Early Asthma Prediction Using XGBoost: A Machine Learning Approach

1st Raufull Islam Rauf
Department of Computer Science
American International University- Bangladesh
Dhaka, Bangladesh
21-45779-3@student.aiub.edu

2nd Nazmul Hasan Emon
Department of Computer Science
American International University- Bangladesh
Dhaka, Bangladesh
21-45829-3@student.aiub.edu

3rd Sadia Afrose
Department of Computer Science
American International University- Bangladesh
Dhaka, Bangladesh
21-45820-3@student.aiub.edu

4th Md. Taufiqur Rahman
Department of Computer Science
American International University- Bangladesh
Dhaka, Bangladesh
22-46116-1@student.aiub.edu

***Abstract*—Asthma is a prolonged respiratory illness that causes the airways in the lungs to swell and tighten, limiting airflow to the human lungs. While asthma currently has no permanent cure, early detection of asthma can prevent long-term lung damage and reduce the risk of life-threatening diseases, such as lung cancer. This study explores the potential of machine learning models for early asthma prediction. Using the Random Forest feature selection technique, key features such as pollution exposure, lung function metrics (FVC and FEV1), physical activity, diet quality, BMI, pollen exposure, sleep quality, dust exposure, and age were identified for model training. Ten machine learning algorithms was applied, with XGBoost delivering the highest precision and demonstrating superior performance across various evaluation metrics, such as precision, recall,F1-score and AUC scores. Also, XGBoost's ability to handle large datasets and complex relationships within the data made it the ideal choice for early asthma prediction. The primary objective of this research is to enhance early asthma prediction, enabling better prevention of asthma-related complications. Ultimately, the findings contribute to improving asthma diagnosis and treatment, improving significant advancements in healthcare.**

***Index Terms*—Asthma Prediction, Early Diagnosis, Healthcare, XGBoost, Polltion exposure, Lung Functions Force Vital Capacity (FVC), Forced Expiratory Volume in 1 Second (FEV1), Inflammation, Area Under Curve (AUC).**

I. Introduction

Asthma is an incurable lung disease that causes inflammation in the airways and makes it difficult to breathe. It is a long-term condition that evolves over time but can be controlled by predicting its symptoms and by maintaining those symptoms. Due to the wide differential diagnosis of common respiratory symptoms and the lack of a standardized diagnostic approach, asthma is often under-diagnosed and under-treated. This study proposes an asthma prediction model to enable early detection and diagnosis of asthma, for improving patient outcomes and preventing future severe complications. In our study features such as lung functions force vital capacity (FVC) and forced expiratory volume in 1 second (FEV1) are introduced. These features are critical for the prediction of asthma, consistent with previous research that emphasizes family history, environmental allergens, and other physiological influences [1], [3]. For medical research and prediction model, feature selection and classification models play a crucial role in accurate prediction. Asthma Traditional diagnostic methods often lack precision, which can exacerbate patient suffering. Machine learning models can improve diagnostic accuracy, can also reduce false positives and negatives, and provide actionable insights for personalized asthma management. Previous studies have demonstrated the effectiveness of ML techniques in diagnosing asthma. For example, Random Forest models have achieved high accuracy in predicting childhood asthma [6], on the other hand, logistic regression has outperformed other models such as XGBoost and SVM in specific contexts, particularly for scenarios with low incidence of events [3]. Although linear regression models have been shown to be effective in identifying asthma severity predictors in pediatric populations [11]. XGBoost (eXtreme Gradient Boosting), an optimized distributed gradient boosting library, offers high flexibility and portability. However, its performance is sensitive to a multitude of hyperparameters, which requires careful tuning and consideration for optimal model outcomes [13]. XGBoost handles datasets with many features effectively like lung functions force vital capacity (FVC) or Dust Exposure interact in non-linear ways including building non-linear relationships and gradient boosting efficiency.

II. Related Works

Asthma is a chronic repository condition that affects millions all over the world, including children. Early diagnosis

and effective management are crucial to improving the patient's health.

Several research studies have demonstrated the potential of the machine learning model in predicting asthma in an early stage, particularly in children. Random Forest, identified as the superior model within an intelligent system design that detects early asthma in children under 10 years of age [1]. Another study shows the effectiveness of the simpler model decision tree in achieving 81% accuracy in early asthma prediction [2]. Machine learning extends beyond early detection of predicting asthma severity and optimization. Utilizing five-factor linear regression model to assess bronchial asthma severity in children of age 6 to 18, showing the utility of regression-based model approaches in predicting severity levels [4]. Another study developed a multi-factorial regression model including biological markers like thymic stromal lymphopoietin (TSPL) levels, highlighting the potential of complex biological interactions [3]. Additionally, research comparing Conditional Inference Tree (CIT) models with traditional risk scores demonstrated the value of personalized predictions [9]. The performance of machine learning models in predicting asthma outcomes depends heavily on selecting the right algorithm and fine-tuning effectively. Several Support Vector Machines (SVM) consistently outperformed traditional models showing the ability of handling complex data patterns [12]. Fine-tuning XGBoost with genetic algorithms significantly improved its accuracy, increasing it from 82.86% to 91.43% [11]. These models shows lower accuracy compared to our XGBoost model, which achieves higher accuracy with random state of 30. In contrast, other implementations of Linear Regression, Decision Tree, Support Vector Machine, Random Forest, XGBoost have achieved lower accuracy than our model highlighting the challenge of achieving consistently high performance due to issues like suboptimal tuning and lower accuracy.

