COMP90089 - Project Proposal

Group 23
Harish Kannan - 1534410
Nan Sang - 1335480
Yujie Li - 1174055
Zhuyu Wu - 1413859

Semester 2, 2024

1 Research Objective

The aim of our research project is to employ supervised machine learning techniques on the MIMIC-IV dataset to predict the severity of acute pancreatitis in ICU patients based on their admission condition. Acute pancreatitis, characterized by severe upper abdominal pain, elevated pancreatic enzyme levels and/or cross-sectional imaging findings (Szatmary et al., 2022). This disease is the third most common gastrointestinal cause of hospitalization in the U.S., with a global incidence of 30–40 cases per 100,000 population per year. It also has a mortality rate of 1-5% (Petrov and Yadav, 2019), often due to pancreatic necrosis and persistent organ failure. Given the lack of an internationally approved treatment for acute pancreatitis, the ability to accurately predict the duration and severity of treatment at the time of admission is vital. This project seeks to deliver precise information about patients' expected in-hospital conditions and residence, enabling healthcare facilities to optimize care plans. By prioritizing treatments such as critical care, organ support, parenteral nutrition and antibiotics, hospitals can more effectively provide pain management, prevent complications, and reduce the risk of exacerbating the condition (Vege et al., 2024b).

2 Data Source & Phenotyping

Our primary source for this research project will be the MIMIC-IV dataset. MIMIC-IV is a rich and extensive database that contains de-identified health records of hospitalized patients, offering a wide range of their clinical information.

Our approach towards data processing and digital phenotyping involves multiple steps. Initially we gather relevant data, focusing on specifics required for this project, such as **patient admissions**, **demographics**, **vital signs**, **laboratory results**, **comorbid conditions** and **duration of stay**.

To perform digital phenotyping, we plan to use the following data from MIMIC-IV: An elevation in serum lipase or amylase to three times or greater than the upper limit of normal indicates acute pancreatitis. We evaluate only the serum lipase levels, as they remain elevated for a longer duration and have higher specificity when compared to amylase (Vege et al., 2024a).

The normal range for adults under 60 is usually 10 to 140 U/L, while for those 60 and older, the value ranges from 24 to 151 U/L (University of Rochester Medical Center, 2024). We will extract serum lipase measurements from the dataset and identify patients whose levels exceed the calculated thresholds.

Next, we'll carry out essential data preprocessing, verification, and cleaning to handle missing values, remove duplicates, standardized formats, and ensure the quality of the data.

3 Methodology

3.1 Methodology for Severity Prediction of Acute Pancreatitis

Our approach is to develop a predictive model to assess the severity of acute pancreatitis (AP) in ICU patients based on their admission data, including clinical and diagnostic indicators. The methodology will involve the following steps:

- **Handling Missing Data**: Use imputation techniques or exclude records with excessive missing data.
- **Normalization**: Scale numerical features to ensure consistency across different measurements.
- **Exploratory Data Analysis (EDA)**: Visualize the implemented data and discover some relationships between acute pancreatitis and Diagnostic indicators.
- **Addressing Imbalanced Data**: Given that only 1-5% of AP cases result in mortality, there is a significant imbalance in the dataset with the majority of the cases falling into non-fatal categories. To address this problem, techniques like SMOTE and oversampling will be used to help the model learn more effectively from all types of cases.
- **Feature Engineering**: Create interaction terms, and composite scores, and use domain knowledge to select the most relevant features.
- **Data Splitting**: Divide the dataset into three datasets. The training set for model training, a validation set for hyperparameter tuning, and a testing set for final model evaluation.
- **Model Selection**: Utilize Logistic Regression (LR) as a baseline model to evaluate its performance in predicting the severity of acute pancreatitis (AP). Implement Support Vector Machine (SVM), Random Forest (RF) and Neural Networks (ANN) due to their outperformance compared to the LR (Zhou et al., 2022). We aim to determine the most effective approach for accurate prediction.
- **Hyperparameter Tuning**: Optimize the selected model by applying cross-validate techniques to fine-tune hyperparameters like kernel type (SVM), tree depth (RF), learning rate and layer numbers (ANN).

- **Model Evaluation**: Evaluate the performance on the test dataset using metrics of accuracy, precision, recall, F1-score, and ROC-AUC and calibration curve.
- **Visualization and Interpretation**: Use several visualization techniques to interpret the model results. These methods might include confusion matrix, ROC curve, precision recall curve and decision boundary visualization.

3.2 Performance Metrics

To ensure that the model is clinically useful and reliable, the following metrics we will be employed:

- **Accuracy**: Measure the overall correctness of the model's predictions.
- **Precision**: Minimize false positives, especially for predicting severe outcomes like mortality.
- **Recall**: Ensure the model captures all true positive cases of severe outcomes.
- **F1-score**: Balance precision and recall, particularly in the context of unbalanced data where severe outcomes may be less frequent.
- **ROC-AUC**: Evaluate the model's ability to distinguish between different severity levels, with a focus on binary classifications like fatal vs nonfatal outcomes.
- **Confusion Matrix**: Provide a detailed breakdown of correct and incorrect predictions across all severity classes.
- **Calibration curves**: Assess the accuracy of the model's probability estimates, ensuring that the predicted risks are well-calibrated with actual outcomes.

3.3 Expected Outcomes

We expect the following outcomes from this project:

- **Enhanced Prediction of Acute Pancreatitis Severity**: Develop a relatively accurate machine learning model that helps predict the severity of AP in ICU patients at the time of their admission, aiding clinicians making decisions on conducting medical interventions.
- **Categorizing Pancreatitis Severity and Predicting Hospitalization Duration**: An indicator that can be utilized by doctors, including a categorization of patient pancreatitis severity and the expected duration of hospitalization.
- **Optimised Patient Management**: Facilitate better medical resources allocation and treatment planning within hospitals, reducing ICU stay lengths, mortality rates and identifying risks of persistent organ failure for acute pancreatitis patients.

References

Petrov, M. S. and Yadav, D. (2019). Global epidemiology and holistic prevention of pancreatitis. *Nature Reviews Gastroenterology & Hepatology*, 16(3):175–182.

Szatmary, P., Grammatikopoulos, T., Cai, W., Huang, W., Mukherjee, R., Halloran, C., Beyer, G., and Sutton, R. (2022). Acute pancreatitis: Diagnosis and treatment. *Drugs*, 82(12):1251–1276.

University of Rochester Medical Center (2024). Lipase. University of Rochester Medical Center Health Encyclopedia.

Vege, S. S., Adler, D. G., and Law, K. (2024a). Clinical manifestations and diagnosis of acute pancreatitis. *UpToDate*. Last updated March 20, 2024, Literature review current through July 2024.

Vege, S. S., Gardner, T. B., and Law, K. (2024b). Etiology of acute pancreatitis. *UpToDate*. Literature review current through July 2024, Last updated March 18, 2024.

Zhou, Y., Ge, Y.-t., Shi, X.-l., Wu, K.-y., Chen, W.-w., Ding, Y.-b., Xiao, W.-m., Wang, D., Lu, G.-t., and Hu, L.-h. (2022). Machine learning predictive models for acute pancreatitis: A systematic review. *International Journal of Medical Informatics*, 157:104641.