OPEN ACCESS

International Journal of Innovations in Science & Technology

RESEARCH & INNOVATION DIVISION

IJIST

Machine Learning-Based Asthma Diagnosis Prediction Using Lung Function and Demographic Features

Hafiz Gulfam Umar[1], Gul Andam [1] and Urwa Bibi[1]

[1] Ghazi University

*** Correspondence:** gullandam8@gmail.com

Citation | Umar. H. G, Andam. G, Bibi. U, "Machine Learning-Based Asthma Diagnosis Prediction Using Lung Function and Demographic Features", IJIST, Vol. 07 Issue. 01 pp 440-461, Feb 2025

Received | Jan 25, 2025 **Revised |** Feb 22, 2025 **Accepted |** Feb 24, 2025 **Published |** Feb 25, 2025.

Asthma is a prevalent chronic respiratory disease, which poses significant diagnostic challenges because of its multifactorial nature. This study aims to develop a machine-learning approach for predicting asthma diagnosis using key features such as body mass index (BMI), age, lung function parameters (FEV1 and FVC), and demographic information. A dataset containing clinical and demographic records was utilized to train and evaluate models, including Random Forest, Neural Networks, and XGBoost classifiers. The performance of the following models was assessed using metrics such as precision, recall, accuracy, and F1-score, with Random Forest exhibiting the highest predictive performance. In addition to traditional performance metrics, advanced visualization techniques like SHAP (Shapley Additive ex Planation's) values were employed to interpret model predictions and assess feature importance. Results demonstrate that age, BMI, and lung function are key predictors of asthma diagnosis, with lung function parameters showing the strongest correlation with diagnosis outcomes. The study also explores various 3D and interactive visualizations to enhance the interpretability of the models. The proposed approach demonstrates that machine learning models when combined with clinical data, can accurately predict asthma diagnosis and potentially aid healthcare professionals in early detection and personalized treatment plans. This research highlights the potential of data-driven models in improving asthma diagnosis and contributing to better clinical decision-making.

Keywords: Asthma diagnosis, Machine learning models, FVC (Forced Vital Capacity), Lung function parameters, FEV1 (Forced Expiratory Volume in one second)

Introduction:

Asthma is an inflammatory condition of the airways that is affecting millions of people worldwide, leading to significant morbidity and reduced quality of life. Its symptoms, which include shortness of breath, wheezing, and chest tightness, vary greatly among individuals and across different age groups. Early and accurate diagnosis of asthma is critical to prevent exacerbations and manage the disease effectively. However, diagnosing asthma can be challenging due to overlapping symptoms with other respiratory conditions and the heterogeneity of its clinical presentation. In recent years, the integration of different machine-learning techniques into medical research has shown great potential in improving the accuracy and speed of diagnostic processes. Machine learning algorithms identify hidden patterns, process vast amounts of patient data, and make predictions with high precision. These models, when applied to clinical and physiological data, can assist in diagnosing asthma more accurately than traditional methods. By leveraging advanced models such as XGBoost, Neural Networks, and Random Forest, it is possible to automate the diagnosis process while providing insights into the importance of different diagnostic factors. The study focuses on using machine learning models to predict asthma diagnosis based on key clinical features, including lung function parameters like Forced Vital Capacity (FVC) and forced expiratory volume in one second (FEV1), along with demographic characteristics like age and Body Mass Index (BMI). These features are known to influence lung function and asthma severity, making them crucial predictors of the disease. In addition to building accurate models, this research highlights the importance of explainability and interpretability in machine learning. Tools like SHAP (Shapley Additive explanations) are employed to understand the contributions of individual features to the model's predictions, offering insights that can support clinicians in understanding the rationale behind predictions. Furthermore, advanced data visualization techniques, including 3D scatter plots and feature importance heatmaps, are applied to explore the relationships between many clinical features and asthma diagnosis. This paper is organized as follows: Section 2 reviews the dataset and features used for prediction. Section 3 presents the methodology, including the machine learning models implemented. Section 4 covers the results and analysis of the models' performance, in Section 5 there is a conclusion and future work.

Related Work:

To improve predictions of asthma self-management, the focus of the study was machine learning methods to create early warning algorithms using the Asthma Mobile Health Study (AMHS), a publicly available mHealth dataset. To distinguish between stable and unstable periods, we used some popular supervised learning algorithms (classification). We discovered that both naïve Bayes-based classifiers and logistic regression are providing accuracy (AUC > 0. 87) [1]. Asthma does not currently have a cure. Nonetheless, the illness can be managed with current management techniques, such as the use of "preventer" inhalers. An action plan and assisted self-management greatly lower the chance of an asthma attack [2]. This study tests many machine-learning approaches to develop a model for predicting asthma attacks. The bio signals dataset and the online environmental dataset are the two primary components of the dataset that was employed. The study employed a variety of machine-learning approaches, including random forests, support vector machines, logistic regression, gradient boosting models, and decision trees [3].

A different study showed how a model's performance can be enhanced by applying different predictors from other sources. Nearly 37 attributes were selected, including environmental triggers and multiple bio signals. Pattern-Based Decision Trees (PBDT) and pattern-based class-association rules (PBCAR) were the machine learning techniques that were used. The model's accuracy in predicting asthma attacks was 0.87. With 87% and 86% accuracy rates, respectively, PBDT outperformed PBCAR by a small margin. Another model for predicting asthma attacks exists. In addition to environmental triggers, they gathered a dataset that contained bio signal data of three subjects. The model gives an accuracy of 93% when the SVM algorithm is applied. However, this model was just trained and tested only on three individuals, so, this was unclear how it would function if tested on a larger population [4]. Asthma is a chronic, fluctuating illness. Since there is currently no treatment for asthma, long-term maintenance is the main focus. Although mobile health (mHealth) holds promise for managing chronic diseases, it must do more than just monitor patients. Therefore, mHealth must use machine learning to deliver customized algorithms and feedback. It's important to comprehend how these machine learning algorithms have been used in the view of mHealth to control asthma [5]. Monitoring breathing and identifying respiratory problems may aid in the early detection of asthma attacks. Portable sleep diagnostic devices to track breathing are among the tools that have been suggested for home monitoring [6].

One of the main signs of an asthma episode is a decline in Peak Expiratory Flow (PEF). Patients occasionally utilize peak flow meters at home to obtain objective values and determine whether any intervention is necessary. Peak flow meters are not as detailed as spirometers, which are another tool for measuring lung function [7]. This study's main goal is to examine the predictive and variable importance evaluation capabilities of machine learning systems. Several performance metrics, such as precision, accuracy, Kappa statistics, recall, F-measure, AUC value. And ROC curve has been taken into account in a thorough comparison. The best algorithms, according to the findings, are C5.0 for asthma, SVM (with non-linear kernel), Random Forest (with CART learner), and GBM for COPD. However, MEF50 for COPD and FEV3 for asthma are the most crucial variables. Among the top five variables, FEV1 and FVC have been prevalent. Tests of statistical significance have validated the variables' rank [8]. In a different study, a machine learning-based approach to early exacerbation detection and subsequent triage was provided. In this application, supervised algorithms were trained using the opinions of physicians in a clinically and statistically thorough sample of patient data. The model's accuracy was assessed using a representative patient validation set and a physician panel that initiates the same cases [9]. In this study, we use cutting-edge machine-learning techniques to suggest a sustainable method of diagnosing asthma. More precisely, we employ the extreme gradient boosting algorithm for classification, feature selection was used to identify the data augmentation that increased the dataset's durability and significant characteristics.

In this suggested method, data augmentation entails creating artificial samples that expand the training dataset, which is subsequently used to improve training data first. That may mitigate this issue of asthma-related data imbalance. The extreme gradient boosting technique is then applied to select important features and increase diagnosis accuracy [10]. In addition to, paroxysmal, intermittent, persistent, or persistent evidence plus acute attacks, asthma symptoms can also be categorized by the duration, frequency, changes between night and day, and nighttime symptoms. Asthma symptoms can also be categorized as seasonal attacks, annual attacks, or a combination of both. It is still unclear how genetic, environmental, viral, and dietary variables contribute to asthma [11]. Another study presented the Adaptive Fuzzy Inference System (ANFIS), a backpropagation technique that lowers asthma diagnostic mistakes [12]. When the asthma dataset is imbalanced, traditional algorithms usually split minority classes into the majority classes, achieving a greater accuracy rate; nevertheless, that makes it challenging for this algorithm to categorize significant minority classes [13]. There is often an imbalance issue in the data on asthma symptoms. It has recently been demonstrated that Generative Adversarial Networks (GANs) offer fresh approaches to data augmentation for the imbalance issue [14] .In the present investigation, we introduce a machine learning-based algorithm designed to forecast the risk of asthma. Leveraging Internet-of-Things resources, this technology was comprehensively implemented on a mobile phone as a smartphone health application. Peak Expiratory Flow Rates (PEFR), which are widely recognized as significant asthma risk determinants, are conventionally assessed utilizing external devices such as peak flow meters. The results of this PEFR were classified into three different risk categories: "Yellow" (Moderate Risk), "Green" (Safe), and "Red" (High Risk), and are compared against the optimal peak flow measurement attained by all individuals [15]. The primary goal of this research was to identify user requirements and explore methodologies for integrating electronic components into inhalers to address these needs and enhance both the provision of services and the user experience. Consequently, the electronic systems incorporated within the inhaler will be controlled by a mobile application.

This application will provide notifications regarding the appropriate times for usage as well as the geographical location of the inhaler [16]. In the proposed system, we minimized the reliance on the inhaler, thereby facilitating the prevention of asthma exacerbations. In this research endeavor, a cloud-based framework that adeptly integrates multimodal data along with a user-friendly web interface was established. The apparatus is designed for portability and is equipped with a sensor for temperature detection. Through the utilization of this asthma management pack, each suffering from asthma has effectively safeguarded their well-being against environmental conditions. The sensors employed in this study quantify various outputs on the respiratory spectrum, which are contingent upon lung volume and lung capacity [16].

