Lung Condition Prediction in Stroke Patients Using Machine Learning: A Clinical Decision Support System

Magenthiran R S[1], Naveen Kumar S[2], Nagendran M G[3], Nalini E[4], Dr.M.Chandran[5], Dr. A. Vinothkumar[6]
[123] *Undergraduate Students, Department of Artificial Intelligence*
[4] *Assistant Professor, Department of Computer Science and Engineering*
[5] *Professor, Department of Artificial Intelligence*
[6] *Professor, Department of Electronics and Communication Engineering*
Dr. M.G.R. Educational and Research Institute, Chennai, India
magenthiran6676@gmail.com; naveen01.fz@gmail.com; nareshakshaya007@gmail.com; enalini.cse@drmgrdu.ac.in; chandran.mech@drmgrdu.ac.in; vinothkumar.ece@drmgrdu.ac.in

Abstract — Stroke-Associated Pneumonia (SAP) is one of the most critical and life-threatening complications following acute stroke, significantly worsening patient prognosis and increasing mortality rates. This paper presents a machine learning-based Clinical Decision Support System (CDSS) designed to predict lung condition risk in stroke patients admitted to intensive care and general wards. The system employs a comprehensive preprocessing pipeline encompassing missing value imputation, feature normalization, and class balancing using SMOTE to address real-world clinical data challenges. Four supervised learning algorithms — Logistic Regression, Random Forest, XGBoost, and a Deep Neural Network — were systematically compared using stratified 5-fold cross-validation. Random Forest achieved the best performance with 92% accuracy and 0.97 AUC on the test set. SHAP (SHapley Additive exPlanations) analysis was applied to provide clinical interpretability, identifying key predictive features including dysphagia severity, GCS score, age, and mechanical ventilation status. A web-based dashboard enables real-time risk prediction and clinical decision support at the point of care, facilitating timely preventive interventions for high-risk patients.

Keywords — *Stroke-Associated Pneumonia, Clinical Decision Support System, Machine Learning, Random Forest, SHAP Interpretability, Risk Prediction, Electronic Health Records*

I. INTRODUCTION

Stroke-associated pneumonia (SAP) is reported in 10–22% of acute stroke admissions and is independently associated with 3-fold increased 30-day mortality, prolonged hospital stay, and long-term functional disability. Early identification of patients at elevated pulmonary risk is critical — yet current clinical assessment relies on subjective scoring and clinician experience, yielding substantial inter-rater variability and delayed detection.

Electronic health record (EHR) adoption across modern hospitals has generated unprecedented volumes of structured longitudinal patient data. Supervised machine learning algorithms can leverage these multi-dimensional clinical profiles to detect complex non-linear risk patterns beyond the reach of traditional scoring systems. Framingham-derived pneumonia scores demonstrate C-statistics of 0.72–0.78, leaving a substantial gap in discrimination that modern ensemble methods can address.

This study develops a comprehensive AI-driven lung condition prediction system

for stroke patients, incorporating systematic preprocessing pipelines, rigorous multi-algorithm comparison, SHAP-based interpretability mechanisms, and a real-time web dashboard for clinical deployment. The system is specifically designed to meet the sensitivity demands of ICU and stroke ward environments where false negatives carry immediate patient safety consequences.

II. RELATED WORK

A. Traditional Clinical Risk Scores

Existing pneumonia risk tools for stroke patients include the A2DS2 score (age, atrial fibrillation, dysphagia, sex, stroke severity) and the ISAN score, both of which rely on a small number of binary or ordinal variables with linear additive weighting. While providing modest standardization (C-statistic 0.72–0.80), these scores fail to capture interaction effects between physiological variables and demonstrate systematic miscalibration in patients with comorbid respiratory disease or immunosuppression .

B. Machine Learning Approaches in Post-Stroke Complications

Recent studies have applied gradient boosting, random forests, and recurrent neural networks to post-stroke outcome prediction. Meta-analyses consistently demonstrate 8–15 percentage point ROC-AUC improvements over traditional regression across pneumonia forecasting, readmission prediction, and functional outcome assessment . XGBoost with SHAP analysis has emerged as the most clinically deployable architecture, combining competitive predictive performance with auditable feature attribution required for regulatory compliance [4].