III. Methodology

A. Data Collection

The dataset for this study was collected from Kaggle, containing data on 2392 patients. Includes demographic details, lifestyle factors, environmental exposures, allergy factors, medical history, clinical measurements, and an indicator of asthma diagnosis for all patients [14] . Its ideal for machine learning tasks such as feature importance analysis, classification and the prediction of asthma-related outcomes.

B. Prepossessing Data

Cleaning and Filtering From the data set, irrelevant columns were removed and missing or null values were handled by excluding such entries to maintain data quality. Outlier removed data was filtered to eliminate noise and remove extreme outlier values for increased accuracy also duplicate instances were discarded. Normalization: Continuous variables were scaled to ensure compatibility with various machine learning models.

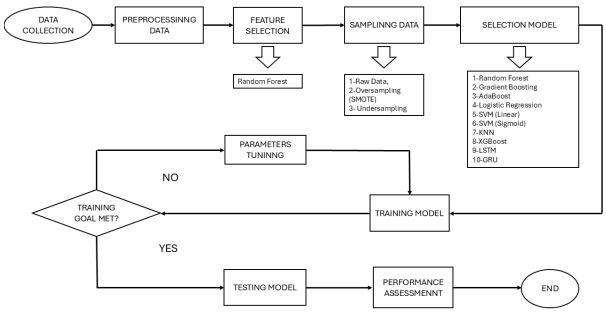

Fig. 1. Methodology Diagram

C. Data Sampling

The dataset shows a huge class imbalance, where 1750 instances labeled as class 0 (no asthma) and only 125 instances labeled as class 1 (asthma). This imbalance posed a challenge for the model, as it could lead models to favor the majority class and perform poorly in identifying the minority class (asthma). To solve this issue, SMOTE (Synthetic Minority Oversampling Technique) was employed to generate synthetic samples for the minority class by interpolating between existing data points, effectively increasing its representation while maintaining diversity in the dataset. Also, random undersampling was applied to the majority class by randomly removing instances to reduce its dominance and achieve a balanced dataset. Both techniques aimed to improve the impact of imbalance on model training and improve classification performance. Finally, the models' performance was compared across three datasets the original imbalanced dataset, the under-sampled dataset, and the SMOTE-over-sampled dataset. This comparison helped assess the effectiveness of these strategies in enhancing accuracy, and the ability to reliably detect asthma cases and the original imbalanced dataset outperformed other sampling datasets.

D. Feature Selection

Two methods were applied for feature selection the Chi-Squared Test and Random Forest Feature Importance. Where the Chi-Squared Test measures the dependency between each feature and the target variable, on the other hand Random Forest provides a more robust evaluation by considering feature importance in the context of an ensemble learning algorithm. Also, Random Forest ranks features based on how much they reduce impurity across multiple decision trees, capturing complex, non-linear interactions between features and the target variable. This makes it more effective in identifying the most predictive features for the model. And its better performance, the Random Forest model, is also highly effective for complex datasets with interdependent variables. Based on this approach, the most significant features identified were Pollution Exposure, Lung Function (FVC and FEV1), Physical Activity, Diet Quality,Body Mass Index (BMI), Pollen Exposure, Sleep Quality, Dust Exposure, and Age shown in the figure 2. Also,

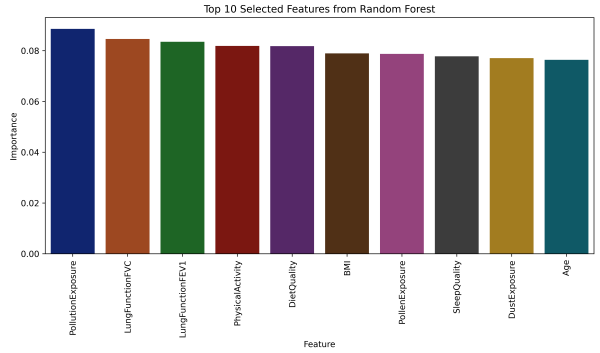

Fig. 2. Selected Features

Random Forests capability to handle multi-col-linearity, its resistance to over-fitting, and its ability to rank features based on their actual impact on predictions made it a superior choice for feature selection in this analysis. That is why Random Forest was chosen for Feature Selection.

