JURNAL EMACS
e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.7 No.3 September 2025: 337-342 DOI: 10.21512/emacsjournal.v7i3.14108

Binary Classification of Asthma for the CAPS Pediatric Dataset in Malawi Using Machine Learning

Jaffarus Sodiq[1*], Syarifah Diana Permai[2]

[1,2]Statistics Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
jaffarus.sodiq@binus.ac.id, syarifah.permai@binus.ac.id

*Correspondence: jaffarus.sodiq@binus.ac.id

Abstract *– Childhood asthma poses a significant public health challenge, especially in low- and middle-income countries. An early intervention is essential for effective management and improved prevention of Childhood asthma. This study aims to develop a predictive model for childhood asthma by applying machine learning (ML) techniques. The dataset includes self-reported information on respiratory symptoms, anthropometric measurements, spirometry data, and personal carbon monoxide (CO) exposure among children aged 6–8 years in rural Malawi. We employed a supervised ML approach, focusing on classification algorithms and handling imbalanced outcomes, including Random Forest, Logistic Regression, and XGBoost. Additionally, this study applied the Synthetic Minority Over-sampling Technique (SMOTE), creating synthetic samples of the minority class to balance the distribution of the outcome variable in the training data. Data preprocessing involved handling missing values, feature selection, and normalization to ensure data quality and model performance. Model evaluation was conducted using cross-validation and performance metrics, including precision, recall, and F1-score. Among the evaluated models, Logistic Regression emerged as the most balanced approach, offering strong precision and the highest F1-score while maintaining a reasonable recall rate. This balance reduces the likelihood of overdiagnosis while still capturing a significant proportion of true positives, making it suitable for early screening applications. Moreover, Logistic regression, with its simple mathematical structure, provides more transparency and explainability, which are vital for clinical adoption and gaining practitioner trust.*

Keywords: *classification; lung; asthma; machine learning; child; health; logistic regression; random forest; XGBoost*

I. INTRODUCTION

Asthma is a chronic respiratory condition that poses a significant health concern for both children and adults. Asthma affects around 358 million people, while COPD affects 174 million people worldwide. (Soriano et al., 2017). A comprehensive multicenter study conducted in 2019 found that children with poorly controlled asthma experienced significantly greater disruptions in school life, including frequent absenteeism, limited participation in physical and extracurricular activities, and impaired peer relationships. (ALBANI et al., 2020). On a broader scale, asthma represents a substantial public health challenge, straining healthcare systems worldwide. Recent studies underscore the alarming rise in asthma prevalence globally, with low- and middle-income countries (LMICs) disproportionately affected, bearing approximately 90% of the global asthma-related disease burden. This disparity highlights the urgent need for targeted research and interventions in these regions to mitigate the growing burden of asthma (Mortimer et al., 2022).

Understanding the factors that contribute to childhood asthma is crucial for improving early diagnosis and prevention. Numerous studies have identified key risk factors associated with the development of asthma. BMC Pediatrics

(2025) analyzed data from over 1.5 million children across 164 studies and identified a range of robust risk factors for childhood asthma. Among those parental smoking, premature birth, cesarean section delivery, lack of breastfeeding, family history of asthma, eczema, rhinitis, pet exposure, residing in areas with high traffic density, and early antibiotic or paracetamol use (Zhou & Tang, 2025). These findings underscore the importance of understanding these determinants for enhancing early diagnosis and preventive measures in pediatric populations. Innovative methodologies have been adopted to analyze these complex interactions in recent years. A study conducted in Morocco utilized machine learning models to predict the occurrence of childhood asthma, showcasing the growing recognition of advanced computational tools in identifying high-risk individuals and enabling timely interventions (Jeddi et al., 2021).

Various studies have explored machine learning models to classify and predict asthma based on clinical features, patient data, and voice analysis. A systematic review of 17 studies on machine learning (ML) models using electronic health records to predict asthma attacks revealed considerable variation in defining attacks and challenges like extreme data imbalance (Budiarto et al, 2023). A scoping review of 15 studies from 2019–2023 found that logistic regression and random forests were the most common methods, with recurrent neural networks and XGBoost showing good performance in predicting exacerbations (Ojha et al, 2024). The integration of machine learning (ML) in asthma management shows significant promise for personalized medicine, particularly in predicting acute exacerbations by considering various patient factors such as medical history, biomarkers, and environmental conditions (Molfino et al, 2024).

