

Examining the Impact of Environmental Factors on Asthma Development Using a Boosting Ensemble Model

Kamran Javed¹, Mumtaz Ahmad², Muhammad Rehan Faheem^{3*}, Rizwana Yasmeen⁴, Hina Shafique⁵, and Muhammad Saeed¹

¹The Department of AI, Faculty of Computing, Lahore Garrison University, Lahore, Pakistan.

²The Department of CS, Faculty of Computing, Lahore Garrison University, Lahore, Pakistan.

³Fakulti Kecerdasan Buatan dan Keselamatan Siber, Universiti Teknikal Malaysia, 76100, Melaka, Malaysia.

⁴Gomal Research Institute of Computing, Gomal University, Dera Ismail Khan, Pakistan.

⁵Department Software Engineering, Faculty of Computer Science, Superior University, Pakistan.

*Corresponding Author: Muhammad Rehan Faheem. Email: rehan@utem.edu.my

Received: May 23, 2026 Accepted: August 27, 2026

Abstract: Asthma is a chronic inflammatory airway disease which is affected by both environmental, demographic, lifestyle and clinical factors. Early prediction of asthma risk is vital for early intervention and better asthma management. . This study suggests the boosting ensemble approach to predict the development of asthma on a dataset of 2,392 people with demographic information, lifestyle information, allergy history, environmental exposures, and clinical attributes. To enhance the reliability of the model, data preprocessing techniques such as data cleaning, encoding categorical features, scaling features, and feature selection were employed. The proposed framework uses a soft-voting approach that averages class probabilities and compares the models using a consistent experimental setup. The experimental results show that the proposed boosting ensemble produced an accuracy of 99%, better than all baseline models and provides highly accurate predictions for asthma and non-asthma classes. The results show that combining some boosting algorithms improves the prediction accuracy and robustness than using the individual classifiers. This proposed framework offers a reliable tool for predicting asthma in early life and may serve as a basis for future clinical decision making, but needs to be validated with other independent and geographically distinct populations in order to establish its generalizability. Because the dataset is synthetic, these results show methodological performance under controlled experimental conditions only; clinical usefulness would require validation on real patient data.

Keywords: Asthma Prediction; Boosting Ensemble; XGBoost; LightGBM; CatBoost; Environmental Factors; Machine Learning

1. Introduction

Asthma is a typical chronic respiratory disorder with inflammation and constriction of the bronchial passages that may result in frequent attacks of wheezing, breathlessness, chest tightness and coughing. It impacts people of all ages and many patients begin to experience it at a young age [1] . Asthma is a major public health problem with an estimated 260 million people suffering from the disease worldwide in 2021 [2] . Chronic respiratory diseases also cause a number of deaths, including asthma. Asthma can be difficult to identify at an early stage because there are many factors at both the genetic, environmental and individual level that can influence the development and progression of asthma. Early prediction of asthma risk is crucial for timely early intervention, disease management and decrease in asthma burden [3].

Asthma onset and severity are greatly influenced by environment. Chronic respiratory disease is caused by increased exposure to air pollutants, extreme temperatures, wildfire and airborne allergens [4].

Eosinophilic pneumonia (EP) affects lung function, causes inflammation and leads to enhanced oxidative stress [5]. Climate change will exacerbate these impacts by an increase in extreme weather events and extended pollen seasons [6]. Children with a low socio-economic status are the most at risk from early life exposure to fine particulate matter for asthma [7]. These are interdependent and have a non-linear relationship. This complexity restricts usual analysis techniques. Clinical and environmental data are often high dimensional and difficult to model with standard statistical models [8]. This is where machine learning is a good alternative. It acquires patterns over a vast space of multiple interacting variables not encompassed by classical methods. Machine learning models can have different strengths for predicting outcomes in pediatric asthma [9]. Machine learning is an ideal tool for asthma risk prediction due to these strengths. As regards the development of asthma models, there are many comparisons between the RF and the SVM, such as in the case of asthma risk prediction [10]. Using a RF approach, severe asthma phenotypes have been identified based on demographic, clinical and inflammatory features [11]. One of the strengths of RF is that it can be used with high dimensional data and is scalable [12]. These studies have concurred that the single models provide a good baseline. But they fail to capture some of the more intricate relationships and are prone to being affected by skewed data. Boosting and ensemble are a newer theme. Combining the strengths of models and smoothing out the errors from each model individually improves prediction with boosting techniques [13].

In the field of diagnosis, the extreme gradient boosting with feature selection has been shown to improve the diagnosis of asthma [14]. Gradient boosting has been consistently shown to be the most successful technique in this area [15]. XGBoost, LightGBM and CatBoost have different strengths to regularize, to speed up and to deal with categorical data [16]. Few studies couple these three complementary boosters together in one model in the context of asthma. The potential of the current state of the art machine learning techniques, there are several challenges. First, the dataset of asthma is imbalanced, which can lead to a bias of the prediction models towards the majority class and make predictions on minority classes unreliable [17]. Secondly, there is a lack of uniformity across existing studies in terms of data sets, data processing methods, validation procedures, and evaluation protocols that make it difficult to compare the results of the various studies and repeat them [18].

The research introduces an average class probability base soft-voting boosting ensemble model with the combination of XGBoost, LightGBM and CatBoost. All models use a common data processing pipeline and evaluation protocol to facilitate fair comparisons and reproducible results. The proposed framework will give an accurate and reliable tool for predicting asthma development from environmental and related risk factors.

This study has made the following main contributions:

1. We propose a heterogeneous boosting ensemble integrating soft voting ensemble that combines XGBoost, LightGBM and CatBoost to learn complex and potentially non-linear relationships between environmental, clinical, demographic and lifestyle factors that influence the development of asthma.
2. In order to provide a solid analytical framework for modeling multifactorial asthma risk, we establish a unified machine-learning framework for asthma development prediction that incorporates systematic data preprocessing, feature selection, heterogeneous boosting, and probability-based decision fusion.
3. We demonstrate the predictive superiority of the proposed ensemble and its potential to support data-driven assessment of asthma development associated with environmental and clinical factors thru a controlled comparative evaluation against five well-known machine-learning classifiers under the same experimental protocol.

The rest of the paper is organized in this manner. The relevant literature is reviewed in Section 2. The section 3 describes the materials and methods, including the proposed model. Section 4 displays the results of the experiment. The results are discussed in Section 5. The study is concluded in Section 6.

