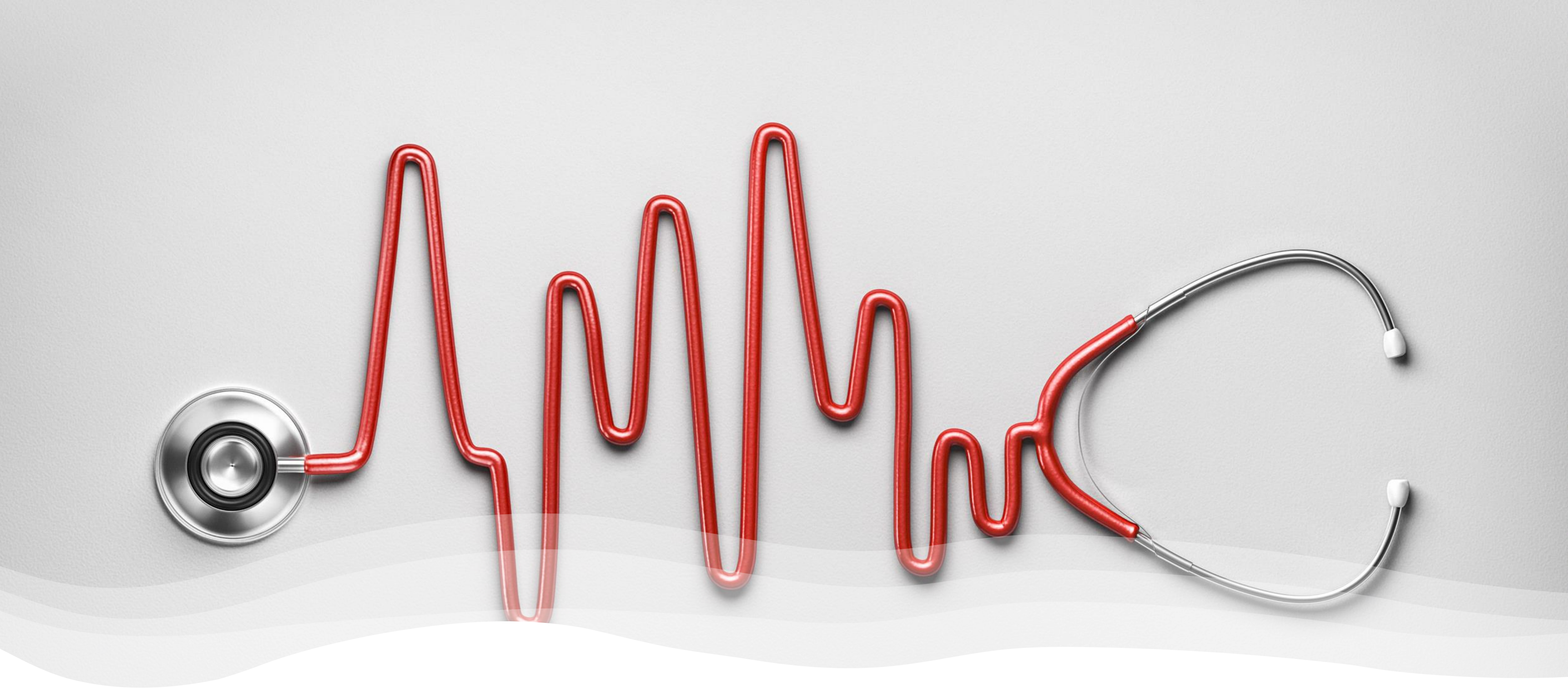


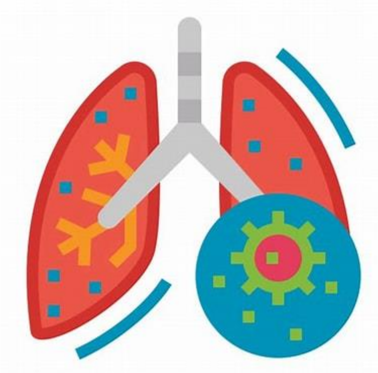
Prediction of Ventilator-Associated Pneumonia in Traumatic Brain Injury Patients Using Machine Learning Algorithms

RAHUL DIPAK TALREJA

Case Background

- **Ventilator-associated pneumonia (VAP):** A common and serious complication in traumatic brain injury (TBI) patients requiring mechanical ventilation.

- **Impact:** Significantly increases morbidity, mortality, hospital stays, and healthcare costs.



Problem Statement:

VAP poses a significant risk to TBI patients, yet its early detection remains a challenge in clinical practice. Traditional methods rely on delayed diagnosis, leading to increased healthcare burden

Objective:

- To explore the potential of machine learning models in predicting VAP early.
- To compare the effectiveness of Random Forest, Logistic Regression, and AdaBoost models for clinical decision support.

Why We Choose This Paper?

The paper tackles a critical clinical problem at the intersection of AI and healthcare.

- **Innovation:** Focuses on applying ML techniques to real-world healthcare data, addressing the limitations of conventional methods.
- **Learning Opportunity:** Provides hands-on insights into building predictive models, evaluating performance metrics, and interpreting results in a healthcare context.

Methodology

Study Population

Traumatic Brain Injury (TBI) patients identified using ICD-9 codes (80,000–80,199; 80,300–80,499; 8500–85,419) .

Exclusion Criteria:

- Missing laboratory test data.
- Missing GCS or vital signs on admission.
- Mechanical ventilation duration <48 hours.

Data Preprocessing

The target variable 'vap' was extracted and merged with the patients data on basis of subject_id and hadm_id. This step comprised of 37 varibales and 3508 observations.

