Severity Analysis of Acute Pancreatitis in Hospitalized Patients using MIMIC-IV Dataset

Harish Kannan, Yujie Li, Zhuyu Wu, Nan Sang
[1]**University of Melbourne, Australia;**

Sem 2, 2024

1. Abstract

Objective: This study is aimed at predicting the severity of Acute pancreatitis in hospitalized patients using machine learning methods, enabling timely intervention by identifying key predictive factors.

Materials and Methods: We used the MIMIC-IV dataset to form the patient cohort. During preprocessing, we addressed missing values and standardized the data. A variety of algorithms were employed, including tree-based, linear, ensemble, and distance-based models, and we used SMOTE to handle class imbalance. Hyperparameter tuning was done using grid search with 10-fold cross-validation to optimize model performance.

Results: CatBoost emerged as the most effective model for predicting the severity of Acute pancreatitis, with the highest accuracy, precision, recall, F1 score, and ROC AUC, indicating that strong predictive power can be achieved without data balancing. In contrast, models such as logistic regression and K-nearest neighbors showed moderate to low performance, while SMOTE had limited impact on improving the overall model effectiveness. Similarly, Random Forest combined with SMOTE achieved the best results in predicting mortality, with the highest F1 Score. The other metrics were also notably high and well-balanced.

Discussion: Our study investigates use of different ML models to predict severity of patients with Acute pancreatitis, emphasizing the need for balance between correctly identifying patients and model usability. CatBoost shows high scores across all metrics for predicting severity, while Random Forest augumented with SMOTE excels in predicting mortality with its high F1 Score. However, reliance on MIMIC-IV data and focusing on specific clinical features may affect the generalizability of the models. This indicates a need for integrating more comprehensive data in future research.

Conclusion: This study demonstrates the importance of precise model selection and featured analysis in predicting the severity and mortality of Acute pancreatitis patients. CatBoost and Random Forest demonstrate promising results for prediction of the same. These models could play a vital role in assisting healthcare providers to reduce AP related complications. Future works should focus on improving robustness and generalizability of the model.

2. Introduction

2.1 Clinical Problem

Acute pancreatitis (AP) is an acute inflammatory disease of the pancreas, frequently presenting with severe epigastric pain, elevated pancreatic enzymes, and cross-sectional imaging findings in the pancreas cite.[1] AP is a leading cause of hospitalization for gastrointestinal diseases, with a global incidence rate of approximately 30–40 cases per 100,000 people per year. Severe cases also have a high mortality rates (1–5%) due to complications like pancreatic necrosis and persistent organ failure.[2]

In addition to identifying AP, assessing the severity of AP is also crucial. Given the lack of an internationally approved treatment for AP, accurately predicting the severity and required treatment duration at an early stage of patient admission could greatly assist healthcare providers in identifying patients who may need intensive interventions (such as critical care, organ support, parenteral nutrition and antibiotics therapy) and optimizing their care plans to play a more vital role in pain management, preventing complications, and slowing disease progression.[3]

2.2 Literature Review

Earlier studies have suggested that the hospital stay duration varies with the severity of AP. One study reported hospital stays of approximately 1.5, 6.9, and 14.2 days for mild, moderate, and severe AP cases, respectively.[4] In another study, mild cases had an average hospital stay ranging from 5 to 13 days, while severe cases required extended stays averaging between 9 and 18 days.[5] These findings indicate that hospital study duration may serve as a useful indicator of AP severity, with notably longer stays observed in patients with severe cases.

Considering the current need to assess the severity of AP, we believe that using machine learning (ML) models holds significant potentials. ML models can mine and analyze large amount of clinical data to uncover deeper relationships and features, offering improved performance compared to traditional AP scoring systems.

In previous studies, ML models have been primarily applied in the following aspects related to AP: complication, mortality, recurrence, severity,and surgical timing.[6] In severity prediction, a study used an Artificial Neural Network (ANN) model to improve the assessment of AP severity, and they found that the ANN model has better performance at capturing nonlinear features and more accurate in distinguishing between mild and severe cases compared to Partial Least Squares Discriminant Analysis (PLS-DA), while their model primarily relied on patients' blood routine parameters and biochemical markers.[7] Another study developed an AP severity prediction model based on logistic regression (LR) by using 11 key blood test features and showed higher sensitivity in detecting severe acute pancreatities cases compared to traditional scoring systems like Ranson's, APACHE-II, and BISAP, reaching an AUC of 0.73.[8] Support vector machine (SVM) was also used to predict AP severity in an early research. They selected some key imaging features to build a contrast-enhanced MRI-based radiomics model for early prediction of AP severity, which significantly outperformed the traditional scoring systems.[9] However, there is no study that incorporate blood marker (White Blood Cell, WBC), biochemical markers, and enzyme markers as features to train multiple ML models for predicting AP severity, although these indicators are significant in traditional scoring systems like APACHE II. Therefore, our study will focus on using these data, along with additional models, to predict the AP severity, which will be primarily distinguished by length of hospital stay and mortality conditions.[10]

3. Methods

This section comprises our entire methodology. The brief summary of our analysis process is shown in Figure 1

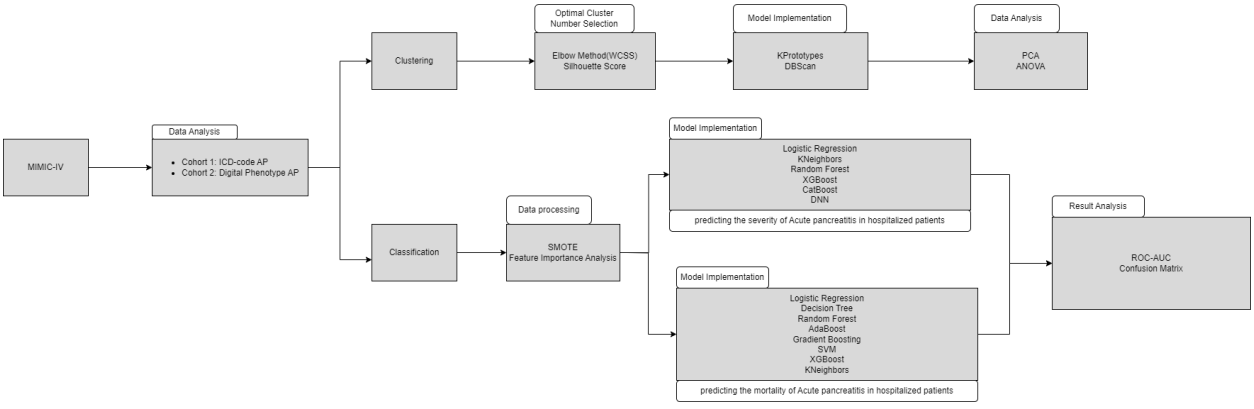

Figure 1. Flowchart summing up the process

3.1 Data sources

The data source is MIMIC-IV, a de-identified ICU and hospital admissions dataset from Beth Israel Deaconess Medical Center.[11] It provides data of patient demographics, lab tests, microbiology, and medication records. MIMIC-IV is HIPAA-compliant and supports clinical research and machine learning applications.

