Proceedings of ICBioMed 2024 Workshop: Workshop on Intelligent Medical Data Analysis for Precision Medicine
DOI: 10.54254/2753-8818/50/2024AU0143

Establishment of a predictive model on asthma diagnosis using environmental and allergy factors

Xinqi Huang

Shanghai High School International Division, Shanghai, China

jasperh07@hotmail.com

Abstract. Asthma, being a leading chronic disease globally, involves molecular mechanisms that are quite complex to facilitate easy diagnosis—more especially in the presence of environmental factors, which play an underrecognized role. The current study intends to assess the relationship shared by four allergy risk factors (pollution, pollen, dust, and pets) with the development of asthma. Logistic regression and random forest models were developed using data available from the "Asthma Disease Dataset" sourced from Kaggle to determine the predictive power of these factors. The SMOTETomek technique was applied for data preprocessing due to class imbalance; while relationships and model performance were assessed using the Pearson correlation matrix, box-and-whisker plots, and confusion matrices. Results indicated that pollen exposure is highly predictive of asthma. Dust and pet allergies are negatively associated with an asthma diagnosis. Pollution exposure does not indicate a clear link to asthma. The logistic regression model was less accurate in distinguishing between classes than the random forest model. Therefore, based on accuracy and class distinguishability, the random forest model outperformed logistic regression. These findings contribute to the understanding of environmental and allergy factors in asthma development and underline the feature of such an aspect in predictive modeling. Longitudinal data should be the focus of future studies, along with more objective measures to ensure enhanced reliability for these models.

Keywords: Asthma, Environmental Factors, Allergy Factors, Predictive Models, SMOTETomek.

1. Introduction

Asthma is the most prevalent chronic non-communicable disease globally, affecting around 5 to 10 percent of the world population [1]. It still faces a major dilemma in the diagnosis process due to its complex molecular mechanisms [2]. Characterized by the fluctuating airflow obstruction that leads to shortness of breath and wheezing [1]. A wide range of predictive and diagnostic models based on phenotypic symptoms and bio signals are being constructed, and a very limited amount of such models are based on environmental factors [3, 4].

Moreover, the existing models' performances are not ideal. As described in Daines' meta-analysis using the Prediction model Risk of Bias Assessment Tool, they carry a risk of bias and a great fluctuation in terms of performance [3]. This can easily result in over-diagnosis or under-diagnosis: the former causing unnecessarily costly treatment that can potentially lead to health issues, according to Akker's retrospective analysis, and the latter misguiding treatments to be inadequate and risking harm, as

DOI: 10.54254/2753-8818/50/2024AU0143

illustrated in José's retrospective statistical analysis [5, 6]. Additionally, understanding the factors contributing to asthma and their relative significance statistically is crucial for developing effective prevention and treatment strategies. This makes the development of a model that can predict asthma, especially through the less-utilized field of environmental factors, an imminent need. Meanwhile, allergy refers to various conditions caused by hypersensitivity of the immune system to typically harmless substances in the environment [7]. Various studies discover that allergy risk factors, including environmental factors such as the level of air pollution pollen, dust, and exposure to pets, have a significant relationship with the development of asthma [8-15]. Nevertheless, though individual correlations with the triggering of asthma are determined, a more holistic study focusing on all of the allergy risk factors, analyzing their respective significance coefficients in the model not be conducted.

This study will statistically analyze the correlation between four allergy risk factors – air pollution, pollen, dust, and exposure to pets – and the development of asthma, attempting to establish two predictive models utilizing logistic regression and random forest. This can bring more understanding to the relationship between allergy and asthma, supplementing current progress on the diagnostic models of asthma based on environmental factors, while also providing an indication of the potential risk factors for asthma in everyday life.

2. Methods

2.1. Data set

The dataset obtained for analysis is the "Asthma Disease Dataset" shared by Rabie El Kharoua on Kaggle [16]. This dataset contains 2392 patients' extensive health information, which includes the environmental and allergy factors, medical history, clinical measurements, symptoms, diagnosis information, and confidential information. The variables mainly referenced in this study are the diagnosis information and the environmental and allergy factors, such as the levels of pollution exposure, pollen exposure, dust exposure, and pet allergy.

Within the dataset, the levels of exposure to pollution, pollen, and dust are assessed with a value from 0 to 10, with the lower values indicating less exposure and vice versa. For the pet allergy and diagnosis of asthma, a binary value of either 0 or 1 is given, with 1 indicating the existence of the conditions and 0 meaning the opposite.