Classification of asthma into three classes, namely Mild Asthma, Moderate Asthma, and Severe Asthma in Indonesia, using the K-Nearest Neighbor (KNN) algorithm has achieved accuracy values above 90% (Muqarrabin et al, 2025). This high accuracy demonstrates KNN's effectiveness in distinguishing varying severity levels of asthma based on relevant patient attributes and clinical features. Other models like Random Forest and XGBoost also demonstrated high accuracy for asthma classification (Kotlia et al, 2025). These results confirm the effectiveness of machine learning algorithms in supporting accurate and reliable asthma diagnosis. Classification of predominantly allergic and non-allergic asthma has been performed using machine learning models, where the Support Vector Machine (SVM) with a linear kernel function showed the best performance, followed by logistic regression models. The SVM model effectively separates the two asthma categories by finding an optimal hyperplane that maximizes the margin between the classes, resulting in superior classification accuracy compared to other models. Logistic regression also provides good results but typically ranks slightly below the linear kernel SVM in accuracy for this task (Bhardwaj et al, 2023). Machine learning has shown great potential in predicting asthma exacerbations. Recent studies using routine clinical data and blood parameters have developed models such as boosting combined with random forest that achieve high diagnostic accuracy (Chen, et al, 2025).

The rapid advancements in machine learning algorithms have revolutionized data analysis, especially in terms of complexity and accuracy. In healthcare, particularly in the classification of lung diseases, it has gained significant attention due to its ability to improve diagnostic accuracy and predictive capabilities.

Pourhomayoun and Shakibi developed a predictive analytics algorithm using machine learning to assess health risks and mortality prediction in COVID-19 patients. Their study employed several algorithms to predict mortality risk (Pourhomayoun & Shakibi, 2021).

Another study using Long Short-Term Memory (LSTM) networks showed superior performance in predictive analytics. The study highlighted the importance of selecting the right algorithm based on the dataset's complexity and the nature of the classification problem (Bansal et al., 2022).

The growing adoption of machine learning algorithms over traditional statistical approaches is due to their ability to process high-dimensional medical data effectively. Machine learning algorithms have a tuning capability that makes the model quite flexible in capturing patterns (Xie & Xu, 2024).

By integrating traditional epidemiological studies with modern methods, a comprehensive understanding of childhood asthma can be developed. This collaboration enhances risk prediction accuracy and facilitates the design of targeted, evidence-based interventions that promote early diagnosis, particularly in underserved and resource-limited areas.

II. METHODS

The data for this study is derived from the research conducted by Rylance et al. (2019), titled 'Lung Health and Exposure to Air Pollution in Malawian Children (CAPS): a cross-sectional study.' This study assessed respiratory health and air pollution exposure among children aged 6–8 years in Malawi. Data collection involved structured questionnaires to assess respiratory symptoms, anthropometric measurements, personal carbon monoxide (CO) exposure, lung function tests, and air pollution exposure. The study also compared children from households participating in the Cooking and Pneumonia Study (CAPS), which examined the impact of cleaner-burning biomass cookstoves on children exposed to them and those who used traditional cooking methods. This comparison aimed to identify an environmental risk factor that can be modified to prevent or reduce asthma in children, making it a valuable early warning system for communities that rely on traditional cooking methods.

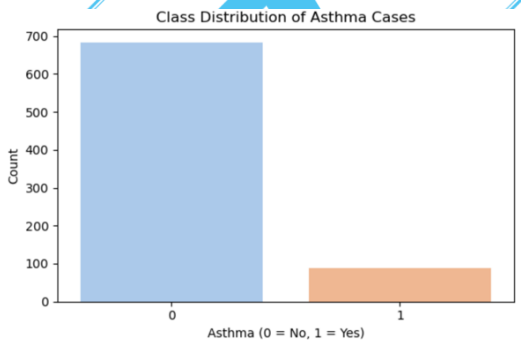

Figure 1. Class Distribution

Subsequent steps focused on feature engineering and selection to refine the dataset for analysis. We generated a new variable that is assumed to have a relevant impact on childhood asthma such as 'missed_school,' which indicates the number of school days missed due to asthma, 'no.adm', which represents the number of hospital admissions due to asthma, and 'no.abx', which signifies the number of antibiotic courses taken due to asthma. Further cleaning involved filtering out records where specific values, such as an outlier category in the 'asthma' variable, could introduce noise into the analysis. Additionally, duplicate variables, if present, were identified and removed to prevent redundancy.

Since the dataset is relatively small and the outcome variable “asthma” is imbalanced (with about 11.5% positive cases), it is essential to address this issue before building a predictive model. Imbalanced data can lead to biased models that tend to predict only the majority class. To handle this challenge, we used machine learning algorithms known for their effectiveness with imbalanced datasets, including Logistic Regression, Random Forest, and XGBoost. Additionally, this study applied the Synthetic Minority Over-sampling Technique (SMOTE). This method creates synthetic samples of the minority class to balance the distribution of the outcome variable in the training data. (Imani et al., 2025). Table 1 summarizes the selected algorithms, their architectures, and hyperparameter tuning configurations.

