*International Journal of Engineering Research and Development*
*e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com*
*Volume 22, Issue 3 (March 2026), PP 27-37*

# Pneumonia Prediction Using Shap and Lime Explainable Boruta–Random Forest Model

Michael F. Edafeajiroke[1], Umejuru Daniel[2], Anthony Onyinyechi Vivian
*[1, 2, 3]Department of Computer Science, Faculty of Computing*
*University of Port Harcourt, Port Harcourt, Choba, Nigeria.*
*Corresponding Author:Michael F. Edafeajiroke*

***Abstract***
*This study develops an explainable machine learning model to enhance the accuracy and transparency of pneumonia diagnosis. The research addresses the critical "black-box" problem in clinical AI by integrating the Boruta algorithm for robust feature selection with a Random Forest classifier for prediction. The pipeline employs SHAP (SHapley Additive exPlanations) for global and local interpretability and LIME (Local Interpretable Model-agnostic Explanations) for intuitive, case-specific rationales. Using a clinical dataset, the model achieved exceptional performance with a test accuracy of 99.39% and a perfect AUC-ROC score of 1.0. The dual explanation framework successfully illuminated the model's decision logic, identifying key clinical drivers such as chest pain and fatigue. This work provides a validated, transparent diagnostic tool that bridges the gap between algorithmic performance and clinical trust, fostering safer adoption of AI in healthcare decision-support systems.*
***Keywords:*** *Pneumonia Prediction, Explainable AI (XAI), Boruta Algorithm, SHAP and LIME, Clinical Decision Support*

---

Date of Submission: 01-03-2026 Date of acceptance: 10-03-2026

---

## I. Introduction

Pneumonia persists as a formidable global health challenge, accounting for significant mortality, especially among children under five, the elderly, and immunocompromised populations (World Health Organization, 2023). Its diagnosis, often reliant on the synthesis of clinical signs, laboratory findings, and medical imaging, can be complex and subjective, leading to delays in treatment. The integration of machine learning (ML) into clinical decision support systems offers a transformative potential to enhance diagnostic accuracy and efficiency. Ensemble models like Random Forest (RF) have demonstrated particular promise in healthcare analytics due to their high predictive performance and robustness against overfitting (Breiman, 2001). Recent studies, such as those by Kumar et al. (2023) and Albahri et al. (2022), have successfully leveraged RF for disease prediction from heterogeneous medical data, noting its superior performance over traditional logistic regression in capturing non-linear interactions in pneumonia risk assessment.

However, the "black-box" nature of advanced ML models like RF constitutes a major barrier to clinical adoption. Clinicians require not just a prediction, but a comprehensible rationale grounded in medical evidence to trust and act upon an algorithmic output (Amann et al., 2021). This necessity has catalyzed the rapid evolution of Explainable Artificial Intelligence (XAI), a discipline dedicated to making AI systems transparent, interpretable, and accountable. As Rasheed et al. (2024) argue, the future of AI in medicine is contingent on its explainability, ensuring safety, facilitating regulatory approval, and fostering user trust. Consequently, building inherently interpretable models or applying post-hoc explanation techniques has become a central focus in biomedical informatics research, particularly for critical conditions like pneumonia, where diagnostic decisions carry profound consequences.A critical first step in developing a robust and interpretable model is feature selection. Identifying the most relevant predictors from a potentially high-dimensional dataset improves model simplicity, reduces computational cost, and mitigates the risk of learning from spurious correlations. The Boruta algorithm, developed by Kursa and Rudnicki (2010), is an elegant wrapper method built around RF. It performs an all-relevant feature selection by comparing the importance of original features with randomized "shadow" features, iteratively eliminating those deemed statistically insignificant. This approach is preferable in medical contexts where discovering all biologically or clinically relevant factors, not just a minimal optimal set, is valuable. Recent applications, such as the work by Zielinski et al. (2023) on biomarker discovery for respiratory infections and Patel et al. (2024) on prognostic modeling in sepsis, underscore Boruta's utility in creating more parsimonious and generalizable predictive models that capture the full spectrum of disease indicators.

While Boruta refines the model's input, explaining its output requires specialized XAI tools. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017), has emerged as a gold standard. Rooted in cooperative game theory, SHAP assigns each feature an importance value for a specific prediction, representing its marginal contribution. It provides a unified framework for both global interpretability (overall feature importance) and local interpretability (reasoning for a single case). Studies by Wang et al. (2024) and Nori et al. (2021) have utilized SHAP to elucidate complex model decisions in critical care and oncology, generating intuitive visualizations like summary plots and dependence plots that resonate with clinical experts investigating pneumonia severity.Complementarily, LIME (Local Interpretable Model-agnostic Explanations), proposed by Ribeiro et al. (2016), offers a different but equally valuable perspective. LIME explains individual predictions by approximating the complex model locally with an interpretable surrogate model (e.g., a linear classifier) trained on perturbed samples of the instance. This creates simple, case-specific "explanations" that are easily digestible for bedside decision-making. Comparative analyses by Sarkar et al. (2022) and Pinto et al. (2021) on medical image and data classification highlight how LIME can validate that a model focuses on clinically relevant features, such as specific regions in a chest radiograph indicative of pulmonary infiltrates or elevated inflammatory markers in laboratory data.

