DOCTORAL THESIS

Explainable artificial intelligence for studying the impact of environmental factors on health

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

Interuniversity PhD Course in Sustainable Land Management

University of Bari and Polytechnic University of Bari

XXXVII cycle

SSD: PHYS-06/A

PhD Candidate:

Donato Romano

Supervisor:

Prof. Sabina Tangaro

Coordinator:

Prof. Francesco Gentile

Academic Triennium: 2021–2024

Abstract

Name of the student: **Donato Romano**

Degree for which submitted: **PhD**

Thesis title: **Explainable artificial intelligence for studying the impact of environmental factors on health**

Thesis supervisor: **Prof. Sabina Tangaro**

Month and year of thesis submission: **March 2025**

This thesis explores the application of machine learning (ML) and explainable artificial intelligence (XAI) to investigate the complex relationship between environmental factors and non-communicable diseases (NCDs), with a particular focus on mortality trends in Italy. The research is motivated by the growing recognition of the impact of environmental exposures, such as air pollution, on human health, particularly in the development and progression of chronic diseases like respiratory cancers, cardiovascular diseases, and metabolic disorders. By leveraging advanced ML techniques and XAI, this work aims to uncover the underlying mechanisms linking environmental stressors to health outcomes, providing actionable insights for public health interventions and policy-making.

The thesis is structured around three key studies, each employing a combination of predictive modeling and XAI to analyze different health outcomes influenced by environmental and socio-economic factors. The first study focuses on the relationship between air pollution and mortality from respiratory system cancers in Italian provinces. Using remote sensing data from the Copernicus Atmosphere Monitoring Service (CAMS) and socio-economic data from the Italian National Institute of Statistics (ISTAT), the study employs a Random Forest regressor to predict standardized mortality ratios (SMRs) and identifies nitrogen monoxide (NO), income levels, and ozone (O_3) concentrations as the

most significant predictors of respiratory cancer mortality. The use of SHAP (Shapley Additive Explanations) values enhances the interpretability of the model, offering insights into the contributions of various environmental and socio-economic factors.

The second study builds on the first by comparing thirteen classification models to predict whether Italian provinces exhibit respiratory cancer mortality rates above or below the national average. The study evaluates the performance of various ML algorithms, including Random Forest, Gradient Boosting, and Support Vector Machines, and identifies the Extra Trees classifier as the most accurate model. SHAP analysis reveals that NOx, benzene, and income levels are the most influential features, reinforcing the findings of the first study and providing a broader methodological perspective for identifying environmental determinants of respiratory cancer mortality.

The third study shifts focus to endocrine, nutritional, and metabolic diseases, examining the relationship between air pollution, socio-economic factors, and mortality rates across Italian provinces. The study employs three regression models—Gradient Boosting (GB), Random Forest (RF), and XGBoost—to predict SMRs for these diseases. Gradient Boosting emerges as the best-performing model, with SHAP analysis highlighting the critical roles of life quality, lifetime, and benzene exposure in influencing mortality rates. The study underscores the interplay between environmental and socio-economic determinants in shaping health outcomes and provides actionable insights for policymakers aiming to reduce health disparities and mitigate risk factors.

Collectively, these studies demonstrate the effectiveness of combining predictive modeling with XAI to provide accurate, interpretable, and actionable insights into the impact of environmental factors on public health. The use of SHAP values ensures transparency and trustworthiness in AI-driven epidemiological research, enabling stakeholders to understand the underlying mechanisms driving model predictions. This approach not only enhances the interpretability of complex ML models but also facilitates the generation of new hypotheses and guides future research directions.

The findings of this thesis have significant implications for public health policy, particularly in the areas of pollution control, urban planning, and targeted interventions. By identifying key environmental and socio-economic predictors of mortality, this work provides a data-driven framework for assessing health risks and guiding preventive measures. The integration of geospatial analysis and robust validation techniques further emphasizes the importance of methodological rigor in environmental health research.

In conclusion, this thesis contributes to the growing field of AI-driven epidemiology by providing a transparent and interpretable framework for assessing environmental health risks. The research highlights the transformative potential of ML and XAI in advancing our understanding of the complex interactions between environmental exposures and human health. By bridging machine learning, explainability, and epidemiology, this work lays the foundation for future advancements in AI-assisted health risk assessment and environmental policy development, ultimately aiming to reduce the burden of non-communicable diseases and promote healthier, more sustainable communities.

Contents

List of Publications v

1 Introduction 1
- 1.1 Background and Motivation . . . 2
- 1.2 Machine Learning models . . . 4
- 1.3 Explainable Artificial Intelligence (XAI) . . . 10
- 1.4 Outline of the thesis . . . 13

2 AI-Driven Insights on Air Pollution and Respiratory Cancer Mortality in Italy 23

3 Machine learning models for higher and lower mortality 34

4 AI-Powered Insights into Environmental and Socioeconomic Influences on Metabolic Disease Mortality in Italy 46

5 Conclusions 58

List of Publications

Publications from Thesis

1. **Romano D.**, et al., Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach; *10.3389/fpubh.2024.1344865*.

2. **Romano D.**, et al., Insights from Explainable Artificial Intelligence of Pollution and Socioeconomic Influences for Respiratory Cancer Mortality in Italy; *10.3390/jpm14040430*.

3. **Romano D.**, et al., Explainable artificial intelligence for endocrine, nutritional and metabolic mortality in Italy with geolocalized pollution data; Under review at Applide System Innovation.

Chapter 1

Introduction

This chapter establishes the foundation of the thesis, highlighting the role of artificial intelligence (AI) and explainable machine learning (XAI) in advancing environmental research. It outlines the research objectives and emphasizes the importance of interpretability in AI models for supporting environmental sustainability and policy development.

1.1 Background and Motivation

In recent decades, the link between environmental factors and human health has become increasingly clear [1]. Air pollution, exposure to toxic substances, climate change, and other environmental factors play a crucial role in the development and progression of numerous diseases, including respiratory, cardiovascular, and chronic illnesses. In this context, the concept of the exposome, introduced to describe the total set of environmental exposures an individual experiences throughout their life, represents a fundamental tool for understanding the complexity of interactions between environment and health [2]. The exposome encompasses not only external factors such as air pollution, chemical exposure, and climatic conditions but also internal factors like the microbiome and metabolic processes. For example, prolonged exposure to fine particulate matter (PM2.5) has been associated with an increased risk of cardiovascular diseases, while the interaction between chemical agents and genetic factors can contribute to the development of chronic diseases [3]. This holistic approach enables integrated analysis of environmental exposures, offering new perspectives for disease prevention and management.

Simultaneously, machine learning (ML) has emerged as an essential tool to address the complexity and multidimensionality of data related to the exposome. ML techniques, such as convolutional neural networks and unsupervised clustering algorithms, are used to analyze large environmental and health datasets, uncover hidden patterns, and predict relationships between exposures and diseases [4, 5]. For instance, ML models have been successfully applied to identify predictive biomarkers of air pollution exposure and to analyze complex interactions between genetic and environmental factors [6, 7]. Beyond this, ensemble learning approaches, such as random forests and gradient boosting machines, have proven effective in modeling nonlinear and interactive effects of environmental exposures on health outcomes [8]. Furthermore, natural language processing (NLP) has been utilized to extract valuable information from unstructured text data, such as medical records and environmental reports, to better understand exposure-disease links [9, 10].

Recent studies have also focused on integrating ML with geospatial analysis to examine the impact of localized environmental factors, such as urban heat islands and industrial emissions, on community health [11, 12]. These models are further enhanced by incorporating temporal dynamics, enabling the prediction of long-term health outcomes based on historical exposure trends [13]. However, understanding these dynamics is a complex challenge due to the multidimensional nature of available data and their origin from heterogeneous sources such as environmental sensors, epidemiological databases, and health monitoring systems [14, 15, 16].

Building upon these advances, the application of Explainable Artificial Intelligence (XAI) is emerging as a crucial step in ensuring transparency and trustworthiness in ML models dealing with environmental and health data [17]. XAI provides techniques to make AI systems more interpretable, allowing researchers and policymakers to understand the underlying mechanisms driving model predictions [18].

In this work about environmental factors, ML models have a crucial in studying the interaction between exposome and human health, in particular with noncommunicable diseases [19]. The exposome plays a crucial role in understanding noncommunicable diseases (NCDs), which are significantly influenced by environmental exposures over a lifetime. NCDs such as cardiovascular diseases, diabetes, and respiratory illnesses are shaped by complex interactions between external exposures like air pollution, chemical toxins, and climate-related stressors, and internal factors including metabolic responses and the microbiome [20, 21, 22]. Explainable artificial intelligence highlights the most significant environmental contributors to NCDs, guiding policies on pollution control, urban planning, and public health interventions. Futhermore, XAI can suggest new hypotheses about how environmental changes influence health, guiding experimental and observational studies.

The goal of this work is to investigate the interplay between environmental factors, encapsulated by the exposome concept, and noncommunicable diseases (NCDs), leveraging advanced machine learning (ML) and explainable artificial intelligence (XAI) methods.

By integrating complex, multidimensional datasets from environmental and health domains, the study aims to uncover and quantify the relationships between environmental exposures and the onset or progression of NCDs such as cardiovascular diseases, diabetes, and respiratory illnesses. Furthermore, it seeks to provide interpretable insights through XAI techniques to ensure transparency and actionable results for stakeholders, including researchers, policymakers, and healthcare practitioners. This approach also facilitates the generation of new hypotheses on the influence of environmental changes on human health, driving future studies and targeted interventions, while supporting evidence-based policy formulation in areas like pollution control, urban planning, and public health.

This thesis explores predictive modeling and Explainable Artificial Intelligence (XAI) to analyze and predict mortality trends in Italy, focusing on transparency and actionable insights for public health. The first study develops a regression model to predict mortality from cancers of the respiratory system, identifying critical predictors such as environmental and socioeconomic factors, with XAI techniques providing insights into feature contributions. The second study compares various classification models to predict whether Italian provinces exhibit respiratory cancer mortality rates above the national average, leveraging XAI tools to enhance interpretability and robustness. The third study evaluates different regression approaches to predict mortality from endocrine, nutritional and metabolic diseases, highlighting key predictors and ensuring model transparency through XAI. Collectively, these studies demonstrate the effectiveness of combining predictive modeling with XAI to provide accurate, interpretable, and actionable insights to address critical public health challenges in Italy.

1.2 Machine Learning models

In recent decades, machine learning (ML) has emerged as a transformative field in artificial intelligence (AI), driving advances in diverse domains such as healthcare, finance, autonomous systems, and natural language processing [23, 24, 25, 26]. At its core, machine

learning is the study and development of algorithms that enable computers to identify patterns and make predictions or decisions without explicit programming. These algorithms leverage data to iteratively improve performance, offering solutions to problems that were previously considered intractable due to their complexity or scale [27].

The success of machine learning algorithms is fundamentally tied to the choice of methodology, the quality and quantity of data, and the computational resources available. Machine learning techniques are broadly categorized into supervised, unsupervised, semi-supervised, and reinforcement learning, each suited to specific types of problem and data characteristics. A key strength of supervised learning lies in its ability to generalize to new, unseen data when sufficient high-quality labeled examples are provided. However, its performance is heavily dependent on the availability of large, accurately labeled datasets.

Supervised learning, one of the most widely used paradigms, operates by training models on labeled datasets where each input is paired with a corresponding output. This approach enables the algorithm to learn a mapping function between inputs and outputs, making it particularly effective for tasks such as classification and regression.

Unsupervised learning, in contrast, does not rely on labeled data. Instead, it seeks to identify hidden patterns, structures, or relationships within the data. This makes it especially useful for exploratory data analysis, clustering, and dimensionality reduction.

Semi-supervised learning bridges the gap between supervised and unsupervised approaches by leveraging both labeled and unlabeled data. This is particularly advantageous when obtaining labeled data is costly, but large volumes of unlabeled data are readily available. In such cases, the algorithm uses the limited labeled examples to infer patterns and relationships in the unlabeled data, effectively amplifying its learning capacity. For instance, semi-supervised learning is widely used in natural language processing tasks, where only a fraction of textual data may have annotated labels, while the majority remains unlabeled. By combining the strengths of both supervised and unsupervised methods, semi-supervised learning provides a practical solution for domains like healthcare, where annotations are expensive and require expert input. However, its effectiveness depends on the assumption

that labeled and unlabeled data share similar underlying structures, which may not always hold true.

Another significant paradigm is reinforcement learning, which diverges from the above categories by focusing on learning through interaction with an environment. An agent learns to take actions in order to maximize cumulative rewards based on feedback from the environment. This approach is particularly well-suited to dynamic, sequential decision-making tasks, such as robotics, game-playing, and autonomous systems. Unlike supervised learning, reinforcement learning does not rely on explicit input-output pairs but instead learns by trial and error, guided by reward signals.

Each of these paradigms has its strengths and limitations, making them suitable for different applications depending on the problem domain, the nature of the data, and the specific requirements of the task. The choice between supervised, unsupervised, semi-supervised, or reinforcement learning is often dictated by the trade-offs between the availability of labeled data, computational resources, and the desired balance between accuracy and interpretability. Furthermore, ongoing advancements in self-supervised learning and hybrid approaches are increasingly blurring the boundaries between these categories, enabling the development of more flexible and robust solutions for real-world challenges.

The application of these methods in environmental health offers enormous potential for identifying complex correlations between ecological factors, pollution, and diseases, enhancing the ability to prevent and respond to health emergencies.

Classification is one of the most widely used machine learning techniques in environmental and health-related studies, as it allows researchers to categorize data points into predefined classes. In air quality monitoring, machine learning models classify pollution levels based on sensor data, helping policymakers implement timely interventions. In healthcare, classification is crucial for diagnosing diseases by analyzing medical records, imaging data, and genetic markers, enabling early detection of conditions such as cancer or respiratory diseases linked to environmental exposure. Additionally, in epidemiology, classification models predict disease outbreaks by analyzing climate, pollution, and demographic data,

improving public health preparedness. These applications demonstrate how classification enhances decision-making, risk assessment, and targeted interventions in critical domains.

Building on classification approaches, regression techniques play a crucial role in environmental and health sciences by predicting continuous numerical outcomes, offering deeper insights into risk factors and trends. In environmental science, regression models estimate pollution levels, temperature variations, and climate-related impacts, aiding in resource management and policy-making. For instance, linear and nonlinear regression methods predict air pollutant concentrations based on meteorological and industrial activity data, helping assess public exposure risks. Similarly, in healthcare, regression is used to analyze disease progression, forecast hospital admissions, and estimate the impact of environmental factors on chronic illnesses. The key difference between classification and regression lies in their outputs: classification assigns discrete labels to data points (e.g., categorizing air quality as "good" or "hazardous"), whereas regression predicts continuous values (e.g., estimating the concentration of pollutants in the air). Together, these approaches provide a comprehensive analytical framework, enabling data-driven decision-making in public health and environmental protection.

In contrast to classification and regression, clustering is an unsupervised learning approach that groups similar data points without predefined labels, making it particularly useful for discovering hidden patterns in environmental and health data. Unlike classification, which assigns known categories, clustering identifies natural groupings within datasets based on feature similarities. In environmental science, clustering techniques help segment geographic regions based on pollution levels, identify climate zones, and detect anomalous environmental patterns, such as sudden spikes in air contamination. For example, k-means clustering can group areas with similar air quality profiles, aiding policymakers in targeted interventions. In healthcare, clustering is used to categorize patients with similar symptoms or risk factors, improving disease diagnosis and personalized treatment strategies. Clustering also plays a crucial role in epidemiology, detecting outbreak patterns in infectious diseases by analyzing spatial and temporal distributions. By uncovering hidden structures in data, clustering provides valuable insights that complement supervised

learning methods, enabling more effective decision making in environmental monitoring and healthcare planning.

Despite its transformative impact, machine learning has several limitations, particularly in complex fields like environmental science and healthcare. One of the biggest challenges is data quality and availability: many models require large, well-labeled datasets, but in real-world applications, data can be noisy, incomplete, or biased [28]. For instance, environmental monitoring systems may have missing sensor readings [29], while medical records often contain inconsistencies that can affect model accuracy [30]. Additionally, interpretability and explainability remain major concerns, especially for deep learning models, which function as "black boxes" [31]. In high-stakes applications like disease diagnosis or air quality prediction, understanding why a model makes a certain decision is crucial for trust and regulatory compliance [32].

Generalization and robustness are also ongoing challenges, as models trained on specific datasets may perform poorly when applied to new environments with different conditions [33]. For example, a pollution prediction model trained in one city may not work well in another due to variations in industrial activities or climate [34]. Furthermore, machine learning models are susceptible to bias and fairness issues, particularly when data reflect historical inequalities, potentially leading to discriminatory outcomes in healthcare or environmental policies [35]. Lastly, computational complexity and energy consumption pose practical limitations, as training large models requires significant computational resources, raising concerns about sustainability and accessibility [36, 37]. Although continuous advances are addressing these limitations, ensuring ethical, interpretable, and efficient machine learning models remains a key challenge in their widespread adoption.

The typical machine learning workflow for analyzing geospatial data consists of several key stages, from data collection to model deployment, ensuring that spatial patterns and relationships are effectively captured and leveraged for decision-making.

The first step is data acquisition and preprocessing, where geospatial datasets are gathered from various sources, such as satellite imagery, remote sensing data, geographic information

systems (GIS), and environmental monitoring stations. This data often comes in different formats (raster, vector, time-series) and requires preprocessing steps like coordinate standardization, missing value imputation, and noise reduction [38]. Feature engineering is particularly crucial in geospatial analysis, as it involves extracting meaningful attributes, such as land use types, elevation, proximity to water bodies, and climate variables, which enhance model performance [39].

Next, exploratory data analysis (EDA) is performed to visualize spatial distributions, detect outliers, and identify patterns using geospatial techniques like spatial autocorrelation (Moran's I), kriging, and kernel density estimation [40]. This step helps researchers determine whether machine learning is appropriate and which algorithms might be most effective.

Following EDA, the model selection and training phase involves choosing suitable machine learning techniques based on the problem type. Supervised learning methods, such as classification and regression, are commonly used for tasks like land cover classification and climate prediction, whereas unsupervised learning methods such as clustering are applied to detect spatial anomalies or segment geographic regions [41]. Deep learning models, particularly convolutional neural networks (CNNs), are widely used for satellite image analysis due to their ability to extract complex spatial features [42].

Once trained, models undergo validation and evaluation using spatial cross-validation techniques, as traditional random splitting may not account for spatial dependencies [43]. Metrics like coefficient of determination (r^2) and mean absolute error (MAE) for regression, accuracy and F1-score for classification, and Silhouette Score for clustering help assess model performance. Overfitting is a common concern in geospatial applications, so regularization techniques and hyperparameter tuning are applied to improve generalization.

Finally, the model deployment and interpretation phase involves integrating predictions into GIS platforms, dashboards, or decision-support systems. Model interpretability is critical, as stakeholders in urban planning, environmental monitoring, and disaster response need transparent insights rather than black-box predictions. Tools like SHAP (Shapley

Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are increasingly used to explain spatial model outputs [44]. Continuous monitoring and retraining are essential, as environmental conditions and geospatial patterns evolve over time.

By following this structured workflow, machine learning enables effective analysis of geospatial data, supporting applications such as deforestation monitoring, climate modeling, and pollution forecasting, ultimately aiding in better decision-making for environmental and public health policies.

1.3 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a field of research dedicated to making machine learning models more transparent, interpretable, and trustworthy. As AI systems become increasingly integrated into critical decision-making processes, such as healthcare, environmental monitoring, finance, and autonomous systems, understanding how and why these models produce specific outputs is essential to ensure fairness, accountability, and reliability [45]. Traditional machine learning models, particularly deep learning architectures, often function as "black boxes", meaning that their internal decision-making processes are not easily understood by humans [46]. This lack of transparency can lead to ethical concerns, particularly when models inadvertently reinforce biases or make incorrect predictions without clear explanations.

To address these challenges, Explainable AI (XAI) techniques can be categorized into local and global interpretability methods, each serving different purposes in understanding model behavior.

Global explainability focuses on providing an overall understanding of how a model makes predictions across an entire dataset. This approach helps researchers and practitioners assess the importance of features, detect biases, and ensure the model aligns with domain knowledge. Methods such as feature importance rankings, partial dependence plots

(PDPs), and SHAP (Shapley Additive Explanations) help analyze how input variables influence predictions on a broader scale [47]. For instance, in environmental applications, a globally interpretable model might reveal that air pollution levels and temperature variations are the strongest predictors of respiratory disease outbreaks.

