INTERNATIONAL JOURNAL OF RESEARCH AND REVIEWS IN SOCIAL AND APPLIED SCIENCES

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

MACHINE LEARNING MODELS FOR TRACING COMMUNICABLE LUNG DISEASES

Chisaa N. Kings –Wali[1], L.N. Onyejegbu[2], B.B. Baridam[3]
Department of Computer science, University of port Harcourt, port Harcourt, Rivers state.
EMAIL: Chisaaking@yahoo.co.uk[1]

ABSTRACT

Communicable lung diseases such as tuberculosis, pneumonia, and COVID-19 remain major global health concerns due to the high transmission and mortality rates, particularly in developing countries like Nigeria, where healthcare systems face persistent challenges with early detection, diagnosis, and effective control. Traditional disease-tracing and monitoring methods often fail to accurately predict real time infection severity, thereby hindering timely response, appropriate treatment, and the efficient allocation of medical resources. This project aims to develop single, unified machine learning model capable to distinguishes disease type (communicable vs non-communicable across multiple lung conditions),classifies severity (mild/moderate/severe) in real time, and triggers public-health actions for automatic isolation built into the real working system To achieve this, a Machine Learning models was designed, combining K-Modes clustering for grouping patients with similar symptoms from categorical datasets with a Random Forest classifier for categorizing disease severity levels into mild, moderate, or severe. This integration enables both unsupervised grouping and supervised prediction within a single analytical framework. Developed using the Object-Oriented Analysis and Design Methodology (OOADM) and implemented in Python, the system utilizes Excel as the primary data source, obtained from the University of Port Harcourt Teaching Hospital (Internal Medicine: Cardiology Department). Excel functions as a simple storage medium accessed and managed through Python libraries, particularly Pandas, which is employed for statistical computation and detailed data analysis. SQLite database engine was used to store user biodata and profile for system authentication and authorization. The system allows real-time monitoring by analyzing patient symptoms and vital parameters upon data entry, providing immediate diagnostic and severity-level feedback without the need for laboratory tests. This facilitates rapid clinical decision-making by healthcare professionals and supports proactive public health interventions. The model achieved an accuracy of 95.6%, precision of 96.1%, recall of 95.2%, and an F1-score of 92.2%, demonstrating reliability in predicting infection severity. Ultimately, the application enhances disease surveillance, enables early isolation of high-risk patients, and strengthens national healthcare response systems.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

INTRODUCTION

Communicable lung diseases such as tuberculosis, pneumonia, and COVID-19 continue to exert major pressure on global health systems, with the burden especially severe in resource-limited settings like Nigeria where rapid diagnosis is essential. Machine learning models have become increasingly valuable for analysing chest X-ray images and clinical records, enabling faster disease detection, contact-tracing support, and the classification of illness severity. Distinguishing patients into mild, moderate, and severe categories is critical for effective triage, resource allocation, and timely clinical decision-making, as severe cases require urgent intervention while mild cases may be safely managed at home.

Accurate diagnosis is further complicated by the rise of Nontuberculous Mycobacteria (NTM), which often mimic pulmonary tuberculosis in both symptoms and imaging. Studies such as Zhiheng Xing et al. (2020) highlight the need for precise, region-specific diagnostic tools to reduce misclassification. Although RT-PCR and next-generation sequencing remain standard confirmation methods, they face drawbacks—including false negatives and limited availability—especially in low-infrastructure environments. CT imaging offers higher diagnostic clarity (Chung et al., 2020; Kwee et al., 2020), but remains costly and inaccessible for many facilities.

Recent research has demonstrated the effectiveness of machine learning in automating lung-disease detection. For example, Rajaraman et al. (2019) used deep learning to distinguish bacterial from viral pneumonia in children, while other models have expanded diagnostic capabilities for diseases such as COVID-19 and NTM. Broader frameworks by Wong et al. (2019) and Ganasegeran & Abdulrahman (2020) emphasize the potential of AI in infectious-disease surveillance and population-level behavioural tracking, but these works do not offer real-time, patient-level severity prediction.

To address this gap, the present study integrates K-Modes clustering with Random Forest classification to enable real-time disease detection, severity prediction, and automated public-health response. By leveraging large clinical datasets and immediate data-processing pipelines, this system aims to strengthen diagnostic accuracy, improve clinical decision-making, and enhance disease-surveillance capacity in environments where rapid response is crucial.

