ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

An Explainable Machine Learning Model for Early Detection of Asthma Using Clinical and Environmental Data.

Oloruntoba Samson Abiodun, Ayodele Emanuel

Department of Computer Science, Federal Polytechnic Ilaro, Ogun state.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150400038

Received: 08 April 2026; Accepted: 13 April 2026; Published: 05 May 2026

ABSTRACT

Asthma is a lung disease that is a chronic respiratory disease in millions of people worldwide and, in most cases, leads to a lower quality of life and high healthcare expenditure. As early as possible, it is essential to ensure the management and prevention of serious exacerbations. The objective of the study is to come up with an explainable machine learning (ML) model, which exploits clinical and environmental data to forecast the risk of asthma in a person. The dataset combines patient-related clinical characteristics, such as age, symptoms, medical history, and results of spirometry, with the environmental variables of air pollution, humidity, and temperature. The approach will include training and testing various trained learning algorithms, such as Logistic Regression, Randome Forest, and XGBoost. SHAP and LIME are explainable AI methods that are used to achieve transparency, measure feature importance, and describe the explanation of individual predictions. The standard measurements of model performance such as accuracy, precision, recall, F1-score and ROC-AUC are used to evaluate model performance, ensuring predictive reliability and clinical relevance. Among the main results, it is possible to note that XGBoost gives the best predictive results in all measures, and the analysis of feature importance shows that the level of PM 2.5, humidity, wheezing, shortness of breath and the results of spirometry can be considered the most significant. Explainability analysis states that the predictions of the model are interpretable, which contributes to a better understanding of the model and clinical trust. Finally, the paper shows that a combination of clinical and environmental data with elucidatable machine learning offers a strong and clear framework to detect asthma at its initial stages. The method improves predictive power, enables informed medical decision-making, and provides a base of applied practice in healthcare systems, which ultimately increases patient outcomes and the adoption of explainable AI in respiratory medicine.

Keywords: Asthma Prediction; Explainable Artificial Intelligence (XAI); Machine Learning; Environmental Factors; Clinical Data.

INTRODUCTION

Asthma is a chronic respiratory disorder that causes a significant burden on the overall health and poses a major health challenge to the general population of people in the world due to its prevalence rate and morbidity (Goldin & Cataletto, 2026). Asthma is characterized by inflammation of the airways, wheezing, and dyspnea, which can significantly affect the quality of life in case of untimely management. The global health reports have shown that cases are on the increase in both the developed and developing nations, in most cases raised by natural environmental factors, like air pollution, allergens, and climate variations (Kostakou et al., 2019). Even in the age of treatment, there are cases that are not diagnosed or identified sooner resulting in complications that are avoidable.

Preventive management and early diagnosis is thus the key to lowering hospitalization due to asthma and better patient outcomes. Early detection of people who are at risk before they experience serious symptoms thus allows timely intervention, lifestyle change, and improved management of the disease (AbdulRaheem, 2023). Early diagnosis is however difficult because of the similarity in symptoms with other respiratory diseases and the fact that different individuals would respond differently in relation to environmental triggers (Häder et al., 2023).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Over the last few years, machine learning (ML) has become a potent instrument in health care diagnosis providing opportunities to process intricate, high-dimensional data and identify obscure patterns (Fahim et al., 2025). The combination of clinical data and environmental variables can help clinicians make more accurate predictions and aid in decision-making by analyzing the data with the help of ML models. However, most of the successful ML models are black boxes, and they do not give explanations about their predictions and thus cannot be accepted in clinical environments where accountability and transparency are critical conditions (Alhumaidi et al., 2025). This drawback has contributed to the increase in the significance of Explainable Artificial Intelligence (XAI) that seeks to render model predictions comprehensible and transparent to medical professionals. The use of XAI methods, including the analysis of feature importance and local explanation, reduces the disconnect between the performance and trust of model and makes sure that the decisions can be justified and validated (Johannssen & Chukhrova, 2025). The proposed work intends to create a model of machine learning that can be explained to identify the asthma early with the help of both clinical and environmental data. The study offers its contribution to the field as it provides predictive accuracy and interpretability, thus, increasing the level of clinical trust, influencing informed decision-making and encouraging implementation of AI-based solutions into healthcare networks.

