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

Team Mataors- *Rahul Dipak Talreja, Ruchira Kanawade, Peivand Mehrabadi*
BANA 650 Healthcare Analytics, class of 2024
California State University Northridge
December 3, 2024

Abstract

***Background*:** Ventilator-associated pneumonia (VAP) is a serious complication among traumatic brain injury (TBI) patients requiring mechanical ventilation, significantly impacting morbidity, mortality, and healthcare costs. Early prediction of VAP is crucial for guiding preventive strategies and improving patient outcomes.

Objective: This study aims to replicate and extend the research on "Machine Learning Algorithms for Prediction of Ventilator-Associated Pneumonia in Traumatic Brain Injury Patients." The objective is to validate the effectiveness of machine learning in forecasting healthcare outcomes, particularly the risk of VAP in TBI patients.

Methods: TBI patients on mechanical ventilation for over 48 hours were selected from the MIMIC-III database, with VAP identified using ICD-9 codes. Machine learning models like Random Forest, Logistic Regression and AdaBoost were built and evaluated based on AUC, sensitivity, specificity, accuracy, and F-score.

Results: 3508 TBI patients were included from the MIMIC-III. Among the models, Random Forest demonstrated the highest predictive performance with an AUC of 0.9799, indicating superior capability in distinguishing VAP cases. AdaBoost achieved an AUC of 0.9373, showing balanced performance, while Logistic Regression lagged with an AUC of 0.8777.

Conclusion: Random Forest emerged as the most effective model for VAP prediction, offering the highest accuracy and discrimination ability. Its robust performance highlights its potential utility in clinical settings for early risk determination in TBI patients.

Introduction

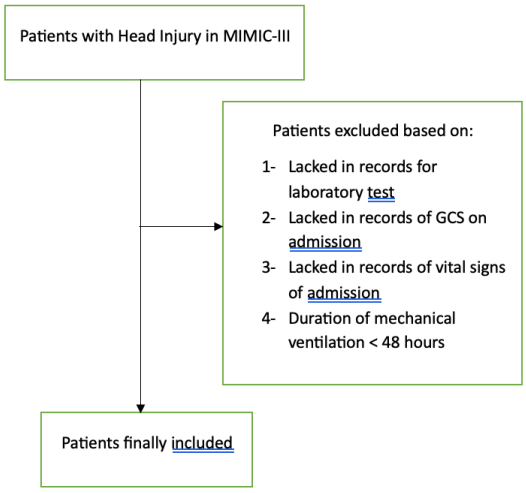
Traumatic brain injury (TBI), a common type of central nervous system damage, remains a global health issue, imposing significant burdens on both the families of affected individuals and societal economies. The outcomes for TBI patients are influenced not only by the severity of the brain injury itself but also by complications outside the brain. These complications can arise from the body's natural physiological responses to the injury or from medical interventions. One such complication is ventilator-associated pneumonia (VAP), a frequent pulmonary issue in TBI patients, with reported incidence rates ranging from 20.4% to 60.6%. Studies have shown that TBI patients who develop VAP face higher

healthcare costs, increased mortality rates, worse neurological outcomes, and longer hospital stays.

Ventilator-associated pneumonia (VAP) remains a significant healthcare challenge, particularly among patients with traumatic brain injury (TBI). Early prediction and intervention can substantially reduce morbidity and mortality rates. Machine learning (ML) has emerged as a promising tool for leveraging electronic health record (EHR) data to predict VAP. The original research explored several ML models for VAP prediction, demonstrating their potential for clinical application. This study aims to replicate and extend the original research, applying and comparing three ML models—Random Forest, Logistic Regression, and AdaBoost—to evaluate their performance in predicting VAP outcomes in TBI patients. Our analysis focuses on model-specific performance metrics, healthcare implications, and limitations.