3.2 Cohort

We primarily used digital phenotyping to identify AP patients. To perform digital phenotyping, we used the following criteria: An elevation in serum lipase or amylase to three times or greater than the upper limit of normal indicates AP. We finally evaluated only the serum lipase levels, as they remain elevated for a longer duration and have higher specificity when compared to amylase.[12] The normal range for adults under 60 is usually 10-140 U/L, while for 60 and above, the value ranges from 24 to 151 U/L.[13] We kept only the initial lipase measurement taken after admission for each patient, disregarding subsequent tests recorded during the same admission.

For all the patients, the feature included in our analysis are summarized in Table 1. Weight, height, and vital signs such as heart rate, arterial BP systolic and diastolic, respiratory rate, and SpO2 were not included, as they were only recorded in the ICU data, potentially limiting our dataset size. Additionally, data on CT scans (considered as an important indicator for AP diagonisis) was also excluded due to limited available data.

Besides, previous studies have found that the serum creatinine to albumin ratio can assist in predicting mortality in AP patients. We believe this indicator may also contribute to the prediction of AP severity. Consequently, these two measures have been combined into a single ratio indicator. A total of 2,299 records, including 2,104 unique patients, were used for subsequent data processing. We also retrieved patients' comorbidity ICD codes to get their comorbidity information.

3.3 Data Processing

There are some missing data as well large-value data contributed from multiple lab test. Hence, different data preprocessing methods are used to handle the raw data.

3.3.1 Missing Data : Our initial observation revealed a significant presence of NaN values across many features in the dataset. To address this, we eliminated any features where over 50% of the data were missing. Subsequently, for continuous features, we imputed missing values with the feature's mean. For categorical data, we applied one-hot encoding and then removed any rows that still contained NaN values.

3.3.2 Scaling data : StandardScaler is applied to continuous features to normalize their range, such as lipase levels, which exhibit wide variance and could disproportionately influence feature correlation and subsequent model performance.

3.3.3 Multicollinearity : Addressing multicollinearity is crucial in providing high-quality data for classification models in healthcare settings, as emphasized in past studies.[14] We use the Variance Inflation Factor (VIF) to assess collinearity among numerical features. Most features display VIF values in the desirable range of 0-5, indicating low inter-correlations. However, features like actual_age and Charlson_Comorbidity_Index (CCI) exhibit high VIF values, suggesting potential adverse impacts on the performance of classification models. These features are flagged for possible exclusion during later stages of feature selection.

Variable	Unit	Category
Gender	-	-
Race	-	-
Age	Years	-
In-hospital death	Binary (Yes/No)	-
Length of stay	Days	-
Confirmed AP with ICD code	Binary (Yes/No)	-
Charlson Comorbidity Index (CCI)	Score	-
White blood cell count (WBC)	$10^3/\mu L$	Biochemistry
Creatinine albumin ratio	Ratio	Biochemistry
Anion gap	mmol/L	Biochemistry
Bicarbonate	mmol/L	Biochemistry
Blood urea nitrogen (BUN)	mg/dL	Biochemistry
Calcium	mg/dL	Biochemistry
Chloride	mmol/L	Biochemistry
Glucose	mg/dL	Biochemistry
Sodium	mmol/L	Biochemistry
Potassium	mmol/L	Biochemistry
Total bilirubin	mg/dL	Biochemistry
Lipase level	U/L	Enzyme
Amylase	U/L	Enzyme
Alanine aminotransferase (ALT)	U/L	Enzyme
Alkaline phosphatase (ALP)	U/L	Enzyme
Aspartate aminotransferase (AST)	U/L	Enzyme

Table 1. List of Variables, Units, and Categories

3.4 Feature Selection

We explore the importance of feature for the tree-based model in the digital phenotype, as shown in Figure 2 and the importance of feature for the tree-based model in the ICD code Figure 3. WBC, potassium, glucose, and bicarbonate were considered the most important factors in predicting the severity of acute pancreatitis in hospitalised patients. Both figures showed similar importance of the features with only minor differences in the order of importance.

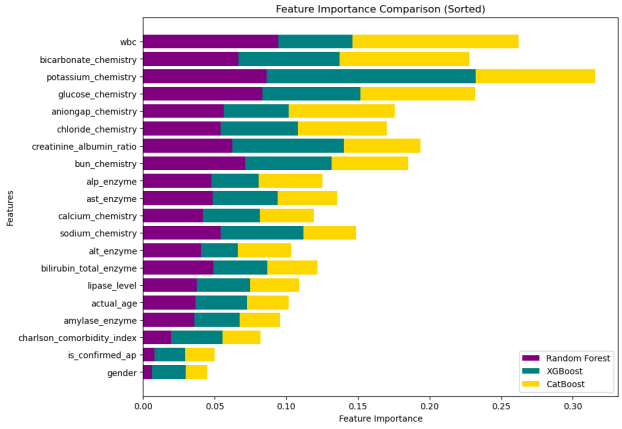

Figure 2. Feature importance for tree-based model on digital phenotype

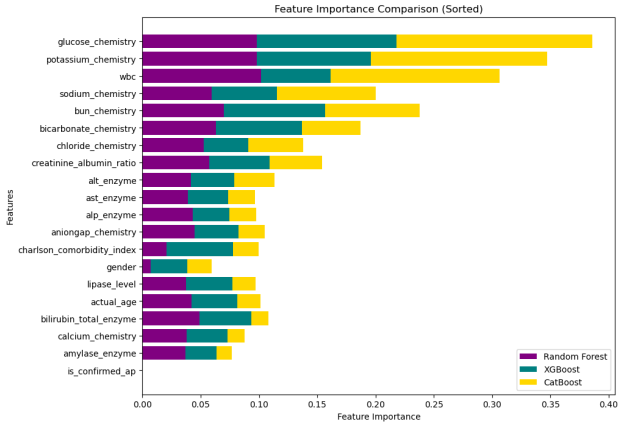

Figure 3. Feature importance for tree-based model on ICD code

It's just that the amount of data containing ICD is much smaller than that using digital phenotypes. For the model, the more data that helps the model, the better the model performance. At the same time, we removed some features that contributed less to the model to ensure that the model is not too complex to learn the model noise. And keep the model as simple and funny as possible, so we finally removed the "charlson comorbidity index", "lipase level", "actual age", "gender", "is confirmed ap", "amylase enzyme" features, and the model results were significantly improved.