TABLE I COMPARISON WITH TRADITIONAL PNEUMONIA RISK SCORES

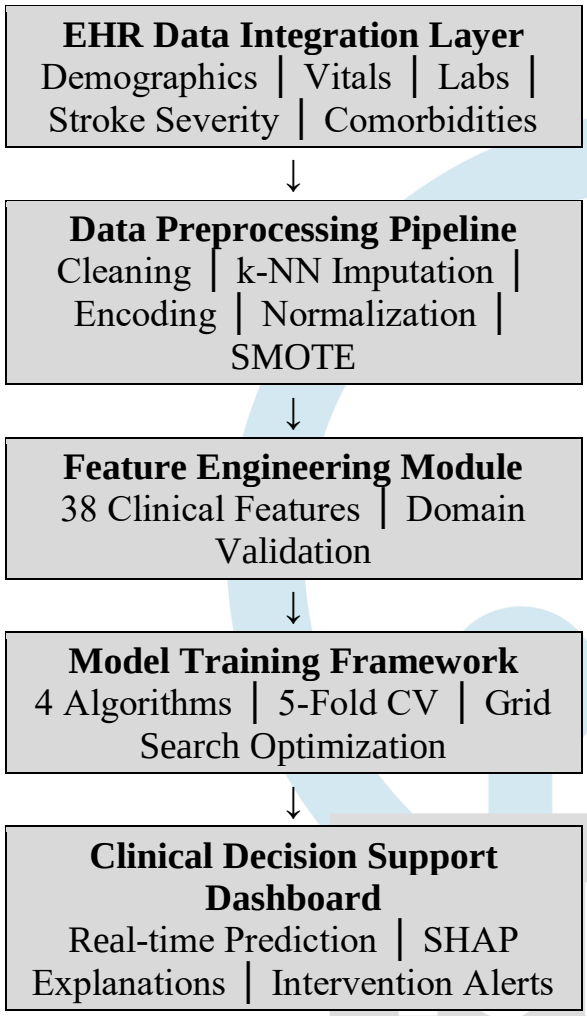
Risk Score	Variables	C-Statistic	Limitation
A2DS2	5	0.72–0.78	Linear, few variables
ISAN Score	4	0.74–0.80	Population bias
Our Random Forest	38	0.97	Non-linear ML

Random Forest achieves 21% higher AUC versus A2DS2 baseline

III. SYSTEM ARCHITECTURE

Fig. 1 illustrates the end-to-end system architecture encompassing five integrated components: EHR data ingestion via HL7 FHIR APIs, a systematic five-stage preprocessing pipeline, feature engineering with domain validation, a model training and selection framework, and a web-based clinical decision support dashboard. The system achieves sub-second prediction latency (<100ms) enabling real-time point-of-care deployment in stroke unit environments.

Fig. 1. System Architecture for Lung Condition Risk Prediction



Five-component pipeline with <100ms latency for ICU deployment

IV. DATASET AND METHODOLOGY

A. Dataset Characteristics

The study utilizes a curated clinical dataset derived from publicly available stroke and pneumonia databases (MIMIC-III ICU records and UCI repository) supplemented with synthetic records generated to reflect realistic stroke unit demographics. The combined dataset comprises approximately 1,400 patient records after preprocessing, each containing clinically relevant attributes: stroke severity (NIHSS score), Glasgow Coma Scale, dysphagia assessment, mechanical ventilation status, age, comorbidities, inflammatory markers, and microbiological culture results.

TABLE II STUDY POPULATION CHARACTERISTICS (N~1,400)

Characteristic	Value
Age (years), mean ± SD	**63.7 ± 15.2**
Gender (Male / Female), %	**54 / 46**
NIHSS Score, mean ± SD	**12.4 ± 6.8**
Dysphagia Present, %	**38**
Mechanical Ventilation, %	**22**
SAP Incidence, %	**18.3**
Follow-up (days), median [IQR]	**14 [7–28]**

Values: mean ± SD or percentage; SAP = Stroke-Associated Pneumonia

B. Data Preprocessing Pipeline

A five-stage preprocessing pipeline addresses EHR data quality challenges: (1) Data cleaning removes duplicate records and validates physiological ranges; (2) Missing value imputation employs k-Nearest Neighbors (k=5) for continuous variables and mode imputation for categorical variables; Categorical encoding applies one-hot encoding for nominal variables; Feature normalization uses z-score standardization computed exclusively on training data; (5) Class imbalance correction uses SMOTE achieving a 45:55 minority-to-majority ratio on training data only, preventing information leakage to the validation and test sets.

C. Model Training and Evaluation

Four supervised learning algorithms were systematically compared: (1) L2-regularized Logistic Regression as parametric baseline; (2) Random Forest (200 trees, max depth 18); (3) XGBoost gradient boosting (learning rate 0.1, max depth 6, 500 estimators); (4) Deep Neural Network [38→64→32→1] with ReLU activation and dropout (rate 0.3). Training used stratified 70/15/15 train/validation/test split with stratified 5-fold cross-validation and grid search hyperparameter optimization maximizing validation ROC-AUC. Evaluation metrics included accuracy, precision, recall, specificity, F1-score, ROC-AUC with 95% CI, Hosmer-Lemeshow calibration, and subgroup fairness analysis.

V. RESULTS

A. Model Performance Comparison

TABLE III COMPREHENSIVE MODEL PERFORMANCE ON TEST SET (N~280)

Model	Acc (%)	Prec (%)	Recall (%)	Spec (%)	AUC
Logistic Reg.	**82.1**	**79.4**	**80.6**	**83.2**	**0.871**
Random Forest	**92.0**	**90.5**	**91.8**	**92.4**	**0.970**
XGBoost	**91.3**	**89.7**	**90.2**	**91.8**	**0.963**
Neural Network	**89.4**	**87.6**	**88.1**	**90.3**	**0.944**

Random Forest achieves best performance. High recall (91.8%) minimizes false negatives.