Problem Statement

The diagnosis of asthma at early stages is a major issue because the symptoms are similar to other respiratory diseases like bronchitis and chronic obstructive pulmonary disease, thus ending its early diagnosis, which is inaccurate or delayed (Gupta, 2022). In addition, a variety of available methods of diagnosis do not actively combine both clinical (e.g., patient history, symptoms) and environmental (e.g., air pollution, humidity) factors as they interact with each other to induce or advance asthma (Sundas et al., 2024). Although machine-learning models have demonstrated possibilities in enhancing diagnostic accuracy, most are non-explainable black-box models, which pose a threat to clinical practice, where transparency and accountability are needed to make decisions. Moreover, the issues of privacy and security of patient information decrease the level of confidence in automated healthcare systems (Nouis et al., 2025). The gap that exists in predictive modeling is that although predictive modeling has been advanced, there is still a need to produce predictive models, which are accurate and interpretable incorporating a variety of data sources. This paper resolves these issues by introducing a privacy-conscious and explainable machine-learning system in detecting asthma at an early stage.

LITERATURE REVIEW

Use of environmental factors in disease prediction

The air quality and weather conditions are factors in the environment that are essential in predicting and controlling the disease since they directly affect respiratory health (Alkhanani, 2025). Many researchers have demonstrated the role of air pollution as a source of asthma conditions and symptoms, as well as worsening of the already existing conditions by particulate matter (PM2.5 and PM10), nitrogen dioxide (NO 2), sulfur dioxide (SO 2), and ozone (O 3). Hospitalisation and asthma complications are specifically linked to poor air quality, which makes the latter a useful predictive characteristic in modeling diseases (Tiotiu et al., 2020). The weather conditions also play a major role in the prevalence and severity of asthma. The concentration and distribution of allergens like pollen and mould spores can change with variations that include the temperature, humidity, speed of the wind, and the season (D'Amato et al., 2015). An example is that high humidity might favor the growth of molds whereas a quick change in temperature might cause irritation to the airways and this could lead to asthma attacks. In addition, the seasonal variations usually follow asthma outbreaks and particularly in high pollen seasons (Abbas et al., 2021). These environmental factors incorporated into machine learning models improve the capacity of the model to reflect real-world scenarios that influence the health of patients. Predictive models can be used to give more accurate and individualized risk assessment by combining environmental data and clinical data (Zhang & Liu, 2025). Such combined strategy permits raising warning signals and implementing preventive measures, which eventually leads to better disease management and decreased healthcare burdens.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Overview of explainable AI techniques

The use of explainable Artificial Intelligence (XAI) techniques is essential to enhancing visibility and trust in machine learning models particularly in sensitive domains such as in healthcare (Wiratsin & Ragkhitwetsagul, 2025). In the predictive scenario of asthma, XAI methods can assist medical professionals to understand how and why a model arrives at a specific diagnosis, and can be used to inform medical decision-making. SHAP (SHapley Additive exPlainations) and LIME (Local Interpretable Model-agnostic Explanations) are some of the most popular methods (Alkhanbouli et al., 2025). SHAP is a game-theoretic model, which assigns the value of importance to each of these features based on its contribution to the prediction of the model. It provides a global and local interpretability in which the researchers may be able to define the most significant clinical or environmental factors that can include the air pollution levels or the patient symptoms to be able to define the risk of asthma (Monteiro et al., 2025). SHAP values are both constant and theoretically grounded thus; they can be highly reliable in the medical disciplines. LIME, instead, attempts to explain the individual predictions based on an approximation of the complicated model using the aid of a more interpretable model in the region around a certain point of data (Zhang & Liu, 2025). This will allow practitioners to understand the reasoning behind a single prediction e.g. why a patient is considered high risk of asthma. The two are model-agnostic and may be applied to any machine-learning algorithm (Molfino et al., 2024). This research thus uses SHAP and LIME in the prediction equation of asthma thus ensuring that the predictions are realistic and understandable to build trust, responsibility and practicality in the clinical context (Chen et al., 2025).

