Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 1, April (2025), Pages 20-29
e-ISSN: 2829-310X | 10.29207/joseit.

Machine Learning Models for Air Pollution Health Risk Assessment

Lipatova A.V[1*], Potapchenko T.D[2]
[12]Moscow Technical University of Communications and Informatics, Moscow, Russian Federation

[*]*myagkova.anastasia@mail.ru*

Abstract. This study explored the application of machine learning (ML) models and artificial neural networks (ANNs) in the assessment of public health concerns associated with air pollution. Utilizing a dataset comprising over 12,000 records from India and Nepal, encompassing both quantitative measurements and visual data, several classification models have been constructed and evaluated to predict air quality index (AQI) categories indicative of varying health risk levels. The implemented models comprise a decision tree (DT), support vector machine (SVM), random forest (RF), XGBoost, and deep neural networks (both convolutional and recurrent). The methodology entailed data preprocessing, feature significance analysis, and model assessment using accuracy metrics and ROC curves. The findings revealed a high classification accuracy across all models (>90%), with ensemble-based methods demonstrating enhanced performance. XGBoost attained superior accuracy with optimal resource efficiency; however, artificial neural network (ANN) models, particularly long short-term memory (LSTM), obtained accuracy levels of 98% by the 15th training epoch. Feature significance analysis revealed that AQI, PM2.5, and PM10 were the primary predictors of health risk categorization. Correlation analysis demonstrated robust associations between particulate matter measures (PM2.5, PM10), underscoring their significance in air quality evaluation. This study proposes a methodological framework for automating risk assessment procedures using machine learning approaches to facilitate more effective environmental health monitoring. The findings suggest that ensemble models offer an optimal balance between precision and computing efficiency for real-time air quality classification systems with potential applications in early warning systems and public health intervention techniques.

Keywords: *air pollution classification; public health risk assessment; machine learning; ensemble models; environmental monitoring*

How to cites:
Lipatova A.V, & Potapchenko T.D. (2025). Machine Learning Models for Air Pollution Health Risk Assessment. Journal of Systems Engineering and Information Technology (JOSEIT), 4(1). https://doi.org/10.29207/joseit.v4i1.6544

Received by the Editor: April 15, 2025
Final Revision: April 24, 2025
Published: 2025-04-30

SDGs contributed:

1. Introduction

The need to implement effective measures to prevent the development of human diseases caused by a variety of anthropogenic factors has led to an increased demand for the automation of processes involved in the analysis of heterogeneous and large volumes of environmental data [1], [2]. This necessity is the primary driver of demand for such solutions. Modern approaches to data analysis, underpinned by artificial intelligence (AI) technologies, machine learning (ML), and deep learning (DL) algorithms and models [3], [4], [5], including artificial neural networks (ANN), have emerged as a consequence of existing risk-oriented approaches to assessing the consequences of anthropogenic factors and their impact on public health. These approaches have been shown to enhance the feasibility. The merits of these approaches lie in their capacity to facilitate the construction of the generalizability of models, unification of predictive algorithms, and the provision of visual interpretations of findings, thereby offering advantages over prevailing statistical and mathematical techniques.

The issue of data analysis in the domain of public health risk assessment in the context of environmental contamination has been the focus of numerous scientific publications [6], [7]. In light of this, an examination of the most common methods and approaches aimed at automating the processes of intellectual problem solving is warranted [8], [9]. An overview of artificial intelligence and machine learning techniques for forecasting the impact of air pollution on health, particularly chronic respiratory disorders, can be found in [10], [11]. Researchers emphasize the high accuracy of hybrid models that incorporate a range of prediction methods. The researchers assessed models using accuracy measures, such as the root mean square error (RMSE) and mean absolute error

(MAE). They found that while these models are beneficial in providing early alerts for health threats, substantial modifications are made to address imbalances in the input datasets, resulting in incomplete models.

The present study [12] is predicated on the utilization of a random forest model to examine the effects of human emissions and climatic conditions on long-term changes in air pollution levels in eastern China. The objective of this study was to investigate the relationship between these two elements. The analysis revealed a substantial decrease in pollution levels coinciding with a decline in various anthropogenic air emissions across multiple regions. The model developed by the Ministry of Environment (MOE) identifies patterns of seasonal oscillations in air pollution, enabling a more precise evaluation of the associated health risks at various pollutant concentrations. The model's accuracy was noteworthy, with a percentage exceeding 86%.

An evaluation of various alternative GO models, including LSTM and Bi-LSTM, was conducted by the authors [13] to forecast the amounts of PM2.5 and CO impurities present in the air. The researchers discovered that the Stacked LSTM model exhibited a superior degree of accuracy for PM2.5, whereas the encoder–decoder LSTM model demonstrated a higher level of accuracy for CO values. A thorough examination of both models was performed. The findings of this study can be utilized in the communication of short-term health risk values, with the prediction horizon ranging from one to three days. Furthermore, the study incorporated the use of classification algorithms (Support Vector Machine [SVM]) to ascertain the association between air pollution levels and illnesses affecting the cardiovascular and pulmonary systems.

