

Intelligent decision support system for assessing and analyzing population health risks

Nickolay Rudnichenko^{1*}, Denys Shvedov¹, Natalia Shibaeva¹ and Igor Petrov²

¹ Odesa Polytechnic National University, Shevchenko Avenue 1, Odesa, 65001, Ukraine

³ National University "Odesa Maritime Academy", Didrichson street 8, Odesa, 65029, Ukraine

Abstract

The paper presents the development of an intelligent decision support system for assessing and analyzing population health risks associated with air pollution using a hybrid machine learning and deep learning framework. The study is motivated by the growing impact of atmospheric contaminants, including PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and O₃, on public health, especially in densely populated urban areas. The necessity of applying intelligent data-driven approaches for accurate, scalable, and interpretable environmental risk assessment is substantiated. A review of existing studies is carried out to identify the limitations of traditional statistical methods and to demonstrate the potential of neural networks and ensemble algorithms for capturing nonlinear relationships and spatiotemporal patterns in environmental datasets. The proposed system is based on a six-stage analytical workflow incorporating decision trees, support vector machines, random forests, XGBoost, convolutional neural networks, and long short-term memory networks. These models are integrated into a unified process covering data preparation, classification, forecasting, and risk-level determination. Model training and validation are performed using the Air Pollution Image Dataset collected in India and Nepal, supplemented with structured air quality data. The software solution follows a client-server architecture with a Python and Flask backend and a JavaScript-based frontend. UML diagrams are used to describe system components and their interactions, while the analytical dashboard provides tools for data visualization, prediction generation, and comparative model analysis. Experimental evaluation demonstrates the applicability of the proposed approach, with ensemble and deep learning techniques delivering the most accurate results. Comparative model assessment is conducted using accuracy, precision, recall, F1-score, and average training time metrics.

Keywords

decision support system, health risk assessment, machine learning; deep learning, data mining, predictive analytics

1. Introduction

In recent years, environmental conditions have become an increasingly important determinant of population health. Rapid urbanization, industrial development, intensive transportation, and other forms of anthropogenic activity have contributed to the deterioration of atmospheric air quality, especially in densely populated regions. Long-term exposure to pollutants such as particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃), and carbon monoxide (CO) is associated with a broad spectrum of adverse health outcomes, including respiratory and cardiovascular diseases, complications during pregnancy, impaired neurodevelopment, and premature mortality. The increasing complexity of environmental health interactions creates a need for intelligent tools that can transform heterogeneous observations into reliable population health risk assessments. Public health institutions and environmental authorities require not only continuous monitoring of pollutant concentrations but also analytical mechanisms for identifying

ADT-ISMDDC 2026: *Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control*, September 22–24, 2026, Odesa, Ukraine.

^{1*} Corresponding author.

✉ nickolay.rud@gmail.com (N. Rudnichenko); studylearner@gmail.com (D. Shvedov); nati.sh@gmail.com (N. Shibaeva), firmn@gmail.com (I. Petrov);

ORCID: 0000-0002-7343-8076 (N. Rudnichenko); 0009-0002-4823-8782 (D. Shvedov); 0000-0002-7869-9953 (N. Shibaeva), 0000-0002-8740-6198 (I. Petrov);



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dangerous trends, vulnerable population groups, and territories with elevated health risks [1]. However, there remains a considerable gap between the results of epidemiological studies and their practical implementation in intelligent decision support systems. Existing approaches often lack the ability to integrate environmental and health-related information, process rapidly changing data, account for spatial and temporal dependencies, and provide decision-oriented interpretations of identified risks [2]. A significant challenge is the limited availability of integrated and standardized datasets combining environmental measurements with population health characteristics. Environmental monitoring systems, medical databases, demographic statistics, and epidemiological surveys are usually maintained independently and differ in structure, temporal frequency, spatial resolution, and accessibility. In addition, real-world datasets frequently contain missing observations, inconsistent measurements, temporal gaps, and incomplete geographic coverage. These limitations can reduce the reliability and generalization capability of analytical and predictive models used for population health risk assessment [3]. Data integration also requires substantial preprocessing because information may originate from fixed monitoring stations, satellite observations, mobile sensors, healthcare institutions, public surveys, and other sources. Such data can differ considerably in format, scale, accuracy, and sampling frequency. Their effective combination requires data cleaning, normalization, feature construction, missing-value treatment, and alignment of spatial and temporal characteristics. The use of individual-level medical information additionally introduces requirements related to privacy protection, anonymization, and responsible data management [4]. Despite these challenges, intelligent decision support technologies provide considerable opportunities for improving environmental health risk management. Statistical analysis, machine learning, anomaly detection, and predictive methods can be combined to identify complex relationships between environmental exposure and population health indicators, estimate current risk levels, and detect potentially hazardous changes at an early stage. Such capabilities can support evidence-based preventive measures, resource allocation, healthcare planning, and targeted interventions. Therefore, intelligent analytical platforms can become an important component of smart and sustainable public health infrastructure [5]. In this context, the present research focuses on the development of an intelligent decision support system for assessing and analyzing population health risks associated with air pollution. The proposed system is intended to integrate heterogeneous environmental and health-related data, identify significant patterns and dependencies, evaluate population risk levels, and provide analytical information to support informed public health decisions.