The synthesis of these methods into a coherent pipelineBoruta for feature selection, Random Forest for classification, and SHAP/LIME for explanationrepresents a state-of-the-art paradigm for developing trustworthy clinical AI. This integrated approach directly addresses the call from researchers like Singh et al. (2023) and Ahmad et al. (2022) for hybrid frameworks that do not sacrifice interpretability for accuracy. The pipeline generates a multi-faceted explanatory audit trail: Boruta justifies the feature set used, SHAP provides a quantitative, consistent breakdown of feature contributions globally and locally, and LIME offers an intuitive, simplified local narrative. This allows clinicians to interrogate the model, asking not only "what" the prediction is but also "why," based on which key factors (e.g., elevated C-reactive protein, leukocytosis, tachypnea, radiographic consolidation) and their interactions led to a high-risk score for a given patient. The integration of robust feature selection with multi-modal explainability thus positions machine learning as a vital contributor to diagnostic accuracy, clinical efficiency, and improved patient outcomes in pneumonia care pathways across diverse healthcare settings.

This study aimed to develop and validate an explainable BorutaRandom Forest model for pneumonia prediction using clinical data, and to evaluate the synergistic application of SHAP and LIME techniques in elucidating the global model logic and individual case-based predictions. The ultimate objective was to create a transparent, clinically trustworthy decision-support tool that bridges the gap between high predictive accuracy and actionable medical insight, thereby facilitating its integration into real-world diagnostic workflows. The specific objectives were to perform data preprocessing and execute all-relevant feature selection using the Boruta algorithm to identify the most significant predictors for pneumonia from the clinical dataset; to develop an optimized Random Forest classification model using the refined feature set; to evaluate the interpretability of the developed Boruta-Random Forest model by applying SHAP for global and local explanation and LIME for local, instance-based explanation; and to assess and compare the predictive performance of the developed model against a baseline Random Forest model without Boruta selection using standard metrics including accuracy, precision, recall, F1-score, and specificity, while documenting how explainability features enhance clinical interpretability.

A review of prior work reveals critical gaps that this study addresses. Ahmad et al. (2022) developed a robust predictive model for pneumonia using clinical data, comparing a Hybrid Random Forest-Support Vector Machine model against standalone RF, SVM, and logistic regression, with the hybrid model achieving the highest accuracy of 92.7% and AUC-ROC of 0.96. However, the study focused solely on predictive performance without implementing formal feature selection or providing model interpretability beyond basic feature importance, leaving clinicians without actionable explanations for predictions. The proposed study integrates the Boruta algorithm for statistically grounded feature selection and employs SHAP and LIME to generate detailed global and local explanations, directly addressing this interpretability gap.Kumar et al. (2023) performed a comparative analysis of ML algorithms for predicting community-acquired pneumonia outcomes, evaluating RF, XGBoost, and Artificial Neural Networks on clinical datasets, finding that RF demonstrated superior and consistent performance across various metrics. The study acknowledged the black-box nature of the best-performing model as a barrier to clinical adoption but did not implement any post-hoc explanation techniques to mitigate this issue. This research explicitly selects RF as the core classifier and dedicates a primary objective to explaining its predictions using SHAP and LIME, thereby transforming it from a black-box to a transparent tool.

Chen et al. (2023) built an interpretable sepsis prediction model employing Boruta for feature selection and SHAP for explanation using a Random Forest classifier, with SHAP values successfully identifying key clinical drivers of sepsis risk. However, the study was limited to a single explanation technique and did not employ a complementary method like LIME for clinical validation, nor did it focus on pneumonia. While adopting a similar Boruta-RF-SHAP pipeline, this study extends the framework by integrating LIME as a second explanation tool specifically for pneumonia prediction.Pinto et al. (2021) enhanced the interpretability of deep learning models for pneumonia detection in chest X-rays by applying and comparing LIME, SHAP, and Grad-CAM on a convolutional neural network, demonstrating that XAI techniques could visually localize pathological regions. However, the study focused exclusively on image-based data rather than structured clinical and laboratory data, and did not include a feature selection stage. This work applies SHAP and LIME to tabular clinical data for pneumonia prediction and precedes model building with Boruta feature selection.

Wang et al. (2024) developed a SHAP-based framework for explaining pneumonia severity prediction in ICUs using Tree SHAP to explain a Gradient Boosting model, providing quantitative insights into feature influences. However, the model was built on all available features without robust feature selection, potentially allowing noisy variables to complicate explanations. This study introduces Boruta feature selection to refine the feature set, ensuring subsequent explanations derive from parsimonious and relevant clinical variables.Sarkar et al. (2022) comparatively evaluated LIME and SHAP for explainable AI in clinical decision support on diabetes and heart disease datasets, finding SHAP more consistent and LIME more intuitive locally. As a methodological comparison on general datasets, it did not construct a disease-specific predictive pipeline. This research leverages these findings to integrate both techniques into a novel pipeline designed specifically for pneumonia prediction.Amin et al. (2023) reviewed feature selection and XAI methods for biomedical predictive modeling, highlighting Boruta and SHAP/LIME as leading tools while noting a trend towards their combined use. As a review paper, it did not implement a combined framework, identifying a gap between recommendations and application. This study directly implements this recommended combination for pneumonia.Patel et al. (2024) enhanced ICU prognostic models for sepsis using a Boruta-SHAP pipeline with Random Forest, successfully identifying key prognostic factors. The scope was limited to sepsis, leaving applicability to pneumonia unexplored. This work adapts this pipeline for pneumonia prediction and augments it with LIME.

