Shaik Rahena, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Intelligent Bronchiectasis Risk Prediction Using Explainable Machine Learning Models

Shaik Rahena, Associate Professor Mrs.M.Radhika
Department of CSE, Mam Women's Engineering College

Abstract- There is a substantial clinical and health care burden associated with bronchiectasis, a chronic respiratory disorder characterised by frequent exacerbations, persistent symptoms, and frequent hospitalisations. Conventional treatment techniques rely on clinician opinion and assessment based on guidelines, which may not adequately reflect the unique risk heterogeneity of each patient. This research presents a paradigm for bronchiectasis risk classification and intervention assistance using machine learning and routinely gathered clinical data. The purpose of analysing a structured dataset that includes demographic, clinical, etiological, lifestyle, and treatment-related variables is to divide patients into groups with varying degrees of risk. Stratified cross validation testing is used for a variety of machine learning models, including feedforward neural networks, Random Forest, XGBoost, and Logistic Regression. Based on the data, it seems that the baseline clinical characteristics are generally linearly separable, as Logistic Regression provides the greatest prediction performance. Ensemble and neural models work well together and compete well in terms of interpretability and decision-support. Applying explainable AI with SHAP helps to make sense of model predictions and isolate key risk variables. Further patient variables uncovered by unsupervised clustering included exacerbation load and disease duration. The suggested approach combines phenotyping, predictability, and explainability to help with individual-level respiratory health management and data utilisation for treatment decision-making in bronchiectasis.

Keywords— Bronchiectasis, ML, LR, Explainable AI, STI, Phenotyping of Patients, Machine Learning.

I. INTRODUCTION

Permanent airway dilatation, persistent coughing up of mucus, sputum, and recurrent infections and exacerbations are hallmarks of bronchiectasis, a chronic and progressive lung disease. Patients' functional decline and repeated readmissions to the hospital cause high healthcare costs, low quality of life, and an increase in healthcare use. Clinical management of bronchiectasis remains challenging due to variations in illness aetiology, severity, and responsiveness to therapy, despite advancements in respiratory medicine [1]. Current therapeutic approaches are mostly reliant on physician discretion and recommendations derived from guidelines, neither of which may accurately represent patients' risk profiles. Traditional statistical approaches are ill-suited to describe the complicated and nonlinear connections between variables such as age, illness duration, aetiology, smoking history, chronic cough, antibiotic usage, physiotherapy, and oxygen treatment [2]. Consequently, there is an increasing need to use data-driven methods that can integrate multi-dimensional clinical data in order to provide individualised respiratory therapy [3].

Disease prediction, risk prioritisation, and decision assistance are three areas where machine learning (ML) has shown promising results in healthcare [4]. However, it has not been extensively used in bronchiectasis, particularly in creating interpretable models to direct therapeutic therapies. This research begins to fill that void by presenting a system driven by machine learning that can forecast a patient's risk and provide interpretable data and phenotyping to guide targeted treatments.

Mucus builds up in the airways due to bronchiectasis, a chronic heterogeneous respiratory condition that becomes worse with time and requires more medical treatment. Because of their worse prognosis and under-representation in conventional severity indices and evaluation based on guidelines, high-risk and frequent-exacerbator phenotypes must be differentiated, according to recent clinical research [5, 6]. But traditional severity indices and guidelines-based measurements don't allow for customisation and can't handle complex interactions across demographic, etiological, and treatment-related phenotypes [7]. Thanks to recent developments in ML, it is now much simpler to use patient data to stratify those with chronic lung disorders and to predict their prognosis. Gradient Boosting, Logistic Regression, and Random Forest are supervised ML models that, when trained on routinely collected clinical data, often outperform more traditional statistical approaches in terms of prediction accuracy [8-9]. Research based on ML is essential for safe clinical deployment, but there is a dearth of literature on the subject at the moment. Furthermore, the majority of bronchiectasis investigations have disregarded characteristics related to interpretability, calibration reliability, and leakage-free validation [10]. An increasing number of machine learning (ML) models have been using SHAP-based Explainable Artificial Intelligence (XAI) techniques to make them more visible and credible to doctors in recent years. Recent studies have shown that interpretable models are important for assisting high-stakes medical decision-making [11–13], and they may often provide performance that is equivalent to complicated black box models. At the same time, a growing number of studies are using unsupervised learning to patient phenotype, which has led to the identification of subgroups with therapeutic relevance, such as those with unique patterns of disease and those who have frequent exacerbations [14, 15].