2.2. Data preprocessing

To perform data cleaning, a Python script is executed to validate each value within the dataset. The script identifies and removes all rows (representing individual patients) that contain missing values or outliers, ensuring that each variable adheres to its respective criteria. In this case, no rows are removed, indicating that the dataset is complete and robust.

Additionally, due to the issue of class imbalance of cases with versus without asthma diagnosis within the dataset, the Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTETomek) is implemented to boost the size of the minor class. This process will be discussed in detail in the oversampling subsection.

2.3. Data analysis

2.3.1. Descriptive analysis

To acquire a comprehensive understanding of the dataset, a descriptive analysis of the environmental and allergic-related characteristics, along with the patients' diagnostic results, has been performed.

A Pearson correlation matrix is first plotted between the variables of pollution exposure, pollen exposure, dust exposure, pet allergy, and diagnosis. This step aims to assess the relationships between the environmental and allergic variables and to give an overview of their relationship with the result of diagnosis.

DOI: 10.54254/2753-8818/50/2024AU0143

Next, box-and-whisker plots are employed to plot the basic characteristics of the pollution exposure, pollen exposure, and dust exposure data. Since all three variables are measured in the range of 0 to 10 and are continuous, the box-and-whisker plot stands out as a good medium of demonstration. For the binary variables “pet allergy” and “diagnosis,” a bar diagram that shows their respective quantities of 0s and 1s is being plotted, visually demonstrating and comparing the cases with the quality present and absent.

2.3.2. Oversampling

Due to an imbalance of the number of cases diagnosed with asthma and the number of cases not diagnosed with asthma, oversampling is required for the minor class. This is to prevent the majority class bias from taking place: since the diagnostic results would eventually be the dependent variable for the models, by the imbalanced sample, the models may neglect the minor class, giving rise to an accuracy paradox and poor predictive performance for the minor class. This process is done using the SMOTETomek method, which integrates the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek links that remove noisy and borderline instances, resulting in a clean and balanced dataset. The oversampled dataset is then separately saved for future usage in the two regression models. During this process, information for the patient ID and doctor in charge in the dataset is being dropped due to their lack of relevance for the analysis.

2.3.3. Regression models

A logistic regression is first performed, using all the available patient characteristics on the dataset (except for patient ID, the doctor in charge, and diagnosis results) as the independent variables and the diagnosis results as the dependent variable. This ensures the stableness and completeness of the model as compared to only taking the environmental and allergic factors as the independent variables. Three-fourths of the oversampled dataset is being extracted as the training data, which is being used for fitting the logistic regression model, while the rest is used as the test data.

After fitting and testing the logistic regression model, a confusion matrix is presented, along with the F1 score, the recall, and the overall accuracy for all the test cases. A Receiver Operating Characteristic (ROC) curve is also graphed with the Area Under Curve (AUC) shown to further tell how well the model discriminates between the two diagnosis classes. At last, odds ratios for the four environmental and allergic characteristics are calculated from their respective coefficients in the logistic regression model. The four odds ratios are then plotted along with their 95% Confidence Intervals (CIs). The CIs are calculated by adding and subtracting the corresponding z-value of 1.96 multiplied by the standard errors from the odds ratios.

A random forest model is also fitted with the same set of variables and techniques for splitting the dataset applied. Identical parameters are being visually demonstrated except for the odds ratios, which have been replaced by the feature importance values for the four corresponding characteristics within the random forest model.

All of the above processes are being implemented using Python’s sklearn library, along with the pandas library to manage the dataset and imblearn library to perform SMOTETomek. Graphs are plotted using the Seaborn library and the Matplotlib library.