On the other hand, local explainability provides insights into individual predictions rather than the overall model behavior. This is particularly useful in high-stakes decision-making, where understanding why a specific output was generated is crucial. SHAP local values allows users to interpret specific decisions by approximating how the model would behave with slight modifications to the input data [48]. In healthcare, for example, local XAI can help explain why an AI model classified a particular patient as high-risk for heart disease, identifying whether factors like age, blood pressure, or cholesterol levels played a significant role. Similarly, in geospatial analysis, local explanations can clarify why a machine learning model predicts a specific region to be at high risk for deforestation, allowing policymakers to take targeted action.

XAI methods can be further categorized into model-specific and model-agnostic approaches, each offering different levels of flexibility and interpretability depending on the machine learning model used.

Model-specific methods are designed to work exclusively with certain types of models and leverage their internal structures to provide explanations. These methods are often more precise and computationally efficient because they directly interpret the model's learned representations. For example, in decision trees and random forests, feature importance scores and decision path analysis can be used to explain how the model arrived at a particular prediction [49]. However, the main limitation of model-specific methods is that they do not generalize well to different types of models, meaning a technique developed for neural networks may not be applicable to tree-based models or support vector machines. In contrast, model-agnostic methods are more flexible, as they can be applied to any machine learning model without requiring access to its internal workings. These methods typically approximate model behavior using surrogate models or perturbation-based

techniques. These techniques allow users to gain insight into complex models without sacrificing flexibility, making them widely used in fields like healthcare, finance, and environmental sciences. While model-agnostic methods provide greater generalization, they often involve higher computational costs and may introduce approximation errors, reducing the fidelity of explanations. On the other hand, model-specific methods offer more precise insights, but are restricted to certain model architectures. To balance interpretability and accuracy, researchers often combine both approaches, selecting the most suitable XAI technique based on the complexity of the model and the requirements of the application domain.

SHAP (Shapley Additive Explanations) values are a powerful method for interpreting machine learning models by assigning an importance score to each feature in a prediction. This technique is based on Shapley values, a concept from cooperative game theory introduced by Lloyd Shapley in 1953, which provides a mathematically fair way to distribute credit among players in a game [50]. In the context of machine learning, each feature is treated as a "player" that contributes to the final model prediction, and SHAP values quantify the average marginal contribution of each feature across all possible subsets of features [51]. This method provides both local and global interpretability, allowing users to understand individual predictions as well as overall model behavior. Unlike traditional feature importance measures, SHAP values maintain consistency and fairness, ensuring that features contributing more to a prediction receive higher importance scores. Additionally, SHAP is available in both model-agnostic and model-specific variants, making it applicable across different machine learning models while offering optimized versions for tree-based models to enhance computational efficiency. The method also enables intuitive visualizations, such as SHAP summary plots, dependence plots, and force plots, which help in uncovering how features influence predictions.

This thesis work is structured around three key studies that apply SHAP values to analyze the impact of environmental factors on health outcomes. The first two studies focus on mortality from respiratory cancers, examining how air pollution and socioeconomic conditions contribute to increased risks. The third study, instead, investigates mortality due

to metabolic, endocrine, and nutritional diseases, highlighting the role of environmental stressors such as air quality and exposure to endocrine disruptors.

1.4 Outline of the thesis

This thesis is structured as a collection of published or submitted research articles, each focusing on the application of machine learning and explainable artificial intelligence to investigate the relationship between environmental factors and non-communicable diseases. Each chapter corresponds to a distinct study, presenting a detailed analysis of specific health outcomes influenced by environmental exposures. The first chapter provides an overview of the theoretical foundations of ML and XAI, discussing their relevance in environmental health research. The subsequent chapters contain individual research articles, each applying ML models and SHAP-based interpretability techniques to examine different non-communicable diseases, such as respiratory cancers and metabolic disorders, in relation to air pollution, socioeconomic conditions, and other environmental stressors. Finally, the thesis concludes with a synthesis of the key findings, highlighting their implications for public health policies and future research directions in the field of environmental epidemiology and AI-driven health analytics.

The following list summarizes each case study in detail:

- **Chapter 2**. This chapter presents a study on the relationship between air pollution and mortality due to respiratory system cancers in Italy, using machine learning and explainable artificial intelligence. The study leverages remote sensing and socio-economic data to analyze Standardized Mortality Ratio (SMR) across Italian provinces. The methodology involves hierarchical spatial clustering, a Random Forest regressor, and SHAP values to identify the most relevant environmental and socio-economic factors influencing mortality rates. The results highlight nitrogen monoxide (NO), income levels, and ozone (O_3) concentrations as the key predictors of respiratory cancer mortality, providing valuable insights for public health

interventions and environmental policy. This work demonstrates how XAI enhances transparency in AI-driven epidemiological research, offering a data-driven framework for assessing environmental health risks and guiding targeted preventive measures.

- **Chapter 3**. Building on the previous study, which applied a regression-based approach, Chapter 3 contains a work that explores a comparative analysis of classification models to predict whether Italian provinces have above- or below-average Standardized Mortality Ratios for respiratory cancers. By evaluating thirteen machine learning algorithms, including Random Forest, Support Vector Machines and Gradient Boosting, this study aims to determine the most effective classification approach. The analysis highlights differences in predictive performance, demonstrating how various models handle complex interactions between air pollution, socioeconomic conditions, and urbanization. SHAP values are then used to interpret the best performing model, reinforcing previous findings and expanding on the role of NOx, benzene, and income levels as significant risk factors. The results further validate the importance of XAI in epidemiological studies, offering a broader methodological perspective for identifying environmental determinants of respiratory cancer mortality.

- **Chapter 4**. The final chapter, titled "Explainable artificial intelligence for endocrine, nutritional and metabolic mortality in Italy with geolocalized pollution data", examines the risk of death for endocrine, nutritional and metabolic diseases for environmental factors. As the work presented in the Chapter 2, this study exploited hierarchical spatial clustering to identify clusters of provinces with similar mortality and compare the performances of three different machine learning models evaluated with a leave-on-cluster-out cross-validation. Finally, the SHAP values of the best model were evaluated, identifying which features had the most impact on the prediction of the model.

In conclusion, this thesis explores the application of machine learning and explainable artificial intelligence (XAI) in understanding the complex relationship between environmental

factors and mortality rates in Italy. Through a series of interconnected studies, this work demonstrates how advanced AI techniques can reveal key insights into the influence of air pollution, socio-economic conditions, and urbanization on public health outcomes, specifically focusing on respiratory and endocrine-related diseases. The use of SHAP values not only aids in model interpretability but also enhances the transparency and trustworthiness of AI-driven epidemiological research. Ultimately, this thesis contributes valuable data-driven frameworks for guiding public health interventions, informing environmental policy, and advancing the integration of XAI in health risk assessment.

Bibliography

[1] D. Rojas-Rueda, E. Morales-Zamora, W. A. Alsufyani, C. H. Herbst, S. M. AlBalawi, R. Alsukait, and M. Alomran, "Environmental risk factors and health: an umbrella review of meta-analyses," *International journal of environmental research and public health*, vol. 18, no. 2, p. 704, 2021.

[2] C. P. Wild, "Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 8, pp. 1847–1850, 2005.

[3] C. Krittanawong, Y. K. Qadeer, R. B. Hayes, Z. Wang, G. D. Thurston, S. Virani, and C. J. Lavie, "Pm2. 5 and cardiovascular diseases: State-of-the-art review," *International Journal of Cardiology Cardiovascular Risk and Prevention*, p. 200217, 2023.

[4] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC public health*, vol. 17, pp. 1–19, 2017.

[5] X. Wu, Q. Zhou, L. Mu, and X. Hu, "Machine learning in the identification, prediction and exploration of environmental toxicology: Challenges and perspectives," *Journal of Hazardous Materials*, vol. 438, p. 129487, 2022.

[6] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. Loncar Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, *et al.*, "Applications of machine learning in human microbiome studies: a review on feature

selection, biomarker identification, disease prediction and treatment," *Frontiers in microbiology*, vol. 12, p. 634511, 2021.

[7] R. B. Ghannam and S. M. Techtmann, "Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1092–1107, 2021.

[8] L. Yu, W. Liu, X. Wang, Z. Ye, Q. Tan, W. Qiu, X. Nie, M. Li, B. Wang, and W. Chen, "A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture," *Environmental Pollution*, vol. 306, p. 119356, 2022.

[9] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review," *Journal of biomedical informatics*, vol. 73, pp. 14–29, 2017.

[10] B. G. Patra, M. M. Sharma, V. Vekaria, P. Adekkanattu, O. V. Patterson, B. Glicksberg, L. A. Lepow, E. Ryu, J. M. Biernacka, A. Furmanchuk, *et al.*, "Extracting social determinants of health from electronic health records using natural language processing: a systematic review," *Journal of the American Medical Informatics Association*, vol. 28, no. 12, pp. 2716–2727, 2021.

[11] F. Li, T. Yigitcanlar, M. Nepal, K. Nguyen, and F. Dur, "Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework," *Sustainable Cities and Society*, vol. 96, p. 104653, 2023.

[12] D. Koldasbayeva, P. Tregubova, M. Gasanov, A. Zaytsev, A. Petrovskaia, and E. Burnaev, "Challenges in data-driven geospatial modeling for environmental research and practice," *Nature Communications*, vol. 15, no. 1, p. 10700, 2024.

[13] F. A. Prodhan, J. Zhang, S. S. Hasan, T. P. P. Sharma, and H. P. Mohana, "A review of machine learning methods for drought hazard monitoring and forecasting: Current

research trends, challenges, and future research directions," *Environmental modelling & software*, vol. 149, p. 105327, 2022.

[14] C. P. Wild, "The exposome: from concept to utility," *International journal of epidemiology*, vol. 41, no. 1, pp. 24–32, 2012.

[15] G. W. Miller and D. P. Jones, "The nature of nurture: refining the definition of the exposome," *Toxicological sciences*, vol. 137, no. 1, pp. 1–2, 2014.

[16] R. Vermeulen, E. L. Schymanski, A.-L. Barabási, and G. W. Miller, "The exposome and health: Where chemistry meets biology," *Science*, vol. 367, no. 6476, pp. 392–396, 2020.

[17] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Information Fusion*, vol. 96, pp. 156–191, 2023.

[18] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (xai)," *IEEe Access*, 2023.

[19] W. Kanyongo and A. E. Ezugwu, "Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives," *Informatics in Medicine Unlocked*, vol. 38, p. 101232, 2023.

[20] R. E. Norman, D. O. Carpenter, J. Scott, M. N. Brune, and P. D. Sly, "Environmental exposures: an underrecognized contribution to noncommunicable diseases," *Reviews on environmental health*, vol. 28, no. 1, pp. 59–65, 2013.

[21] M. Dhimal, T. Neupane, and M. Lamichhane Dhimal, "Understanding linkages between environmental risk factors and noncommunicable diseases—a review," *FASEB BioAdvances*, vol. 3, no. 5, pp. 287–294, 2021.

[22] L. Di Renzo, P. Gualtieri, G. Frank, R. Cianci, M. Caldarelli, G. Leggeri, G. Raffaelli, E. Pizzocaro, M. Cirillo, and A. De Lorenzo, "Exploring the exposome spectrum:

Unveiling endogenous and exogenous factors in non-communicable chronic diseases," *Diseases*, vol. 12, no. 8, p. 176, 2024.

[23] K. Shailaja, B. Seetharamulu, and M. Jabbar, "Machine learning in healthcare: A review," in *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pp. 910–914, IEEE, 2018.

[24] N. Nazareth and Y. V. R. Reddy, "Financial applications of machine learning: A literature review," *Expert Systems with Applications*, vol. 219, p. 119640, 2023.

[25] R. G. Goriparthi, "Ai-driven predictive analytics for autonomous systems: A machine learning approach," *Revista de Inteligencia Artificial en Medicina*, vol. 15, no. 1, pp. 843–879, 2024.

[26] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, *et al.*, "Machine learning and natural language processing in mental health: systematic review," *Journal of medical Internet research*, vol. 23, no. 5, p. e15708, 2021.

[27] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[28] N. J. Isaac, A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy, "Statistics for citizen science: extracting signals of change from noisy ecological data," *Methods in Ecology and Evolution*, vol. 5, no. 10, pp. 1052–1060, 2014.

[29] E. Elnahrawy and B. Nath, "Cleaning and querying noisy sensors," in *Proceedings of the 2nd ACM international conference on Wireless sensor networks and applications*, pp. 78–87, 2003.

[30] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.

[31] A. Rocha, J. P. Papa, and L. A. Meira, "How far do we get using machine learning black-boxes?," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 02, p. 1261001, 2012.

[32] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[33] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mechanical systems and signal processing*, vol. 100, pp. 439–453, 2018.

[34] A.-L. Balogun, A. Tella, L. Baloo, and N. Adebisi, "A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science," *Urban Climate*, vol. 40, p. 100989, 2021.

[35] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data & Society*, vol. 4, no. 2, p. 2053951717743530, 2017.

[36] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications surveys & tutorials*, vol. 18, no. 1, pp. 732–794, 2015.

[37] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges," *Journal of Cleaner Production*, p. 143090, 2024.

[38] T. Yu, L. Li, L. Chen, and W. Song, "Low-quality multivariate spatio-temporal serial data preprocessing," *Cluster Computing*, vol. 22, pp. 2357–2370, 2019.

[39] R. R. Vincent, E. Sakthivel, M. Kumari, F. Nisha, and A. Rohini, "Machine learning for geospatial analysis: Enhancing spatial understanding and decision-making," in *Ethics, Machine Learning, and Python in Geospatial Analysis*, pp. 168–195, IGI Global, 2024.

[40] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan, "Identifying patterns in spatial information: A survey of methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 193–214, 2011.

[41] J. Wang and F. Biljecki, "Unsupervised machine learning in urban studies: A systematic review of applications," *Cities*, vol. 129, p. 103925, 2022.

[42] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018.

[43] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren, "Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing," *Machine Learning*, vol. 111, no. 7, pp. 2715–2740, 2022.

[44] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost," *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022.

[45] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[46] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[47] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[48] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021.

[49] D. Deb and R. M. Smith, "Application of random forest and shap tree explainer in exploring spatial (in) justice to aid urban planning," *ISPRS International Journal of Geo-Information*, vol. 10, no. 9, p. 629, 2021.

[50] L. S. Shapley, "A value for n-person games," *Contribution to the Theory of Games*, vol. 2, 1953.

[51] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

Chapter 2

AI-Driven Insights on Air Pollution and Respiratory Cancer Mortality in Italy

This chapter explores the relationship between air pollution and mortality from cancer of the respiratory system in Italian provinces. The study employs explainable artificial intelligence (XAI) and machine learning (ML) techniques to identify key environmental and socio-economic factors influencing mortality rates.

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 07 May 2024
DOI 10.3389/fpubh.2024.1344865

Check for updates

OPEN ACCESS

EDITED BY
Uttara Saran,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Arlette Setiawan,
Padjadjaran University, Indonesia
Karthikeyan Bagavathy Shanmugam,
Emory University, United States

*CORRESPONDENCE
Roberto Cilli
✉ roberto.cilli@uniba.it
Sabina Tangaro
✉ sabina.tangaro@uniba.it

†These authors have contributed equally to this work

RECEIVED 26 November 2023
ACCEPTED 08 April 2024
PUBLISHED 07 May 2024

CITATION
Romano D, Novielli P, Cilli R, Amoroso N, Monaco A, Bellotti R and Tangaro S (2024) Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach.
Front. Public Health 12:1344865.
doi: 10.3389/fpubh.2024.1344865

COPYRIGHT
© 2024 Romano, Novielli, Cilli, Amoroso, Monaco, Bellotti and Tangaro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach

Donato Romano[1,2], Pierfrancesco Novielli[1,2], Roberto Cilli[2,3]*, Nicola Amoroso[2,4], Alfonso Monaco[2,3], Roberto Bellotti[2,3†] and Sabina Tangaro[1,2*†]

[1]Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti Universita' degli Studi di Bari Aldo Moro, Bari, Italy, [2]Istituto Nazionale di Fisica Nucleare Sezione di Bari, Bari, Italy, [3]Dipartimento Interateneo di Fisica, "M. Merlin" Universita' degli Studi di Bari Aldo Moro, Bari, Italy, [4]Dipartimento di Farmacia Scienze, del Farmaco Universita' degli Studi di Bari Aldo Moro, Bari, Italy

Respiratory system cancer, encompassing lung, trachea and bronchus cancer, constitute a substantial and evolving public health challenge. Since pollution plays a prominent cause in the development of this disease, identifying which substances are most harmful is fundamental for implementing policies aimed at reducing exposure to these substances. We propose an approach based on explainable artificial intelligence (XAI) based on remote sensing data to identify the factors that most influence the prediction of the standard mortality ratio (SMR) for respiratory system cancer in the Italian provinces using environment and socio-economic data. First of all, we identified 10 clusters of provinces through the study of the SMR variogram. Then, a Random Forest regressor is used for learning a compact representation of data. Finally, we used XAI to identify which features were most important in predicting SMR values. Our machine learning analysis shows that NO, income and O3 are the first three relevant features for the mortality of this type of cancer, and provides a guideline on intervention priorities in reducing risk factors.

KEYWORDS

explainable artificial intelligence, air pollution, lung cancer, respiratory disease, socio-economic indices, public health, remote sensing 2010 MSC: 00-01, 99-00

1 Introduction

This study aims to investigate the relationship between air pollution and respiratory system cancer mortality in Italian provinces. Air pollution is a pressing issue of the modern world, impacting human health and the environment in numerous ways (1). One of its significant consequences is its link to respiratory system cancer, a malignancy that claims millions of lives globally each year (2). Air pollution, which mainly consists of fine particulate matter and toxic gases, can enter the lung tissue and trigger a series of chronic inflammatory and oxidative damage to the cells, leading to malignant transformation.

Respiratory system cancer is a type of cancer that affects the lungs, bronchi and trachea. The most common types of respiratory system cancer are lung cancer and bronchial cancer. (3). While smoking is a primary risk factor, there are other important contributing factors, such as secondhand smoke, occupational exposure, and air pollution. This sets the stage for the focus of the study (3). It is estimated that 14% of lung cancer deaths are attributable

to environmental air pollution (4). Symptoms of respiratory system cancer can include persistent cough, chest pain, shortness of breath, hoarseness, fatigue, and weight loss.

A number of works have dealt with long and short-term effect of air pollution and cancer (5, 6), highlighting a multitude of contributing factors such as the type of pollutant, exposure time and frequency, and individual susceptibility (7): an increase of 10 $\mu g/m3$ in PM10 concentration raises the average likelihood of developing lung cancer by 20%, whereas a 5 $\mu g/m3$ rise in PM2.5 elevates the risk by 30% (8). Recognizing and addressing the link between air pollution and respiratory system cancer is critical for protecting public health on a global scale. It informs evidence-based policies, encourages international collaboration, and empowers individuals and communities to take actions that can mitigate the impact of air pollution on respiratory health.

The goal of this work was to implement a Machine Learning (ML) algorithm predicting the standard mortality ratio (SMR) for lung, bronchi and trachea cancer of the Italian provinces by using air pollution data downloaded from Copernicus Atmosphere Monitoring Service (CAMS) and socio-economic data downloaded from ISTAT. ML is the discipline dealing with the replication of the learning mechanisms of the human brain through statistical algorithms (9). Random forest is a machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that creates several decision trees and combines their predictions to make a final decision or prediction. Recent advancements in ML techniques resulted in the introduction of eXplainable Artificial Intelligence (XAI) which allows for the identification of the crucial attributes for each instance (10–12). Explainable AI provides clarity and understanding into the decision-making processes of AI models. It helps improve transparency, trust, accountability, and compliance with regulations. XAI methods has been applied to ML algorithms to provide a clear picture of the relevant features affecting the performance of the models, their relations with the outcomes, with each other's and both their local and global effects. The main intent of our study was to present a framework to determine which pollution indices, based on remote sensing data, are most associated to the mortality from cancer of the respiratory system. We studied the mortality in the Italian provinces given its heterogeneity in terms of ecological and environmental features. In order to do this, we evaluated to what extent mortality from cancer of the respiratory system can be predicted based on environmental pollution and socio-economic indices.