Subsequent to the implementation of MO, the authors of another study [14] developed two MS2Quant models for estimating the ionization efficiency and an MS2Tox model for evaluating the toxicity of aquaculture products. The development of both models was predicated on the utilization of the MO. The developed models are applicable for the identification of potentially hazardous chemicals in water based on the analysis of mass spectral data. Furthermore, these models facilitate expeditious identification and classification of pollutants in wastewater, thereby enabling health risk assessment. The authors of [15] conducted research in the Guangzhong Basin area to examine ensemble learning approaches for evaluating groundwater quality. Specifically, they employed LightGBM models in conjunction with uncertainty analysis and the SHAP technique to forecast polluted water quality parameters. These models are capable of considering the impact of both natural and anthropogenic factors, thereby facilitating the identification of significant health concerns. However, the accuracy of these models depends on the values of the input hyperparameters.

In the contemporary scientific milieu, significant emphasis is placed on the implementation of metamaterial optics (MO) and carbon nanostructures (CS) to automate the analysis of ecologically significant data pertinent to population health. This focus is predicated on recognizing this subject as a contemporary and imperative area of research. The objective of this study is to develop an analytical system that utilizes machine learning algorithms to evaluate the prevalence of threats to public health.

2. Methods

The problem under consideration can be conceptualized as the categorization of numerous classes [16]. Multiclass classification, a predictive task in machine learning, entails the model's determination of the observed item's belonging to one of multiple potential classes [17]. This undertaking is formally designated as the "classification task." The mathematical formulation of this problem in the context of public health risk assessment can be described as follows. The representation of the input dataset can be expressed as $X=\{x_1, x_2,\ldots,x_n\}$, where each x_i represents a feature vector originating from the space $\Re^{(d)}$. In the context of our study, the number of distinct classes is six, and each object x_i is associated with a class label y_i that belongs to the set $y_i \in \{1,2,\ldots,K\}$. The total number of classes, denoted by K, is a crucial element of our model.

To address the unique challenge of incorporating visual data along with numerical measurements, our methodology employed a dual-stream approach. The visual component of our dataset comprised atmospheric images captured under varying pollution conditions, which required specialized pre-processing techniques. These images underwent a systematic transformation process beginning with standardization to a uniform resolution of 224×224 pixels, followed by normalization to ensure consistent pixel intensity distributions across the dataset. To extract meaningful features from these visual inputs, we implemented a transfer learning strategy utilizing a pre-trained ResNet-50 architecture, fine-tuned on our specific air pollution visual dataset [15], [18], [19]. The convolutional layers of this network extracted high-level features that captured subtle visual indicators of air quality, including atmospheric opacity, color shifts, and particulate matter visibility patterns.

It is necessary to construct a function $f:R^d \rightarrow \{1,2,\ldots,K\}$ that for any input object x will predict a class label y (population health risk). The MO model is built using a training sample $\{(x1,y1),(x2,y2),\ldots,(xn,yn)\}$ that can be generated from informative input features. The task is to find an approximation of function f based on these data. If $P(y=k|x)$ is the probability that object x belongs to class k, then the function $f(x)$ predicts the class with the maximum posterior probability:

$$f(x) = \arg\max_{k \in \{1,2,\ldots,K\}} P(y = k \qquad (1)$$

In our hybrid approach, the visual features extracted from the CNN are concatenated with numerical air quality measurements to form an enriched feature vector. This fusion of visual and numerical data provides a more comprehensive representation of environmental conditions, allowing our models to leverage both quantitative measurements and qualitative visual cues. The integration process involved careful feature scaling to ensure that neither the visual nor numerical components dominated the learning process. A weighted fusion strategy was employed, where the relative importance of visual versus numerical features was determined through cross-validation experiments, ultimately assigning a 0.4 weight to visual features and 0.6 to numerical measurements based on their respective contributions to prediction accuracy.

A loss function, which measures the discrepancy between the predicted and actual class labels, is used to train the model. One commonly used loss function is cross-entropy.

$$L(y, y') = -\sum_{k=1}^{K} y_k \log\left(y'_k\right) \qquad (2)$$

where *yk* is a binary indicator (0 or 1) indicating whether the object belongs to class *k and y' = P (y=k|x)* is the probability that the object belongs to class *k* predicted by the model. The model is optimized by minimizing the loss function L using optimization techniques such as gradient descent. The final prediction was performed as a maximum-likelihood class selection based on the trained model. The overall pipeline of the system operation for this study is shown in Fig.1.

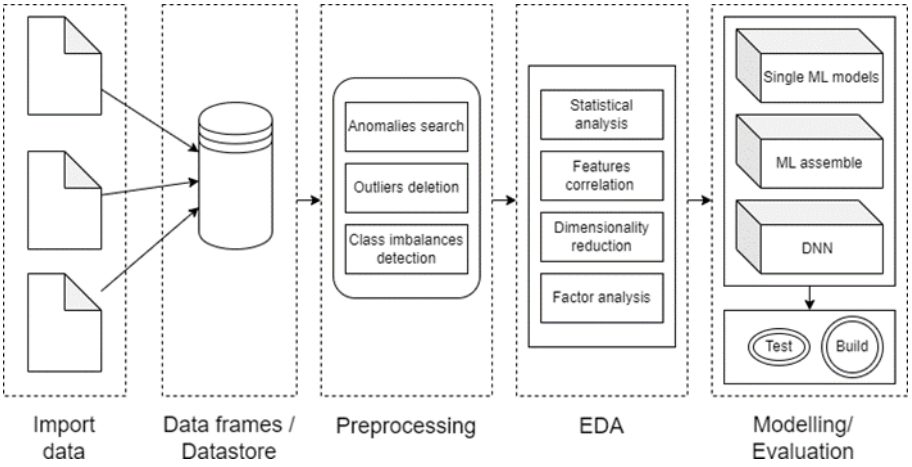

Figure 1. General pipeline of system operation

The imported datasets were saved as dataframe objects by utilizing the functions of the Pandas library. Subsequently, the preprocessing procedure is executed, which involves the identification of anomalies and outliers as well as the elimination of the output class imbalance. Subsequently, a series of exploratory data analysis procedures were carried out, which included statistical, correlation, and factor analyses, as well as optional data dimensionality reduction in the event that there were a large number of input features. Consequently, separate data were created as part of the modeling stage.