3.5 Severity of AP

The histograms depict the distribution of hospital stay lengths for patients with acute pancreatitis, categorized by a binary outcome (survival vs. death). The data shows a positive skew for both groups, particularly among survivors. This skew suggests that while most survivors are discharged relatively quickly, a subset experiences prolonged hospital stays, which may indicate complications or more severe comorbidity of the disease. Conversely, the distribution for deceased patients, characterized by generally shorter stays, hints at rapid clinical deterioration or admission for end-of-life care. To quantify the severity of AP and improve predicting the level of AP progress, we have integrated both the length of stay and mortality outcomes into a severity scoring system. This system classifies severity into categories: 0-1.5 days as mild, 1.5-6.9 days as moderate, and 6.9-14.2 days as severe, with an additional category for deceased patients, reflecting findings from prior studies.[4]

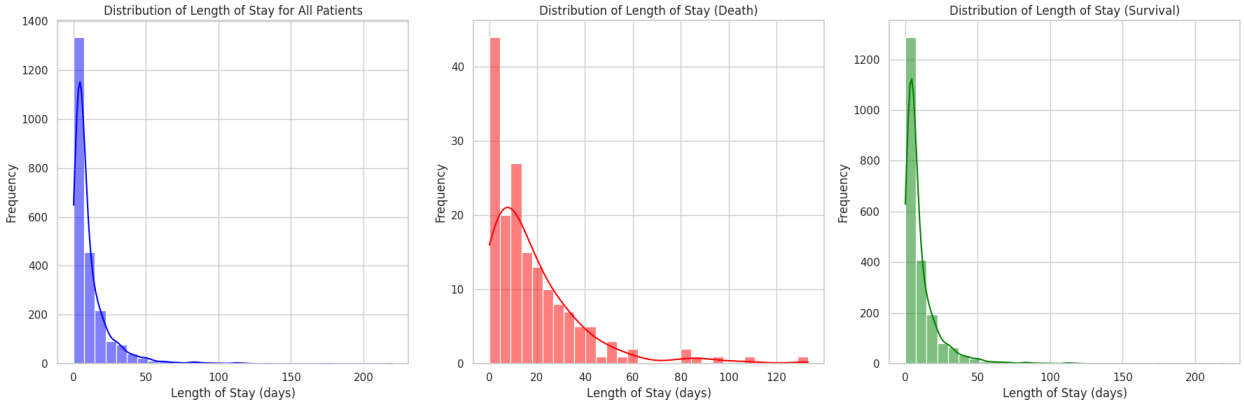

Figure 4. Histogram of Length of Stay for Different Patients

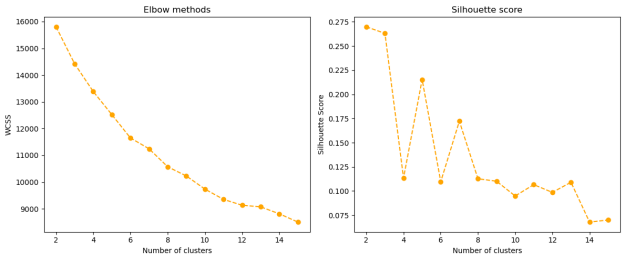

Figure 5. Evaluation result of K-Prototype on digital phenotype cohort

3.6 Clustering analysis

To gain the symptoms that impact the severity of patients, clustering methodologies were employed. The K prototype algorithm with initialization of 'Cao'10 was used. This algorithm is designed to manage datasets comprising both numerical and categorical features by integrating distinct similarity measures corresponding to various data types. In the analysis process, the attributes of gender and severity were considered categorical variables. Meanwhile, other numerical features were preserved, except for the unique hadm_id, which was excluded from the analysis.

Furthermore, the density - based DBSCAN algorithm ($\epsilon = 3$) was also adopted. This algorithm is based on its density-based non parametric approach in cluster discovery. By comparing the parametric K - Prototype clustering method with the non - parametric DBSCAN, the objective was to obtain a deeper understanding of the robustness and reliability of these two types of clustering techniques within this specific context.

$$\text{WCSS} = \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} d(x_j, m_i)$$

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}$$

$$a(x_i) = \frac{1}{|C_i| - 1} \sum_{\substack{x_j \in C_i \\ x_j \neq x_i}} d(x_i, x_j)$$

$$b(x_i) = \min_{j \neq i} \left(\frac{1}{|C_j|} \sum_{x_k \in C_j} d(x_i, x_k) \right)$$

$$\begin{aligned} n_i &: \text{ number of points in cluster } C_i \\ m_i &: \text{ centroid of cluster } C_i \\ C_j &: \text{ the } j\text{-th cluster} \\ d(x_i, x_j) &: \text{ distance between data points } x_i \text{ and } x_j \end{aligned}$$

3.7 Exploratory Data Analysis (EDA) Results

The Pie chart (Figure 2) reveals a higher mortality rate among non-AP patients with elevated lipase and amylase levels compared to those with similar lab results who are diagnosed with AP via ICD codes. This

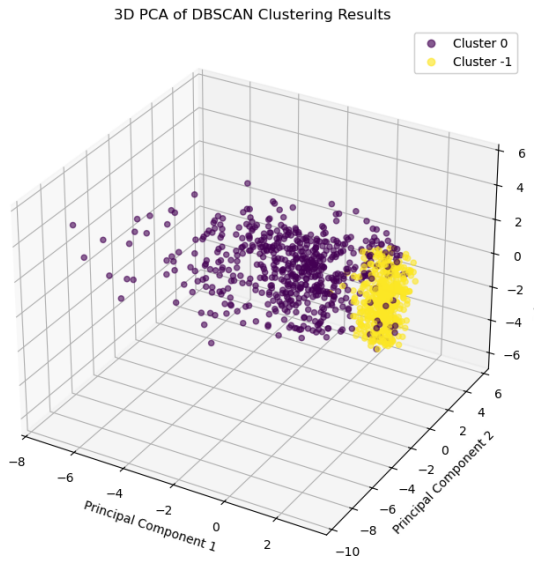

Figure 6. Distribution of clusters from DBSCAN on phenotype

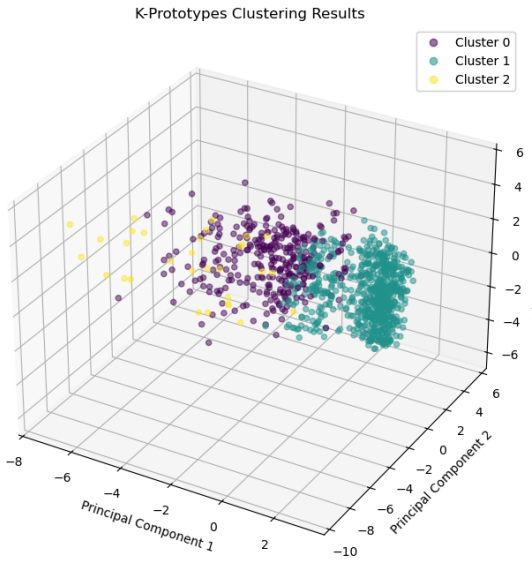

Figure 7. Distribution of clusters from K-Prototype on phenotype

observation prompts further investigation into whether comorbidity also contribute to the severity of AP, particularly among deceased patients, which is discussed more thoroughly in the results section.