2. Literature Review

The increasing amount of environmental and healthcare data available, machine learning techniques have been adopted for asthma prediction. In this study [19], air pollution, an LSTM model on meteorological and air pollution data of Seoul showed the Pearson correlation coefficient of 0.984, which means that the meteorological data contains significant predictive information of air pollution. The

cumulative 15- and 30-day exposure to PM10 was modeled by randomly combining 3,609 asthma outpatient records with hourly PM10 measurements, with best performance of the Random Forest model ($R = 0.92$) the use of the data from an outpatient hospital in one city may limit the generalizability of the results, the limited number of environmental and patient-level variables used, and the data from one year [20]. In another study which used the NHANES 2017–2018 data set (8,888 adults), several ML algorithms were compared, and Random Forest and SVM had an accuracy of 84.53% and an AUC of 0.73 respectively but the study was limited by relying on self-reported asthma, cross-sectional nature and no external validation [21]. Another study included 1,473 cases of asthma in 350 neighborhoods of Tehran and correlated environmental, demographic, socioeconomic and GIS variables with Gradient Boosting and obtained testing R^2 . It was limited in applicability, however, because of a lack of individual-level information, spatial constraints, and single city validation, with an applicability of 0.76 [22]. In general, these studies suggest that ML provides effective assistance in asthma prediction, and simultaneously point out the necessity for integrated models that include clinical, demographic and environmental characteristics, spatial characteristics and larger datasets that can be validated with external data.

A study in Tehran combined Geographic Information Systems (GIS), remote sensing and ensemble learning to locate high-risk asthma zones, using the Gradient Boosting Machine (GBM) for the highest predictive performance ($R^2 = 0.76$) with the use of environmental, socioeconomic and weather parameters [23]. The model showed good performance, it was validated in a single geographic area, which reduces its applicability. In a Canadian birth cohort study, five machine learning models and two ensemble models were tested for their predictive ability for asthma in 3-year-old children, with the best models having an AUC higher than 0.90 [24]. While both studies have shown potential in the application of machine learning for asthma prediction, the results need to be verified in different geographical areas and populations to increase the generalizability of the results.

Using clinical and environmental data from 1,368 children, the Isle of Wight birth cohort research created seven machine learning models to predict asthma; the SVM had the best predictive ability (AUC = 0.82) [25]. Using climate, air pollution, and hospitalization data, machine learning has also been used to predict asthma-related healthcare consumption. RF, XGBoost, Support Vector Regression (SVR), KNN and Multiple Linear Regression models were assessed. After adding delayed hospitalization data, the top-performing model obtained an R^2 of 0.84, indicating the usefulness of past healthcare usage data in enhancing predictive performance [26]. A Gradient Boosting model was applied to a pediatric study using clinical and lung-function data from 900 children, and yielded an F-score of 0.974 [27]. The model was developed within a single clinical center, the performance of the model outside the population studied is not guaranteed. Similar limitations have been noted in previous research conducted by survey means, but with larger data sets. One study compared the performance of the XGBoost, Neural Network (NN), RF, SVM, and Logistic Regression (LR) models, based on the 2023 NHIS data set, which had 9,716 children [28]. LR achieved the best AUC score of 0.7654 following under-sampling, and the F1 score of 0.3452 illustrated the impact of class imbalance on the performance of the prediction. All of these results lend themselves to the idea that while more data may help overcome data imbalance and generalizability issues in the model, it is not a cure-all.

In Sri Lanka, 13 MLAs were compared and a hybrid Logistic Regression–LightGBM model with a symptom self-reported data obtained an AUC of 0.9062, highest among all the algorithms tested [29]. The results show that ensemble learning algorithm can be more accurate than the single algorithm. With the emergence of some more sophisticated ensemble frameworks, interest has continued to be renewed in boosting methods. In the study [30], a CatBoost model with feature selection, class balancing, data preprocessing and hyperparameter optimization demonstrated an accuracy of 93% for asthma attack prediction, outperforming several popular machine learning algorithms. In another research, CatBoost was fused with acoustic and clinical parameters to classify asthma based on AIRS Kaggle dataset with the accuracy of 90.3% and F1 score of 0.915 [31]. In this Study [32], used a Generative Adversarial Network (GAN) model combined with XGBoost to tackle the issues of class imbalance and feature selection, with an AUC of 0.929 and an accuracy of 94.03%. The results show that the boosted algorithms are able to learn complex data related to asthma and to mitigate imbalanced data effects.

Previous studies have shown that machine learning could be used to predict asthma, few have considered the effect of incorporating environmental, demographic, lifestyle and clinical characteristics

into a single boosting ensemble and explored the role played by environmental factors as contributors to asthma risk. Previous research often analyzes the predictive performance of a model, the performance of only one machine learning model, or the performance data with a geographically restricted dataset. There is a need for an integrated and interpretable boosting framework that can accurately predict asthma development and delineate the environmental factors that have the strongest association with the prediction. The overall summary of recent state of the art for asthma prediction data is described in Table 1.

Table 1. Summary of recent state-of-the-art on machine learning for asthma prediction data

Ref.	Methodology	Dataset(s)	Key Results	Limitations
[19]	LSTM-FC neural network	Seoul meteorological + air pollution data	Pearson correlation 0.984	City-specific, limited transfer
[20]	RF + other ML models	Cumulative PM10 data, Adiyaman, Turkey (post-earthquake)	R-value 0.92	Single-city post-disaster setting
[21]	SVM, RF, AdaBoost, XGBoost, KNN, NB, MLP	NHANES 2017–2018 (8,888 adults).	SVM achieved AUC = 0.72.	Excluded environmental factors
[22]	Random Forest, GBM, XGBoost with GIS	1,473 patients + air quality/weather/socioeconomic, Tehran	GBM achieved $R^2 = 0.76$	Limited to Tehran-specific data.
[23]	XGBoost-based asthma severity prediction	Air pollutant + demographic data, Uttarakhand, India	Accuracy = 97%	Evaluated a single boosting algorithm.
[24]	5 ML models + 2 ensembles	CHILD birth cohort, children $\leq$ age 4 (Canada)	AUC > 0.90 at age 3	Poor prediction before age 1
[25]	7 ML algorithms (SVM best)	Isle of Wight birth cohort, 1,368 children	SVM AUC 0.82	Modest sample size
[26]	RF, XGBoost, SVR, KNN, MLR	Climate, air pollution, and hospitalization records	Best $R^2 = 0.84$	Single-region study; no patient-level clinical variables.
[27]	Gradient boosting (+ 7 others)	900 children, clinical + lung-function features	F1 0.974	Single-clinic; limited external validity
[28]	XGBoost, NN, RF, SVM, LR + sampling	9,716 youth, NHIS 2021–2022 (validated on 2023)	AUC 0.7654, F1 0.3452	Severe imbalance; low F1
[29]	13 MLAs compared; hybrid Logistic Regression–LightGBM	Self-reported symptom data from Sri Lanka	AUC = 0.9062, highest among tested models	Self-reporting bias and limited generalizability
[30]	CatBoost ensemble with feature selection and class balancing	Clinical asthma dataset	Accuracy = 93%	Focused on asthma attack prediction rather than disease development.
[31]	Class-weighted CatBoost ensemble	AIRS Kaggle dataset (clinical + acoustic features)	Accuracy = 90.3%, F1 = 0.915	Dependence on acoustic inputs limits clinical applicability
[32]	GAN + XGBoost	Asthma dataset	94.03% Accuracy, AUC = 0.929	Focused on classification, limited validation across populations

