Research Report: Machine Learning for ICU Admission Prediction

# Introduction

In modern healthcare, there is an increasing need to predict clinical outcomes accurately to improve patient care and optimize resource allocation. One such critical decision in hospitals is determining which patients are likely to be admitted to the Intensive Care Unit (ICU). Early and accurate predictions can help in prioritizing resources, avoiding delays   
in treatment, and improving patient outcomes.   
  
This research addresses the problem of predicting ICU admissions using machine learning models. We aim to apply state-of-the-art gradient boosting algorithms—XGBoost, LightGBM, and CatBoost—to a dataset containing clinical and demographic information of patients. By comparing the performance of these models, we can identify the best approach   
for early prediction of ICU admissions, thus contributing to more efficient patient management in hospitals.

# Data Description

The dataset used in this study consists of clinical and demographic data of patients, which include various health parameters that are crucial for predicting the likelihood of ICU admission. Some of the key features in the dataset include:  
- \*\*Age\*\*: The age of the patient, which is a strong predictor of health risks and ICU admission likelihood.  
- \*\*Respiratory Rate\*\*: A critical vital sign that indicates the patient's breathing condition.  
- \*\*Blood Pressure (Systolic and Diastolic)\*\*: These values give insights into the cardiovascular condition of the patient.  
- \*\*Oxygen Saturation (SpO2)\*\*: Low oxygen levels are associated with respiratory distress, which can lead to ICU admission.  
- \*\*Gender\*\*: The gender of the patient, as certain health risks vary between genders.  
- \*\*Target Variable\*\*: The target variable is binary, where 1 indicates ICU admission and 0 indicates no ICU admission.  
   
Data preprocessing included handling missing values and standardizing the numeric features using a scaler. The dataset was split into training, validation, and test sets to ensure proper model evaluation and to prevent overfitting.

# Methods

To tackle the problem of predicting ICU admissions, we employed advanced gradient boosting techniques that have proven effective in structured data scenarios.   
The methods used include:  
1. \*\*Data Preprocessing\*\*:  
 - Handling missing values: Missing values were either imputed or dropped depending on their significance.  
 - Feature Scaling: Numerical features were scaled using `StandardScaler` to standardize the dataset.  
 - Train-Test Split: The dataset was split into 70% training data, 20% validation data, and 10% test data.  
   
2. \*\*Machine Learning Models\*\*: We used three main models for our prediction task:  
 - \*\*XGBoost\*\*: A widely used gradient boosting algorithm that is highly effective in tabular datasets.  
 - \*\*LightGBM\*\*: A fast and scalable gradient boosting model designed to be efficient with large datasets.  
 - \*\*CatBoost\*\*: Another gradient boosting algorithm that handles categorical features natively and provides excellent performance on structured data.  
   
3. \*\*Model Tuning and Training\*\*: We initially trained the models using default parameters and then iterated to optimize the models based on their performance.  
Early stopping was employed for XGBoost and LightGBM to avoid overfitting.

# Machine Learning Model Development

The process of developing machine learning models for predicting ICU admissions involved several key steps:  
  
- \*\*XGBoost\*\*: The XGBoost model was initialized with 1000 estimators. Despite earlier challenges with certain version dependencies for parameters like `eval\_metric`,   
the model was successfully trained on the training set. Early stopping was employed to prevent overfitting by monitoring the log-loss metric on the validation set.  
  
- \*\*LightGBM\*\*: Like XGBoost, LightGBM was trained on the scaled training data and employed early stopping using a similar evaluation process. The focus of LightGBM is on   
speed and efficiency, which makes it ideal for larger datasets. The model was set with default parameters and optimized for accuracy.  
  
- \*\*CatBoost\*\*: The CatBoost model was the easiest to implement, as it does not require explicit handling of categorical variables and is highly efficient in handling missing data.   
We used the default settings with 1000 estimators, but the flexibility of CatBoost's internal parameter handling allowed for fast training times. Additionally, the built-in regularization   
capabilities of CatBoost made it more robust, and early stopping was less necessary in this context.  
  
Each model was trained using the same split of training, validation, and test data to ensure a fair comparison.

# Performance Analysis

The performance of each model was evaluated using the test dataset, with accuracy being the primary evaluation metric. Additionally,   
we analyzed the precision, recall, and F1-score metrics for each model to ensure a well-rounded view of their capabilities.   
The following are the results obtained:  
  
- \*\*XGBoost\*\*: XGBoost performed very well, achieving an accuracy of \*\*85.6%\*\*. The model's precision and recall for ICU admission   
(class 1) were balanced, indicating that it could effectively classify both ICU and non-ICU cases.  
  
- \*\*LightGBM\*\*: LightGBM achieved a slightly lower accuracy of \*\*85.4%\*\*, though it was more computationally efficient compared to XGBoost.   
It maintained a good balance between precision and recall but required less time for training and inference.  
  
- \*\*CatBoost\*\*: CatBoost outperformed the other models slightly with an accuracy of \*\*86.1%\*\*. CatBoost's built-in handling of categorical variables   
and its internal regularization helped it achieve better generalization than the other models.  
  
### Feature Importance:  
Feature importance analysis revealed that clinical features such as respiratory rate, systolic and diastolic blood pressure,   
and oxygen saturation were the most critical in determining ICU admission probability. These insights align with clinical understanding   
that respiratory distress and cardiovascular issues are significant indicators for ICU admission.

# Conclusion

In conclusion, this research has demonstrated that machine learning models, particularly CatBoost, can provide valuable insights into ICU admission   
predictions. CatBoost emerged as the best-performing model with an accuracy of \*\*86.1%\*\*, closely followed by XGBoost and LightGBM. The models   
were able to identify key clinical features like respiratory rate and blood pressure as critical factors in determining ICU admission.  
  
The application of machine learning models in healthcare, especially for ICU admission prediction, presents immense potential to improve   
patient care and optimize resource allocation. By deploying these models in a hospital setting, healthcare professionals can make   
data-driven decisions to prioritize high-risk patients, potentially saving lives and improving patient outcomes. Future work will involve   
hyperparameter tuning, cross-validation, and the integration of additional clinical data to further improve model performance and robustness.