’

E. Model Selection and Evaluation

Ten machine learning models were tested on the preprocessed for the dataset: 1. Random Forest 2. Gradient Boosting 3. AdaBoost 4. Logistic Regression 5. SVM (Linear) 6. SVM (Sigmoid) 7. KNN 8. XGBoost 9. LSTM 10. GRU

F. Rationale for XGBoost Selection

After comprehensive evaluations, XGBoost was chosen as the best-performing model because of Boosting Efficiency: Iteratively minimizes errors for complex patterns and imbalanced datasets. Handling Missing Data: Effectively learns from missing values. Feature Importance Insights: Provides detailed predictors for asthma outcomes. Scalability and Speed: Optimized parallel processing enables fast training. Under utilization in Asthma Research: XGBoost’s potential in this domain remains relatively unexplored.

G. Implementation of XGBoost

The selected top 10 features were used for training the XGBoost model, which was evaluated using a traditional 70:30 train-test split with 30 random states and metrics such as Precision, Recall, F1-Score, AUC, MSE, and MAE were recorded for performance comparison for each model. A function for predicting asthma risk based on user input was also implemented where can user give input for these 10 important features Pollution Exposure, LungFunctionFVC, LungFunctionFEV1, Physical Activity, Diet Quality, BMI, Pollen Exposure, Sleep Quality, Dust Exposure After that, the trained XGBoost model will be leveraged to provide a prediction and a confidence score.

IV. Result

After using ten models including machine learning, deep learning and ensemble learning models with different sampling techniques, XGBoost has the highest precision, recall, F1-score and Area Under Curve (AUC) value with minimum Mean Square Error (MSE) and Mean Absoluate Error (MAE). As shown in the Table 1 bellow, XGBoost achieved highest performance having precision = 0.96, recall = 0.96, AUC = 0.61 with MSE = 0.045 and MAE = 0.045.

TABLE I
Models Performance Analysis

Model	Precision	Recall	F1-Score	AUC
Random Forest	0.91	0.95	0.93	0.61
Gradient Boosting	0.91	0.95	0.93	0.52
AdaBoost	0.96	0.96	0.93	0.37
Logistic Regression	0.96	0.96	0.93	0.49
SVM(Linear)	0.96	0.96	0.93	0.55
SVM(Sigmoid)	0.96	0.96	0.93	0.43
KNN	0.96	0.96	0.93	0.54
XGBoost	0.96	0.96	0.93	0.61
LSTM	0.96	0.96	0.93	0.53
GRU	0.96	0.96	0.93	0.56

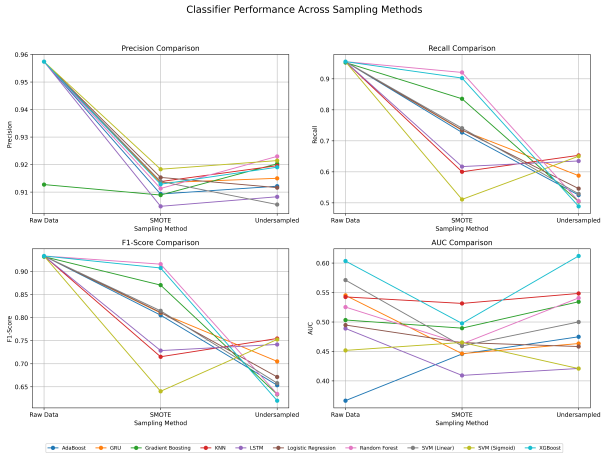

Fig. 3. Performance Analysis

Table I summarizes the performance of various machine learning, deep learning and ensemble learning models applied to the raw data, oversampling and under sampling techniques to evaluated based on key matrices such as precision, recall, F1-score and AUC. Among the models tested, XGBoost achieved the highest AUC = 0.61, demonstrating its superior ability to distinguish between classes compared to other models. XGBoost works as sequentially adding decision trees to an ensemble, where each new tree focuses on correcting the errors made by the previous tree, effectively building a predictive model by combining multiple week learner. However, models like AdaBoost, Logistic Regression and KNN showed slightly lower AUC values than XGBoost ranging between 0.37 to 0.55, regardless of maintaining similar F1-score, precision and recall results shown in the figure 3. These variations shows the limitations of certain models in handling data effectively, particularly when evaluated using AUC as a criterion for predictive model performance. Interestingly, deep learning models such as Long short-term memory (LSTM),Gated recurrent unit (GRU) also demonstrated competitive results with achieving F1-score of 0.93 with fine-tuned parameters selecting random

state of 30. Finally, the results shows the effectiveness of XGBoost in handling complex data and achieving reliable prediction with 96% of accuracy and 98% 99.8% of confidence, making it a strong candidate for further refinement and application in asthma prediction.