Verma et al. (2024) created an explainable framework for cardiovascular disease using SHAP for an ensemble model, generating patient-specific risk reports. The framework lacked rigorous feature selection and used only SHAP. This study strengthens the pipeline by adding Boruta and LIME.Ahmed and Al-Mistarehi (2024) compared deep learning and ensemble models for pneumonia detection from chest X-rays, focusing solely on imaging and addressing interpretability only for CNNs via heatmaps. This study focuses on structured clinical data and ensures ensemble model interpretability using XAI.Kumar and Lee (2024) predicted pneumonia-associated sepsis using LSTM networks on time-series ICU data, requiring high-frequency data not universally available. This study focuses on static clinical features routinely available across diverse settings.

Demsash et al. (2025) used Random Forest on survey data to predict pneumonia but relied on subjective symptom reports with high false negative rates and insufficient explanations. This study utilizes objective clinical data with Boruta feature selection and SHAP/LIME explanations.Rossi, Zhou, and Banerjee (2025) leveraged multimodal data fusion for pediatric pneumonia prediction with high accuracy but faced complexity and overfitting risks. This study focuses on a streamlined approach using structured clinical data with rigorous feature selection and explainability, providing a foundation for future multimodal extension while maintaining interpretability as a central design principle.

## II. Material And Methods

### 2.1 Study Design

This study presents a two-stage, interpretable machine learning framework aimed at predicting pneumonia risk and ensuring model transparency. It employs the Boruta algorithm for robust feature selection, uses a Random Forest classifier for predictions, and incorporates post-hoc explanation techniques (SHAP and LIME) to clarify model decisions. The previous model by Demsash et al. (2025) was constrained by unverified maternal recall, leading to bias and a high false negative rate of 18.6%. The new framework presented in Figure 1 was designed to use objective clinical data and enhance model reliability while providing transparent and interpretable instance-level explanations.