Methods

The patients in this study were sourced from the Medical Information Mart for Intensive Care-III (MIMIC-III) database, a publicly available resource developed by the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology. This database contains medical records of patients admitted to Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012 and has received ethical approval from the institutional review boards of both MIT and BIDMC. All patient data in MIMIC-III is anonymized and de-identified to ensure privacy. TBI patients were identified using ICD-9 codes (80,000–80,199; 80,300–80,499; 8500–85,419) and considered eligible for the study. These eligible TBI patients in the MIMIC-III were excluded from this study according to the following criteria: (1) Lacked in records of laboratory test; (2) Lacked in records of GCS on admission; (3) Lacked in records of vital signs on admission; (4) Duration of mechanical ventilation < 48 hours. A total of 3508 TBI patients receiving mechanical ventilation were finally included into the analysis.



Data Preprocessing: The preprocessing phase involved multiple steps to prepare the dataset for analysis and ensure its quality for machine learning modeling. First, the target variable, *VAP*, was extracted and merged with the main dataset using subject_id and hadm_id through a left join. This process resulted in a complete dataset comprising 37 variables and 3,508 rows (based on subject_id).

A thorough assessment of missing values revealed null entries in several columns, including *heart rate*, *systolic blood pressure* (systolic_bp), Glasgow Coma Scale (GCS), tracheostomy, vasopressor use, white blood cell count, hemoglobin, *and* platelet count. To address missing data in categorical variables such as *gender* and *ethnicity*, one-hot encoding was applied using a mapping dictionary to convert categories into numeric values.

Next, a correlation matrix was computed to evaluate the relationships between each feature and the target variable, *VAP*. Variables with weak correlation (correlation values close to 0) and redundant columns with repeated or nearly identical values were removed. Weakly correlated features such as hadm_id_x, hadm_id_y, hyperlipidemia, chronic_liver_disease, heart_rate, chronic_renal_disease, and white_blood_cell were excluded. Similarly, redundant columns like cancer, diabetes_percentage, vasopressor, and others representing percentage values of chronic diseases were also dropped. This refinement resulted in a more concise dataset, retaining only the most relevant features for predicting *VAP*.

To handle missing values, different imputation strategies were applied. Columns with high proportions of missing data, such as *systolic_bp*, *GCS*, and *tracheostomy* (more than 30% missing), were imputed using an ensemble learning model. This ensemble combined predictions from KNN, Random Forest, and Linear Regression, ensuring robust and accurate estimations. For variables with fewer missing values, including *hemoglobin*, *platelet*, and *icustay_length*, median imputation was used as it is less sensitive to outliers.

Outliers in numerical variables were identified and removed using the interquartile range (IQR) method to enhance data quality. Finally, Min-Max Scaling was applied to normalize numerical features, standardizing the data to a consistent range and improving the performance and convergence of machine learning algorithms.

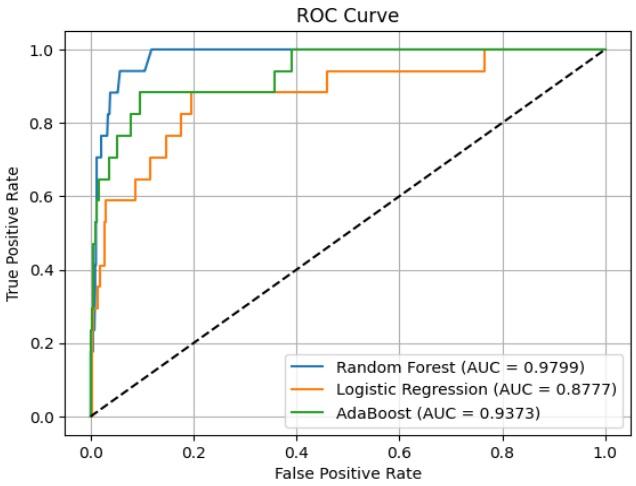
By following these preprocessing steps, the dataset was cleaned, optimized, and prepared for the subsequent development of predictive models. This comprehensive approach ensured the reliability and accuracy of the machine learning analyses.