2 Materials

The study was conducted using mortality data from ISTAT, that is the Italian Institute of Statistics, responsible for collecting, analyzing, and disseminating official statistics on the country's population, economy, and society.[1] ISTAT's main functions include conducting population censuses, compiling and publishing official statistics on topics such as, employment, and economic indicators, and providing support and expertise to other public institutions and organizations in the field of statistics. In particular, the respiratory system cancer mortality at the provinces level in 2019 has been considered.

In this work, we mainly exploited pollutants' concentration from the Copernicus Air Monitoring Service (CAMS). An outline of the data preparation workflow is presented in Figure 1. In order to get an analysis-ready data table we firstly collected daily air quality maps from the Copernicus Data Store for the year 2019.[2]

2.1 Input data preparation

The pollution data of year 2019 has been downloaded from Copernicus Atmosphere Monitoring Service (CAMS). It is a European Union program that provides comprehensive information on air quality and the Earth's atmosphere.

It aims to improve air quality forecasts and support decision-making related to air quality management and environmental policy-making in Europe and around the world. CAMS uses a wide range of independent monitoring and modeling systems to collect and analyze data on atmospheric composition, air quality, and weather patterns. CAMS provides annual air quality reanalyzes for Europe based on both unvalidated and validated observations.[3] Since the downloaded data covered a larger area than that of interest, only pollution data of the Italian peninsula were extracted using the Python library GeoPandas.[4] Then, for each selected pollutant, both annual average and standard deviation were computed. The polluting substances considered were: carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter 10 (pm10), particulate matter 2.5 (pm2.5) and sulfur dioxide (SO_2). Pollutant values are the result of an ensemble median of 11 state-of-the-art numerical air quality models developed in Europe: CHIMERE from INERIS (France) (13), EMEP from MET Norway (Norway) (14), EURAD-IM from Jülich IEK (Germany) (15), LOTOS-EUROS from KNMI and TNO (Netherlands) (16), MATCH from SMHI (Sweden) (17), MOCAGE from METEO-FRANCE (France) (18), SILAM from FMI (Finland) (19), DEHM from AARHUS UNIVERSITY (Denmark) (20), GEM-AQ from IEP-NRI (Poland) (21), MONARCH from BSC (Spain) (22) and MINNI from ENEA (Italy) (23). The yearly ensemble reanalyzes are available with a time resolution of 1 h from step 1st January to the 31st of December, while the horizontal resolution of ensemble reanalyzes is on a $0.1^\circ \times 0.1^\circ$ regular latitude-longitude grid. The analysis was conducted using 2019 data since pollution data available on the CAMS platform starts from April 2018. Additionally, years following 2019 were not considered due to the COVID epidemic that

1 Available online at: https://www.istat.it/it/archivio/222527.

2 Available online at: https://atmosphere.copernicus.eu/.

3 Available online at: https://ads.atmosphere.copernicus.eu/#!/search?text=andtype$=$datasetandkeywords$=$((%20%22Product%20type:%20Reanalysis%22%20)%20AND%20(%20%22Variable%20domain:%20Atmosphere%20(composition)%22%20)%20AND%20(%20%22Spatial%20coverage:%20Europe%22%20)%20AND%20(%20%22Temporal%20coverage:%20Past%~22%20)).

4 Available online at: https://geopandas.org/en/stable/.

Romano et al. 10.3389/fpubh.2024.1344865

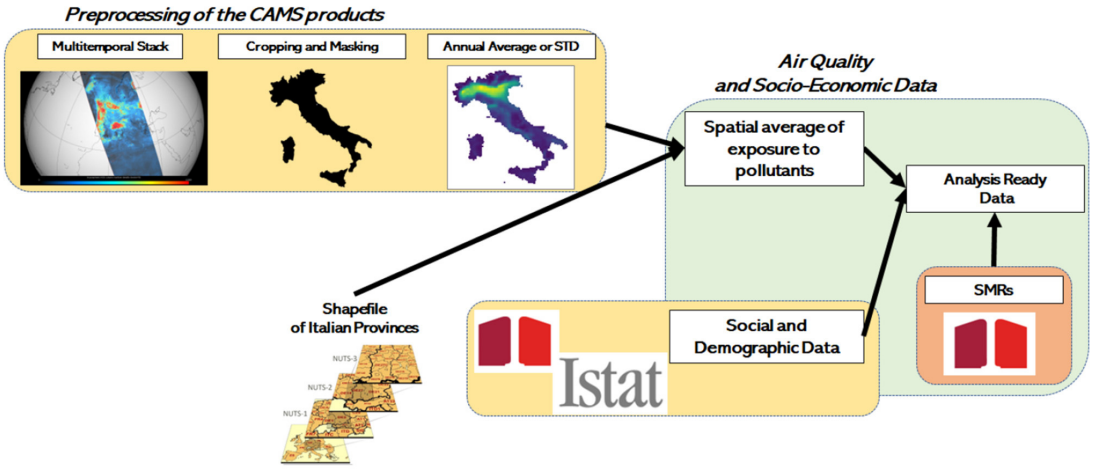

FIGURE 1
Data preparation workflow.

TABLE 1 Mean and standard deviation values of air pollution data by province and pollutant.

Pollutants	Northern Italy	Center Italy	Southern Italy
Pm2.5$(\frac{\mu g}{m^3})$	15.55 ± 5.42	10.94 ± 1.11	11.37 ± 2.15
pm10 $(\frac{\mu g}{m^3})$	19.09 ± 6.03	15.56 ± 1.37	16.33 ± 2.29
CO$(\frac{\mu g}{m^3})$	194.04 ± 45.17	156.58 ± 9.24	150.85 ± 20.63
NO$(\frac{\mu g}{m^3})$	1.23 ± 1.21	0.34 ± 0.16	0.33 ± 0.3
$NO_2(\frac{\mu g}{m^3})$	13.15 ± 6.49	7.32 ± 2.55	6.17 ± 3.5
$SO^2(\frac{\mu g}{m^3})$	1.56 ± 0.66	1.01 ± 0.27	1.33 ± 0.51
$O^3(\frac{\mu g}{m^3})$	58.88 ± 7.79	67.82 ± 5.88	69.63 ± 7.29

began in the early months of 2020, a disease that increased deaths from respiratory illnesses. The average values and standard deviations of pollutants for the provinces of Northern, Central, and Southern Italy are reported in Table 1. Subsequently, in order to get a structured data table to be added to socio-demographic descriptors, a spatial average of these intermediary maps have been computed within the boundaries of each Italian province. Other pollution-related and socio-economic variables were considered to enrich the data set, in particular: cultivated areas, urban areas, benzene, temperature, N fertilizer, P4010 fertilizer, microelement fertilizer, organic fertilizer, bed number, which represents the number of hospital beds available, lifetime, income, life quality, instruction, vehicles total, urban traffic, photovoltaic panel, green urban, electric consumption, noise and wastes (24).

The complete list of the selected descriptors is reported in Table 2.

2.2 Output data preparation: the indirect standardization

Indirect standardization is a statistical method used to compare the rates of events or conditions in two or more populations while controlling for differences in population characteristics such as age, sex or socioeconomic status. This method calculates expected rates for each population by applying the distribution of a particular population characteristic (for example age) to a reference population with known rates (25). The observed rates of a particular outcome in each population are then compared to the expected rates, which have been adjusted for any differences in population characteristics. In this work we computed the standard mortality ratio (SMR), the ratio between the deaths observed in a territory and those expected in the same. The expected deaths were calculated by applying the corresponding specific mortality ratios of the population assumed as standard (the national one in this case) to the average annual population by age classes of each territorial unit. The average standardized mortality ratio (SMR for cancer of the respiratory system) distribution within Italian provinces is shown in Figure 2A. Moreover, the province of Sondrio was removed from our analysis since it appears to be an outlier in the SMR distribution, as shown in the Figure 2B.

3 Methods

The main goal of this work is to find through an XAI algorithm, which air pollutants and/or socio-economic index contribute most to mortality from lung, trachea and bronchial cancer in Italian provinces. A cross-validation framework has been implemented to train a Random Forest regressor (RF) (26) of the standard mortality ratio for respiratory cancer in Italian provinces; then, by using SHapley Additive exPlanations (27), a method based on cooperative game theory and used, we performed an explain ability analysis

Romano et al. 10.3389/fpubh.2024.1344865

TABLE 2 Overview of the variables selected for this study.

Data	Type	Source	Value	Coding
Pollutant	Time series $\left(\frac{\mu g}{m^3}\right)$	CAMS	Mean $\left(\frac{\mu g}{m^3}\right)$, Standard deviation $\left(\frac{\mu g}{m^3}\right)$	mean pm2.5, std pm2.5, mean pm10, std pm10, mean CO, std CO, mean NO, std NO, mean NO_2, std NO_2, mean O_3, std O_3, mean SO_2, std SO2
Anthropogenic		ISTAT	Mean	N fertilizer, P410 fertilizer, Microelement fertilizer, Organic fertilizer, bed number, life time, income, life quality, instruction, Vehicles total, urban traffic, Electric Consumption, noise, wastes
Environment		ISTAT	Mean	cultivated areas, urban areas, benzene, temperature, photovoltaic panel, green urban

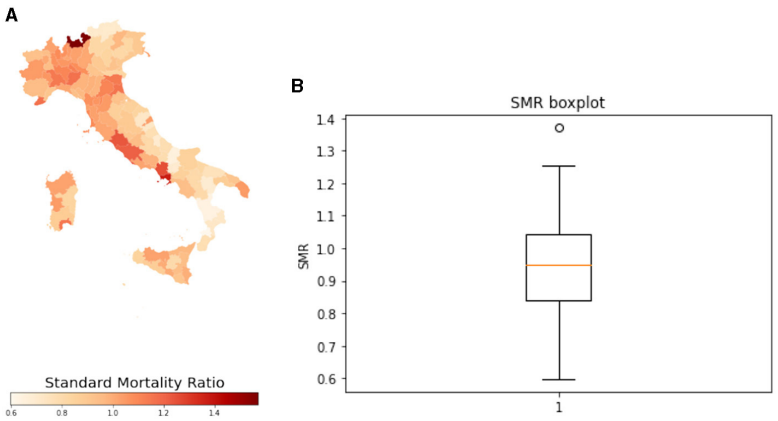

FIGURE 2
(A) Standardized mortality ratio distribution in 2019 within Italian provinces. **(B)** Boxplot of standardized mortality ratio for cancer of the respiratory system of the Italian provinces in 2019.

to increase transparency and interpretability of machine learning model. The Figure 3 shows a schematic overview of the methods adopted in the present work.

3.1 Random forest regressor of standard mortality ratio

Random Forest (RF) operates as an ensemble learning classifier rooted in the concept of classification trees. RF essentially creates a collection of classification trees, wherein each tree is trained on a bootstrapped sample from the available data. To prevent biased estimations, one-third of the available examples are excluded and utilized for the out-of-bag error estimation.

In the process of tree growth, the ideal split at each node relies on a random selection of M/3 descriptors where M represents the total available descriptors. It has been shown that the classification error is influenced by two primary factors: the interdependence among trees in the forest and the individual predictive power of each tree. Managing these factors involves adjusting the number of trees in the forest and the quantity of features sampled per split. The accuracy of RF models substantially depends on two parameters, the number of sampled features f and the number of the forest trees T.

In this study, RF was implemented using the random Forest routine from the scikit-learn package (v 1.2.1) (28) with its default configuration.

3.2 Hierarchical spatial clustering and model performance assessment

In order to avoid information leakage between spatially dependent observations, we preliminary found a partition of clusters of adjacent provinces to keep apart during the training and validation steps.

The spatial clusters of provinces adopted in the cross-validation scheme were found through a combined use of a hierarchical clustering algorithm applied to the matrix of the euclidean distances between provinces, and the semivariogram plot of the SMRs. A semivariogram plot is a tool used in geostatistics to assess the spatial dependence of a variable. It is usually plotted on a graph where the x-axis corresponds to the distance between a pair of selected locations (also known as spatial lag), while the y-axis corresponds to the average

squared difference of the variable values computed for those pairs of location within a given distance interval. In this work, we estimated the spatial range of the semivariogram, i.e., the distance after which the observations are supposed to be no longer correlated, using a discretionary approach by a visual identification of the plateau (i.e., when no further increase in variance is observed).

Establishing the range of the empirical semivariogram of the SMRs allowed us to delineate ten clusters after thresholding the dendrogram obtained by the hierarchical clustering of the geographical distances of the Italian provinces (Figure 4).

We utilized a leave-one-cluster-out cross-validation approach to mitigate spatial data bias when assessing the regression performance of an ensemble Random Forest regressor. This validation scheme was implemented to prevent overestimation of performance. In fact, spatial autocorrelation in two adjacent provinces, one selected in the training set and the other in validation, may lead to overoptimistic results.

Finally, we evaluated the performance of our machine learning model, by computing the coefficient of determination r^2 (Equation 1) and the mean absolute error *MAE* (Equation 2), whose definition is provided in the following:

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|(y_i - \hat{y}_i)\right| \quad (2)$$

where $\hat{y}_i$ are the predicted values, $\bar{y}$ is their average, while y_i are the observed values of the SMR in the validation set.

3.3 Features explain ability

Epidemiological studies are crucial in understanding and controlling the spread of diseases, but often require large sets of data that are difficult for humans to analyze efficiently. Artificial intelligence (AI) has emerged as a useful tool in epidemiological studies, with promising applications in predicting disease outbreaks, identifying risk factors, and developing targeted interventions. However, as AI becomes more prevalent in the field, there is a growing concern about its lack of transparency and explain ability, which can limit its utility and undermine the trust in its results. Explainable artificial intelligence (XAI) can address these concerns by providing interpretable models, transparent decision-making processes, and clear explanations of the AI's predictions and recommendations. In this work, we used SHapley Additive exPlanations (SHAP) method, a XAI algorithm borrowed from game theory (29, 30). The SHapley Additive exPlanations (SHAP) method is a model-agnostic approach to interpret the output of any machine learning model. It provides a unified framework for interpreting the predictions of any model by assigning a feature importance value to each input feature (Equation 3).

SHAP method values how a feature affects the performance of the model on the validation set by including and removing it from the model:

$$\Phi j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!\,(|S| - |F| - 1)!}{|S|!}\left[fx(F \cup j) - fx(F)\right] \quad (3)$$

where x is an instance, the sum is over all the subsets S of features which include the feature j, $\frac{|F|!(|S|-|F|-1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that doesn't belong to S, while $fx(F \cup j)$ and $fx(F)$ denote respectively the regression score obtained by including and non-including feature j.

4 Results

The goal of this work was to evaluate, through explainable machine learning models, whether and to what extent air pollutants and socio-economic descriptors are associated with mortality due to respiratory cancer.

Following the procedures described in the methods section, we delineated 10 clusters when using the spatial range value from the semivariogram plot to split the dendrogram of the geographical distances between the Italian provinces. Then, a leave-one cluster-out validation scheme was adopted in order to get a robust assessment of the model performance. We quantitatively assessed the model performance in terms of the average metrics obtained on the validation set; our model achieved an $r2$ value of 0.28 and a mean absolute error (*M AE*) value of 0.10. Moreover, the importance of all available variables included in the analysis was computed by exploiting the permutation feature importance algorithm and plotted in Figure 5A. Finally, a scatter plot displaying the mutual agreement between the actual and predicted Standard Mortality Ratios is shown in Figure 5B. The provinces characterized by the highest SMR in the 2019 are located in the right-most part of the scatterplot; these are the provinces of Napoli, Caserta, Viterbo, Roma, Piacenza, Imperia, Ravenna, Cagliari, Alessandria and Ferrara.

Figure 6 shows the most important features for regression according to the SHAP algorithm. This summary plot offers an overview of the varying degrees of influence of every feature on the model's predictions, thereby enabling a better grasp of the comprehensive significance and effect of distinct features in the analysis.

According to the SHAP summary plot, the top 5 most important variables include three related to pollution (std NO, mean O_3, mean NO_2), one associated with climate (temperature), and another tied to social factors (income).

5 Discussion

The performances obtained in terms of r2 and MAE are respectively 0.28 and 0.10. This result shows that environmental pollution is associated to the considered type of cancer. From the analysis conducted, it emerges that there is an association between certain pollutants and the incidence of respiratory system cancer.

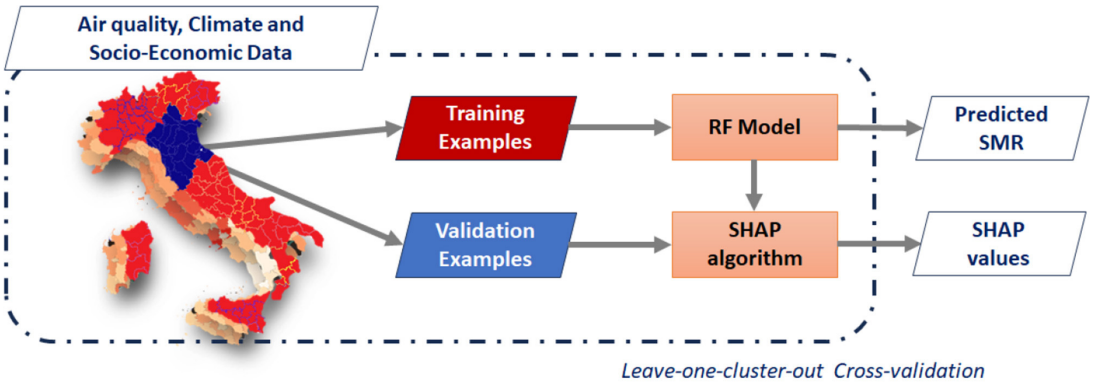

FIGURE 3
Methodological overview: climate, annual average and standard deviation pollution concentrations and socio-economic variables are used to model the standard mortality ratios (SMRs) in the Italian provinces due to cancer of the respiratory system. Then, the SHAP algorithm is used to compute the contribution of each input variable to the predicted SMRs.

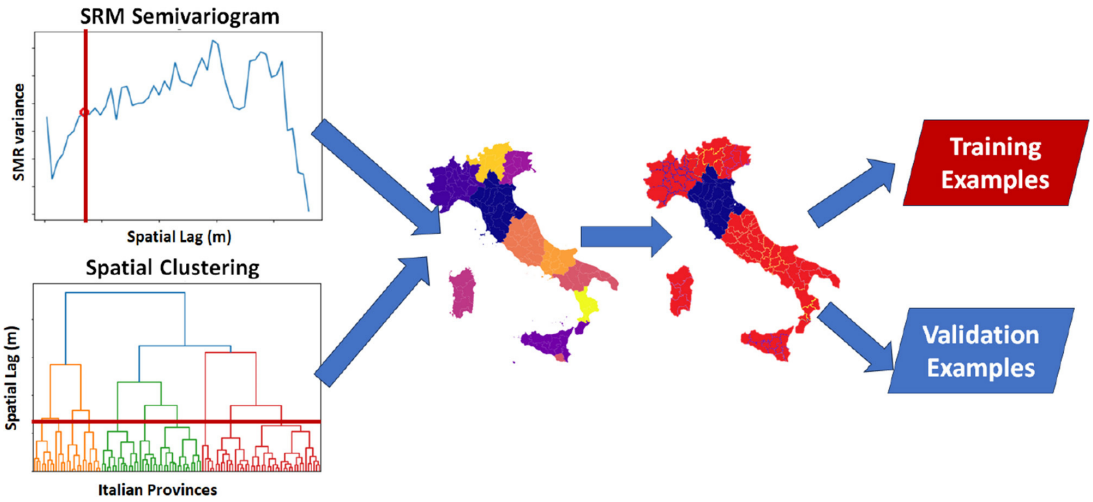

FIGURE 4
The hierarchical clustering algorithm used in combination with the semivariogram plot of the SMR rates delineated 10 spatially-contiguous community that alleviated the spatial data leakage that might occur in epidemiological studies.

This information should be included alongside other risk factors in studies investigating risk factors for individual subjects, including personalized ones (31, 32).

The spatial auto-correlation analysis, performed by using the variogram of the standard mortality ratio, shown ten clusters of provinces. The cross validation performed by using these clusters, obtained an r2 lower than one with a random cross-validation, by showing a spatial bias that overestimated the random forest regressor performance.