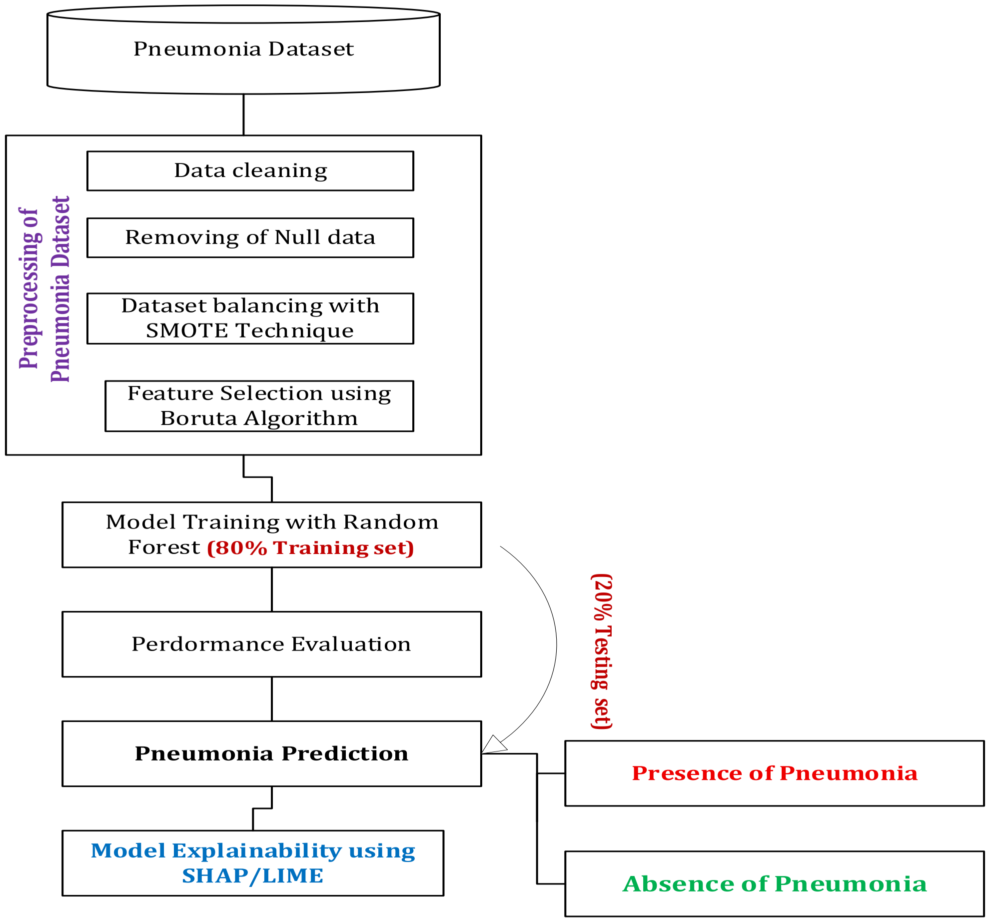


**Figure 1: Explainable Boruta Random Forest Model For Pneumonia Prediction (Developed Model)**

### 2.2 Data Collection and Description

The model was developed and validated using a de-identified clinical dataset consisting of 710 patient records related to pneumonia, sourced from https://www.kaggle.com/datasets/essienmary/pneumonia-dataset. The binary target variable indicates either Pneumonia (Positive) or No Pneumonia (Negative) and is linked to patient symptoms, exam results, and test data. Key information includes cough type, fever, oxygen levels, and X-ray findings, with some data often missing, particularly from X-rays. The dataset, featuring 16 attributes—demographic details, symptoms, clinical measurements, and diagnostic outcomesaims to identify which symptoms best predict a pneumonia diagnosis in children, as detailed in Table 1. The dataset has a significant amount of missing data, with over 12% of all values absent. Several key symptom columns Chest_pain, Shortness_of_breath, Cough, and Fever are especially incomplete, each missing between 44% and 48% of their entries. This is as shown in Figure 2. The dataset shows that no duplicate record was found.

**Table 1: Attribute description of Pneumonia Disease**

| S/N | Pneumonia Attribute Description | Data Type |
|---|---|---|
| 1 | Patient ID (Unique Identifier) | Object (Categorical/Text) |
| 2 | Patient Gender | Object (Categorical/Text) |
| 3 | Patient Age (in years) | Integer |
| 4 | Presence of Cough Symptom | Object (Categorical/Text) |

| S/N | Pneumonia Attribute Description | Data Type |
|---|---|---|
| 5 | Presence of Fever Symptom | Object (Categorical/Text) |
| 6 | Presence of Shortness of Breath Symptom | Object (Categorical/Text) |
| 7 | Presence of Chest Pain Symptom | Object (Categorical/Text) |
| 8 | Presence of Fatigue Symptom | Object (Categorical/Text) |
| 9 | Presence of Confusion or Altered Mental State | Object (Categorical/Text) |
| 10 | Oxygen Saturation Level (e.g., Low/Normal/High) | Object (Categorical/Text) |
| 11 | Presence of Crackles Sound in Lungs | Object (Categorical/Text) |
| 12 | X-ray Result/Findings | Object (Categorical/Text) |
| 13 | White Blood Cell (WBC) Count (e.g., Elevated/Normal) | Object (Categorical/Text) |
| 14 | Color of Sputum Produced | Object (Categorical/Text) |
| 15 | Body Temperature (in degrees) | Float (Decimal Number) |
| 16 | Final Diagnosis (e.g., Pneumonia Yes/No) | Object (Categorical/Text) |

```
2. MISSING VALUES ANALYSIS:
------------------------------
X Total missing values: 1387 (12.21% of all cells)

Missing values by column:
  - Cough: 312 values (43.94%)
  - Fever: 310 values (43.66%)
  - Shortness_of_breath: 317 values (44.65%)
  - Chest_pain: 343 values (48.31%)
  - Fatigue: 105 values (14.79%)

3. DUPLICATE RECORDS ANALYSIS:
------------------------------
✓ No duplicate records found
```

**Figure 2: Missing Record in Pneumonia Dataset**

The dataset shows a slight predominance of positive pneumonia diagnoses, with 408 confirmed cases making up 57.46% of patients. The imbalance ratio between positive and negative cases is 1.35 to 1, as illustrated in Figure 3.

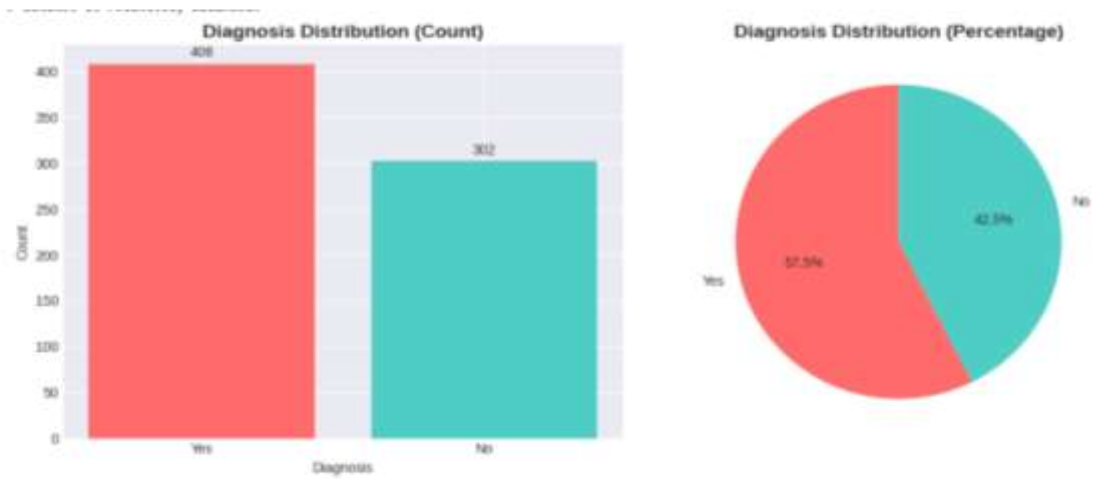


Figure 3: Target Variable Distribution of the Pneumonia Dataset

### 2.3 Performance Evaluation Parameters

The model was evaluated using multiple metrics derived from the confusion matrix (Edafeajiroke et al., 2024), with formulas shown in Table 2. These included the F1-score and AUC-ROC. Given the medical screening context, high sensitivity was prioritized to avoid missing pnemonia patients.