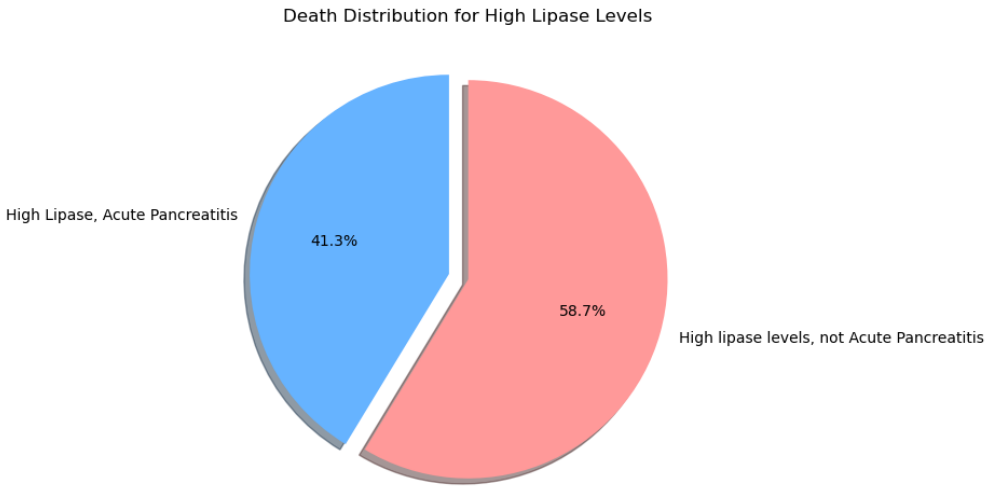

Figure 8. Comparison of mortality rate between AP and non-AP patients

Then Top 10 most occurring diseases based on ICD Code can for patients with high lipase and amylase levels can be observed from the countplot (Figure 3).

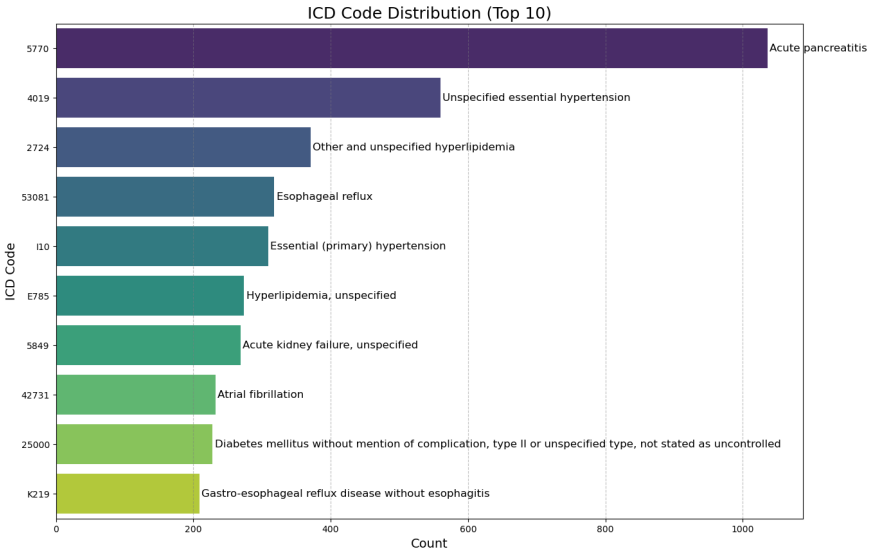

Figure 9. Distribution of Top 10 Diseases using ICD Code

Additionally, we observed a significant label imbalance in the data (Figure 8), particularly in the under-representation of mild AP cases after preprocessing. To address this imbalance in the modeling stage, techniques like SMOTE will be employed alongside models such as K-nearest neighbors (KNN), which struggle to handle imbalanced datasets.

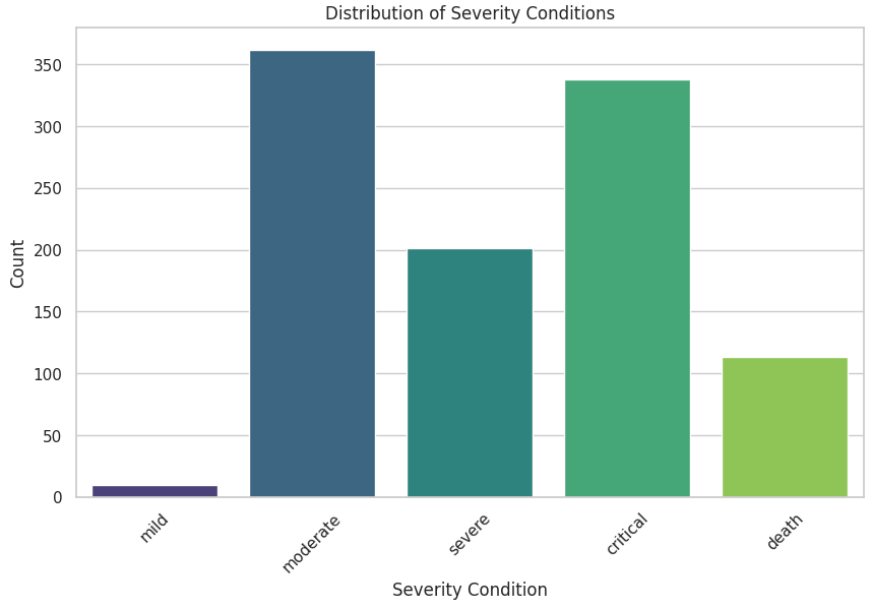

Figure 10. Distribution of Severity of AP after preprocessing

Models

The selected models comprise a mix of linear, tree-based, ensemble, distance-based and deep learning algorithms to achieve an optimal balance between precision and recall, enabling robust predictive performance across various tasks.