XAI estimated which are the most important global and local features in predicting SMR. In particular, for the present case, the main role played by the standard deviation of NO and mean O_3 were revealed. This result is consistent with the literature, as it is known that a greater exposure to NO_2 is correlated with a greater risk of developing lung cancer (33). According to our study, a low average O_3 concentration is linked to a greater SMR for tumors of the respiratory system. Such a spurious association have been observed before in previous works, including Dutton et al. (34) and Travaglio et al. (35), and we believe that a plausible interpretation is that increased ozone exposure is acting a as proxy of residing in rural areas (36, 37). Moreover, it is also known that the concentration of O_3 is anti-correlated to the concentration

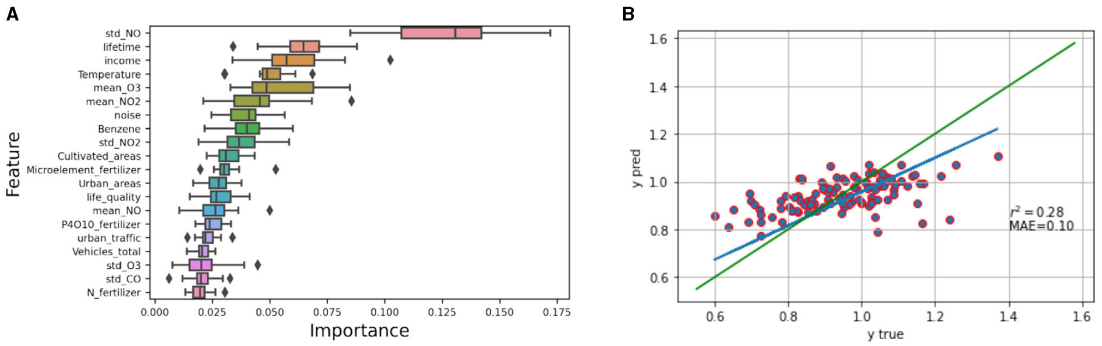

FIGURE 5
(A) Random forest features importance. **(B)** Scatter plot of predicted Standard Mortality Ratio (y pred) vs. true Standard Mortality Ratio (y true). The blue line represents the best fit for the plotted points, whereas the green one represents where the points would fall if all predicted values perfectly matched the observed ones.

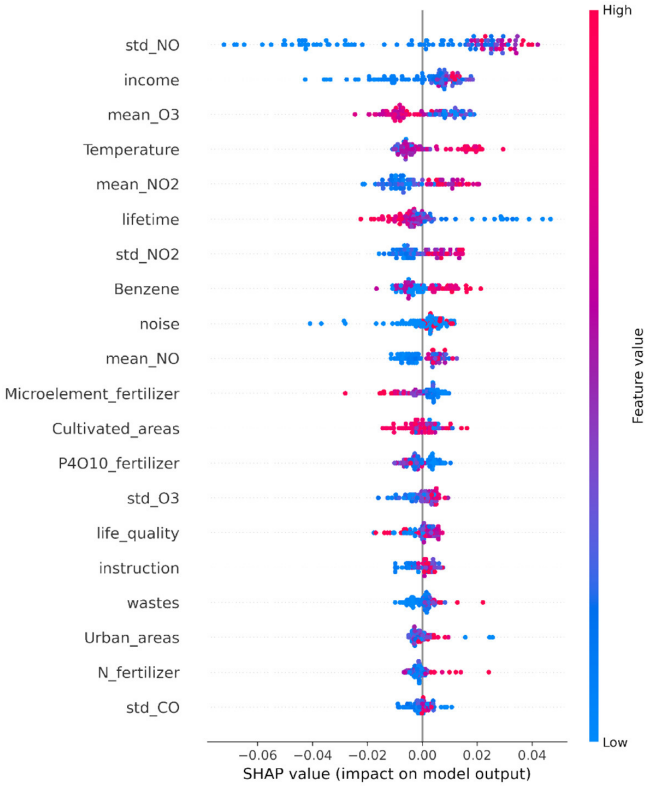

FIGURE 6
SHAP summary plot illustrating the SHAP values for each feature. Each point on the plot represents a Shapley value of a provinces, with the y-axis indicating the corresponding feature and the x-axis representing the Shapley value itself. The color gradient reflects the feature value, ranging from low to high. The features are ordered based on their mean importance, with more important features positioned toward the top.

of NO_2 (36–42). This spatial pattern between rural areas and O_3 concentration is rather general, as O_3 in troposphere is a secondary pollutant that is produced after NO_2 reacting with UV light (40, 43). We also believe that the observed negative correlation between NO_2 and O_3 can be largely attributable to this causal link.

Concerning the socio-demographic descriptors, income and lifetime were the most important. Higher income appears to be positively associated with SMR, implying greater cancer exposure among wealthier individuals. Such association is in stark contrast with the part of the current literature supporting the evidence of how social inequalities may increase exposure to poor air quality and vulnerability to respiratory diseases (34, 44, 45). Richardson et al. (44) claimed evidence of income-related inequalities in exposure to pollutants. Studies conducted to individual-level demographic data that accounts social exclusion and ethnicity (34, 45), evidenced that individual belonging to ethnic minorities are disproportionately exposed to poor air quality. However, this pattern has not been observed in Italy. Germani et al. (46), provided an empirical analysis conducted on the Italian provinces (NUTS3) claiming that air pollution increases with the average income per administrative unit. Since our study relies on provinces, coarse air quality products and no individual-level data is included to account for social and gender, we believe that the average income per province may act as a proxy for both human activities and may not point to social exclusion due to the coarse resolution of the proposed study.

Lifetime duration shows a consistent trend with expected associations: shorter lifespans correspond to higher mortality rates from respiratory diseases. We might speculate that respiratory diseases significantly impact life expectancy in Italy, particularly in certain provinces, potentially influencing the overall health of the population.

The presented study has limitations that we aim to overcome in future research. In particular, the database considered here examined the remote sensing observations of 2019 of exposure to air pollutants. An extension of the time range of our study would guarantee greater robustness to the analyses. Moreover, another limitation of this study is the failure to account for the population density differences within the provinces, as it varies between small urban centers and large cities, implying uniform exposure to air pollutants for the entire population of a province.

6 Conclusion

The presented analysis reveals a correlation between specific pollutants and socio-economic indices and the occurrence of respiratory system cancer. These finding could still offer valuable insights for further epidemiological studies as our results may suggest which variables to gather to perform analyses on individual level dataset that could lead to stronger and more conclusive results. Moreover, this study opens up future prospects for similar research on other types of cancer related to environmental pollution, as well as other types of diseases such as neurodegenerative ones.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

DR: Writing—review & editing, Writing—original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. PN: Writing—review & editing, Validation, Writing—original draft, Methodology, Investigation, Conceptualization. RC: Writing—review & editing, Writing—original draft, Validation, Methodology. NA: Writing—review & editing. AM: Writing—review & editing, Validation, Methodology. RB: Writing—review & editing, Supervision, Methodology, Funding acquisition. ST: Writing—review & editing, Writing—original draft, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This paper has been supported by the TEBAKA (TErritorial BAsic Knowledge Acquisition project Avviso MIUR n.1735 del 13/07/2017, the National Institute for Nuclear Physics (INFN), next AIM (Artificial Intelligence in Medicine: next steps) research project (INFN-CSN5), https://www.pi.infn.it/aim (accessed on 30 October 2023); The National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4-Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union-NextGenerationEU, award number: Project code: CN00000013, Concession Decree no. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: National Centre for HPC, Big Data and Quantum Computing.

Acknowledgments

Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the PON Ricerca e Competitivita' 2007–2013-Azione I-Interventi di rafforzamento strutturale *PONa*3 00052, Avviso 254/Ric, University of Bari. The authors would like to thank Dr. Nicola Carelli of ARPA-Puglia for discussion of presented analysis results.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Fajersztajn L, Veras M, Barrozo LV, Saldiva P. Air pollution: a potentially modifiable risk factor for lung cancer. *Nat Rev Cancer.* (2013) 13:674–8. doi: 10.1038/nrc3572

2. Huang Y, Zhu M, Ji M, Fan J, Xie J, Wei X, et al. Air pollution, genetic factors, and the risk of lung cancer: a prospective study in the UK Biobank. *Am J Respir Crit Care Med.* (2021) 204:817–25. doi: 10.1164/rccm.202011-4063OC

3. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and health impacts of air pollution: a review. *Front Pub Health.* (2020) 8:505570. doi: 10.3389/fpubh.2020.00014

4. GBD 2016 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet.* (2017) 390:1345. doi: 10.1016/S0140-6736(17)32366-8

5. Pershagen G. Air pollution and cancer. *IARC Sci Pub.* (1990) 104:240–51.

6. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y et al. *Extended Follow-up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality.* Boston, MA: Health Effects Institute (2009).

7. Cohen AJ, Pope CA. Lung cancer and air pollution. *Environ Health Perspect.* (1995) 103:219–24. doi: 10.1289/ehp.95103s8219

8. Santos UDP, Arbex MA, Braga ALF, Mizutani RF, Cançado JED, Terra-Filho M, et al. Environmental air pollution: respiratory effects. *J Brasileiro de Pneumologia.* (2021) 47:e20200267. doi: 10.36416/1806-3756/e20200267

9. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer (2009).

10. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMR, Bellotti R, et al. Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci.* (2021) 15:674055. doi: 10.3389/fnins.2021.674055

11. Lombardi A, Diacono D, Amoroso N, Biecek P, Monaco A, Bellantuono L, et al. A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease. *Brain Inf.* (2022) 9:17. doi: 10.1186/s40708-022-00165-5

12. Amoroso N, Quarto S, La Rocca M, Tangaro S, Monaco A, Bellotti R. An explainability artificial intelligence approach to brain connectivity in Alzheimer's disease. *Front Aging Neurosci.* (2023) 15. doi: 10.3389/fnagi.2023.1238065

13. Thunis P, Degraeuwe B, Pisoni E, Meleux F, Clappier A. Analyzing the efficiency of short-term air quality plans in European cities, using the CHIMERE air quality model. *Air Q Atmosp Health.* (2017) 10:235–48. doi: 10.1007/s11869-016-0427-y

14. Hass H, Ebel A, Feldmann H, Jakobs HJ, Memmesheimer M. Evaluation studies with a regional chemical transport model (EURAD) using air quality data from the EMEP monitoring network. *Atmosp Environ Part A Gen Topics.* (1993) 27:867–87. doi: 10.1016/0960-1686(93)90007-L

15. Duarte EDSF, Franke P, Lange AC, Friese E, da Silva Lopes FJ, da Silva JJ, et al. Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution. *Atmos Pollut Res.* (2021) 12:451–69. doi: 10.1016/j.apr.2020.12.006

16. Hinestroza-Ramirez JE, Lopez-Restrepo S, Yarce Botero A, Segers A, Rendon-Perez AM, Isaza-Cadavid S, et al. Improving air pollution modelling in complex terrain with a coupled WRF–LOTOS–EUROS approach: a case study in Aburrá Valley, Colombia. *Atmosphere.* (2023) 14:738. doi: 10.3390/atmos14040738

17. Persson C, Langner J, Robertson L. Air pollution assessment studies for Sweden based on the MATCH model and air pollution measurements. *Air Pollut Modeling Appl.* (1996) 9:127–34. doi: 10.1007/978-1-4615-5841-5_15

18. Joly M, Josse B, Plu M, Arteta J, Guth J, Meleux F. *High-Resolution air Quality Forecasts With MOCAGE Chemistry Transport Model. Air Pollution Modeling and its Application XXIV 2016.* Cham: Springer International Publishing (2016). p. 563–5.

19. Ots R, Loot A, Kaasik M. Scale-dependent and seasonal performance of SILAM model in Estonia. In: *Air Pollution Modeling and its Application XXII 2014.* Geneva: Springer Netherlands (2014). p. 593–7.

20. Van Loon M, Vautard R, Schaap M, Bergström R, Bessagnet B, Brandt J, et al. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmospheric Environ.* (2007) 41:2083–97. doi: 10.1016/j.atmosenv.2006.10.073

21. Neary L, Kaminski JW, Lupu A, McConnell JC. Developments and Results From a global multiscale air quality model (GEM-AQ). In: *Air Pollution Modeling and Its Application XVII.* Cham: Springer (2007). p. 403–10.

22. Klose M, Jorba O, Gonçalves Ageitos M, Escribano J, Dawson ML, Obiso V, Di Tomaso E. Mineral dust cycle in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (MONARCH) version 2.0. *Geosci Model Dev Discussions.* (2021) 2021:1–59. doi: 10.5194/gmd-14-6403-2021

23. Mircea M, Zanini G, Briganti G, Cappelletti A, Pederzoli A, Vitali L, et al. Modeling Air Quality Over Italy With MINNI Atmospheric Modeling System: From Regional to Local Scale. In: *Air Pollution Modeling and its Application XXI 2012.* Cham: Springer (2012). p. 491–8.

24. Gatti RC, Di Paola A, Monaco A, Velichevskaya A, Amoroso N, Bellotti R. The spatial association between environmental pollution and long-term cancer mortality in Italy. *Sci Total Environ.* (2023) 855:158439. doi: 10.1016/j.scitotenv.2022.158439

25. Wilcox AJ, Russell IT. Birthweight and perinatal mortality: III. Towards a new method of analysis. *Int J Epidemiol.* (1986) 15:188–96. doi: 10.1093/ije/15.2.188

26. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

27. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* (2020) 58:82–115. doi: 10.1016/j.inffus.2019.12.012

28. Bisong E, Bisong E. *Introduction to Scikit-learn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 215–229. doi: 10.1007/978-1-4842-4470-8_18

29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. (2017) p. 30.

30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Int.* (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

31. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *Eur Resp J.* (2016) 48:889–902. doi: 10.1183/13993003.00359-2016

32. Parikh AR. Lung Cancer Genomics. *Acta Medica Academica.* (2019) 48:244. doi: 10.5644/ama2006-124.244

33. Nyberg F, Gustavsson P, Järup L, Bellander T, Berglind N, Jakobsson R, et al. Urban air pollution and lung cancer in Stockholm. *Epidemiology.* (2000) 11:487–95. doi: 10.1097/00001648-200009000-00002

34. Dutton A. *Coronavirus (COVID-19) Related Mortality Rates and the Effects of Air Pollution in England.* London: Office of National Statistics (2020).

35. Travaglio M, Yu Y, Popovic R, Selley L, Leal NS, Martins LM. Links between air pollution and COVID-19 in England. *Environ Pollut.* (2021) 268:115859. doi: 10.1016/j.envpol.2020.115859

36. Melkonyan A, Kuttler W. Long-term analysis of NO, NO_2 and O_3 concentrations in North Rhine-Westphalia, Germany. *Atmos Environ.* (2012) 60:316–26. doi: 10.1016/j.atmosenv.2012.06.048

37. Hagenbjörk A, Malmqvist E, Mattisson K, Sommar NJ, Modig L. The spatial variation of O_3, NO, NO_2 and NO x and the relation between them in two Swedish cities. *Environ Monit Assess.* (2017) 189:1–12. doi: 10.1007/s10661-017-5872-z

38. Domínguez-López D, Adame JA, Hernández-Ceballos MA, Vaca F, De la Morena BA, Bolívar JP. Spatial and temporal variation of surface ozone, NO and NO 2 at urban, suburban, rural and industrial sites in the southwest of the Iberian Peninsula. *Environ Monitor Assessm.* (2014) 186:5337–51. doi: 10.1007/s10661-014-3783-9

39. Clapp LJ, Jenkin ME. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NOx in the UK. *Atmos Environ.* (2001) 35:6391–405. doi: 10.1016/S1352-2310(01)00378-8

40. Gorrochategui E, Hernandez I, Tauler R. A model for simultaneous evaluation of NO_2, O_3, and PM 10 pollution in urban and rural areas: handling incomplete

data sets with multivariate curve resolution analysis. *Atmospheric Chem Phys.* (2022) 22:9111–27. doi: 10.5194/acp-22-9111-2022

41. Fernández-Guisuraga JM, Castro A, Alves C, Calvo A, Alonso-Blanco E, Blanco-Alegre C, et al. Nitrogen oxides and ozone in Portugal: trends and ozone estimation in an urban and a rural site. *Environ Sci Pollut Res.* (2016) 23:17171–82. doi: 10.1007/s11356-016-6888-6

42. Safieddine S, Clerbaux C, George M, Hadji-Lazaro J, Hurtmans D, Coheur PF, et al. Tropospheric ozone and nitrogen dioxide measurements in urban and rural regions as seen by IASI and GOME-2. *J Geophys Res Atmos.* (2013) 118:10–555. doi: 10.1002/jgrd.50669

43. Reeves CE, Penkett SA, Bauguitte S, Law KS, Evans MJ, Bandy BJ, Kley D. Potential for photochemical ozone formation in the troposphere over the North Atlantic as derived from aircraft observations during ACSOE. *J Geophys Res Atmos.* (2002) 107:ACH-14. doi: 10.1029/2002JD002415

44. Richardson EA, Pearce J, Tunstall H, Mitchell R, Shortt NK. Particulate air pollution and health inequalities: a Europe-wide ecological analysis. *Int J Health Geogr.* (2013) 12:1–10. doi: 10.1186/1476-072X-12-34

45. van den Brekel L, Lenters V, Mackenbach JD, Hoek G, Wagtendonk A, Lakerveld J, et al. Ethnic and socioeconomic inequalities in air pollution exposure: a cross-sectional analysis of nationwide individual-level data from the Netherlands. *The Lancet Planetary Health.* (2024) 8:e18–29. doi: 10.1016/S2542-5196(23)00258-9

46. Germani AR, Morone P, Testa G. Environmental justice and air pollution: a case study on Italian provinces. *Ecol Econ.* (2014) 106:69–82. doi: 10.1016/j.ecolecon.2014.07.010

Chapter 3

Machine learning models for higher and lower mortality

This chapter compares thirteen classification models to identify provinces with higher or lower mortality from respiratory system cancer. The study explores how features contribute to the prediction of the best model with shapley values.

Journal of *Personalized Medicine*

MDPI

Article

Insights from Explainable Artificial Intelligence of Pollution and Socioeconomic Influences for Respiratory Cancer Mortality in Italy

Donato Romano [1,2], Pierfrancesco Novielli [1,2], Domenico Diacono [2], Roberto Cilli [2,3], Ester Pantaleo [2,3], Nicola Amoroso [2,4], Loredana Bellantuono [2,5], Alfonso Monaco [2,3], Roberto Bellotti [2,3] and Sabina Tangaro [1,2,*]

1 Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, 70126 Bari, Italy; donato.romano@uniba.it (D.R.); pierfrancesco.novielli@uniba.it (P.N.)
2 Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70126 Bari, Italy; domenico.diacono@ba.infn.it (D.D.); roberto.cilli@uniba.it (R.C.); ester.pantaleo@uniba.it (E.P.); nicola.amoroso@uniba.it (N.A.); loredana.bellantuono@uniba.it (L.B.); alfonso.monaco@uniba.it (A.M.); roberto.bellotti@uniba.it (R.B.)
3 Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari Aldo Moro, 70126 Bari, Italy
4 Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, 70126 Bari, Italy
5 Dipartimento di Biomedicina Traslazionale e Neuroscienze, Università degli Studi di Bari Aldo Moro, 70126 Bari, Italy
* Correspondence: sabina.tangaro@uniba.it

Abstract: Respiratory malignancies, encompassing cancers affecting the lungs, the trachea, and the bronchi, pose a significant and dynamic public health challenge. Given that air pollution stands as a significant contributor to the onset of these ailments, discerning the most detrimental agents becomes imperative for crafting policies aimed at mitigating exposure. This study advocates for the utilization of explainable artificial intelligence (XAI) methodologies, leveraging remote sensing data, to ascertain the primary influencers on the prediction of standard mortality rates (SMRs) attributable to respiratory cancer across Italian provinces, utilizing both environmental and socioeconomic data. By scrutinizing thirteen distinct machine learning algorithms, we endeavor to pinpoint the most accurate model for categorizing Italian provinces as either above or below the national average SMR value for respiratory cancer. Furthermore, employing XAI techniques, we delineate the salient factors crucial in predicting the two classes of SMR. Through our machine learning scrutiny, we illuminate the environmental and socioeconomic factors pertinent to mortality in this disease category, thereby offering a roadmap for prioritizing interventions aimed at mitigating risk factors.