Class imbalance is implemented based on the class weighting technique [20], [21], whereby the values of weights are calculated as the inverse of the frequency of a class in the sample. Consequently, this results in an increase in the level of model penalty for classes that are less prevalent in the dataset. The absence of a provision for the fabrication of synthetic data contributes to the enhancement of categorization reliability. The models that were developed underwent a review process based on the metrics chosen to assess their quality and accuracy. Following this evaluation, the models underwent testing, and their final objects were serialized into separate files for subsequent downloading and potential usage on fresh data. This process enables the evaluation of potential hazards to a population.

2.1 Dataset

During the course of the investigation into datasets pertaining to the influence of various anthropogenic variables on public health, it was ascertained that no exhaustive datasets reflecting the diverse elements of environmental contamination are available for gratuous access. A substantial number of datasets pertaining to air mass pollution in various regions worldwide, including India, are available on data-analysis platforms and open repositories. This phenomenon was particularly pronounced. Air Pollution resources from India and Nepal (APD) will serve as the foundational resource for this study [22].

This composite dataset comprised text datasets accompanied by comprehensive descriptions of data by significant attributes, along with images collected in India and Nepal. These images are intended to describe and characterize the level of risk of harm to people caused by the level of air pollution from a variety of harmful substances under a variety of conditions. The incorporation of visual depictions from diverse geographical regions was instrumental in elucidating the regional specificity of the dataset. This dataset complements the comprehensive information presented in tabular format in comma-separated values (CSV) format, which delineates the characteristics of pollution distribution in various locations across India and Nepal.

One of the curious aspects of the dataset was that the photos were taken with varying degrees of contamination. These images can be analyzed using computer vision and MO methods, both of which are separate approaches. The visual dataset comprised 12,000 images captured under standardized conditions, with each image corresponding to a specific set of numerical air quality measurements. These images were collected using high-resolution digital cameras at fixed locations across multiple cities to ensure consistent perspectives and framing. The temporal alignment between visual data collection and numerical measurements was maintained within a 15-minute window to ensure data coherence. Each image was labeled with the corresponding AQI category, creating a supervised learning scenario in which visual features could be mapped to pollution severity levels. The sample size was approximately 12,000 records and the distribution of the target classes was provided in proportions that were approximately equivalent to each other, ranging from 13 to 21 percent. The statistics are also aggregated on the basis of information collected from two separate states, India and Nepal, each with a different socioeconomic and environmental situation. This allows for a comparative study of data from different locations.

This dataset has the capacity to analyze not only data from air composition measurements but also visual pollution signals (which allow for a wider feature space and complicated formalizable elements), which may be beneficial for a more thorough evaluation. It is important to emphasize that this dataset has the potential to study both quantitative and qualitative data. The visual component adds critical contextual information that numerical sensors alone cannot capture, such as the presence of smog, haze density, and visibility reduction, all of which directly correlate with the human perception of air quality and potential health impacts. This multimodal approach allows our models to learn the complex relationships between measurable pollutant concentrations and their visible manifestations in the environment.

This means that the data can be used in conjunction with weather information and public health measures to perform integrated risk assessment. In terms of its organizational structure, the dataset consists of two catalogs: Combined_Dataset and the Country_wise_ dataset catalogs. The following cities in India were included in the dataset: Delhi, Nagaland, Bangalore, Greater Noida, Faridabad, Mumbai, and Tamil Nadu. The collection also included information on the city of Biratnagar, Nepal. There is a CSV format file that stores the input properties of the dataset. These attributes include information about the location, filename (image), date (year, month, day, hour), and air pollution indicators (PM2.5, PM10, O3, CO, SO2, NO2) as well as the target class AQI_Class.

Six different classifications of air pollution were included in the dataset as target attributes. The term "good" refers to the numerical range from 0 to 50, which indicates that the air quality is excellent and that there is little to no risk of pollution in the general population. In this class, air quality is considered acceptable; however, for certain pollutants, there may be moderate health problems for a very small number of people who are unusually sensitive to air pollution. In other words, the risks to the general population were minimal. The numerical range corresponding to this class is 51-100. The term "unhealthy for sensitive groups" corresponds to the number range (101-150).

In this scenario, individuals who are members of sensitive groups may be at a risk of adverse health outcomes. However, the general population is highly unlikely to be at high risk of developing chronic diseases. Therefore, the class can be understood as having low risk. According to this output characteristic category, more than half of the general population may be experiencing health problems and aggravation of illnesses, and members of vulnerable groups may be experiencing major health problems. Unhealthy corresponds to numerical values in the range 150–200. The threat level was medium. Very Unhealthy, which corresponds to the numerical range (201-300), where the risk of adverse health effects that cannot be reversed is significant for all categories. Hazardous/Severe (Severe) is a numerical number corresponding to the range (301-500) and is a feature of urgent and emergency situations, such as accidents, where there is a high probability of irreparable damage to public health. Thus, the level of risk is critical.