2. Analysis existing researches and approaches

Given the complexity of environmental determinants of population health, intelligent decision support systems have become an important direction for transforming heterogeneous data into actionable risk information. Recent advances in machine learning (ML) and deep learning (DL) [6] provide effective tools for modeling complex relationships between environmental exposure and health-related outcomes. These methods can identify nonlinear dependencies, process heterogeneous data sources, automatically select informative features, and adapt to changing environmental and population conditions [7]. In comparison with conventional statistical approaches, ML methods are less dependent on predefined relationships between variables and can reveal latent patterns within high-dimensional datasets. A wide range of architectures, including deep neural networks, convolutional and recurrent models, attention mechanisms, and ensemble learning approaches [8], has demonstrated considerable potential for environmental and population health analysis. Their applications include pollutant concentration forecasting, exposure assessment, disease incidence prediction, and evaluation of environmental risk dynamics. The integration of such models with geospatial, temporal, demographic, and socioeconomic information makes it possible to obtain more detailed representations of population vulnerability.

Nevertheless, the transition from experimental ML applications [9] to operational intelligent decision support systems remains challenging. Many existing studies rely on relatively small, synthetic, or highly specialized datasets and therefore require additional validation before practical deployment. Reliable implementation also depends on adaptation to regional characteristics, compatibility with existing health information systems, and consideration of local public health policies. Furthermore, cooperation between data scientists, epidemiologists, environmental specialists, and public health professionals is required to ensure that computational models produce meaningful and ethically justified results. Issues of model interpretability, fairness, uncertainty, and potential bias are particularly important when analytical results can influence health-related resource allocation and risk communication. For example, the study [10] investigates high-accuracy classification of air pollution levels using ensemble ML algorithms and neural network architectures. XGBoost and LSTM demonstrated particularly strong performance, achieving AQI prediction accuracy above 90% in the reported experiments.

The research presented in [11] provides a closer connection between environmental exposure and health outcomes by analyzing the relationship between $PM_{2.5}$ and PM_{10} concentrations and outpatient visits for respiratory diseases. The use of a multilayer perceptron demonstrates the possibility of predicting health-related indicators for specific age groups, which is relevant to differentiated population risk assessment. The use of real medical observations and seasonal and demographic characteristics strengthens the practical value of the approach. However, the relatively simple model architecture and restricted geographic coverage limit its applicability to larger-scale intelligent decision support. These limitations can be addressed through more advanced architectures, including LSTM and attention mechanisms, together with broader spatial coverage and additional environmental and demographic indicators. The hybrid DL methodology described in [12] combines spatial convolutional neural networks, temporal recurrent components, and attention mechanisms to forecast pollutant concentrations. Its ability to process incomplete observations and represent spatiotemporal dependencies is particularly relevant to intelligent environmental monitoring. Nevertheless, the resulting model remains focused on forecasting pollution rather than determining its consequences for population health.

Another relevant approach is presented in [13], where U-Net neural networks are used as a surrogate model for rapid assessment of health impacts associated with short-term $PM_{2.5}$ exposure. By replacing computationally expensive atmospheric simulations with a conditional DL model, the authors achieve substantially faster processing and demonstrate the feasibility of scalable assessment. However, the proposed methodology primarily reproduces the outputs of a physical simulation model and does not constitute a complete health-risk assessment pipeline. In particular, its integration with epidemiological, demographic, or clinical information remains limited. The GIS-based methodology proposed in [14] represents another important direction by combining satellite-derived $PM_{2.5}$ measurements, spatial interpolation, and exposure–response functions to estimate population exposure in urban areas with insufficient ground-based monitoring. The major advantage of this approach is its ability to provide detailed spatial representations of pollution exposure. However, limited consideration of temporal variability and the absence of advanced ML-based predictive mechanisms restrict its potential for dynamic decision support [15].

Overall, existing research demonstrates the increasing role of ML and DL in environmental data preprocessing, feature extraction, classification, forecasting, and risk modeling. However, these capabilities are frequently implemented as separate analytical solutions rather than as components of an integrated intelligent decision support environment. A more comprehensive approach requires the joint processing of environmental observations and health outcome indicators, followed by risk estimation, interpretation, and presentation of results in a form suitable for decision-makers. The objective of the proposed research is therefore to develop a scalable and interpretable framework

that combines these capabilities and supports not only accurate population health risk prediction but also the identification of factors and conditions requiring targeted preventive interventions [16].

The scientific novelty of the research consists in developing an integrated end-to-end approach for intelligent population health risk assessment that combines heterogeneous environmental and health-related information within a unified analytical framework. Particular attention is given to the problems of data incompleteness, heterogeneity, spatial and temporal variability, and model interpretability. The practical significance of the proposed system is associated with its potential application in public health monitoring, environmental risk management, and urban planning. By combining predictive analytics with interpretable decision-support mechanisms, the proposed approach provides a technological basis for more timely identification of population health risks and evidence-based selection of appropriate preventive measures.

3. System concept, implementation and technical aspects

The proposed intelligent framework combines several ML and DL techniques into a sequential decision-support architecture, where each model addresses a specific analytical component of population health risk assessment [17]. Such a layered organization allows the system to combine interpretable statistical learning with more complex predictive models, improving both the accuracy of risk estimation and the transparency of the resulting decisions. It also provides flexibility for processing heterogeneous environmental and population data under different spatial, temporal, and contextual conditions [18].

The first processing layer performs initial data screening and risk-oriented categorization using a decision tree (DT). The main advantage of DT at this stage is its transparent hierarchical structure, which makes it possible to identify the environmental indicators and threshold conditions most strongly associated with changes in population health risk.