Keywords: explainable artificial intelligence; machine learning; remote sensing; air pollution; exposome; respiratory disease; socioeconomic indices; public health

check for updates

Citation: Romano, D.; Novielli, P.; Diacono, D.; Cilli, R.; Pantaleo, E.; Amoroso, N.; Bellantuono, L.; Monaco, A.; Bellotti, R.; Tangaro, S. Insights from Explainable Artificial Intelligence of Pollution and Socioeconomic Influences for Respiratory Cancer Mortality in Italy. *J. Pers. Med.* **2024**, *14*, 430. https://doi.org/10.3390/jpm14040430

Academic Editors: Yi Guo, Zhe He and Xing He

Received: 4 March 2024
Revised: 10 April 2024
Accepted: 11 April 2024
Published: 18 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Success in mapping the human genome has stimulated the complementary concept of the *exposome*, i.e., the measure of the complete environmental exposure of an individual or a population to their surrounding environment and the study of how those exposures relate to health [1]. Research on the exposome in the context of noncommunicable diseases (NCDs), or diseases that cannot be transferred from one person to the other, is relatively novel and advancing impressively [2]. Although NCDs are associated with a genetic predisposition (the genome), exposure to health-affecting environmental parameters has a strong impact on their risk. With the use of Earth observation data, recent research studies in this field are exploring possible associations of NCDs with environmental parameters to identify relevant factors and prioritize intervention strategies to mitigate their effects, in line with one of the targets of the 2030 United Nations Agenda for Sustainable Development [3], i.e., reducing premature mortality from NCDs.

Among NCDs, respiratory tract cancers, which include tracheal, bronchus, and lung cancer, are the leading causes of cancer death [4], registering the highest age-standardized mortality rate among all cancers [5]. Studies focusing on respiratory diseases and/or cancer using Earth observation data include factors such as air pollution, atmospheric factors, and land surface data in their analysis. They recognize air pollution parameters as important factors, such as NO_2 and O_3, and identify prominent associations with parameters such as PM 2.5 exposure, asbestos exposure, UV irradiation, and light pollution [6–12].

In our analysis, we use air quality estimates by the CAMS reanalysis dataset of atmospheric composition [13] and spatial census data at provincial resolution (environmental and socioeconomic variables) to predict standard cancer mortality rates (SMRs) attributable to respiratory tract cancers as provided by the Italian National Institute of Statistics (ISTAT). More specifically, using data from 2019, we estimate the national average SMR and define a binary variable with a value of 1 for provinces with an SMR higher than the average and 0 otherwise. We then use Earth observation and census data to predict this binary variable, comparing the performance of thirteen different models. Finally, we apply an explainable artificial intelligence strategy, thus providing model-agnostic insights into how the best model calculates predictions with the aim of facilitating the determination of feature importance and offering a roadmap for prioritizing interventions [14,15]. Our approach is therefore aimed at contributing to a comprehensive understanding of the relationships between the environment and human health for the specific disease under consideration.

This paper is structured as follows. In Section 2, we describe the input data and the adopted methodologies. More specifically, in Section 2.1, we define our input factors; in Section 2.2, we define our output binary variable; in Section 2.4, we reduce the features of the dataset through a multicollinearity analysis; in Section 2.5, we introduce the 13 machine algorithms and the evaluation metrics used; in Section 2.6, we describe the explainable artificial intelligence approach adopted. In Section 3, we present the results of our analysis that we discuss in Section 4. Finally, in Section 6, we draw our conclusions.

2. Materials and Methods

2.1. Pollutants Data and Socioeconomic Indices

The Copernicus Atmosphere Monitoring Service (https://ads.atmosphere.copernicus.eu/#!/search?text=&type=dataset&keywords=((%20%22Product%20type:%20Reanalysis%22%20)%20AND%20(%20%22Variable%20domain:%20Atmosphere%20(composition)%22%20)%20AND%20(%20%22Spatial%20coverage:%20Europe%22%20)%20AND%20(%20%22Temporal%20coverage:%20Past%22%20) (accessed on 10 April 2024) (CAMS) platform managed by the European Union's Copernicus Program provides information on air pollution and air composition globally. It utilizes data from various sources such as satellites, aircraft, ground stations, and numerical models to monitor and analyze real-time air quality and provide forecasts. The pollutants data provided by the CAMS platform are the result of an ensemble median over 9 numerical air quality models [13,16]: CHIMERE (CH) from INERIS (France) [17], EMEP (EM) from MET Norway (Norway) [18], EURAD-IM (EU) from Jülich IEK (Germany) [19], LOTOS-EUROS (LO) from KNMI and TNO (the Netherlands) [20], MATCH (MA) from SMHI (Sweden) [21], MOCAGE (MO) from Meteo-France (France) [22], SILAM (S) from FMI (Finland) [23], DEHM (DE) from Aarhus University (Denmark) [24], and GEM-AQ (GE) from IEP-NRI (Poland) [25]. For the purposes of the present study, which aims at investigating the association between pollution and mortality for cancer with the respiratory system, only the concentrations of CO, NO, NO_2, O_3, PM 10, PM 2.5, and SO_2 were considered. Since the data downloaded from the CAMS platform covered an area larger than that of our interest, the Python library for analyzing geolocalized data https://geopandas.org/en/stable/ (accessed on 10 April 2024) Geopandas (version 0.14.1) was used to extract the pollution data of the Italian peninsula. From the daily coverage over all of 2019, we retrieved the mean values for the concentrations of the selected pollutants.

In order to maximize data consistency, other pollution-related and socioeconomic variables were considered: cultivated areas, urban areas, benzene, temperature, N fertilizer,

J. Pers. Med. **2024**, 14, 430

P4010 fertilizer, microelement fertilizer, organic fertilizer, the number of hospital beds available (bed number), lifetime, income, life quality, instruction, vehicles per km^2 of land area (vehicles total), vehicles per km^2 of urbanized surface (urban traffic), photovoltaic panel, percentage of urban greenery density on the surface (green urban), electric consumption, exceeding the limits detected following noise checks (noise), and municipal waste collection in tons (wastes) [26].

2.2. Standardized Mortality Ratio

The Italian National Institute of Statistics (ISTAT) is the main official statistical institute in Italy, responsible for collecting, analyzing, and disseminating statistical information about the country.

The standardized mortality ratio (SMR) is a measure of the mortality rate of a specific population compared with a standard or reference population (e.g., the national population). It is calculated by dividing the number of observed deaths in the study population by the number of expected deaths based on the mortality rates of the standard population. In our study, the values of the SMR of the Italian provinces were calculated starting from the number of observed deaths for respiratory system cancer in 2019 by age class provided by ISTAT. The mean value of the SMR was used to divide the Italian provinces into two classes: all the provinces with an SMR lower than the mean value were assigned to class 0, while all the others were assigned to class 1. An outline of the data preparation workflow is presented in Figure 1. A graphical representation of Italian provinces with a higher or lower SMR than the mean value for cancer to the respiratory system is shown in Figure 2.

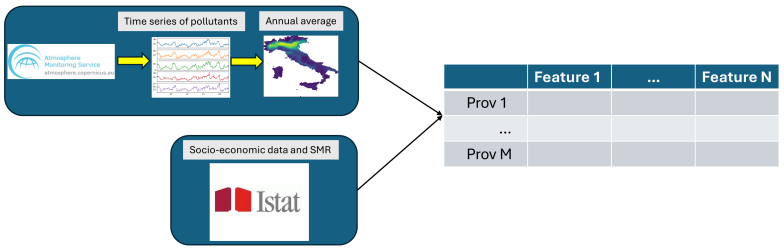

Figure 1. Dataset construction workflow.

2.3. Analysis Flowchart

The diagram presented in Figure 3 serves as a visual representation encapsulating the comprehensive analytical journey delineated within this paper. At the heart of this endeavor lies the curation of a dataset, meticulously constructed to encompass a rich tapestry of data streams. The foundational pillars of this dataset were erected through a data acquisition process, commencing with the extraction of air pollution metrics from diverse provinces across the Italian peninsula. These statistics were sourced from the Copernicus Atmosphere Monitoring Service (CAMS) platform, renowned for its robust and comprehensive environmental monitoring capabilities. In tandem with air quality metrics, a plethora of socioeconomic indicators were judiciously incorporated into the dataset. These encompassed a spectrum of socioeconomic variables, ranging from income distribution patterns to urbanization indices, providing a holistic perspective on the societal fabric under scrutiny. Furthermore, the dataset was augmented with mortality data attributable to respiratory system cancers, gleaned from the extensive repository maintained by the Italian National Institute of Statistics (ISTAT). The analysis commenced with the exploration of multicollinearity dynamics inherent within the dataset. Leveraging the Variation Inflation Factor (VIF) as a diagnostic tool, we dissected intervariable relationships to discern and mitigate the pernicious effects of collinearity. Through a systematic iterative process, variables exhibiting VIF indices surpassing the critical threshold of 10 were pruned from the dataset, thereby ensuring the integrity and robustness of subsequent analyses. Subsequently, the refined dataset served as fertile ground for the cultivation of predictive

J. Pers. Med. **2024**, *14*, 430

models leveraging state-of-the-art machine learning techniques. Employing the Python (3.10.12 version) library https://pycaret.org (accessed on 10 April 2024) PyCaret (version 3.3.0) as a conduit, a diverse ensemble of 13 distinct machine learning models was meticulously trained and fine-tuned and tasked with the classification of provinces exhibiting mortality rates from respiratory system cancers either surpassing or falling below the national average. The efficacy of these predictive models was rigorously scrutinized through a comprehensive validation process, employing a 10-fold cross-validation methodology by evaluating the performance obtained from the validation folds. Once the model with the largest ROC AUC was identified, the most important features for the model to classify provinces with higher or lower mortality due to respiratory system cancer were calculated using the explainable artificial intelligence algorithm SHAP.

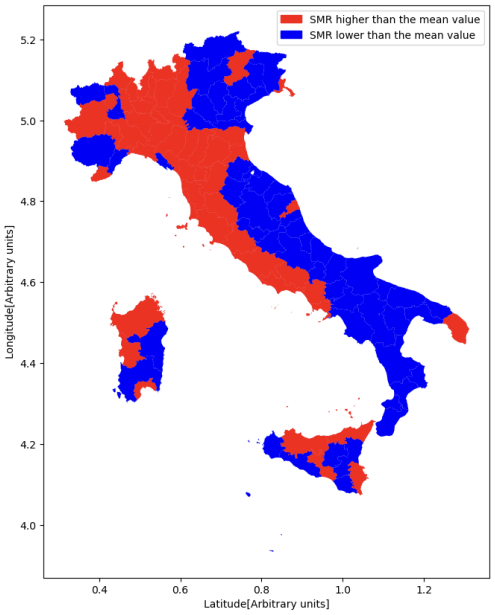

Figure 2. Colormap of the Italian provinces with higher or lower SMR than the mean value for cancer to the respiratory system.

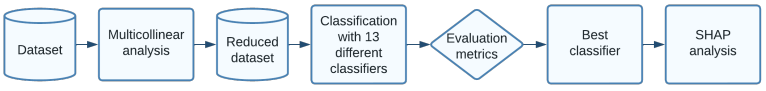

Figure 3. Dataset analysis workflow.

2.4. Feature Correlation Analysis

The Variation Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity among independent variables. The VIF provides an indication of the strength of correlation among independent variables, detecting the effect of inflation on the variance in estimated coefficients in the regression model.

In this study, the VIF was calculated to identify and address the issue of multicollinearity within our dataset. An iterative approach was adopted, where the VIF was computed repeatedly, eliminating the variable with the highest VIF at each iteration, provided its value exceeded the threshold of 10. This process effectively reduced multicollinearity in the model, thereby improving the stability of coefficient estimates.

The surviving features are mean NO, cultivated areas, benzene, P4O10 fertilizer, microelement fertilizer, organic fertilizer, income, instruction, vehicles total, urban traffic, green urban, and noise.

2.5. Comparison of Classification Models

Thirteen different classification models, Gradient Boosting Classifier (gbc) [27], Light Gradient Boosting Machine (lgb) [28], Random Forest Classifier (rf) [29], Extra Trees Classifier (et) [30], K Neighbors Classifier (knn) [31], Extreme Gradient Boosting (xgb) [32], Linear Discriminant Analysis (lda) [33], Ada Boost Classifier (ada) [34], Decision Tree Classifier (dt) [35], Naive Bayes (nb) [36], Quadratic Discriminant Analysis (qda) [37], Logistic Regression (lr) [38], and Dummy Classifier (dum) [39], a classification model that does not learn anything from the training data but is particularly useful for assessing the performance of more complex models and understanding the difficulty of the classification task, were trained on the whole dataset in a 10-stratified cross-validation.

The performance of our machine learning models was evaluated on each validation set through the following metrics:

- Accuracy:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

- AUC ROC: The area under the Receiver Operating Characteristic (ROC) curve;
- Recall:

$$REC = \frac{TP}{TP + FN} \tag{2}$$

- Precision:

$$PREC = \frac{TP}{TP + FP} \tag{3}$$

- F1-score:

$$F1 = 2 \cdot \frac{PREC \cdot REC}{PREC + REC} \tag{4}$$

- Kappa:

$$k = \frac{2 \cdot (TP \cdot TN - FN \cdot FP)}{(TP + FP) \cdot (TP + FN) \cdot (FN + TN)} \tag{5}$$

2.6. Explainable Algorithm

We adopted the SHAP algorithm [40,41] to explain the decisions of the Gradient Booster Classifier models on each test sample. It provides insight into the contribution of each feature to the model's prediction for a specific instance. SHAP values are based on game theory concepts and specifically on Shapley values, which originated in cooperative game theory. One of the key advantages of SHAP is its model-agnostic nature. It can be applied to any machine learning model, whether it is a black-box or white-box model, making it versatile and widely applicable. The measurement of how a feature affects the performance of the classification model on the validation set is computed by including and removing it from the model:

$$\Phi_j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} [f_x(F \cup j) - f_x(F)] \tag{6}$$

where $\Phi_j(x)$ represents the SHAP value of feature j for the prediction of the model f given the input x, S is the set of all features, $F \subseteq S - \{j\}$ represents all possible subsets of features excluding feature j, and $\frac{|F|!(|S|-|F|-1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that does not belong to S, while $f_x(F \cup j)$ and $f_x(F)$ denote, respectively, the model's prediction when feature j is added to the subset F and when it is absent. We also averaged the ten

J. Pers. Med. **2024**, 14, 430

realizations of SHAP values in order to obtain a single representative SHAP vector. The whole analysis workflow is shown in Figure 3.

3. Results

The aim of this study was to assess, using explainable machine learning models, the extent to which air pollutants and socioeconomic descriptors are associated with higher or lower mortality from respiratory system cancer.

3.1. Performance of Classification Models

Table 1 shows the average performance of the 13 machine learning classifiers calculated as the mean performance obtained during the 10-fold cross-validation. The classifiers are sorted in descending order according to the accuracy values.

Table 1. Performance of different classification models.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
et	0.78 ± 0.19	0.75 ± 0.17	0.73 ± 0.26	0.82 ± 0.18	0.75 ± 0.23	0.57 ± 0.37
lgb	0.74 ± 0.10	0.86 ± 0.13	0.71 ± 0.16	0.80 ± 0.18	0.73 ± 0.10	0.47 ± 0.21
rf	0.72 ± 0.21	0.80 ± 0.16	0.70 ± 0.25	0.78 ± 0.23	0.70 ± 0.22	0.44 ± 0.42
xgb	0.70 ± 0.16	0.78 ± 0.18	0.68 ± 0.24	0.75 ± 0.18	0.67 ± 0.19	0.39 ± 0.32
gbc	0.68 ± 0.18	0.77 ± 0.19	0.68 ± 0.18	0.72 ± 0.21	0.68 ± 0.16	0.36 ± 0.36
ada	0.67 ± 0.28	0.72 ± 0.17	0.68 ± 0.21	0.67 ± 0.17	0.67 ± 0.17	0.34 ± 0.36
qda	0.65 ± 0.12	0.77 ± 0.14	0.50 ± 0.17	0.73 ± 0.23	0.58 ± 0.16	0.30 ± 0.26
lr	0.64 ± 0.13	0.68 ± 0.17	0.56 ± 0.22	0.71 ± 0.20	0.59 ± 0.17	0.28 ± 0.26
knn	0.64 ± 0.16	0.69 ± 0.15	0.58 ± 0.14	0.69 ± 0.21	0.62 ± 0.15	0.27 ± 0.32
lda	0.62 ± 0.13	0.66 ± 0.21	0.50 ± 0.18	0.70 ± 0.22	0.55 ± 0.16	0.23 ± 0.26
dt	0.61 ± 0.21	0.61 ± 0.21	0.61 ± 0.21	0.64 ± 0.22	0.61 ± 0.19	0.21 ± 0.42
nb	0.55 ± 0.08	0.47 ± 0.17	0.17 ± 0.13	0.60 ± 0.44	0.25 ± 0.18	0.09 ± 0.14
dummy	0.50 ± 0.04	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

The Extra Trees classifier emerged as the best in predicting an SMR higher or lower than the mean value of mortality from respiratory system cancer, with a mean accuracy of 0.74 ± 0.13. The mean ROC AUC of the best model is shown in Figure 4. Each fold of the cross-validation process yielded a distinct AUC value, allowing us to compute the mean AUC across all folds and assess the model's overall performance.

3.2. Interpreting Model Predictions: Insights from SHAP Analysis

In this section, we present a summary plot of the SHAP values obtained from a binary classification model, shedding light on the influential features driving the model's predictions. Within the 10-fold cross-validation, SHAP values of the features of the provinces present in the validation fold were calculated after training the classification model with the nine folds of training data. This procedure was repeated for each of the validation folds. The summary plot provides a comprehensive overview of the impact of each feature on model predictions, revealing both the direction and magnitude of their influence. Figure 5 illustrates the summary plot generated from the SHAP values computed for the binary classification model. Each point in the summary plot represents a feature, with its position on the y-axis indicating the feature's importance in terms of absolute SHAP value. The color of each point represents the feature's value, with red indicating high values and blue indicating low values. Additionally, the horizontal bars represent the impact of each feature on model predictions, with longer bars signifying a stronger influence.

J. Pers. Med. **2024**, *14*, 430

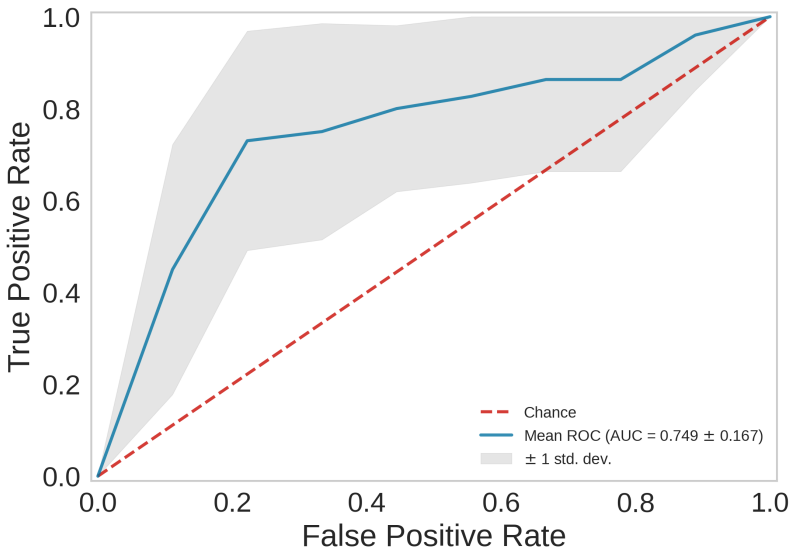

Figure 4. Average ROC curve with standard deviation of the Extra Trees classifier over 10 validation sets.

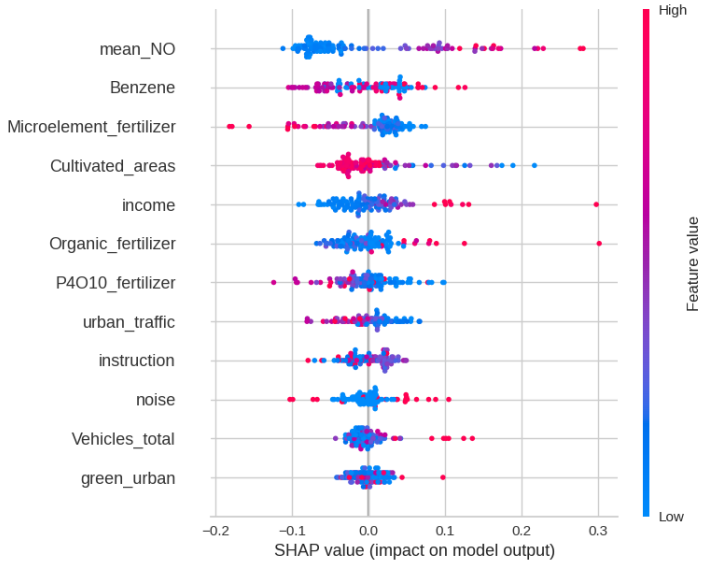

Figure 5. Shap summary plot. It reports the SHAP values, represented along the x-axis, of the dataset features, reported on the y-axis. Each point on the graph represents a province.

