Ensemble Machine Learning Model to Analyse the Correlation Between Environmental Features and Respiratory Admissions in the Emergency Room

Vito Telesca and Maríca Rondinone(✉)

University of Basilicata, 85100 Potenza, Italy
{vito.telesca,marica.rondinone,marica.rondinone}@unibas.it

Abstract. This study proposes an ensemble machine learning model to analyse the association between respiratory Emergency Room (ER) admissions and environmental factors, such as air pollution and weather-climatic conditions. The analysed climatic variables include air temperatures Tmin, Tmax, Taverage, atmospheric pressure P, relative humidity RH, and levels of CO, O_3, PM_{10}, and NO_2. The data were processed as daily averages to ensure consistency and comparability in the analyses. Data on ER, provided by the Policlinico of Bari, cover the period from 2013 to 2023.

The analysis was conducted using ensemble learning techniques, applying three regression models: Random Forest, XGBoost, and Adaboost. The models were trained on a pre-processed database using a 7-day exponential moving average (EMA7) to obtain a more stable time series. Model hyperparameters were optimized through Bayesian optimization. Among the analysed models, XGBoost showed high predictive capacity in test sets. In particular, the R^2 value was 0.772, while the MAE was 0.049 cases/day.

Applying SHAP (SHapley Additive exPlanations) analysis to the XGBoost model allowed us to identify the most important variables influencing hospital admissions and their related patterns. The most relevant features, ranked by importance, were: low values of average air temperature and atmospheric pressure, and high values of CO. The SHAP method, and in particular the use of Bee Swarm plots, were used to globally interpret the results obtained by the model and allowed us to reach the above results. Furthermore, in order to determine for the most important features the values that cause an increase in admission to the emergency room for respiratory diseases, a local analysis was carried out by applying the LIME model which allowed us to say that the greater onset of respiratory diseases is associated with average temperatures lower than 12.28 °C, atmospheric pressure values lower than or equal to 1006.81 hPa and CO concentrations greater than 0.84 mg/m^3.

Keywords: Respiratory Diseases · Random Forest · XGBoost · AdaBoost · Global and Local Interpretability

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15898, pp. 417–429, 2026.
https://doi.org/10.1007/978-3-031-97657-5_27

1 Introduction

Respiratory diseases are a cause of mortality globally and are related to climatic and environmental factors such as air quality, temperature and air pollutants [1, 2]. Scientific studies have shown that fine particulate matter (PM2.5, PM10), nitrogen dioxide (NO2) and sulphur dioxide (SO2) are responsible for the onset of respiratory diseases, especially in areas of high industrialisation and vehicular traffic [3, 4]. Exposure to these pollutants causes impairment of lung function [5, 6], while climatic variables such as intense cold and humidity favour viral infections and allergic reactions [7].

To analyze these dynamics, the study uses ensemble machine learning techniques (Random Forest, XGBoost, AdaBoost) [8] and interprets the results through global and local interpretability methods (SHAP and LIME graphs), in order to identify significant correlations between environmental data and respiratory health. Specifically, after an initial preprocessing of the data and a preliminary correlation analysis between climatic and environmental factors and the target variable, the Random Forest, XGBoost and AdaBoost models (optimized through Bayesian optimization) were applied, whose performances were compared through cross-validation, analysis of metrics and errors and their distribution.

From the comparison between the three models, XGBoost was found to be the one with the best performance and, consequently, the SHAP model was applied to identify the features that best conveyed the model's predictions. Furthermore, the LIME model was used to determine, for the most significant parameters, the threshold values of the quantities.

2 Study Area

In this study, the city of Bari, the capital of Puglia, is examined. Located in Southern Italy and bathed by the Adriatic Sea, the city with a population of 320,000 inhabitants [9], has a Mediterranean climate with hot and sunny summers and mild and rainy winters [10]. To carry out the analyses discussed in this manuscript, hospital data provided by the Policlinico di Bari, one of the most important hospitals in the city, were used, while environmental and climatic data were obtained from local monitoring cabins.