DOI: 10.54254/2753-8818/50/2024AU0143

3. Results

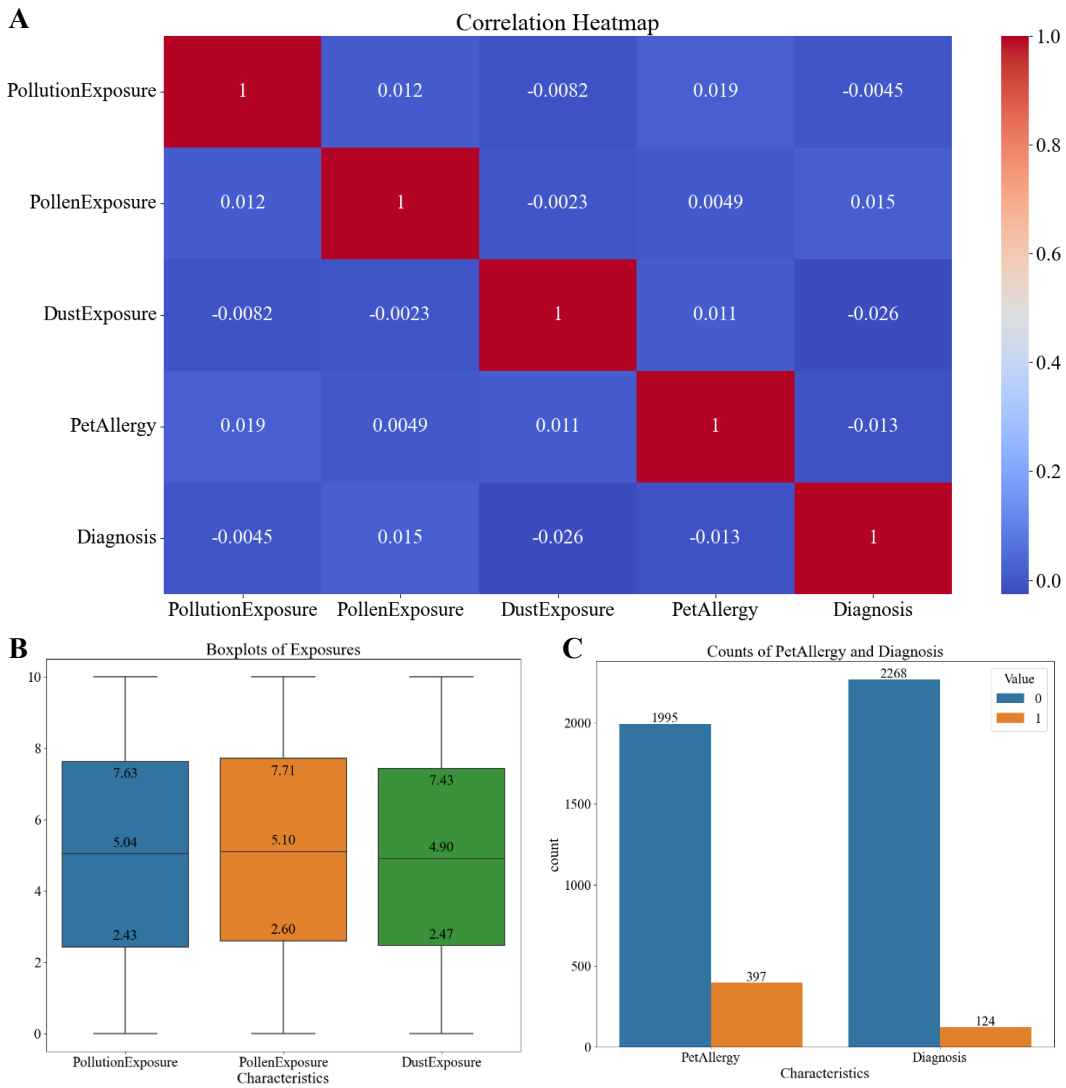

Figure 1. Descriptive Analysis of Environmental and Allergic Factors and Diagnostic Results for Asthma within the Dataset (n=2392). A demonstrates the correlation matrix of the factors. B demonstrates the boxplots for pollution, pollen, and dust exposure. C demonstrates the bar graphs for pet allergy and diagnosis values.

The results of the descriptive analysis are shown in Figure 1. In Figure 1A, the correlation matrix, none of the four characteristics and diagnosis results carry a pair-wise correlation with an absolute value over 0.3, indicating minimal or no relationship between the characteristics. Figure 1B portrays that the mean values for the level of pollution exposure, pollen exposure, and dust exposure are 5.04, 5.10, and 4.90, respectively. The box-and-whisker diagram also demonstrates that the values are mostly evenly distributed across the range of 0 to 10 since all three characteristics' third quintiles are around the value of 7.5/10 and the first quintiles are around 2.5/10. In Figure 1C, it may be observed a domination of patients without pet allergies over people with pet allergies with 1995 cases to 397 cases, and similarly a domination of patients without the diagnosis of asthma versus ones that have, with a value of 2268 to 124. This indicates that the size of the major class is about 18.3 times the size of the minor class.

DOI: 10.54254/2753-8818/50/2024AU0143

After the dataset is oversampled, the number of cases with and without a diagnosis of asthma both becomes 2267. The decrease by one for the major group may be allocated to the filtration process of border values of the Tomek links.

