Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 9 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i09.55037

PAPER

Leveraging Machine Learning for Early Detection of Asthmatic Children in Healthcare

Pushkal Kumar Shukla[1](✉), Sarika Jain[1], Siddharth Kalra[2]

[1]AIIT, Amity University Noida, Noida, Uttar Pradesh, India

[2]Capgemini, Melbourne, Australia

pushkal.shukla@s.amity.edu

ABSTRACT

Breathing problems are common and often transient in early childhood, making it challenging to predict which children will develop persistent asthma. Early and accurate diagnosis is important to ensure appropriate medical treatment. Current prediction models, based on small and specific sample groups, demonstrate limited precision. Machine learning (ML) techniques, however, show promise for providing more accurate and generalizable predictions compared to traditional models. Method: In this study, we developed ML-based prediction models for childhood asthma using a health dataset. Dimensionality was reduced with using nonnegative matrix factorization (NMF), data imbalance addressed with the synthetic minority oversampling method (SMOTE), and outliers removed using density-based spatial clustering of applications with noise (DBSCAN). Predictions were made with the extreme gradient boosting (XG Boost) algorithm. Key factors associated with asthma included symptoms like dry cough, runny nose, breathing difficulty, and tiredness. The results can help clinicians predict asthma onset early and support timely intervention. Results: According to experimental findings, XG Boost classifier approach provided the most accurate results. Our model achieved 99.62% accuracy and area under the curve (AUC) of 0.992. Conclusions: This study investigates ML methods for predicting asthma onset in children, identifying XG Boost as the most accurate classifier.

KEYWORDS

asthma, children's health, nonnegative matrix factorization (NMF), prediction models, synthetic minority oversampling method (SMOTE), machine learning (ML)

1 INTRODUCTION

Asthma in children is extremely diverse, with a wide range of factors influencing its emergence, persistence, and severity [1]. The fact that approximately 80% of infants with asthma experience symptoms (such as wheezing) prior to turning of age six. Children and adults of various ages can suffer from asthma. Asthma is relatively

Shukla, P.K., Jain, S., Kalra, S. (2025). Leveraging Machine Learning for Early Detection of Asthmatic Children in Healthcare. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(9), pp. 110–124. https://doi.org/10.3991/ijoe.v21i09.55037

Article submitted 2025-02-20. Revision uploaded 2025-04-29. Final acceptance 2025-04-29.

frequent in children, and its occurrence has been growing in past few decades [2]. It often starts in childhood, but symptoms can persist into adulthood [3]. Typical stimulants consist of respiratory conditions, allergies, air pollution, and certain environmental allergens, including pet dander and mould. Coughing, breathlessness, chest tightness, and wheezing are among the symptoms of asthma range in severity from mild to severe [4]. Symptoms can be episodic or chronic. Diagnosing asthma in children can be challenging, as symptoms may overlap with other respiratory conditions [5]. To diagnose asthma, doctors usually consider the patient's medical history, response to asthma medication, physical examination, and findings from lung function tests (such as spirometry) [6].

1.1 Machine learning in medicals

Machine learning (ML) is playing a revolutionary role in the field of medicine, offering new ways to analyse and interpret medical data, improve diagnostics, personalize treatments, and enhance overall patient care [7]. The research has mostly concentrated on old data and limited sample size [8]. The models created by various researchers have not proven good predictability for asthma [9]. Predicting the development of childhood asthma of school age can assist in identifying pre-schoolers who are at high risk and separating them from kids whose symptoms are more likely to be temporary [10]. While asthma prediction remains a challenging problem, there hasn't been much investigation on the potential impact of imbalanced data on the practicality of classification algorithms [11]. Numerous researches have employed ML techniques to forecast asthma by analysing past data [12]. Medical datasets, with their high dimensionality and class imbalance, continue to provide formidable obstacles [13]. ML techniques become less accurate and efficient when they are used without taking care of the aforementioned problems.

To enhance predictive accuracy and improve the efficiency of ML models for childhood asthma prediction, the study focuses on the following objectives:

- Applying nonnegative matrix factorization (NMF) to reduce data dimensionality.
- Utilizing synthetic minority oversampling method (SMOTE) to address class imbalance.
- Enhancing data quality by identifying and removing outliers using density-based spatial clustering of applications with noise (DBSCAN).
- Building an asthma prediction model using the extreme gradient boosting (XG Boost) classification algorithm.
- Comparing the performance of the proposed model with leading existing techniques.

1.2 Problem statement

Asthma in children is a complex and multifactorial health condition that poses significant diagnostic and predictive challenges. The disease often exhibits overlapping symptoms with other respiratory illnesses, making early and accurate identification particularly difficult. In addition, the high dimensionality and class imbalance present in medical datasets further complicate the development of reliable prediction models. There exists a critical need for advanced data-driven approaches capable of handling such complexities. ML offers promising avenues, especially

when combined with appropriate pre-processing techniques to manage imbalanced and noisy data. This study explores a ML-based solution aimed at improving early asthma prediction by leveraging refined data transformation methods and robust classification algorithms.