3. Materials and Methods

This section presents the materials, dataset, and methodology designed for the proposed boosting ensemble framework for classification asthma. It illustrates the overall framework architecture consisting of data collection, preprocessing, feature engineering, model training, and probabilistic ensemble learning. The methodology exploits three complementary gradient boosting algorithms Boost, LightGBM and CatBoost and uses a soft-voting strategy to combine their prediction probabilities to improve classification performance and robustness. This section describes the experimental setup and the evaluation metrics used to measure the effectiveness of the proposed model.

3.1. Dataset

In this study, we used the publicly available Asthma Disease Dataset. It includes information on 2,392 individuals, each represented by at least one record. The dataset was extracted from Kaggle data. The dataset consists of 2,392 synthetic individuals with demographic, lifestyle, environmental, allergy-related, medical history, clinical, and symptom-related variables. The original dataset is highly imbalanced (2,268 non-asthma and 124 asthma records). To balance the classes, SMOTE was applied to the full dataset, giving 4,536 records (2,268 per class). This balanced set was then divided in a 70:30 ratio into 3,175 training records and 1,361 test records (703 non-asthma and 658 asthma). These are the observations shown in Figure 1. The full list of predictor variables and their associated feature groups are described in Table 2. The outcome variable, Diagnosis, was constructed as a binary variable, with 0 = no asthma and 1 = asthma. PatientID and DoctorInCharge were excluded from the predictive analysis as they are identifiers or non-predictive fields. The possible influence of clinical and symptom-related variables on model performance was taken into account when interpreting the results, as some of these may be reflecting characteristics measured simultaneously to asthma diagnosis.

Table 2. Feature Groups in the Asthma Dataset

Feature group	Features
Demographics	Age, Gender, Ethnicity, Education Level
Lifestyle	Smoking, diet quality, physical activity
Environmental and allergy	PollutionExposure, PollenExposure, DustExposure, PetAllergy
Clinical measurements	LungFunctionFEV1, LungFunctionFVC
Medical history	FamilyHistoryAsthma, HistoryOfAllergies, Eczema, HayFever, GastroesophagealReflux
Symptoms	Wheezing, ShortnessOfBreath, ChestTightness, Coughing, NighttimeSymptoms, ExerciseInduced
Target	Diagnosis (0 = No asthma, 1 = Asthma)

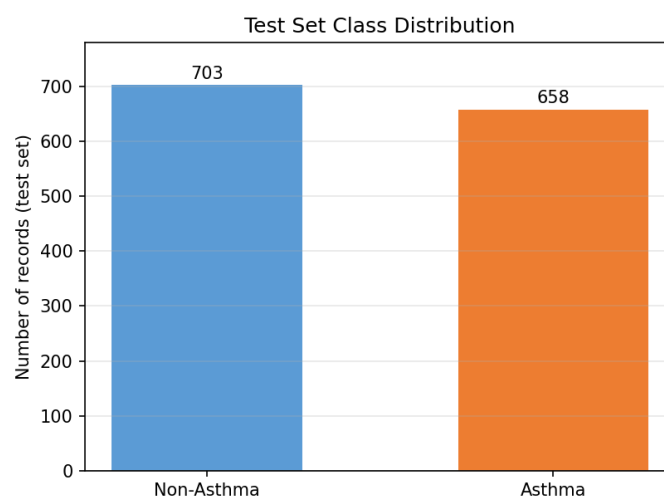


Figure 1. Test set class distribution after SMOTE balancing (703 non-asthma, 658 asthmas)

3.2. Data Preprocessing

The dataset of asthma was preprocessed to ensure the consistency of data and improve the reliability of learning process. The preprocessing pipeline included duplicate record removal, missing-value imputation, categorical feature transformation, feature normalization, and feature selection. To avoid duplicate patient records, duplicate observations were eliminated. Missing numeric and categorical attributes were filled in with the median and mode of the respective training features. With the original feature matrix missing numerical observations were assigned as in equation (1).

$$(\tilde{x})_{ij} = x_{ij}1_{x_{ij} \neq NA} + \text{median}(X_j)1_{x_{ij} = NA} \quad (1)$$

Missing categorical observations were replaced by the most frequently occurring category as described in equation (2).

$$\tilde{x}_{ij} = \text{mode}(X_j) \quad (2)$$

Categorical variables were then transformed into numeric versions using one-hot encoding for nominal variables and ordinal encoding for ordered variables. This transformation preserves the categorical information and allows efficient processing by the learning algorithms. The one-hot encoding of a categorical attribute with K different classes are expressed as in equation (3).

$$\phi(x_{ij}) = [I(x_{ij} = c_1), I(x_{ij} = c_2), \dots, I(x_{ij} = c_K)] \quad (3)$$

To reduce the impact of disparate measurement scales, continuous variables were normalized using z-score normalization. Every numerical characteristic was altered in accordance with in equation (4).

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (4)$$

Numerical stability during model training is enhanced by the transformation that results in standardized variables with zero mean and unit variance.