Missing values were identified n categorical variables like gender and ethnicity were were transformed into numeric varibales using **one- hot encoding.**

- **Corelation matrix** was computed for each varibles to the target variable and the variables with weaker corelation (close to 0) and redundunt variables were also removed.
  (chronic_liver_disease, chronic_renal_disease, and white_blood_cell were excluded.)

- Then to **handle missing values**, the variables with more than 30% missing values were treated with ensamble learning- consist of RF, LR and KNN. Variables with lower missing values were imputed using median as it is less sensitive to outliers.

- **Outliers** in numerical variables were deducted using IQR method and removed. Finally, min-maxx scaling was aplied to normalize numerical features to improve the perfomance.

ML Algorithms

The dataset was randomly divide into 80% for the training set and 20% for validation set to execute these 3 ML models- Random Forest, Logistic Regression and AdaBoost were implemented to predict VAP in TBI patients

Performance Metrics:

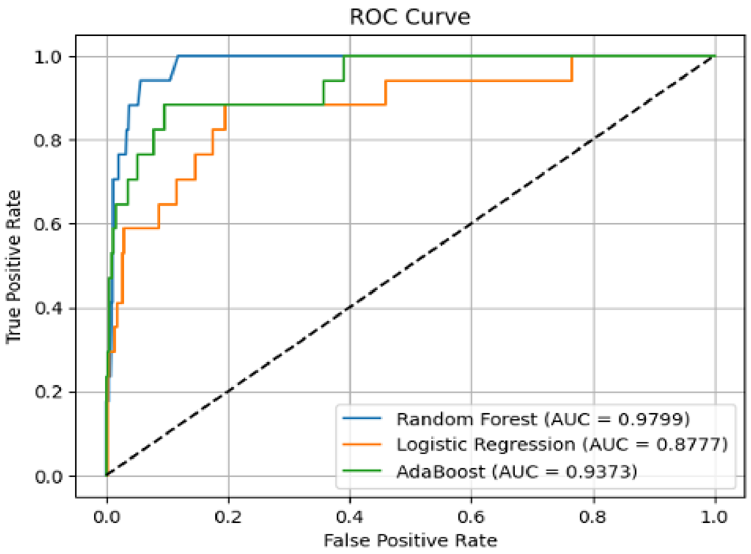
Area Under the Curve (AUC)., Accuracy, Sensitivity, Specificity, and F1 score.

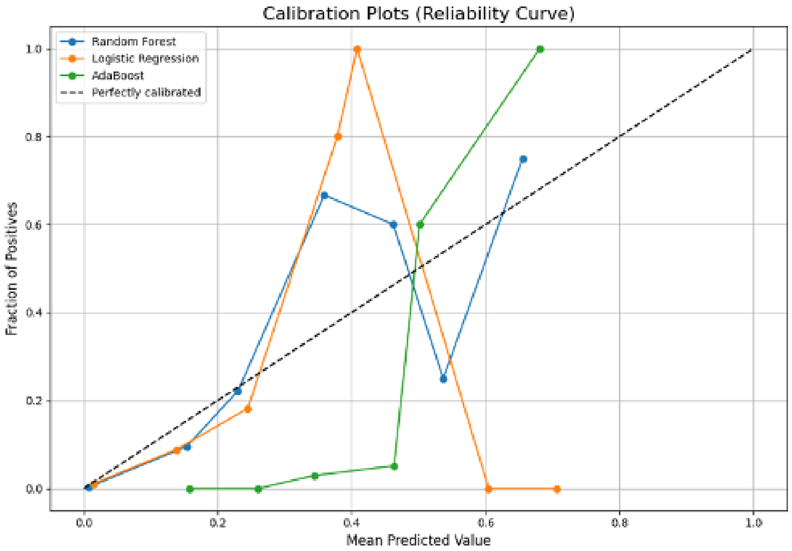
Additional Evaluations:

- Calibration curves: Agreement between predicted and observed probabilities.
- Decision curve analysis: Clinical utility and net benefits.

Results

- **Random Forest** Outperformed all models with the highest AUC (0.9799), demonstrating excellent discriminatory power for identifying VAP cases.
- **Logistic Regression** Struggled with the imbalanced dataset AUC (0.8777) , failing to correctly classify VAP cases, underscoring the importance of handling data imbalance.
- **AdaBoost** Showed moderate performance (AUC: 0.9373) but was sensitive to hyperparameter tuning, leaving room for optimization.





Findings: Why is this important for healthcare?

Sensitivity Challenges:

- While accuracy was high across models, sensitivity remained low, highlighting difficulties in correctly identifying VAP cases.
- Emphasizes the need for improved methods to better detect rare but critical outcomes.

Class Imbalance:

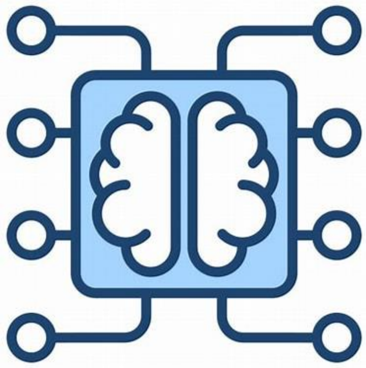
- Imbalanced datasets are common in healthcare and can skew predictions. Addressing this is critical to ensure reliable performance, especially for rare events like VAP.

Value of Ensemble Learning:

- Models like Random Forest and AdaBoost demonstrated robustness with proper tuning and preprocessing, making them suitable tools for clinical decision-making.

What did we learn about data and models?

Good data is key: Having clean and accurate data is essential for building good models.



Choosing the right model: Ensemble models like Random Forest often perform better than simpler linear models in healthcare.

Beyond Accuracy: Sensitivity, specificity, and AUC are essential for evaluating real-world clinical applicability.

THANK YOU