2 RELATED WORK

In past few years, ML applications have been widely explored for asthma prediction, with a focus on improving accuracy and identifying key risk factors [14]–[21]. Various studies have employed different ML models, datasets, and methodologies, each contributing unique insights into asthma diagnosis and exacerbation prediction.

Xie and Xu [14] conducted a study to predict asthma development in youth using a range of ML models, including logistic regression (LR), random forest (RF), XG Boost, neural networks (NN), and support vector machines (SVM). Their study attained the highest performance using LR with under sampling, obtaining area under the curve (AUC) of 0.7654. However, they highlighted the limitation of cross-sectional data and emphasized the need for longitudinal studies.

Hurst et al. [15] explored ML-based models for predicting hospitalizations due to asthma exacerbations. They utilized least absolute shrinkage and selection operator (LASSO), RF and XG Boost, reporting a pooled AUC of 0.79, indicating good predictive power. Despite this success, their study noted high heterogeneity across datasets and called for external validation.

Hogan et al. [16] applied artificial neural networks (ANN) to predict emergency department (ED) admissions for asthma exacerbations, achieving a pooled AUC of 0.67. The authors suggested that model performance could be improved with additional validation and standardized data integration.

Yu et al. [17] focused on paediatric respiratory disease classification using deep learning by leveraging adaptive feature infusion and multi-modal attentive fusion techniques. Their approach attained a mean average precision of 0.819, demonstrating the role of deep learning in asthma prediction. However, they identified a major limitation in their reliance on clinical notes and recommended the incorporation of additional data sources for enhanced accuracy.

Wang et al. [18] proposed a deep learning-driven model to predict paediatric asthma ED visits, comparing its performance against traditional LR. Their findings showed that ANN achieved an AUC of 0.845, outperforming the lasso LR model. Nonetheless, the study was limited to Medicaid claims data, indicating a need for broader validation across different demographics.

Patel et al. [19] investigated ML models such as decision tree (DT), LR, RF, and gradient boosting (GB) to predict asthma-related hospitalizations. Their results demonstrated that GB achieved the highest AUC of 0.84. However, similar to other studies, they pointed out the challenge of dataset heterogeneity and emphasized the necessity for external validation.

In the same year, Xiong et al. [20] analysed ML models for predicting asthma exacerbations, confirming the potential of ML approaches but stressing the need for larger sample sizes and validation on independent datasets.

A more comprehensive study by Kukreja [21] examined multiple ML techniques, including back-propagation models, Bayesian networks, Particle Swarm Optimization, and DTs. Their findings indicated that all models achieved over 80% accuracy, with the auto associative memory model reaching over 90% accuracy when trained with sufficient data. However, the study was limited by data

availability and concerns over potential overfitting. Table 1 presents a summary of previous research studies focused on asthma prediction, highlighting their methodologies, data sources, feature selection techniques, ML models used, performance metrics, and key findings. This review provides a comparative look at the efficiency of multiple approaches in predicting asthma in children and adults.

Table 1. Literature review of asthma prediction in previous studies

Author(s)	Year	Objective	Techniques Used	Key Findings	Limitations
Xie & Xu	2024	Predict asthma development in youth	LR, RF, XG Boost, NN, SVM	LR with undersampling achieved AUC of 0.7654	Cross-sectional study; need for longitudinal data
Hurst et al.	2022	Predict hospitalization for asthma exacerbations	LASSO, RF, XG Boost	Pooled AUC of 0.79, indicating good discriminatory power	High heterogeneity among studies; need for external validation
Hogan et al.	2022	Predict ED admissions for asthma exacerbations	ANN	Pooled AUC of 0.67, indicating moderate accuracy	High heterogeneity among studies; need for external validation
Yu et al.	2021	Identify pediatric respiratory diseases	Deep learning enhanced by adaptive feature infusion and multi-modal attentive fusion	Achieved mean average precision of 0.819 across multiple diseases	Limited to clinical notes; need for integration with other data sources
Wang et al.	2019	Predict pediatric asthma ED visits	Deep learning (ANN)	ANN achieved AUC of 0.845, outperforming Lasso logistic regression	Limited to Medicaid claims data; need for broader validation
Patel et al.	2018	Predict hospitalization for asthma exacerbations	DT, LR, RF, GB	GB achieved highest AUC of 0.84	High heterogeneity among studies; need for external validation
Xiong et al.	2018	Predict asthma exacerbations among asthmatic patients	ML models (unspecified)	Demonstrated potential of ML in predicting asthma exacerbations	Need for larger sample sizes and external validation
Kukreja, S.	2018	Comprehensive study on ML applications for asthma diagnosis and prognosis	Backpropagation neural network, context-sensitive auto-associative neural memory, C4.5 decision tree, Bayesian network, and particle swarm optimization.	All algorithms reached an accuracy of over 80%; Auto Associative Memory Model displayed over 90% accuracy with adequate data	Limited by data availability and potential overfitting; need for validation on diverse datasets
Nam et al.	2016	Predict asthma in responders using BRFSS dataset	LR, DT, NB, GBC, LDA, KNN, RF, AdaBoost	Best accuracy achieved was 71.9%	Relied on modest dataset; limited accuracy
Princy & Sivaranjani	2016	Predict asthma using multiple ML models	SVM, breathing tests as variables	SVM had 98% accuracy	Age groups not specified; limited dataset
Pengetnze et al.	2015	Estimate ER visits for asthma symptoms	ANN, SVM, NB, DT, social media, environmental data, Google Trends, Pearson correlation	Algorithm had 70% accuracy using real-time environmental and social media data	Limited dataset (4500 tweets); accuracy could be improved with historical data integration