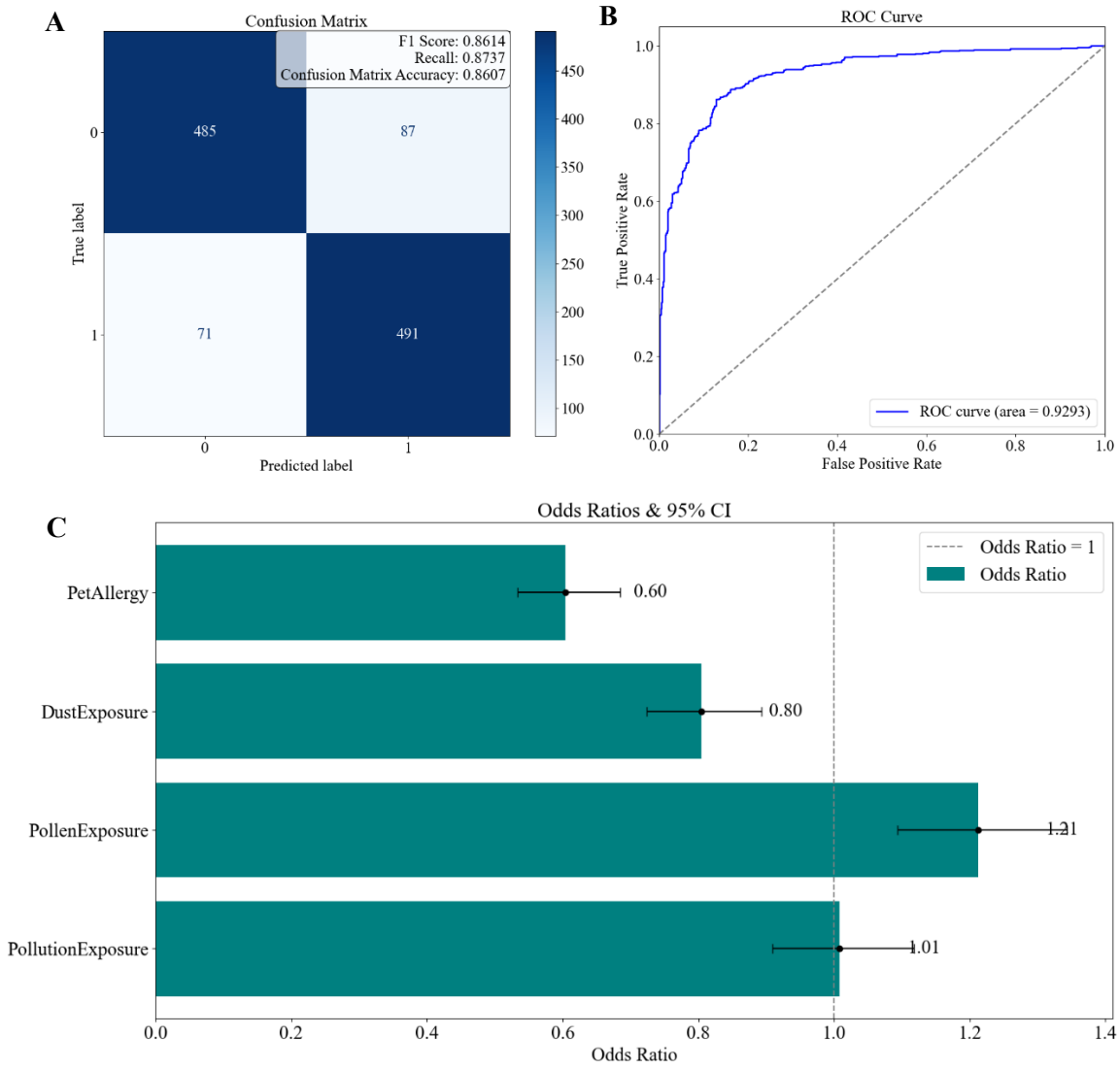

Figure 2. Logistic Regression Results for Patient Characteristics and Diagnostic Results for Asthma. The confusion matrix and the F1, Recall, and Accuracy values. B the ROC curve and AUC. C the bar graph of the odds ratios of the environmental and allergic factors and their 95% CI.

The results of the logistic regression analysis are shown in Figure 2. Figure 2A demonstrates the overall accuracy for the model to be 86.07%, with an F1 score of 0.8614 and a recall of 0.8737, while Figure 2B presents a near-ideal ROC curve with an AUC of 0.9293. As these values are all relatively close to 1, it indicates that the model is balanced and can effectively identify and distinguish the positive instances while minimizing false positives and false negatives. The odds ratios are presented in Figure 2C, with the characteristic of Pet Allergy having an odds ratio of 0.60, deviating from 1 the most, and then Pollen Exposure, Dust Exposure, and Pollution Exposure following it with values of 1.21, 0.80, and 1.01, respectively.