3 Materials and Methods

3.1 Methodology

The workflow of the applied methodology is illustrated in Fig. 1; the steps are the following:

i) after an initial data collection phase, the data underwent preprocessing to handle missing values and align them temporally;

ii) a correlation analysis was then conducted to identify relationships between climatic variables and hospital admissions. The correlation analysis was conducted by calculating both the Pearson coefficient (r) and the relative p-value: the first allows us to understand how much two variables are correlated with each other and presents values that oscillate between 1 (positively correlated variables), 0 (absence of correlation) and -1 (negative correlation, as one variable increases, the other decreases). The p-value, on the other hand, was used to understand if the correlation obtained was statistically significant, in fact, a p-value lower than a threshold of 0.05 supports the conclusion that the correlation is not random [11]. In the case in question, the p-value was found to be lower than the threshold value and therefore the correlation obtained was considered statistically significant;

iii) the data were then normalized to make them comparable;

iv) an Exponential Moving Average (EMA7) filter was applied to reduce daily variability and identify significant trends in the time series. This method smooths the data and improves the quality of predictions;

v) For predictive analysis, three ensemble learning algorithms [12] were utilized:

- Random Forest
- XGBoost
- AdaBoost

These models were chosen for their capability to manage complex data and identify nonlinear relationships between meteorological variables and hospital admissions. In particular, the Random Forest is an algorithm that is based on the use of decision trees built using subsets of data. The final prediction is obtained from the average of the predictions that are provided by the different trees (in the case of regression analysis) or by majority vote (in the case of classification, in which the class that has emerged most from the analysis carried out by the different decision trees is chosen) [13]. The XGBoost (Extreme Gradient Boosting) model is also based on the use of decision trees that work sequentially. The characteristic of this model lies in the ability of the subsequent trees to correct the errors of the previous trees: in this way the model has a high precision and can also be used for very complex analyses [8]. Finally, the AdaBoost (Adaptive Boosting) model is a model that combines simple models in which the subsequent ones will correct the errors of the previous ones, ensuring that in the end, even when combining the results of all the simple models, those that have made the most accurate predictions will have greater weight, thus obtaining, in fact, an extremely reliable and precise prediction [14]. To improve their predictive performance, the models were optimized through hyperparameter tuning using Bayesian optimization. Additionally, a model validation phase was conducted that included [15]:

- Cross-validation to ensure the robustness of the results.
- Calculation of performance metrics (e.g., R^2, MAE, RMSE).
- Analysis of error distribution to identify potential biases in the models;

vi) to ensure transparency in the results and understand the contribution of meteorological variables to predictions, Feature Importance and Explainable Artificial Intelligence (XAI) techniques were adopted [16]. Specifically:

- SHAP (Shapley Additive Explanations) was used to determine feature importance.
- Global and local interpretability was applied to analyze the model's behavior both on a general scale and for individual case studies.