Literature Review / Related Work

Machine learning has become central to supporting the diagnosis, classification, and severity assessment of lung diseases, especially in settings where rapid evaluation is essential. Extensive research has focused on applying deep learning and ensemble models to medical imaging and clinical datasets. Hybrid deep-learning frameworks such as those proposed by Bharati et al. (2020) and Rajinikanth (2020) demonstrated strong performance in detecting lung abnormalities from chest X-ray and CT images. Similarly, Kumar and Ravi (2023) evaluated multiple deep-learning models for respiratory disease classification, highlighting continued advancement in model accuracy for pneumonia, tuberculosis, and COVID-19 detection. Additional improvements have come from AI-driven optimization techniques, such as knowledge distillation approaches by Roy and Bhattacharyya (2024) and segmentation-assisted pipelines like Nair et al. (2024), which have

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

further enhanced detection reliability across diverse imaging conditions. In tuberculosis detection, Showkatian et al. (2022) showed that deep learning techniques significantly outperform conventional radiographic interpretation, establishing the importance of automated tools in early infectious-disease detection.

Beyond imaging, Random Forest has gained prominence as a robust classifier for disease prediction due to its ability to model heterogeneous, high-variance clinical datasets. Early work by Khalilia et al. (2011) validated Random Forest's ability to handle imbalanced medical data, while Islam et al. (2018) applied Random Forest and decision-tree methods to clinical pneumonia datasets, reporting high predictive accuracy. More recent studies expanded its application to complex respiratory-disease contexts. For instance, Yang et al. (2025) demonstrated the effectiveness of Random Forest in predicting respiratory diseases using electronic health data, whereas Hong et al. (2022) and Yar Muhammad et al. (2021) used the algorithm to classify disease severity and risk in acute and chronic conditions. In population-level analytics, Kristensen et al. (2023) showed that Random Forest improved spirometry-based respiratory assessment, and Narasimhan and Victor (2025) integrated metaheuristic optimization to enhance classification precision in chronic disease prediction. Even outside medical imaging, Random Forest's adaptability has been documented; Fife (2021) and Rahimian (2019) illustrated its utility in diverse predictive-modelling tasks ranging from psychological assessment to emergency-admission forecasting, while Roysden and Wright (2020) applied Random Forest with NLP to predict healthcare utilization.

The broader AI landscape further contextualizes these model-specific contributions. Wong et al. (2019) proposed an AI-based big-data analytics framework for infectious-disease surveillance, highlighting the importance of automated systems for outbreak detection and public-health decision support. Ganasegeran and Abdulrahman (2020) emphasized the role of AI in tracking population behaviour during epidemics, demonstrating how intelligent systems can support government preparedness and public-health communication. Meanwhile, imaging-specific innovations—such as Kotter et al. (2020), who developed AI-assisted tools that improved radiologist detection of malignant lung nodules—show the growing acceptance of machine-learning augmentation in clinical workflows. Zhang (2020) also contributed by demonstrating high-accuracy pneumonia prediction using convolutional deep-neural networks, reinforcing the importance of automated tools where radiological expertise is limited.

Despite these advancements, gaps remain in the integration of unsupervised clustering with supervised classification for real-time patient-level severity prediction. While existing studies excel in imaging interpretation, risk prediction, or big-data surveillance, none combine K-Modes clustering for disease-type grouping with Random Forest for severity classification in a unified, real-time framework. Likewise, prior work has not fully addressed automated alerting and decision-support functions that link patient diagnosis to actionable public-health responses. This gap forms the motivation for the present study, which uses K-Modes to cluster diseases into communicable and non-communicable categories and Random Forest to predict severity levels through a real-time machine-learning pipeline designed for rapid diagnosis, monitoring, and epidemic-response support.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

MATERIALS AND METHODS

Methodology

This study adopts an Object-Oriented Analysis and Design (OOAD) methodology to develop a machine learning system capable of tracing communicable lung diseases and predicting severity levels. OOAD provides a structured technological approach for analyzing and designing software systems by identifying key objects—such as patients, health centers, and locations and defining their attributes (e.g., age, symptoms, identity number) and behaviours (e.g., updating patient status). By modelling the system around interacting objects and their relationships, OOAD supports clarity, maintainability, and reusability while reducing development time and cost.

Given the complexity of real-time disease tracing, the methodology integrates iterative and modular development. The system is decomposed into functional modules such as data preprocessing, feature extraction, clustering, and classification. These modules are designed as reusable objects and subsystems, allowing the system to evolve smoothly and accommodate new data sources, algorithms, or workflows. The iterative nature of OOAD ensures continuous refinement, enabling the system to adapt to changing project requirements and public-health needs.

Object-oriented design enhances the performance and scalability of the proposed machine learning framework. By organizing components into classes, subclasses, and hierarchical relationships, the model simplifies system updates, supports parallel processing across multi-core environments, and allows multiple developers to work efficiently. This methodological approach provides robustness for handling large clinical datasets, ensures encapsulation of complex disease-tracing logic, and offers an extensible foundation for integrating future improvements in communicable-disease surveillance and automated public-health response.

Input Data and Feature Preparation

Patient-level data of 20501 in table 1 were obtained from hospital information systems, including demographic attributes, clinical symptoms, vital signs, and comorbidities, as summarized in Table 3.1. These records represent individual hospital visits and contain routinely collected clinical information such as age, sex, residence, cough, fever, chest pain, temperature, heart rate, respiratory rate, and oxygen saturation.