DOI: 10.54254/2753-8818/50/2024AU0143

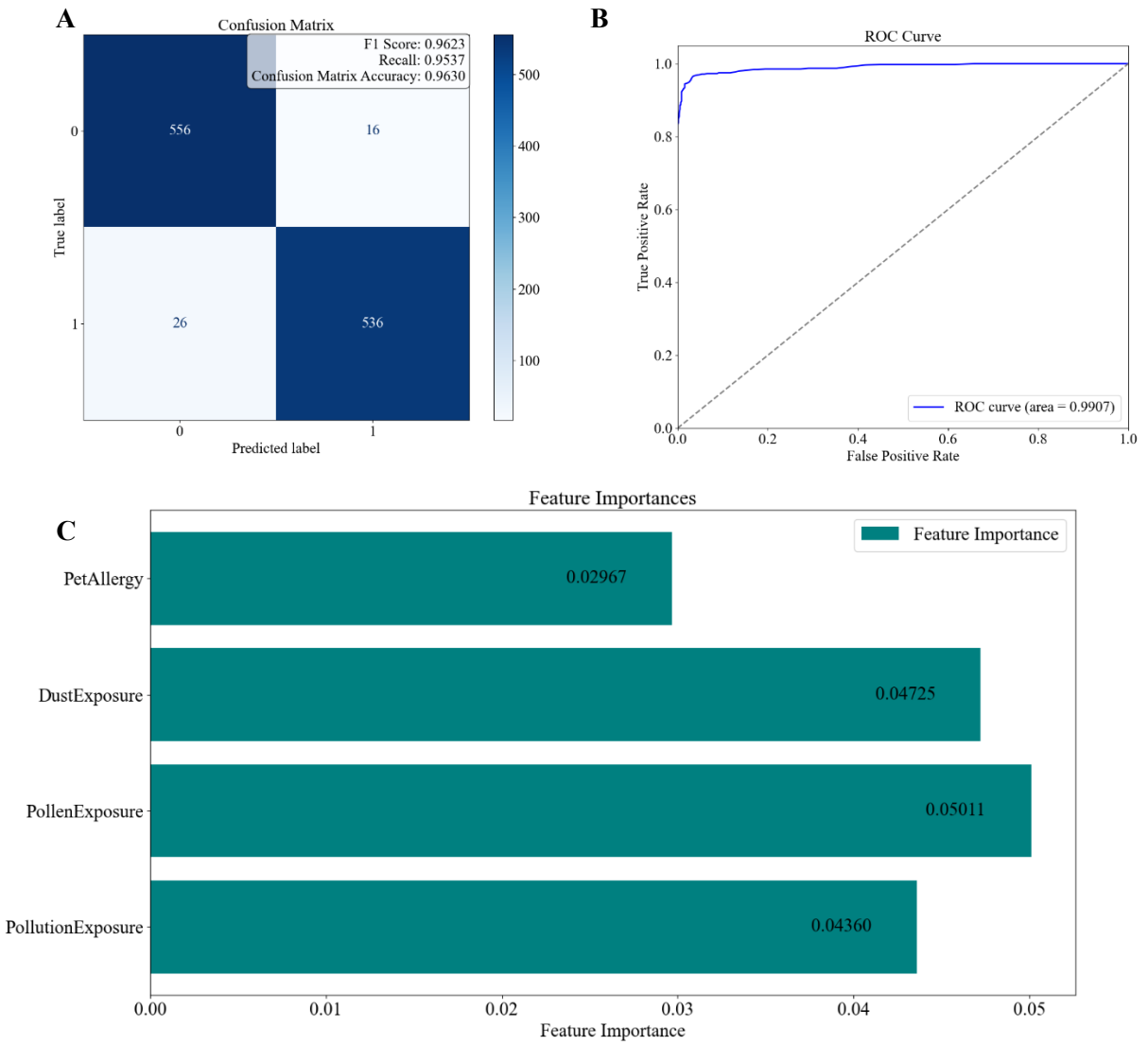

Figure 3. Random Forest Results for Patient Characteristics and Diagnostic Results for Asthma. the confusion matrix and the F1, Recall, and Accuracy values. B the ROC curve and AUC. C the bar graph of the feature importance values of the environmental and allergic factors.