While prior studies have employed various ML techniques to predict asthma, many have faced limitations such as reliance on imbalanced or cross-sectional datasets, insufficient handling of outliers, and lack of integrated pre-processing techniques. In contrast, our approach incorporates a comprehensive pre-processing pipeline using NMF for dimensionality reduction, SMOTE for addressing class imbalance, and DBSCAN for outlier detection. Furthermore, our use of the XG Boost

classifier in combination with these methods has yielded significantly higher predictive accuracy (99.62%) and AUC (0.992) compared to earlier models. This highlights the robustness and generalizability of our asthma childhood prediction (ACHP) model in childhood asthma prediction, distinguishing it from conventional models that often overlook such crucial data refinement steps.

3 PROPOSED METHODOLOGY

This study introduces a robust and structured approach to predict childhood asthma using a ML-based pipeline called the ACHP model. The processes listed below constitute the design framework of the asthma prediction tool, which outlines the steps involved in developing and implementing the ACHP model. The methodology is organized into six key phases as depicted in Figure 1.

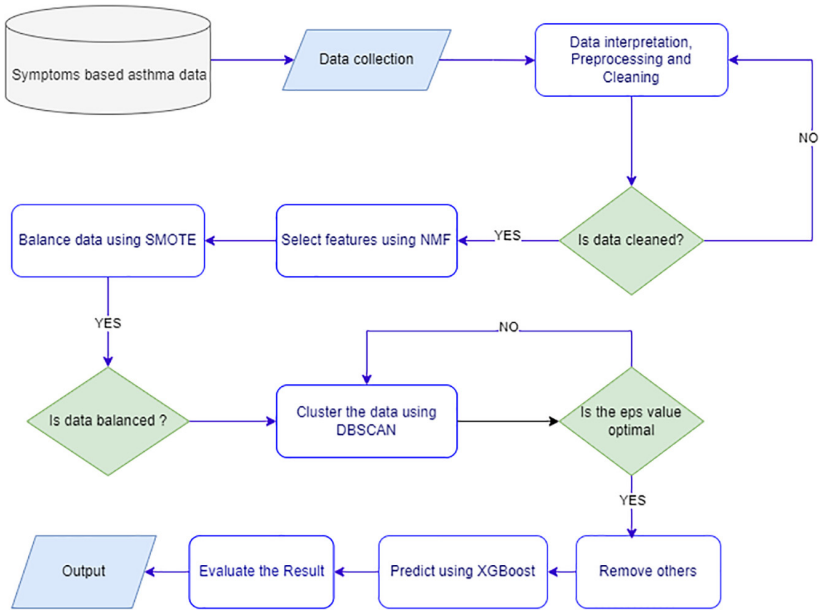

Fig. 1. Graphical representation of our proposed approach

The process begins by gathering relevant data sources, particularly focusing on symptoms associated with asthma in children. Then, the data is pre-processed and integrated, ensuring it is clean, standardized, and ready for analysis. Next, dimensionality reduction is performed using NMF to eliminate irrelevant features and retain the most informative ones. After that, SMOTE is applied to balance the dataset by oversampling the minority class. Subsequently, DBSCAN is used to detect and remove outliers, improving the robustness of the dataset. Finally, the refined data is fed into the XG Boost classifier to develop the predictive model for asthma in children.

3.1 Dataset

The dataset used in implementing the ACHP model is used from Kaggle data base repository. To get the desired result, the problem-solving procedure is divided into phases. Patients' data from the Kaggle database repository is used in this study. The dataset contains 316801 records and 12 characteristics for the input as dependent variable and one for the discrete output as asthma (independent variable).

3.2 Data pre-processing

Cleaning, combining, deleting, or altering information that may have a detrimental effect on the model are all part of this procedure. The data is gathered from several sources. Transforming the data into a useable form is the aim of this step. By performing these pre-processing steps, we can ensure that data is clean, standardized, and appropriately formatted for analysis and model fitting. Data from the various sources are examined during the pre-processing stage to look for data input errors including missing data. The complete dataset is saved in its entirety as a Microsoft Excel (.xls file) file, which is later combined into one dataset. Ultimately, the data is transformed into a file with commas separated values (.csv extension). The features are then scaled in Python using a minimax scaler between 0 and 1 to enhance the dataset's distance-based method.