***Machine Learning Algorithms*:**The dataset was randomly divided into a training set (80%) and a validation set (20%) to develop and test the predictive models. Three machine learning algorithms, including Random Forest, Logistic Regression, and AdaBoost, were implemented to predict the occurrence of *VAP* in TBI patients. These algorithms were trained using the data from the training set, leveraging their unique strengths to handle complex patterns and relationships in the dataset. The trained models were then prepared for evaluation against unseen data in the validation set.

Evaluation Metrics: The performance of the predictive models was assessed using several evaluation metrics to ensure reliability and robustness. Key indices included the area under the receiver operating characteristics curve (AUC), accuracy, sensitivity, specificity, and F1 score. These metrics provided a comprehensive understanding of each model's predictive power and balance between true positives and false positives. Additionally, calibration curves were generated to assess the agreement between predicted and observed probabilities, while decision curve analyses were used to evaluate the clinical utility and net benefits of the models in practical scenarios. This multifaceted evaluation ensured that the models were both statistically sound and clinically meaningful.

Results

The performance of the three machine learning models—Random Forest, Logistic Regression, and AdaBoost—was assessed using various metrics, as shown in the results and the ROC curve. Random Forest achieved the highest AUC (0.9799), demonstrating superior discriminatory power compared to AdaBoost (AUC = 0.9373) and Logistic Regression (AUC = 0.8777). In terms of accuracy, all three models performed similarly, with Random Forest achieving 97.58%, AdaBoost achieving 97.86%, and Logistic Regression achieving 97.29%. However, significant differences were observed in sensitivity and specificity. While all models had high specificity (≥ 99%), their sensitivity varied, with Random Forest and AdaBoost achieving a sensitivity of 24% but Logistic Regression failing to identify any true positives (sensitivity = 0%).



Random Forest Model

Accuracy: 0.9758 **AUC:** 0.9799 **Specificity:** 0.99 **Sensitivity:** 0.24

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.98	0.99	0.99	685
1	0.50	0.24	0.32	17

Confusion Matrix

Actual/ Predicted	0	1
0	681	4
1	13	4

Logistic Regression Model

Accuracy: 0.9729 **AUC:** 0.8777 **Specificity:** 1.0 **Sensitivity:** 0.0

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.98	1.00	0.99	685
1	0.00	0.00	0.00	17

Confusion Matrix:

Actual/ Predicted	0	1
0	683	2
1	17	0

AdaBoost Model

Accuracy: 0.9786 **AUC:** 0.9373 **Specificity:** 1.0 **Sensitivity:** 0.24

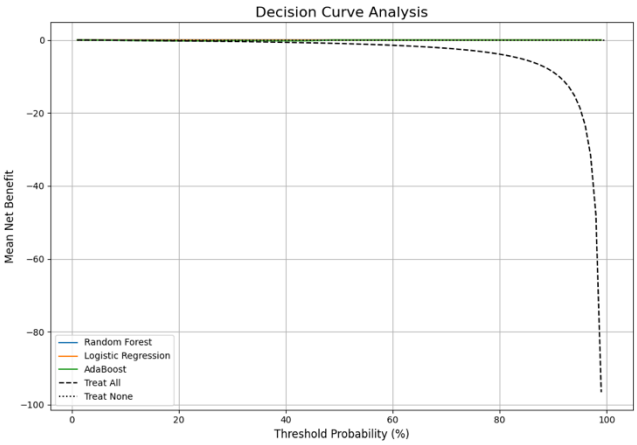
Classification Report:

