,” *Annals of the American Thoracic Society*, vol. 15, no. 7, pp. 846–853, Jul 2018.

[5] C. R. Fehnel, Y. Lee, L. C. Wendell, B. B. Thompson, N. S. Potter, and V. Mor, “Post–acute care data for predicting readmission after ischemic stroke: a nationwide cohort analysis using the minimum data set,” *Journal of the American Heart Association*, vol. 4, no. 9, p. e002145, 2015.

[6] Y.-C. Chen, J.-H. Chung, Y.-J. Yeh, S.-J. Lou, H.-F. Lin, C.-H. Lin, H.-H. Hsien, K.-W. Hung, S.-C. J. Yeh, and H.-Y. Shi, “Predicting 30-day readmission for stroke using machine learning algorithms: A prospective cohort study,” *Frontiers in Neurology*, vol. 13, p. 875491, 2022.

[7] N. Darabi, N. Hosseinichimeh, A. Noto, R. Zand, and V. Abedi, “Machine learning-enabled 30-day readmission model for stroke patients,” *Frontiers in neurology*, vol. 12, p. 638267, 2021.

[8] C. M. Lineback, R. Garg, E. Oh, A. M. Naidech, J. L. Holl, and S. Prabhakaran, “Prediction of 30-day readmission after stroke using machine learning and natural language processing,” *Frontiers in Neurology*, vol. 12, p. 649521, 2021.

[9] J. Lv, M. Zhang, Y. Fu, M. Chen, B. Chen, Z. Xu, X. Yan, S. Hu, and N. Zhao, “An interpretable machine learning approach for predicting 30-day readmission after stroke,” *International journal of medical informatics*, vol. 174, p. 105050, 2023.

[10] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, “Mimic-iv, a freely accessible electronic health record dataset,” *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[11] American Stroke Association, “Common diagnosis methods,” 2023, accessed: 2024-11-02. [Online]. Available: https://www.stroke.org/en/about-stroke/types-of-stroke/common-diagnosis-methods

[12] M. I. Miller, A. Orfanoudaki, M. Cronin, H. Saglam, I. So Yeon Kim, O. Balogun, M. Tzalidi, K. Vasilopoulos, G. Fanaropoulou, N. M. Fanaropoulou *et al.*, “Natural language processing of radiology reports to detect complications of ischemic stroke,” *Neurocritical care*, vol. 37, no. Suppl 2, pp. 291–302, 2022.

[13] C. C. Aggarwal, *Outlier Analysis*. Springer, 2015.

[14] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, 2004.

[15] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[16] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.

[17] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

[18] V. Y. Kulkarni, P. K. Sinha, and M. C. Petare, “Weighted hybrid decision tree model for random forest classifier,” *Journal of The Institution of Engineers (India): Series B*, vol. 97, pp. 209–217, 2016.

[19] K. Lin, Y. Hu, and G. Kong, “Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model,” *International journal of medical informatics*, vol. 125, pp. 55–61, 2019.

[20] D. Wang, J. Li, Y. Sun, X. Ding, X. Zhang, S. Liu, B. Han, H. Wang, X. Duan, and T. Sun, “A machine learning model for accurate prediction of sepsis in icu patients,” *Frontiers in public health*, vol. 9, p. 754348, 2021.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[22] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[23] S. M. Lundberg and S.-I. Lee, “Explainable machine learning predictions with shap (shapley additive explanations),” *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2018.

Appendix

The code for this project can be found at the following GitHub repository: `https://github.com/tssarathi/comp90089`