All these studies used only lung health data while we are using some demographic features with it to improve early prediction of asthma.

Material and Methods:

This section highlights the methodology that is used to develop, train, and check the machine learning model's performance for predicting asthma diagnosis. This methodology comprises several key stages, which include data preprocessing, feature selection, model implementation, and performance evaluation. That is shown in Figure 1 below.

Dataset Description:

The dataset for this study was taken from Kaggle which consists of clinical and demographic data from patients, with a specific focus on predicting asthma diagnosis. The dataset consists of two target classes (0 = No Asthma, 1 = Asthma). This dataset consists of features like age, BMI, lung function parameters (e.g., LungFunctionFEV1, Lung Function FVC), and several categorical indicators for symptoms (e.g., dry cough, difficulty in breathing). Key features include:

- **Age:** (in years)
- **BMI:** (Body Mass Index)
- **Lung Function FEV1:** (Forced Expiratory Volume in one second)
- **Lung Function FVC:** (Forced Vital Capacity)
- **Diagnosis:** (Target variable: 0 = No Asthma, 1 = Asthma)

The dataset is highly imbalanced, with fewer instances classified as (1 = Asthma) and a greater number classified as (0 = No Asthma). In addition, other demographic features and medical history, such as gender and medications, were available in the dataset but were not directly used in this analysis, as the focus was on age, BMI, and lung function parameters.

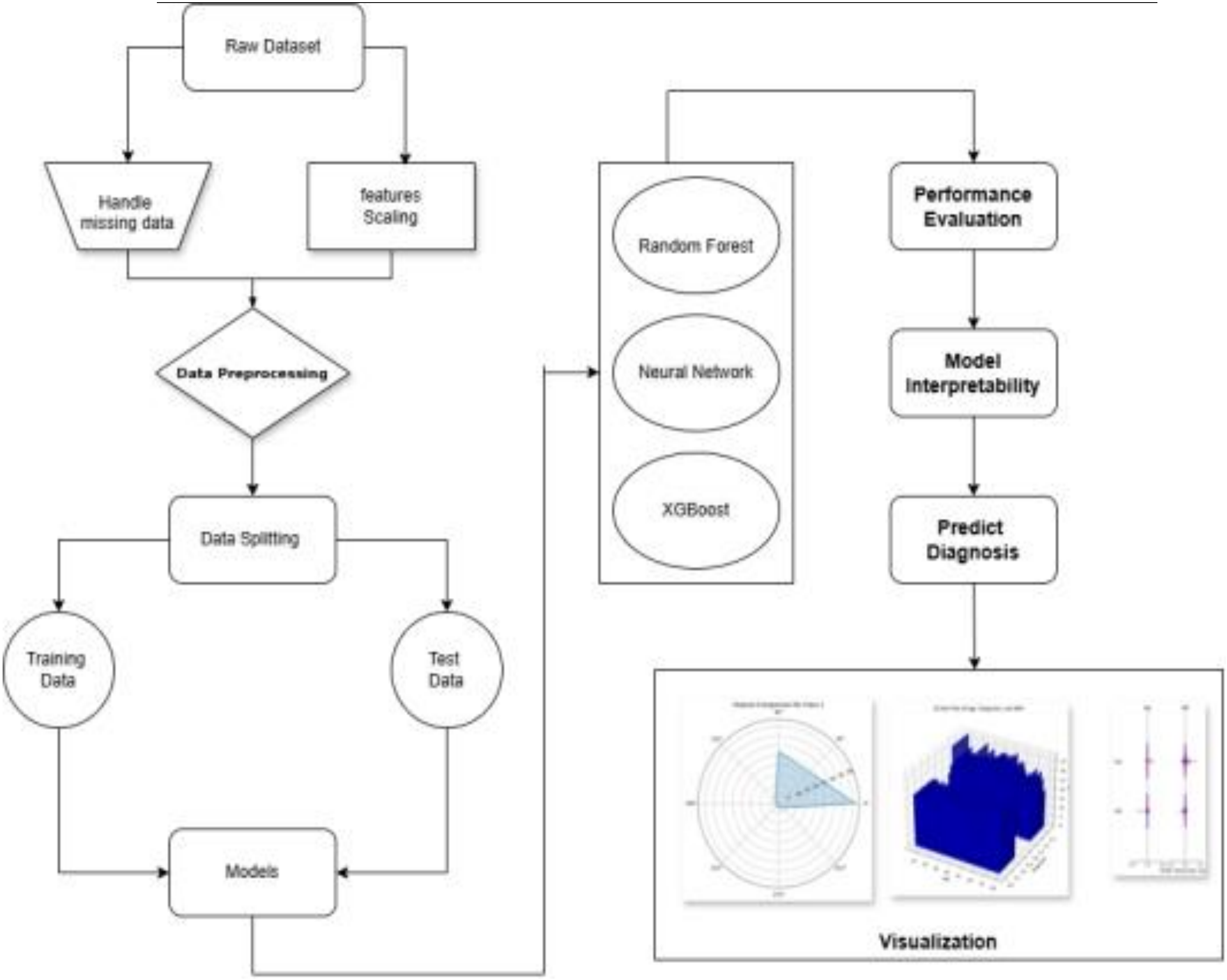

Figure 1 Block Diagram

Data Preprocessing:

This was a critical step that ensured the models were trained on clean and standardized data. The following steps were performed here:

- **Handling Missing Data**: By using imputation techniques problem of Missing values was solved, such as replacing missing lung function values with the median or mean values of the given feature.
- **Feature Scaling**: Continuous variables like age, BMI, FEV1, and FVC were standardized to have a standard deviation of one and a mean of zero. This step was compulsory for models like Neural Networks that were sensitive to feature scaling.
- **Data Splitting**: This dataset was split into testing (20%) and training (80%) sets. This training set was used to develop models, while the test dataset was used to check their performance. All this workflow is shown in figure 2, which is given below.

Feature Selection:

The features selected for model training were chosen based on their relevance to asthma diagnosis and lung function assessment:

- **Age**: Age is a significant factor influencing lung function and the likelihood of asthma.

- **BMI**: Obesity is associated with worse asthma outcomes and is included as a predictor.
- **Lung FunctionFEV1**: FEV1 is a critical measure of airway obstruction.
- **LungFunctionFVC**: FVC helps assess the volume of air a patient can forcibly exhale.

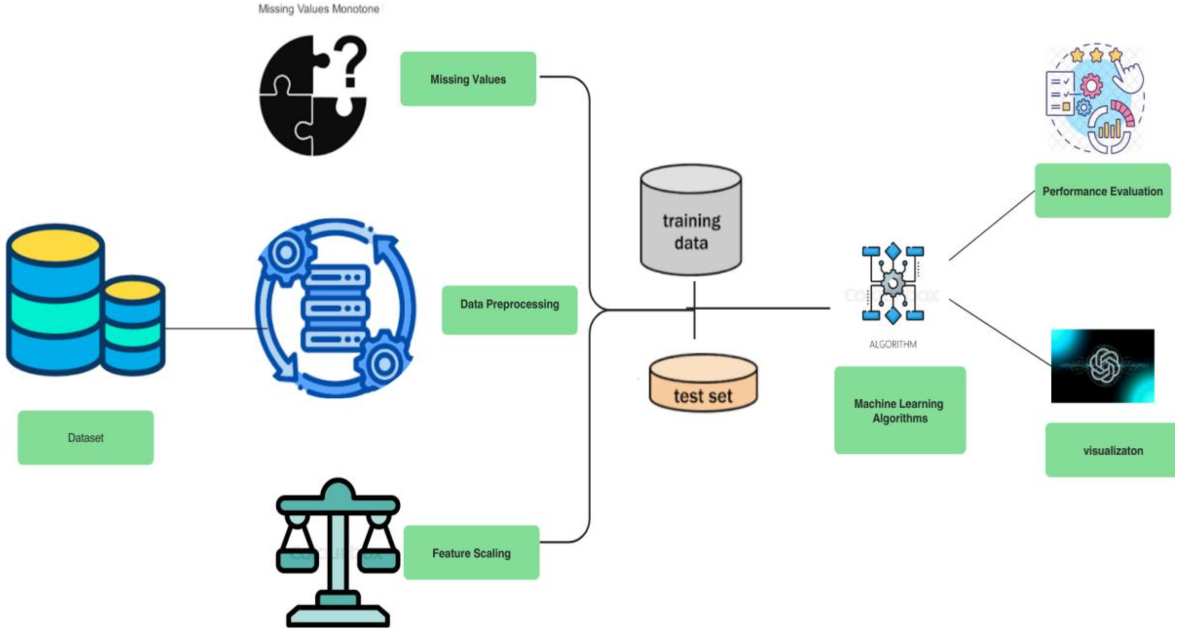

Figure 2. Workflow

Radar Charts for Both Classes:

The radar charts in figure 3 for **Class 1** and **Class 0** represent the feature profiles for every class based on a dataset involving features like Age, BMI, Lung Function FEV1, Lung Function FVC, etc.