The second layer applies a support vector machine (SVM) to investigate more complex relationships between environmental exposure and medical-demographic characteristics. By operating in a multidimensional feature space, SVM can identify nonlinear boundaries between observations associated with different levels of health risk.

The third layer employs a random forest (RF) ensemble to consolidate and refine the information obtained from previous processing stages. By constructing and aggregating multiple decision trees based on different subsets of observations and features, RF reduces the influence of individual observations and random noise.

The fourth stage incorporates XGBoost to further improve predictive performance. In contrast to the preceding models, this algorithm sequentially corrects prediction errors generated by previous trees, allowing complex dependencies between environmental, temporal, spatial, and socioeconomic variables to be represented more effectively.

The fifth stage introduces a convolutional neural network (CNN) for extracting spatial and contextual characteristics from image-based or geospatial information. This component is particularly important when pollution observations are represented through environmental images, spatial layers, or distributed sensor structures. CNN can identify local spatial patterns and correlations that may be difficult to represent through conventional tabular models, thereby supporting the detection of areas with potentially elevated environmental and health risks.

The final analytical component is based on a long short-term memory (LSTM) model, which is responsible for representing temporal dependencies in environmental exposure and health-related indicators. Its ability to retain information from previous observations makes LSTM suitable for

analyzing prolonged exposure effects and identifying temporal patterns that may not be apparent from individual measurements.

The combination of these models creates a multidimensional decision-support pipeline capable of integrating complementary analytical capabilities rather than relying on a single predictive algorithm. Interpretable ML models provide an initial explanation of risk factors, ensemble methods improve the robustness of classification, while CNN and LSTM extend the analysis toward spatial and temporal dimensions.

The primary dataset selected for the study is the Air Pollution Image Dataset from India and Nepal [19]. To improve the reliability and generalization of the analytical results, the image dataset was supplemented and cross-validated using structured environmental measurements obtained from two authoritative sources [20, 21].

Within the proposed research, population health risk assessment is considered not as a conventional classification problem solved by a single ML algorithm, but as a hybrid decision-making process that combines several complementary analytical mechanisms. The central idea is to transform heterogeneous environmental and population-related observations into an integrated representation of health risk through successive stages of classification, refinement, spatial analysis, and temporal interpretation.

Each observation represents a specific environmental and population state and may contain information about pollutant concentrations, demographic characteristics, geographic conditions, temporal parameters, and other factors associated with population health. The available observations are divided into six predefined risk categories, representing different levels of potential health impact. The task of the intelligent system is therefore to determine the most appropriate risk category for a new observation while preserving information about the factors and analytical stages that contributed to the final decision.

Unlike a conventional approach in which the prediction of a single ML model is directly treated as the final result, the proposed hybrid methodology introduces a sequential refinement mechanism. At the initial level, DT is used to establish an interpretable preliminary categorization of observations and identify the most influential environmental conditions. The resulting information is subsequently analyzed by SVM to distinguish more complex combinations of risk factors. RF then performs ensemble-based refinement of the obtained classification, reducing the sensitivity of the assessment to individual observations and measurement noise. XGBoost further enhances the decision process by learning residual patterns and incorporating additional contextual characteristics.

The hybrid architecture is extended through the integration of CNN and LSTM, which introduce spatial and temporal dimensions into the assessment process. CNN extracts relevant spatial characteristics from pollution-related images or geographic representations, while LSTM analyzes the evolution of environmental and health-related conditions over time. Their outputs are used as additional evidence for the final risk assessment rather than as independent predictions. Consequently, the system combines structural, statistical, spatial, and temporal information before producing the final population health risk category.

The proposed novelty therefore consists in the organization of ML and DL models as cooperating analytical components of a unified decision-support mechanism. Each model contributes a specific type of knowledge and compensates for limitations of the other approaches. Interpretable models provide an initial explanation of risk conditions, ensemble algorithms increase classification stability, while CNN and LSTM extend the assessment toward spatial and temporal dependencies.

The training process is consequently oriented toward minimizing not only classification errors but also inconsistencies between the different analytical stages. The quality of the final system is evaluated by comparing its predicted risk categories with reference observations and by analyzing the stability of predictions across different data conditions. Particular attention is given to the ability of the hybrid system to correctly distinguish neighboring risk categories, maintain reliable performance when individual indicators are incomplete or noisy, and preserve an interpretable connection between input characteristics and the resulting population health risk assessment.

The model is optimized by minimizing the loss function L using optimization techniques such as gradient descent. The final prediction is made by selecting the class with the highest probability based on the trained model. The overall system pipeline used for conducting the study is presented in Fig. 1.