All raw data were consolidated into a shared repository and subjected to cleaning, normalization, and structured organization to ensure consistency prior to supervised learning. Patients identified as communicable through the earlier K-Modes clustering stage shown in table 3.2 were retained and assigned a ClusterID, which served as an additional feature for downstream analysis.

This refined dataset forms the input for the Random Forest classification module. By leveraging the unsupervised clustering output, the supervised model begins training with patient groups that already reflect underlying disease patterns, improving its ability to predict disease type, assess severity levels, and support early triage decisions.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Table 3.1, Source Data

record_id	symptoms	heart_rate	blood_pressure_s	blood_pressure_d	fever	body_temperatu	duration_of_infe	oxygen_saturatio	encounter_datetime	age_years	sex	BMI	residence_type	household_fuel	smoking_statu
REC2-0000	Cough: Sudden, with phl	96	101	61	TRUE	38.1	5	95	18/02/2024 23:51	51	male	19.9	rural	biomass	forme
REC2-000X	Cough: Sudden, with phl	95	101	61	TRUE	38.1	5	95	18/02/2024 23:51	51	male	19.9	rural	biomass	forme
REC2-0000	Cough: Persistent (2+ we	87	130	85	TRUE	38.1	36	92.7	03/01/2023 08:34	41	male	18.8	urban	mixed	forme
REC2-0000	Cough: Intermittent, ofte	78	139	83	FALSE	36.9	4	95.9	14/02/2025 07:56	18	male	23.4	rural	biomass	never
REC2-0000	Cough: Sudden onset; Sp	92	80	57	TRUE	37.5	6	94.4	07/11/2024 01:38	65	male	27.1	urban	gas/electri	forme
REC2-0000	Cough: Sudden onset; Sp	129	84	64	TRUE	37.5	2	89.4	23/03/2024 16:18	57	female	31.6	urban	biomass	currer
REC2-0000	Cough: Intermittent, ofte	123	123	73	FALSE	37.2	12	97.3	03/07/2023 21:00	49	male	29.2	peri-urba	gas/electri	never
REC2-0000	Cough: Chronic, product	97	127	102	FALSE	37.3	6	92.4	09/04/2023 05:32	49	female	29.3	peri-urba	biomass	currer
REC2-0000	Cough: Sudden, with phl	101	107	73	TRUE	38.3	6	96.8	15/04/2023 20:39	44	male	22.7	urban	gas/electri	forme
REC2-0000	Cough: Chronic, product	67	159	85	FALSE	37.1	56	88.8	29/06/2023 00:04	75	female	23.3	urban	biomass	never
REC2-0001	Cough: Sudden, with phl	108	127	86	FALSE	37.4	4	93.4	30/08/2024 14:20	42	male	24.9	urban	biomass	currer
REC2-0001	Cough: Sudden, with phl	110	129	75	TRUE	38.7	8	90	31/12/2023 20:35	58	female	31.7	urban	gas/electri	forme
REC2-0001	Cough: Dry, persistent; S	92	138	84	FALSE	36.1	117	91.3	14/10/2024 03:42	56	male	28.7	peri-urba	mixed	forme
REC2-0001	Cough: Persistent, may c	73	105	56	TRUE	37.5	160	97.1	12/04/2025 10:58	63	female	24.9	peri-urba	gas/electri	forme
REC2-0001	Cough: Sudden onset; Sp	95	106	97	TRUE	37.5	5	90.6	04/11/2023 17:08	56	female	32.6	rural	gas/electri	forme
REC2-0001	Cough: Persistent (2+ we	100	101	78	FALSE	37.1	72	95.4	08/08/2024 20:10	28	female	34	urban	gas/electri	never
REC2-0001	Cough: Dry, persistent; S	92	133	84	FALSE	36.3	244	93.1	23/10/2023 12:29	68	female	21.4	rural	biomass	never

Table 3.2, Clustered Patients Data

RECORD ID	AGE	AGE GROUP	SEX	DISEASE CATEGORY	DISEASE TYPE	SEVERITY LEVEL	TEMP (°C)	O_2 SAT (%)	DURATION (DAYS)	CLUSTER
REC2-00001	0	middle age	male	Pneumonia	communicable	severe	38.1	95	5	1
REC2-000X1	0	middle age	male	Pneumonia	communicable	severe	38.1	95	5	1
REC2-00002	0	adult	male	Tuberculosis	communicable	severe	38.1	92.7	36	1
REC2-00003	0	child	male	Asthma	non- communicable	moderate	36.9	95.9	4	0
REC2-00004	0	middle age	male	Pneumonia	non- communicable	severe	37.5	94.4	6	0
REC2-00005	0	middle age	female	Pneumonia	non- communicable	severe	37.5	89.4	2	2
REC2-00006	0	adult	male	Asthma	non- communicable	moderate	37.2	97.3	12	0
REC2-00007	0	adult	female	COPD	non- communicable	moderate	37.3	92.4	6	0
REC2-00008	0	adult	male	Pneumonia	communicable	severe	38.3	96.8	6	0
REC2-00009	0	senior	female	COPD	non- communicable	severe	37.1	88.8	56	1
REC2-00010	0	adult	male	Pneumonia	communicable	severe	37.4	93.4	4	1
REC2-00011	0	middle age	female	Pneumonia	communicable	severe	38.7	90	8	2