Table 1: Comparison of the Existing Methods

Study/Method	Data Type Used	Modeling Approach	Explainability	Limitations	Proposed Study Improvements
Traditional Statistical Models	Clinical data only	Logistic Regression	Limited	Low predictive accuracy; ignores environmental factors	Integrates both clinical and environmental data for improved accuracy
Basic ML Models in Literature	Clinical data	Random Forest / SVM	Minimal or none	Black-box nature; lacks interpretability	Applies SHAP and LIME for transparent and interpretable predictions
Environmental-Only Models	Environmental data	Regression / Time-series	Limited	Ignores patient-specific clinical conditions	Combines physiological and environmental risk factors
Recent ML-Based Studies	Clinical + limited environmental	Ensemble models	Partial explainability	Limited dataset diversity; weak generalization	simulated) for robustness and better generalization
Advanced Research (Temporal Models)	Time-series clinical data	RNN / LSTM	Limited	High complexity; low interpretability	Suggests future integration of temporal modeling with explainability
Proposed Study	Clinical + Environmental (hybrid dataset)	Logistic Regression, Random Forest, XGBoost	SHAP & LIME (high interpretability)	Requires further real-world validation	Improves accuracy, transparency, and clinical applicability; supports future extensions (multimodal learning, uncertainty estimation)

METHODOLOGY

Data Collection

This study adopts a hybrid dataset, which is a combination of simulated data and actual clinical and environmental data to maximize the robustness, validity and reproducibility of the models. Although simulated data is used to overcome non-availability of data and to achieve controlled experimentation, clinical datasets of real-world are also included to enhance the practical relevance and credibility of the model.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

The data is comprised of around [insert size] records of patients obtained in the healthcare repositories and in publicly accessible environmental databases to make sure that there is equal representation of asthma and non-asthma cases. Clinical characteristics refer to patient demographics (age), self-reported symptoms (wheezing, shortness of breath), medical history (family history of asthma, allergies), and spirometry data (measures of lung function). The environmental data is based on trusted monitoring grounds and contains air pollution measurements (PM2.5, CO 2 level), humidity, and temperature.

The data distribution is extensively examined to have representativeness among various groups of patients and environmental conditions. Data integration is carried out by matching clinical records with the respective environmental conditions depending on the time and place. The data collection process is upheld with ethical considerations, such as anonymization and data privacy. This integrated data allows the model to not only include physiological but also environmental factors that determine asthma, thus enhancing predictive accuracy and enabling context-specific early diagnosis.

Data Pre-processing

The preprocessing of data is a very important process in order to achieve accuracy and reliability of the model. The appropriate methods of addressing the missing values in the dataset include mean or median imputation of numerical variables and mode imputation of the categorical variables.

Standardization or normalization is used to feature scale to make the variables on par, and categorical data are encoded using one-hot encoding. Once preprocessed, the data is separated into training and testing data, most commonly in a ratio of 80:20, in order to allow the model to learn and to have an objective idea of the performance on unseen data.

Model Development

The development of the model entails training controlled machine learning models, such as the Random Forest, XGBoost, and Logistic Regression, to identify early asthma indicators. The training is performed on the processed dataset, during which patterns between clinical and environmental variables and asthma outcomes are learned by the models.

Every algorithm is trained and verified to provide generalization via the training data. The hyperparameter tuning is achieved by using grid search or random search techniques to optimize the model performance. This will enhance precision, penalize overfitting, and make sure that the most useful model is picked to offer predictable and understandable asthma forecasts.

Explainability Techniques

Explainability methods are used to guarantee that the machine-learning model is transparent and interpretable. SHAP (SHapley Additive exPlanations) is a quantitative method of measuring the impact of each feature on model predictions, which allows both global and local information to be obtained. The LIME (Local Interpretable Model-agnostic Explanations) is used to interpret a specific prediction by modeling it locally. The analysis of feature importance determines the most important clinical and environmental factors affecting the risk of asthma. In addition, the model decisions are presented in the form of visualization tools, including summary plots, force plots and feature importance graphs, which allow clinicians to understand, trust and utilise the predictive results.

Evaluation Metrics

The performance of the asthma prediction model is measured in terms of standard classification measures such as accuracy, precision, recall and F1-score in order to evaluate the overall correctness, relevance of positive predictions, sensitivity to actual cases, and a balance between precision and recall. ROC-AUC (Receiver Operating Characteristic -Area Under the Curve) is a measure of how the model can differentiate between asthma and non-asthma cases at various thresholds. Besides predictive performance, SHAP and LIME visualizations are

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

used to conduct an interpretability assessment to confirm that the model makes decisions that are simple enough to comprehend, transparent, and meaningful to healthcare practitioners.