V. Conclusion

Findings of this research emphasize the potential of machine learning, deep learning and ensemble learning models in predictive analysis of asthma diagnosis. Among the tested models XGBoost demonstrated superior performance, achieving the highest precision, recall, F1-score and AUC while maintaining lower error rates. XGBoost has the ability to handle complex dataset and sequentially corrects errors highlights its robustness and adaptability. Despite of using other models, including Random Forest, Gradient Boosting and deep learning approaches such as LSTM and GRU also ensemble learning models like AdaBoost, showed competitive performance in certain matrices but fell short in terms of AUC, underscoring the challenges in their ability to differentiate between classes effectively. Our study emphasizes the importance of parameter optimization and fine-tuning in increasing model performance. As demonstrated, even a high-performing model like XGBoost can achieve significant gains through thoughtful optimization. In conclusion, our study reinforces the potential of advanced machine learning techniques in healthcare applications, especially chronic disease management also highlights the immense potential of machine learning to revolutionize healthcare analytics.

VI. Future Work

Researchers could focus on further enhancing these models, integrating real-world clinical data, and exploring their applicability in diverse and dynamic environments to make personalized healthcare a reality. Despite of promising results, the lower-than-expected AUC value suggests areas for improvement for further studies with more balanced dataset.

References

[1] P. Kumar, J. Jain, U. K. Jaiswal, D. Yadav, and B. Solanki, "Childhood Asthma Disease Prediction Using Classification Algorithms of Supervised Machine Learning," in *Proc. 2024 1st Int. Conf. Advanced Computing and Emerging Technologies (ACET)*, 2024, pp. 1–6.

[2] T. Soni, D. Gupta, and M. Dutta, "Machine Learning in Healthcare: Decision Trees for Asthma Risk Prediction," in *Proc. 2024 4th Int. Conf. Sustainable Expert Systems (ICSES)*, 2024, pp. 1211–1214.

[3] O. Pihnastyi, O. Kozhyna, and K. Voloshyn, "Linear Regression Models for Bronchial Asthma Severity Prediction based on TSLP," in *Proc. ICST*, 2023, pp. 276–290.

[4] O. Pihnastyi, O. Kozhyna, and T. Kulik, "Linear Regression Approximate Models for Predicting Severe Course of Bronchial Asthma," in *Proc. ITTAP*, 2022, pp. 55–65.

[5] A. A. H. de Hond, I. M. J. Kant, P. J. Honkoop, A. D. Smith, E. W. Steyerberg, and J. K. Sont, "Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations," *Sci. Rep.*, vol. 12, no. 1, pp. 20363, 2022.

[6] D.-D. Li, T. Chen, Y.-L. Ling, Y. Jiang, and Q.-G. Li, "A methylation diagnostic model based on random forests and neural networks for asthma identification," *Comput. Math. Methods Med.*, vol. 2022, no. 1, pp. 2679050, 2022.

[7] A. Rani and H. Sehrawat, "Role of machine learning and random forest in accuracy enhancement during asthma prediction," in *Proc. 2022 10th Int. Conf. Reliability, Infocom Technol. Optimization (Trends Future Directions) (ICRITO)*, 2022, pp. 1–10.

[8] H. Li, X. Zhang, Q. Zhao, X. Bai, and S. Wang, "[Retracted] Assessment of Clinical Diagnostic Efficacy of Pulmonary Function Test Based on DBN-SVM of Pediatric Asthma and Cough Variant Asthma," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1182114, 2022.

[9] A. H. Owora, R. S. Tepper, C. D. Ramsey, and A. B. Becker, "Decision tree-based rules outperform risk scores for childhood asthma prognosis," *Pediatr. Allergy Immunol.*, vol. 32, no. 7, pp. 1464–1473, 2021.

[10] D. M. Kothalawala *et al.*, "Development of childhood asthma prediction models using machine learning approaches," *Clin. Transl. Allergy*, vol. 11, no. 9, pp. e12076, 2021.

[11] Z. Lin *et al.*, "Predicting environmental risk factors in relation to health outcomes among school children from Romania using random forest model-An analysis of data from the SINPHONIE project," *Sci. Total Environ.*, vol. 784, pp. 147145, 2021.

[12] W. Akbar *et al.*, "Predictive analytics model based on multiclass classification for asthma severity by using random forest algorithm," in *Proc. 2020 Int. Conf. Electrical, Commun., Comput. Eng. (ICECCE)*, 2020, pp. 1–4.

[13] J. Chen, F. Zhao, Y. Sun, and Y. Yin, "Improved XGBoost model based on genetic algorithm," *Int. J. Comput. Appl. Technol.*, vol. 62, no. 3, pp. 240–245, 2020.

[14] R. E. Kharoua, "Asthma Disease Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/dsv/8669080.