Correlation analysis and tree-based feature significance were used for feature selection minimize feature redundancy and preserve the most valuable predictors. The Pearson correlation coefficient was first used to identify strongly connected data as formulated in equation (5).

$$\rho_{jk} = \frac{\sum (x_{ij} - \mu_j)(x_{ik} - \mu_k)}{(N-1)\sigma_j\sigma_k} \quad (5)$$

To reduce multicollinearity, correlated predictors that above a predetermined threshold were eliminated. The cumulative split gain from gradient-boosted decision trees, as shown in equation (6) was then used to rank the remaining variables.

$$\text{Imp}(X_j) = \sum_{t=1}^T \sum_{s \in S_{t,j}} \Delta G(s) \quad (6)$$

The The dataset showed a substantial class imbalance which could lead to unreliable model learning due to insufficient minority-class representation. Therefore, SMOTE oversampling was applied to increase minority-class representation before model training. After balancing, the dataset was divided using stratified sampling to preserve class distribution. All subsequent preprocessing, feature transformation, and model training steps were performed using the training data, while model evaluation was conducted on the independent test subset. This strategy was adopted to improve minority-class learning and provide a more reliable assessment of classification performance under imbalanced conditions. All preprocessing parameters, including as feature-selection criteria, normalization coefficients, and imputation statistics, were only computed from the training data before being applied to the testing and validation datasets in order to prevent information leaking.

3.3. Data Splitting

The original dataset is highly imbalanced (2,268 non-asthma and 124 asthma records). To balance the classes, SMOTE was applied to the full dataset, giving 4,536 records (2,268 per class). This balanced set was then divided in a 70:30 ratio into 3,175 training records and 1,361 test records (703 non-asthma and 658 asthma). These are the observations shown in Figure 1. Because oversampling was carried out before partitioning, the test set contains synthetic samples derived from neighboring training records.

3.4. Proposed Boosting Ensemble Framework

The goal of the proposed boosting ensemble framework is to enhance the accuracy of asthma prediction by combining three complementary gradient-boosting algorithms (XGBoost, LightGBM and CatBoost) through soft voting, that is, equal-weight averaging of their predicted class probabilities as defined in Equation (12). No meta-learner is trained, so the method is a soft-voting ensemble and not a

stacking ensemble. The general architecture of the proposed system is shown in Fig. 2. The asthma dataset is preprocessed and features are selected to form a uniform and meaningful feature space. The optimized feature matrix is then fed to the three boosting classifiers, with each of these classifiers acting independently and estimating the posterior probability of asthma. The probabilistic outputs are finally combined by soft voting to get the final prediction. This architecture makes use of the complementary learning property of the individual boosting algorithms, and increases the robustness of the predictions, as they will not rely solely on one algorithm.

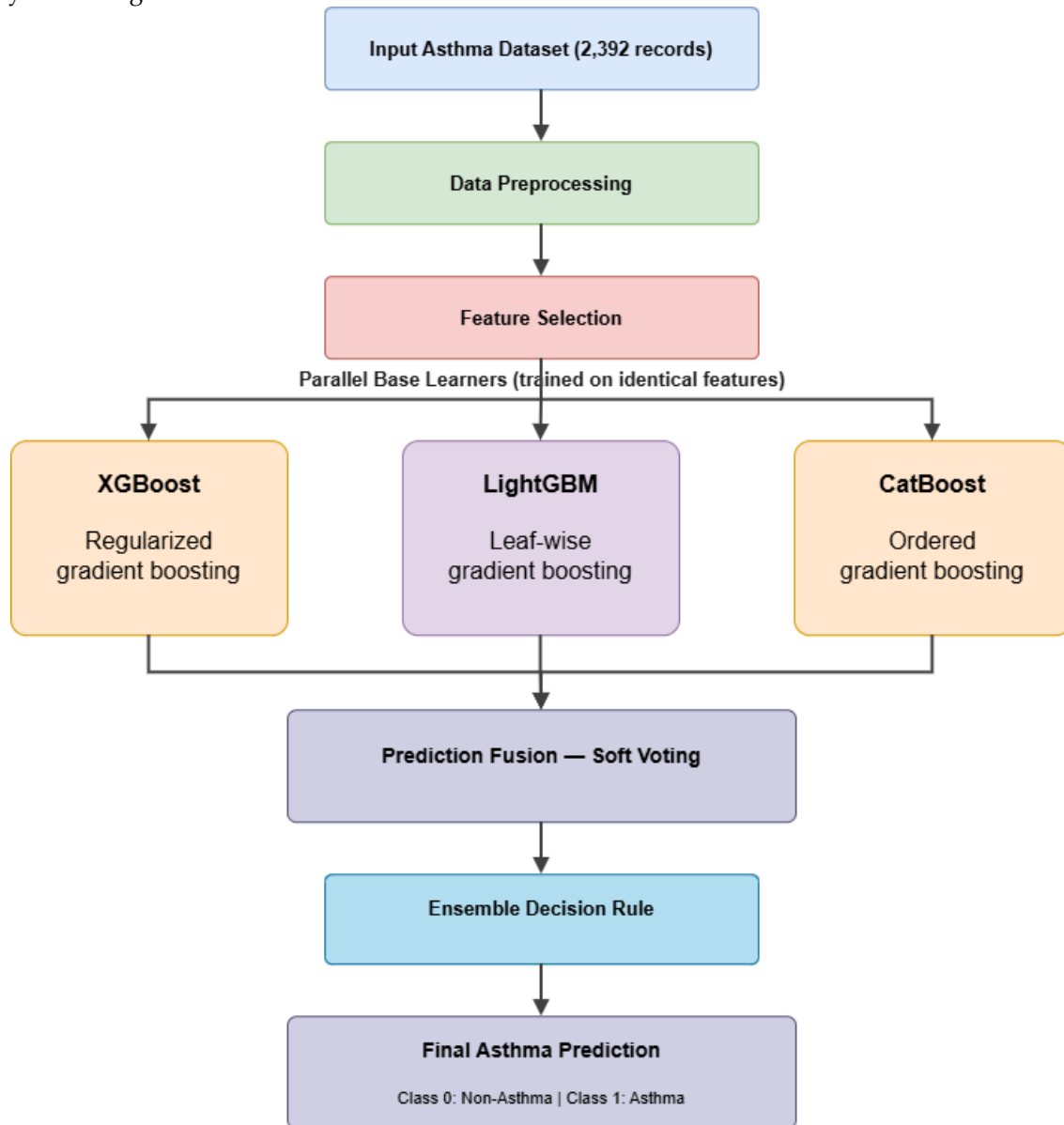


Figure 2. Architecture of the proposed boosting ensemble

The preprocessed dataset be represented in equation (7).

$$D^* = \{(x_i, y_i)\}_{i=1}^N \quad (7)$$

The ensemble consists of three base learners shown in equation (8).

$$M = \{M_{XGB}, M_{LGBM}, M_{CAT}\} \quad (8)$$

The members of the gradient-boosting family, each of the three models uses a distinct approach to tree building and optimization. By employing second-order gradient optimization to build regularized additive trees, XGBoost increases prediction accuracy while managing model complexity. To increase computational efficiency and shorter training times, LightGBM uses a leaf-wise tree growth technique in conjunction with a histogram-based learning algorithm. On the other hand, CatBoost uses target-statistics encoding and ordered boosting to reduce prediction shift and facilitate efficient learning from categorical

data. The three classifiers produce complementary decision limits that increase the ensemble's diversity as a result of these algorithmic variations.