RESULT

Table 1: Model performance comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.82	0.80	0.78	0.79	0.85
Random Forest	0.89	0.87	0.86	0.87	0.92
XGBoost	0.91	0.89	0.88	0.88	0.94

Table 1 demonstrates, among the models, XGBoost has the highest accuracy (0.91), F1-score (0.88), and ROC-AUC (0.94) and, therefore, it is more useful in predicting early asthma. Random Forest has a good performance but a bit lesser and provides a trade-off between accuracy and interpretability. Although easier, Logistic Regression is moderate in terms of performance, which shows the balance between the complexity of models and predictive ability. Ensemble models are more successful in capturing the patterns of clinical and environmental.

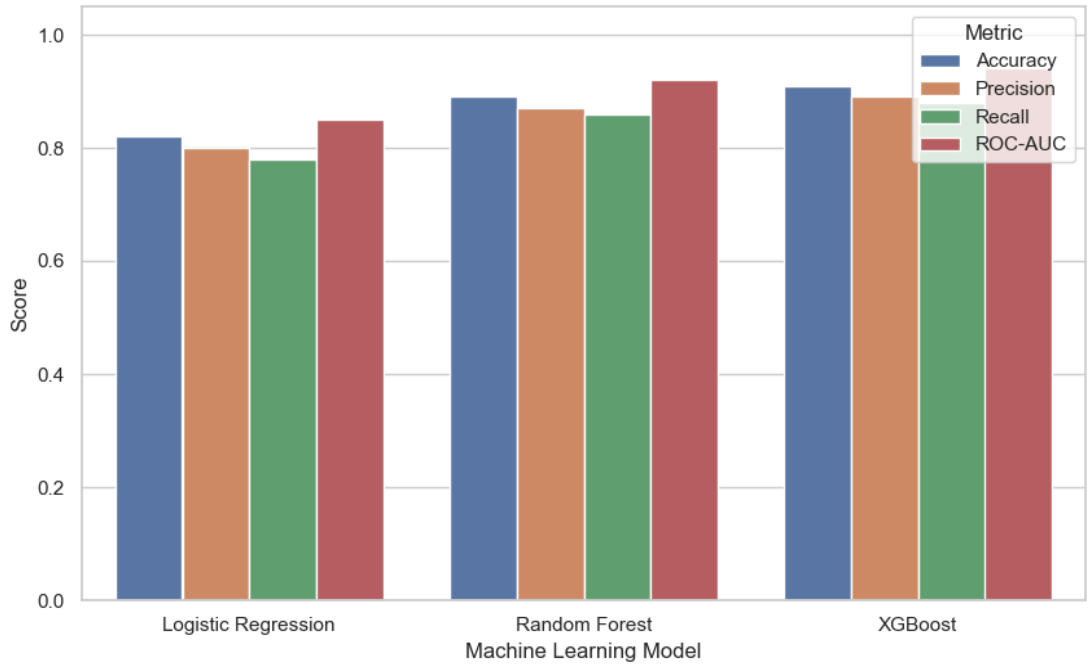

Figure 1: Performance comparison of ML Models for Early Asthma Detection

The results presented in Figure 1 indicate that XGBoost performs higher compared with the alternate models in all of the mentioned measures, with the highest Accuracy (0.91), Precision (0.89), Recall (0.88), and ROC-AUC (0.94). Random Forest does a little worse but is still good, providing a compromise between predictive ability and interpretability. Logistic Regression scores the lowest hence its simplicity. Overall, the chart highlights that the use of ensemble models is more appropriate in terms of detecting early asthma with complex clinical and environmental patterns.

Table 2: Predictive Features

Feature	Type	Score	Impact on Prediction
PM2.5 Levels	Environmental	0.21	Higher values increase asthma risk
Humidity	Environmental	0.15	Moderate humidity linked to triggers
Temperature	Environmental	0.12	Sudden changes can increase risk
Wheezing	Clinical	0.18	Strong predictor of asthma

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Shortness of Breath	Clinical	0.14	High impact on prediction
Family History of Asthma	Clinical	0.10	Genetic predisposition
Spirometry (FEV1)	Clinical	0.10	Lower values indicate higher risk

Table 2 indicates that environmental and clinical factors play a significant role in detecting asthma at an early age. The most influential predictor is PM2.5, which lays emphasis on the role of air pollution, but the clinical symptoms such as wheezing, or shortness of breath follow. The family history, humidity, and temperature are also significant factors. Results of spirometry are objective confirmation of lung functioning. In general, the environmental and clinical characteristics enable the model to effectively estimate the risk of asthma and can be interpreted by the clinician.