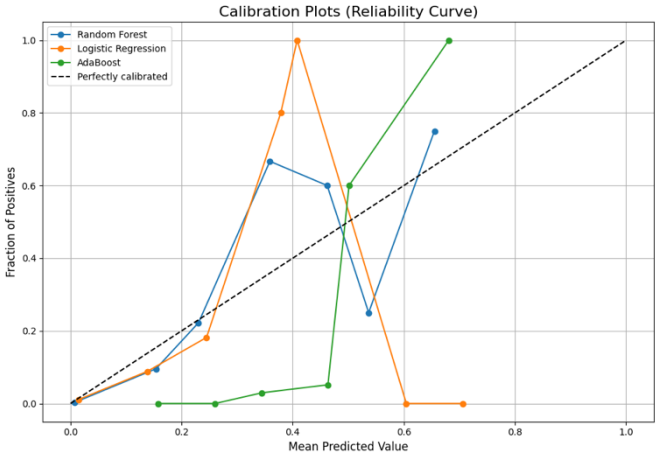
Class	Precision	Recall	F1-Score	Support
0	0.98	1.00	0.99	685
1	0.67	0.24	0.35	17

Confusion Matrix

Actual/ Predicted	0	1
0	683	2
1	13	4

The confusion matrices and classification reports further highlight the trade-off between precision and recall, particularly for classifying positive cases (class 1). Random Forest and AdaBoost demonstrated moderate performance in correctly identifying positive cases, while Logistic Regression struggled to balance precision and recall.





The calibration curves suggest that the Random Forest model is more aligned with observed outcomes, making it the most robust choice for predicting *VAP* in TBI patients. Overall, while Random Forest outperformed other models, improving sensitivity remains critical for practical applications in clinical settings.

Conclusion

Key findings from the replicated study include that the **Random Forest model** outperformed others in terms of AUC, indicating its superior ability to distinguish between VAP and non-VAP cases. Logistic Regression struggled with the imbalanced dataset, failing to classify any VAP cases correctly, which emphasizes the importance of addressing class imbalance in healthcare datasets. AdaBoost showed moderate performance but demonstrated susceptibility to hyperparameter tuning, which could be optimized for better results. These findings highlight the

potential of machine learning models in enhancing predictive accuracy for VAP in TBI patients.

Healthcare Implications

Importance of Sensitivity: While overall accuracy was high, the sensitivity for all models remained low, suggesting challenges in correctly identifying VAP cases. This calls for further work to improve the detection of such critical conditions.

Data Imbalance Challenges: Effective handling of imbalanced datasets is critical to improving predictions for rare but critical outcomes like VAP, ensuring that models can identify even less frequent events accurately.

Ensemble Learning Benefits: Models like Random Forest and AdaBoost can provide robust performance with proper tuning and data preprocessing, making them valuable tools in clinical decision-making, particularly for complex outcomes such as VAP.

Lessons Learned

General Lessons
Data Quality is Paramount: Healthcare datasets often suffer from missing values and class imbalance, requiring robust preprocessing.
Model Selection Matters: Ensemble methods often outperform linear models in complex healthcare scenarios.
Evaluation Beyond Accuracy: Metrics like sensitivity, specificity, and AUC are crucial for understanding real-world applicability.

Analytical Learnings
Random Forest Strengths: This model's ability to handle high-dimensional and imbalanced data makes it a strong candidate for healthcare applications.
Logistic Regression Limitations: Despite its simplicity, logistic regression may not be suitable for imbalanced datasets without significant adjustments.
Boosting Challenges: Hyperparameter tuning significantly impacts boosting models, requiring careful experimentation for optimal performance.

Acknowledgements

Thanks for the contribution of MIMIC-III database's developers.

Instructor- Dr. Akash Gupta
akash.gupta@csun.edu

Heart & Lung - The Journal of Acute and Critical Care 62 (2023) 225–232 Ruoran Wang, MD, Linrui Cai, MD, Yan Liu, BS, Jing Zhang, MD,PHD, Xiaofeng Ou, MD,PHD, Jianguo Xu, MD,PHD

Arize- https://arize.com/blog-course/what-is-calibration-reliability-curve/

Watson K, Reoch J, Heales LJ, et al. The incidence and characteristics of ventilator- associated pneumonia in a regional nontertiary Australian intensive care unit: A retrospective clinical audit study. *Australian Critical Care*. 2022;35(3):294–301