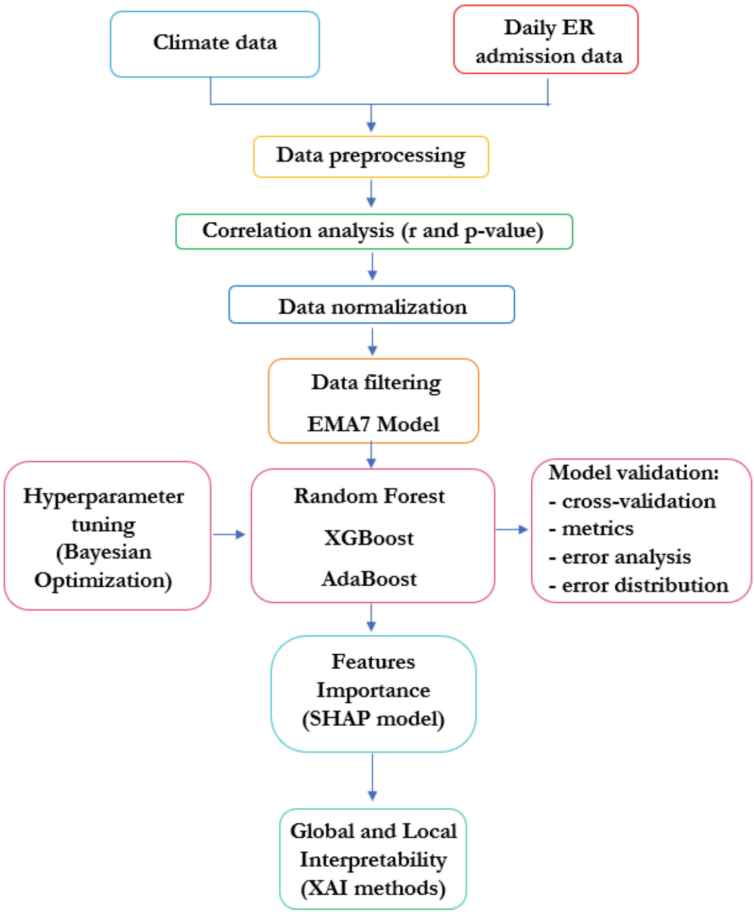

Fig. 1. Methodological framework

The main objective of XAI methodologies is to make artificial intelligence models transparent, thus avoiding the so-called "black-box" processes.

3.2 Meteorological Data

In this study, the data used are of a climatic and environmental nature. In the first case, the average temperature T [°C], the atmospheric pressure P [hPa] and the relative humidity RH [%] were taken into consideration; regarding the environmental parameters, the

carbon monoxide CO [mg/m3], the fine particulate matter PM10 [$\mu g/m^3$], the nitrogen dioxide NO2 [$\mu g/m^3$] and the ozone O3 [$\mu g/m^3$] were taken into consideration. The data considered cover a period of ten years, from 2013 to 2023.

To describe the collected data, statistical parameters such as minimum, maximum, mean, standard deviation, median (50th percentile), first, third and fourth quartile (25th, 75th and 90th percentile respectively) were used (Table 1).

Table 1. Statistical description of the data

	T	P	RH	CO	PM10	NO2	O3	RD
min	~0,11	976,60	25,49	0	0	0	0	0
avg	17,83	1011,77	70,29	0,70	22,69	53,02	82,78	5,08
max	36,28	1039,35	99,00	3,00	117,00	157,00	154,00	25,00
std	6,39	8,20	10,50	0,40	11,06	24,55	20,80	3,23
90th	26,67	1022,09	83,51	1,30	36,00	87,00	110,00	9,00
75th	23,47	1017,12	78,00	0,90	28,00	68,00	97,63	7,00
50th	17,28	1012,32	70,76	0,60	21,00	50,00	82,00	4,00
25th	12,36	1006,20	63,23	0,40	15,00	35,00	67,00	3,00

3.3 Emergency Room (ER) Data

The health data used in this study concern patients affected by respiratory diseases and were provided by the Policlinico of Bari, one of the main hospital facilities in Southern Italy. These data include hospitalized for respiratory diseases (RD), diagnoses of asthma and COPD, and other respiratory complications related to adverse environmental conditions. Integrating this information with meteorological data allows for examining possible correlations between air pollution, climatic conditions, and respiratory health in the Bari population.

4 Application to Case Study

4.1 Correlation Analysis

Following the methodological framework, the first step of the analysis involved studying the correlations between the features and the target variable. This was achieved by calculating the Pearson correlation coefficient, r, and the corresponding p-value.

As can be seen (Fig. 2), a negative correlation between RD and temperature ($r = -0.51$) emerges; this suggests that warmer climatic conditions could reduce the risk of RD. A negative correlation is also observed with P ($r = -0.27$), implying that an increase in atmospheric pressure levels could be associated with a decrease of the number of respiratory diseases cases, although the correlation is not very strong. A negative