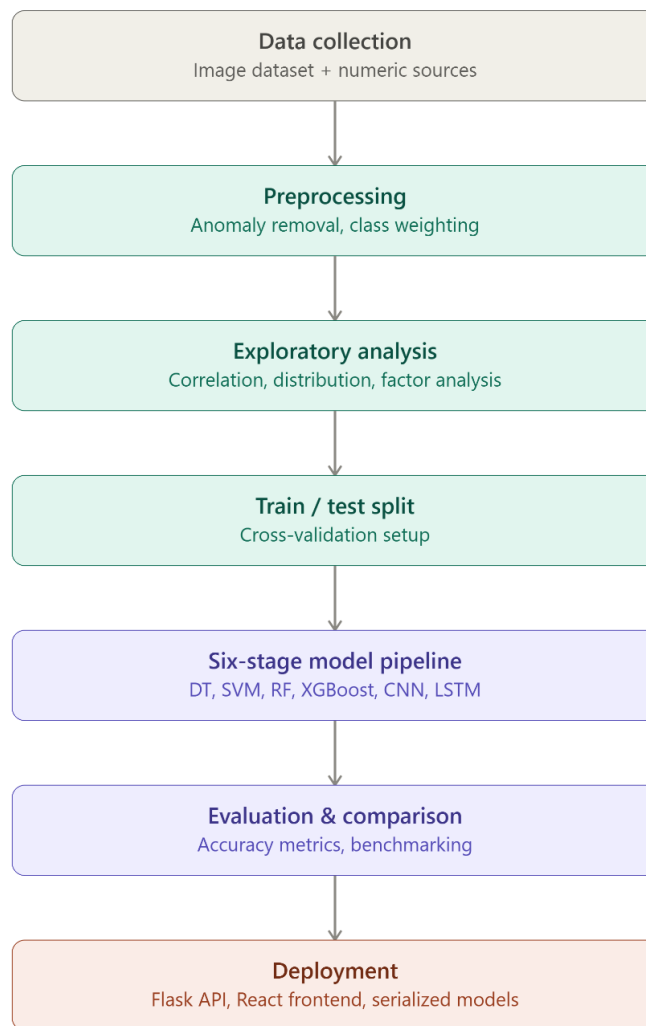


Figure 1: General concept pipeline

The imported datasets are loaded using Pandas and represented as DataFrame objects. The preprocessing stage includes data quality assessment, anomaly and outlier detection, missing-value treatment, feature transformation, and correction of class imbalance. EDA is then performed using statistical, correlation, and factor analysis, with dimensionality reduction applied when required. At the modeling stage, DT, SVM, RF, XGBoost, CNN, and LSTM components are trained on

appropriately prepared training and testing subsets, with cross-validation used to assess the stability of the obtained results.

Class imbalance is handled through class weighting based on the frequency of risk categories, increasing the influence of underrepresented classes without generating synthetic observations. This preserves the original data structure and reduces the possibility of introducing artificial patterns. The individual models are evaluated using selected performance metrics and then integrated according to their roles in the hybrid pipeline. Their outputs provide complementary evidence for the subsequent risk assessment rather than independent final predictions.

After training, the resulting models and preprocessing components can be serialized and reused for new datasets. This enables the developed system to process previously unseen environmental and population information and produce an updated assessment of public health risk. The implementation uses NumPy and Pandas for data processing, Matplotlib and Seaborn for visualization, and relevant modules of the scikit-learn (sklearn) library for encoding, normalization, model construction, cross-validation, and performance evaluation.

In the components diagram (Fig.2) we can see that the proposed system is organized as a three-component architecture that connects the user interface, analytical processing, and data management into a unified decision-support workflow. The Frontend component, implemented using JavaScript, provides the user-facing dashboard and interactive map-based visualization of environmental conditions and population health risks. It receives analytical results and presents them in an accessible form for monitoring and decision-making. The Python + Flask component serves as the central analytical layer of the system. It provides the REST API, performs data preprocessing, executes the integrated ML/DL models, and conducts population health risk analysis. This component acts as an intermediary between the user interface and the underlying data resources, transforming incoming information into structured analytical results. The Data Layer component provides storage and retrieval of environmental observations, structured datasets, and generated model results using SQL and CSV-based resources. It supplies the analytical layer with the required input data and stores processed results for subsequent visualization and analysis.

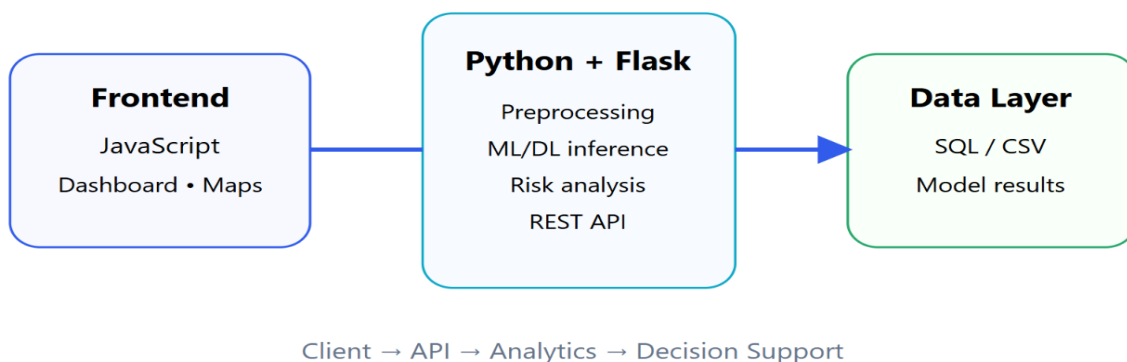


Figure 2: Base components scheme

The use-case model defines the main interaction between the Health analyst and the intelligent decision support system. The analyst can upload air quality data, which initiates data processing and preparation for subsequent analysis. The Request risk assessment use case applies the hybrid ML/DL pipeline to the available environmental and population-related data and generates an assessment of the corresponding health risk level. The View analytical dashboard use case provides access to processed results through visual representations, allowing the analyst to examine risk levels, environmental indicators, spatial patterns, and changes in population health risk. The Compare model performance use case supports evaluation of the individual and hybrid analytical components, helping determine their reliability and contribution to the overall assessment. Together, these use

cases define the main decision-support cycle from data acquisition and risk analysis to result interpretation and model evaluation, supporting evidence-based assessment of population health risks associated with air pollution. Base use case diagram is shown in Fig.3.