Notable outcomes derived from this This paper's main arguments are as follows: first, that a leakage-free ML framework can be built to stratify respiratory risk using real-world clinical features; second, that various ML models can be compared using stratified cross-validation; third, that significant drivers of respiratory risk can be identified through explainable AI integration; and finally, that clinically meaningful patient phenotypes can be discovered through unsupervised learning.

II. INFORMATION ABOUT THE DATASET AND ITS PRECESSING

Medical professionals can only benefit from machine learning and deep learning when they work with trustworthy, clinically relevant patient data. Accurate clinical interpretation of prediction outcomes, effective model training, and bias-free data pre-processing are all dependent on well-characterized datasets. Here we detail the dataset's components and methodology for obtaining it. Additionally, it delves into the ethical issues that were taken into account while gathering data and the methodical processing that was carried out to enhance data quality, eliminate inconsistencies, and prepare the data for model development and testing.

Ethical Compliance and Dataset Source

Following official ethical permission, anonymised patient records were obtained from a tertiary-care hospital and used in this investigation. To guarantee privacy and adhere to accepted ethical standards in research, we eliminated all patient identification before analysis.

Factors Related to Health and Demographics

The dataset includes a wide range of factors pertaining to demographics, clinical aspects, lifestyle, aetiology, and therapy. In this context, "patient characteristics" refer to factors like age, illness duration, frequency of exacerbations, volume of sputum, antibiotic use, history of hospitalisation, factors such as gender, smoking status, the existence of a chronic cough, the use of oxygen, and physical therapy. Clinically meaningful stratification is achieved by classifying etiological variables into four groups: chronic obstructive pulmonary disease (COPD), idiopathic, post-infectious, and post-tuberculosis.

Defining Outcomes and Assigning Risks

In order to classify patients as either low-, moderate, or high-risk, a risk label is created according to

clinical motivation. Commonly recognised as markers of disease severity, such clinical variables as sputum volume, history of hospitalisation, and exacerbation frequency form the basis of this categorisation.

Ensuring Data Quality and Cleaning

It was necessary to consolidate duplicate variables, eliminate index-like or non-informative variables, and fix column inconsistencies as part of the data preprocessing. The goal of these initiatives was to improve the data's consistency, quality, and suitability for machine learning.

Feature Picking and Data Leak Prevention

In order to train the model without leaking data and to evaluate the risk model fairly, we excluded the variables used to create the risk labels from the feature set. Rather than relying on direct inference from outcome-defining factors, our method ensures that predictive performance is grounded on real learning.

Section F: Sorting Data by Category and Applying Numerical Standards

We have achieved uniform scaling by standardising numerical characteristics and encoding categorical variables into binary representations. Balanced learning, better model convergence, and more stable predictions are all benefits of this preprocessing phase.

III. METHODS

Figure 1 is a flow diagram of the whole machine learning system that helps with clinical decision support and risk stratification for bronchiectasis. A clinically generated risk label is derived after data loading, cleaning, and preprocessing; outcome defining factors are omitted to prevent data leaking. Stratified K-fold cross-validation is then used to train and evaluate a feedforward neural network, Random Forest, Logistic Regression, and XGBoost, among other supervised learning models. The final performance is assessed on a concealed test set that includes confusion matrix measurements. Next, the Random Forest model is enhanced using SHAP-based explainable AI to uncover key drivers of risk prediction, while latent patient characteristics are identified via unsupervised K-means clustering. Interventions for the treatment of bronchiectasis may be better planned with the use of data when phenotyping, explainability, and predictive performance are all taken into account.

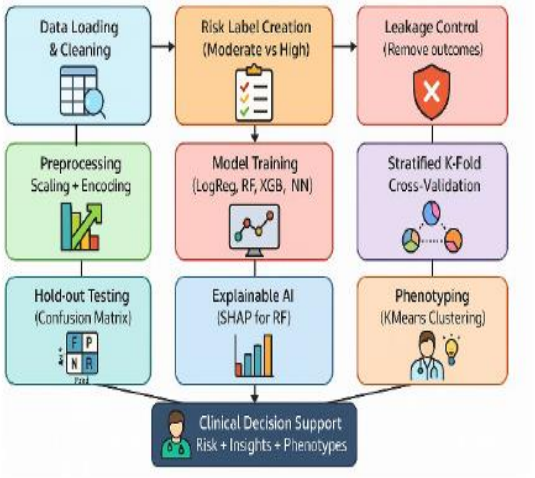

Figure 1: A Proposed Model for the Management of Bronchiectasis Based on Machine Learning

After then, the training data is used to train the suggested model, which allows it to discover useful patterns and correlations. The prediction results are obtained by applying the learned model to the testing data once training is complete. Finally, appropriate performance metrics are used to assess the accuracy, dependability, and generalisability of the proposed framework, which in turn determines the sufficiency of predictions.