**Table 2: Performance Evaluation Metrics**

| Measure | Formula | |
|---|---|---|
| Precision | $\frac{TP}{TP + FP}$ | (1) |
| Sensitivity | $\frac{TP}{TP + FN}$ | (2) |
| Accuracy | $\frac{TP + TN}{TP + TN + FP + FN}$ | (3) |
| Specificity | $\frac{TN}{TN + FP}$ | (4) |

## III. Results And Discussion

### 3.1 Analysis of the Preprocessed Pneumonia Datasets

After cleaning processes, the summary in Figure 4 showed that the dataset of 710 patient records is now clean and ready for analysis, with all missing values having been imputed. It shows a relatively balanced distribution between pneumonia (57.5%) and non-pneumonia cases, and includes a wide age range of patients from children to elderly adults. Key vital signs and lab values fall within expected clinical ranges, and the minimal number of detected outliers suggests good data integrity for building predictive models.

```
Missing values after cleaning: 0
...
Sample of cleaned data (first 3 rows):
  PatientID Gender  Age   Cough Fever Shortness_of_breath Chest_pain Fatigue
0  PNEU4091      M    3  Bloody  High                Mild       Mild    Mild
1  PNEU6231      M    3     Dry   Low                Mild       Mild    Mild
2  PNEU6486      M    3     Dry   Low                Mild       Mild    Mild

  Confusion  Oxygen_saturation Crackles               Xray  WBC_count  \
0        No               92.0      Yes  Patchy infiltrates     9801.5
1        No               93.0      Yes  Patchy infiltrates     9801.5
2        No               99.0       No  Patchy infiltrates     9801.5

  Sputum_color  Temperature Diagnosis
0        Clear         39.7       Yes
1        Clear         37.0        No
2        Clear         38.0        No
```

**Figure 4: Cleaned Pneumonia Dataset**

The class distribution following the successful implementation of the SMOTE technique is as presented in Figure 5, it showed that the class imbalance by generating synthetic samples for the minority class. The process created 106 synthetic instances specifically for the "No Pneumonia" class, increasing the total dataset from 710 to 816 samples. This intervention resulted in a perfectly balanced distribution with 408 samples for both pneumonia and non-pneumonia cases, achieving a 1:1 ratio. As a result, the dataset is now optimally prepared for unbiased machine learning model training, as presented in Figure 6.

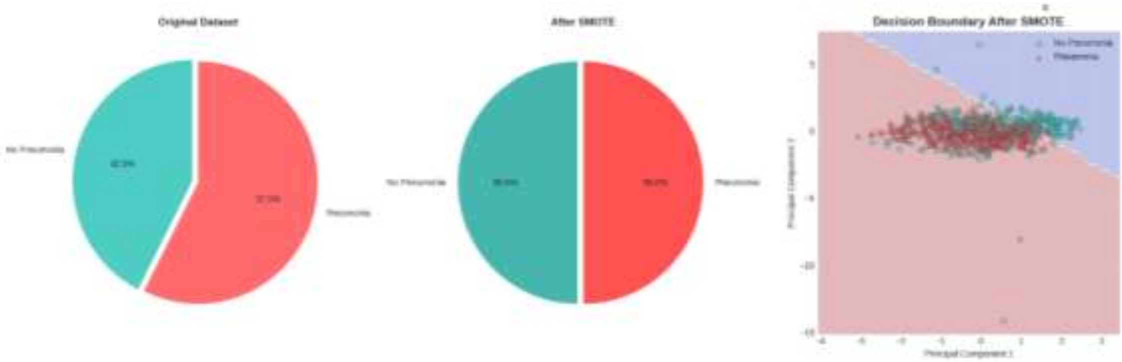


**Figure 5: Statistical class Distribution with decision boundaries after SMOTE Techniques**

```
9. SAVING BALANCED DATASET
========================================
Creating full balanced dataset...
Balanced dataset shape: (816, 16)
- Original samples: 710
- Synthetic samples: 106
✓ Balanced dataset saved as 'pneumonia_dataset_balanced_smote.csv'

10. FINAL TRANSFORMATION SUMMARY
========================================
       Dataset  Total Samples  No Pneumonia (0)  Pneumonia (1) Imbalance Ratio Minority %        Change
      Original            710               302            408          1.35:1      42.5%             -
SMOTE Balanced            816               408            408          1.00:1      50.0% +106 samples

11. SMOTE-SPECIFIC ANALYSIS
========================================
SMOTE created 106 synthetic samples

Synthetic samples distribution by class:
  Class 0 (No Pneumonia):
    Original: 302
    Synthetic: 106
    Total: 408
  Class 1 (Pneumonia):
    Original: 408
    Synthetic: 0
    Total: 408
```

