COMP90089 - Project Proposal

Group 17

Joseph Suckling

Sarathi Thirumalai Soundararajan

Nadya Ulibasa

Xuan Wu

1 Research Objective

The goal of our project is to utilize supervised machine learning technologies to predict short term (less than 30 days) hospital readmission rates for patients diagnosed with ischemic stroke.

Stroke is the second leading cause of death worldwide (Kuriakose & Xiao, 2020), while ischemic stroke accounts for over 87% of strokes in the US (Zhou, Lansberg, & de Havenon, 2023). Short term readmission after ischemic stroke is common – occurring in around 17.4% of patients – costly, and often detrimental to patient health outcomes (Zhou, Lansberg, & de Havenon, 2023). Improved understanding of readmission occurrence and likelihood will help clinicians improve patient outcomes and avoid unnecessary hospital expenditure.

2 Data Source & Phenotyping

The MIMIC-IV dataset will be our primary data source for this project. Due to the severe nature of ischemic stroke, our project will identify stroke patients from within the intensive care unit (ICU) records. MIMIC-IV contains data for over 65,000 ICU patients, including a vast array of relevant clinical data.

We will establish a stroke cohort through the use of ICD codes assigned by clinicians. We will narrow this to an ischemic stroke cohort through the creation of a digital phenotype. Despite few direct biomarkers existing to distinguish ischemic stroke, text-mining techniques are reliable in identifying ischemic stroke patients (Sung, Lin, & Hu, 2020), this will be augmented by

complementary biomarkers such as D-dimer (Soomro, Guerchicoff, Nichols, Suleman, & Dangas, 2016).

Readmission data for this cohort will be determined from the emergency department dataset.

We will build a feature set utilizing the following pertinent categories (examples features are not exhaustive):

- Infection: WBC count, CRP levels
- Comorbidities: hypertension, diabetes – glycosylated hemoglobin (Lv et al., 2023), AF
- Demographics: age, gender
- Cardiovascular laboratory data: serum levels of cardiac enzymes such as creatine kinase-MB (CK-MB) component and troponin T (Radhakrishnan et al., 2021).

Data cleaning and preprocessing will be performed on all data to standardize and ensure consistency. Duplicates will be removed, missing values handled, and features will be scaled.

3 Methodology

Our approach is to develop a predictive model for hospital readmission in intensive care (ICU) patients diagnosed with ischemic stroke (brain attack). Our methodology is as follows:

1. Data Pre-processing: Extract and clean data including outlier detection, missing data filling, data normalisation and format encoding.
2. Exploratory Data Analysis (EDA): Analyse and plot data distributions, then identify correlations and trends.
3. Addressing Class Imbalance: Since readmission might be a rare event, address techniques such as Synthetic Minority Over-sampling Technique (SMOTE) in the minority class and undersampling in the majority class.
4. Feature Selection: Compare the correlation extent among features to reduce multicollinearity and apply Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) (Goyal et al., 2020) to select and create relevant features.

5. Data Splitting: Divide the dataset into training, validation, and testing subsets. Use k-fold cross-validation to ensure the model is not overfitting and also check the class balance in all subsets with stratified splitting.
6. Model Selection: Conduct experiments with supervised learning models, such as Logistic Regression, Random Forest, Support Vector Machine, and Neural Network, to select the best model in terms of performance and generalisation.
7. Hyperparameter Tuning: Apply cross-validation and Bayesian Optimization techniques to find optimal hyperparameters for our selected model.
8. Model Evaluation and Comparison: Assess the model's performance on the test dataset using metrics including accuracy, precision, recall, F1-score, ROC-AUC score, confusion matrix and calibration curves. Compare performance among models.

9. Visualisation and Interpretation: Visualise results of the model and interpret feature importance by SHapley Additive exPlanations (SHAP) Values to identify significant predictors of readmission (Gebreyesus et al., 2022).

3.1 Performance Metrics

We will evaluate the performance of our model using the following metrics:

- Accuracy: Show overall correctness of the prediction model.
- Precision: Helps minimise wrongly predicting readmission.
- Recall: Ensure the model has a high true positive rate.
- F1-score: Balances precision and recall, suitable for unbalanced data.
- ROC-AUC: Evaluates the model's ability to distinguish potential readmission.
- Calibration curves: Visualize if the model under/overestimates the readmission risk.

3.2 Expected Outcomes

We expect the following outcomes from this project:

- A model with improved predictive accuracy for readmission risk.
- Enhancement of Clinical Decision Support Systems (CDSS) within the ICU setting: The insights generated by the model could also guide early identification of high-risk patients and post-discharge care plans, helping to tailor follow-up care.

References

[1] Kuriakose, D., & Xiao, Z. (2020). Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. International journal of molecular sciences, 21(20), 7609. https://doi.org/10.3390/ijms21207609

[2] Zhou, L. W., Lansberg, M. G., & de Havenon, A. (2023). Rates and reasons for hospital readmission after acute ischemic stroke in a US population-based cohort. PloS one, 18(8), e0289640.

[3] Sung, S. F., Lin, C. Y., & Hu, Y. H. (2(Sung, Lin, & Hu, 2020)020). EMR-Based Phenotyping of Ischemic Stroke Using Supervised Machine Learning and Text Mining Techniques. IEEE journal of biomedical and health informatics, 24(10), 2922–2931. https://doi.org/10.1109/JBHI.2020.2976931

[4] Soomro, A. Y., Guerchicoff, A., Nichols, D. J., Suleman, J., & Dangas, G. D. (2016). The current role and prospects of D-dimer biomarker. European Heart Journal-Cardiovascular Pharmacotherapy, 2(3), 175–184.

[5] Lv, J., Zhang, M., Fu, Y., Chen, M., Chen, B., Xu, Z., Yan, X., Hu, S., & Zhao, N. (2023). An interpretable machine learning approach for predicting 30-day readmission after stroke. International Journal of Medical Informatics, 174, 105050. https://doi.org/10.1016/j.ijmedinf.2023.105050

[6] Radhakrishnan, S., Moorthy, S., Gadde, S., & Madhavan, K. (2021). Role of Cardiac Biomarkers in the Assessment of Acute Cerebrovascular Accident. Journal of neurosciences in rural practice, 12(1), 106–111. https://doi.org/10.1055/s-0040-1721198

[7] Goyal, J., Khandnor, P., & Aseri, T.C. (2020). Analysis of Parkinson's disease diagnosis using a combination of Genetic Algorithm and Recursive Feature Elimination. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (pp. 268-272). IEEE.

[8] Gebreyesus, Y., Dalton, D., Nixon, S., & De Chiara, D. (2022). Machine learning for data center optimizations: Feature selection using SHAP. *MDPI Journal of Data Science*, 9(3), 1-15. https://doi.org/10.3390/jds9030155.