Formulating the Problem

The primary objective of our study is to classify bronchiectasis patients into distinct risk categories using demographic and clinical data collected at the outset. The formulation facilitates the early identification of patients at high risk and their proactive intervention planning.

Models for Machine Learning

This research looks at the categorisation performance of four different supervised learning models. Because of its user-friendliness and clear results, Logistic Regression (LR) is a popular choice for basic linear classifiers. When dealing with nonlinear characteristics and interactions, an

ensemble tree called Random Forest (RF) is used. This tree uses many decision trees to learn the features. Its capacity to handle complex patterns with high prediction power and accuracy has led to Extreme Gradient Boosting's (XGBoost) widespread use. To further records non-linear relationships between characteristics and depends on higher-level data, feedforward Neural Networks (NNs) are used.

Section C: Cross-Validation Plan

To ensure that all folds in K-fold cross-validation have an equal representation of classes, it is stratified. To ensure statistical validity and prevent class omission, the number of folds is adaptively selected based on the smallest class size.

Section D. Patient Diagnosis

Standardised clinical characteristics, including outcome-related variables, are employed in unsupervised K-means clustering to uncover hidden patient phenotypes. Differences in age, duration of illness, exacerbation rate, and hospitalisation are described using cluster level analysis.

IV. RESULTS AND DISCUSSION

Following the completion of the text update, the document is made ready for the template. Make an additional duplicate of the form. file by selecting "Save As" from the file menu. When naming your paper, follow the guidelines set forth by your conference. After you've made the new file, import your prepared text file and make sure to highlight everything in it. To apply styles to your document, go to the Formatting toolbar in Microsoft Word and look for the scroll box on the left.

Section A: Dataset Features and Risk Classification

Demographic, clinical, etiological, lifestyle, and treatment-related characteristics were among the nineteen variables included in the dataset, which included one thousand records of patients with bronchiectasis. Based on factors such as the frequency of exacerbations, volume of sputum, and hospitalisation history, patients are classified into two risk groups: High risk (55.6% of cases) and Moderate risk (44.4% of cases). As is typical of respiratory data from tertiary care facilities, the data is mostly a symptomatic clinical cohort due to the absence of low risk patients. Patients requiring intensive care and those whose illness was reasonably stable might be identified using this binary risk model.

How Well Machine Learning Models Fare in Cross-Validation

To generate a reliable and objective performance estimate, we used Stratified K-fold cross-validation with K = 5, maintaining the class percentage within the folds. Table 1 summarises the findings of the accuracy and macro-F1 scores of the evaluated models. The best overall performance was achieved by Logistic Regression, with a mean accuracy of 0.731 and a macro-F1 score of 0.720. This provides further evidence that a linear combination of normal clinical characteristics may be able to discriminate between persons at moderate and high risk of bronchiectasis. Another proof of Logistic Regression's great generalisability is its uniformity in the folds. Random Forest's intrinsically interpretable nature and competitive outcomes are shown in its 0.694 average accuracy and 0.687 macro-F1 score. Although Logistic Regression is more accurate at making predictions, Random Forest is more practical in clinical settings because to its explainability and ability to depict feature interactions. In comparison to the other approaches, XGBoost's mean accuracy was 0.670 and its macro-F1 score was 0.665. The dataset seems to have been too simple and there are too many binary clinical indications for boosting-based nonlinear modelling to be of much use. A feedforward neural network achieved 0.715 average accuracy and 0.704 macro-F1 score, which is comparable to Logistic Regression. However, there may not be a substantial performance boost from deep nonlinear representations, which makes them unusable with the current dataset size and feature composition. Finally, our results demonstrate that simpler models may achieve competitive performance in bronchiectasis risk classification if clinically essential variables are appropriately selected and leakage is avoided.

Examining the Effectiveness of Different Machine Learning Models via Cross-Validation

Model	*Mean Accuracy*	*Mean Macro-F1*
Logistic Regression	0.731	0.720
Random Forest	0.694	0.687
XGBoost	0.670	0.665
Neural Network	0.715	0.704

Tabulated in Table I are the average macro-F1 scores and classification accuracy achieved by stratified 5-fold cross-validation. The most accurate prediction was made by Logistic Regression, while ensemble and deep learning models were close behind.