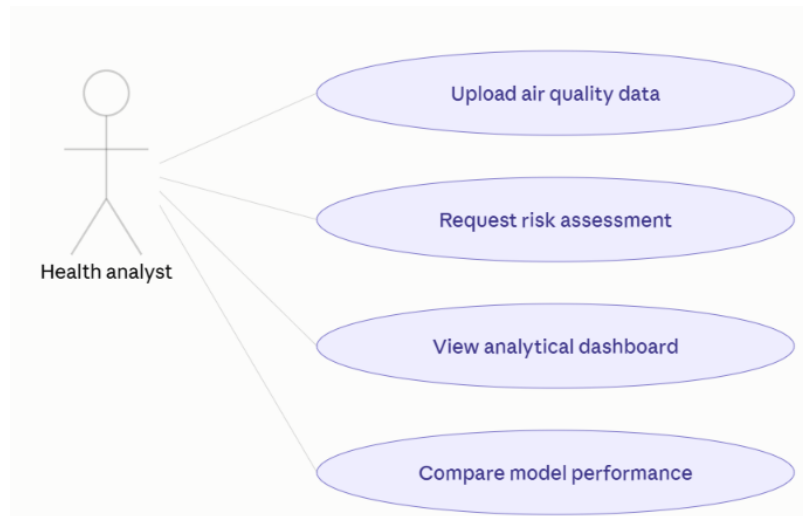


Figure 3: Base use case diagram

4. Experiments and results analysis

The locally deployed application provides an interactive environment through which users can access the principal functions of the proposed decision support system. The interface integrates data submission, analytical processing, predictive modeling, and risk interpretation within a coherent workflow. Its design is intended to make the principal computational stages accessible while preserving the relationship between the underlying environmental data and the resulting analytical outputs. The data acquisition section presents the requirements for the input dataset and provides an interactive control labeled “Upload file.” After a file has been selected, the interface displays its name and relevant metadata, enabling the user to verify the correctness of the uploaded resource. Following this verification procedure, the user can select “Confirm data” to initiate the subsequent analytical stage.

Once the dataset has been confirmed, it is transferred to the backend processing service. The exploratory data analysis module processes the uploaded information and generates statistical and graphical outputs associated with the characteristics of the environmental observations. These outputs may include pollutant distribution plots, correlation heatmaps, descriptive summaries, anomaly visualizations, and dimensionality reduction representations. The analytical results are subsequently transmitted to the frontend and displayed for interpretation by the user. The interface also presents the six-stage analytical pipeline and the risk decision logic, enabling users to examine the relationship between data acquisition, preprocessing, exploratory analysis, model development, model evaluation, and decision support. This visual representation contributes to the transparency of the system by providing an overview of the procedures through which environmental observations are transformed into predictive and risk-related outputs.

. The predictive modeling service is responsible for managing the trained machine learning and deep learning models. Conventional machine learning estimators may be serialized in the .pkl format, while neural network models may be stored using the .keras format. The serialized models are maintained within the server-side file system and are loaded when a corresponding prediction request is received. During inference, the backend identifies the appropriate model, retrieves the stored model representation, and applies the preprocessing operations required for the submitted input data. The processed input is then supplied to the selected model, which generates the

corresponding prediction or forecasting result. The output is returned to the frontend as a structured JSON response containing the relevant predictive information and, where applicable, the associated risk classification. This backend configuration separates the analytical and predictive services from the presentation layer, thereby supporting the modular organization of the application. It also provides a foundation for the integration of additional analytical procedures and predictive architectures as the system evolves. The proposed system combines quantitative environmental analysis with an interactive decision support interface, allowing users to examine pollutant distributions, investigate relationships among environmental variables, evaluate alternative predictive models, and interpret air-quality-related risk categories. The resulting architecture is intended to support environmental health monitoring, the identification of potentially hazardous pollution conditions, and the communication of analytical information to relevant stakeholders. System’s user’s prototype interface fragment is shown in Fig.4.