3. Results and Discussion

Initially, the libraries were imported for data processing, creation of structures (numpy, pandas collections) to provide the necessary manipulations with input features, data visualization (matplotlib, seaborn), and a number of packages from the sklearn library to perform procedures for converting categorical data (string or text labels) into numerical values, normalization of data, and connection of model and object evaluation metrics for their direct creation (e.g., DecisionTreeClassifier). These libraries were used at the inception of this process. The results of

the correlation analysis of quality are shown in Figure 2. The subsequent phase of the study will entail the execution of the investigative procedures.

Attributes	Location	Filename	Year	Month	Day	Hour	AQI	PM2.5	PM10	O3	CO	SO2	NO2
Location	1	0.976	-0.241	0.205	0.213	0.185	0.737	0.751	0.654	-0.008	-0.026	0.422	0.606
Filename	0.976	1	-0.224	0.222	0.295	0.144	0.699	0.740	0.656	-0.033	0.009	0.451	0.610
Year	-0.241	-0.224	1	-0.976	0.317	0.208	-0.105	-0.034	-0.084	0.249	0.065	0.075	0.104
Month	0.205	0.222	-0.976	1	-0.340	-0.263	0.045	-0.007	0.069	-0.260	-0.008	-0.057	-0.111
Day	0.213	0.295	0.317	-0.340	1	0.142	0.046	0.278	0.124	0.092	0.162	0.371	0.267
Hour	0.185	0.144	0.208	-0.263	0.142	1	0.256	0.139	0.096	0.582	-0.390	-0.191	0.147
AQI	0.737	0.699	-0.105	0.045	0.046	0.256	1	0.806	0.664	0.054	-0.216	0.224	0.487
PM2.5	0.751	0.740	-0.034	-0.007	0.278	0.139	0.806	1	0.813	0.035	-0.062	0.281	0.709
PM10	0.654	0.656	-0.084	0.069	0.124	0.096	0.664	0.813	1	0.137	-0.059	0.169	0.571
O3	-0.008	-0.033	0.249	-0.260	0.092	0.582	0.054	0.035	0.137	1	-0.349	-0.314	0.106
CO	-0.026	0.009	0.065	-0.008	0.162	-0.390	-0.216	-0.062	-0.059	-0.349	1	0.398	-0.007
SO2	0.422	0.451	0.075	-0.057	0.371	-0.191	0.224	0.281	0.169	-0.314	0.398	1	0.332
NO2	0.606	0.610	0.104	-0.111	0.267	0.147	0.487	0.709	0.571	0.106	-0.007	0.332	1

Figure 2. Correlation assessment table for dataset input characteristics

It is imperative to acknowledge that, with the exception of the non-informative attributes of the image file name, high correlation values are characteristic of the attributes that characterize air pollution caused by harmful impurities (PM2.5, PM10). This phenomenon can be attributed to the inherent characteristics of these impurities and the proximity of the measurement equipment. The decision to exclude the Filename property from the dataframe object was informed by its perceived lack of informative value. This decision was made within the context of the preliminary analysis, data cleaning, and preparation for the job under consideration. In the context of data exploration, statistical descriptions of the input characteristics were acquired by leveraging the Pandas description () function. This function outputs the quantity, mean, standard deviation, and range of data, as illustrated in Figure 3.

```
# Describe data
dataframe.describe()
```

	Year	Month	Day	AQI	PM2.5	PM10	O3	CO	SO2	NO2
count	10281.000000	10281.000000	10281.000000	10281.000000	10281.000000	10281.000000	10031.000000	9807.000000	9034.000000	9920.000000
mean	2022.948254	2.683980	12.116526	167.517848	142.942723	145.403790	39.393336	101.412380	13.305071	37.899775
std	0.221524	1.764717	8.277427	102.798851	130.398412	103.952927	33.371867	116.346153	9.876013	39.627358
min	2022.000000	2.000000	1.000000	15.000000	4.000000	7.000000	1.000000	0.000000	2.000000	0.670000
25%	2023.000000	2.000000	3.000000	97.000000	35.000000	64.000000	12.000000	4.000000	4.400000	7.000000
50%	2023.000000	2.000000	13.000000	152.000000	70.080000	113.000000	31.000000	52.000000	10.000000	20.000000
75%	2023.000000	3.000000	20.000000	230.000000	257.000000	198.000000	59.660000	174.000000	20.000000	64.000000
max	2023.000000	10.000000	28.000000	450.000000	500.000000	480.000000	225.000000	410.000000	57.000000	169.000000

Figure 3. Result of the assessment of statistical metrics

In the context of data preparation, a conversion method known as label encoding has been developed. This procedure involves converting categorical data into numerical values using LabelEncoder class. The rationale behind this procedure is to ensure compatibility of the data with mathematical operations and models. The subsequent analysis revealed a notable correlation between the signs month and year, which led to the exclusion of these variables from the final dataset. During the analysis of gaps using the isnull() method, it was discovered that there were more than 2000 gaps in the O3, CO, SO2, and NO2 attributes. To address this, missing values were filled in by calculating and substituting the mean values using the mean() function. This approach ensures a comprehensive resolution of the missing values.