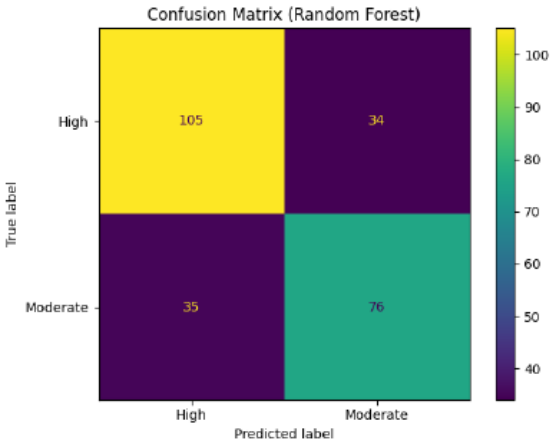

Figure 2: Random Forest model's global feature relevance as measured by SHAP

Figure 2 shows the confusion matrix that summarises the Random Forest model's performance on the independent test set for classifying bronchiectasis patients as either high-risk or moderate-risk. The confusion matrix study shows that out of 105 high-risk patients, the model got it right, but it got 34 high-risk cases wrong, labelling them as moderate risk. On top of that, 35 instances with a moderate risk were mistakenly labelled as high risk, whereas 76 patients with a moderate risk were appropriately identified (true negatives). A recall of around 0.76 and a precision of about 0.75 were reached by the high-risk class, according to the class-based performance analysis. This indicates that the clinically important patients were successfully detected with acceptable alert reliability. The recall for the moderate-risk class was around 0.68, with most of the misclassifications falling into the high-risk category. This is better from a clinical standpoint since it reduces the likelihood of underestimating the severity of the condition.

With its great sensitivity, the Random Forest model can identify high-risk bronchiectasis patients and almost all severe cases. As a therapeutically acceptable trade-off, the increased frequency of misclassifications—such as the identification of moderate-risk patients as high-risk—allows for earlier intervention and closer monitoring. When it comes to managing respiratory illness with underlying chronic disorders, the model fits well to risk-averse clinical decision-making overall.

With most misclassifications skewed toward overestimation of hazards, the Random Forest classifier allows for better clinical decision-making, and it is highly sensitive to instances of severe bronchiectasis.

Last Test on Hold-Out Set for Model Evaluation

The Random Forest model was also tested on the held-out test set, which constitutes 25% of the dataset (n = 250), so that the whole model analysis could be carried out and the results could be explained. With findings that are similar in both risk categories, the model yields an accuracy of 0.72.

Most people who need immediate medical attention have been correctly identified, thanks to the high-risk category and a recall of 0.76. This sensitivity to severe events is especially crucial in clinical decision-support systems, since a catastrophic outcome might ensue from failing to detect high-risk individuals. With a recall of 0.68, most of the misclassifications in the moderate-risk group were categorised as overestimating risk, which is clinically safer than underestimating.

This study demonstrates that the suggested framework achieves a satisfactory compromise between sensitivity and specificity, making it suitable for use in risk-based patient triage.

Calibration Analysis

Combining testing of distinguishing measures with model calibration defines the quality of risk probability prediction. Close agreement between the

expected and actual high-risk frequencies across probability bins indicates that the Random Forest model is well-calibrated, as shown by the calibration (reliability) curve in Fig. 3. A decent estimate of the likelihood may also be supported by the matching Brier score.

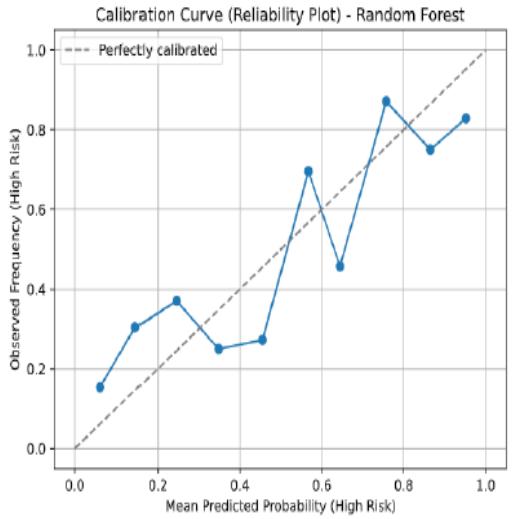

Figure 3 shows the Random Forest model's calibration (reliability) curve on the hold-out test set. When making clinical decisions based on risk, intervention planning is guided by probability thresholds.

these results provide justification for using model outputs. Maximum calibration is shown by the dotted diagonal. The fact that the curve is so near to the diagonal indicates that the forecasts and actual findings are quite congruent.