Equation (9) illustrates how each classifier individually determines the posterior probability of asthma for an input feature vector x .

$$P_m(y = 1|x), m \in \{1,2,3\} \quad (9)$$

Each boosting classifier constructs an additive ensemble of decision trees is shown in equation (10).

$$F_{m(x)} = \sum_{t=1}^{T_m} f_{t(x)} \quad (10)$$

The corresponding posterior probability is obtained using the sigmoid transformation as defined in equation (11).

$$P_m(y = 1|x) = \frac{1}{1+\exp(-F_{m(x)})} \quad (11)$$

The proposed framework uses a soft-voting technique to aggregate the posterior probability produced by XGBoost, LightGBM, and CatBoost instead of depending on the prediction of a single classifier. The final posterior probability is calculated using equation (12) where a unified confidence score is obtained by averaging the probabilities given by the three boosting classifiers with equal weights. The decision process stated in equation (13) is then used to decide the final class label. If the averaged posterior probability is greater than or equal to the predetermined threshold, the sample is categorized as asthma; if not, it is classed as non-asthma.

$$P(y = 1 | x) = \left(\frac{1}{3}\right) \sum_{m=1}^3 P_m(y = 1|x) \quad (12)$$

$$\hat{y} = \begin{cases} 1, P(y = 1 | x) \geq 0.5 \\ 0, P(y = 1 | x) < 0.5 \end{cases} \quad (13)$$

The proposed model seeks to increase generalization and prediction accuracy over individual boosting classifiers by combining the complementing characteristics of XGBoost, LightGBM, and CatBoost into a single soft-voting framework. The boosting ensemble framework was proposed with three heterogeneous gradient-boosting classifiers such as XGBoost, LightGBM and CatBoost. After data preprocessing and feature selection, the same set of features optimized was fed into all the base learners to ensure the same learning process. All classifiers were trained independently with the same feature representation, so that each classifier learned complementary decision boundaries based on its different boosting algorithm. The posterior probabilities were then combined using the soft voting method described in equation (12) and the final asthma prediction was obtained using the decision rule defined by equation (13).

The same hyper-parameter values were used for all the constituent learners to make an honest comparison. The boosting classifiers were trained for 40 boosting iterations with a learning rate of 0.1 and a maximum tree depth of six. In the case of binary classification, binary cross-entropy (also known as log loss) was used to optimize. The model robustness and generalization performance was evaluated through k-fold cross validation during training. Table 3 summarizes the full ensemble boosting implementation details for the proposed boosting ensemble.

Table 3. Configuration of the proposed boosting ensemble framework

Parameter	Configuration
Prediction task	Binary classification
Input features	Preprocessed environmental, demographic, lifestyle, clinical, and medical history features
Base learner 1	XGBoost (gradient boosting decision trees)
Base learner 2	LightGBM (leaf-wise gradient boosting)
Base learner 3	CatBoost (ordered boosting with categorical feature handling)
Ensemble strategy	Soft voting (probability averaging)
Number of base learners	3
Number of estimators	40 per model
Learning rate	0.1
Maximum tree depth	6
Objective function	Binary cross-entropy (Log Loss)
Validation strategy	(k)-fold cross-validation

Probability output	Posterior probability ($P(y=1 \mid \mathbf{x})$)
Decision threshold	0.5
Final output	Class 0 (Non-asthma), Class 1 (Asthma)

The same training subset and feature representation were used to independently optimize each of the three boosting classifiers during model training. In order to minimize the binary cross-entropy goal, each method built its own ensemble of decision trees in accordance with its own optimization strategy. A single confidence score was obtained during inference by averaging the posterior probabilities computed by XGBoost, LightGBM, and CatBoost using the soft-voting technique. The final asthma prediction was then determined by comparing the averaged likelihood with a 0.5 decision threshold. By utilizing the complementing capabilities of the individual boosting algorithms, this probability-level fusion allows the suggested ensemble to outperform individual boosting models in terms of predictive resilience and generalization.

3.5. Performance Evaluation

The proposed boosting ensemble framework's diagnostic ability asthma prediction is a binary classification problem, so the confusion matrix was used to calculate the accuracy, precision, recall and f1 score. Moreover, the receiver operating characteristic (ROC) curve and the area under the curve (ROC-AUC) were used to assess the discriminative power of the proposed model for various decision thresholds. The metrics collectively measure classification accuracy, the reliability of prediction, the sensitivity to positive cases, and the strength of the proposed framework in general. The confusion matrix has four predictions outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These quantities denote the following evaluation metrics. The overall classification accuracy is given by in equation (14).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Equation (15) expresses precision, which is the percentage of projected asthma cases that are properly diagnosed.

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

Equation (16) computes recall, also known as sensitivity, which is the percentage of real asthma cases that the suggested model accurately detected.

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

The F1-score is the harmonic mean of precision and recall, combining both into a single metric, as given in Equation (17). Because it accounts for both false positives and false negatives, it is particularly informative for imbalanced datasets. The formula for calculating the F1-score is:

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (17)$$

The receiver operating characteristic (ROC) curve was examined in order to assess the suggested boosting ensemble's discriminative power across all classification levels. The definition of the area under the ROC curve (ROC-AUC) is in equation (18).

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (18)$$

After model training was finished, all evaluation metrics were calculated using the independent testing dataset. To provide a fair and impartial performance evaluation, the suggested boosting ensemble was compared with the baseline machine learning classifiers using the same evaluation criteria.

4. Results

This section presents the empirical evaluation of the proposed soft-voting boosting ensemble and its five single-model baselines. All models were trained and tested using the same protocol and performance is reported at both aggregate and per-class levels, supplemented by confusion-matrix and ROC analysis. In clinical practice, reporting per-class metrics in addition to overall accuracy is important because the costs of missed asthma cases (false negatives) and false alarms (false positives) are asymmetric and must be explicitly examined.

4.1. Experimental Setup

The scikit-learn, XGBoost, LightGBM, and CatBoost libraries were used in a Google Colab environment to develop the models in Python. All models were trained and assessed using the same dataset partition, preprocessing pipeline, and evaluation criteria to guaranty a fair and consistent comparison. Google Colab's computational resources were used to carry out the studies in table 4.