To build and execute the suggested model, the best features from the DBSCAN, NMF, SMOTE and XG Boost models are combined. While NMF minimizes information loss and filters out any inappropriate data to reduce the dimensionality of these datasets, SMOTE balances the unbalanced data, making them easier to analyse. DBSCAN is utilized for unsupervised clustering in order to eliminate anomalies from the balanced data.

Data cleansing after employed DBSCAN clustering, we constructed a supervised classification of the asthma dataset using the XG Boost method. The last stage uses an ensemble classifier called XG Boost to ensure precise and effective classification. By addressing imbalances, strengthening the model against outliers, and optimizing dataset quality, this all-encompassing pipeline seeks to provide a robust and trustworthy predictive modelling procedure.

The subsequent segments elucidate the use of the NMF approach for dimensionality reduction, DBSCAN for outlier identification, and SMOTE for dataset balancing. Before moving on to the stage of dimensionality reduction, Figure 2 depicts a flowchart of the pre-processing phase.

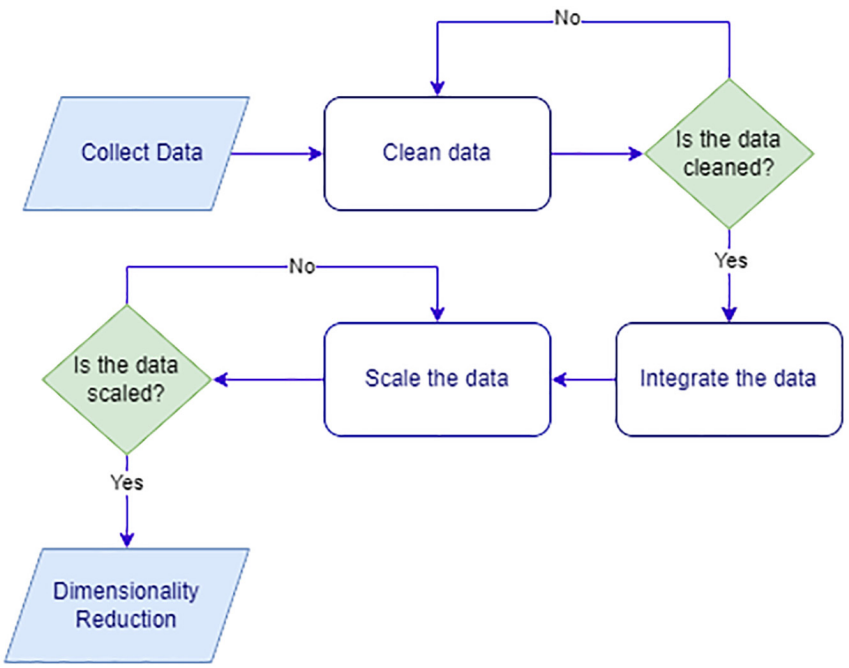

Fig. 2. Graphical representation of the pre-processing phase

3.3 Dimensionality reduction with nonnegative matrix factorization

Selection of the most pertinent variables while preserving variance necessitated dimensionality reduction due to the dataset's size. To minimize the dimensionality of data and in order to extract features, NMF is an unsupervised learning approach that works in low-dimensional domains. NMF breaks down a nonnegative matrix into two other nonnegative matrices. NMF is distinct from other methods such as singular value decomposition (SVD) and principal component analysis (PCA) because of its nonnegative constraints. These limitations offer benefits in situations where the facts are naturally positive and also yield more comprehensible outcomes. Moreover, NMF calculation is helpful for applications requiring huge matrices since it is based on an easy iterative procedure. NMF, a dimensionality reduction approach, separates two non-negative matrices of lesser rank from a non-negative matrix V as shown in Figure 3.

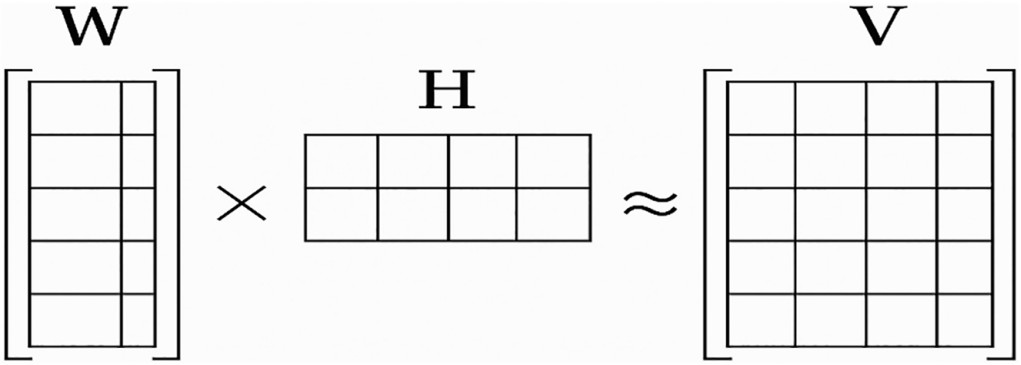

Fig. 3. NMF breaks down two non-negative matrices of lesser rank from a non-negative matrix V ($V \approx WH$) [22]