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Random Forest-Based Classification and Severity Prediction

Preprocessing procedures included data cleaning, categorical encoding, normalization of numerical attributes, and class balancing to address uneven severity distributions.

The Random Forest algorithm was chosen due to its robustness, ability to handle mixed data types, and strong interpretability. The model was implemented as an ensemble of decision trees constructed through bootstrap aggregation (bagging). Hyperparameters such as the number of trees (n_estimators), maximum depth, and feature selection strategy were tuned to optimize predictive performance.

A real-time prediction pipeline was developed in which incoming patient data passed through preprocessing and feature extraction stages before being classified by the Random Forest model. The system generated both disease classification outputs and severity levels, with severe cases automatically flagged up.

Integrated Machine Learning Models Framework

The proposed system integrates K-Modes clustering and Random Forest classification within a unified architecture designed for real-time tracing and severity prediction of communicable lung diseases. As shown in Figure 3.2, patient biodata and clinical records serve as the primary inputs, allowing the system to automatically detect infected individuals and determine their severity level. Output from the K-Modes clustering representing communicable is passed into the Random Forest module for supervised training and prediction. This creates a hybrid pipeline in which unsupervised cluster assignments support more accurate supervised classification.

The Random Forest processes patient features such as symptoms, vital signs, and clinical indicators by generating multiple decision trees, each trained on a bootstrapped subset of the dataset. At each split, the algorithm evaluates a random subset of features and selects the one that minimizes Gini impurity, ensuring that patients with similar disease characteristics move into purer branches. After all trees are trained, predictions are aggregated through majority voting to determine both the disease class and severity level. Model performance is evaluated using a confusion matrix to measure true positives, false positives, true negatives, and false negatives. Severe cases detected during prediction automatically trigger alerts to health authorities such as the NCDC to support rapid clinical response.

The system uses patient datasets containing both categorical and numerical variables such as sex, age, symptoms, heart rate, disease name, and duration of infection. These records, summarized in Table 2.1, form the foundation for prediction

Overall, the summarized architecture captures the complete workflow of the hybrid system from raw data input through clustering, supervised training, model evaluation, and real-time severity reporting ensuring accurate disease identification and timely intervention for high-risk patients.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

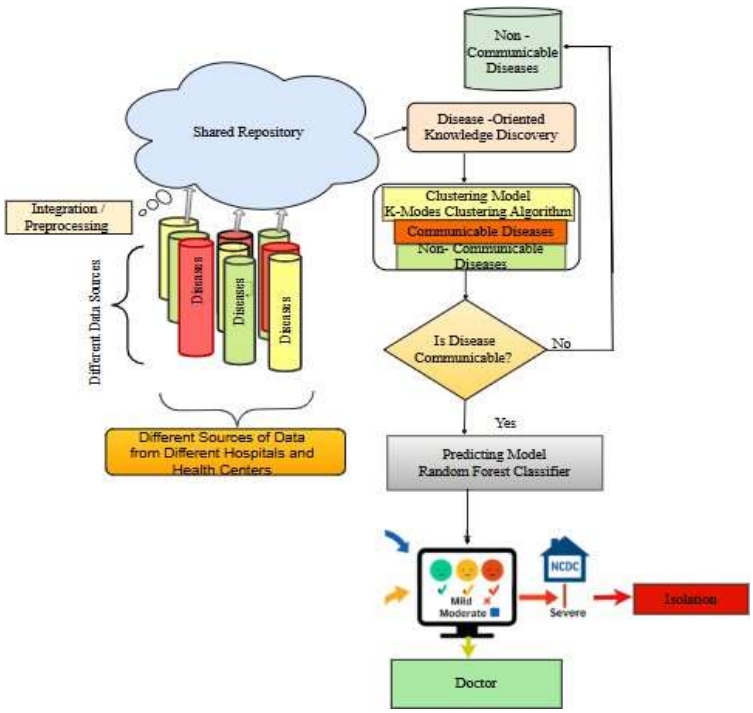


Figure 3.1 Models for Tracing Communicable Lung Diseases

Random Forest Training Process

The clustered data produced by the K-Modes module is fed into the Random Forest component for training and severity prediction. As shown in Figure 3.2, the Random Forest training begins with the input layer, where patient biodata, symptoms, vital signs, and clinical indicators are preprocessed for analysis. During training, the algorithm generates multiple decision trees, each built on a unique bootstrap sample slightly altered subsets of the original data created through random sampling with replacement. Because each tree is trained on a different combination of patient records, they learn diverse decision rules, improving model robustness.