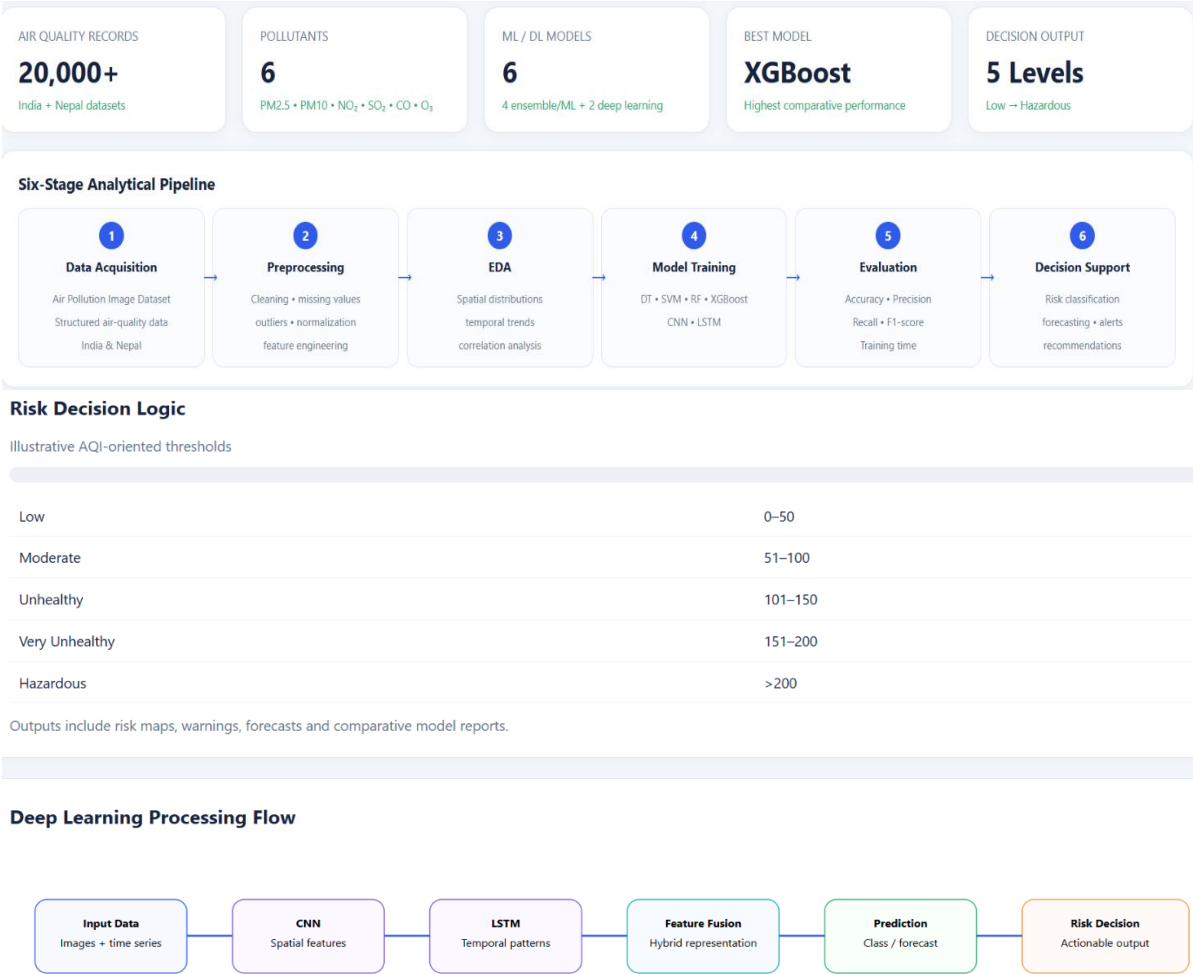


Figure 4: System’s user’s prototype interface fragment

To initiate the forecasting procedure, the user selects the required model and task type, either classification or regression. The “Start Prediction” button then becomes available, allowing the user to submit a request to the backend and execute the corresponding prediction pipeline. Upon completion, the system returns the generated visualizations and a detailed CSV report to the frontend. For machine learning model evaluation, the dataset was divided into training and testing subsets using a 75:25 ratio, respectively. Comparative assessment of classification performance was supported by confusion matrix visualizations generated using the Seaborn library. The resulting confusion matrices for some ML models are presented in Fig. 5.

SVM						Random forest						...					
935	32	24	0	0	0	958	18	15	0	0	0	958	18	15	0	0	0
18	958	12	6	0	0	9	972	7	6	0	0	9	972	7	6	0	0
14	150	1580	18	0	22	9	8	1752	11	0	4	9	8	1752	11	0	4
0	3	42	1585	9	24	0	0	19	1633	7	4	0	0	19	1633	7	4
0	0	0	15	105	745	0	0	0	9	835	42	0	0	0	9	835	42
0	0	0	24	9	1358	0	0	0	10	9	1373	0	0	0	10	9	1373

GRU						MLP					
967	17	6	0	0	0	928	40	23	0	0	0
9	970	8	0	0	0	22	948	18	6	0	0
0	9	1762	13	0	0	18	160	1560	20	0	26
0	0	20	1655	5	3	0	4	50	1560	12	28
0	0	0	11	1352	18	0	0	0	20	95	700
0	0	0	4	14	1878	0	0	0	28	12	1330

Figure 5: The confusion matrices for the implemented models

The comparative evaluation demonstrates substantial differences in the predictive performance of the developed models. As illustrated in Figure 6, all classifiers achieve ROC curves substantially above the random-classification baseline, indicating effective discrimination between air-quality risk classes. XGBoost exhibits the highest area under the ROC curve (AUC = 0.998), followed by CNN (0.995), LSTM (0.992), Random Forest (0.990), SVM (0.980), and Decision Tree (0.910). The near-unity AUC values of the ensemble and deep learning models indicate strong classification capability, whereas the comparatively lower AUC of the Decision Tree suggests limited generalization performance.

The results presented in Table 1 confirm the superiority of XGBoost, which achieves the highest accuracy (0.993), precision (0.991), recall (0.989), and F1-score (0.988). Random Forest demonstrates the strongest performance among the remaining conventional machine learning models, with an accuracy of 0.968 and an F1-score of 0.962. CNN and LSTM also achieve high predictive performance, reaching accuracies of 0.987 and 0.981, respectively. SVM provides moderate performance, whereas Decision Tree records the lowest values across all evaluation metrics, with an accuracy of 0.912 and an F1-score of 0.906.