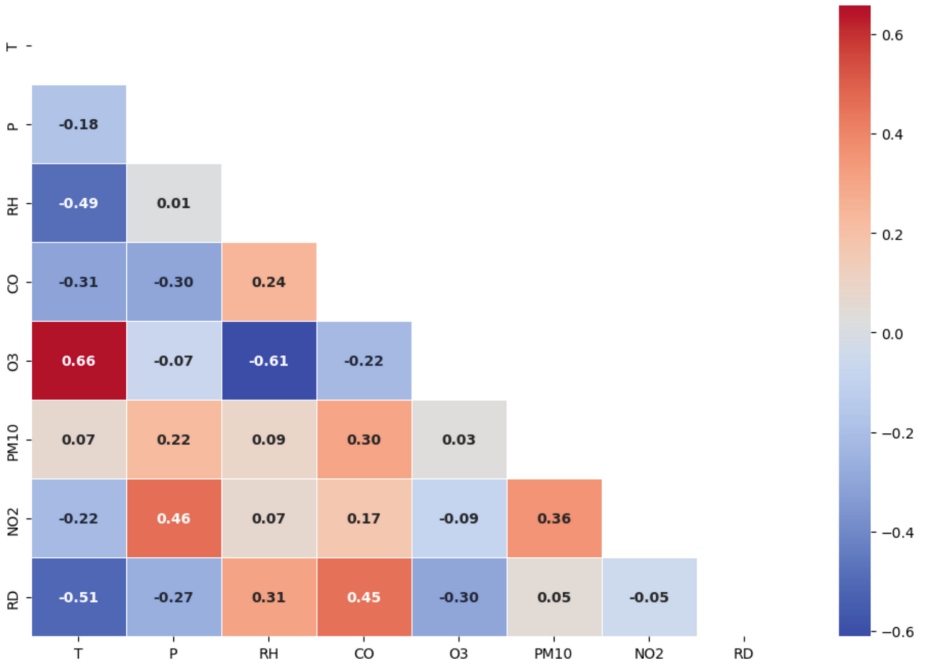

Fig. 2. Correlation matrix

correlation is also found with O_3 ($r = -0.30$), indicating that an increase in ozone concentration in the air could be associated with a slight reduction in the number of RD cases. Positive correlations are found with RH ($r = 0.31$), suggesting that an increase in humidity may promote the onset or worsening of respiratory diseases. Finally, a positive correlation is also observed with CO ($r = 0.45$), suggesting that exposure to these pollutants could significantly increase the risk of developing otitis. Normalization helps to prevent variables with larger numerical values from dominating those with smaller values, ensuring a more balanced assessment of the relationships between variables. Subsequently, to enhance the analysis quality and minimize daily data fluctuations, smoothing was applied using a 7-day Exponential Moving Average (EMA7).

4.2 Machine Learning Models

Once the primary correlations and data were defined, an ensemble model was implemented to enhance the prediction of RD cases, utilizing the combination of multiple predictive algorithms. As mentioned earlier, the models employed were Random Forest, Extreme Gradient Boost, and Adaptive Boost. These three models were fine-tuned using Bayesian optimization, and their performances were then compared (Table 2).

Training a machine learning model with the Random Forest algorithm yielded strong results, with the R^2 coefficient for the test set reaching 0.740. This indicates that the model showed good adaptability and accuracy when predicting a new dataset not seen during training fase. The performance of the model was also validated through the Mean

Table 2. Model performance

Random Forest	XGBoost	AdaBoost post
R2 training (-) = 0.963	R2 training (-) = 0.995	R2 training (-) = 0.999
R^2 test (-) = 0.740	R^2 test (-) = 0.772	R^2 test (-) = 0.789
Mae training (case/day) = 0.02	Mae training (case/day) = 0.007	Mae training (case/day) = 0.001
MAE test (case/day) = 0.053	MAE test (case/day) = 0.049	MAE test (case/day) = 0.045

Absolute Error (MAE), which for the test set was 0.053 cases per day, reflecting a reasonable prediction accuracy. In comparison, the XGBoost model also delivered competitive results, with an R^2 of 0.772 for the test set. The MAE for XGBoost was even lower, at 0.049 cases per day, suggesting a more precise prediction compared to Random Forest. The AdaBoost model achieved the best performance during the training phase, with an R^2 of 0.999 and a MAE of 0.001 cases per day, indicating high accuracy and precision. However, such outstanding results raised concerns about overfitting, potentially making the model less reliable during generalization. In addition to evaluating R^2 and MAE, the error distribution was also examined. Although the XGBoost model showed both overestimation and underestimation (errors within the $\lesssim 20\%$ range), it showed a more favourable error distribution, closely resembling a Gaussian curve (Fig. 3), a key characteristic for enhancing the robustness and generalizability of predictions.