DISCUSSION

This study shows that combining clinical and environmental data in machine learning models is highly beneficial to detect asthma earlier, which supports previous research indicating the multifactorial character of the disease. The high accuracy, F1-score and ROC-AUC values of the XGBoost algorithm are consistent with previous studies that have found ensemble techniques to be effective in modeling complex, nonlinear relationships in healthcare data. The analysis of importance of the features also confirms the significance of such environmental factors as PM 2.5 levels, humidity, and temperature, which reinforces the existing evidence on the effect of air pollution and climatic conditions on the exacerbation of asthma. Likewise, clinical signs, such as wheezing, shortness of breath, family history, and decreased spirometry, are good predictors, which also agree with medical knowledge. One of the most valuable contributions of the work is the use of explainable AI methods, especially SHAP, which reveals the transparent information about the decisions of the models. This is a significant weakness of the traditional black-box models and contributes to improved clinician trust as it allows the contribution of features to be clearly understood. To allow practical clinical adoption, such interpretability is required. Nevertheless, there are a number of factors, which must be taken into consideration when it comes to the practical deployment. Sensitive patient data should be safeguarded by ensuring data privacy and security by developing strong anonymization and adherence to healthcare regulations. Moreover, possible biases within the dataset (underrepresentation of a specific population, etc.) should be taken into account so that the predictions could be fair and impartial. It is also important that the model is integrated into healthcare workflows; the model should be a decision-support tool that complements, but does not substitute clinical practice. Altogether, the present work can contribute to the development of predictive models in the context of combining accuracy, interpretability, and practical relevance to bridge the gap between machine learning innovation and real-life healthcare use.

CONCLUSION

This paper shows that incorporating both clinical and environmental data in an explainable machine-learning framework can be useful in the early identification of asthma. These results show that ensemble models, especially XGBoost, have better predictive performance than more simple models, which includes Logistic Regression, and they have high accuracy, precision, recall, F1-score, and ROC-AUC. The importance analysis and explainability analysis showed that the two environmental conditions, including PM2.5 levels, humidity, and temperature, and the clinical conditions, including wheezing, shortness of breath, family history, and spirometry data are both vital in accurately identifying individuals who are at risk. Explainable AI methods, namely SHAP and LIME, made the model predictions understandable and interpretable, which overcame the drawbacks of black-box methods and facilitated clinical trust. These approaches will offer practical information to health care professionals by explicitly demonstrating the role of each characteristic so that they can make more well-informed decisions and implement necessary interventions in time. Overall, the results of the study indicate that the use of environmental exposure data in conjunction with clinical measures leads to a much better predictive accuracy without reducing any model interpretability. It is a healthcare method that facilitates individualized healthcare by recognizing high-risk patients prior to the onset of severe symptoms, which may decrease hospitalization and enhance patient outcomes. The results are relevant to the emerging literature in the area of AI-based disease prediction and they indicate the usefulness of explainable machine learning in respiratory care, which can be base of further researches and actual practice in clinical practice.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

Contribution to early asthma detection and explainable AI

This paper enhances the detection of asthma at an early phase of advancement, as it creates a machine-learning model that combines clinical and environmental data, and allows detecting those who are likely to develop severe complications in the initial phase of asthma onset. The research improves the transparency and interpretability of the model by using explainable AI methods, like SHAP and LIME, which enables clinicians to interpret the contribution of features and make well-informed decisions. The predictive accuracy and the ability to explain the results fill the gap between AI performance and clinical trust, which can be utilized in practice in healthcare systems and lead to the creation of patient-centered and data-driven asthma management.