The results of the random forest model analysis are shown in Figure 3. Figure 3A demonstrates the overall accuracy for the model to be 96.30%, higher than the case of the linear regression, while the F1 score, recall, and AUC values are 0.9623, 0.9537, and 0.9907, being exceptionally high. This indicates that the model can fit the dataset perfectly and can distinguish between classes well. Figure 2C presents the feature importance for the four environmental and allergic factors, with Pollen Exposure being the highest, taking 0.05011, and Dust Exposure, Pollution Exposure, and Pet Allergy being 0.04725, 0.04360, and 0.02967, respectively. This order generally corresponds with the results of the magnitude of deviation of the odds ratios for the four characteristics within the logistic regression, except for the case of pet allergy, which stands to have the largest deviation in the logistic regression and the smallest importance in the random forest model.

4. Discussion

In the linear regression model, it may be observed that the odds ratios' CIs of pollen exposure, dust exposure, and pet allergy do not contain 1. This indicates a positive correlation between the extent of

DOI: 10.54254/2753-8818/50/2024AU0143

pollen exposure with the diagnosis of asthma and a negative correlation between the extent of dust exposure and the presence of pet allergy with the diagnosis of asthma. Meanwhile, no conclusions can be made for the correlation between pollution exposure and the diagnosis of asthma with its CI for odds ratio ranging both above and below 1.

The correlation of the values of the feature importances of the characteristics of pollen exposure, dust exposure, and pollution exposure with the extent of deviation from 1 of the odds ratios of the characteristics serves the cross-verify the validity of the results. It also directly demonstrates the weight of each characteristic in predicting asthma, implying their relative strength of relationship with the development of asthma, with pollen exposure as the most crucial factor. Moreover, the disagreement between the odds ratios and feature importance for pet allergy may be explained by the fact that the nature of this variable is binary. Due to the difference in nature of the logistic regression and the random forest model to treat the data values, it would be meaningless to compare the feature importance for the binary value along with other continuous variables ranging from 0 to 10 while referencing the odds ratio values.

The results for the positive correlation and significance of pollen exposure agree with current mainstream research. It is currently found that exposure to pollen may lower lung function in the case of teenagers [11]. Exposure to high levels of birch pollen, for example, may exacerbate respiratory symptoms and reduce peak expiratory flow in individuals with allergic asthma [17].

The ambiguous relationship between pollution exposure and the development of asthma also aligns with the progress of current studies. For example, it has been found that pollutions, especially air pollution, which stands as the major source of pollution in daily life, are associated with asthma incidents in children and adolescents [8, 9]. In various cohort studies, it is stated that a causal relationship between air pollution and the development of adult asthma has yet to be established [18]. Since the average age of the patients within the dataset is above 40, an adult population is being resembled, thus explaining the absence of associations between pollution exposure and asthma development in this case.

For pet allergies, various studies agree that generally, exposure to pet allergens [19] and cats and dogs at a young age would decrease the chances of developing asthma [15]. This provides a possible explanation for the result of the study that pet allergies carry a negative correlation with asthma development since consistent exposure to pet allergens may also lead to the development of pet allergies. However, it should be taken into consideration that the majority of samples in this study are adults, with 1978 out of 2392 samples (82.69%) with an age larger than or equal to 18, and there are no traces of information on whether they had pets when they were young. With the limited number of studies in this field, a comprehensive explanation may not be provided.

On the other hand, the result that dust exposure has a negative correlation with asthma development deviates from prevailing scientific evidence. It is demonstrated that house dust mites, a key component of dust, along with particulate matter as a whole, can worsen asthma symptoms and outcomes. However, one possible reason for the presented negative correlation may be the hygiene hypothesis [18, 20]. The Hygiene hypothesis suggests that exposure to specific microorganisms within the dust may stimulate the immune system to protect against diseases, which may include allergic ones such as asthma. Indeed, it has been shown that rural areas, from which a wider range of microbes are present, do have a lower prevalence of asthma and other allergic conditions as compared to urban areas [21]. Nevertheless, it is important to note that the range of microbes is not explicitly linked with dust and is still in need of future studies.

There are several potential limitations of this study. The first limitation is the nature of the dataset. The dataset is cross-sectional, meaning that it only captures the state of the patients at a single point in time. This limits the ability of the model to establish a causal relationship between the environmental and allergy factors and the diagnosis of asthma since the process of development of asthma and the change of the variables over time may not be observed. To resolve this issue, longitudinal measurements can be made in follow-up studies. Additionally, the dataset is also quite imbalanced with the people not diagnosed with asthma being 18.3 times the ones diagnosed with asthma. While the use of oversampling methods is needed, it may introduce synthetic instances that do not perfectly represent real-world cases.

DOI: 10.54254/2753-8818/50/2024AU0143

This negatively affects the generalizability of the models, as well as their ecological validity. In future studies, more samples with the diagnosis of asthma may be incorporated into the data by a more targeting sampling approach.