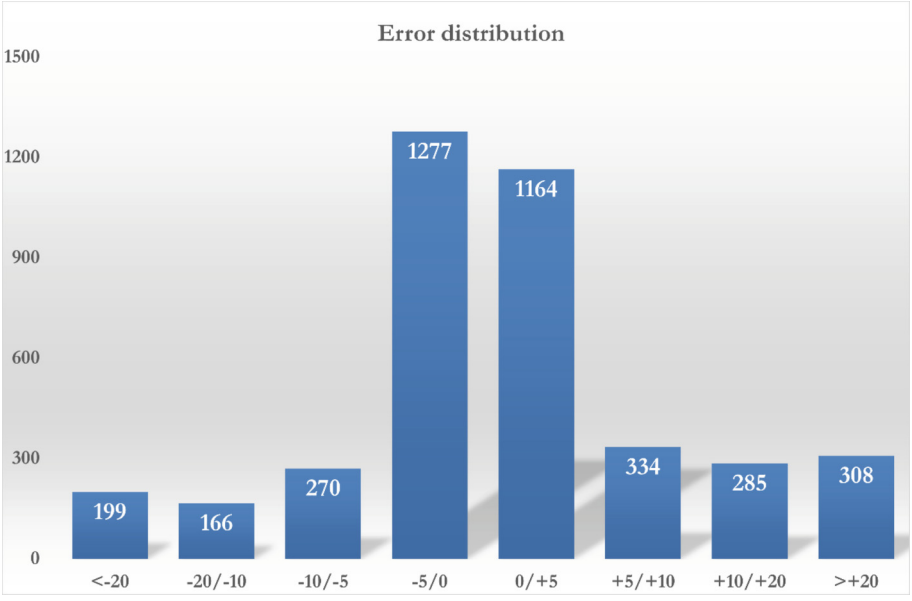

Fig. 3. Error distribution of XGBoost model

The results in Fig. 3 show that the majority of the simulated values (61%) fall within the relatively small error range of ± 5%. Specifically, 31.9% of the simulated values are within the error range of -5% to 0%, while 29.1% fall within the range of 0% to + 5%. This distribution illustrates a strong alignment of the simulation model with the actual values, with a significant percentage of simulated values deviating only slightly from the actual values within the error margin of ± 5%. As you can see, most of the errors fall within the ± 5% range, which allows us to say that the XGBoost model is able to accurately simulate the real values. Furthermore, the robustness of the model was evaluated by applying the k-fold cross-validation procedure, with k = 10 [17] (Table 3).

Table 3. Average Results of Cross Validation

Cross Validation (k-fold = 10)	
R^2 test (-)	0.772
MAE test (case/day)	0.049
MSE test (case/day)^2	0.005
RMSE test (case/day)	0.068

From the results obtained, it can be observed that the model shows good results even on subsets of data unknown to it, allowing us to affirm that it shows a good capacity to generalization. On the basis of this, it is possible to continue the analysis with the application of Features Importance, in order to identify the variables that most influence the model's predictions in terms of access to the emergency room for respiratory diseases.

4.3 SHAP Feature Importance and Global Interpretability

The SHAP method belongs to the so-called XAI methods, that is, methods used to clearly interpret the results obtained from machine learning models. Through this method, the features that have most influenced the prediction of the model used are identified [18–20].

In this study, SHAP was employed to identify the most influential variables affecting the target outcome, as shown in the ranked bar chart and the corresponding table (Fig. 4 and Table 4).

The bar chart and the relative Table (Fig. 4 and Table 4) shows that the average temperature is the variable with the greatest impact, contributing around 38% to the model's predictions. Further analysis reveals that the contributions of the other features are more balanced; since SHAP analysis assumes that the most relevant features are those whose combined contribution reaches at least 70–80%, the other important features in this case are temperature, atmospheric pressure, and carbon monoxide.

The primary objective of XAI techniques is to promote transparency in the behaviour of artificial intelligence models, thereby avoiding the so-called black-box mechanisms.

Global interpretability was further enhanced through the SHAP Bee Swarm plot (Fig. 5), which illustrates both the magnitude and direction of the three most important feature's effects across all predictions.