A consistent trend is observed across the evaluation metrics, as accuracy, precision, recall, and F1-score increase in a broadly similar order. The relatively small differences between these measures suggest a balanced classification performance, although the slightly lower recall values indicate that some risk categories remain more difficult to identify correctly. The computational results reveal a trade-off between predictive effectiveness and training time. Decision Tree requires approximately 3 seconds for training, compared with approximately 15 seconds for XGBoost, 45 seconds for CNN, and 67 seconds for LSTM. Despite its higher computational cost than the conventional tree-based models, XGBoost provides the most favorable combination of predictive accuracy and computational efficiency among the evaluated high-performing models. These findings support its selection as the principal predictive model for integration into the proposed intelligent decision support system.

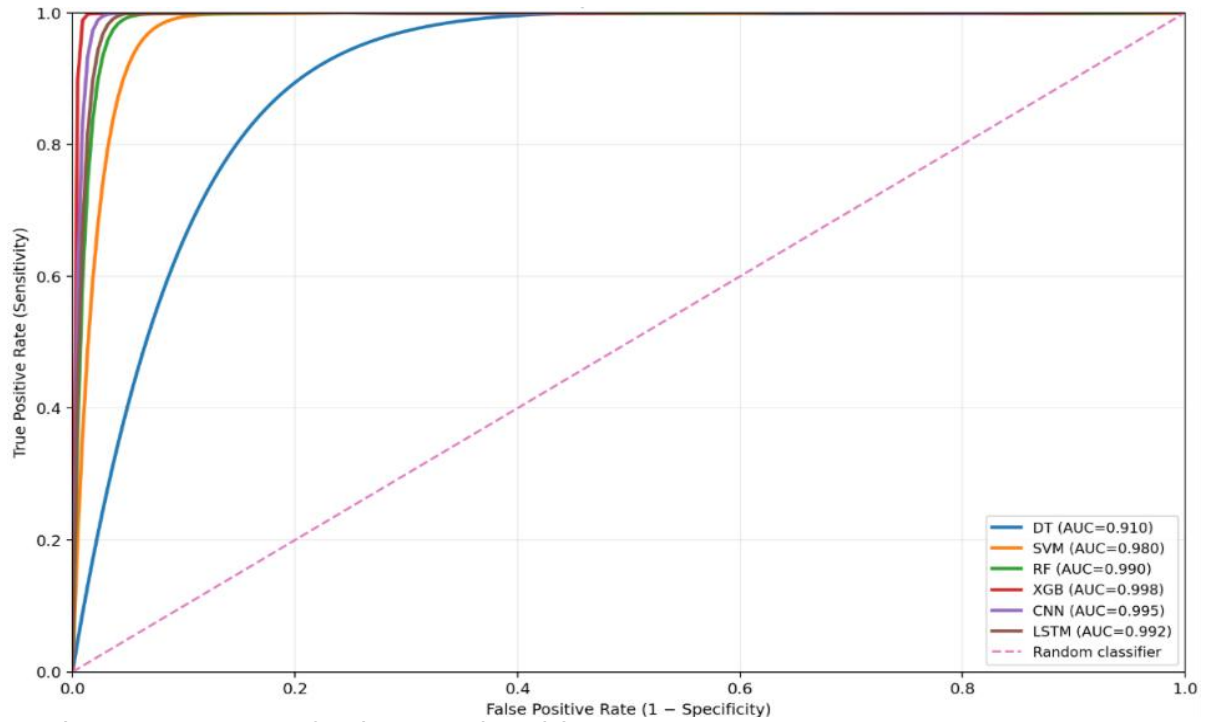


Figure 6: ROC curve for the created models

Table 1

Created models metrics comparative analysis results

Model	Accuracy	Precision	Recall	F1-score	Avg. training time
DT	0.912	0.908	0.905	0.906	~3 s
SVM	0.934	0.929	0.926	0.927	~12 s
RF	0.968	0.964	0.961	0.962	~7 s
XGBoost	0.993	0.991	0.989	0.988	~15 s
CNN	0.987	0.984	0.982	0.983	~45 s
LSTM	0.981	0.978	0.975	0.976	~67 s

Overall, the experimental results demonstrate that the developed models provide a high level of accuracy in classifying population health risks associated with air pollution. Among the evaluated architectures, XGBoost achieved the strongest overall performance, with an accuracy of 0.993, precision of 0.991, recall of 0.989, F1-score of 0.988, and an AUC of 0.998. CNN and LSTM also demonstrated high classification capability, reaching accuracies of 0.987 and 0.981, respectively, while their AUC values of 0.995 and 0.992 confirm strong class discrimination. These results indicate that both ensemble and deep learning approaches are capable of capturing complex nonlinear relationships between pollutant concentrations and population health risk categories.

The experimental comparison also demonstrates a clear relationship between predictive performance and computational requirements. XGBoost required approximately 15 s for training, whereas CNN and LSTM required approximately 45 s and 67 s, respectively. This indicates that the higher computational complexity of the neural network architectures does not result in a corresponding improvement over XGBoost under the considered experimental conditions. The

introduction of the average training time indicator $\backslash(t_{avg})\backslash$ therefore provides an additional criterion for assessing model applicability in an operational decision support environment.