4. Discussion

Our findings show that most of the input variables included in this study are weakly correlated to the target variable, the standard mortality ratio, while our Extra Trees model accurately predicts provinces withan SMR higher than the mean value. Therefore, it is likely that a nonlinear relationship between air quality and SMR exists, thus suggesting the beneficial role of including machine learning tools in this analysis. This is also consistent when comparing the classification performance of the tree-based algorithms with other shallow classification algorithms, including the LDA (and others). In ecological studies,

J. Pers. Med. 2024, 14, 430

linear models should be preferred since they are easy to interpret; however, Extra Trees combined with the SHAP is a robust choice that does not require any prior assumptions, while providing local feature importance values that are completely intelligible.

We also used the distribution of the SHAP values as a tool to estimate the global feature importance of the variables included in our analysis. Accordingly, this analysis confirms that average exposure to NOX is the most important feature in predicting the SMR of respiratory system cancer that is higher than the mean value. This result is in agreement with previous analyses confirming the positive association between respiratory diseases and exposure to NO when conducting studies on populations [42,43].

According to this study, the second most important feature was the average benzene concentration per province. Chronic exposure to benzene has been linked to leukemia [44,45]; however, there is limited evidence of a link with respiratory cancer in the scientific literature [45–47]. Nevertheless, benzene pollution is typically generated because of biomass combustion due to wildfires [48] and vehicle exhaust [49]. Hence, we believe this variable is a proxy for population exposure to poor air quality due to smoke from combustion from human and natural activities.

Minor effects can also be imputed to microelement fertilizer consumption and extension of the cultivated areas. As can be seen in Figure 5, the distribution of the related SHAP values suggests that these two variables act as a proxy for living in a rural environment.

Another important feature was the mean income per province. Although it is known that socioeconomic inequalities have a role in respiratory cancer outcomes [50], this effect cannot be assessed through a population study. Besides, it is likely that the average income per province is acting as a proxy for residing in high-populated areas, where most of the air pollution is emitted [42], while not accounting for social disparities.

5. Limitations

Key limitations in the proposed workflow stem from the dataset's limited size and the utilization of standard mortality ratios aggregated at provincial levels to explore associations between respiratory cancer mortality and air pollution. In fact, the dataset's limited size hampers the ability to discern reliable nonlinear relationships due to the significant influence of outliers and noise. Consequently, we employed a k-fold cross-validation framework and then computed confidence intervals for both predictions and local feature importance. On the other hand, the reliance on aggregated mortality ratios for individual provinces exposes the study to the ecological fallacy, as well as to potential omitted variable bias [51,52], meaning that conclusions drawn at the group level may not accurately represent individual-level associations or causality. Finally, it should be stressed that most of the important variables according to the SHAP algorithm may act as a proxy for the high density of urban activities. If a causality exists, this might be related to the role of NOx in contributing to the development of asthma and respiratory infections, causing a range of harmful effects on lungs [53–55]. On the other hand, NO is also related to proximity to residential areas and human activities, so the measured positive association could be partly attributable to other air pollutants, including particulate matter [42]. This work presents a first attempt to extensively evaluate the statistical association between air quality and respiratory cancer mortality over the Italian provinces while accounting for effects from confounding socioeconomic variables. Our conclusions are in agreement with previous cohort studies and meta-analyses [42,56]. The database considered here examined a single year of observations; thus, extending the temporal range of our study could grant increased robustness to the analyses and design a regression approach to model the SMR due to respiratory cancer.

6. Conclusions

This paper explores the intricate relationship between air pollution and respiratory cancer, with a focus on cancers of the lung, trachea, and bronchi. Recognizing the urgent public health challenge posed by respiratory cancers, we address the need to identify the

most harmful pollutants to inform targeted policy interventions. Our study introduces a novel approach using explainable artificial intelligence based on remote sensing data and socioeconomic data to predict Italian provinces with respiratory system cancer mortality rates higher or lower than the average, which offers several implications for formulating regulations and allocating resources. The results of the research can provide an empirical basis for adopting or strengthening regulations regarding air quality. Identifying provinces with higher mortality rates for respiratory system cancers can highlight areas with higher levels of air pollution and provide an incentive for implementing policies aimed at reducing atmospheric pollution. The ability to predict provinces with respiratory system cancer mortality rates higher or lower than the average can guide more efficient allocation of resources. Areas identified with higher risk may require additional investments in prevention programs, early diagnosis, and treatment of respiratory system cancer, as well as interventions to improve air quality. The adoption of stricter regulations on air pollution and the targeted allocation of healthcare resources can improve the respiratory health of a population and reduce the incidence and mortality of respiratory system cancers.

Author Contributions: Conceptualization, D.R., R.C. and S.T.; methodology, D.R., P.N., D.D. and S.T.; software, D.R. and D.D.; validation, D.R., P.N., D.D., R.C., E.P., R.B. and S.T.; investigation, D.R. and P.N.; data curation, D.R. and R.C.; writing—original draft preparation, D.R., P.N., R.C., E.P. and S.T.; writing—review and editing, D.R., P.N., D.D., R.C., E.P., N.A., L.B., A.M., R.B. and S.T.; visualization, D.R., D.D., R.C. and S.T.; supervision, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This paper has been supported by the TEBAKA (TErritorial BAsic Knowledge Acquisition project "Avviso MIUR n.1735 del 13/07/2017", the National Institute for Nuclear Physics (INFN), next_AIM (Artificial Intelligence in Medicine: next steps) research project (INFN-CSN5), https://www.pi.infn.it/aim (accessed on 30 October 2023); the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number: Project code: CN00000013, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: "National Centre for HPC, Big Data and Quantum Computing".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are either publicly available on databases cited in the bibliography or available from the corresponding author on request.

Acknowledgments: The authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the "PON Ricerca e Competitività 2007–2013-Azione I-Interventi di rafforzamento strutturale" PONa3 _00052, Avviso 254/Ric, University of Bari.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wild, C.P. The exposome: From concept to utility. *Int. J. Epidemiol.* **2012**, *41*, 24–32. [CrossRef] [PubMed]
2. Sogno, P.; Traidl-Hoffmann, C.; Kuenzer, C. Earth observation data supporting non-communicable disease research: A review. *Remote Sens.* **2020**, *12*, 2541. [CrossRef]
3. United Nations Department of Economic and Social Affairs. *The Sustainable Development Goals Report 2023: Special Edition*; United Nations Department of Economic and Social Affairs: New York, NY, USA, 2023.
4. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
5. Mahesh, P. Implementing precision medicine in best practices of chronic airway diseases. *Indian J. Med. Res.* **2019**, *149*, 802. [CrossRef]
6. Hystad, P.; Demers, P.A.; Johnson, K.C.; Brook, J.; van Donkelaar, A.; Lamsal, L.; Martin, R.; Brauer, M. Spatiotemporal air pollution exposure assessment for a Canadian population-based lung cancer case-control study. *Environ. Health* **2012**, *11*, 22. [CrossRef] [PubMed]

7. Tomczak, A.; Miller, A.B.; Weichenthal, S.A.; To, T.; Wall, C.; van Donkelaar, A.; Martin, R.V.; Crouse, D.L.; Villeneuve, P.J. Long-term exposure to fine particulate matter air pollution and the risk of lung cancer among participants of the Canadian National Breast Screening Study. *Int. J. Cancer* **2016**, *139*, 1958–1966. [CrossRef] [PubMed]
8. Consonni, D.; Carugno, M.; De Matteis, S.; Nordio, F.; Randi, G.; Bazzano, M.; Caporaso, N.E.; Tucker, M.A.; Bertazzi, P.A.; Pesatori, A.C.; et al. Outdoor particulate matter (PM10) exposure and lung cancer risk in the EAGLE study. *PLoS ONE* **2018**, *13*, e0203539. [CrossRef] [PubMed]
9. Di Gilio, A.; Catino, A.; Lombardi, A.; Palmisani, J.; Facchini, L.; Mongelli, T.; Varesano, N.; Bellotti, R.; Galetta, D.; de Gennaro, G.; et al. Breath analysis for early detection of malignant pleural mesothelioma: Volatile organic compounds (VOCs) determination and possible biochemical pathways. *Cancers* **2020**, *12*, 1262. [CrossRef] [PubMed]
10. Kamis, A.; Cao, R.; He, Y.; Tian, Y.; Wu, C. Predicting lung cancer in the United States: A multiple model examination of public health factors. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6127. [CrossRef]
11. Ahmed, Z.U.; Sun, K.; Shelly, M.; Mu, L. Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA. *Sci. Rep.* **2021**, *11*, 24090. [CrossRef]
12. Monaco, A.; Lacalamita, A.; Amoroso, N.; D'Orta, A.; Del Buono, A.; di Tuoro, F.; Tangaro, S.; Galeandro, A.I.; Bellotti, R. Random forests highlight the combined effect of environmental heavy metals exposure and genetic damages for cardiovascular diseases. *Appl. Sci.* **2021**, *11*, 8405. [CrossRef]
13. Casciaro, G.; Cavaiola, M.; Mazzino, A. Calibrating the CAMS European multi-model air quality forecasts for regional air pollution monitoring. *Atmos. Environ.* **2022**, *287*, 119259. [CrossRef]
14. Ladbury, C.; Zarinshenas, R.; Semwal, H.; Tam, A.; Vaidehi, N.; Rodin, A.S.; Liu, A.; Glaser, S.; Salgia, R.; Amini, A. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: A narrative review. *Transl. Cancer Res.* **2022**, *11*, 3853. [CrossRef] [PubMed]
15. Roussel, C.; Böhm, K. Geospatial xai: A review. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 355. [CrossRef]
16. Marécal, V.; Peuch, V.H.; Andersson, C.; Andersson, S.; Arteta, J.; Beekmann, M.; Benedictow, A.; Bergström, R.; Bessagnet, B.; Cansado, A.; et al. A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. *Geosci. Model Dev.* **2015**, *8*, 2777–2813. [CrossRef]
17. Thunis, P.; Degraeuwe, B.; Pisoni, E.; Meleux, F.; Clappier, A. Analyzing the efficiency of short-term air quality plans in European cities, using the CHIMERE air quality model. *Air Qual. Atmos. Health* **2017**, *10*, 235–248. [CrossRef] [PubMed]
18. Hass, H.; Ebel, A.; Feldmann, H.; Jakobs, H.; Memmesheimer, M. Evaluation studies with a regional chemical transport model (EURAD) using air quality data from the EMEP monitoring network. *Atmos. Environ. Part Gen. Top.* **1993**, *27*, 867–887. [CrossRef]
19. Duarte, E.D.S.F.; Franke, P.; Lange, A.C.; Friese, E.; da Silva Lopes, F.J.; da Silva, J.J.; dos Reis, J.S.; Landulfo, E.; e Silva, C.M.S.; Elbern, H.; et al. Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution. *Atmos. Pollut. Res.* **2021**, *12*, 451–469. [CrossRef]
20. Hinestroza-Ramirez, J.E.; Lopez-Restrepo, S.; Yarce Botero, A.; Segers, A.; Rendon-Perez, A.M.; Isaza-Cadavid, S.; Heemink, A.; Quintero, O.L. Improving Air Pollution Modelling in Complex Terrain with a Coupled WRF–LOTOS–EUROS Approach: A Case Study in Aburrá Valley, Colombia. *Atmosphere* **2023**, *14*, 738. [CrossRef]
21. Persson, C.; Langner, J.; Robertson, L. Air pollution assessment studies for Sweden based on the MATCH model and air pollution measurements. In *Air Pollution Modeling and Its Application XI*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 127–134.
22. Joly, M.; Josse, B.; Plu, M.; Arteta, J.; Guth, J.; Meleux, F. High-Resolution Air Quality Forecasts with MOCAGE Chemistry Transport Model. In *Air Pollution Modeling and Its Application XXIV*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 563–565.
23. Ots, R.; Loot, A.; Kaasik, M. Scale-dependent and seasonal performance of SILAM model in Estonia. In *Air Pollution Modeling and its Application XXII*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 593–597.
24. Van Loon, M.; Vautard, R.; Schaap, M.; Bergström, R.; Bessagnet, B.; Brandt, J.; Builtjes, P.; Christensen, J.; Cuvelier, C.; Graff, A.; et al. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmos. Environ.* **2007**, *41*, 2083–2097. [CrossRef]
25. Neary, L.; Kaminski, J.W.; Lupu, A.; McConnell, J.C. Developments and results from a global multiscale air quality model (GEM-AQ). In *Air Pollution Modeling and Its Application XVII*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 403–410.
26. Cazzolla Gatti, R.; Di Paola, A.; Monaco, A.; Velichevskaya, A.; Amoroso, N.; Bellotti, R. The spatial association between environmental pollution and long-term cancer mortality in Italy. *Sci. Total Environ.* **2023**, *855*, 158439. [CrossRef]
27. Mayr, A.; Binder, H.; Gefeller, O.; Schmid, M. The evolution of boosting algorithms. *Methods Inf. Med.* **2014**, *53*, 419–427. [PubMed]
28. Abdurrahman, M.H.; Irawan, B.; Setianingsih, C. A review of light gradient boosting machine method for hate speech classification on twitter. In Proceedings of the 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Kuala Lumpur, Malaysia, 28 November 2020; pp. 1–6.
29. Parmar, A.; Katariya, R.; Patel, V. A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 758–763.
30. Baby, D.; Devaraj, S.J.; Hemanth, J. Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach. *Turk. J. Electr. Eng. Comput. Sci.* **2021**, *29*, 2742–2757. [CrossRef]
31. Kataria, A.; Singh, M. A review of data classification using k-nearest neighbour algorithm. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, *3*, 354–360.

32. Azmi, S.S.; Baliga, S. An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. *Int. Res. J. Eng. Technol.* **2020**, *7*, 6867–6870.
33. McLachlan, G.J. Discriminant analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2012**, *4*, 421–431. [CrossRef]
34. An, T.K.; Kim, M.H. A new diverse AdaBoost classifier. In Proceedings of the 2010 International conference on artificial intelligence and computational intelligence, Sanya, China, 23–24 October 2010; Volume 1, pp. 359–363.
35. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [CrossRef]
36. Saritas, M.M.; Yasar, A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *Int. J. Intell. Syst. Appl. Eng.* **2019**, *7*, 88–91. [CrossRef]
37. Tharwat, A. Linear vs. quadratic discriminant analysis classifier: A tutorial. *Int. J. Appl. Pattern Recognit.* **2016**, *3*, 145–180. [CrossRef]
38. Nick, T.G.; Campbell, K.M. Logistic regression. *Top. Biostat.* **2007**, 273–301.
39. Barletta, L.; Giusti, A.; Rottondi, C.; Tornatore, M. QoT estimation for unestablished lighpaths using machine learning. In *Optical Fiber Communication Conference*; Optica Publishing Group: Washington, DC, USA, 2017; p. Th1J-1.
40. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
41. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
42. Hamra, G.B.; Laden, F.; Cohen, A.J.; Raaschou-Nielsen, O.; Brauer, M.; Loomis, D. Lung Cancer and Exposure to Nitrogen Dioxide and Traffic: A Systematic Review and Meta-Analysis. *Environ. Health Perspect.* **2015**, *123*, 1107–1112. [CrossRef]
43. Amoroso, N.; Cilli, R.; Maggipinto, T.; Monaco, A.; Tangaro, S.; Bellotti, R. Satellite data and machine learning reveal a significant correlation between NO_2 and COVID-19 mortality. *Environ. Res.* **2022**, *204*, 111970. [CrossRef] [PubMed]
44. Snyder, R. Leukemia and benzene. *Int. J. Environ. Res. Public Health* **2012**, *9*, 2875–2893. [CrossRef] [PubMed]
45. Loomis, D.; Guyton, K.Z.; Grosse, Y.; El Ghissassi, F.; Bouvard, V.; Benbrahim-Tallaa, L.; Guha, N.; Vilahur, N.; Mattock, H.; Straif, K. Carcinogenicity of benzene. *Lancet Oncol.* **2017**, *18*, 1574–1575. [CrossRef] [PubMed]
46. Ferrero, A.; Esplugues, A.; Estarlich, M.; Llop, S.; Cases, A.; Mantilla, E.; Ballester, F.; Iñiguez, C. Infants' indoor and outdoor residential exposure to benzene and respiratory health in a Spanish cohort. *Environ. Pollut.* **2017**, *222*, 486–494. [CrossRef] [PubMed]
47. D'Andrea, M.A.; Reddy, G.K. Health Risks Associated With Benzene Exposure in Children: A Systematic Review. *Glob. Pediatr. Health* **2018**, *5*, 2333794X18789275. [CrossRef] [PubMed]
48. O'Dell, K.; Hornbrook, R.S.; Permar, W.; Levin, E.J.; Garofalo, L.A.; Apel, E.C.; Blake, N.J.; Jarnot, A.; Pothier, M.A.; Farmer, D.K.; et al. Hazardous air pollutants in fresh and aged western US wildfire smoke and implications for long-term exposure. *Environ. Sci. Technol.* **2020**, *54*, 11838–11847. [CrossRef]
49. Jo, W.K.; Song, K.B. Exposure to volatile organic compounds for individuals with occupations associated with potential exposure to motor vehicle exhaust and/or gasoline vapor emissions. *Sci. Total Environ.* **2001**, *269*, 25–37. [CrossRef]
50. Redondo-Sánchez, D.; Petrova, D.; Rodríguez-Barranco, M.; Fernández-Navarro, P.; Jiménez-Moleón, J.J.; Sánchez, M.J. Socio-economic inequalities in lung cancer outcomes: An overview of systematic reviews. *Cancers* **2022**, *14*, 398. [CrossRef] [PubMed]
51. Wilms, R.; Mäthner, E.; Winnen, L.; Lanwehr, R. Omitted variable bias: A threat to estimating causal relationships. *Methods Psychol.* **2021**, *5*, 100075. [CrossRef]
52. Clarke, K.A. The phantom menace: Omitted variable bias in econometric research. *Confl. Manag. Peace Sci.* **2005**, *22*, 341–352. [CrossRef]
53. Pilotto, L.S.; Douglas, R.M.; Wilson, S.R. Respiratory effects associated with indoor nitrogen dioxide exposure in children. *Int. J. Epidemiol.* **1997**, *26*, 788–796. [CrossRef] [PubMed]
54. Gamble, J.; Jones, W.; Minshall, S. Epidemiological-environmental study of diesel bus garage workers: Acute effects of NO_2 and respirable particulate on the respiratory system. *Environ. Res.* **1987**, *42*, 201–214. [CrossRef] [PubMed]
55. Kubota, K.; Murakami, M.; Takenaka, S.; Kawai, K.; Kyono, H. Effects of long-term nitrogen dioxide exposure on rat lung: morphological observations. *Environ. Health Perspect.* **1987**, *73*, 157–169. [CrossRef]
56. Atkinson, R.W.; Butland, B.K.; Anderson, H.R.; Maynard, R.L. Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies. *Epidemiology* **2018**, *29*, 460. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Chapter 4

AI-Powered Insights into Environmental and Socioeconomic Influences on Metabolic Disease Mortality in Italy

This chapter investigates the relationship between air pollution, socio-economic factors, and mortality from endocrine, nutritional, and metabolic diseases in Italian provinces using machine learning (ML) and explainable artificial intelligence (XAI).