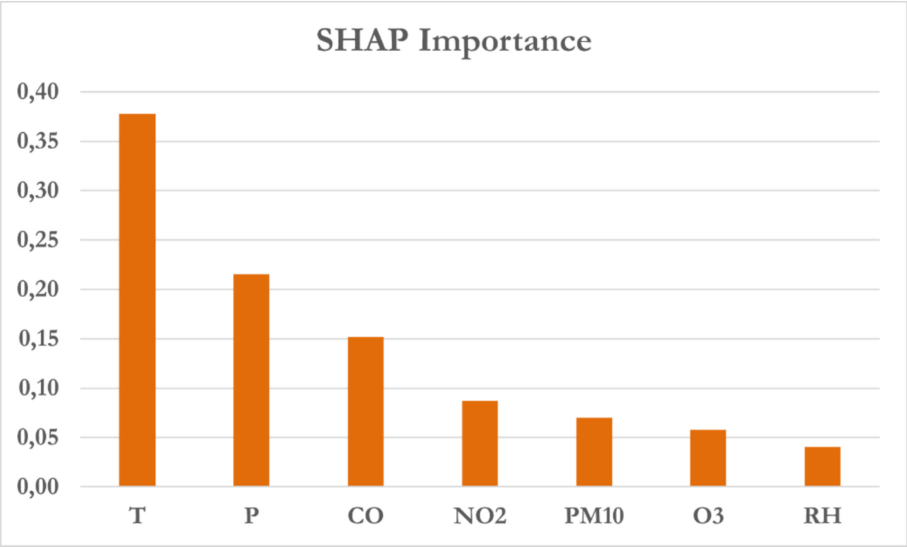

Fig. 4. SHAP feature importance based on SHAP values

Table 4. Percentage contribution of each feature

Feature	% Contribution
T	**0,38**
P	**0,22**
CO	**0,15**
NO2	0,09
PM10	0,07
O3	0,06
RH	0,04

As a visualization tool based on XAI techniques, the Bee Swarm plot provides a comprehensive overview of how input variables influence the model's behaviour on a global scale; the plot clearly shows that the onset of respiratory diseases is generally associated with low values of average temperature and atmospheric pressure, while high levels of carbon monoxide tend to increase the incidence of RDs.

4.4 Local Interpretability

LIME (Local Interpretable Model-agnostic Explanations) is a used technique to enhance the interpretability of complex machine learning models. On the other hand, LIME offers a local approximation of complex models by fitting a surrogate model to a small subset of perturbed data [21]. This simplified local model is interpretable and can be used to

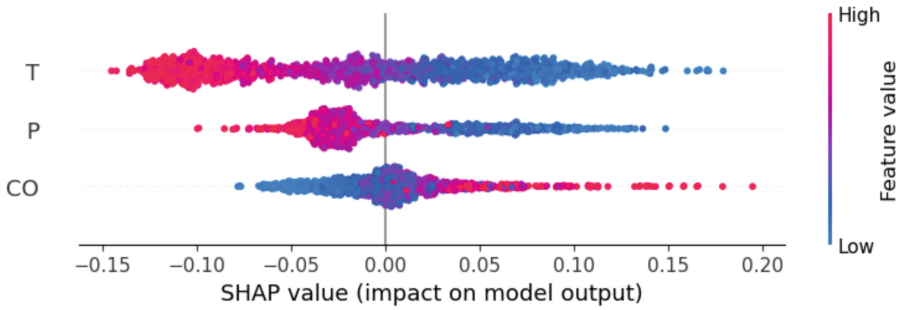

Fig. 5. Bee Swarm plot showing the distribution and impact of T, P and CO

explain how the original model behaves around a specific data point, providing clarity about why a certain prediction was made. In this study, LIME is used to analyse the role of average temperature, atmospheric pressure, and carbon monoxide in the model's predictions. The results obtained are shown in Fig. 6.

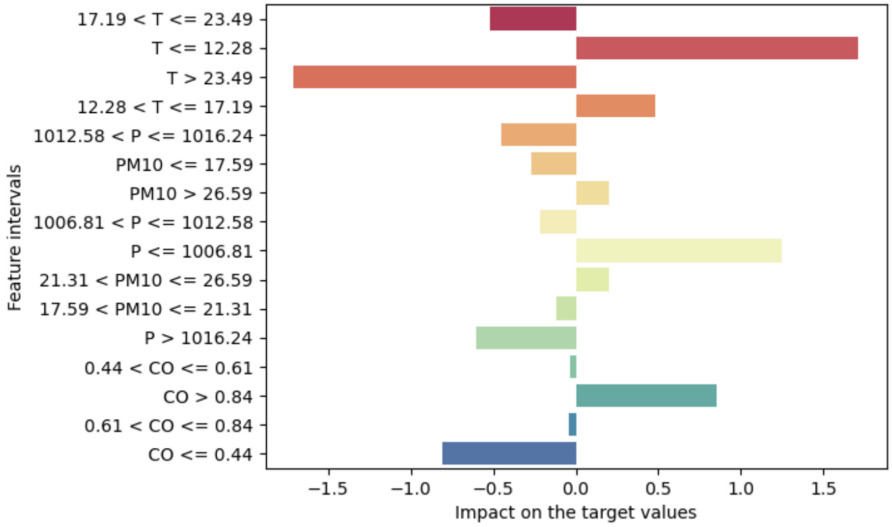

Fig. 6. Significant average temperature, atmospheric pressure and carbon monoxide intervals