Predicted Probability Distribution

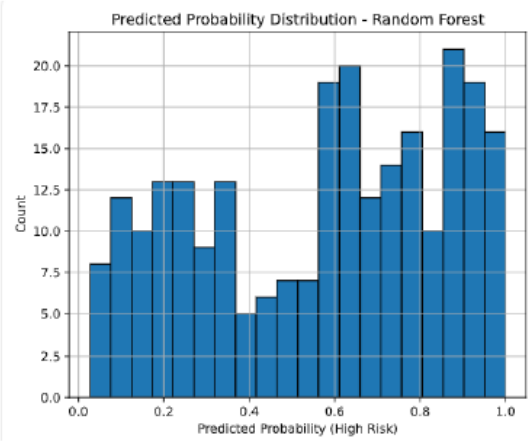

Fig.4. Distribution of predicted high-risk probabilities generated by the Random Forest model on the hold-out test set

One way to see how the model handles risk stratification is in Fig.4, which displays the distribution of the anticipated probability. A large range of probability estimates, with clear differences between low and high confidence levels, are provided by the random forest model. Consistent with the model's sensitivity to clinically severe situations, a big proportion of patients get high probability on their estimated risks. In place of rigid binary categorisation, this probability model permits adaptive intervention planning and aids in the implementation of threshold-based clinical decision-making standards.

SHAP-Based Explainable AI Analysis

More transparency and clinical confidence are achieved by enhancing the Random Forest model with SHAP-based explainability, which allows one to get insight into the algorithm's decision-making process. Based on the results of the SHAP summary analysis, the following characteristics are most predictive of bronchiectasis risk: age, length of illness, aetiology indicators, presence of persistent cough, smoking history, and therapy-related variables (oxygen and physiotherapy usage). Additional criteria added to this list. In addition, data at the patient level was supplied by directional SHAP values, which show how certain feature profiles affected the high- or moderate-risk patient categorisation. This interpretability is crucial for clinical adoption because it allows clinicians to better tailor treatments to specific patients by evaluating model behaviour in light of their domain expertise.

We used accuracy, recall, and F1-score to measure the model's performance. Table 2 shows that the classifier successfully recognises clinically relevant occurrences in the high-risk class (support = 139) with an F1-score of 0.75, which is the result of an accuracy of 0.75 and a recall of 0.76. In the group at intermediate risk, the results demonstrated a recall of 0.68 and a precision of 0.69, with an F1-score of 0.69 indicating a greater level of misclassification (support = 111). A general classification The

predictive power of the Random Forest model was examined using a distinct hold-out test set, as seen in Table II. Precision, recall, and F1-score were the metrics used to determine the outcomes. The model demonstrated excellent performance in identifying patients with critical clinical needs, with an F1-score of 0.75 in the high-risk category (support = 139) and an accuracy of 0.75 and a recall of 0.76. The moderate-risk group (support = 111) achieved an F1-score of 0.69, a recall of 0.68, and an accuracy of 0.69. Based on 250 samples, the performance was equally distributed across classes, as shown by an F1-score of 0.72 and a total classification accuracy of 0.72. Its recall was 0.72 and its macro-averaged accuracy was 0.72 as well. To minimise the danger of incorrectly underestimating the severity of the ailment while still improving sensitivity, the majority of misclassifications tend to identify patients as high-risk.

Hold-Out Test Set Performance of the Random Forest Model

Class	*Precision*	*Recall*	*F1-Score*	*Support*
High Risk	0.75	0.76	0.75	139
Moderate Risk	0.69	0.68	0.69	111
Overall Accuracy	–	–	**0.72**	250
Macro Average	0.72	0.72	0.72	250

The Random Forest-based bronchiectasis risk prediction model's SHAP summary map is shown in Figure 5. It provides an overview of the key aspects and the ways they influence the patient population on a worldwide scale.

Leading causes of adverse clinical outcomes: Figure 5 displays the SHAP summary plot, which reveals that the most important variables in predicting the risk of bronchiectasis are age, the number of years that the illness has been present, indicators linked to the aetiology, the presence of a persistent cough, and the use of oxygen and physiotherapy as treatments. Additionally, these characteristics possess the most elevated SHAP magnitudes. This provides support for the hypothesis that baseline patient characteristics and the burden of long-term disease are crucial for risk categorisation.