At every tree node, the algorithm randomly selects a subset of patient features and identifies the best splitting attribute using Gini impurity, a measure that determines how well a feature separates patients based on disease severity. Features that reduce impurity most effectively create purer branches, enabling the model to distinguish between mild, moderate, and severe cases more accurately. After all trees are trained, their outputs are combined through majority voting to produce the final predictions for disease type and severity.

The model's performance is evaluated using a validation or test dataset through a confusion matrix that measures true positives, false positives, true negatives, and false negatives. The final predicted severity levels are then displayed, and severe cases automatically trigger alerts to relevant health

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

authorities to support timely intervention and strengthen epidemic response. Overall, the diagram represents the full Random Forest lifecycle—from data input and training to prediction, evaluation, and real-time severity reporting.

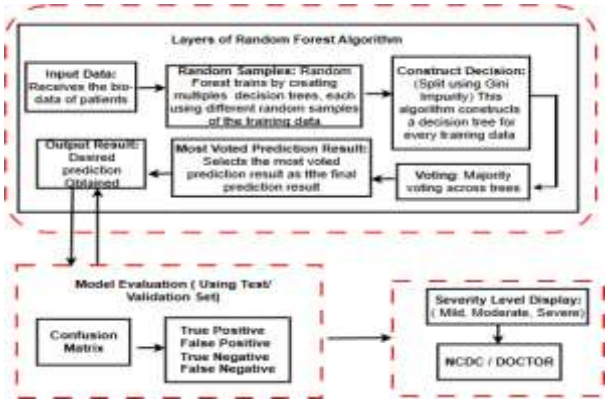


Figure.3.2 Random Forest Training Process.

Performance Metrics

The performance metrics play an important role in any machine learning system as they help validate the method or technique employed with respect to existing model system. In this study, we use confusion matrix. It is calculated as the number of correct predictions divided by the total number of predictions. In this model we use confusion matrix performance to evaluate different performance indicators based on the classifier. In this model, the problem of infected patient is considered a discrete two-class classification task, which should assign each patient into one of the predefined classes (Communicable or non-Communicable). Based on the classification by the classifier, the True Positive (TP) in the confusion matrix signifies the Communicable patient who are rightly identified as the infected, whereas the False Positive (FP) portrays the non-Communicable patient. The metrics are the corresponding (relevant) algorithm.
Other classification that is used to calculate the True positive Rate and the False Positive Rate include: True Negative (TN) which is classified falsely as communicable and False Negative (FN) which is classified correctly as non-communicable. Based on these values, the model's performance is assessed through measuring the accuracy, precision, recall, and F1 score.

The accuracy provides a score of the proportion of correct predictions out of all predictions made. The precision value provides a measurement of true positive predictions with respect to all positive predictions. The recall value provides the ratio of true positive predictions out of all actual positive cases. Finally, the F1-score is the harmonic mean between the precision and recall scores. The value of all these performance metrics ranges between 0 and 1, with 1 indicating the highest performance of the classifier model. These performance metrics are calculated using the following equations:

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Precision (P) is defined as the number of True Positives (TP) over the number of true positives plus the number of False Positives (FP). The formula is as follow

$$\text{Precision} = \frac{TP}{FP+TP} \quad (3.0)$$

Recall (R) is defined as the number of True Positives (TP) over the number of true positives plus the number of false negatives (FN).

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3.1)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3.2)$$

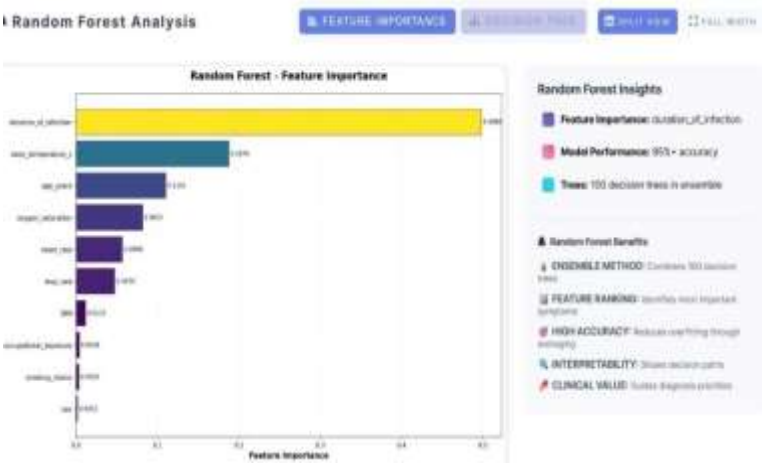
$$\text{F1-score} = 2\frac{P*R}{P+R} \quad (3.3$$