3.1 Model Development and Research

To develop the MO models, the Python programming language, as well as the sklearn, matplotlib, seaborn, keras, and tensorflow libraries [12], [13] were utilized. These libraries served as the foundation for the formation of separate Jupiter Notebook modules. Within each module, the procedures entailed the importation of program dependencies (libraries), incorporation of input data (training and test samples), creation of advisory models, and evaluation of their effectiveness based on the metrics described above. Finally, the models were serialized into pickle object files. The implementation of MO models entailed the utilization of the following models: decision tree (DT), support vector machine (SVM), random forest (RF), XGBoost, and deep artificial neural network (convolutional CNN and recurrent LSTM).

Seventy-five percent of the data sample was allocated for training purposes, while the remaining 25% was designated for the evaluation of machine learning (ML) models. The data samples were meticulously partitioned into distinct training and test subsets. To achieve the objective of distinguish the values of the metrics into their own logs, it was determined that the storage of these values would be implemented in the variables that correspond to them. To conduct a summary analysis of the results of the model evaluations and determine the accuracy of the

solution to the classification problem, visualization in the form of an error matrix was utilized through the utilization of seaborn. The results of constructing such matrices for all models of the MN are shown in Figure 4. To enhance the comprehensibility of the results, the output risk classes were converted into a numerical range from 0 to 5, in a specific order.

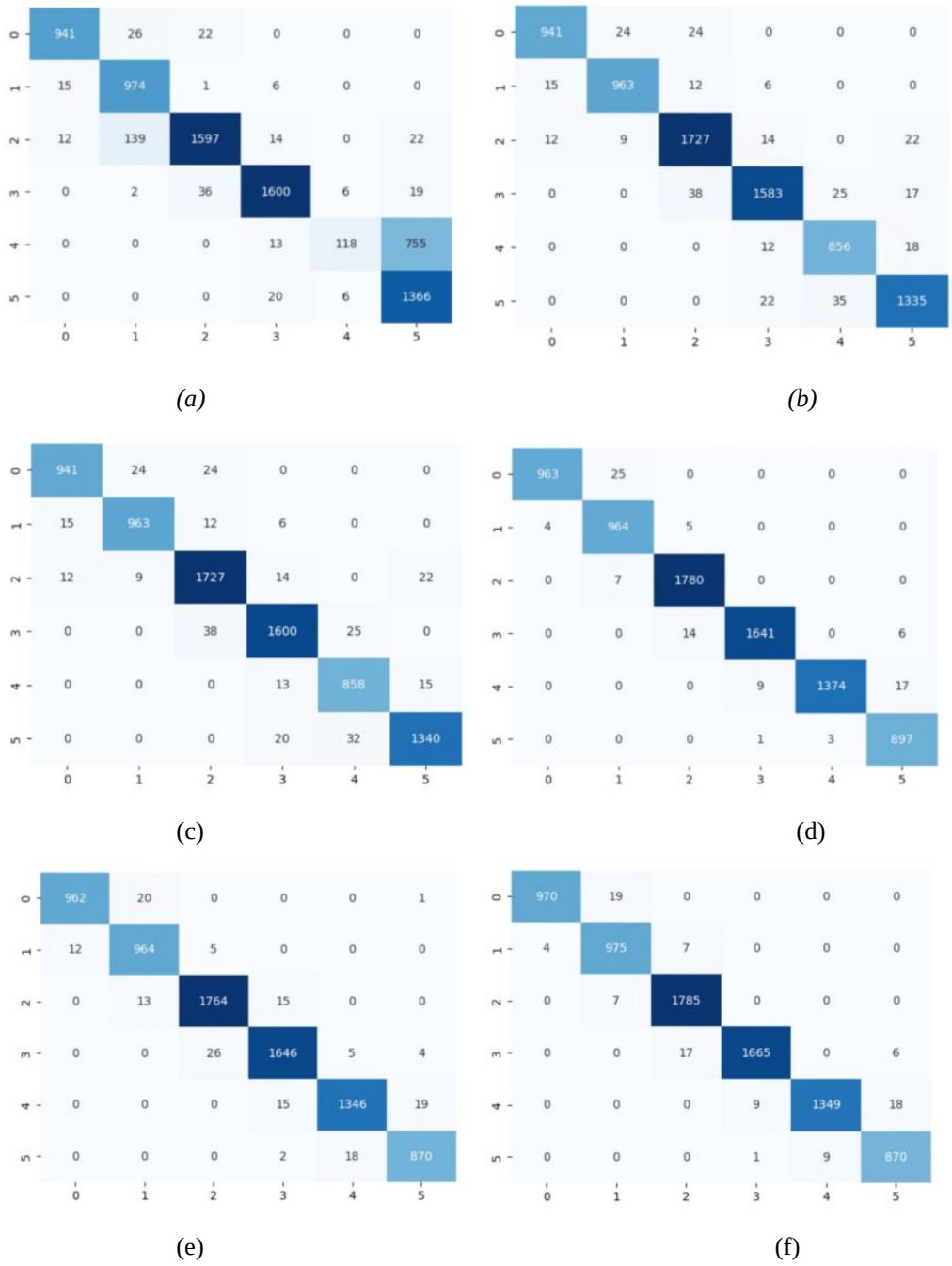

Figure 4. Error matrices of decision tree (*a*), SVM (*b*), random forest (*c*), XGBoost (*d*), recurrent (*e*) and convolution (*f*) ANN models

The findings indicated that the models exhibited a high level of accuracy, with XGBoost demonstrating the greatest success in terms of classification accuracy. To facilitate a comprehensive comparison of the classification accuracy across all models in multiclass form, a visualization was developed. This approach enables a more profound examination of the models through visualization. Figure 5 shows the mean dependencies on the ROC curves of the models presented as detached visualizations. The nature of these curves varies across different ranges, as can be observed. The estimates of ensemble models (random forest and XGBoost) are the ones that are closest to the ideal values (smoothed and closer to 1), suggesting that these values are closest to the ideal.