The graph in Fig. 6 highlights that temperatures at or below $T \leq 12.28$ °C are those most strongly associated with an increased risk of respiratory disease onset. An appreciable influence is also observed within the range $12.28\ °C < T \leq 17.19\ °C$. On the other hand, for $17.19\ °C < T \leq 23.49\ °C$ and especially for higher temperatures ($T > 29.49\ °C$), the predicted number of RD cases tends to decrease. With regard to atmospheric pressure, values equal to or below $P \leq 1006.81$ hPa appear to contribute most significantly to the onset of RD. A transitional effect is noted in the intermediate range $1006.81 < P \leq 1012.58$ hPa. Beyond that, for $1012.58 < P \leq 1016.24$ hPa and

$P > 1016.24$ hPa, the model forecasts a progressive decline in the number of cases. For carbon monoxide concentrations, the data suggest that lower levels ($CO \leq 0.44$ mg/m^3) are associated with a reduced number of predicted RD events. Slightly higher values, specifically within the ranges $0.44 < CO \leq 0.61$ mg/m^3 and $0.61 < CO \leq 0.84$ mg/m^3, show a moderate contribution, whereas the predicted incidence continues to decline for CO levels exceeding 0.84 mg/m^3.

5 Conclusion

This study explored the interaction between meteorological variables and respiratory health in the population of Bari, with the aim of identifying environmental factors that influence the incidence of respiratory diseases (RD). The analysis was conducted using an artificial intelligence approach, which integrated meteorological data from environmental monitoring stations with healthcare information related to emergency department admissions for respiratory diseases. The correlation analysis revealed significant associations between various environmental variables and the frequency of respiratory diseases. The study used an ensemble learning approach that involved the use of the optimized Random Forest, XGBoost and AdaBoost models, whose performance comparison highlighted the XGBoost model as the best performing, since it showed a determination coefficient R^2 equal to 0.772 and a MAE of 0.049 both for the test phase. The subsequent application of the SHAP model with the use of the Bee Swarm Plot allowed for a global interpretation of the results, revealing low temperature and pressure values, and high CO values as the most significant in predicting visits to the ER for respiratory diseases.

The threshold values were then determined through the application of the LIME model which allowed for the local analysis of the results obtained and which showed as threshold values that temperatures ≤ 12.28 °C, atmospheric pressures ≤ 1006.81 hPa and carbon monoxide concentrations > 0.84 mg/m^3.

In the future, the application of other Machine e Deep Learning algorithms will be developed to obtain higher performances and new global and local interpretability methods to investigate the complex correlations between environmental factors and Emergency Room admissions. In this way it will be possible to link the model results to specific meteorological and environmental predictors, allowing us to know in advance any increases in hospitalizations for cases of respiratory diseases.

A limitation to be taken into consideration concerns the geographical origin of the data. The study is based on the use of environmental and climate data relating to the city of Bari. The health data come from the Policlinico of the same city; however, it is possible that some of the patients examined are not residents of Bari but come from other areas of Puglia, characterized by different climate and environmental conditions from those under study. This discrepancy could cause a bias in the analysis phase since the climate and environmental data used may not reflect the real exposure of the patients. In this regard, in the future, a data selection could be carried out based on the municipality of residence of the patients, in order to obtain an accurate estimate of individual exposure.

Conflicts of Interest.. The authors have no conflicts of interest to declare related to the content of this article.

Ethics Approval and Consent to Participate.. The Emergency Department visit database is fully anonymized according to the privacy code. It is a completely de-identified data set that, as such was not subject to the approval of the ethics committee. No patient contact was made, and patients could not be traced. The data provided not contains any personal information about patients.