Logistic Regression: An ideal baseline model for its quick training and interpretability. It provides probabilistic outputs, making it valuable for understanding the likelihood of clinical outcomes. [15]

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

where $P(y = 1|X)$ is the probability of $y = 1$, β are the regression coefficients, and X are the input features.

Decision Tree: Known for its simplicity and interpretability, this model is particularly useful in clinical decision support systems. [16] It effectively captures non-linear relationships, making it suitable for varied medical data scenarios.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the proportion of class i.

Support Vector Machine (SVM): SVM excels in high-dimensional spaces and is adept at defining complex, higher-order separation planes between diagnostic categories using kernel functions. This capability makes it effective for distinguishing between different health states.

$$\text{maximize } \frac{2}{\|\mathbf{w}\|}$$

subject to the constraint:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i$$

where $\mathbf{w}$ is the weight vector, b is the bias, and y_i is the label.

Ensemble Models (Random Forest and Gradient Boosting): These models are resilient against outliers and help in reducing variance, avoiding overfitting, and enhancing prediction accuracy, especially in handling non-linear medical data complexities. [17]

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

where $h_t(x)$ is the prediction of the t-th tree, and T is the number of trees.

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x)$$

where α is the learning rate, and $h_m(x)$ is the current model's prediction.

K-Nearest Neighbors (KNN) + SMOTE: KNN, paired with the Synthetic Minority Oversampling Technique (SMOTE), addresses class imbalance by generating synthetic examples based on nearest neighbor relationships. This combination is effective for datasets with irregular decision boundaries and is crucial for diagnosing medical cases by identifying patterns in complex datasets. [18]

$$f(x) = \frac{1}{k} \sum_{i=1}^{k} y_i$$

where y_i is the label of the i-th nearest neighbor.

AdaBoost: This model adjusts to challenging data by altering the weights of incorrectly classified instances, enhancing its effectiveness in medical predictions where scenarios can vary widely.

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right)$$

XGBoost: Renowned for its speed and performance, XGBoost handles large, complex datasets efficiently, making it a strong candidate for robust medical data analysis.

$$f(x) = \sum_{k=1}^{K} f_k(x), \quad f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the set of trees, and f_k represents the k-th tree's prediction.

Deep Neural Networks (DNN): DNNs leverage multiple layers of learning to capture intricate patterns in data, beneficial for high-dimensional datasets typically found in medical research.

$$a^{[l]} = g(W^{[l]}a^{[l-1]} + b^{[l]})$$

where $a^{[l]}$ is the activation output of layer l, $W^{[l]}$ is the weight matrix, and $b^{[l]}$ is the bias vector.

CatBoost: This gradient boosting algorithm specializes in handling categorical data with minimal preprocessing and is known for its high accuracy and speed, making it suitable for medical datasets where category-rich features are prevalent.[19]

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x)$$

Evaluation Metrics

Train-Test-Split: The dataset was split into 80% for training and 20% for testing using the train_test_split function, and all the models were trained and tested using the same splits.

The following metrics were employed to evaluate the models to ensure robust and reliable detection of severity of Acute pancreatitis[20]:

Accuracy: This metric provides a general overview of the model's performance across all predictions. Given the imbalance in the dataset, it was not solely relied upon.

Precision: Precision is crucial because a higher precision rate is essential for reducing false positives, which in turn enhances clinical reliability.

Recall (Sensitivity): Recall is vital for accurately identifying all relevant cases of each severity levels of AP, as accurate and timely identification can be crucial for patient management.

F1 Score: F1 Score is the harmonic mean of precision and recall. It serves as a balanced metric, making it suitable for imbalanced datasets.

ROC-AUC: The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) is used to assess the model's ability to discriminate the different level of severity effectively.[21]

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = \frac{FP}{FP + TN}$$

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, d(\text{FPR})$$

Confusion Matrix: This is crucial as it quantifies the accuracy of the predictive models in classifying severity levels, helping clinicians with reducing misdiagnosis.

The model's performance were validated using the testing data to ensure its reliability and generalizability across different patients.

Ethics Statement

This study is carried out using data from MIMIC-IV dataset, which is publicly available but requires users to complete modules in human subjects research to obtain access. We adhered to all data user agreements and ethical guidelines related to MIMIC to ensure responsible handling of sensitive patient data. No sensitive or identifiable information is involved in this study, preserving the anonymity and privacy of individuals. Data generated or processed during this study will not be shared, and all analyses have been conducted with care to respect the confidentiality of the medical records.

Results

To assess the level of severity in patients, we applied six machine learning models: LR, KNN, Random Forest (RF), XGBoost, CatBoost and DNN in the specified phenotype and ICD cohort. Among them, decision tree was used as a baseline model to compare with advanced ensemble learning methods such as XGBoost, RF and CatBoost, as well as neural network models. Before model training, hyper-parameters were optimized by grid search cross-validation. In addition, SMOTE was used for oversampling to address the class imbalance problem in the dataset.

Method	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
LogisticRegression	**0.69**	0.66	0.69	0.66	0.73
LogisticRegression + SMOTE	0.49	**0.69**	0.49	0.55	0.72
KNeighbors	0.59	0.58	0.59	0.58	0.66
KNeighbors + SMOTE	0.44	**0.63**	0.44	0.49	0.64
Random Forest	**0.70**	0.67	0.70	0.67	0.54
XGBoost	**0.70**	0.67	0.70	**0.68**	0.54
CatBoost	**0.75**	0.72	0.75	**0.73**	**0.84**
DNN	0.53	0.42	0.53	0.45	0.75
DNN + SMOTE	0.56	0.59	0.56	0.56	0.74

Table 2. Comparison of model performance metrics for handling class imbalance through synthetic data generation.

Table 2 summarises the performance metrics of each model in predicting the severity of acute pancreatitis. Among the models that we tested, CatBoost scored the highest in all metrics, with an accuracy of 0.75, precision of 0.72, recall of 0.75, F1 score of 0.73, and ROC AUC of 0.84, indicating that it does not require oversampling to obtain strong predictive power.

In contrast, LR and KNN showed moderate result, with a slight improvement in accuracy after applying SMOTE. However a low overall accuracy and F1 score. RF and XGBoost both had an accuracy of 0.70 and an F1 score of around 0.67, but a relatively low ROC AUC of 0.54. It indicating moderate classification strength. DNN have limited effects by Smote, with a slight improvement, increasing accuracy from 0.53 to 0.56, while ROC AUC remains around 0.74-0.75.