MDPI

Article

Explainable artificial intelligence for endocrine, nutritional and metabolic mortality in Italy with geolocalized pollution data

Donato Romano [1,2], Michele Magarelli [1], Pierfrancesco Novielli [1,2], Domenico Diacono [2], Pierpaolo Di Bitonto [1,2], Nicola Amoroso [2,4], Alfonso Monaco [2,3], Roberto Bellotti [2,3], Sabina Tangaro [1,2,*]

[1] Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy
[3] Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari Aldo Moro, Bari, Italy
[4] Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari Aldo Moro Bari, Italy
* Correspondence: sabina.tangaro@uniba.it

Abstract: This study investigates the predictive performance of three regression models—Gradient Boosting (GB), Random Forest (RF), and XGBoost—in forecasting mortality due to endocrine, nutritional, and metabolic diseases across Italian provinces. Utilizing a dataset encompassing air pollution metrics and socio-economic indices, the models were trained and tested to evaluate their accuracy and robustness. Performance was assessed using metrics such as coefficient of determination (r^2), mean absolute error (MAE) and root mean squared error (RMSE), revealing that GB outperformed both RF and XGB, offering superior predictive accuracy and model stability (r^2= 0.54, MAE = 0.17 and RMSE = 0.05). To further interpret the results, SHAP (SHapley Additive exPlanations) analysis was applied to the best-performing model to identify the most influential features driving mortality predictions. The analysis highlighted the critical roles of specific pollutants, including benzene and socio-economic factors such as life quality and instruction, in influencing mortality rates. These findings underscore the interplay between environmental and socio-economic determinants in health outcomes and provide actionable insights for policymakers aiming to reduce health disparities and mitigate risk factors. By combining advanced machine learning techniques with explainability tools, this research demonstrates the potential for data-driven approaches to inform public health strategies and promote targeted interventions in the context of complex environmental and social determinants of health.

Keywords: remote sensing; machine learning; nutritional disease; endocrine disease; metabolic disease; exposome; explainable artificial intelligence; air pollution; public health; socioeconomic indices;

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Copyright: © 2025 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Non-communicable diseases, particularly endocrine, nutritional, and metabolic disorders, represent a growing public health concern worldwide. These conditions, which include diabetes, thyroid disorders, and malnutrition-related complications, are influenced by complex interactions between genetic, environmental, and socio-economic factors [1]. In Italy, substantial regional disparities in mortality rates from such diseases suggest a significant role for external determinants, including air pollution and socio-economic conditions [2]. The incidence of these disorders varies across age groups, with type 1 diabetes often emerging in childhood or adolescence, posing lifelong management challenges [3], while type 2 diabetes and thyroid dysfunctions predominantly affect middle-aged and elderly populations due to aging, lifestyle factors and dietary habits [4]. Malnutrition-related complications, though less frequent in high-income countries, disproportionately impact vulnerable groups such as infants, the elderly, and socio-economically disadvantaged populations [5]. These conditions carry severe health consequences, including chronic

complications like cardiovascular disease, kidney failure, cognitive impairments, and heightened susceptibility to infections, which significantly reduce quality of life and life expectancy [6]. Socio-economic and social costs are equally profound, covering direct expenses for medical care, diagnostics, and treatment, as well as indirect costs arising from productivity losses, absenteeism, early retirement, and financial strain on families and caregivers [7]. In Italy, diabetes management alone consumes a substantial portion of healthcare resources, highlighting the immense pressure these diseases place on public health systems [8]. Socio-economically disadvantaged areas bear a disproportionate burden, exacerbating health inequalities and underscoring the need for comprehensive public health strategies that integrate preventive measures, early detection programs, and equitable access to care and support systems [9].

Air pollution has been widely recognized as a major contributor to various health conditions, including endocrine and metabolic disorders. Studies have linked exposure to pollutants such as particulate matter (PM10, PM2.5), nitrogen oxides (NO and NO_2), and ozone (O_3) to systemic inflammation, hormonal disruptions, and metabolic imbalances [10–12]. Concurrently, socio-economic variables, such as income, education levels, healthcare access, and urbanization, are known to modulate health outcomes, acting as either risk amplifiers or protective buffers [13]. However, the interplay between these factors remains underexplored, particularly in the context of mortality associated with endocrine, nutritional and metabolic diseases.

The advent of machine learning (ML) offers powerful tools to analyze large, complex datasets and uncover hidden patterns that traditional statistical methods may overlook. ML models can integrate diverse data sources, ranging from environmental pollution metrics to socio-economic indicators, to predict health outcomes with high accuracy [14]. Moreover, explainable artificial intelligence (XAI) frameworks, such as Shapley Additive Explanations (SHAP), enable the interpretation of model predictions, shedding light on the relative importance and interplay of contributing factors [15]. This combination of predictive capability and interpretability makes ML a valuable approach for addressing public health challenges [16–18].

This study applies ML techniques to assess the likelihood of mortality from endocrine, nutritional, and metabolic diseases across Italian provinces. By integrating air pollution data sourced from the Copernicus Atmosphere Monitoring Service (CAMS) platform with socio-economic and mortality statistics obtained from ISTAT (Italian National Institute of Statistics), we aim to identify key predictors and quantify their contributions to health outcomes.

This paper has the following structure: Section 2 provides the description of the data used and the adopted methodologies. In particular, Section 2.1 defines the dataset composition and its preparation to ML analysis ; Section 2.2 presents the reduction of the features of the dataset through the Variation Inflation Factor; Section 2.3 describes the leave-one-group-out cross validation; Section 2.4 introduces the 3 machine algorithms and the evaluation metrics used; Section 2.5, presents the explainable artificial intelligence approach adopted. Section 3 shows the results of our analysis that are discussed in Section 4. Finally, Section 5 contains our conclusions.

2. Materials and Methods

2.1. Data preparation description

The data used refer to those of the previously conducted study [16], on air quality and atmospheric composition acquired by Copernicus Atmosphere Monitoring Service (CAMS). In particular, the pollutant data provided by CAMS are derived from the median ensemble of nine numerical air quality models [19–26]. In this study, which examines the relationship between air pollution and mortality for endocrine nutritional and metabolic disease, specific pollutant concentrations such as CO, NO, NO_2, O_3, PM10, PM2.5, and SO_2 were analyzed, using the Python library Geopandas. To enhance data consistency,

additional variables related to pollution and socio-economic factors were included. Figure 1 presents an outline of the data preparation workflow.

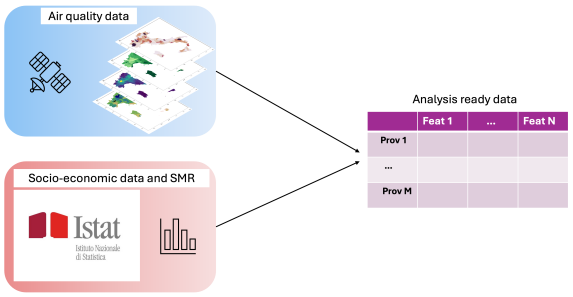

Figure 1. Dataset preparation workflow.

In this study, we calculated the standardized mortality ratio (SMR), which represents the ratio of observed deaths in a given area to the number of deaths expected in the same area. Expected deaths were determined by applying the age-specific mortality rates from a reference population (in this case, the national population) to the average annual population, stratified by age groups, within each territorial unit. Figure 2 illustrates the distribution of the average SMR for endocrine, nutritional and metabolic diseases across Italian provinces in 2019.

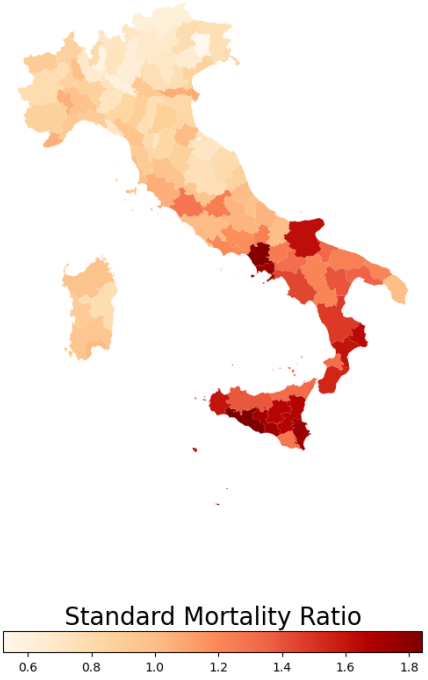

Figure 2. Standardized mortality ratio distribution in 2019 for endocrine, nutritional and metabolic diseases within Italian provinces.

Figure 3 provides a visual overview of the analytical process carried out in this study. The first step of this work has been to create a well-structured dataset of the Italian provinces with multiple types of data: the air pollution data were collected from the CAMS platform, while the socio-economic indices such as life quality distribution and density of urban green areas, and the mortality data related to endocrine, nutritional and metabolic diseases have been obtained from the database of the Italian National Institute of Statistics (ISTAT). During the preliminary phase of the analysis, the multicollinearity of the dataset variables was assessed by calculating Variation Inflation Factor (VIF). All variables with a VIF greater than 10 were iteratively removed from the dataset.

To prevent information leakage between spatially related observations, we first identified groups of neighboring provinces (clusters) that were kept separate during the training and validation phases. These spatial clusters were determined using a combination of methods. A hierarchical clustering algorithm was applied to the matrix of Euclidean distances between provinces, and the semivariogram plot of the SMRs was used to identify the point at which to stop hierarchical clustering. The spatial range of the semivariogram, which indicates the distance beyond which observations are assumed to be uncorrelated, was visually estimated. This was done by identifying the point where the plot levels off (the plateau), indicating that the variance no longer increases significantly with distance.

We applied a leave-one-cluster-out cross-validation method to reduce spatial bias while evaluating the regression performance of three ensemble models: Gradient Boosting [27], Random Forest [28], and XGBoost [29]. To compare the performance of these machine learning models, we calculated metrics such as the coefficient of determination (r^2), the mean absolute error (MAE) and the root mean squared error (RMSE). In addition, we identified the most influential features for the model using the SHAP algorithm, an approach from explainable artificial intelligence.

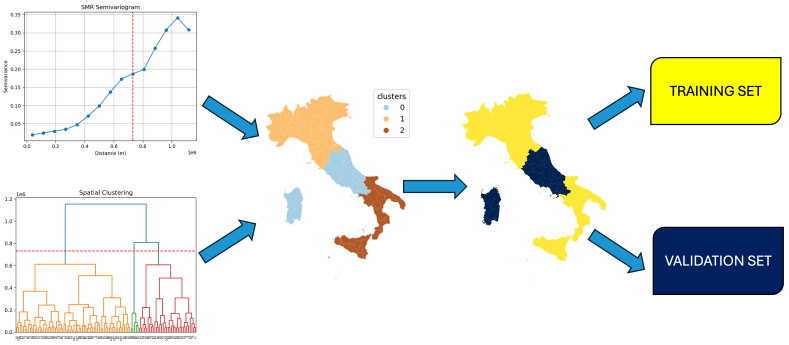

Figure 3. The hierchical clustering algorithm used in combination with the semivariogram plot of the SMR rates delineated three spatially-contiguous community that alleviated the spatial data leakage that might occurr in epidemiological studies.

2.2. Multicollonearity analysis

The Variation Inflation Factor (VIF) is a tool used in regression analysis to check for multicollinearity among independent variables. It measures how strongly the independent variables are correlated and indicates how much the variance of the regression coefficients is inflated due to this correlation.

In this study, we computed VIF to identify and address multicollinearity in the dataset. We applied an iterative method, where the VIF was recalculated multiple times, removing

the variable with the highest VIF in each step if its value exceeded 10. This approach helped to minimize multicollinearity and improved the reliability of the coefficient estimates.

The variables removed after this analysis are: mean pm2.5, income, bed number, mean CO and mean NO.

2.3. Leave-one-group-out cross-validation

To prevent information leakage between observations with spatial dependencies, we divided the provinces into clusters of neighboring regions. These clusters were kept separate during the training and validation phases. The spatial clusters used in the cross-validation process were identified by applying the hierarchical clustering algorithm to the matrix of Euclidean distances between provinces, and the semivariogram plot of the SMRs.

A semivariogram is a geostatistical tool for evaluating spatial dependence in a variable. It is represented on a graph where the abscissa shows the distance (spatial lag) between pairs of locations, and the ordinate displays the average squared difference in the variable's values for those pairs within a specified distance range. In this study, we visually estimated the spatial range of the semivariogram defined as the distance beyond which observations are no longer correlated by identifying the plateau, where variance stops increasing significantly.

Using the spatial range from the semivariogram of the SMRs, we defined three clusters by applying a threshold to the dendrogram created from the hierarchical clustering of geographic distances between Italian provinces.

2.4. Machine learning models

The goal of our analysis was to compare the performance of three different machine learning models, Gradient Boosting (GB), Random Forest (RF) and XGBoost (XGB), to predict mortality for endocrine, nutritional and metabolic diseases through geolocalized data.

Gradient Boosting is an ensamble machine learning model that combines the strengths of multiple weak learners, typically decision trees, to create a strong learner. The algorithm works iteratively, where each subsequent model is trained to minimize the residual errors (i.e., the differences between predicted and actual values) of the preceding model using a gradient descent approach. By sequentially adding models, Gradient Boosting reduces bias and variance, improving overall performance. Hyperparameters such as the learning rate, maximum tree depth, and number of estimators play a critical role in controlling the model's complexity and preventing overfitting. In this work we set the learning rate to 0.1, the maximum tree depth to 3 and the number of estimators to 100, a typical parameter setting.

Random Forest regressor is an ensemble learning algorithm known for its robustness and ability to handle high-dimensional data. It operates by constructing a multitude of decision trees during training and combines their outputs to produce a final prediction through averaging. The algorithm introduces randomness by using bootstrapped samples of the data to train each tree and by selecting a random subset of features at each split. This approach reduces overfitting and improves generalization compared to individual decision trees. We chose 100 trees to build the forest, a common setting.

XGBoost (Extreme Gradient Boosting) is a high-performance implementation of the gradient boosting algorithm, designed to optimize both computational efficiency and predictive accuracy. XGBoost builds an ensemble of decision trees sequentially, where each tree aims to reduce the residual errors of its predecessors using a gradient descent method. Its enhancements include tree pruning, regularization terms in the loss function, and support for parallel and distributed computation, making it faster and more robust than traditional gradient boosting techniques.

We evaluated the perfomances of the three machine models by computing the coefficient of determination r^2, the mean absolute error MAE and the root mean squared error $RMSE$:

- Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

- Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \quad (2)$$

- Root mean squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where $\hat{y}_i$ are the predicted values, $\bar{y}$ is their average, while y_i are the observed values of the SMR in the validation set.

2.5. Explanation of the features

To interpret the results of our machine learning models and identify the most influential predictors, we utilized SHAP (SHapley Additive exPlanations), a state-of-the-art algorithm for explainable artificial intelligence (XAI). SHAP provides a unified framework for quantifying the contribution of each feature to the model's predictions, offering both global and local interpretability.

The SHAP method is based on cooperative game theory, where each feature is treated as a "player" in a predictive game, and its contribution to the final prediction is assessed in terms of Shapley values. These values are calculated by considering the marginal contribution of a feature across all possible subsets of features, ensuring fairness and consistency. Unlike traditional feature importance methods, SHAP assigns feature contributions in a manner that accounts for interactions and dependencies among predictors:

$$\Phi_j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!}[f_x(F \cup j) - f_x(F)] \quad (4)$$

where x is an instance, the sum is over all the subsets S of features which include the feature j, $\frac{|F|!(|S|-|F|-1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that doesn't belong to S, while $f_x(F \cup j)$ and $f_x(F)$ denote respectively the regression score obtained by including and removing feature j.

In this study, we used SHAP to generate explanations for the predictions of our ensemble models. Specifically, SHAP summary plots were employed to visualize the overall importance of each feature across the dataset. This analysis provided valuable insights into how different variables, such as pollution levels, socio-economic indicators, and mortality data, influenced the model's predictions.

3. Results

The goal of this work was to evaluate, using explainable machine learning models, whether and to what extent air pollutants and socio-economic descriptors are associated with mortality due to endocrine, nutritional and metabolic diseases.

3.1. Regression performance

In order to validate the regression performance, by using the spatial range estimated from the semivariogram plot, we divided the dendrogram of geographic distances between Italian provinces into three distinct clusters. To ensure a reliable evaluation of model performance, we implemented a leave-one cluster-out validation approach on each of the three different models. The performance of the models were evaluated quantitatively by

calculating the average metrics obtained from the validation set. The results of the model metrics are shown in Table 1.

Table 1. Performance of different regression models.

Model	r^2	MAE	RMSE
Gradient Boosting	0.54	0.17	0.05
Random Forest	0.46	0.19	0.05
XGBoost	0.04	0.26	0.10

The Gradient Boosting regressor emerged as the best in predicting SMR for endocrine, nutritional and metabolic diseases with a r^2 of 0.54. The scatter plot of the predicted SMR vs the real ones is shown in figure 4.

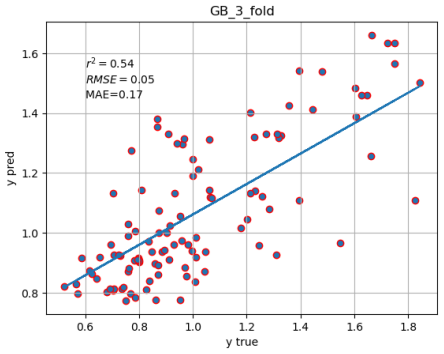

Figure 4. Scatter plot of predicted Standard Mortality Ratio (y pred) vs true Standard Mortality Ratio (y true). The blue line represents the best fit for the plotted points.

3.2. Interpretation of models through SHAP values

Figure 5 shows the feature importance obtained with SHAP from the best regression model GB. The analysis identified life quality and lifetime as the most influential features. Features such as instruction, benzene and temperature also contributed to the predictions, although with lower SHAP values. The plot showed both the magnitude and direction of the impact for each feature, with higher SHAP values indicating stronger influence on the predicted mortality rates. The x-axis represents the SHAP values, showing the range and direction of the impact a feature has on the model's predictions. Positive SHAP values indicate an increase in the prediction of a higher value of outcome, while negative values increase in the prediction of an lower value of outcome. Points are color-coded to reflect the feature's value for the observation. Red is used for high feature values and blue for low values, helping to show how feature value correlates with its impact.

4. Discussion

This study analyzed the association between air pollution and mortality from endocrine, nutritional, and metabolic diseases, including geolocalized data on CO, NO, NO_2, O_3, PM10, PM2.5, and SO_2, and considering demographic and socio-economic variations such as age and gender. The standardized mortality ratio (SMR) was calculated as the ratio between observed deaths in a provincial area and expected ones. Our findings show that most of the input variables included in this study are weakly correlated with the target variable, the standard mortality ratio, while our gradient boosting model accurately predicts provinces with an SMR higher than the mean value. The performances obtained in terms of r² and MAE are 0.54 and 0.17. This result shows that environmental pollution is associated with endocrine, nutritional, and metabolic diseases. These findings should be

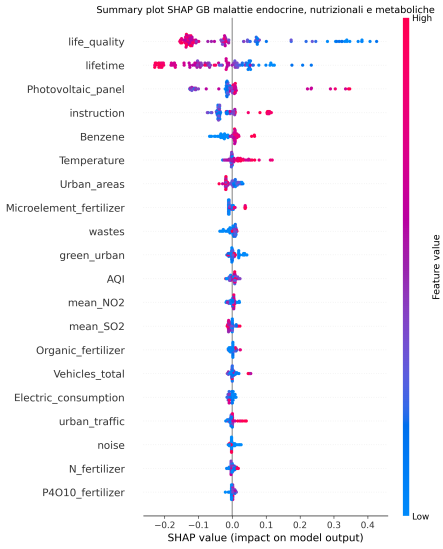

Figure 5. SHAP summary plot showing the Shapley values for each feature. Each point represents the Shapley value for a specific province, with the y-axis displaying the associated feature and the x-axis showing the Shapley value. The color gradient indicates the feature's value, ranging from low to high. Features are ranked by their average importance, with the most significant ones appearing at the top.

included, along with other risk factors, in studies investigating risk factors for individuals, even in personalized contexts. Explainable AI (XAI) estimated the most important global and local variables in predicting the SMR. The SHAP algorithm was used to identify the most influential variables in the model, highlighting the link between environmental, socio-economic factors, and population health. Quality of life emerges as the most important factor: in areas where it is higher, lower mortality rates are observed [30,31]. This result emphasizes how a healthy lifestyle, access to services, and a favorable environment can positively impact health. At the same time, life expectancy is closely linked to mortality rates: in areas where people live shorter lives, the impact of endocrine diseases seems to be particularly relevant. High values of the variable photovoltaic panel significantly influence the GB regressor's predictions of elevated mortality rates for endocrine, nutritional, and metabolic diseases. However, this pattern can be attributed to the geographic distribution of photovoltaic panels in Italy. Provinces with a density of solar panels per square kilometer exceeding the national average are predominantly located in southern Italy, the same region with a higher SMR for these diseases. Although no scientific studies have examined a direct correlation between the percentage of students who did not complete primary education (variable instruction) and deaths from endocrine, nutritional, and metabolic diseases, it is well established that lower education levels are associated with less healthy behaviors and reduced access to health information factors that can negatively impact overall health [32,33]. Education plays a crucial role in the prevention and management of chronic diseases, including endocrine, nutritional, and metabolic disorders. Therefore, while there is no direct evidence of a specific correlation, it is plausible that a higher dropout rate may indirectly contribute to an increase in mortality from these conditions. Among the pollutants analyzed, benzene stands out for its correlation with higher mortality rates. This is concerning, considering that benzene is often produced by activities such as vehicular traffic and combustion [34]. A higher average annual temperature is associated with a greater probability of mortality from endocrine, nutritional and metabolic diseases according to the model used, in line with the results present in the literature [35]. Air quality

and the average concentration of NO_2 are also strongly linked to mortality, confirming the role of air pollution in the development of chronic diseases [36–38]. As for socio-economic and agricultural factors, such as the use of fertilizers, especially those with microelements, they suggest a relationship with rural environments and chemical exposures linked to agriculture. Research has revealed that individuals working in the fertilizer industry tend to have elevated levels of Cd and Pb in their biological samples, compared to healthy individuals who are either unexposed or exposed across all age groups. This outcome is likely linked to the buildup of toxic metals stemming from fertilizer contamination [39]. The study in question has limitations that may influence the interpretation of the results and their applicability in a broader context. Below are the main critical aspects explained. The analysis is based on data collected at the provincial level, treating each province as a homogeneous entity. However, within provinces, there are significant differences in population density, such as those between densely populated urban areas and sparsely inhabited rural areas. In this work it is assumed that the entire population of a province is exposed to the same levels of air pollutants. While the study is observational in nature and identifies associations between air pollutants and endocrine/metabolic diseases, but however it cannot establish a direct causal relationship. This means that the results could be influenced by confounding factors not considered, necessitating further analysis to confirm the actual causes. Finally, feature such as benzene, identified as a pollutant with a significant impact on mortality, are not delve into the specific sources (e.g., vehicle traffic or industrial activities) nor does it consider other exposures that could interact with benzene, making it difficult to assign precise responsibility to this pollutant.