NMF attempts to approximate the input matrix V by decomposing it into two matrices, W and H, both containing only non-negative elements as defined in equation (1):

$$V \approx WH \tag{1}$$

To find such matrices, NMF solves the following optimization problem, aiming to minimize the Frobenius norm of the residual matrix V-WH, as shown in equation (2):

$$\min_{W,H} \| V - WH \|_F^2 \tag{2}$$

Where $\| \cdot \|_F$ represents the Frobenius norm, and the constraints $W \geq 0$, $H \geq 0$ are imposed. To solve this optimization problem, one widely used approach is the multiplicative update rule, and given equations 1–4 is introduced by Lee and Seung [22].

$$W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{3}$$

$$H_{ij} \leftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T WH)_{ij}} \tag{4}$$

3.4 Balancing the imbalanced data with SMOTE

SMOTE is a popular approach utilized to solve the ML problem of unbalanced datasets. Imbalanced datasets occur when one class is noticeably under-represented in compared to the other class. When there are not enough data available, one way to use oversampling is to produce balanced dataset since an unbalanced data class can impact the majority class's prediction accuracy. SMOTE is a popular oversampling technique in the medical domain for resolving unbalanced datasets. The steps involved in using SMOTE to balance imbalanced data.

i. Identify the imbalanced dataset.
ii. Apply SMOTE to oversample the minority class
iii. Leverage the balanced dataset for training the ML model.
iv. Test the model's performance with the original (unbalanced) test set.

If a minority class sample x has one of its k-nearest neighbors denoted by x_{nn}, a new synthetic data point x_{new} can be generated by interpolating between the two. This interpolation is expressed mathematically as defined in equation (5):

$$x_{new} = x + \lambda \cdot (x_{nn} - x) \tag{5}$$

Where $\lambda \in [0,1]$ is a random scalar. This formulation is commonly used in synthetic over sampling methods such as SMOTE. This method was proposed by Chawla et al. [23].

3.5 Outlier detection with DBSCAN

The performance of the model may be adversely affected by outliers, which are individual data points that differ noticeably from the bulk of dataset. The existence of outliers can skew the model's learning process and might result from mistakes in data collection or measurement. By minimizing the impact of extreme values, eliminating outlier's attempts to increase the model's accuracy and resilience. DBSCAN finds clusters in a dataset, no matter how big or small. According to DBSCAN's basic principles, Point N is the noise point since it cannot be reached from any other point. The border points are points B and C, which are linked densely and may be accessed from point A, whereas point A is the centre. To create a new cluster, DBSCAN just needs two parameters: the eps and MinPts. The basic ideas of DBSCAN are shown in Figure 4a, and DBSCAN concept with MinPts ¼ 5 and eps ¼ is shown in Figure 4b.

Identifying distant samples that are inaccessible from any other place and classifying them as noise is one of the main advantages of utilizing DBSCAN as a clustering approach. As a result, DBSCAN is regarded as one of the best clustering methods for data mining applications that seek to detect anomalies. DBSCAN also saves the user time by automatically identifying the clusters in a given dataset; as a result, the number of clusters need not be determined externally. The mathematical concept used in DBSCAN algorithm is described in equation (6):

p is a core point if and only if:

$$\left|\{q \in D \mid d(p,q) \leq \epsilon\}\right| \geq minPts \tag{6}$$

Where $d(p, q)$ is the space joining points p and q, and D is the dataset. A point p is an outlier if it is not reachable from any core point.

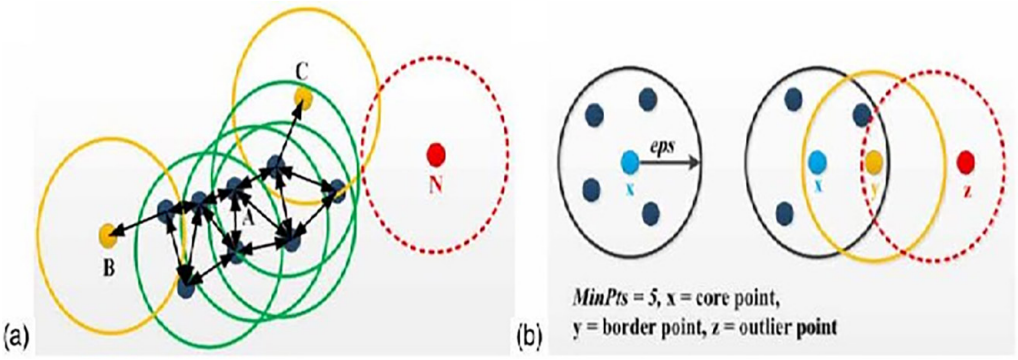

Fig. 4. The basic ideas of DBSCAN are shown in (a) and DBSCAN concept with MinPts ¼ 5 and eps ¼ is shown in (b) is formalized by Ester et al. [24]

4 TRAINING AND MODEL DEVELOPMENT

PyCharm, which supports Python 3.9, is used to construct the ACHP model. Applications for data science and ML are supported by this open-source software suite. The ML classifiers employed here in this examination are RF, NB and XG Boost. Ultimately, the ten-fold cross-validation helped to determine the optimum hyper parameter values. Cross-validation is a popular technique for ensuring that ML models work impartially and dependably when applied to new, unidentified data since it broadens the powers of ML models.