Secondly, some of the variables in the dataset are self-reported. For example, the scores in the dataset for "diet quality" and "sleep quality" are arbitrarily determined by the patient, while some characteristics such as "shortness of breath" and exposure to pollution factors may not be fully measured by the doctors and depend on the patient's descriptions. Self-reported data can be subject to recall bias and inaccuracies, which may affect the reliability of the final results. This, then, may be resolved by more objective and effortful ways of measuring, such as using devices to keep track of the participants' diet and sleeping status over time and monitoring the air quality data in the participants' daily environments.

5. Conclusion

This study discusses and highlights the relationship between several environmental and allergy factors and the diagnosis of asthma. While pollen exposure plays a huge role in triggering and exacerbating asthma, dust exposure, and pet allergies are found to have a negative correlation with the diagnosis of asthma, and the relationship between pollution exposure and asthma diagnosis is still ambiguous. In the study, a logistic regression model and a random forest model are established to assess the odds ratios and feature importance of the characteristics. Aside from a pet allergy, all three other exposure factors in both models follow the sequence with pollen exposure carrying the most influence, dust exposure following it, and pollution exposure having the least significance. The random forest model carries a higher accuracy, 96.30%, than the logistic regression model, 86.07%, while both models demonstrate a satisfying quality regarding the distinguishment between different classes and minimizing false positives and negatives. None of the environmental and allergy factors are found to have a pair-wise correlation with each other.

The study has a broad application in the field of medical research and public health. The correlations presented between exposure to specific factors can be used to hypothesize the patients' reason for asthma development and provide suggestions for exacerbation of symptoms, reducing asthma incidence. A pathway is also provided for researchers to further study the reason why dust and pet allergies carry such negative correlations. In future studies, more advanced mediums of measurement, ones that can measure more comprehensive and accurate quotative data, may be incorporated into the data collection process to yield results with higher reliability and validity. As the understanding of the environmental and allergy factors themselves grows, more links with a theoretical base can also be provided between them and asthma development.

References

[1] Porsbjerg, C., Melén E, Lehtimäki L and Shaw D. (2023). Asthma. The Lancet, 401(10379), 858–73.

[2] Armeftis, C., Gratziou, C., Siafakas, N. M., Paraskevi Katsaounou, Zoi Dorothea Pana, & Petros Bakakos. (2023). An update on asthma diagnosis. *Journal of Asthma*, *60*(12), 1–7. Taylor and Francis Online.

[3] Daines, L., McLean, S., Buelo, A., Lewis, S., Sheikh, A., & Pinnock, H. (2019). Systematic review of clinical prediction models to support the diagnosis of asthma in primary care. *Npj Primary Care Respiratory Medicine*, *29*(1).

[4] Alharbi, E. T., Nadeem, F., & Cherif, A. (2021). Predictive models for personalized asthma attacks based on patient's biosignals and environmental factors: a systematic review. *BMC Medical Informatics and Decision Making*, *21*(1).

[5] Akker, I. L. den, Luijn K van and Verheij T. (2016). Overdiagnosis of asthma in children in primary care: a retrospective analysis. *British Journal of General Practice*, *66*(644), e152–e157.

DOI: 10.54254/2753-8818/50/2024AU0143

[6] José, B. P. de S., Camargos, P. A. M., Cruz Filho, Á. A. S. da, & Corrêa, R. de A. (2014). Diagnostic accuracy of respiratory diseases in primary health units. *Revista Da Associação Médica Brasileira*, *60*(6), 599–612. SciElo Brazil.

[7] Akdis, C.A., Mübeccel Akdis, Boyd, S. D., Sampath, V., Galli, S. J., & Nadeau, K. C. (2023). Allergy: Mechanistic insights into new methods of prevention and therapy. *Science Translational Medicine*, *15*(679).

[8] Zanobetti, A., Ryan, P. H., Coull, B. A., Luttmann-Gibson, H., Datta, S., Blossom, J., Brokamp, C., Lothrop, N., Miller, R. L., Beamer, P. I., Visness, C. M., Andrews, H., Bacharier, L. B., Hartert, T., Johnson, C. C., Ownby, D. R., Khurana Hershey, G. K., Joseph, C. L. M., Mendonça, E. A., & Jackson, D. J. (2024). Early-Life Exposure to Air Pollution and Childhood Asthma Cumulative Incidence in the ECHO CREW Consortium. *JAMA Network Open*, *7*(2), e240535.