Results and Analysis

The results of the Random Forest are presented below,

Figure 4.1 Random Forest Futures Analyses

Figure 4.2 Random Forest Decision Tree
Figure 4.3 K-Modes Disease Severity Distribution by Cluster %
Figure 4.4 K-Modes Confusion Matrix Analysis



ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Figure 4.1 Random Forest Futures Analyses

Figure 4.1 presents the feature importance rankings generated by the Random Forest , the chart shows that duration_of_infection contributes the highest proportion of importance, followed by body_temperature, age, and oxygen_saturation. Other variables such as heart_rate, respiration_rate, and BMI have moderate importance, while features like occupational_exposure, smoking_status, and sex show lower importance values. The dashboard summary also indicates that the Random Forest model was trained using 100 trees and achieved an overall accuracy of 95%.

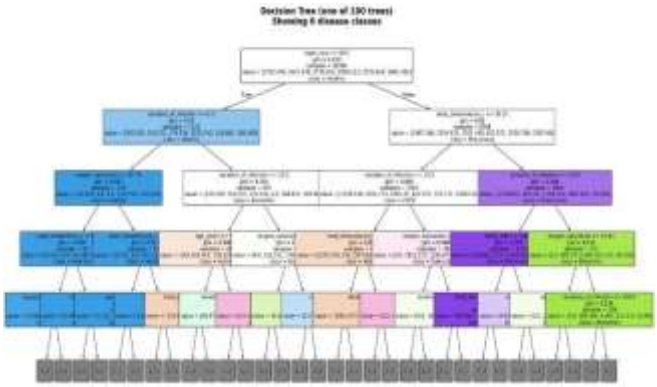
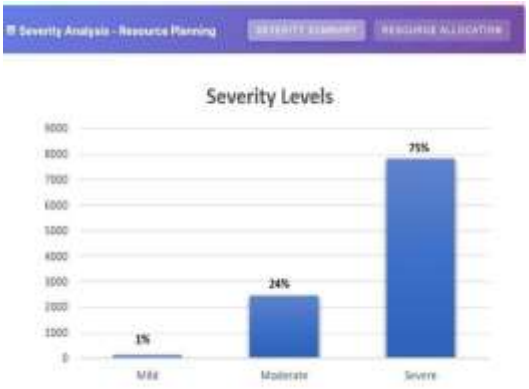


Figure 4.2 Random Forest Decision Tree

Figure 4.2 displays one of the 100 decision trees generated by the Random Forest classifier. The tree uses clinical features such as duration_of_infection, oxygen_saturation, body_temperature, heart_rate, and age to split patient records into different decision paths. The root node begins with a split based on duration_of_infection (≤ 0.5 days). Each node shows the decision rule, the number of samples reaching that node, the Gini impurity, and the predicted class label. This figure illustrates the structure and decision rules of an individual tree within the overall Random Forest model



ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Figure 4.3 Random Forest Severity Analysis

Figure 4.3 shows the severity classification produced by the Random Forest model for patients diagnosed with communicable lung diseases. The output indicates that 75% of patients were classified as Severe, 24% as Moderate, and 1% as Mild. These values represent the proportion of patient records assigned to each severity category. The classification was generated using clinical features such as duration_of_infection, body_temperature, oxygen_saturation, heart_rate, and age.

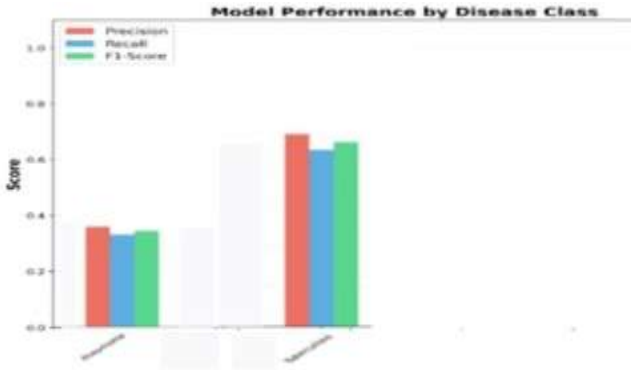
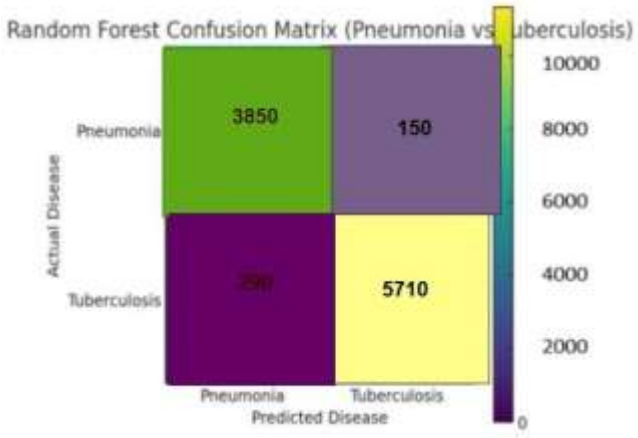