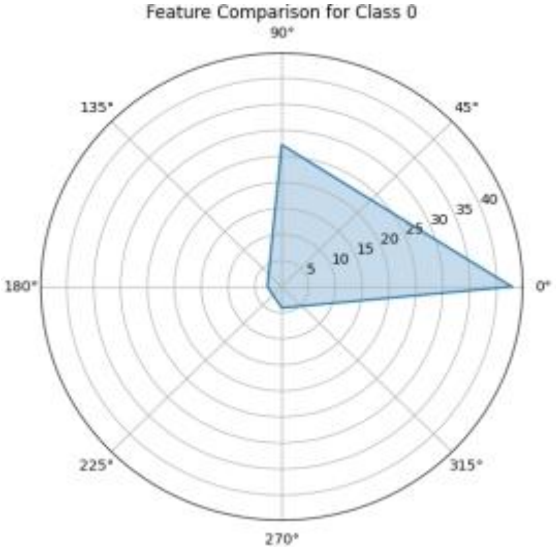

Figure 3. Feature Comparison for Class “O

Radar charts help in comparing multiple features simultaneously across different classes, making it easier to spot feature-level differences. The variations in patterns and lengths along each axis suggest key differences in how features are distributed for Class 0 versus Class 1**.** Data-driven decisions can be made based on this visualized comparison, such as identifying which features are more relevant for separating or distinguishing the two classes.

Table1. Differences in Radar Chart

Feature	**Class 0** (Radar Chart)	**Class 1** (Radar Chart)	**Comparison/Observation**
Age	Moderate to high values	Moderate values	Class 0 shows a broader range, indicating a higher average.
BMI	High values	Lower to moderate values	Class 0 has a higher BMI compared to Class 1.
Lung Function FEV1	Moderate to high distribution	Low to moderate distribution	Class 0 has higher average lung function values.
Lung Function FVC	Higher values	Lower values	FVC values are higher for Class 0 compared to Class 1.

Table 1 summarizes the differences observed in the radar charts for **Class 1** and **Class 0** based on the features in the dataset:

Machine Learning Models:

Many machine learning models which are shown in figure 4 were implemented to predict asthma diagnosis:

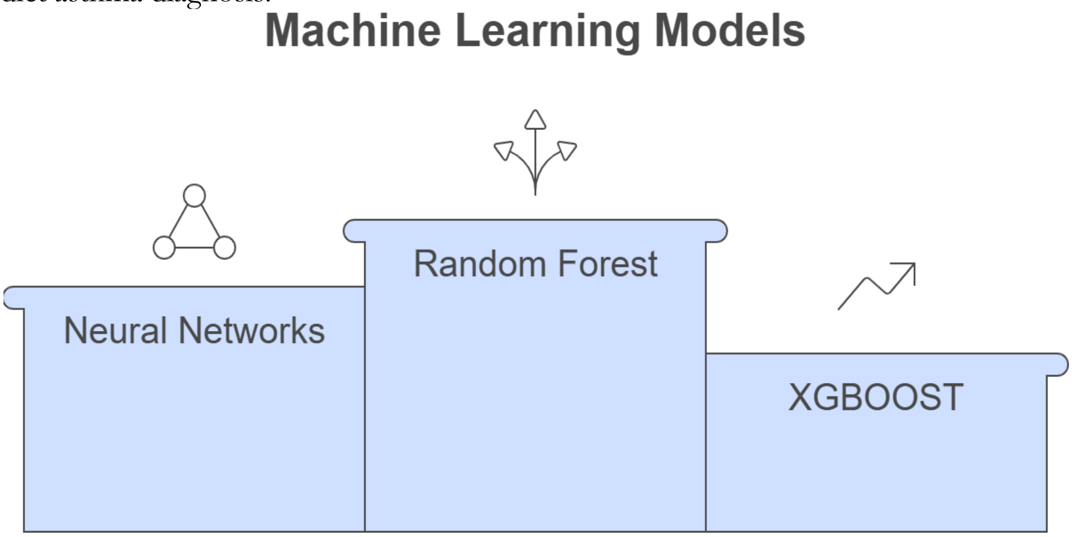

Figure 4. Machine Learning Model

Random Forest Classifier: The tree-based model for regression and classification is called RF. Using the majority voting criterion, the ensemble model aggregated the number of decision trees [17]. By enhancing the model's accuracy and stability, the Random Forest technique contributes to better generalization [18]. A robust, ensemble-based model that uses various decision trees which make predictions. Random Forest is particularly used for handling complex interactions between features and is resistant to overfitting.

The mathematical model can be written as:

$$y^{\wedge} = f(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

Here:

- T gives the total number of trees.
- Ft (x) predicts the t-th decision tree.
- The final prediction y^ is obtained by taking the majority vote or averaging the predictions of all decision trees.

Neural Networks (Multilayer Perceptron): A deep learning model that was employed to capture non-linear relationships in the data. This architecture consists of an input layer that has four nodes (one for each feature), a hidden layer, and an output layer for binary classification (asthma or no asthma). The rectified linear unit (ReLU) function was used for hidden layers, here output layer used the sigmoid function.

$$\mathbf{y}^{\wedge} = \boldsymbol{\sigma}\left(\mathbf{W}^{(\mathbf{L})} . \mathbf{h}^{(\mathbf{L}-\mathbf{1})} + \mathbf{b}^{(\mathbf{L})}\right)$$

Here:

- h(l)= $\sigma\left(W^{(l)} . h^{(l-1)} + b^{(l)}\right)$ it is the activation function that is used at layer l.
- $W^{(l)}$, $b^{(l)}$ these are the biases and weights at layer l.
- σ is an activation function (commonly sigmoid for binary classification).
- L is the number of layers in the network.

This model learns the parameters W and b by reducing a loss function (e.g., cross-entropy loss).

XGBoost:

One well-liked tree learning algorithm is XGBoost. The foundation of the system is second-order Taylor expansion of the objective function, that XGBoost employs to evaluate model validity. By referring to this central part of the system, it matches the residuals of the previous forecast. It repeatedly splits a new tree to match the residuals of the previous forecast [19] It reduce s overfitting and is based on the DT method. It stands for eXtreme Gradient Boosting [20].

An advanced gradient boosting algorithm that builds decision trees sequentially to minimize prediction error. XGBoost is highly effective for structured data and is known for its superior performance and speed.

$$\mathbf{y}^{\wedge} = \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}} \boldsymbol{\alpha}_{\mathbf{k}}\, \mathbf{f}_{\mathbf{k}}(\mathbf{x})$$

Where:

- k represents boosted tree numbers.
- $f_k(x)$ represent k^th decision tree.
- α_k is a weight for the k^th tree.
- The model is trained by reducing a loss function (e.g., cross-entropy loss) and adding a new tree that corrects the errors of the previous trees.

Mathematical Model for Asthma Diagnosis Using Machine Learning:

In this problem, we are using many machine learning models (Neural Network, Random Forest, and XGBoost) to predict asthma diagnosis (binary classification: Class 0 = No Asthma, Class 1 = Asthma) based on several features like lung function (FEV1, FVC), BMI, and Age. The general mathematical formulations for the machine learning models and performance metrics used are as follows:

General Machine Learning Formulation:

Given a dataset D={(xi,yi)}, here:

- $x_i=[x_{i1},x_{i2},\ldots,x_{im}]$ it represents the feature vector of the $_i$-th sample, with m features (e.g., Age, BMI, LungFunctionFEV1, LungFunctionFVC).
- $y_i \in$ {0,1} is the binary class label for asthma diagnosis (0 = No Asthma, 1 = Asthma).

The aim is to find a function f(x_i) which maps input x_i to output $y^{\wedge}{}_{I}$=f(x_i) such that $y^{\wedge}{}_{I} \approx y$

Model Training:

The models were trained on the preprocessed dataset. For each model, the following procedure was used:

Hyper-parameter Tuning: Cross-validation and Grid search were used to perform an optimized hyper-parameter for each model. For Random Forest, the number of trees and the maximum depth were tuned. For Neural Networks, some hidden layers and nodes per layer were adjusted, and for XGBoost, learning rate, max depth, and number of boosting rounds were optimized.

Cross-Validation: To mitigate over-fitting, cross-validation k-fold (k=5) was used during model training. This technique ensures that the models generalize well to unseen data by training on multiple subsets of the training data.

Performance Evaluation:

After training, the models were evaluated on the test set using various classification metrics:

Accuracy: It measures the proportion of correct predictions over total predictions. Figure 5 shows the accuracy of all models.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

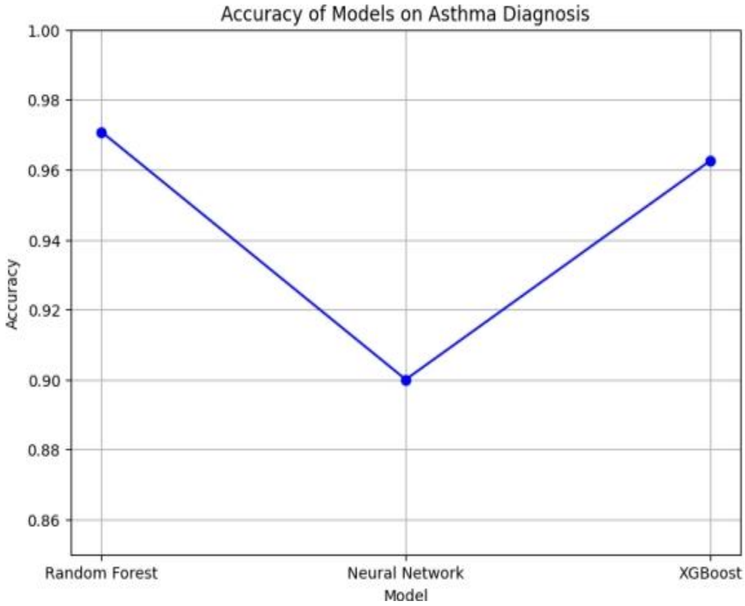

Figure 5. Models Accuracy

Precision: It was determined by adding all of the given model's predictions. Next, the proportion of accurate forecasts is divided by the given number of predictions [21]. Indicates the proportion of true positive asthma diagnoses from all predicted positives.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

Recall: Recall, sometimes referred to as the sensitivity or true positive rate, is a second important statistic [22].Assesses the ability of the model that correctly identify patients with asthma.