Figure 12 shows the ROC curves of multiple models for digital phenotype classification. The ROC curves plot the True Positive rate against the False Positive rate at various threshold settings, providing a comparative visualization of the discriminative power of each model.

Overall, CatBoost emerged as the most effective model for pancreatitis severity classification in this study, while DNN was another promising model. In contrast, logistic regression provided moderate prediction accuracy, while KNN, XGBoost, and RF performed poorly, and the ROC curves indicated limited classification

utility for this dataset. Although SMOTE improves model accuracy, its implementation does not result in a significant improvement in the model's overall performance. These findings demonstrate the effectiveness of CatBoost in handling the complexity of digital phenotype data in this application.

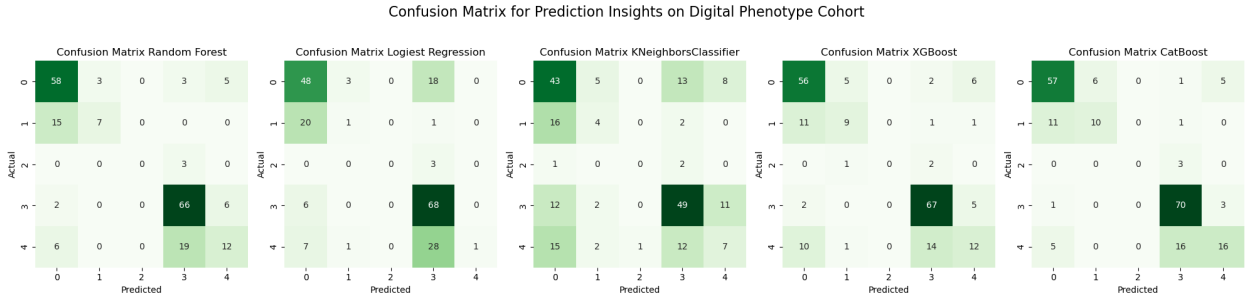

Figure 11. Confusion matrix for prediction severity of acute pancreatitis on digital phenotype cohort

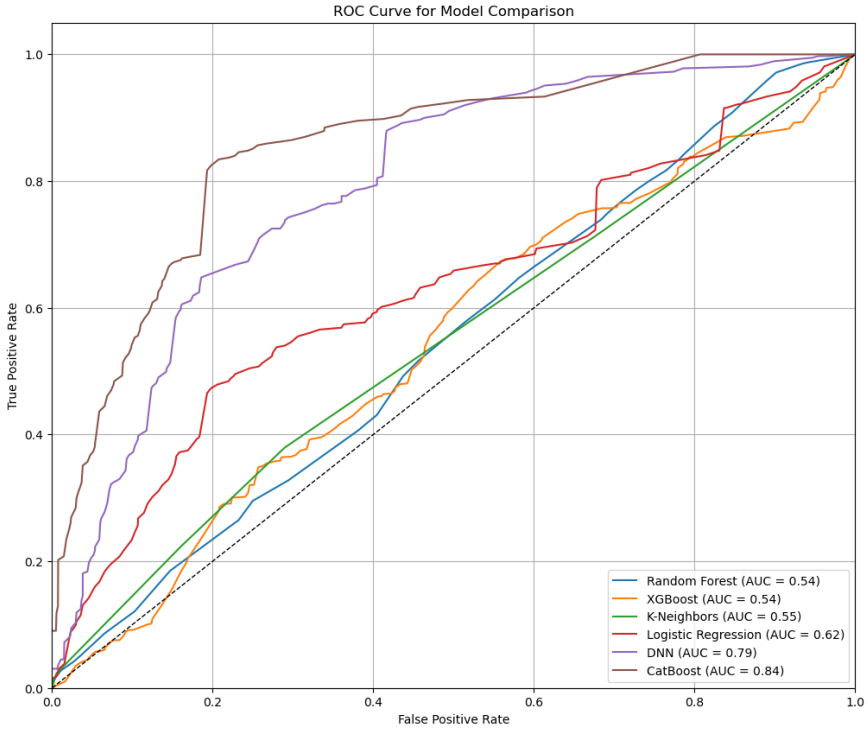

Figure 12. ROC curve for model on digital phenotype

The study also aims to predict mortality among hospitalized patients with Acute pancreatitis and other comorbidities, identifying deadly combinations of diseases with AP using key predictive features through Machine Learning. The prediction task was approached as a supervised learning problem to categorize patients into alive or dead statuses, as well as assess their likelihood of death.

Table 3 summarizes the performance metrics of each model in predicting the mortality status of patients with acute pancreatitis (AP). Among the tested models, RF with SMOTE achieved the highest F1 Score at 0.66, demonstrating an optimal balance between precision and recall. The baseline model LR, on the other hand, recorded an F1 score of 0.64, the highest testing accuracy, and ROC-AUC value at 0.95. It also achieved the highest precision at 0.81, indicating that it was able to capture the patterns well despite its simplistic nature.

Method	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	**0.95**	**0.81**	0.53	0.64	**0.95**
Random Forest + SMOTE	**0.95**	0.69	0.63	**0.66**	0.94
Gradient Boosting + SMOTE	0.94	0.67	0.50	0.57	0.94
SVM + SMOTE	**0.95**	0.68	0.53	0.60	**0.95**
KNN + SMOTE	0.91	0.44	**0.69**	0.54	0.82
Decision Tree + SMOTE	0.91	0.45	0.59	0.51	0.88
AdaBoost + SMOTE	0.93	0.55	0.66	0.60	0.93
XGBoost + SMOTE	**0.95**	0.70	0.59	0.64	**0.95**

Table 3. Comparison of model performance metrics for mortality prediction.

KNN with SMOTE achieved the highest recall score of 0.69, but also recorded the lowest ROC-AUC value of 0.82 among all the models tested.

Figure 13 shows the ROC curves of multiple models for predicting mortality. All models are shown to outperform the baseline (yellow dashed line) representing a model with no discriminative ability (AUC = 0.5), with most providing strong discrimination.