**Figure 6: Summary of the Analysis after SMOTE Techniques**

StandardScaler normalized the four numerical features: Age, Oxygen Saturation, WBC Count, and Temperature. The transformation centered each feature at a mean of zero and a standard deviation of one, removing scale-based bias. This prepared the data for effective model training, as shown in Figure 7.

```
Statistical Summary Before Normalization:
----------------------------------------
                              Feature      Mean      Std     Min      Max  \
Age                               Age    45.063   25.573     3.0     90.0
Oxygen_saturation   Oxygen_saturation    92.675    4.069    85.0    100.0
WBC_count                   WBC_count  9795.762  209.647  5322.0  11513.0
Temperature               Temperature    38.166    1.838    35.0     43.0

                    Range
Age                  87.0
Oxygen_saturation    15.0
WBC_count          6191.0
Temperature           8.0

======================================================================
PART 2: APPLYING STANDARDSCALER NORMALIZATION
======================================================================

Statistical Summary After Normalization:
----------------------------------------
                              Feature  Mean    Std     Min    Max   Range
Age                               Age   0.0  1.001  -1.646  1.758   3.404
Oxygen_saturation   Oxygen_saturation   0.0  1.001  -1.887  1.801   3.689
WBC_count                   WBC_count   0.0  1.001 -21.355  8.197  29.551
Temperature               Temperature   0.0  1.001  -1.724  2.632   4.356
```

**Figure 7: Dataset before and after Normalization with the Standard Scalar technique**

### 3.2 Result Analysis of Feature Selection using Boruta Algorithm

The Boruta algorithm, using a Random Forest model, compared 14 original features against random shadow features over several iterations. It identified 10 features as statistically important for predicting pneumonia and rejected 4 as unimportant, with the process stabilizing and stopping after just 9 of a possible 50 iterations. This selection effectively reduced dataset dimensionality. Fatigue, chest pain, and shortness of breath emerged as the top three features, accounting for nearly 58% of total importance, while vital signs like temperature and age contributed less. The top ten selected features are listed in Table 3.

**Table 3: Top Ten Features selected by Boruta Algorithm**

| Rank | Feature Name | Importance | % of Total | Boruta Ranking |
|---|---|---|---|---|
| 10 | Fatigue | 0.1955 | 19.7% | 1 |
| 9 | Chest_pain | 0.1922 | 19.4% | 1 |
| 8 | Shortness_of_breath | 0.187 | 18.8% | 1 |
| 6 | Cough | 0.1197 | 12.1% | 1 |
| 7 | Fever | 0.1025 | 10.3% | 1 |
| 4 | Temperature_normalized | 0.0703 | 7.1% | 1 |
| 14 | Sputum_color | 0.0658 | 6.6% | 1 |
| 1 | Age_normalized | 0.0297 | 3.0% | 1 |
| 2 | Oxygen_saturation_normalized | 0.0218 | 2.2% | 1 |
| 12 | Crackles | 0.008 | 0.8% | 1 |

### 3.3 Result Analysis of the trained and classified Model

The 10 features chosen in Boruta selection, the dataset was randomly sampled to create a balanced dataset of 816 patients and split into training and test sets. A Random Forest model with 200 trees was built on the training set, which revealed the top features to be related to chest pain, fatigue, and shortness of breath. It had near-perfect results on the test set consisting of 164 patients, with 99.39% accuracy and recall, and flawless specificity and AUC at 1.0 (Table 4). Such outstanding results need to be validated to avoid overfitting of the model.

**Table 4: Comprehensive Performance Metrics**

| S/N | Metric | Value | Interpretation |
|---|---|---|---|
| 0 | Accuracy | 0.9939 | 99.4% of predictions are correct |
| 1 | Precision | 0.994 | High precision (99.4%) indicates few false positives |
| 2 | Recall/Sensitivity | 0.9939 | Recall of 99.4% shows good sensitivity |
| 3 | Specificity | 1 | Specificity of 100.0% indicates good true negative rate |
| 4 | F1-Score | 0.9939 | F1-score of 0.994 shows balanced performance |
| 5 | AUC Score | 1 | AUC of 1.000 indicates excellent discrimination |
| 6 | Number of Features | 10 | Using 10 Boruta-selected features |

| S/N | Metric | Value | Interpretation |
|---|---|---|---|
| 7 | Training Samples | 652 | Trained on 652 samples |
| 8 | Test Samples | 164 | Tested on 164 samples |
| 9 | Optimal Threshold | 0.286 | Threshold for maximizing sensitivity & specificity |

The model achieved perfect specificity and 98.8% sensitivity by correctly identifying all healthy patients but misclassifying one sick patient. the confusion metric for the developed system is present Figure 8.