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

F1-Score: Precision and recall are averaged to determine the F1 score. Generally speaking, F1-Score is regarded as a trustworthy way that compare how well different classifiers perform, significantly when data is uneven [23]. The harmonic mean of the precision and the recall provides a balanced measure of this model's performance.

$$\mathbf{F1 - Score} = 2 \times \frac{\textit{Percision} \times \textit{Recall}}{\textit{Percision} + \textit{Recall}}$$

Area Under the ROC Curve (AUC-ROC): ROC curves in figure 6 are employed to assess how well classification algorithms perform. Plotting the True Positive Rate (TPR), which is known as recall, against the False Positive Rate (FPR) at different threshold values is what the curve does [24]. Providing the balance between specificity and sensitivity by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds.

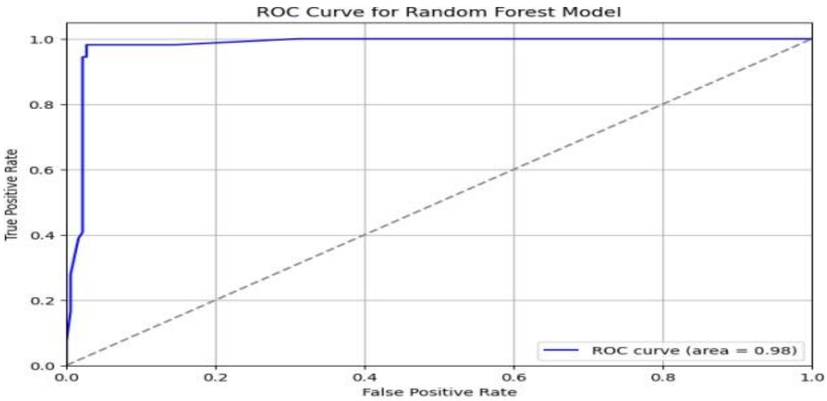

Figure 6. ROC Curve

- The area under the ROC curve is 0.98, which indicates excellent performance.
- An AUC of 0.98 means that the model will differentiate a positive instance from a negative instance correctly.

The Random Forest model demonstrates outstanding predictive power, as indicated by the high AUC of 0.98, meaning it's very effective at differentiating between two classes in the dataset.

Model Interpretability:

To evaluate the performance of the model, interpretability was a key focus of this study. SHAP (Shapley Additive explanations) values were employed to tell model predictions by quantifying the portion of each feature towards the prediction. This step helps to provide insights into how specific features, such as age, BMI, and lung function, influence the likelihood of asthma disease. For evaluation, we employed a confusion matrix, which is composed of TP (true positive), FP (false positive), TN (true negative), and FN (false negative). [25]

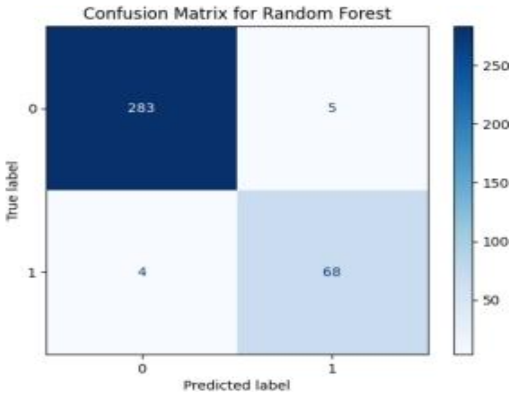

Figure 7. Confusion Matrix of RF

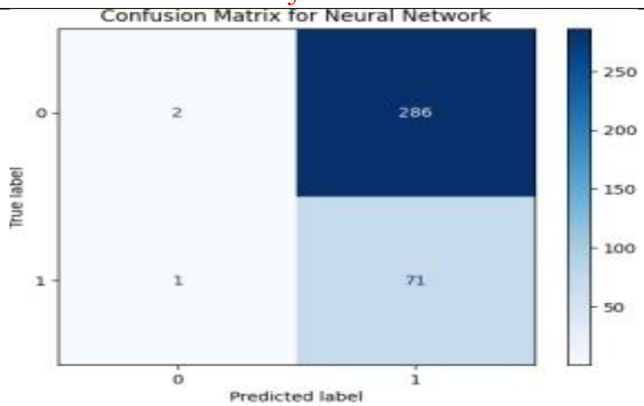

Figure 8. Confusion Matrix of NN

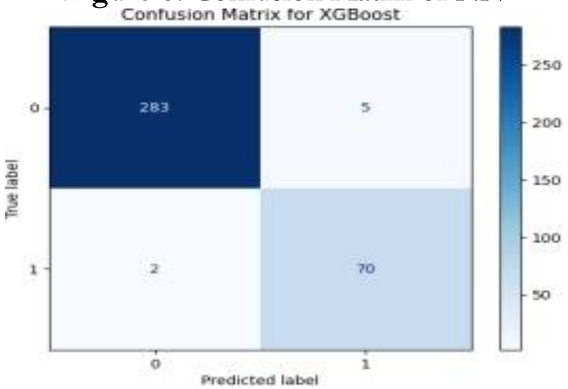

Figure 9. Confusion Matrix of XGBOOST

Table 2 Metric of All Models

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)	Model
Random Forest	283	5	4	68	Random Forest
Neural Network	2	286	1	71	Neural Network
XGBoost	283	5	2	70	XGBoost

Table 2 summarizes the performance of all three models that are based on their confusion matrices:

Random Forest confusion matrix in figure 7 and XGBoost confusion matrix in figure 8 offer strong performance with balanced metrics, suitable for real-world applications requiring high accuracy and reliability in predictions. Neural Network confusion matrix in figure 9 demonstrates high sensitivity but at the cost of very low specificity, which makes it less practical in this context unless adjusted to reduce false positives.

Visualization Techniques:

We used advanced visualization techniques to find out the relationships between the target variable and features:

- **3D Scatter Plots**: Visualized the interaction between age, BMI, and lung function parameters. These show the pairwise relationships between two features in figure 10, with the color coding separating the diagnosis groups.
- The blue group (Diagnosis = 0) covers a wider range of values for both LungFunctionFEV1 and FVC. The orange group (Diagnosis = 1) tends to be clustered around lower lung function values, with some higher-age participants.

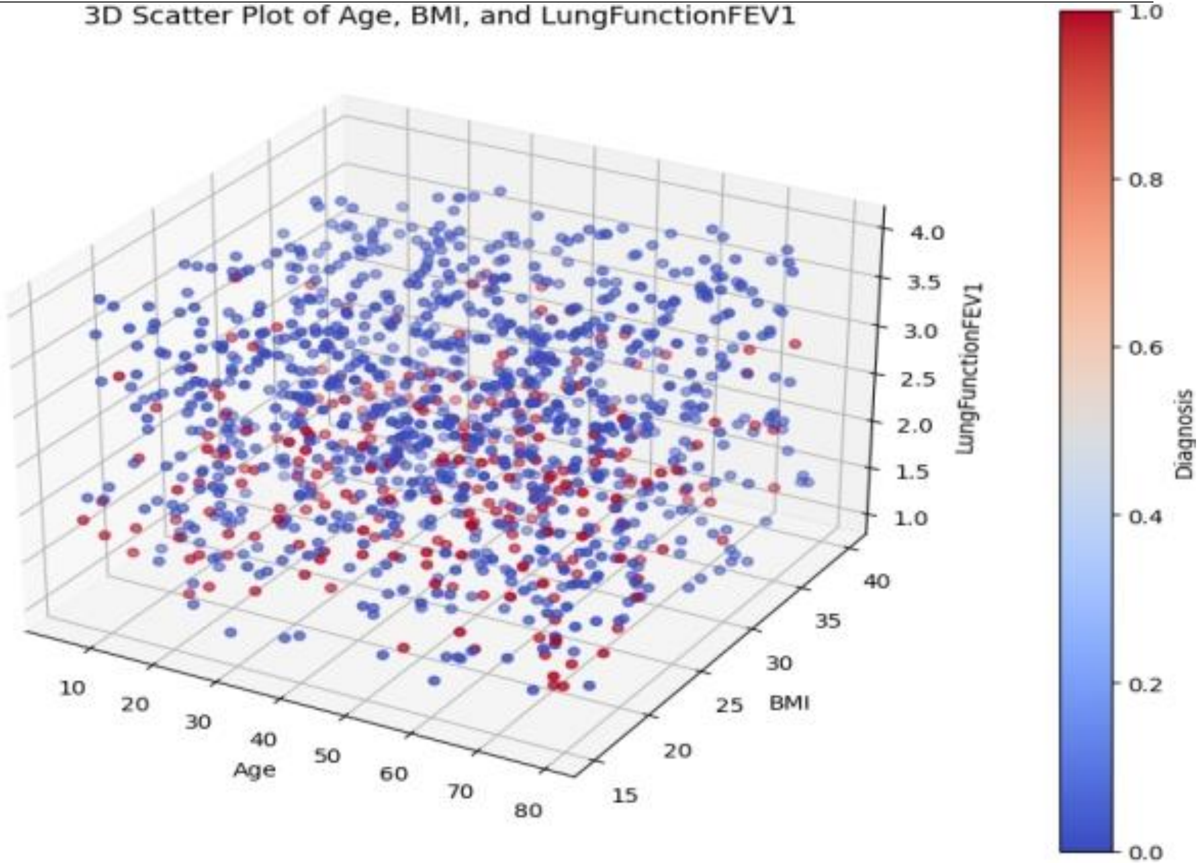

Figure 10. 3D Scatter Plot

- **SHAP Summary Plots**: SHAP Summary Plots in figure 11 showed the global importance of features in predicting asthma diagnosis across the dataset.