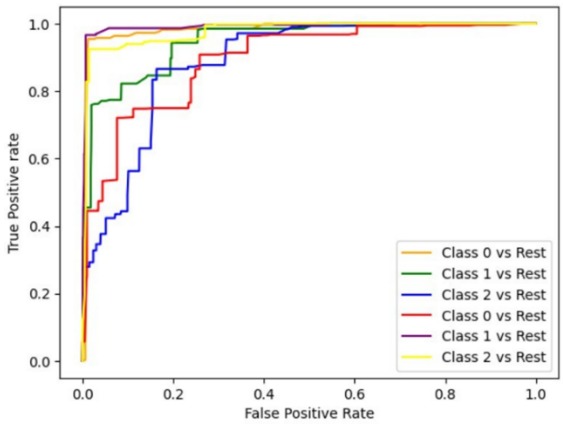

Figure 5. Averaged dependencies of models on ROC curves

To conduct an in-depth investigation of the nature of the training process for artificial neural networks (ANN) models, graphical dependencies were constructed that included estimates of accuracy and loss (see Figure 6). As shown in these figures, the ANN models demonstrated a high degree of accuracy. Specifically, the LSTM model attains accuracy values of approximately 0.98 by the 15th epoch, surpassing the CNN model's accuracy of 0.98 at the 25th epoch. Beyond this point, the growth rate slows. Periodic fluctuations in performance indicative of the risk of overtraining were observed. However, the selected regularization values effectively mitigated this adverse effect. A similar trend was observed in the CNN model, where the initial error values were higher than those of the LSTM model; however, the learning rate of the convolutional model was considerably faster than that of the LSTM model.

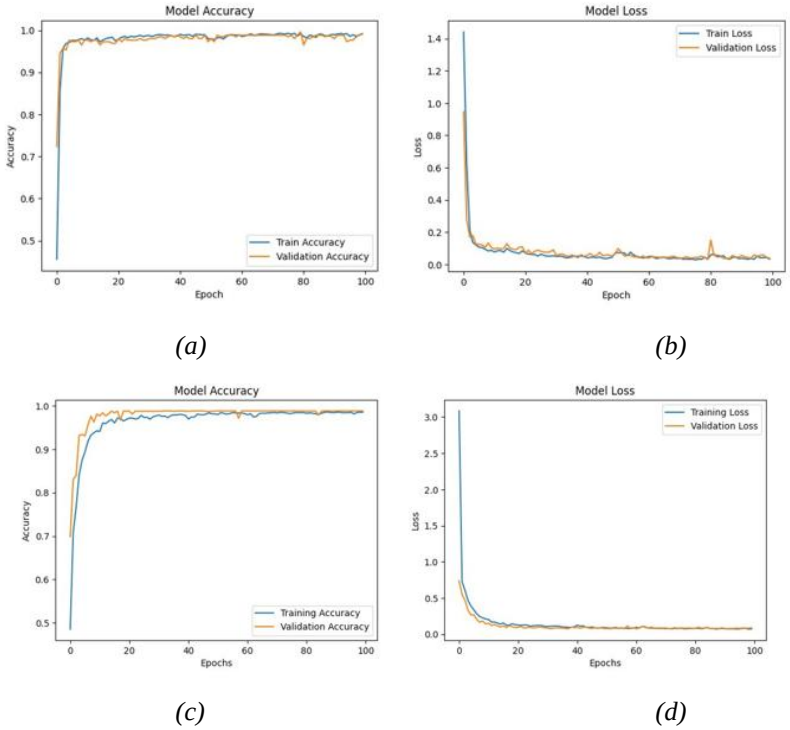

Figure 6. Dependencies of accuracy and loss values on training epochs of recurrent ANN (a, b) and convolutional ANN (c, d)

The outcomes of the comparison study of the metrics of the developed MO models are presented in Figure 7. A histogram of feature importance evaluation of dataset characteristics was generated to conduct additional analysis of the findings obtained from the use of models, as shown in Figure 9.

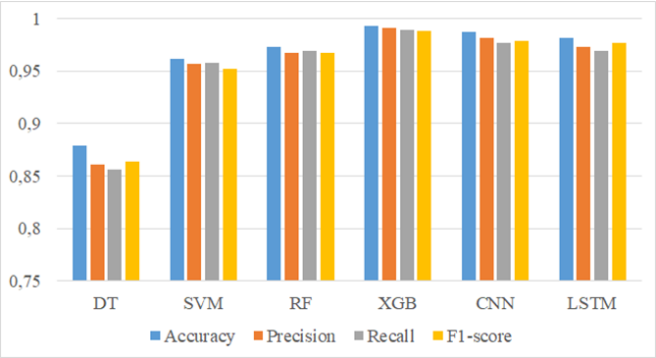

Figure 7. Histogram comparing metrics of MOE models

The feature significance evaluation allows for the identification of the characteristics with the greatest impact on the model's predictions. This approach enables refinement of model quality by eliminating uninformative or redundant characteristics, thereby enhancing the interpretability of the model's predictions. Given the pronounced relevance of AQI, PM2.5, and PM10, it is imperative to prioritize their consideration during the development and subsequent optimization of models. It is noteworthy that the decision tree model is the fastest, but also the least accurate. Conversely, the support vector model is the most accurate model, but its time cost is five to six times greater than that of the decision tree model, and two times more than that of ensemble-based models.