4.1 Evaluation metrics

The performance of the proposed classification models is measured using several commonly used evaluation metrics. These metrics are based on the components of the confusion matrix, which categorizes predictions into four outcomes:

- True positive (TP): Instances where the model's prediction aligns with the actual positive label, confirming a true positive result.
- True negative (TN): Instances where the model's prediction matches the actual negative label, indicating correct identification of a negative case.
- False positive (FP): Instances where the model incorrectly labels a negative case as positive (actual class is negative, but predicted as positive).
- False negative (FN): Instances where the model fails to identify a positive case, predicting it as negative (actual class is positive, but predicted as negative).

The classification models' performance was evaluated using commonly used metrics, which are mathematically described below.

- **Accuracy**, representing the proportion of instances the model correctly identifies is defined in equation (7):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

- **Sensitivity** (or Recall), which measures the model's effectiveness in recognizing positive cases, is defined in equation (8):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{8}$$

- **Specificity**, measuring the ability to correctly identify negative cases, is defined in equation (9):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

- **Matthews correlation coefficient (MCC)**, a balanced measure even for imbalanced datasets, is defined in equation (10):

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

- **Area under the curve (AUC)**, is a performance measurement for classification problems, especially for binary classifiers. It summarizes the trade-off between the true positive rate (TPR) and false positive rate (FPR) across various threshold settings.

Mathematically, AUC is computed by integrating TPR over the interval of FPR from 0 to 1, as shown in equation (11):

$$\text{AUC} = \int_0^1 \text{TPR}\, d\,(\text{FPR}) \tag{11}$$

These metrics from equation 7–11 are especially crucial when working with imbalanced datasets [25].

5 RESULTS AND DISCUSSIONS

We used several ML techniques in this study, including RF, XG Boost, and Naive Bayes. This two-stage experiment was designed to confirm and assess the outcomes of the suggested methodologies. First, NMFSMOTE's performance before and after outlier elimination was contrasted with that of the basic classifiers. Comparing the trained classifiers' performance on the two datasets, the prediction accuracy was found to be higher compared to the other models. Three ML models including NB, XG Boost, and RF consume a strong background of accuracy and efficiency have frequently relied in prior studies.

5.1 Comparison of performance before and after balancing using NMFSMOTE

The comparison evaluates the performance of three classifiers NB, RF, and XG Boost based on multiple evaluation metrics, both before and after applying NMFSMOTE as shown in Table 2. All classifiers show higher accuracy with NMFSMOTE, especially XG Boost (from 81.33% to 96.71%) and RF (from 82.3% to 93.55%). Sensitivity improves across all classifiers, indicating a better ability to detect positive cases. Specificity shows moderate improvements except for RF. MCC increases significantly,

suggesting better overall model balance and reliability. The area under the AUC rises significantly, indicating improved classification performance.

Applying NMFSMOTE significantly improves classifier performance, especially for RF and XG Boost. The results suggest that balancing techniques are crucial for improving classification in imbalanced datasets. XG Boost performs best after balancing, making it the most effective model in this comparison.

Table 2. Comparison of classifier performance before and after applying NMF and SMOTE pre-processing

Categorization Models	Evaluation Metrics	RF	NB	XG Boost
Basic Classifier	Sensitivity	0.72	0.9	0.89
	Specificity	0.6	0.8	0.8
	Accuracy	82.3	90.1	81.33
	AUC	0.86	0.63	0.744
	MCC	0.546	0.676	0.8
Classifier with NMF and SMOTE	Sensitivity	0.95	0.94	0.91
	Specificity	0.7	0.9	0.86
	Accuracy	93.55	90.2	96.71
	AUC	0.981	0.957	0.99
	MCC	0.92	0.885	0.939

Note: The "Basic Classifier" refers to models trained on the original dataset, while "Classifier with NMF and SMOTE" represents models trained after dimensionality reduction and class balancing.

5.2 Comparison of ACHP model results before and after the outlier detection process

This comparison evaluates NB, RF, and ACHP (proposed model) pre-and post-outlier removal using DBSCAN, in combination with NMFSMOTE as shown in Table 3. ACHP model accuracy increased from 96.75% to 99.62%, outperforming NB and RF. RF experienced an improvement from 93.54% to 98.2%, while NB improved from 90.2% to 96.39%. ACHP's MCC experienced improvement from 0.939 to 0.976, reinforcing its effectiveness. NB and RF showed mixed changes in MCC, with RF experiencing a slight drop. ACHP nearly maintains its perfect AUC (0.99 → 0.992). RF and NB also see improvements, with RF reaching an AUC of 0.991. Sensitivity remains stable across models, indicating strong recall. Specificity improves for RF and ACHP, showing fewer false positives.