Each dot represents an interaction between the two features (Age and BMI). The color represents the value of one feature (blue for lower values, pink for higher values), while the x-axis shows how this interaction affects the model's prediction (positive or negative SHAP interaction values). The vertical lines show how the SHAP interaction value for one feature (on the y-axis) varies across the dataset.

This plot indicates that the interaction between Age and BMI doesn't have a large or consistent impact on the model's prediction, as the SHAP interaction values cluster close to zero.

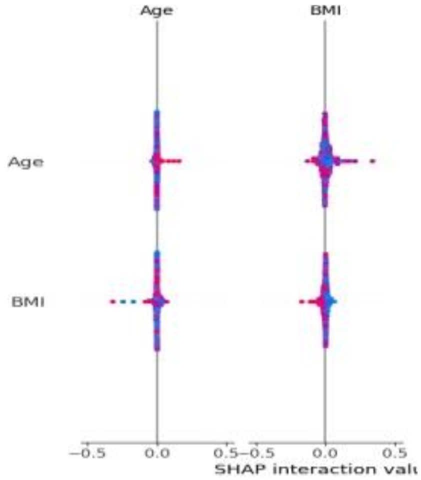

Figure 11. SHAP Plots

Correlation heatmap: From this heatmap in figure 12, we can conclude that there is very little linear dependency between Age, BMI, and lung function parameters (LungFunctionFEV1, LungFunctionFVC). These low correlations suggest that none of these features strongly influence each other in this dataset. Therefore, they could be considered independent factors affecting the prediction of diagnosis, or their combined effects may need more sophisticated modeling techniques like SHAP to fully understand their interaction and importance in the classification models.

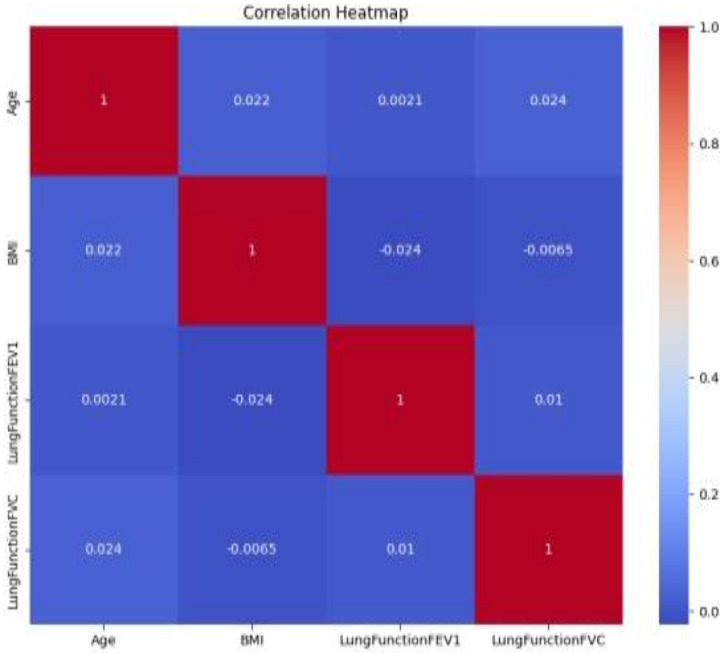

Figure 12. Correlation heatmap

Boxplot: Lung Function (FEV1) Distribution by Diagnosis:

Boxplot in figure 13 suggests that FEV1 is a distinguishing feature between the two groups, with lower FEV1 values often associated with asthmatic patients.

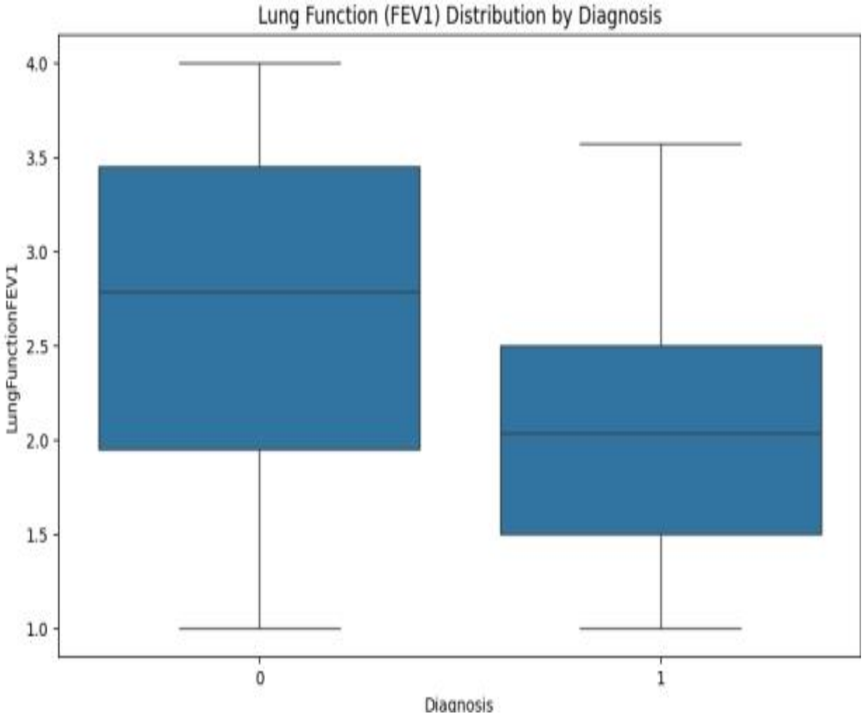

Figure 13. Boxplot of Lung Function (FEV1)

Boxplot: Lung Function (FVC) Distribution by Diagnosis:

Boxplot in figure 14 suggests that similar to FEV1, the lower FVC values in asthmatic patients indicate that these lung function parameters are vital for classifying asthma.

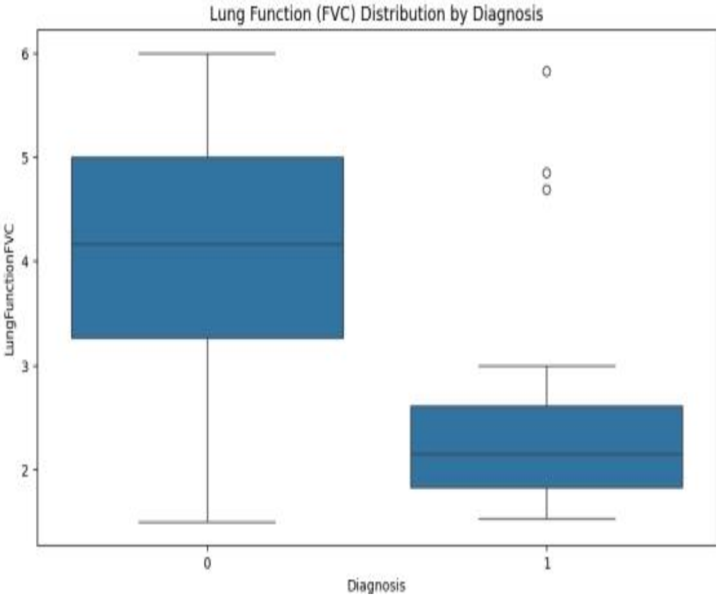

Figure 14. Boxplot of Lung Function (FVC)

These two plots suggest that lung function metrics, particularly FEV1 and FVC, are significantly lower for patients diagnosed with asthma compared to those without. These features contribute strongly to distinguishing between asthmatic and non-asthmatic patients.

Violin Plot:

A violin plot in figure 15 is a technique that is used for data visualization and also to show the distribution of a numerical variable for different categories of data, combining aspects of a box plot and kernel density plot. It helps to provide insight into the probability density of the dataset at different values, offering a more abstract view of data distribution as compared to a simple box plot.

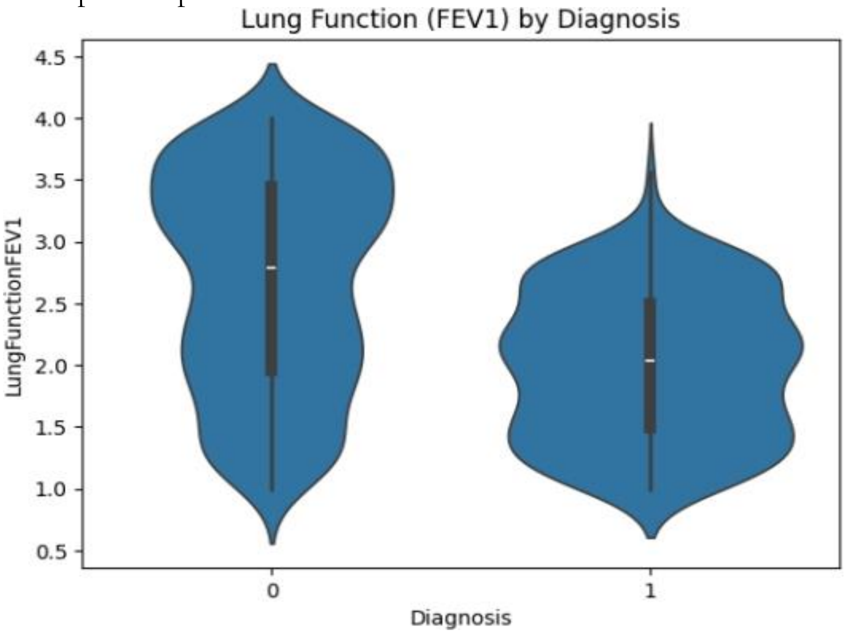

Figure 15. Violin Plot Lung Function (Fev1)

The violin plot in figure 16 shows that non-asthmatic patients tend to have higher and more variable FEV1 values, while asthmatic patients generally have lower FEV1 values with less variation. This reinforces that FEV1 is a significant feature in differentiating between asthmatic and non-asthmatic individuals.