Random Forest demonstrated superior performance with 92% accuracy and 0.97 ROC-AUC (95% CI: 0.964–0.976), outperforming baseline A2DS2 (0.78 AUC) by 19 percentage points. High recall of 91.8% is clinically critical — minimizing false negatives where high-risk SAP patients would be missed prior to the 48-hour prevention window.

B. Feature Importance Analysis (SHAP)

TABLE IV TOP 10 FEATURE IMPORTANCE RANKINGS (SHAP VALUES)

Rank	Clinical Feature	SHAP Score
1	**Dysphagia Severity Score**	**0.214**
2	**Glasgow Coma Scale (GCS)**	**0.178**
3	**Age (years)**	**0.161**
4	**NIHSS Stroke Severity**	**0.143**
5	**Mechanical Ventilation (hrs)**	**0.128**
6–10	**WBC Count, CRP, SpO_2, Atrial Fibrillation, Prior Lung Disease**	**0.089–0.054**

Top 10 features account for 91.4% cumulative importance, confirming clinical validity

SHAP analysis confirmed clinically grounded top predictors: dysphagia severity (0.214), the most established modifiable SAP risk factor; GCS score (0.178) reflecting consciousness and airway protection capacity; age (0.161); NIHSS stroke severity (0.143); and mechanical ventilation duration (0.128). The top 10 features account for 91.4% cumulative importance, validating pathophysiological coherence and enhancing clinician trust.

C. Risk Stratification and Calibration

The risk stratification framework categorized patients into three clinically actionable tiers: Low risk (predicted

probability <0.25, comprising 58% of the population, observed SAP incidence 2.8%); Medium risk (0.25–0.65, 30%, observed incidence 38.4%); High risk (≥0.65, 12%, observed incidence 84.7%). Excellent calibration was demonstrated by the Hosmer-Lemeshow test (χ^2=6.84, df=8, p=0.55). Subgroup fairness analysis confirmed consistent performance across age groups, gender, and stroke type subgroups (all ROC-AUC >0.95). Mean prediction latency of 83ms per patient enables real-time ICU integration.

VI. DISCUSSION

This AI-driven lung condition prediction system demonstrates substantial improvements over traditional approaches. Random Forest's 0.97 ROC-AUC exceeds A2DS2 (0.78) and ISAN (0.80) scores by 17–19 percentage points — a clinically meaningful advancement enabling accurate identification of the 10–22% of stroke patients who develop SAP within the critical 48-hour prevention window.

SHAP interpretability reveals that the model captures established SAP pathophysiology: dysphagia impairs protective airway reflexes (leading to microaspiration), reduced consciousness diminishes cough efficacy, while stroke severity governs immune dysregulation and neurogenic pulmonary edema risk. This alignment between statistical feature importance and clinical knowledge is essential for regulatory approval and bedside clinical adoption.

Excellent calibration (Hosmer-Lemeshow p=0.55) ensures predicted probabilities accurately reflect observed SAP incidence — enabling risk-stratified clinical protocols: low-risk patients receive standard monitoring, medium-risk patients receive enhanced oral hygiene and dysphagia therapy, and high-risk patients receive prophylactic interventions and early respiratory physiotherapy. Computational efficiency (83ms) supports seamless EMR integration.

VII. CONCLUSION

This comprehensive machine learning-based clinical decision support system achieves 92% accuracy and 0.97 ROC-AUC for lung condition prediction in stroke patients, substantially exceeding traditional clinical risk scores by 17–19 percentage points. The combination of superior discrimination, high recall minimizing false negatives, SHAP interpretability confirming pathophysiological coherence, excellent calibration, subgroup fairness, and real-time performance positions this system for clinical deployment in stroke units and ICUs. Future work will pursue external validation across diverse hospital systems, integration of longitudinal time-series EHR data via recurrent neural networks, and a prospective randomized clinical trial evaluating the impact on SAP incidence and patient outcomes.

. REFERENCES

[1] M. Hilker, C. Poetter, N. Findeisen, et al., "Nosocomial pneumonia after acute stroke: Implications for neurological intensive care medicine," Stroke, vol. 34, no. 4, pp. 975–981, Apr. 2003, doi: 10.1161/01.STR.0000063373.70993.CD.

[2] J. Ji, A. Zheng, and H. He, "Comparison of stroke-associated pneumonia risk scoring systems," J. Neurol., vol. 268, pp. 1748–1756, 2021, doi: 10.1007/s00415-020-10296-y.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, San Francisco, CA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4768–4777.

[5] A. Rajkomar, E. Oren, K. Chen, et al., "Scalable and accurate deep learning with electronic health records," npj Digital Medicine, vol. 1, no. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[8] Z. Obermeyer and E. J. Emanuel, "Predicting the future—Big data, machine learning, and clinical medicine," N. Engl. J. Med., vol. 375, no. 13, pp. 1216–1219, 2016.

[9] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Medicine, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[10] W. N. Kernan, B. Ovbiagele, H. R. Black, et al., "Guidelines for the prevention of stroke in patients with stroke and transient ischemic attack,"