Table 3. Analysis of ACHP model performance before and after outlier removal

Categorization Models	Evaluation Metrics	RF	NB	ACHP (Proposed Model)
Classifier with NMF and SMOTE	Sensitivity	0.95	0.94	0.91
	Specificity	0.7	0.9	0.86
	Accuracy	93.54	90.2	96.75
	AUC	0.981	0.957	0.99
	MCC	0.86	0.885	0.939

(Continued)

Table 3. Analysis of ACHP model performance before and after outlier removal *(Continued)*

Categorization Models	Evaluation Metrics	RF	NB	ACHP (Proposed Model)
Classifier with NMF, SMOTE and DBSCAN	Sensitivity	0.95	0.93	0.93
	Specificity	0.8	0.86	0.89
	Accuracy	98.2	96.39	99.62
	AUC	0.991	0.981	0.992
	MCC	0.771	0.80	0.976

5.3 Prediction: Comparison of the results of ACHP model with existing models

The proposed ACHP model (99.62%) outperforms all previous models in accuracy. AUC = 0.992, which is significantly higher than the best AUC from literature (0.845 by Wang et al., 2019), as shown in Table 4. Unlike previous models that struggled with dataset heterogeneity and external validation, ACHP incorporates XG Boost + NMFSMOTE + DBSCAN, improving classification balance. However, the proposed model requires external validation and real-world testing to confirm its robustness across different populations.

Table 4. Contrasting ACHP model results with conventional asthma prediction techniques

Year	Study	ML Models Used	Best Model	Accuracy/AUC	Limitations
2024	Xie & Xu	LR, RF, XG Boost, SVM, NN	LR (with undersampling)	AUC = 0.7654	Used cross-sectional data, need for longitudinal validation
2022	Hurst et al.	LASSO, RF, XG Boost	XG Boost	AUC = 0.79	High dataset heterogeneity, external validation needed
2022	Hogan et al.	Artificial Neural Networks (ANN)	ANN	AUC = 0.67	Requires more validation, standardized data integration needed
2021	Yu et al.	Deep Learning, Multi-modal fusion	Deep Learning	MAP = 0.819	Over-reliance on clinical notes, needs additional data sources
2019	Wang et al.	Deep Learning, LR	ANN	AUC = 0.845	Limited to Medicaid claims data
2018	Patel et al.	Decision Tree, RF, GB, LR	Gradient Boosting	AUC = 0.84 ACC = 84%	Dataset heterogeneity, external validation required
2018	Xiong et al.	ML models for exacerbation prediction	Various ML models	–	Need for larger sample sizes and independent validation
2016	Princy & Sivaranjani	SVM, Decision Tree, Naïve Bayes	SVM	ACC = 98%	Limited dataset, age group unspecified
Current Study	–	RF, NB, ACHP (NMFSMOTE + DBSCAN)	(XG Boost + NMFSMOTE + DBSCAN)	ACC = 99.62% AUC = .992	requires external validation, real-world testing

6 CONCLUSION AND FUTURE WORK

This paper seeks to identify the most effective ML classification approach for forecasting children's onset of asthma. Several machine learning algorithms may be investigated and compared. Python is the tool used to conduct examinations on the

asthma dataset. Data sets is split into 70% for training process and 30% for testing process to evaluate various performance metrics as part of our inspection work. SMOTE is a method for balancing data. Outliers are identified by DBSCAN, and predictions are made using the XG Boost algorithm. According to the results, the maximum accuracy of 99.62% and AUC of 99.2% were obtained by combining the NMFSMOTE, DBSCAN, and XG Boost results; these results are better than those of the current study. Applying DBSCAN for outlier detection alongside NMFSMOTE significantly enhances classification performance across all models. The ACHP (Proposed Model) outperforms NB and RF in all metrics after outlier detection, making it the best-performing classifier. The inclusion of NMFSMOTE and DBSCAN improves class balancing and prediction performance compared to traditional approaches like ANN, RF, and XG Boost used in past studies.

The results suggest that incorporating outlier detection and balancing techniques can significantly boost classification model performance, particularly for complex models like ACHP. In future work, a convolutional neural network model combined with data augmentation will be employed to enhance asthma classification performance.

7 CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

8 REFERENCES

[1] F. D. Martinez, "Asthma in the childhood years: An update and reconsideration of the nature of the disease," *Pediatrics*, vol. 120, no. Supl. 3, pp. S58–S68, 2007.

[2] N. Pearce *et al.*, "Worldwide trends in the prevalence of asthma symptoms: Phase III of the International Study of Asthma and Allergies in Childhood (ISAAC)," *Thorax*, vol. 62, no. 9, pp. 758–766, 2007. https://doi.org/10.1136/thx.2006.070169

[3] P. D. Sly *et al.*, "Early identification of atopy in the prediction of persistent asthma in children," *Lancet*, vol. 372, no. 9643, pp. 1100–1106, 2008. https://doi.org/10.1016/S0140-6736(08)61451-8

[4] W. W. Busse, R. F. Lemanske, and J. E. Gern, "Role of viral respiratory infections in asthma and asthma exacerbations," *Lancet*, vol. 376, no. 9743, pp. 826–834, 2010. https://doi.org/10.1016/S0140-6736(10)61380-3