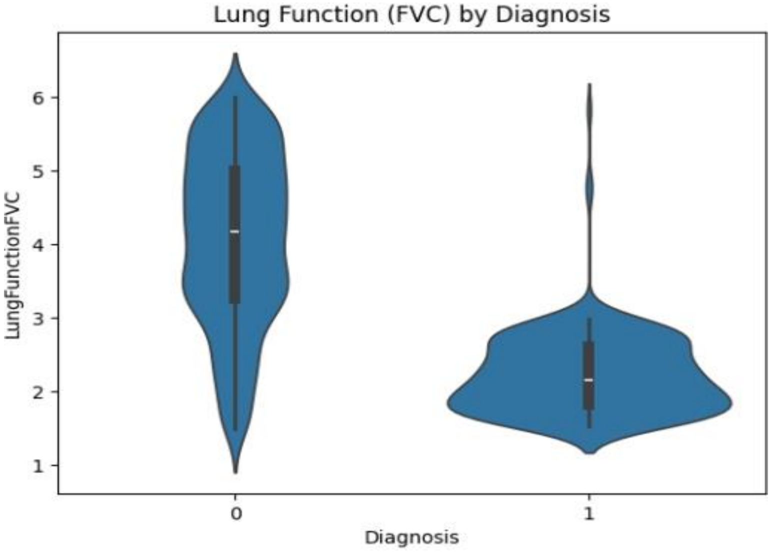

Figure 16. Violin Plot Lung Function (FVC)

This violin plot provides visualization of the distribution of FVC data for two different diagnosis categories (0 and 1). The plot shows that FVC is generally higher for individuals with a diagnosis of 0 (negative) and lower for those with a 1 (positive) diagnosis. This visualization emphasizes that lung function may be significantly impacted for individuals diagnosed positively, potentially making FVC a distinguishing feature for classification or prediction.

Bar Chart:

The bar chart titled "Smoking Status Distribution in the Dataset" in figure 17 displays the distribution of smoking status across the dataset. X-axis represents two smoking status categories, here:

- 0 is represented by non-smokers,
- 1 is represented by smokers.

The Y-axis shows the count of instances for each category.

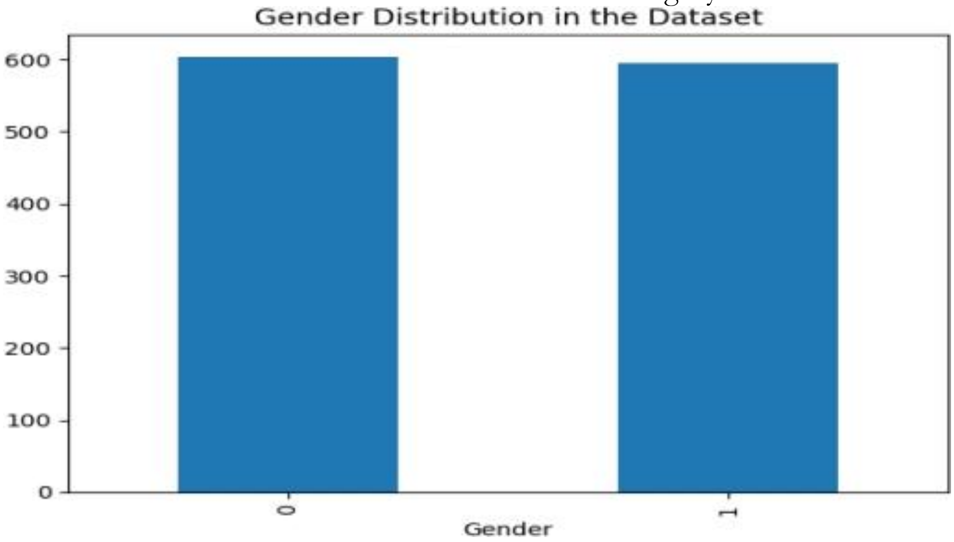

Figure 17. Bar Chart

Partial Dependence Plots:

This plot in figure 18 shows the partial dependence of several features (BMI, FEV1, and FVC on the prediction made by the model. Partial dependence plots (PDPs) help us understand how each feature influences the target variable (diagnosis) while keeping other features constant.

(FEV1 and FVC) have an important impact on diagnosis, and the model is highly sensitive to these features. A decrease in these values strongly correlates with a higher likelihood of diagnosis.

BMI, however, seems to have minimal impact on the model's predictions, as indicated by a flat curve.

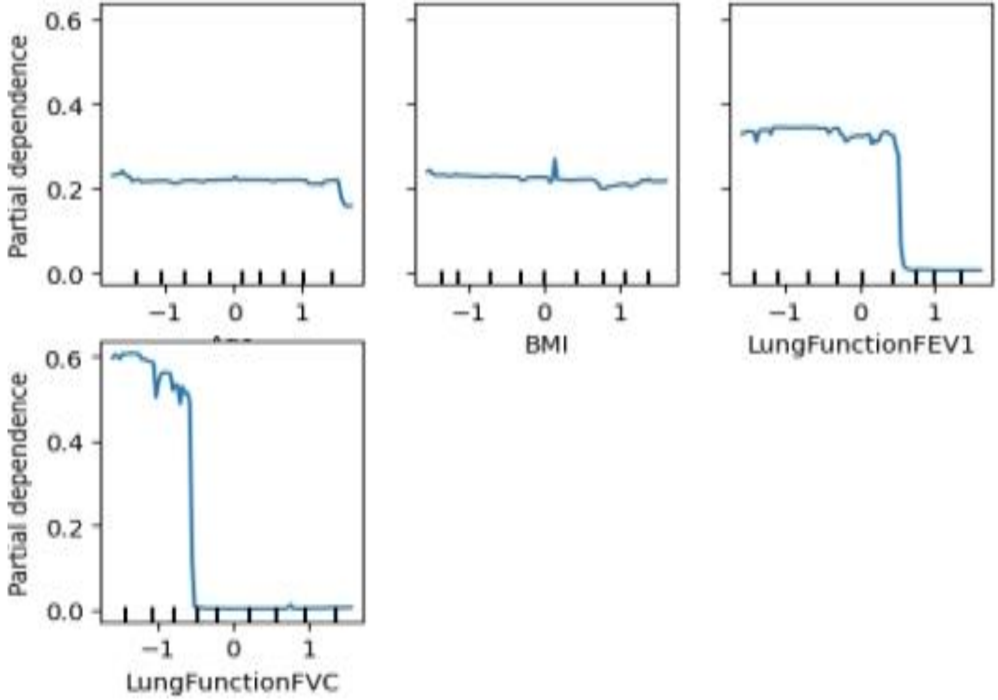

Figure 18. Partial Dependence Plots

PCA (Principal Component Analysis) Plot: This is a 2D PCA (Principal Component Analysis) plot, which projects the dataset into two principal components, allowing us to visualize the data in a simplified two-dimensional space.

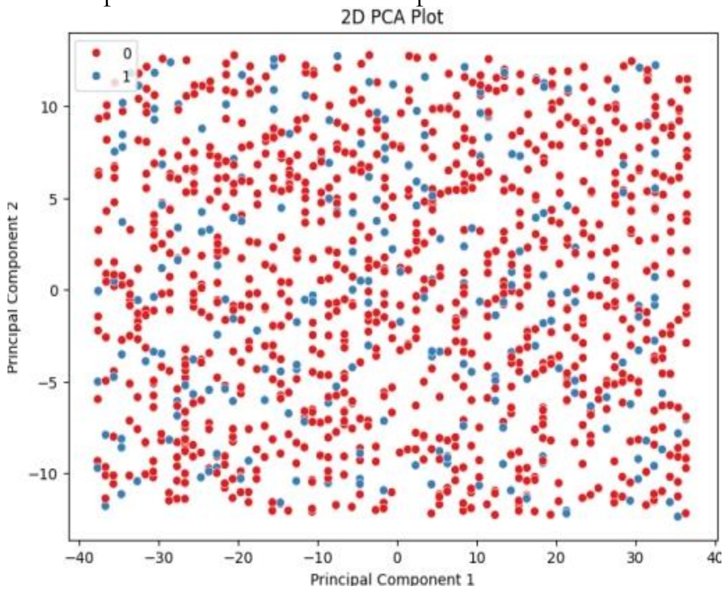

Figure 19. PCA

This PCA plot in figure 19 shows that the dataset may be complex, and the two classes are not linearly separable in this 2D projection. More sophisticated models or additional features may be needed to better separate these classes.

Pair Plot:

This is a pair plot in figure 20 showing the relationships between four features in your dataset: Age, BMI, FEV1, and FVC, grouped by the Diagnosis variable (0 or 1). The pair plot provides a way to visualize both distributions and potential correlations between these variables.