Table 4. The baseline model configuration and hyperparameters

Model	Hyperparameter Settings
Support Vector Machine (SVM)	RBF kernel; C = 1.0; gamma = "scale"; probability = True; random_state = 42
Decision Tree	Gini criterion; max_depth = None; random_state = 42
Random Forest	60 trees; max_depth = None; random_state = 42
K-Nearest Neighbors (KNN)	k = 5; Euclidean distance metric; uniform weights
XGBoost	40 boosting iterations; learning rate = 0.1; maximum depth = 6; binary log-loss objective; seed = 42
LightGBM	40 boosting iterations; learning rate = 0.1; maximum depth = 6; binary log-loss objective; seed = 42
CatBoost	40 boosting iterations; learning rate = 0.1; maximum depth = 6; binary log-loss objective; seed = 42

4.2. Proposed Model Performance

On the test set, the proposed ensemble correctly classified 1,347 of 1,361 records, an overall accuracy of 0.9897. For the non-asthma class, precision, recall and F1-score were all 0.9900 (support = 703). For the asthma class, precision, recall and F1-score were all 0.9894 (support = 658). This corresponds to a sensitivity of 0.9894 and a specificity of 0.9900. Table 5 all computed from the same test labels and the same predictions.

High and well-balanced precision, recall and F1-scores for both the classes. The proposed ensemble model has not shown any class bias in generalizing to the unseen data. The results of these research studies support the strength and validity of the proposed method to predict asthma. Table 5 summarizes the proposed boosting ensemble's performance across the assessed metrics.

Table 5. Performance evaluation of the proposed boosting ensemble

Class	Precision	Recall	F1-score	Support
Non-Asthma	0.9900	0.9900	0.9900	703
Asthma	0.9894	0.9894	0.9894	658
Accuracy			0.9897	1,361
Macro Average	0.9897	0.9897	0.9897	1,361

The proposed boosting ensemble model's per-class performance is shown in Figure 3. The accuracy, recall, and F1-score values for both the asthma and non-asthma classes were consistently high (0.98–0.99), indicating balanced classification performance and the model's efficacy in correctly detecting both groups with little misclassification.

The confusion matrix of the proposed boosting ensemble model on the test dataset is shown in Figure 4. Just seven non-asthma patients were improperly identified as asthma, while seven asthma cases were inaccurately forecasted as non-asthma. The model correctly classified 696 non-asthma and 651 asthma cases. The extremely low number of misclassifications verifies the efficacy of the suggested model in correctly detecting both asthma and non-asthma patients and highlights its great discriminative abilities.

The Receiver Operating Characteristic (ROC) curves for the asthma and non-asthma classes of the suggested boosting ensemble model are shown in Figure 5. Excellent discriminative performance was demonstrated by the model's AUC of 0.994 for the non-asthma class and 0.983 for the asthma class. The model's high ability to differentiate between asthma and non-asthma patients across various classification criteria is demonstrated by the ROC curves, which stay toward the upper-left corner and are significantly above the random classifier baseline.

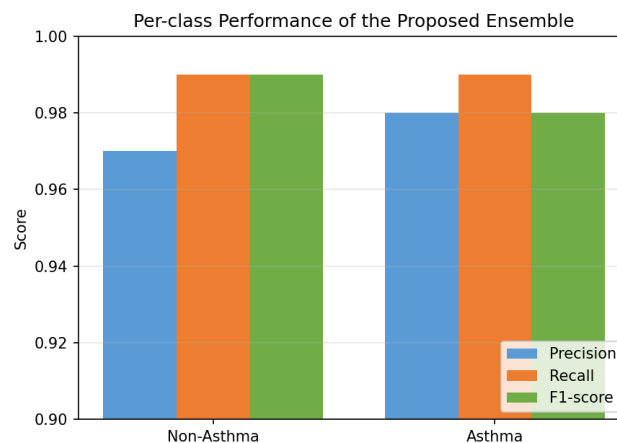


Figure 3. Per-class precision, recall, and F1-score of the proposed ensemble

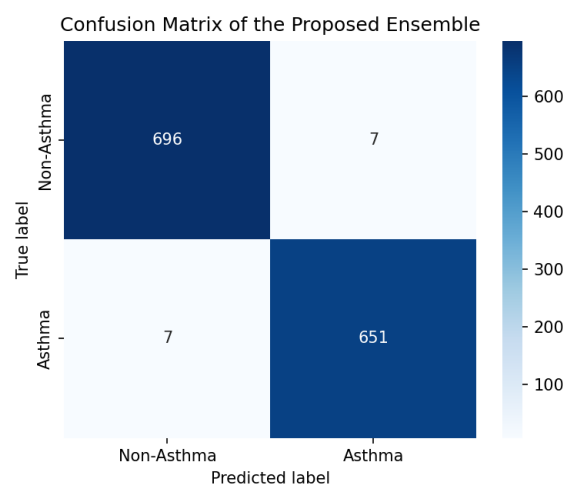


Figure 4. Confusion matrix of the proposed boosting ensemble

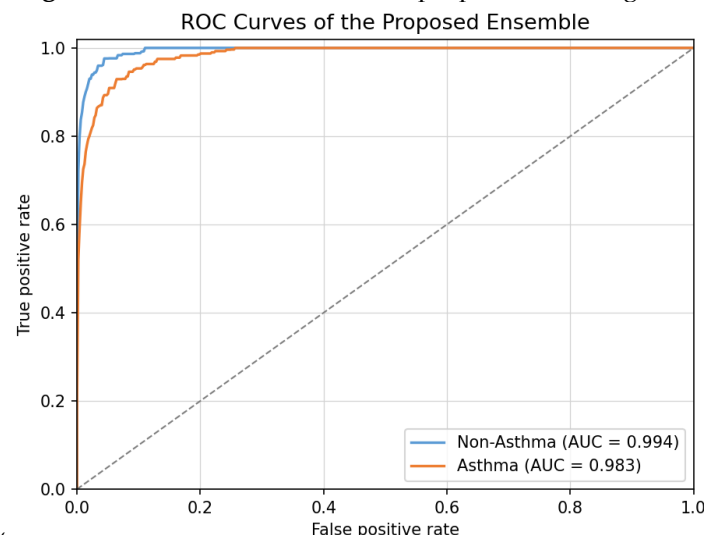


Figure 5. ROC curves of the proposed boosting model

4.3. Baseline Model Performance

To have a fair comparison of the boosting ensemble with the other classifiers, five baseline classifiers were first developed. Precision, recall, F1-score and overall accuracy were used to evaluate their test-set performance. Class-wise evaluation results are presented in Table 6. The test-set accuracy for the five base models was between 84% and 98%. The highest accuracy is 98% for XGBoost, SVM and RF models with 96% and 97% each and the least accuracy is of KNN model at 84%. To further illustrate, the results in the class-wise Table 6 show significant variation in the predictive consistency of the baseline classifiers.