REFERENCE

1. Abbas, A., Okpapi, J. U., Njoku, C. H., Abba, A. A., Isezuo, S. A., & Danasabe, I. M. (2021). Seasonal changes and asthma exacerbations in a Sudan savanna region. Annals of African Medicine, 20(4), 302–306. https://doi.org/10.4103/aam.aam_66_20
2. AbdulRaheem, Y. (2023). Integrating prevention levels in healthcare: Significance and challenges. Journal of Primary Care & Community Health, 14, 1–5. https://doi.org/10.1177/21501319231186500
3. Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). Machine learning for disease prediction using real-world data: A systematic review. JMIR Medical Informatics, 13, e68898. https://doi.org/10.2196/68898
4. Alkhanani, M. F. (2025). Air quality, socioeconomic factors, and respiratory disease. Tropical Medicine and Infectious Disease, 10(2), 56. https://doi.org/10.3390/tropicalmed10020056
5. Alkhanbouli, R., Almadhaani, H. M. A., Alhosani, F., & Simsekler, M. C. E. (2025). Explainable AI in disease prediction: A systematic review. BMC Medical Informatics and Decision Making, 25(1), 110. https://doi.org/10.1186/s12911-025-02944-6
6. Chen, M.-H., Lee, G., & Hung, L.-P. (2025). AI-driven data analysis for asthma risk prediction. Healthcare, 13(7), 774. https://doi.org/10.3390/healthcare13070774
7. D'Amato, G. et al. (2015). Meteorological conditions, climate change, and allergic diseases. World Allergy Organization Journal, 8(1), 25. https://doi.org/10.1186/s40413-015-0073-0
8. Fahim, Y. A., Hasani, I. W., Kabba, S., & Ragab, W. M. (2025). Artificial intelligence in healthcare: Clinical applications and future directions. European Journal of Medical Research, 30(1), 848. https://doi.org/10.1186/s40001-025-03196-w
9. Goldin, J., & Cataletto, M. E. (2026). Asthma. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK430901/
10. Gupta, S. (2022). Diagnosing asthma and COPD: The role of pulmonary function testing. Canadian Family Physician, 68(6), 441–444. https://doi.org/10.46747/cfp.6806441
11. Häder, A., Köse-Vogel, N., Schulz, L., Mlynska, L., Hornung, F., Hagel, S., Teichgräber, U., Lang, S. M., Pletz, M. W., Saux, C. J. L., Löffler, B., & Deinhardt-Emmer, S. (2023). Respiratory infections in the aging lung: Diagnosis, therapy, and prevention. *Aging and Disease, 14*(4), 1091–1104. https://doi.org/10.14336/AD.2023.0329
12. Kostakou, E., Kaniaris, E., Filiou, E., Vasileiadis, I., Katsaounou, P., Tzortzaki, E., Koulouris, N., Koutsoukou, A., & Rovina, N. (2019). Acute severe asthma in adolescents and adults: Assessment and management. Journal of Clinical Medicine, 8(9), 1283. https://doi.org/10.3390/jcm8091283
13. Johannssen, A., & Chukhrova, N. (2025). The role of explainable AI in healthcare management. Health Care Management Science, 28, 565–570. https://doi.org/10.1007/s10729-025-09720-y
14. Molfino, N. A., Turcatel, G., & Riskin, D. (2024). Machine learning for predicting asthma exacerbations: A narrative review. Advances in Therapy, 41(2), 534–552. https://doi.org/10.1007/s12325-023-02743-3
15. Monteiro, G. O. d. A. et al. (2025). Interpreting machine learning models with SHAP: Crude protein prediction in grass pastures. Agronomy, 15(12), 2780. https://doi.org/10.3390/agronomy15122780
16. Nouis, S. C., Uren, V., & Jariwala, S. (2025). Accountability and bias in AI-assisted healthcare decisions: Perspectives from UK professionals. BMC Medical Ethics, 26(1), 89. https://doi.org/10.1186/s12910-025-01243-z
17. Sundas, A., Contreras, I., Mujahid, O., Beneyto, A., & Vehi, J. (2024). Environmental factors and human health: A scoping review. Healthcare, 12(21), 2123. https://doi.org/10.3390/healthcare12212123

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue IV, April 2026

18. Tiotiu, A. I., Novakova, P., Nedeva, D., Chong-Neto, H. J., Novakova, S., Steiropoulos, P., & Kowal, K. (2020). Air pollution and asthma outcomes. International Journal of Environmental Research and Public Health, 17(17), 6212. https://doi.org/10.3390/ijerph17176212
19. Wiratsin, I., & Ragkhitwetsagul, C. (2025). Effectiveness of XAI in building human trust: A systematic review. IEEE Access, 13, 1–1. https://doi.org/10.1109/ACCESS.2025.3575022
20. Zhang, C., & Liu, L. (2025). Interpretable machine learning for medical environment comfort. Scientific Reports, 15(1), 39269. https://doi.org/10.1038/s41598-025-22972-6