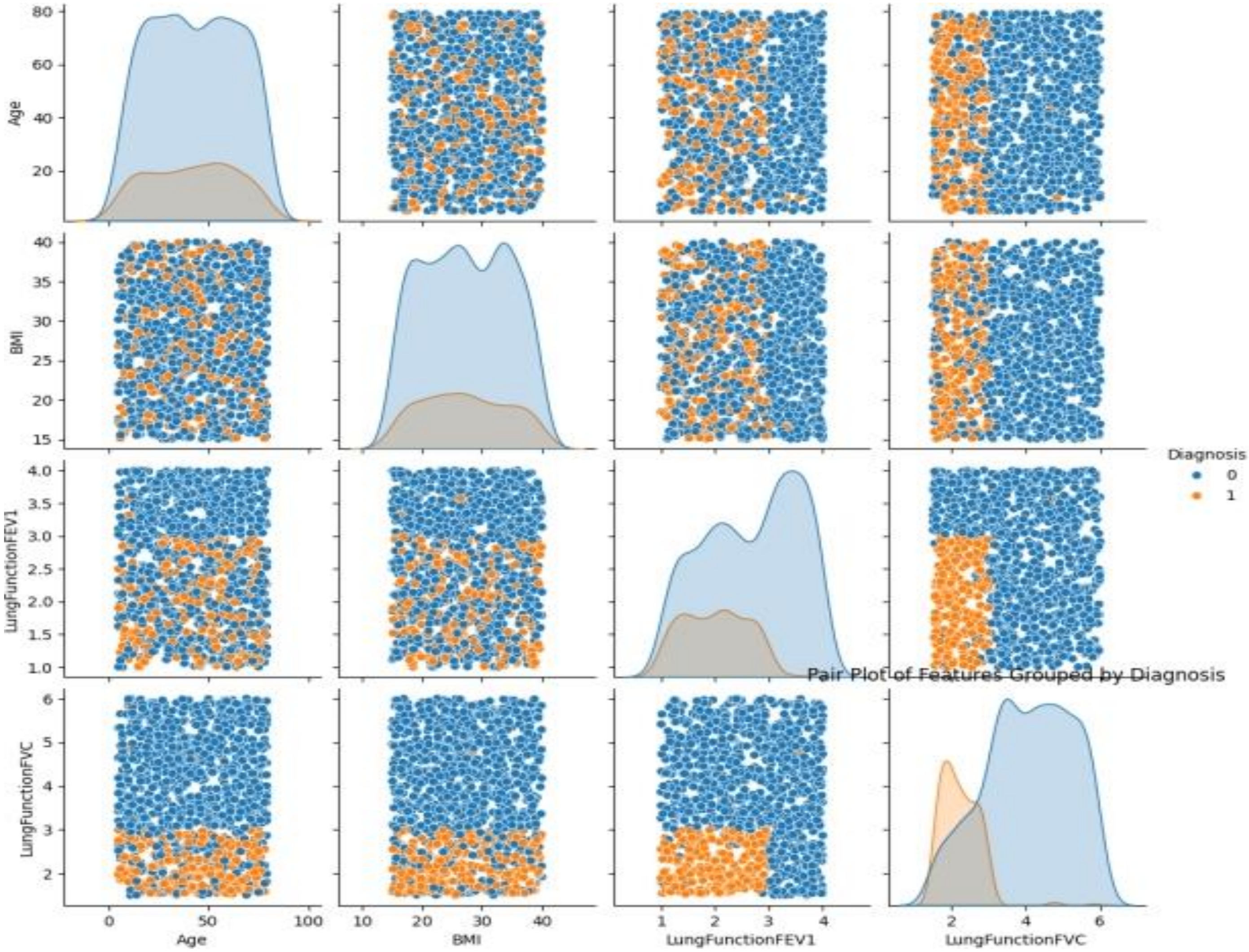

Figure 20. Pair Plot

This plot shows a clearer distinction between the two classes, with more separation between the blue and orange points compared to other pairs. This might indicate that the combination of lung function metrics is more predictive of the diagnosis.

T-SNE Plot:

It is a t-SNE plot, which is a technique that is used for dimensional reduction and provides visualization of high-dimensional data in the lower-dimensional space (in this case, 2D). The plot in figure 21 represents how your data points, which are labeled by the diagnosis variable (0 or 1), are clustered or spread across two components generated by t-SNE.

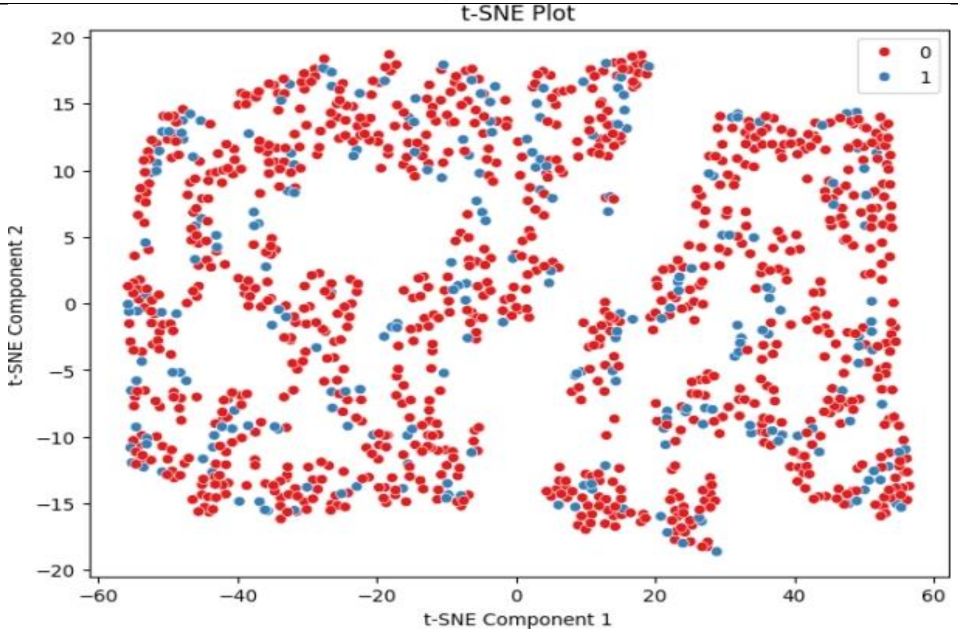

Figure 21. T-SNE Plot

The t-SNE plot reveals how well the two classes (Diagnosis = 0 and Diagnosis = 1) are separated or overlapping.

In this plot, the red points (0) are much more dominant, with the blue points (1) scattered across the plot. However, there are small clusters where the blue points group together.

There is no clear linear separation between the two classes. This could indicate that data is complex and provides a non-linearly separable relation.

3D Bar Plot:

This is a 3D bar plot in figure 22 showing the relationship between Age, Diagnosis, and BMI.

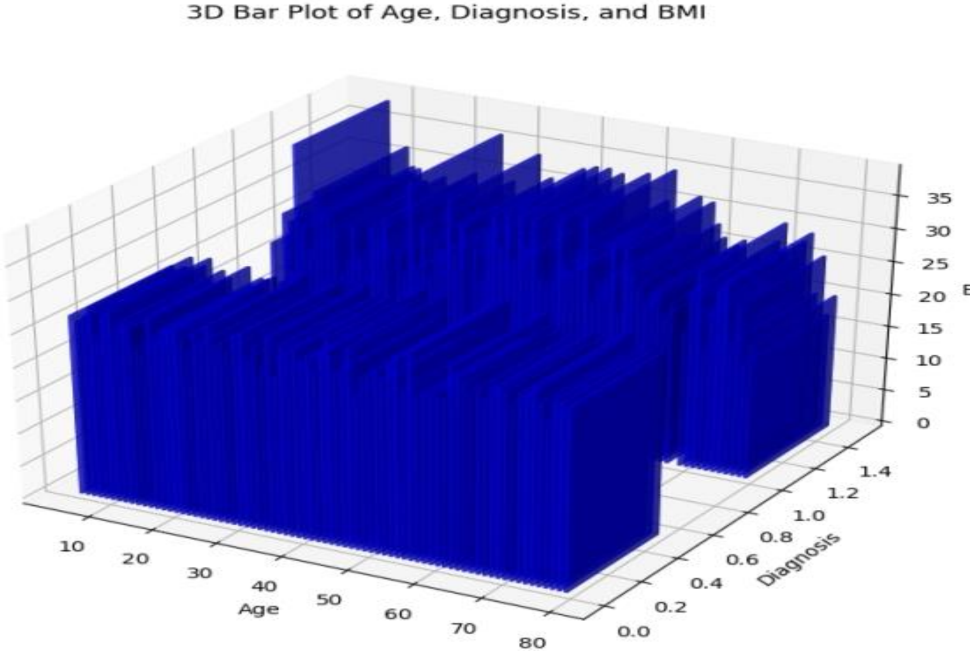

Figure 22. 3D Bar Plot

Each bar in this 3D plot represents the combination of Age, BMI, and Diagnosis.The height of each bar corresponds to the BMI values for the respective age groups and diagnosis label (Diagnosis = 0 or Diagnosis = 1).

There is a dense cluster of bars, showing how the values distribute across the three variables (Age, BMI, and Diagnosis).

Result and Discussion: The machine learning models for asthma diagnosis prediction were evaluated based on the selected features (Age, BMI, Lung Function FEV1, Lung Function FVC). Three models—Random Forest, Neural Network, and XGBoost—were trained and tested on the dataset. Below are the results for each model:

Performance Comparison:

- Random Forest performed the best overall, achieving the highest recall, precision, and F1-score across both classes, with exceptional accuracy in predicting non-asthmatic patients. **XGBoost** followed closely behind a strong balance between recall and precision for both classes and makes it suitable for cases where class imbalance needs careful handling.
- **Neural Network** had slightly lower performance compared to the other two models, but it still delivered strong results, especially for classifying asthmatic patients.

Table 3. Metric of All Models

Metric	Random Forest	Neural Network	XGBoost
Accuracy	97%	95%	96%
Precision (Class 0)	0.99	0.97	0.98
Precision (Class 1)	0.91	0.88	0.91
Recall (Class 0)	0.97	0.96	0.97
Recall (Class 1)	0.98	0.91	0.93
F1-Score (Class 0)	0.98	0.97	0.98
F1-Score (Class 1)	0.95	0.89	0.92
AUC-ROC	High (especially Class 0)	Moderate (slightly lower than Random Forest)	Balanced and high across both classes

This table 3 provides a concise and clear comparison of the three models' performance across the most critical metrics: recall, F1-score, accuracy, precision, and AUC-ROC.

- SHAP (Shapley Additive exPlanations) dataset values were used to predict the feature importance across all models. The results indicated that LungFunctionFEV1 and LungFunctionFVC were the most critical features influencing asthma diagnosis, followed by BMI and Age.

These results indicate that the selected features are relevant for predicting asthma diagnosis, and machine learning models, particularly Random Forest and XGBoost, provide high classification performance for this task.