References

1. Tran, H.M., Tsai, F.J., Lee, Y.L., Chang, J.H., Chang, L.T., Chang, T.Y., et al.: The impact of air pollution on respiratory diseases in an era of climate change: a review of the current evidence. Sci. Total. Environ. **898**, 166340 (2023). https://doi.org/10.1016/j.scitotenv.2023.166340
2. Salve, H.R., Nawaz, H., Sharma, P., Dey, S., Madan, K., Krishnan, A.: Effectiveness of community and household level Interventions for reducing impact of air pollution on health outcomes - A systematic review. ISEE Conf. Abstract. (2022). https://doi.org/10.1289/isee.2022.P-0949
3. Li, J., et al.: Double trouble: the interaction of PM2.5 and O3 on respiratory hospital admissions. Environ. Pollut.Pollut. **338**, 122665 (2023). https://doi.org/10.1016/j.envpol.2023.122665
4. Mebrahtu, T.F., et al.: The effects of exposure to NO2, PM2.5 and PM10 on health service attendances with respiratory illnesses: a time-series analysis. Environ. Pollut.Pollut. **333**, 122123 (2023). https://doi.org/10.1016/j.envpol.2023.122123
5. Su, J.G., Aslebagh, S., Shahriary, E., Barrett, M., Balmes, J.R.: Impacts from air pollution on respiratory disease outcomes: a meta-analysis. Front. Public Health **12**, 1417450 (2024). https://doi.org/10.3389/fpubh.2024.1417450
6. Zhang, Y., Li, J., Wang, Y., Zhang, L., Li, J., Li, Y., et al.: Association between short-term exposure to air pollution and respiratory diseases among children in China: a systematic review and meta-analysis. Environ. Pollut.Pollut. **268**, 115682 (2021). https://doi.org/10.1016/j.envpol.2020.115682
7. Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., et al.: Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet **386**(9991), 369–375 (2015). https://doi.org/10.1016/S0140-6736(15)60585-6
8. Chen, T., Guestrin, C.: *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
9. ISTAT - Italian National Institute of Statistics. (in Italian) https://www.istat.it/. Accessed 1 April 2025
10. ISPRA, Italian National Institute for Environmental Protection and Research: https://www.isprambiente.gov.it/it. Accessed 1 April 2025
11. Gogtay, N.J., Thatte, U.M.: Principles of correlation analysis. J. Assoc. Phys. India **65**(3), 78–81 (2017). https://pubmed.ncbi.nlm.nih.gov/28462548/
12. Mienye, I.D., Sun, Y.: A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access **10**, 99129–99149 (2022). https://doi.org/10.1109/ACCESS.2022.3207287

13. Yang, X., Li, Y., Liu, L., Zang, Z.: Prediction of respiratory diseases based on random forest model. Front. Public Health (2025). https://doi.org/10.3389/fpubh.2025.1537238
14. Gutierrez-Tobal, G.C., Alvarez, D., del Campo, F., Hornero, R.: Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow. IEEE Trans. Biomed. Eng. **63**(3), 636–646 (2016). https://doi.org/10.1109/TBME.2015.2467188
15. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science **7**, e623 (2021). https://doi.org/10.7717/peerj-cs.623
16. Arias-Duart, A., Parés, F., Garcia-Gasulla, D., Giménez-Ábalos, V.: Focus! rating XAI methods and finding biases. In: Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882821
17. Nti, I.K., Nyarko-Boateng, O., Aning, J.: Performance of machine learning algorithms with different k values in k-fold crossvalidation. Int. J. Inform. Technol. Comput. Sci. **13**(6), 61–71 (2021). https://doi.org/10.5815/ijitcs.2021.06.05
18. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al.: Explainable artificial intelligence: a survey. Inform. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
19. Shaikhina, T., Bhatt, U., Zhang, R., Georgatzis, K., Xiang, A., Weller, A.: Effects of uncertainty on the quality of feature importance explanations. In: AAAI Workshop on Explainable Agency in Artificial Intelligence, AAAI Press (2021). https://umangsbhatt.github.io/reports/AAAI_XAI_QB.pdf
20. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with Shapley-value-based explanations as feature importance measures. In: Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pp.5491–5500 (2020). https://proceedings.mlr.press/v119/kumar20e.html
21. Salih, A., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Menegaz, G., et al.: A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. arXiv preprint arXiv:2305.02012 (2023). https://doi.org/10.48550/arXiv.2305.02012