SVM gave fairly good performance for both classes with F1-scores of 0.95 and 0.96, respectively. The performance of RF was also found to be very consistent with class-wise performance with an F1 score of 0.96 for both the classes. XGBoost achieved the highest accuracy with F1-scores of 0.98 for the two classes, non-asthma and asthma, showing that the accuracy of this model was not significantly different between the two outcome classes. The F1-scores of 0.94 and 0.92 for non-asthma and asthma, respectively, show a moderate decrease in accuracy when applied to the task of asthma classification. The performance of KNN was the poorest with F1-scores of 0.80 for non-asthma and 0.88 for asthma. It has weak discrimination power between the two classes of outcomes as measured by lower precision and recall for the non-asthma class. Relatively small differences in class-wise metrics of SVM, RF and XGBoost imply that there is no strong dominance of either class. The results here set the benchmark of the individual baseline of XGBoost very well and give a clear reference to see if the proposed ensemble can improve performance.

Table 6. Class-wise and overall performance of the five baseline classifiers on the test set

Model	Class	Precision	Recall	F1-score	Acc.
Support Vector Machine	Non-Asthma	0.98	0.93	0.95	0.96
Support Vector Machine	Asthma	0.93	0.99	0.96	
Decision Tree	Non-Asthma	0.98	0.90	0.94	0.92
Decision Tree	Asthma	0.91	0.94	0.92	
Random Forest	Non-Asthma	0.96	0.97	0.96	0.97
Random Forest	Asthma	0.95	0.98	0.96	
K-Nearest Neighbors	Non-Asthma	0.81	0.80	0.80	0.84
K-Nearest Neighbors	Asthma	0.88	0.89	0.88	
XGBoost	Non-Asthma	0.99	0.98	0.97	0.98
XGBoost	Asthma	0.97	0.98	0.98	

5. Discussion

This work shows how boosting-based ensemble learning may be used to predict asthma by combining clinical, environmental, demographic, and lifestyle factors. Strong predictive potential was demonstrated by the suggested XGBoost, LightGBM and CatBoost ensemble's 99% accuracy on the test set. These results demonstrate that including a variety of predictor factors can improve prediction accuracy and support the use of sophisticated boosting techniques for modeling the complex nature of asthma.

5.1. Comparison with Baseline models

The proposed ensemble also consistently outperformed all individual baselines. The best baseline accuracy was the XGBoost with 98% followed by SVM with 96%, DT with 92 and KNN with 84%. Therefore, the ensemble improved the score by 2 percentage points compared to the best individual learner, and 3–10 percentage points above the other baseline learners. The relatively good performance of XGBoost further validates the appropriateness of boosting for the feature space and the incremental improvement of XGBoost, LightGBM, and CatBoost demonstrates that their complementarity in learning brings better predictive power than a single machine learning model. The proposed ensemble's accuracy in comparison to the baseline classifiers is shown in Table 7.

Table 7. Comparison of model accuracy

Model	Accuracy
K-Nearest Neighbors	84%
Decision Tree	92%
Support Vector Machine	96%
Random Forest	97%

XGBoost	98%
Proposed Boosting Ensemble	99%

5.2. Comparison with Previous Published Studies

The proposed framework also exhibits better performance than the more recent boosting based frameworks in the literature. The accuracy of asthma severity prediction model presented in [22] is 98% by using XGBoost and in [30] is 93% by using CatBoost with feature selection and class balancing method. Similarly, in [31] the class weighted CatBoost model gave an accuracy of 90.3% and F1 score of 0.915. The Gradient Boosting Machine integrated in GIS in [23] was dedicated to the estimation of the risk of developing asthma, with an R^2 of 0.76, so it is not a convenient comparison with the classification models. The proposed ensemble outperformed these studies consistently, indicating that the combination of boosting complementary algorithms is better than using a single boosting model. The results showed that ensemble boosting not only increases the predictive robustness but also increases the ability to capture more complex patterns related to the development of asthma.

Table 8. Comparison with previous published studies

Ref.	Methodology	Dataset(s)	Key Results	Limitations
[22]	XGBoost-based asthma severity prediction	Air pollutant + demographic data, Uttarakhand, India	Accuracy = 97%	Evaluated a single boosting algorithm
[23]	Random Forest, GBM, XGBoost with GIS	1,473 patients + air quality/weather/socioeconomic, Tehran	GBM achieved $R^2 = 0.76$	Limited to Tehran-specific data
[30]	CatBoost ensemble with feature selection	Clinical asthma dataset	Accuracy = 93%	Focused on asthma attack prediction
[31]	Class-weighted CatBoost ensemble	AIRS Kaggle dataset (clinical + acoustic features)	Accuracy = 90.3%, F1 = 0.915	Dependence on acoustic inputs limits clinical applicability

5.3. Limitations

A large asthma dataset was preprocessed and evaluated under a standardized processing and evaluation environment to develop and assess the proposed boosting ensemble. The model showed good predictive ability, but testing was performed on data sets from various healthcare institutions and geographical areas would provide stronger evidence of the model's generalizability to varying populations. This study employed structured patient records, while the inclusion of longitudinal health records, alongside continually updated environmental data, may offer a better picture of asthma over time. The above factors do not reduce the effectiveness of the proposed approach and can be used for increases in further research and validation of the approach.

6. Conclusion

In this study, a boosting ensemble framework for asthma prediction was developed based on environmental, demographic, lifestyle and clinical factors. The ensemble also performed well, with precision, recall, and F1-scores of 0.98–0.99 for the asthma and non-asthma classes, indicating strong and stable classification results. The results suggest that there is benefit in using complementary boosting learners as a way of increasing the predicting power compared to the individual learners and also of capturing complex relationships between the asthma related factors. These findings are limited to a synthetic benchmark and should be read as a methodological comparison of classifiers, not as evidence of clinical performance.