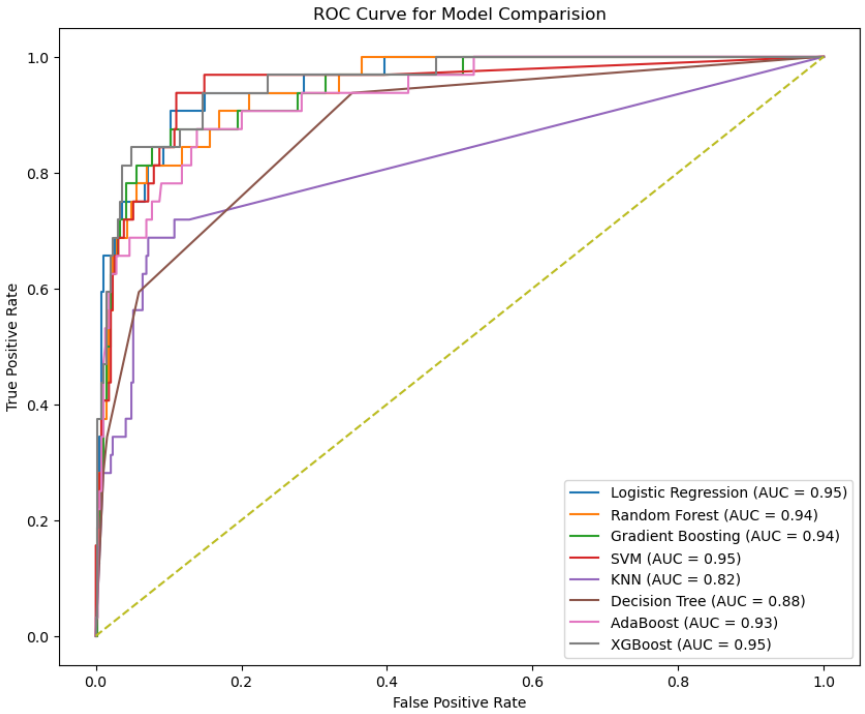

Figure 13. ROC curve for multiple models

In summary, Random Forest augumented with SMOTE is the best performing model. SMOTE with all other models expect Logistic Regression had a significant improvement in F1 Score,but no significant improvement in accuracy and ROC-AUC value.

XGBoost with SMOTE and Logistic Regression demonstrate strong overall performance. These models are particularly effective when handling datasets with numerous features.

The histogram (Figure 7) reveals comorbidity conditions that could significantly increase the mortality rate when they occur together in patients with high lipase and amylase levels diagnosed with AP. In identifying

the top 10 co-occurring medical conditions for patients diagnosed with AP and a mortality rate of 30% and above, acidosis was found to have the highest potential to put the patient's life at risk.[22] This was followed by several variants of acute kidney failure and sepsis. Being aware of these conditions can greatly aid in the treatment of patients initially diagnosed with AP, especially if the patient shows signs of any of these severe medical conditions, or vice versa. This knowledge is crucial to improving the medical outcomes of patients diagnosed with AP.

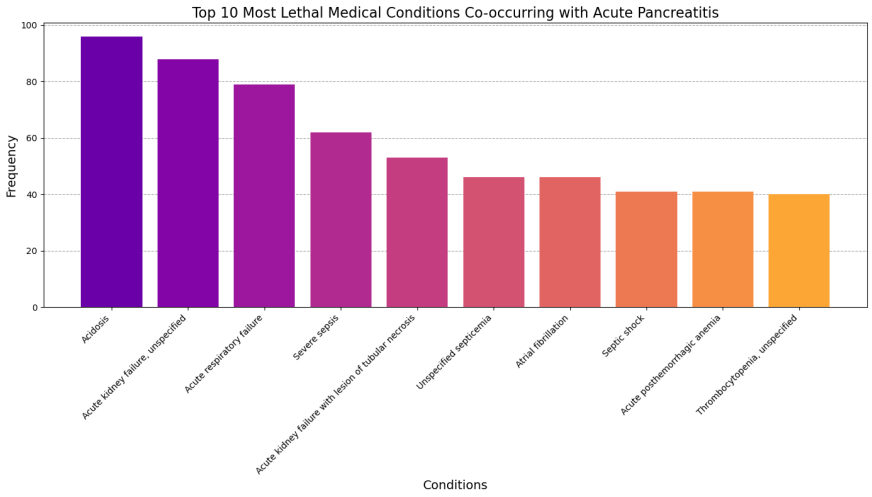

Figure 14. Distribution of Top 10 lethal Diseases with AP

Discussion

In this research, we analyzed the severity of AP in patients using digital phenotype lab results and assessed mortality risk when severe comorbidities were present. For AP severity prediction, the best-performing model was CatBoost, achieving an F1 score of 0.73, while for mortality prediction, RF with SMOTE performed best, with an F1 score of 0.66.

Random Forest models are considered more interpretable than neural networks, as they allow for some understanding of decision paths, which can provide help in explaining predictions. It can also help clinicians understand why the decision was made in this way. Deep Neural Networks (DNN) are generally regarded as 'black box' models because they lack straightforward interpretability. They are more complex and less transparent. Logistic regression could be considered a baseline model due to its simplicity and ease of interpretability, serving as a comparison point for more complex models.

This predictive model holds potential for future adaptation into a real-time application to assess patient risk upon hospital admission. Overall, SMOTE did not consistently enhance model performance, likely due to the limited sample size. Excessive synthetic data in a small dataset can lead to over-fitting, reducing generalizability to new, real-world data.

During EDA, a significant associations between AP severity and comorbidity that increase fatality risk. Hence, this study also led to the discovery of the top 10 co-occurring fatal medical conditions revealing that acidosis along with acute pancreatitis poses a significant threat to the life of patients. This is followed by acute kidney failure due to unspecified reasons and acute respiratory failure which occupy the 2nd and 3rd places, respectively.

Limitation & Future Study

Our research encounters limitations due to label imbalance, which potentially affects the precision of our predictive models for AP severity. While we have implemented the SMOTE technique to mitigate this

issue, the small dataset size still requires further exploration using other techniques, such as under sampling. Additionally, our analysis would benefit from including diverse medical diagnostics, such as abdominal CT scans, to reduce the FPR for AP severity diagnoses. Currently, our dataset is limited to MIMIC-IV, reflecting a single hospital setting and does not include much effective CT scan results. Future studies could enhance generalizability by incorporating real-time patient data from various healthcare settings, such as more ICU data, to broaden the scope and applicability of our findings.

In this study, we developed a predictive model for AP severity using the MIMIC-IV dataset, along with a generic model to estimate fatality rates by analyzing comorbidities. CatBoost emerged as the most effective model for severity prediction, while another model Random Forest + SMOTE demonstrated superior accuracy in predicting mortality rates associated with AP. These findings have the potential to productionize into an online assessor that could predict the eventual development of a patient's AP and significantly enhance clinical decision-making by providing reliable prognostic information. Furthermore, we have showcased an example in the appendix that predicts fatality rates by integrating comorbidity data with specific AP lab results, offering an example of potential practical applications that can predict the possible outcomes of patient outcomes and hence improve the treatment in healthcare settings.