Age is the most important factor as it is consistently associated with higher SHAP values, which indicates that age significantly contributes to being classified as high-risk. Ageing has a multiplicative impact on lung function decline and comorbidity burden in bronchiectasis patients. Sickness duration (in years) is therefore the second most relevant feature since the model predicts a higher risk with a longer illness history. By taking into consideration greater SHAP contributions, the model accurately depicts the increasing airway damage and repeated inflammation seen in patients with extended disease duration. It is evident that the presence of a persistent cough has a direct and beneficial effect on high-risk forecasts. This finding is clinically relevant since a persistent cough is an indicator of disease activity, mucus retention, and persistent airway irritation. For factors associated to aetiology, there is a lot of SHAP variety, which means that various subgroups of aetiology (such as COPD, post-infectious syndrome, and post-TB) contribute to risk in different ways. The underlying sickness aetiology must be considered when bronchiectasis severity is evaluated. For certain aetiologies, the predictions are more accurate than others. Finally, treatment usage variables have large SHAP values, particularly for physiotherapy and oxygen therapy. Patients with advanced or poorly controlled illness are more likely to get these medications, which is why these features serve as indirect markers of disease severity. A larger contribution to the high-risk class indicates this.

Ultimately, the SHAP analysis confirms that the model's predictions align with recognised risk factors for bronchiectasis and are grounded in clinically meaningful variables. This further enhances the validity and interpretability of the proposed ML framework.

The direction of the features' influence: a higher concentration of larger values (red dots) on the positive SHAP axis for many factors indicates a higher probability of being classified as high risk. Older age and longer illness duration, for instance, always push projections into the high-risk area. An

increase in risk is likely if a persistent cough is present and if certain etiological variables are present. Conversely, the lower values (blue dots) focused on the negative axis of SHAP reflect the moderate-risk group.

Patient-specific variability and effects: inter-patient heterogeneity is represented by the vertical propagation of SHAP values for individual variables. Because the same attribute could have varying magnitude affects on different people, the risk of bronchiectasis might vary widely. Machine learning models outperform rule-based scoring systems because they can tailor their forecasts to individual profiles. This is the reason for the disparity.

Clinical interpretability and trustworthiness give models their credibility; SHAP analysis adds to what is already known about bronchiectasis from a clinical perspective. If a patient is classified as high risk based on the values of their unique traits, the SHAP framework may explain why this is the case at the patient level so that clinicians can understand why. This openness is critical for the effectiveness of clinical implementation and collaborative decision-making.

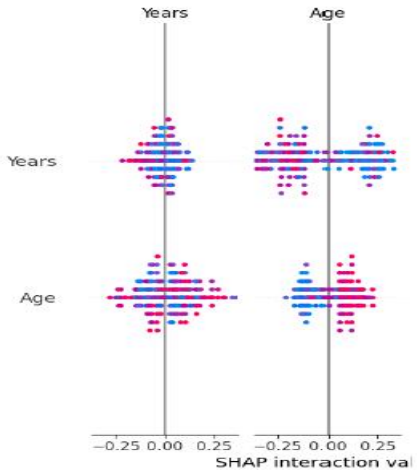

Fig.5. SHAP-based global feature importance of the Random Forest model

Figure 2 demonstrates that the suggested ML model does not rely on random correlations for prediction but instead makes use of therapeutically relevant information. With the use of the SHAP summary graphic, we can measure the significance and direction of effect of each characteristic, allowing us to optimise treatment or intensify surveillance for high-risk patients.

Integrating SHAP Findings into Machine Learning-Based Intervention Strategies: SHAP's interpretability analysis yields valuable insights for personalised intervention strategies in bronchiectasis management. Older patients with long-term diseases need more proactive treatment and follow-ups to prevent hospitalisations and exacerbations due to the strong influence of age on illness duration. The major role that chronic cough plays calls for symptomatic management of patients with persistent cough, with a focus on early airway clearing techniques. Since etiology-related aspects are important, intervention tactics should be altered according to the illness's origin. Patients with bronchiectasis associated to post-tuberculosis or chronic obstructive pulmonary disease (COPD) may need more intense and coordinated treatment lines. It is also possible to utilise the model as a clinical decision support tool to find patients who would do better with early disease treatment or other supportive measures. This is due to the fact that SHAP is greatly affected by oxygen treatment and physiotherapy. With the proposed framework, clinicians can move away from reactive care and toward anticipatory, evidence-based intervention planning, optimisation of therapy allocation, intensity of follow-up, and resource expenditure for bronchiectasis patients. This is achieved by combining predictive risk stratification with explainable insights.