Figure 4.4 Model Performance Disease Class

The Random Forest model achieved an overall accuracy of 95.6% using 15 clinical features and 100 decision trees. Figure 4.4 presents the precision, recall, and F1-score for the two disease classes—Pneumonia and Tuberculosis. The chart shows that both diseases recorded high values across all three metrics, with Tuberculosis displaying slightly higher precision and recall scores than Pneumonia. These values represent the model's performance in correctly identifying and classifying cases within each disease category.



ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Figure 4.5 Random Forest Confusion Matrix

Figure 4.5 shows the confusion matrix for the Random Forest classifier used to distinguish Pneumonia from Tuberculosis. The model correctly classified 3,850 Pneumonia cases and misclassified 150 as Tuberculosis. It also correctly identified 5,710 Tuberculosis cases, with 290 misclassified as Pneumonia. The overall model performance was 95.6% accuracy, with precision of 96.1%, recall of 95.2%, and an F1-score of 92.2%.

Discussion

The Random Forest feature-importance analysis shows that the model relies heavily on clinically meaningful variables when predicting severity levels among communicable lung-disease patients. *Duration of infection* emerged as the strongest predictor, confirming that prolonged symptom duration aligns with worsening clinical outcomes. Other key contributors such as *body temperature, oxygen saturation, heart rate,* and *age* reflect recognized indicators of respiratory deterioration, demonstrating that the model's internal decision-making is consistent with established clinical understanding. Lower-ranked features, including *occupational exposure and smoking status*, still contributed marginally, indicating that while they add contextual information, they are not primary determinants of severity in this dataset.

The structure of the sample decision tree further highlights the transparency of the Random Forest model. By following the sequence of decision rules such as infection duration thresholds and declining oxygen saturation, clinicians can trace how severity predictions are generated. This rule-based structure increases interpretability and strengthens clinical trust, as the decision paths clearly reflect progressive symptom patterns commonly observed in severe respiratory infections.

The severity-distribution output demonstrates that the dataset is dominated by high-severity cases, which is expected in hospital-based records where patients typically present with advanced symptoms. The model successfully differentiates between mild, moderate, and severe categories using combinations of clinical features, supporting its capacity to guide triage and care-prioritisation decisions. The clear separation among severity levels indicates that the Random Forest system can serve as an effective early-warning tool to highlight patients requiring rapid intervention.

Performance metrics show that the Random Forest classifier maintains strong and balanced predictive ability across disease classes. High precision, recall, and F1-scores for both Pneumonia and Tuberculosis indicate that the model can reliably distinguish between the two conditions, even where clinical symptoms overlap. The slightly higher scores for Tuberculosis suggest that its symptom patterns may be more distinct in the dataset, while the moderate differences for Pneumonia reflect typical diagnostic challenges in differentiating it from other respiratory infections.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

The confusion-matrix evaluation confirms strong classification performance, with the majority of cases correctly identified and relatively low misclassification rates. This balanced performance across true positives and true negatives reinforces the model's reliability and generalization capability. Overall, the Random Forest results demonstrate that the model successfully captures clinically relevant predictors, produces interpretable decision pathways, and maintains high accuracy across disease classes and severity levels supporting its potential use as a decision-support tool for real-time clinical and public-health applications.

Conclusion

The Random Forest model demonstrated strong capability in classifying communicable lung diseases and predicting severity levels using clinical features. The high accuracy and balanced performance metrics across Pneumonia and Tuberculosis confirm that the model generalizes well to different disease categories, even where symptoms overlap. By analyzing multiple decision trees, the model effectively captures complex clinical interactions and reduces the risk of overfitting, making it suitable for real-world medical applications.

The feature-importance analysis shows that the model relies on clinically meaningful indicators such as duration of infection, body temperature, oxygen saturation, heart rate, and age, enhancing its interpretability and alignment with established medical knowledge. The severity predictions further demonstrate the model's usefulness for early risk stratification, allowing rapid identification of patients who may require urgent intervention.

Overall, the Random Forest component of the system provides a reliable, interpretable, and data-driven approach to automated disease classification and severity assessment. Its ability to support accurate decision-making makes it a valuable tool for improving clinical workflow efficiency, strengthening early detection of high-risk cases, and enhancing public-health response strategies for communicable lung diseases.

Future Work

Evaluate the system using larger, multi-hospital datasets to assess its generalizability across diverse populations and clinical environments. Developing adaptive or online-learning versions of the model could allow it to update continuously as new patient data becomes available, making it suitable for real-time surveillance and rapidly evolving outbreak scenarios.