For future research, it is recommended to diversify the data sources to include a patient demographic to enhance the robustness and generalizability of the models. Incorporating a model validation stage could also help in assessing the effectiveness of the predictive models in real-world settings. Moreover, exploring additional clinical features and potentially developing capabilities for real-time severity prediction based on live clinical data could further improve the practical application and effectiveness of the models.

Conclusion

In conclusion, this study demonstrates the potentials of ML models in predicting both the severity and mortality of AP in hospitalized patients. Our results indicate that CatBoost excels in predicting severity by effectively capturing complex feature interactions even without the data balancing techniques like SMOTE. For RF, it performs well in mortality prediction by balancing accuracy and F1 scores when combined with SMOTE. Using a broad range of clinical and biochemical data from the MIMIC-IV dataset, these models provide an early-warning system to help clinicians prioritize and customize care for AP patients. Future work should focus on validating these models across diverse clinical settings (e.g. ICU) and incorporating additional clinical features to enhance model robustness and generalizability, improving patient outcomes in intensive care.

Contributions

Table 4. Author Contributions by Section

Section	Zhuyu Wu	Yujie Li	Harish Kannan	Nan Sang
Conceptualization	✓	✓	✓	✓
Visualisation	✓	✓	✓	✓
Query	✓			
Dataset Analysis	✓	✓	✓	✓
Model Coding		✓	✓	✓
Results Analysis		✓	✓	✓
Phenotyping	✓			

References

1. Szatmary P, Grammatikopoulos T, Cai W, Huang W, Mukherjee R, Halloran C, et al. Acute pancreatitis: Diagnosis and treatment. Drugs. 2022 Aug;82(12):1251–1276.

2. Petrov MS, Yadav D. Global epidemiology and holistic prevention of pancreatitis. Nature reviews Gastroenterology & hepatology. 2019;16(3):175-84.

3. Vege SS, Gardner TB, Law K. Etiology of Acute Pancreatitis; 2024. Literature review current through July 2024, Last updated March 18, 2024. UpToDate.

4. Banday IA, Gattoo I, Khan AM, Javeed J, Gupta G, Latief M. Modified computed tomography severity index for evaluation of acute pancreatitis and its correlation with clinical outcome: a tertiary care hospital based observational study. Journal of clinical and diagnostic research: JCDR. 2015;9(8):TC01.

5. Karim T, Jain A, Kumar V, Kumar RB, Kumar L, Patel M. Clinical and severity profile of acute pancreatitis in a hospital for low socioeconomic strata. Indian Journal of Endocrinology and Metabolism. 2020;24(5):416-21.

6. Zhou Y, Ge Yt, Shi Xl, Wu Ky, Chen Ww, Ding Yb, et al. Machine learning predictive models for acute pancreatitis: a systematic review. International journal of medical informatics. 2022;157:104641.

7. Jin X, Ding Z, Li T, Xiong J, Tian G, Liu J. Comparison of MPL-ANN and PLS-DA models for predicting the severity of patients with acute pancreatitis: An exploratory study. The American Journal of Emergency Medicine. 2021;44:85-91.

8. Sun HW, Lu JY, Weng YX, Chen H, He QY, Liu R, et al. Accurate prediction of acute pancreatitis severity with integrative blood molecular measurements. Aging (Albany NY). 2021;13(6):8817.

9. Lin Q, JI Yf, Chen Y, Sun H, Yang Dd, Chen Al, et al. Radiomics model of contrast-enhanced MRI for early prediction of acute pancreatitis severity. Journal of Magnetic Resonance Imaging. 2020;51(2):397-406.

10. Matull W, Pereira S, O'donohue J. Biochemical markers of acute pancreatitis. Journal of clinical pathology. 2006;59(4):340-4.

11. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data. 2023;10(1):1.

12. Vege SS, Adler DG, Law K. Clinical Manifestations and Diagnosis of Acute Pancreatitis; 2024. Last updated March 20, 2024, Literature review current through July 2024. UpToDate.

13. University of Rochester Medical Center. Lipase; 2024. Accessed: Month Day, 2024. University of Rochester Medical Center Health Encyclopedia. Available from: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=lipase.

14. Sundus KI, Hammo BH, Al-Zoubi MB, Al-Omari A. Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset. Informatics in Medicine Unlocked. 2022;33:101088.

15. Lynam AL, Dennis JM, Owen KR, Oram RA, Jones AG, Shields BM, et al. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagnostic and prognostic research. 2020;4:1-10.

16. Srivastava A, Samanta S, Mishra S, Alkhayyat A, Gupta D, Sharma V. Medi-Assist: A Decision Tree based Chronic Diseases Detection Model. In: 2023 4th International Conference on Intelligent Engineering and Management (ICIEM). IEEE; 2023. p. 1-7.

17. Ansyari MR, Mazdadi MI, Indriani F, Kartini D, Saragih TH. Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Feature Selection. Journal of Electronics, Electromedical Engineering, and Medical Informatics. 2023;5(4):250-60.

18. Karamti H, Alharthi R, Anizi AA, Alhebshi RM, Eshmawi A, Alsubai S, et al. Improving prediction of cervical cancer using knn imputed smote features and multi-model ensemble learning approach. Cancers. 2023;15(17):4412.

19. Safaei N, Safaei B, Seyedekrami S, Talafidaryani M, Masoud A, Wang S, et al. E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database. Plos one. 2022;17(5):e0262895.

20. Owusu-Adjei M, Ben Hayfron-Acquah J, Frimpong T, Abdul-Salaam G. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. PLOS Digital Health. 2023;2(11):e0000290.

21. Stern RH. Interpretation of the area under the ROC curve for risk prediction models. arXiv preprint arXiv:210211053. 2021.

22. Rumbus Z, Toth E, Poto L, Vincze A, Veres G, Czako L, et al. Bidirectional relationship between reduced blood pH and acute pancreatitis: a translational study of their noxious combination. Frontiers in physiology. 2018;9:1360.

Appendix

The code and data used for this study can be found at the following GitHub repository: https://github.com/NanSang2000/COMP90089_MLAH. If you want to visit this GitHub repository, please contact nssan@student.unimelb.edu.au.

Repository Structure

The repository contains the following directories:

1. `dataset/`: This directory contain the cohort dataset extracted from MIMIC-IV.
2. `exploratory/`: Jupyter notebooks for preprocessing and EDA.
3. `predictive_analysis/`: Jupyter notebooks for classfication analysis, including feature engineering, modeling, and evaluation.
4. `clustering_analysis/`: Jupyter notebooks for clustering analysis, including modeling, and evaluation.
5. `output/`: Save trained models, evaluation metrics, and visualizations analysis.