Future studies could confirm the framework with larger, multi-center datasets to determine the framework's generalizability across different populations. The use of real time environmental factors like air quality, weather etc. Can further enhance the predictive performance. The incorporation of explainable AI techniques would also make the models more transparent, and validation on real, multi-centre patient cohorts would be a prerequisite for any future clinical evaluation.

Reference

1. Kim, T.H., et al., Global burden of asthma among children and adolescents with projections to 2050: a comprehensive review and forecasted modeling study. *Clinical and experimental pediatrics*, 2025. 68(5): p. 329.
2. Oh, J., et al., Global, regional, and national burden of asthma and atopic dermatitis, 1990–2021, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *The lancet respiratory medicine*, 2025. 13(5): p. 425-446.
3. Yuan, L., et al., Global, regional, national burden of asthma from 1990 to 2021, with projections of incidence to 2050: a systematic analysis of the global burden of disease study 2021. *EClinicalMedicine*, 2025. 80.
4. Xu, J., et al., Climate change, air pollution and chronic respiratory diseases: understanding risk factors and the need for adaptive strategies. *Environmental health and preventive medicine*, 2025. 30: p. 7-7.
5. Soler-Segovia, D., et al., Consequences of climate change for asthma: reviewing the evidence. *Current Opinion in Environmental Science & Health*, 2025. 46: p. 100630.
6. Burbank, A.J., Climate change and the future of allergies and asthma. *Current Allergy and Asthma Reports*, 2025. 25(1): p. 20.
7. Cilluffo, G., et al., Machine learning approaches for asthma exacerbation predictions: a systematic review. *Artificial Intelligence Review*, 2026. 59(4): p. 122.
8. Ojha, T., et al., Exploring machine learning applications in pediatric asthma management: scoping review. *Jmir Ai*, 2024. 3: p. e57983.
9. Reddy, S.S. and P. Pramila. Assessment of asthma risk prediction using random forest compared over accuracy of support vector machine. in *AIP Conference Proceedings*. 2025. AIP Publishing LLC.
10. Bağcı, M.F., et al., Detection and prediction of real-world severe asthma phenotypes by application of machine learning to electronic health records. *Journal of Allergy and Clinical Immunology: Global*, 2025. 4(3): p. 100473.
11. Kaur, A., et al. Revolutionizing Asthma Control Using Random Forest Machine Learning Enhancement Technique. in *2024 International Conference on Integration of Emerging Technologies for the Digital World (ICIETDW)*. 2024. IEEE.
12. Lisik, D., et al., Machine learning-derived asthma and allergy trajectories in children: a systematic review and meta-analysis. *European Respiratory Review*, 2025. 34(175): p. 240160.
13. Alhaj, Z. and R.H. Green, Update in Asthma Management: A Narrative Review of Recent Advances, Challenges, and Future Directions. *Clinical Medicine*, 2026: p. 100623.
14. Budiarto, A., et al., Machine learning-based asthma attack prediction models from routinely collected electronic health records: systematic scoping review. *Jmir Ai*, 2023. 2: p. e46717.
15. Druvo, M.M.R., et al., Explainable artificial intelligence-driven ensemble learning for asthma risk prediction using machine and deep learning. *Journal of International Medical Research*, 2026. 54(4): p. 03000605261440309.
16. Jayamini, W.K.D., et al., Data-efficient machine learning approach for predicting asthma attack risk. *Scientific Reports*, 2026.
17. Budiarto, A., et al. Handling class imbalance in machine learning-based prediction models: A case study in asthma management. in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2023. IEEE.
18. Darsha Jayamini, W.K., et al., Investigating machine learning techniques for predicting risk of asthma exacerbations: a systematic review. *Journal of Medical Systems*, 2024. 48(1): p. 49.
19. Chang, M. and Y. Ku, LSTM model for predicting the daily number of asthma patients in Seoul, South Korea, using meteorological and air pollution data. *Environmental Science and Pollution Research*, 2023. 30(13): p. 37440-37448.
20. Özbeyaz, A., M. Yıldırım, and F. Tufaner, Prediction of asthma outpatients using cumulative particulate matter and machine learning algorithms: a case study in adiyaman, Turkey. *Discover Applied Sciences*, 2024. 7(1): p. 19.
21. Kohandel Gargari, O., et al., Assessing the diagnostic accuracy of machine learning algorithms for identification of asthma in United States adults based on NHANES dataset. *Scientific Reports*, 2025. 15(1): p. 4537.
22. Mohammadi, A., E. Pishgar, and J. Aguilera, Spatial prediction of high-risk areas for asthma in metropolitan areas: a machine learning approach applied to Tehran, Iran. *ISPRS international journal of geo-information*, 2025. 14(3): p. 105.
23. Kotlia, P., J. Pant, and M.C. Lohani, Empowering data analysis and machine learning to predict asthma intensity using air pollutants in conjunction with environmental factors. *Journal of Air Pollution and Health*, 2026. 11(2): p. 183-200.

24. He, P., et al., Early prediction of pediatric asthma in the Canadian Healthy Infant Longitudinal Development (CHILD) birth cohort using machine learning. *Pediatric research*, 2024. 95(7): p. 1818-1825.
25. Gu, C., et al., Disentangling childhood asthma risk structure via Bayesian network topology. *Physica A: Statistical Mechanics and its Applications*, 2026: p. 131306.
26. Reis, J.S.d., et al., Predicting asthma hospitalizations from climate and air pollution data: a machine learning-based approach. *Climate*, 2025. 13(2): p. 23.
27. Yorusun, G., et al., Performance analysis of artificial intelligence-based classification models for diagnosing asthma in children. *Pediatric Allergy and Immunology*, 2025. 36(12): p. e70258.
28. Xie, M. and C. Xu, Predicting the risk of asthma development in youth using machine learning models. *PloS one*, 2025. 20(11): p. e0336591.
29. Gunawardana, J., et al., Machine learning approaches for asthma disease prediction among adults in Sri Lanka. *Health Informatics Journal*, 2024. 30(3): p. 14604582241283968.
30. Yadav, S., et al., A novel effective forecasting model developed using ensemble machine learning for early prognosis of asthma attack and risk grade analysis. *Scalable Computing: Practice and Experience*, 2025. 26(1): p. 398-414.
31. Rashid, J., et al., Respiratory sound-based AI screening of asthma and COPD via multi-feature fusion and CatBoost classification. *Scientific Reports*, 2026.
32. Lee, Z.-J., M.-R. Yang, and B.-J. Hwang, A sustainable approach to asthma diagnosis: classification with data augmentation, feature selection, and boosting algorithm. *Diagnostics*, 2024. 14(7): p. 723.