Patient Phenotyping via Unsupervised Clustering

Unsupervised K-means clustering, on the other hand, uncovered three distinct patient phenotypes, illuminating previously unknown heterogeneity in the bronchiectasis community. Three bronchiectasis phenotypes have been identified by unsupervised grouping. A patient with phenotypic 0 would be 65 and above, have a moderate exacerbation burden, and be more likely to need hospitalisation. In phenotype 1, the clinical load is smaller, the sickness lasts less, and the age distribution is younger. Symptoms of phenotype 2, a frequent-exacerbator phenotype, include a high incidence of

exacerbations, advanced age, and a prolonged disease duration. Based on these findings, a personalised therapeutic program and risk stratification based on phenotypes are appropriate approaches to treat bronchiectasis.

A more conservative approach can be taken for the lower-risk phenotypes, while a more intensive monitoring, physiotherapy, and pro-antibiotic strategy may be useful for the high-exacerbation-linked subgroup. This way, these phenotypes can be used to create a practical intervention.

Table 3: Patient Phenotype Characteristics Identified Using K-Means Clustering

Phenotype	Sample Size	Mean Age (years)	Mean Disease Duration (years)	Mean Exacer-bations
Phenotype 0	487	67.58	7.69	1.11
Phenotype 1	347	32.43	4.29	1.02
Phenotype 2	166	65.57	8.49	2.48

Using unsupervised clustering, three phenotypes of bronchiectasis patients were found, and their clinical features are shown in Table III. The illness load is largest among those with phenotype 2, who are frequent exacerbators.

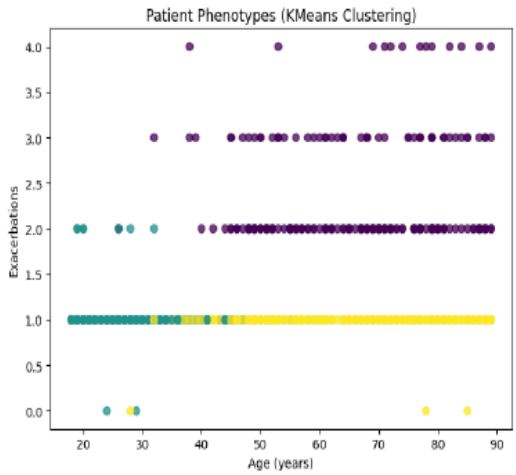

Fig.6 Patient Phenotypes Identified Using K-Means Clustering

The values obtained quantitatively corroborate the visual differentiation as shown in Figure 6, and Table III summarises the quantitative aspects of the observed phenotypes. The biggest subgroup, Phenotype 0 (n=487), has a moderate illness burden, an average annual hospitalisation of 1.69, and a mean exacerbation rate of 1.11. The average age of this grouping is 67.58 years.

With a mean exacerbation rate of 1.02 and decreased number of hospitalisations (1.25), phenotype 1 patients (n = 347) are the youngest with a mean age of 32.43 years. They have the shortest disease period (4.29 years) and the lowest clinical burden. In contrast, phenotype 2 (n = 166), which is linked with a rather stable bronchiectasis phenotype, shows the worst clinical picture, the longest duration of disease (8.49 years), and the largest number of exacerbations (2.48). The patients in this cluster tend to be older (65.57 years).

As a clinically relevant frequent-exacerbator phenotype, Phenotype 2 warrants closer monitoring and, more precisely, therapeutic measures that are targeted to its very high exacerbation load, despite the fact that its hospitalisation rate is comparable to Phenotype 0. These numerical discrepancies provide credence to the theory that clustering has successfully identified subgroups of bronchiectasis patients with distinct disease severity patterns, which may be useful in clinical practice.

By producing patients with distinct patterns of exacerbation and illness duration, clustering results show that unsupervised learning may effectively uncover the hidden structure of bronchiectasis data. The importance of data-based phenotyping for risk reduction and intervention prediction is shown by the high frequency of the frequent-exacerbator group.