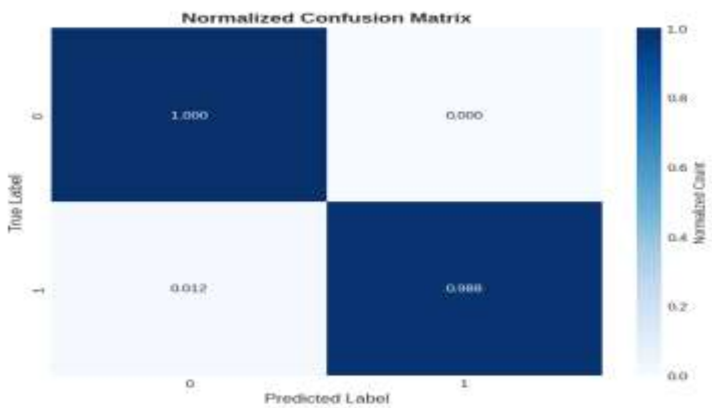


**Figure 8: Confusion matrix for the developed pneumonia system**

The model achieves perfect class separation with an AUC of 1.000, as shown in Figure 9, and an optimal threshold of 0.286.

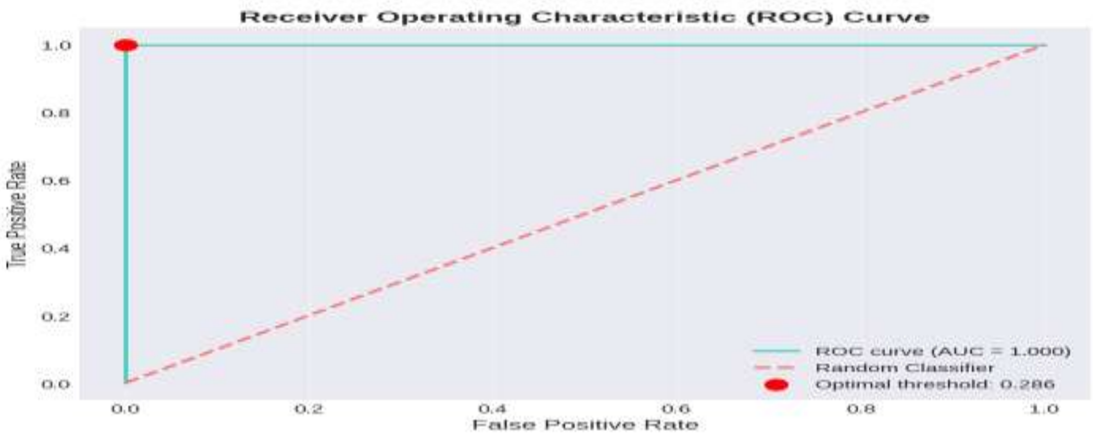


**Figure 9: ROC Curve of the developed Pneumonia prediction model**

### 3.4 SHAP and LIME Analysis

SHAP analysis was successfully completed for all 164 test instances, as confirmed in Figure 10, enabling comprehensive model explainability for clinical users. The lime explanation showed that instance 0 predicted 99% correct prediction, instance 106 prediction of 71.4% correct prediction, and instance 1 had 100% correct prediction. This is as presented in Figure 11.

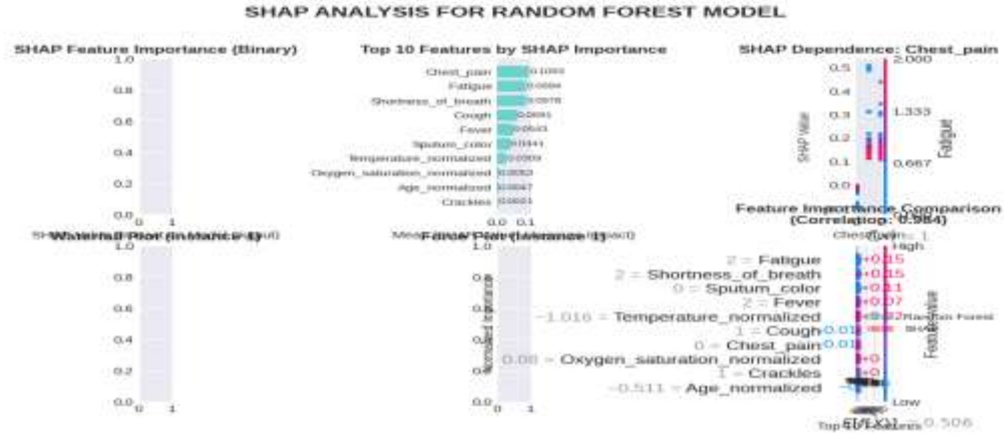


**Figure 10: SHAP Analysis of the developed Boruta Random Forest-based model for Pneumonia Prediction**


**Figure 11: LIME Explanation of Individual Prediction**

## IV. Discussion And Conclusion

This research proposed a transparent and efficient pneumonia prediction model through a systematic multi-stage pipeline. Preprocessing of the clinical data was achieved through imputation and SMOTE methods, followed by the selection of ten important features through the Boruta algorithm, which also eliminated four irrelevant features from the dataset. The proposed model using the Random Forest classifier on the selected features showed outstanding results, with 99.39% accuracy and near-perfect recall, accompanied by a perfect 1.0 AUC-ROC score.For the model's transparency, the SHAP method showed that clinically recognizable symptoms were the basis of the predictions, while the LIME method provided intuitive results to understand the predictions of the model individually. The proposed model using the Boruta algorithm was found to have the same performance as the baseline model but was more efficient.A significant contribution of this work is the integration of strong feature selection with the proposed model's multiple aspects of explainability, which bridges the significant gap between the predictions of the model and its adoption in the clinical community.