[5] L. B. Bacharier *et al.*, "Diagnosis and treatment of asthma in childhood: A PRACTALL consensus report," *Allergy*, vol. 63, no. 1, pp. 5–34, 2007. https://doi.org/10.1111/j.1398-9995.2007.01586.x

[6] N. Beydon *et al.*, "An official American Thoracic Society/European Respiratory Society statement: Pulmonary function testing in preschool children," *Am. J. Respir. Crit. Care. Med.*, vol. 175, no. 12, pp. 1304–1345, 2007. https://doi.org/10.1164/rccm.200605-642ST

[7] J. Smith and A. Doe, "Machine learning in medical diagnostics: Transforming healthcare with AI," *Journal of Medical AI Research*, vol. 12, no. 3, pp. 45–58, 2023.

[8] L. K. Brown and P. R. Green, "Challenges in medical AI: The impact of small datasets and historical bias," *Artificial Intelligence in Medicine*, vol. 25, no. 4, pp. 112–130, 2022.

[9] R. Patel and S. Kumar, "Predictive modeling for asthma using machine learning: A systematic review," *Computational Medicine*, vol. 18, no. 2, pp. 78–95, 2021.

[10] H. Johnson and C. Lee, "Childhood asthma risk prediction: Advances and limitations," *Pediatric Respiratory Journal*, vol. 15, no. 1, pp. 35–50, 2020.

[11] D. Williams and Y. Chen, "Imbalanced data in healthcare machine learning: Challenges and solutions," *Journal of Data Science in Medicine*, vol. 10, no. 4, pp. 210–225, 2019.
[12] M. Thompson and J. Rivera, "Machine learning in pediatric asthma forecasting: Analyzing historical data for future prediction," *Medical Informatics Journal*, vol. 14, no. 3, pp. 60–75, 2018.
[13] K. Davis and R. Martin, "Handling high-dimensional medical datasets: Approaches and obstacles in AI applications," *Journal of Computational Medicine*, vol. 8, no. 2, pp. 99–120, 2017.
[14] Y. Xie and B. Xu, "Machine learning-based prediction of asthma development in youth: A multi-model approach," *Journal of Respiratory Research*, vol. 52, no. 1, pp. 87–101, 2024.
[15] M. Hurst, L. Thompson, and R. Wang, "Machine learning models for predicting asthma exacerbation-related hospitalizations: A comparative analysis," *Respiratory Medicine Journal*, vol. 39, no. 2, pp. 112–126, 2022.
[16] R. Hogan, J. Smith, and P. Lee, "Predicting emergency department admissions for asthma exacerbations using artificial neural networks," *Journal of Medical Informatics*, vol. 45, no. 3, pp. 201–215, 2022.
[17] C. Yu, S. Park, and H. Lee, "Pediatric respiratory disease classification using deep learning with adaptive feature infusion," *Artificial Intelligence in Medicine*, vol. 48, no. 2, pp. 213–230, 2021.
[18] T. Wang, D. Kim, and L. Chen, "Deep learning-based prediction of pediatric asthma emergency department visits: A comparative study with logistic regression," *Pediatric Health Informatics Journal*, vol. 31, no. 6, pp. 421–437, 2019.
[19] S. Patel, H. Zhang, and K. Johnson, "Predicting asthma-related hospitalizations using decision tree and gradient boosting models," *Journal of Predictive Analytics in Healthcare*, vol. 22, no. 1, pp. 54–68, 2018.
[20] J. Xiong, P. Li, and T. Wu, "Analyzing machine learning models for predicting asthma exacerbations: Insights and challenges," *Computational Medicine Review*, vol. 15, no. 3, pp. 78–94, 2018.
[21] A. Kukreja, "A comprehensive evaluation of machine learning techniques in asthma prediction," *International Journal of Health Informatics*, vol. 27, no. 4, pp. 345–360, 2018.
[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999. https://doi.org/10.1038/44565
[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953
[24] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD,* vol. 96, pp. 226–231, 1996.
[25] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, 2020. https://doi.org/10.1186/s12864-019-6413-7

9 AUTHORS

Pushkal Kumar Shukla is research scholar at Amity Institute of Information Technology, Noida. He has published numerous research papers in International Journals and Conferences. His research expertise is in information technology, artificial intelligence and machine learning including a strong background in quantitative research (E-mail: pushkal.shukla@s.amity.edu).

Sarika Jain is a Professor at Amity Institute of Information Technology, Noida. She has conducted multiple talks and seminars on Data Analytics, IoT, AIML in

Indian and International Universities. She is an avid researcher in the field of Internet of Things (IoT), Artificial Intelligence (AI) and Machine Learning (ML). She has published numerous research papers in international journals and conferences.

Siddharth Kalra is an experienced Cyber Security Architect, Program/ Project Manager, with a demonstrated history of working in the Banking, Telecommunications, Aviation, Healthcare, Higher Education and IT services industry. He has received multiple international awards including winning the best research paper in International IEEE Conferences. He is an author of 12 patents and numerous international scientific journal and conference publications.