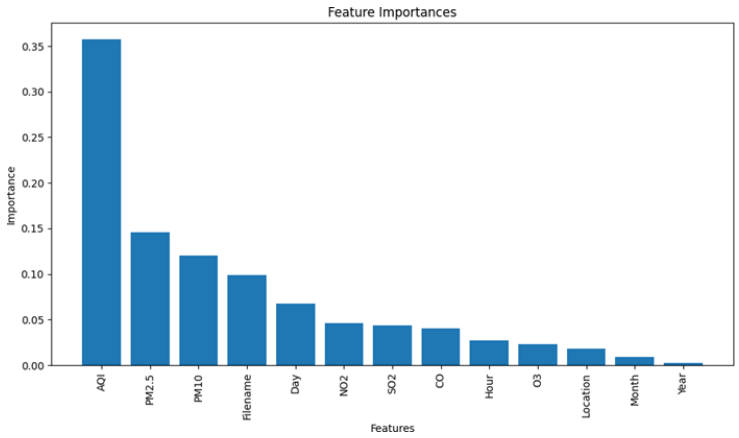

Figure 8. Histogram of estimation of significance of dataset features

ANN models have been shown to be accurate; however, they require a greater amount of resources for training owing to the complexity of their structure. This complexity includes nested (hidden) layers and a high number of neural connections. The LSTM model was more resource-intensive for both models. Conversely, ensemble-based models demonstrate superior performance in terms of accuracy and processing speed. Among these, the XGBoost model is noteworthy for its positive performance.

The implications of our findings extend far beyond the immediate realm of air quality monitoring, weaving a tapestry of potential applications spanning multiple domains of public health and environmental management. As we navigate the increasingly complex landscape of urban environmental challenges, our multimodal approach opens new vistas for integrated health-surveillance systems that transcend traditional boundaries. The fusion of visual and numerical data analysis creates a powerful paradigm that can revolutionize how we conceptualize and respond to environmental health threats.

Consider, for instance, the potential integration of our methodology with a smart city infrastructure. Urban planners and policymakers can leverage these models to create dynamic, real-time health risk maps that adapt to changing environmental conditions. These maps can inform everything from daily commute recommendations to long-term urban development strategies, creating a feedback loop between environmental monitoring and urban design. The visual component of our approach is particularly valuable, as it enables the detection of localized pollution hotspots that might escape traditional sensor networks, such as illegal burning activities or unauthorized industrial emissions.

In the realm of public health emergency response, our models offer a sophisticated early warning system that can fundamentally transform how communities prepare for and respond to environmental health crises. During events, such as forest fires or industrial accidents, the rapid assessment capabilities of our system could provide crucial minutes or hours of additional warning time, enabling more effective evacuation procedures and medical

resource allocation. The visual analysis component adds a critical layer of spatial intelligence, allowing emergency responders to visualize the spread of pollutants and predict their trajectories with unprecedented accuracy.

The educational application of this research presents another fascinating frontier. By translating complex air quality data into visually intuitive risk assessments, our models can serve as powerful tools for environmental education and community engagement. Schools could utilize simplified versions of our system to teach students about environmental science and data literacy, whereas community organizations could employ it to advocate for environmental justice in historically underserved areas. The visual nature of our approach makes it particularly accessible to diverse audiences, thus bridging the gap between scientific complexity and public understanding.

Perhaps most intriguingly, our methodology holds promise for cross-disciplinary applications in fields as diverse as epidemiology, climate science, and urban ecology. Epidemiologists can integrate our air quality risk assessments with disease surveillance data to uncover subtle patterns in respiratory illness outbreaks. Climate scientists might adapt our visual analysis techniques to study the relationship between air pollution and local climate phenomena, whereas urban ecologists could use our models to investigate the impact of air quality on urban biodiversity.

4. Conclusions

The findings of this study indicate that the utilization of diverse MO and CS models to address the classification issue and assess the risks posed by air pollution to public health is a commendable approach. In general, the constructed models are highly accurate, with over 90% accuracy. However, a discrepancy was observed between the training and testing speeds of these models. Ensemble models such as Random Forest and XGBoost exhibit the most robust correlation between accuracy and performance. The limitations of the system are evident in several respects. Specifically, model training and tuning are constrained to the sequential mode, CUDA architecture is not supported, and input data are limited to text files, excluding interactive user interfaces. It is imperative to select the appropriate models and hyperparameter values for each dataset. One potential approach to address this challenge is to utilize optimization methods such as the grid search approach, which could be explored in future research in this area.

References

[1] K. Roell *et al.*, "Development of the InTelligence And Machine LEarning (TAME) Toolkit for Introductory Data Science, Chemical-Biological Analyses, Predictive Modeling, and Database Mining for Environmental Health Research," *Frontiers in Toxicology*, vol. 4, Jun. 2022, doi: 10.3389/ftox.2022.893924.