Section H. Clinical Consequences

A complete machine learning-based intervention strategy for bronchiectasis care may be achieved by integrating explainable AI, predictive modelling, and patient phenotyping. Clinicians will be able to make better use of healthcare resources and optimise treatment methods by using individualised risk projections, insights into interpretability, and stratification by phenotype rather than relying on

generic recommendations. Logistic Regression had the greatest cross-validation performance in this investigation, suggesting that linear combinations of routinely recorded clinical features might best describe the risk of bronchiectasis. The Random Forest model provided clinically acceptable accuracy while also being more interpretable using SHAP-based explanation, which supports the principle of openness in decision-making. Although they were competitive, neural network models were not significantly better than very basic approaches. An further finding from unsupervised phenotyping was the identification of a subset of patients who were regular exacerbators, which has important therapeutic implications. The results show that the framework is a reliable and useful tool for helping people with bronchiectasis make better decisions about their respiratory health.

V. CONCLUSION

The results of this study demonstrate that bronchiectasis management strategies based on machine learning intervention planning are practical and productive. A comprehensive and therapeutically relevant decision-support system is created by integrating patient phenotyping with explainable AI and robust predictive analytics. Improved respiratory outcomes and individualised treatment for bronchiectasis are possible thanks to the method described in this article. The research has some limitations, such as its retrospective character and the absence of longitudinal follow-up data. Incorporating time-series data, validating the framework on multi-center cohorts, and expanding the system to real-time clinical decision assistance will form the foundation of future work.

REFERENCES

1. Chalmers, J. D., & Chotirmall, S. H. (2018). Bronchiectasis: new therapies and new perspectives. The lancet Respiratory medicine, 6(9), 715-726.
2. Leigh, L. Y., Vannelli, P., Crow, H. C., Desai, S., Lepore, M., Anolik, R., & Glick, M. (2021). Diseases of the respiratory tract. Burket's Oral Medicine, 469-504.
3. Osamika, D., Adelusi, B. S., Kelvin-Agwu, M. C., Mustapha, A. Y., & Ikhalea, N. (2023). Predictive analytics for chronic respiratory diseases using big data: Opportunities and challenges. International Journal of Multidisciplinary Research and Growth Evaluation, 6(2), 445-462.
4. Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. Med, 2(6), 642-665.
5. Chalmers, J. D., Aliberti, S., Filonenko, A., Shteinberg, M., Goeminne, P. C., Hill, A. T., ... & McDonnell, M. J. (2018). Characterization of the "frequent exacerbator phenotype" in bronchiectasis. American journal of respiratory and critical care medicine, 197(11), 1410-1420.
6. Finch, S., McDonnell, M. J., Abo-Leyah, H., Aliberti, S., & Chalmers, J. D. (2015). A comprehensive analysis of the impact of Pseudomonas aeruginosa colonization on prognosis in adult bronchiectasis. Annals of the American Thoracic Society, 12(11), 1602-1611.
7. Aliberti, S., Lonni, S., Dore, S., McDonnell, M. J., Goeminne, P. C., Dimakou, K., ... & Chalmers, J. D. (2016). Clinical phenotypes in adult patients with bronchiectasis. European Respiratory Journal, 47(4), 1113-1122.
8. Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930.
9. Chandrasekaran, R., Mac Aogáin, M., Chalmers, J. D., Elborn, S. J., & Chotirmall, S. H. (2018). Geographic variation in the aetiology, epidemiology and microbiology of bronchiectasis. BMC pulmonary medicine, 18(1), 83.
10. Aliberti, S., Lonni, S., Dore, S., McDonnell, M. J., Goeminne, P. C., Dimakou, K., ... & Chalmers, J. D. (2016). Clinical phenotypes in adult patients with bronchiectasis. European Respiratory Journal, 47(4), 1113-1122.
11. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67
12. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.

13. Molnar, C. (2020). Interpretable machine learning. Lulu. com.
14. Hou, X., Zhishan, D., Zhou, Y., & Ran, P. (2025). Data-Driven Clustering Approach to Identify Novel Phenotypes in Acute Exacerbations of Chronic Obstructive Pulmonary Disease: A Retrospective, Multicentre Cohort Study. American Journal of Respiratory and Critical Care Medicine, 211(Abstracts), A6156-A6156
15. De Angelis, A., Johnson, E. D., Sutharsan, S., & Aliberti, S. (2024). Exacerbations of bronchiectasis. European Respiratory Review, 33(173).
16. P. Ramesh Babu, "Machine Learning Techniques in Lung Nodule Diagnosis of Medical Health Care Data," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 11, pp. 1458–1463, Sep. 2019.
17. P. Ramesh Babu, "The Robust Computer Aided Diagnostic System for Lung Nodule Diagnosis," International Journal of Recent Technology and Engineering (IJRTE), pp. 5670–5675.