Table 1. ML Algortihma

Algorithma	Hypertuning Parameter
Logistic Regression	Regularization strength (C) [0.01, 0.1, 1, 10]
Random Forest	n_estimator : [100, 200] max_depth [None,10,20]
XGBoots	n_estimators: [100, 200] max_depth: [3, 6] learning_rate: [0.01, 0.1]

Evaluation metrics are vital for understanding and comparing the performance of classification models, especially in high-stakes fields like healthcare. A standard tool for binary classification is the confusion matrix, which summarizes the model’s predictions against actual outcomes with four components: true positives, true negatives, false positives, and false negatives. These values form the basis for calculating more detailed evaluation metrics. In medical prediction tasks, relying only on accuracy can be misleading, as it does not fully capture how well a model detects

actual cases of a condition (Müller et al., n.d.). Given the severe repercussions of false negatives, this study highlights the use of additional metrics that better reflect the clinical importance of correctly identifying individuals with the condition, thereby ensuring a more reliable and meaningful evaluation of the model.

Table 2. Evaluation Metrics

Matrices	Definition
Recall	True Postive Rate
Precission	Postive predictive value
F1-Score	Balances Recall & Precission
ROC AUC	Overall model performance

III. RESULTS AND DISCUSSION

During the model development phase, we employed three machine learning algorithms, each with hyperparameter tuning, to identify the optimal model architecture for predicting asthma. To address class imbalance in the outcome variable, we used stratified k-fold cross-validation to select the best hyperparameters robustly. Table 3 presents the performance of each model on the testing dataset using the optimal hyperparameter configurations.

Table 3. Experiment Results

Algorithm	Hyper parameters	Precisio/ Recall	F1-Score / ROC
Logistic Regression	C = 0.01	0.55 / 0.611	0.578 / 0.802
XG Boots	max_depth = 6 n_estimators = 100 learning rate = 0.1	0.44 / 0.66	0.53 / 0.81
Random Forest	max_depth = 20, n_estimators = 200	0.52 / 0.55	0.54 / 0.83

All models showed promising results, with overall performance reflected in relatively high ROC AUC scores exceeding 80%. This suggests that the models are reasonably effective at distinguishing between children with asthma and those without. While not perfect, some overlap between the classes remains; their discriminative ability is sufficient for screening or risk stratification.

Among the three models, Logistic Regression achieved the highest precision and the best F1-score balance. Specifically, among children predicted by the model to have asthma, 55% were confirmed cases, which helps reduce the risk of overdiagnosis. Furthermore, the balance between precision and recall (F1-score of 0.578) demonstrates a favourable trade-off, outperforming the other models in this regard.

The XGBoost model also demonstrated potential, with high recall showing its ability to detect most actual asthma cases. However, its very low precision indicates a high rate of false positives, meaning many children identified by the model might not have asthma. The lower F1-score reflects this trade-off when compared with the other models.

Conversely, the Random Forest model had the lowest recall, identifying only about 55% of children who have asthma. In early screening scenarios, recall is important because it is better to catch as many actual positive cases as possible, even if some healthy children are wrongly flagged for further testing. Therefore, a model with higher recall is generally preferred in this context to reduce missed diagnoses.

IV. CONCLUSION

Among the evaluated models, Logistic Regression emerged as the most balanced approach, offering strong precision and the highest F1-score while maintaining a reasonable recall rate. This balance reduces the likelihood of overdiagnosis while still capturing a significant proportion of true positive cases, making it suitable for early screening applications.

Although model selection is crucial in clinical decision-making, the implementation demands careful consideration of factors beyond precision and recall. The practical usefulness of a model depends on how well it integrates into healthcare systems, how easily medical professionals can interpret it, and its computational efficiency for real-time or large-scale use. Logistic regression, with its simple mathematical structure, provides more transparency and explainability, which are vital for clinical adoption and gaining practitioner trust.

Future research should focus on enhancing predictive accuracy by leveraging larger datasets, cost-sensitive analysis, and ensemble approaches that combine the strengths of different algorithms. Additionally, continuous model retraining with the new patient data will be critical to maintaining accuracy and adaptability in diverse clinical environments.

Beyond statistical performance, integration into the health care system must be prioritized to ensure the algorithm's adoption. While logistic regression currently provides the most balanced and clinically relevant performance in this study, advancements in machine learning and broader data availability hold promise for even more effective asthma detection tools.