In this study, we analyzed demographic and clinical features, including age, BMI, and lung function metrics, to predict a specific health diagnosis using machine learning models. Key patterns differentiating diagnosed and non-diagnosed groups were identified through various visualizations, revealing important predictors like lung function values. Our model comparisons showed that the Random Forest classifier achieved the highest accuracy at 97%, followed by Neural Networks at 95%, and XGBoost at 96%. Naive Bayes, while performing slightly lower at 76%, provided useful insights into feature distributions.

These findings demonstrated the value of leveraging machine learning for accurate prediction, diagnosis, and personalized interventions, emphasizing the importance of integrating diverse health features into predictive modeling frameworks for better patient outcomes.

Conclusion:

This study demonstrates that machine learning models can effectively predict asthma diagnosis by leveraging clinical and demographic data, specifically focusing on age, BMI, and lung function parameters (FEV1 and FVC). Among the evaluated models, Random Forest exhibited the highest predictive performance, achieving an accuracy of 97%, precision values of 0.99 (Class 0) and 0.91 (Class 1), recall values of 0.97 (Class 0) and 0.98 (Class 1), and F1-scores of 0.98 (Class 0) and 0.95 (Class 1). This model showed superior capability in identifying patients without asthma (Class 0) while maintaining high performance for asthmatic patients (Class 1). The Neural Network model, with an accuracy of 95%, demonstrated strong but slightly lower performance compared to Random Forest, particularly excelling at handling cases of asthma diagnosis (Class 1) with a recall of 0.91 and a precision of 0.88 for this class. Similarly, XGBoost provided balanced and robust predictions with an accuracy of 96%, maintaining high recall and precision values across both classes, making it suitable for cases involving potential class imbalances. SHAP (Shapley Additive exPlanations) dataset values were employed to enhance the interpretability of these models, revealing that lung function parameters (FEV1 and FVC) were the most influential predictors, followed by BMI and age. The results underlined the importance of lung function measures in predicting asthma and indicate that integrating machine learning techniques with clinical data can facilitate early detection and personalized treatment strategies for asthma patients. Overall, this research supports the efficacy of data-driven models in providing asthma diagnosis, offering healthcare professionals a valuable tool for better clinical decision-making and patient management. The imbalanced dataset, with a larger proportion of non-asthmatic cases (0) compared to asthmatic cases (1), further emphasizes the robustness of the proposed models in handling such data distributions effectively.

Acknowledgment. Acknowledgments are considered necessary.

Author's Contribution. The corresponding author should explain the contribution of each co-author completely.

Conflict of interest. Authors are advised to explain that there exists no conflict of interest for publishing this manuscript in IJIST.

Project details. If this research was conducted as a result of a project, please give details like project number, project cost completion date, etc.

References

[1] S. A. S. Tsang Kevin C. H, Hilary Pinnock, Andrew M Wilson, "Application of Machine Learning to Support Self-Management of Asthma with mHealth," *2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. conjunction with 43rd Annu. Conf. Can. Med. Biol. Eng. Soc.*, 2020, doi: 10.1109/EMBC44109.2020.9175679.

[2] S. J. C. T. Hilary Pinnock, Hannah L Parke, Maria Panagioti, Luke Daines, Gemma Pearce, Eleni Epiphaniou, Peter Bower, Aziz Sheikh, Chris J Griffiths, "Systematic meta-review of supported self-management for asthma: a healthcare perspective," *Prism. RECURSIVE groups*, vol. 15, no. 1, p. 64, 2017, doi: 10.1186/s12916-017-0823-7.

[3] E. Alharbi, A. Cherif, F. Nadeem, and T. Mirza, "Machine Learning Models for Early Prediction of Asthma Attacks Based on Bio-signals and Environmental Triggers," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2022-December, 2022, doi: 10.1109/AICCSA56895.2022.10017305.

[4] N. Kaffash-Charandabi, A. A. Alesheikh, and M. Sharif, "A ubiquitous asthma monitoring framework based on ambient air pollutants and individuals' contexts,"

Environ. Sci. Pollut. Res., vol. 26, no. 8, pp. 7525–7539, Jan. 2019, doi: 10.1007/S11356-019-04185-3.

[5] S. A. S. Kevin C H Tsang, Hilary Pinnock, Andrew M Wilson, "Application of Machine Learning Algorithms for Asthma Management with mHealth: A Clinical Review," *J Asthma Allergy*, vol. 15, pp. 855–873, 2022, doi: 10.2147/JAA.S285742.

[6] A. M. Joseph Prinable, Peter Jones, David Boland, Cindy Thamrin, "Derivation of Breathing Metrics From a Photoplethysmogram at Rest: Machine Learning Methodology," *JMIR Mhealth Uhealth*, vol. 8, no. 7, p. e13737, 2020, doi: 10.2196/13737.

[7] "[PDF] Spirometry: step by step | Semantic Scholar." Accessed: Feb. 15, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Spirometry%3A-step-by-step-Moore/b51f939cd70887ba7b24225bdc26b5c796af1869

[8] E. Bolat, H. Yildirim, S. Altin, and E. Yurtseven, "A COMPREHENSIVE COMPARISON OF MACHINE LEARNING ALGORITHMS ON DIAGNOSING ASTHMA DISEASE AND COPD," *PONTE Int. J. Sci. Res.*, vol. 76, no. 3, 2020, doi: 10.21506/J.PONTE.2020.3.17.

[9] A. N. G. Sumanth Swaminathan , Klajdi Qirko, Ted Smith, Ethan Corcoran, Nicholas G. Wysham, Gaurav Bazaz, George Kappel, "A machine learning approach to triaging patients with chronic obstructive pulmonary disease," *plos 1*, 2017, doi: https://doi.org/10.1371/journal.pone.0188532.

[10] B.-J. H. Zne-Jung Lee, Ming-Ren Yang, "A Sustainable Approach to Asthma Diagnosis: Classification with Data Augmentation, Feature Selection, and Boosting Algorithm," *Diagnostics*, vol. 14, no. 7, p. 723, 2024, doi: https://doi.org/10.3390/diagnostics14070723.

[11] C. H. Lee, J. C. Y. Chen, and V. S. Tseng, "A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring," *Comput. Methods Programs Biomed.*, vol. 101, no. 1, pp. 44–61, Jan. 2011, doi: 10.1016/J.CMPB.2010.04.016.

[12] A. Q. Ansari, N. K. Gupta, and Ekata, "Automatic diagnosis of asthma using neurofuzzy system," *Proc. - 4th Int. Conf. Comput. Intell. Commun. Networks, CICN 2012*, pp. 819–823, 2012, doi: 10.1109/CICN.2012.55.

[13] M. T. K. Mohammed Temraz, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Mach. Learn. with Appl.*, vol. 9, no. 15, p. 100375, 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100375.

[14] F. K. Farnaz Farahanipad, Mohammad Rezaei, Mohammad Sadegh Nasr, "A Survey on GAN-Based Data Augmentation for Hand Pose Estimation Problem," *Technologies*, vol. 10, no. 2, p. 43, 2022, doi: 10.3390/technologies10020043.

[15] K. Nagarajaiah, "ASTHAMA PREDICTION APP USING DJANGO," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 6, no. 6, pp. 221–225, 2024, [Online]. Available: https://www.researchgate.net/publication/382068471_ASTHAMA_PREDICTION_APP_USING_DJANGO

[16] M. U. N Dinesh Kumar, "Digitalized Dust Alert System for Asthma Patients and their Care Takers," *IOP Conf. Ser. Mater. Sci. Eng*, vol. 1057, p. 012092, 2021, doi: 10.1088/1757-899X/1057/1/012092.

[17] M. P. Murugesan Dhasagounder, M Kaviyarasan, S Matheswari, "IOT based Assist Device for Pulmonary Diseased Patients Monitoring Framework," *Int. J. New Innov. Eng. Technol.*, 2023, [Online]. Available: https://www.researchgate.net/publication/371043675_IOT_based_Assist_Device_for_Pulmonary_Diseased_Patients_Monitoring_Framework

[18] G. S. C. Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif

Mehmood, "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," *PLoS One*, vol. 16, no. 2, p. e0245909, 2021, doi: https://doi.org/10.1371/journal.pone.0245909.

[19] T. Wang, Y. Bian, Y. Zhang, and X. Hou, "Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm," *Comput. Geosci.*, vol. 170, p. 105242, Jan. 2023, doi: 10.1016/J.CAGEO.2022.105242.

[20] V. F. Varsha Nemade, "Machine Learning Techniques for Breast Cancer Prediction," *Procedia Comput. Sci.*, vol. 218, pp. 1314–1320, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.110.

[21] M. kumari and P. Ahlawat, "DCPM: an effective and robust approach for diabetes classification and prediction," *Int. J. Inf. Technol.*, vol. 13, no. 3, pp. 1079–1088, Jun. 2021, doi: 10.1007/S41870-021-00656-4.

[22] P. Biswas and T. Samanta, "Anomaly detection using ensemble random forest in wireless sensor network," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 2043–2052, Oct. 2021, doi: 10.1007/S41870-021-00717-8/METRICS.

[23] S. P. Nadikatla Chandrasekhar, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, p. 1210, 2023, doi: https://doi.org/10.3390/pr11041210.

[24] A. Sahu, H. Gm, M. K. Gourisaria, S. S. Rautaray, and M. Pandey, "Cardiovascular risk assessment using data mining inferencing and feature engineering techniques," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 2011–2023, Oct. 2021, doi: 10.1007/S41870-021-00650-W/METRICS.

[25] K. D. Rustam F, Mehmood A, Ahmad M, Ullah S, "Classification of Shopify App User Reviews Using Novel Multi Text Features," *IEEE Access*, vol. 99, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.2972632.