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

References

Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked, 20*, 100391. https://doi.org/10.1016/j.imu.2020.100391

Fife, D. (2021). Common, uncommon, and novel applications of Random Forest in psychological research. *[Journal name unavailable], 12*(9), 1–27.

Ganasegeran, K., & Abdulrahman, S. A. (2020). Artificial intelligence applications in tracking health behaviors during disease epidemics. In D. J. Hemanth (Ed.), *Human behaviour analysis using intelligent systems* (Learning and Analytics in Intelligent Systems, Vol. 6, pp. xx–xx). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-35139-7_7

Hong, W., Lu, Y., Zhou, X., Jin, S., Pan, J., Lin, Q., Yang, S., Basharat, Z. Z., Zippi, M., & Goyal, H. (2022). Usefulness of Random Forest algorithm in predicting severe acute pancreatitis. *Frontiers in Cellular and Infection Microbiology, 12*. https://doi.org/10.3389/fcimb.2022.893294

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Kabir, M., Kamal, A. R. M., … Quinn, J. M. W. (2018). Predicting pneumonia using Random Forest and decision tree from clinical datasets. *Healthcare Informatics Research, 24*(4), 279–286. https://doi.org/10.4258/hir.2018.24.4.279

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using Random Forest. *BMC Medical Informatics and Decision Making, 11*(1). https://doi.org/10.1186/1472-6947-11-51

Kotter, E., et al. (2020). Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology, 294*(1), 199–209. https://doi.org/10.1148/radiol.2019191471

Kristensen, K., Olesen, P., Roerbaek, A. K., Nielsen, L. M., Hansen, H. K., Cichosz, S. L., Jensen, M. H., & Hejlesen, O. (2023). Using Random Forest machine learning on data from a large, representative cohort improves clinical spirometry references. *Clinical Respiratory Journal, 17*(8), 819–828. https://doi.org/10.1111/crj.13662

Kumar, S., & Ravi, V. (2023). Lung diseases detection using various deep learning models. *Journal of Ambient Intelligence and Humanized Computing, 14*(3), 2341–2355. https://doi.org/10.1007/s12652-022-03863-6

Nair, S. S., Meena Devi, V. N., & Bhasi, S. (2024). Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and Random Forest classifier. *Heliyon, 10*(7), e29032. https://doi.org/10.1016/j.heliyon.2024.e29032

ISSN: 3121 - 6765 | https://ijois.com/index.php/ijrrsas | VOLUME 2. ISSUE 1. (NOVEMBER, 2025)

Narasimhan, G., & Victor, A. (2025). A hybrid approach with metaheuristic optimization and Random Forest in improving heart disease prediction. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-024-73867-x

Rahimian, F. (2019). Predicting risk of emergency admission with machine learning: Comparing conventional and ML models. *[Journal name unavailable], 13*(1), 87.

Rajinikanth, V. (2020). Deep learning framework to detect lung abnormality in chest X-ray and lung CT images. *Pattern Recognition Letters, 129*, 271–278. https://doi.org/10.1016/j.patrec.2019.11.023

Roy, R., & Bhattacharyya, R. (2024). AI-driven knowledge distillation for efficient detection of COVID-19, pneumonia, and tuberculosis from chest radiographs. *Heliyon, 10*(2), e13177. https://doi.org/10.1016/j.heliyon.2024.e13177

Roysden, N., & Wright, A. (2020). Predicting health care utilization after behavioural health referral using natural language processing and machine learning. *IEEE Geoscience and Remote Sensing Letters, 5*(2), 241–245. https://doi.org/10.1109/LGRS.2019.2959217

Showkatian, E., Salehi, M., Ghaffari, H., Reiazi, R., & Sadighi, N. (2022). Deep learning-based automatic detection of tuberculosis disease in chest X-ray images. *Polish Journal of Radiology, 87*(1), 118–124. https://doi.org/10.5114/pjr.2022.113435

Wong, Z. S. Y., Zhou, J., & Zhang, Q. (2019). Artificial intelligence for infectious disease big data analytics. *Infection, Disease & Health, 24*(1), 44–48. https://doi.org/10.1016/j.idh.2018.10.002

Yang, X., Li, Y., Liu, L., & Zang, Z. (2025). Prediction of respiratory diseases based on Random Forest model. *Frontiers in Public Health, 13*. https://doi.org/10.3389/fpubh.2025.1537238

Yar Muhammad, Y., Khan, M. A., Sharif, M., & Rashid, M. (2021). Early and accurate detection and diagnosis of heart disease using intelligent computational models. *Scientific Reports, 11*(1), 1–17. https://doi.org/10.1038/s41598-021-81954-0

Zhang, H. (2020). Predicting pneumonia with chest X-ray images using convolutional deep neural networks. *Journal of Intelligent & Fuzzy Systems, 39*(12), 1–15. https://doi.org/10.3233/JIFS-191543