Author Contributions

This study was compiled, designed, and analyzed by Jaffarus Sodiq and Syarifah Diana Permai. The writing was done jointly. All authors have read and approved the manuscript.

Availability of Data and Materials

The data in this manuscript are confidential and not open to the public. Regarding this consideration, we can not open and show the data in this manuscript.

REFERENCES

Albani, E., Michalopoulos, E., Strakadouna, E., Sakka, A., Triga, E., Saridi, M., Karali, M., & Tzenalis, A. (2020). The impact of asthma on children's school life aged 6 to 12 years. *International Journal of Medical Reviews and Case Reports*, *0*, 1. https://doi.org/10.5455/ijmrcr.asthma-children

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071. https://doi.org/10.1016/j.dajour.2022.100071

Bhardwaj, P., Tyagi, A., Tyagi, S., Antão, J., Deng, Q. (2023). Machine learning model for classification of predominantly allergic and non-allergic asthma among preschool children with asthma hospitalization. *J Asthma*, 60(3), 487 – 195.

Budiarto, A., Tsang, K.C.H., Wilson, A.M., Sheikh, A., Shah, S.A. (2023). Machine Learning–Based Asthma Attack Prediction Models From Routinely Collected Electronic Health Records: Systematic Scoping Review. *JMIR AI*, 2, e46717.

Chen, Y., Sun, J., Chen, Y., Li, E., Lu, J., Tang, H., Xie, Y., Zhang, J.,Peng, L., Wu, H., Cheng, Z. J., Sun, B. (2025). Machine learning-based model for acute asthma exacerbation detection using routine blood parameters. *World Allergy Organization Journal*, 18(7), 101074.

Jeddi, Z., Gryech, I., Ghogho, M., Hammoumi, M. E. L., & Mahraoui, C. (2021). Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. *Healthcare (Switzerland)*, *9*(11). https://doi.org/10.3390/healthcare9111464

Kotlia, P., Pant, J., Lohani, M. C. (2025). Identifying Asthma Risk Factors and Developing Predictive Models for Early Intervention Using Machine Learning. *Biomed Pharmacol Journal*, 18.

Molfino, N.A., Turcatel, G., Riskin, D. (2024). Machine Learning Approaches to Predict Asthma Exacerbations: A Narrative Review. *Adv Ther*, 41(2), 534-552.

Mortimer, K., Reddel, H. K., Pitrez, P. M., & Bateman, E. D. (2022). Asthma management in low and middle income countries: case for change. *European Respiratory Journal*, *60*(3). https://doi.org/10.1183/13993003.03179-2021

Muqarrabin, K. A., Fadlisyah, Safari, T. M. (2025). Classification of Asthma Diseases Using Machine Learning Models at Arun Hospital. *Journal of Advanced Computer Knowledge and Algorithms*, 2(2), 30 – 34.

Ojha, T., Patel, A., Sivapragasam, K., Sharma, R., Vosoughi, T., Skidmore, B., Pinto, A.D., Hosseini, B. (2024). Exploring Machine Learning Applications in Pediatric Asthma Management: Scoping Review. JMIR AI, 3, e57983.

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health*, *20*. https://doi.org/10.1016/j.smhl.2020.100178

Rylance S, Nightingale R, Naunje A, Mbalume F, Jewell C, Balmes JR, Grigg J, Mortimer K. (2019). Lung health and exposure to air pollution in Malawian children (CAPS): a cross-sectional study. Thorax, 74(11), 1070-1077. doi: 10.1136/thoraxjnl-2018-212945

Soriano, J. B., Abajobir, A. A., Abate, K. H., Abera, S. F., Agrawal, A., Ahmed, M. B., Aichour, A. N., Aichour, I., Eddine Aichour, M. T., Alam, K., Alam, N., Alkaabi, J. M., Al-Maskari, F., Alvis-Guzman, N., Amberbir, A., Amoako, Y. A., Ansha, M. G., Antó, J. M., Asayesh, H., … Vos, T. (2017). Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Respiratory Medicine*, *5*(9), 691–706. https://doi.org/10.1016/S2213-2600(17)30293-X

Xie, M., & Xu, C. (2024). *Predicting the Risk of Asthma Development in Youth Using Machine Learning Models*. https://doi.org/10.1101/2024.06.24.24309438

Zhou, W., & Tang, J. (2025). Prevalence and risk factors for childhood asthma: a systematic review and meta-analysis. *BMC Pediatrics*, *25*(1). https://doi.org/10.1186/s12887-025-05409-x