5. Conclusions

This study demonstrates the potential of machine learning (ML) models to identify the likelihood of mortality from endocrine, nutritional, and metabolic diseases by analyzing environmental pollution and socio-economic data at the provincial level in Italy. By integrating diverse data streams, including air quality metrics, socio-demographic variables, and health records, we provided a comprehensive framework for investigating complex relationships between environmental exposures and public health outcomes.

Our findings reveal that key predictors of mortality are not limited to air pollutant concentrations but also encompass socio-economic factors such as life quality. The use of interpretable ML techniques, such as Shapley Additive Explanations (SHAP), allowed us to uncover the relative importance of each feature and provided valuable insights into the underlying mechanisms linking environmental and social factors to health risks.

The performance of the models, while promising, highlights the need for further refinement. Future work should focus on incorporating additional datasets, such as longitudinal data, to improve the robustness and generalizability of the models. Moreover, enhancing spatial resolution and integrating behavioral and genetic data could provide a more granular understanding of individual and community-level risks.

This work underscores the importance of leveraging data-driven approaches to address pressing public health challenges. Policymakers and stakeholders can use these findings to design targeted interventions aimed at reducing health disparities and mitigating the effects of environmental pollution on vulnerable populations. By advancing our understanding of the interplay between environment, socio-economic conditions, and health, this research lays the groundwork for more informed and equitable public health strategies.

Author Contributions: Conceptualization, D.R., S.T.; methodology, D.R., P.N., D.D., and S.T.; software, D.R. and D.D.; validation, D.R., P.N., D.D., P.D., R.B. and S.T.; investigation, D.R., P.N.; data curation, D.R.; writing—original draft preparation, D.R., M.M., P.N. and S.T.; writing—review and editing, D.R., M.M., P.N., P.D., D.D., N.A., A.M., R.B. and S.T.; visualization, D.R., D.D. and S.T.; supervision, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This paper has been supported by the METROFOOD-IT project, a project with funding received from the European Union—NextGenerationEU, PNRR—Mission 4 "Education and Research" Component 2: from research to business, Investment 3.1: Fund for the realization of an integrated system of research and innovation infrastructures - IR0000033 (D.M. Prot. n.120 del 21/06/2022), next_AIM (Artificial Intelligence in Medicine: next steps) research project (INFN-CSN5), https://www.pi.infn.it/aim (accessed on 30 October 2023); The National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number: Project code: CN00000013, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: "National Centre for HPC, Big Data and Quantum Computing."

Institutional Review Board Statement: Not applicable

Data Availability Statement: The data that support the findings of this study are either publicly available on databases cited in the bibliography, or available from the corresponding author on request.

Acknowledgments: Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the "PON Ricerca e Competitività 2007–2013-Azione I-Interventi di rafforzamento strutturale" *PONa*3_00052, Avviso 254/Ric, University of Bari.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Organization, W.H. Noncommunicable Diseases Progress Monitor 2018. *World Health Organization* **2018**.
2. of Statistics (ISTAT), I.N.I. Italian National Health Statistics Report 2021. *ISTAT* **2021**. Available at: https://www.istat.it/en/health.
3. Michels, A.W.; Redondo, M.J.; Atkinson, M.A. The pathogenesis, natural history, and treatment of type 1 diabetes: time (thankfully) does not stand still. *The lancet Diabetes & endocrinology* **2022**, *10*, 90–92.
4. Kim, S.H.; Lee, S.Y.; Kim, C.W.; Suh, Y.J.; Hong, S.; Ahn, S.H.; Seo, D.H.; Nam, M.S.; Chon, S.; Woo, J.T.; et al. Impact of socioeconomic status on health behaviors, metabolic control, and chronic complications in type 2 diabetes mellitus. *Diabetes & metabolism journal* **2018**, *42*, 380.
5. Myers, S.S.; Smith, M.R.; Guth, S.; Golden, C.D.; Vaitla, B.; Mueller, N.D.; Dangour, A.D.; Huybers, P. Climate change and global food systems: potential impacts on food security and undernutrition. *Annual review of public health* **2017**, *38*, 259–277.
6. Yen, F.S.; Wei, J.C.C.; Chiu, L.T.; Hsu, C.C.; Hwu, C.M. Diabetes, hypertension, and cardiovascular disease development. *Journal of Translational Medicine* **2022**, *20*, 1–12.
7. Bommer, C.; Heesemann, E.; Sagalova, V.; Manne-Goehler, J.; Atun, R.; Bärnighausen, T.; Vollmer, S. The global economic burden of diabetes in adults aged 20–79 years: a cost-of-illness study. *The lancet Diabetes & endocrinology* **2017**, *5*, 423–430.
8. Pagano, E.; De Rosa, M.; Rossi, E.; Cinconze, E.; Marchesini, G.; Miccoli, R.; Vaccaro, O.; Bonora, E.; Bruno, G. The relative burden of diabetes complications on healthcare costs: the population-based CINECA-SID ARNO Diabetes Observatory. *Nutrition, Metabolism and Cardiovascular Diseases* **2016**, *26*, 944–950.
9. Di Cesare, M.; Khang, Y.H.; Asaria, P.; Blakely, T.; Cowan, M.J.; Farzadfar, F.; Guerrero, R.; Ikeda, N.; Kyobutungi, C.; Msyamboza, K.P.; et al. Inequalities in non-communicable diseases and effective responses. *The Lancet* **2013**, *381*, 585–597.
10. Pope, C.A.; Dockery, D.W. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air & Waste Management Association* **2006**, *56*, 709–742.
11. Thurston, G.D.; et al. A joint ERS/ATS policy statement: what constitutes an adverse health effect of air pollution? An analytical framework. *European Respiratory Journal* **2017**, *49*.
12. Monaco, A.; Lacalamita, A.; Amoroso, N.; D'Orta, A.; Del Buono, A.; di Tuoro, F.; Tangaro, S.; Galeandro, A.I.; Bellotti, R. Random forests highlight the combined effect of environmental heavy metals exposure and genetic damages for cardiovascular diseases. *Applied Sciences* **2021**, *11*, 8405.
13. Braveman, P.A.; Egerter, S.; Williams, D.R. The social determinants of health: coming of age. *Annual Review of Public Health* **2011**, *32*, 381–398.
14. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *New England Journal of Medicine* **2019**, *380*, 1347–1358.
15. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2017, Vol. 30.
16. Romano, D.; Novielli, P.; Diacono, D.; Cilli, R.; Pantaleo, E.; Amoroso, N.; Bellantuono, L.; Monaco, A.; Bellotti, R.; Tangaro, S. Insights from Explainable Artificial Intelligence of Pollution and Socioeconomic Influences for Respiratory Cancer Mortality in Italy. *Journal of Personalized Medicine* **2024**, *14*, 430.

17. Novielli, P.; Romano, D.; Magarelli, M.; Diacono, D.; Monaco, A.; Amoroso, N.; Vacca, M.; De Angelis, M.; Bellotti, R.; Tangaro, S. Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence. *Iscience* **2024**, *27*.
18. Romano, D.; Novielli, P.; Cilli, R.; Amoroso, N.; Monaco, A.; Bellotti, R.; Tangaro, S. Air pollution and mortality for cancer of the respiratory system in Italy: an explainable artificial intelligence approach. *Frontiers in Public Health* **2024**, *12*, 1344865.
19. Thunis, P.; Degraeuwe, B.; Pisoni, E.; Meleux, F.; Clappier, A. Analyzing the efficiency of short-term air quality plans in European cities, using the CHIMERE air quality model. *Air Quality, Atmosphere & Health* **2017**, *10*, 235–248.
20. Hass, H.; Ebel, A.; Feldmann, H.; Jakobs, H.; Memmesheimer, M. Evaluation studies with a regional chemical transport model (EURAD) using air quality data from the EMEP monitoring network. *Atmospheric Environment. Part A. General Topics* **1993**, *27*, 867–887.
21. Duarte, E.D.S.F.; Franke, P.; Lange, A.C.; Friese, E.; da Silva Lopes, F.J.; da Silva, J.J.; dos Reis, J.S.; Landulfo, E.; e Silva, C.M.S.; Elbern, H.; et al. Evaluation of atmospheric aerosols in the metropolitan area of São Paulo simulated by the regional EURAD-IM model on high-resolution. *Atmospheric Pollution Research* **2021**, *12*, 451–469.
22. Hinestroza-Ramirez, J.E.; Lopez-Restrepo, S.; Yarce Botero, A.; Segers, A.; Rendon-Perez, A.M.; Isaza-Cadavid, S.; Heemink, A.; Quintero, O.L. Improving Air Pollution Modelling in Complex Terrain with a Coupled WRF–LOTOS–EUROS Approach: A Case Study in Aburrá Valley, Colombia. *Atmosphere* **2023**, *14*, 738.
23. Persson, C.; Langner, J.; Robertson, L. Air pollution assessment studies for Sweden based on the MATCH model and air pollution measurements. *Air pollution modeling and its application XI* **1996**, pp. 127–134.
24. Joly, M.; Josse, B.; Plu, M.; Arteta, J.; Guth, J.; Meleux, F. High-Resolution Air Quality Forecasts with MOCAGE Chemistry Transport Model. In Proceedings of the Air Pollution Modeling and its Application XXIV. Springer, 2016, pp. 563–565.
25. Van Loon, M.; Vautard, R.; Schaap, M.; Bergström, R.; Bessagnet, B.; Brandt, J.; Builtjes, P.; Christensen, J.; Cuvelier, C.; Graff, A.; et al. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmospheric Environment* **2007**, *41*, 2083–2097.
26. Neary, L.; Kaminski, J.W.; Lupu, A.; McConnell, J.C. Developments and results from a global multiscale air quality model (GEM-AQ). In *Air Pollution Modeling and Its Application XVII*; Springer, 2007; pp. 403–410.
27. Mayr, A.; Binder, H.; Gefeller, O.; Schmid, M. The evolution of boosting algorithms. *Methods of information in medicine* **2014**, *53*, 419–427.
28. Parmar, A.; Katariya, R.; Patel, V. A review on random forest: An ensemble classifier. In Proceedings of the International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer, 2019, pp. 758–763.
29. Azmi, S.S.; Baliga, S. An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. *Int. Res. J. Eng. Technol* **2020**, *7*.
30. Webb, S.M.; Crespo, I.; Santos, A.; Resmini, E.; Aulinas, A.; Valassi, E. Management of endocrine disease: quality of life tools for the management of pituitary disease. *European Journal of Endocrinology* **2017**, *177*, R13–R26.
31. Sonino, N.; Fava, G.A.; Fallo, F.; Boscaro, M. Psychological distress and quality of life in endocrine disease. *Psychotherapy and Psychosomatics* **1990**, *54*, 140–144.
32. Klinker, C.D.; Aaby, A.; Ringgaard, L.W.; Hjort, A.V.; Hawkins, M.; Maindal, H.T. Health literacy is associated with health behaviors in students from vocational education and training schools: a Danish population-based survey. *International journal of environmental research and public health* **2020**, *17*, 671.
33. Zajacova, A.; Lawrence, E.M. The relationship between education and health: reducing disparities through a contextual approach. *Annual review of public health* **2018**, *39*, 273–289.
34. Cui, Y.; Mo, Z.; Ji, P.; Zhong, J.; Li, Z.; Li, D.; Qin, L.; Liao, Q.; He, Z.; Guo, W.; et al. Benzene exposure leads to lipodystrophy and alters endocrine activity in vivo and in Vitro. *Frontiers in Endocrinology* **2022**, *13*, 937281.
35. Li, Y.; Cheng, Y.; Cui, G.; Peng, C.; Xu, Y.; Wang, Y.; Liu, Y.; Liu, J.; Li, C.; Wu, Z.; et al. Association between high temperature and mortality in metropolitan areas of four cities in various climatic zones in China: a time-series study. *Environmental Health* **2014**, *13*, 65.
36. Saleem, A.; Awan, T.; Akhtar, M.F. A comprehensive review on endocrine toxicity of gaseous components and particulate matter in smog. *Frontiers in Endocrinology* **2024**, *15*, 1294205.
37. Darbre, P.D. Overview of air pollution and endocrine disorders. *International journal of general medicine* **2018**, pp. 191–207.
38. Gea, M.; Fea, E.; Racca, L.; Gilli, G.; Gardois, P.; Schilirò, T. Atmospheric endocrine disruptors: A systematic review on oestrogenic and androgenic activity of particulate matter. *Chemosphere* **2024**, *349*, 140887.
39. Ning, J.; Akhter, T.; Sarfraz, M.; Afridi, H.I.; Albasher, G.; Unar, A. The importance of monitoring endocrine-disrupting chemicals and essential elements in biological samples of fertilizer industry workers. *Environmental Research* **2023**, *231*, 116173.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Chapter 5

Conclusions

The final chapter synthesizes the findings, discussing AI's transformative role in environmental research. It highlights the contributions of explainable ML in enhancing transparency and ethical decision-making and suggests future directions to broaden AI applications in environmental science.

This thesis has demonstrated the power of machine learning (ML) and explainable artificial intelligence (XAI) in advancing environmental epidemiology. By systematically analyzing the relationship between environmental factors and non-communicable diseases (NCDs), it provides valuable insights into the role of air pollution, socio-economic conditions, and urbanization in shaping public health outcomes. The findings from the three key studies presented in this work offer actionable insights for policymakers and public health officials, highlighting the urgent need for targeted interventions to mitigate environmental health risks.

The first study, focusing on respiratory system cancers, identified nitrogen monoxide (NO), income levels, and ozone (O_3) concentrations as the most significant predictors of mortality. These findings suggest that reducing NO emissions, particularly in urban areas with high traffic density, should be a priority for public health interventions. Additionally, policies aimed at improving air quality, such as stricter emissions standards and the promotion of green urban spaces, could help mitigate the impact of ozone on respiratory health. The study also highlighted the role of socio-economic factors, indicating that income disparities may exacerbate exposure to harmful pollutants, particularly in wealthier provinces where industrial activities are more concentrated. This underscores the need for equitable environmental policies that address both pollution control and socio-economic inequalities.

The second study, which compared thirteen classification models to predict respiratory cancer mortality rates above or below the national average, reinforced the importance of NOx, benzene, and income levels as key risk factors. The Extra Trees classifier emerged as the most accurate model, providing a robust framework for identifying high-risk provinces. These results can guide resource allocation, enabling policymakers to prioritize regions with higher mortality rates for targeted interventions, such as enhanced healthcare access, pollution monitoring, and public awareness campaigns. The study also demonstrated the value of XAI in interpreting complex ML models, offering transparency and trustworthiness in AI-driven epidemiological research.

The third study, which examined endocrine, nutritional, and metabolic diseases, revealed that life quality, lifetime, and benzene exposure were the most influential predictors of mortality. These findings emphasize the interplay between environmental and socio-economic determinants of health. For instance, improving life quality through better access to healthcare, education, and social services could significantly reduce mortality rates. In addition, reducing benzene exposure, particularly in industrial and urban areas, should be a key focus of environmental regulations. The study also highlighted the potential of Gradient Boosting as a powerful tool for predicting health outcomes, further validating the use of ML in public health research.

While the studies presented in this thesis provide valuable insights, they are not without limitations. One major limitation is the reliance on provincial-level data, which may mask significant variations in pollution exposure and socio-economic conditions within provinces. For example, urban areas within a province may experience higher pollution levels compared to rural areas, but this granularity is lost in the analysis. Additionally, the studies are based on cross-sectional data from a single year (2019), which limits the ability to assess long-term trends or causal relationships. Future research should incorporate longitudinal data to better understand the temporal dynamics of environmental exposures and health outcomes.

Another limitation is the potential for confounding factors that were not accounted for in the models. For instance, individual-level data on lifestyle factors, genetic predispositions, and access to healthcare could provide a more comprehensive understanding of the determinants of NCDs. Furthermore, while SHAP values offer interpretability, they do not establish causality, and the models may still be influenced by unobserved variables. These limitations highlight the need for caution when translating the findings into policy recommendations.

The findings of this thesis open several avenues for future research. First, expanding the geographic scope of the analysis to include other regions or countries could provide a broader understanding of how environmental and socio-economic factors influence health

outcomes across different contexts. Comparative studies could reveal regional variations in risk factors and inform tailored interventions.

Second, incorporating longitudinal data would allow for the examination of long-term trends and the assessment of the effectiveness of public health interventions over time. For example, tracking changes in air quality and mortality rates following the implementation of pollution control measures could provide valuable insights into the impact of policy changes.

Third, integrating additional data sources, such as genetic information, behavioral data, and more granular environmental monitoring data, could enhance the predictive power of ML models. For instance, combining satellite-based pollution data with ground-level sensor data could improve the accuracy of exposure assessments. Similarly, incorporating data on individual health behaviors, such as smoking or diet, could provide a more holistic understanding of the determinants of NCDs.

Fourth, future research could explore the use of more advanced ML techniques, such as deep learning, to capture complex, non-linear relationships between environmental exposures and health outcomes. While these models may offer higher predictive accuracy, their "black-box" nature poses challenges for interpretability. Therefore, the development of new XAI techniques that can provide insights into the decision-making processes of deep learning models will be crucial for their adoption in public health research.

Finally, the ethical and regulatory implications of AI in public health warrant further investigation. As AI-driven tools become more integrated into policy-making, it will be essential to ensure that these technologies are deployed responsibly and equitably. This includes addressing potential biases in the data, ensuring transparency in model predictions, and engaging stakeholders in the development and implementation of AI-based interventions.

In conclusion, this thesis contributes to the growing field of AI-driven epidemiology by providing a transparent and interpretable framework to assess environmental health risks. The findings have significant implications for public health policies, emphasizing the urgent need for targeted interventions to mitigate the impact of environmental stressors

on population health. By bridging machine learning, explainability, and epidemiology, this work lays the foundation for future advancements in AI-assisted health risk assessment and environmental policy development. The integration of these methodologies into public health practice has the potential to transform our understanding of the complex interactions between environmental exposures and human health, ultimately leading to healthier and more sustainable communities.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione”
e Azione IV.5 “Dottorati su tematiche Green”.