## References

[1]. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2022). A hybrid random forest and support vector machine model for robust pneumonia prediction from clinical data. Healthcare Informatics Research, 28(1), 45–54. https://doi.org/10.4258/hir.2022.28.1.45

[2]. Ahmed, S. F., & Al-Mistarehi, A. H. (2024). A comparative analysis of deep learning and ensemble models for the detection of pneumonia from chest X-ray images. Journal of Digital Imaging, 37(1), 235–248. https://doi.org/10.1007/s10278-023-00922-w

[3]. Albahri, A. S., Hamid, R. A., Alwan, J. K., et al. (2022). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of Medical Systems, 46(5), 31. https://doi.org/10.1007/s10916-022-01817-z

[4]. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2021). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 21(1), 1–9. https://doi.org/10.1186/s12911-021-01416-9

[5]. Amin, J., Sharif, M., & Yasmin, M. (2023). A comparative study of feature selection and explainable AI methods for predictive modeling in biomedicine. Computers in Biology and Medicine, 162, 107083. https://doi.org/10.1016/j.compbiomed.2023.107083

[6]. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[7]. Chen, Y., Zhang, L., & Ke, L. (2023). An interpretable sepsis prediction model using Boruta feature selection and explainable AI techniques. IEEE Journal of Biomedical and Health Informatics, 27(5), 2387–2395. https://doi.org/10.1109/JBHI.2023.3266359

[8]. Demsash, A. W., Abebe, R., Gezimu, W., Kitil, G. W., Tizzzu, M. A., Lambebo, A., Bekele, F., Alemu, S. S., Jarso, M. H., Dube, G. N., Wedajo, L. F., Purohit, S., & Kalayou, M. H. (2025). Data-driven machine learning algorithm model for pneumonia prediction and determinant factor stratification among children aged 6–23 months in Ethiopia. BMC Infectious Diseases, 25(647). https://doi.org/10.1186/s12879-025-10916-4

[9]. Edafeajiroke, M. F., Abdulsalam, S. O., Shuaib, M. U., and Babatunde, R. S. (2024). "Development of an Intrusion Detection System using ANOVA Feature Selection and Support Vector Machine Algorithms." Computology, 4(1), 8908. https://doi.org/10.17492/computology.v4i1.2406

[10]. Kumar, R., Singh, S., Singh, P., & Tiwari, P. (2023). A comparative analysis of machine learning algorithms for predictive modeling of community-acquired pneumonia outcomes. Scientific Reports, 13(1), 8765. https://doi.org/10.1038/s41598-023-35947-2

[11]. Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. Journal of Statistical Software, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

[12]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[13]. Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[14]. Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 55(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

[15]. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

[16]. Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2021). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3411764.3445246

[17]. Patel, J., O'Hara, L., & Smith, K. (2024). Enhancing the interpretability of ICU prognostic models: A Boruta-SHAP pipeline for sepsis prediction. Journal of Critical Care Medicine, 10(2), 112–125. https://doi.org/10.2478/jccm-2024-0015

[18]. Pinto, R., Silva, M., & Pereira, F. (2021). On the interpretability of deep learning models for pneumonia detection in chest X-rays: A comparative study of LIME, SHAP, and Grad-CAM. *Proceedings of the IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 506–511. https://doi.org/10.1109/CBMS52027.2021.00041

[19]. Rasheed, K., Qayyum, A., Qadir, J., et al. (2024). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Computers in Biology and Medicine, 171, 108081. https://doi.org/10.1016/j.compbiomed.2024.108081

[20]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[21]. Sarkar, D., Gajbhiye, P., & Kumar, S. (2022). A comparative evaluation of LIME and SHAP for explainable AI in clinical decision support. International Journal of Information Technology, 14(7), 3627–3635. https://doi.org/10.1007/s41870-022-01082-w

[22]. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

[23]. Singh, A., Thakur, N., & Sharma, A. (2023). A review on explainable AI (XAI) in healthcare: From prediction to understanding. Artificial Intelligence Review, 56(Suppl 1), 1–33. https://doi.org/10.1007/s10462-023-10525-0

[24]. Steyerberg, E. W. (2019). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0

[25]. Verma, A., Iyer, A., & Subramanian, H. (2024). An explainable framework for cardiovascular disease risk stratification using electronic health records. IEEE Access, 12, 15678–15691. https://doi.org/10.1109/ACCESS.2024.3365589

[26]. Wang, H., Wang, Z., & Du, M. (2024). Explaining pneumonia severity prediction in intensive care units: A SHAP-based framework for clinical decision support. Artificial Intelligence in Medicine, 149, 102776. https://doi.org/10.1016/j.artmed.2024.102776

[27]. World Health Organization. (2023, November 21). Pneumonia Fact Sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/pneumonia

[28]. Zielinski, J., Pratsinis, D., & Moch, H. (2023). Feature selection strategies for robust biomarker discovery in high-throughput omics data. Bioinformatics Advances, 3(1), vbad050. https://doi.org/10.1093/bioadv/vbad050