[9] Altman, M. C., Kattan, M., O'Connor, G. T., Murphy, R. C., Whalen, E., LeBeau, P., Calatroni, A., Gill, M. A., Gruchalla, R. S., Liu, A. H., Lovinsky-Desir, S., Pongracic, J. A., Kercsmar, C. M., Khurana Hershey, G. K., Zoratti, E. M., Teach, S. J., Bacharier, L. B., Wheatley, L. M. , Sigelman, S. M., & Gergen, P. J. (2023). Associations between outdoor air pollutants and non-viral asthma exacerbations and airway inflammatory responses in children and adolescents living in urban areas in the USA: a retrospective secondary analysis. *The Lancet Planetary Health*, *7*(1), e33–e44.

[10] Achenyo Peace Abbah, Xu, S., & Johannessen, A. (2023). Long-term exposure to outdoor air pollution and asthma in low-and middle-income countries: A systematic review protocol. *PLOS ONE*, *18*(7), e0288667–e0288667.

[11] Annesi-Maesano, I., Cecchi, L., Biagioni, B., Kian Fan Chung, Clot, B., Martine Collaud Coen, Gennaro D'Amato, Athanasios Damialis, Domínguez-Ortega, J., Galán, C., Gilles, S., Holgate, S. T., Jeebhay, M. F., Stelios Kazadzis, Papadopoulos, N. G., Quirce, S., Sastre, J., Tummon, F., Traidl-Hoffmann, C., & Jolanta Walusiak-Skorupa. (2023). Is exposure to pollen a risk factor for moderate and severe asthma exacerbations? *Allergy*, *78*(8).

[12] Schmidt, C. W. (2016). Pollen Overload: Seasonal Allergies in a Changing Climate. *Environmental Health Perspectives*, *124*(4).

[13] Posa, D., Hofmaier, S., Arasi, S., & Matricardi, P. M. (2017). Natural Evolution of IgE Responses to Mite Allergens and Relationship to Progression of Allergic Disease: a Review. *Current Allergy and Asthma Reports*, *17*(5).

[14] Zhang, Y., Ye, B., Zheng, H., Zhang, W., Han, L., Yuan, P., & Zhang, C. (2019). Association Between Organic Dust Exposure and Adult-Asthma: A Systematic Review and Meta-Analysis of Case-Control Studies. *Allergy, Asthma & Immunology Research*, *11*(6), 818.

[15] Taniguchi, Y., & Kobayashi, M. (2023). Exposure to dogs and cats and risk of asthma: A retrospective study. *Exposure to Dogs and Cats and Risk of Asthma: A Retrospective Study*, *18*(3), e0282184–e0282184.

[16] Kharoua, R. (2024). *Asthma Disease Dataset*. https://doi.org/10.34740/KAGGLE/DSV/8669080

[17] Carlsen, H. K., Haga, S. L., Olsson, D., Behndig, A. F., Modig, L., Meister, K., Forsberg, B., & Olin, A.-C. (2022). Birch pollen, air pollution and their interactive effects on airway symptoms and peak expiratory flow in allergic asthma during pollen season – a panel study in Northern and Southern Sweden. *Environmental Health*, *21*(1).

[18] Tiotiu, A. I., Novakova, P., Nedeva, D., Chong-Neto, H. J., Novakova, S., Steiropoulos, P., & Kowal, K. (2020). Impact of Air Pollution on Asthma Outcomes. *International Journal of Environmental Research and Public Health*, *17*(17), 6212.

[19] O'Connor, G.T., Lynch, S. V., Bloomberg, G. R., Kattan, M., Wood, R. A., Gergen, P. J., Jaffee, K. F., Calatroni, A., Bacharier, L. B., Beigelman, A., Sandel, M. T., Johnson, C. C., Faruqi, A. , Santee, C., Fujimura, K. E., Fadrosh, D., Boushey, H., Visness, C. M., & Gern, J. E. (2018). Early-life home environment and risk of asthma among inner-city children. *Journal of Allergy and Clinical Immunology*, *141*(4), 1468–1475.

DOI: 10.54254/2753-8818/50/2024AU0143

[20] Nahla Mohamed Okasha, Sarhan, A., & Ahmed, E. (2021). Association between house dust mites sensitization and level of asthma control and severity in children attending Mansoura University Children's Hospital. *The Egyptian Journal of Bronchology*, *15*(1).

[21] Murrison, L. B., Brandt, E. B., Myers, J. B., & Hershey, G. K. K. (2019). Environmental exposures and mechanisms in allergy and asthma development. *Journal of Clinical Investigation*, *129*(4), 1504–1515.