[2] S. Mistry, N. O. Riches, R. Gouripeddi, and J. C. Facelli, "Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review," *Artif Intell Med*, vol. 135, p. 102461, Jan. 2023, doi: 10.1016/j.artmed.2022.102461.

[3] R. An, J. Shen, and Y. Xiao, "Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies," *J Med Internet Res*, vol. 24, no. 12, p. e40589, Dec. 2022, doi: 10.2196/40589.

[4] S. Cui *et al.*, "Advances and applications of machine learning and deep learning in environmental ecology and health," *Environmental Pollution*, vol. 335, p. 122358, Oct. 2023, doi: 10.1016/j.envpol.2023.122358.

[5] R. C. Bernardes, L. L. Botina, F. P. da Silva, K. M. Fernandes, M. A. P. Lima, and G. F. Martins, "Toxicological assessment of agrochemicals on bees using machine learning tools," *J Hazard Mater*, vol. 424, p. 127344, Feb. 2022, doi: 10.1016/j.jhazmat.2021.127344.

[6] J. B. Neris, D. M. M. Olivares, F. G. Velasco, F. H. M. Luzardo, L. O. Correia, and L. N. González, "HHRISK: A code for assessment of human health risk due to environmental chemical pollution," *Ecotoxicol Environ Saf*, vol. 170, pp. 538–547, Apr. 2019, doi: 10.1016/j.ecoenv.2018.12.017.

[7] P. Gao, G. Huang, L. Zhao, and S. Ma, "Identification of biological indicators for human exposure toxicology in smart cities based on public health data and deep learning," *Front Public Health*, vol. 12, May 2024, doi: 10.3389/fpubh.2024.1361901.

[8] E. A. Cohen Hubal *et al.*, "Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment," *Journal of Toxicology and Environmental Health, Part B*, vol. 13, no. 2–4, pp. 299–313, Jun. 2010, doi: 10.1080/10937404.2010.483947.

[9] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean Technol Environ Policy*, vol. 21, no. 6, pp. 1341–1352, Aug. 2019, doi: 10.1007/s10098-019-01709-w.

[10] G. Polezer *et al.*, "Assessing the impact of PM2.5 on respiratory disease using artificial neural networks," *Environmental Pollution*, vol. 235, pp. 394–403, Apr. 2018, doi: 10.1016/j.envpol.2017.12.111.

[11] N. Temirbekov, M. Temirbekova, D. Tamabay, S. Kasenov, S. Askarov, and Z. Tukenova, "Assessment of the Negative Impact of Urban Air Pollution on Population Health Using Machine Learning Method," *Int J Environ Res Public Health*, vol. 20, no. 18, p. 6770, Sep. 2023, doi: 10.3390/ijerph20186770.

[12] M. Tao, L. Wang, L. Chen, Z. Wang, and J. Tao, "Reversal of Aerosol Properties in Eastern China with Rapid Decline of Anthropogenic Emissions," *Remote Sens (Basel)*, vol. 12, no. 3, p. 523, Feb. 2020, doi: 10.3390/rs12030523.
[13] X. Cheng, W. Zhang, A. Wenzel, and J. Chen, "Stacked ResNet-LSTM and CORAL model for multi-site air quality prediction," *Neural Comput Appl*, vol. 34, no. 16, pp. 13849–13866, Aug. 2022, doi: 10.1007/s00521-022-07175-8.
[14] F. Melnikov, J. Kostal, A. Voutchkova-Kostal, J. B. Zimmerman, and P. T. Anastas, "Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals," *Green Chemistry*, vol. 18, no. 16, pp. 4432–4445, 2016, doi: 10.1039/C6GC00720A.
[15] K. Chen, Q. Liu, T. Yang, Q. Ju, and M. Zhu, "Risk assessment of nitrate groundwater contamination using GIS-based machine learning methods: A case study in the northern Anhui plain, China," *J Contam Hydrol*, vol. 261, p. 104300, Feb. 2024, doi: 10.1016/j.jconhyd.2024.104300.
[16] A. Fernández, V. López, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowl Based Syst*, vol. 42, pp. 97–110, Apr. 2013, doi: 10.1016/j.knosys.2013.01.018.
[17] S. Wang, H. Lu, A. Khan, F. Hajati, M. Khushi, and S. Uddin, "A machine learning software tool for multiclass classification[Formula presented]," *Software Impacts*, vol. 13, 2022, doi: 10.1016/j.simpa.2022.100383.
[18] T. Berliani, E. Rahardja, and L. Septiana, "Perbandingan Kemampuan Klasifikasi Citra X-ray Paru-paru menggunakan Transfer Learning ResNet-50 dan VGG-16," *Journal of Medicine and Health*, vol. 5, no. 2, 2023, doi: 10.28932/jmh.v5i2.6116.
[19] Indriani, "Applying Transfer Learning ResNet-50 for Tracking and Classification of A Coral Reef in Development The Mobile Application with Scrum Framework," *Journal of Information Technology*, vol. 4, no. 2, 2023, doi: 10.47292/joint.v4i2.90.
[20] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach Learn*, vol. 113, no. 7, 2024, doi: 10.1007/s10994-022-06268-8.
[21] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, no. 3, 2015.
[22] S. K. Guttikunda, K. A. Nishadh, and P. Jawahar, "Air pollution knowledge assessments (APnA) for 20 Indian cities," *Urban Clim*, vol. 27, 2019, doi: 10.1016/j.uclim.2018.11.005.