The scientific novelty of the proposed approach lies in the integration of heterogeneous air-quality indicators with machine learning and deep learning models within a unified population health risk decision support framework. In particular, the joint consideration of $PM_{2.5}$, PM_{10} , NO_2 , SO_2 , CO , and O_3 enables the system to capture complex nonlinear interactions among multiple pollution factors rather than relying on individual pollutant indicators. The combination of predictive modeling with a five-level risk classification mechanism further transforms numerical model outputs into interpretable population health risk assessments. The obtained results demonstrate that XGBoost provides the most effective balance between predictive accuracy, discriminative capability, and computational efficiency, supporting its selection as the principal predictive component of the proposed system.

5. Conclusion

The developed intelligent decision support system provides an integrated framework for the analysis and assessment of population health risks associated with air pollution. The experimental results demonstrate that the considered models provide reliable classification performance, with the proposed XGBoost model achieving the most favorable overall results and an AUC of 0.998. The obtained results confirm the suitability of the framework for distinguishing different levels of air-quality-related risk.

The main contribution of the study is the integration of exploratory analysis, statistical assessment, predictive modeling, and risk classification within a unified decision support environment. The combination of conventional machine learning and deep learning approaches enables the system to account for both the multidimensional structure of pollutant data and temporal characteristics of environmental observations. The inclusion of computational efficiency in model comparison additionally supports the selection of models suitable for practical deployment.

The developed interface and backend architecture provide a basis for transforming environmental measurements into interpretable risk-oriented outputs, including predictions, forecasts, and warning information. This makes the system applicable to environmental monitoring and population health assessment. Further development may focus on real-time sensor integration, extended temporal forecasting, and scalable deployment to support continuous air-quality risk monitoring.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] F. H. Dominski, J. H. L. Branco, G. Buonanno, L. Stabile, M. G. da Silva, A. Andrade, Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses, *Environmental Research* 201 (2021) 111487. DOI:10.1016/j.envres.2021.111487.
- [2] G.-P. Bălă, R.-M. Răjnoveanu, E. Tudorache, R. Motișan, C. Oancea, Air pollution exposure—the (in)visible risk factor for respiratory diseases, *Environmental Science and Pollution Research* 28 (2021) 19615–19628. DOI:10.1007/s11356-021-13208-x.
- [3] S. Al-Janabi, A. Alkaim, E. Al-Janabi, A. Aljeboree, M. Mustafa, Intelligent forecaster of concentrations ($PM_{2.5}$, PM_{10} , NO_2 , CO , O_3 , SO_2) caused air pollution (IFCsAP), *Neural Computing and Applications* 33 (2021) 14199–14229. DOI:10.1007/s00521-021-06067-7.

- [4] R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, F. Li, Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering, *Expert Systems with Applications* 169 (2021) 114513. DOI:10.1016/j.eswa.2020.114513.
- [5] A. Aggarwal, D. Toshniwal, A hybrid deep learning framework for urban air quality forecasting, *Journal of Cleaner Production* 329 (2021) 129660. DOI:10.1016/j.jclepro.2021.129660.
- [6] A. Gilik, A. S. Ogrenci, A. Ozmen, Air quality prediction using CNN+LSTM-based hybrid deep learning architecture, *Environmental Science and Pollution Research* 29 (2022) 11920–11938. DOI:10.1007/s11356-021-16227-w.
- [7] S.-T. Zang, Q.-J. Wu, X.-Y. Li, C. Gao, Y.-S. Liu, Y.-T. Jiang, J.-Y. Zhang, H. Sun, Q. Chang, Y.-H. Zhao, Long-term PM_{2.5} exposure and various health outcomes: An umbrella review of systematic reviews and meta-analyses of observational studies, *Science of the Total Environment* 812 (2022) 152381. DOI:10.1016/j.scitotenv.2021.152381.
- [8] Z. Zhang, B. Xu, W. Xu, F. Wang, J. Gao, Y. Li, M. Li, Y. Feng, G. Shi, Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on PM_{2.5} pollution, *Environmental Research* 212 (2022) 113322. DOI:10.1016/j.envres.2022.113322.
- [9] K. Ravindra, S. S. Bahadur, V. Katoch, S. Bhardwaj, M. Kaur-Sidhu, M. Gupta, S. Mor, Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections, *Science of the Total Environment* 858 (2023) 159509. DOI:10.1016/j.scitotenv.2022.159509.
- [10] M. Méndez, M. G. Merayo, M. Núñez, Machine learning algorithms to forecast air quality: a survey, *Artificial Intelligence Review* 56 (2023) 10031–10066. DOI:10.1007/s10462-023-10424-4.
- [11] T. M. T. Lei, S. C. W. Ng, S. W. I. Siu, Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau, *Sustainability* 15 (2023) 5341. DOI:10.3390/su15065341.
- [12] Z. Zhang, S. Zhang, Modeling air quality PM_{2.5} forecasting using deep sparse attention-based transformer networks, *International Journal of Environmental Science and Technology* 20 (2023) 13535–13550. DOI:10.1007/s13762-023-04900-1.
- [13] Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, Y. Zheng, R. Zimmermann, AirFormer: Predicting Nationwide Air Quality in China with Transformers, *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023) 14329–14337. DOI:10.1609/aaai.v37i12.26676.
- [14] M. Yu, S. Zhang, K. Zhang, J. Yin, M. Varela, J. Miao, Developing high-resolution PM_{2.5} exposure models by integrating low-cost sensors, automated machine learning, and big human mobility data, *Frontiers in Environmental Science* 11 (2023) 1223160. DOI:10.3389/fenvs.2023.1223160.
- [15] I. Jairi, S. Ben Othman, L. Canivet, H. Zgaya, Explainable based approach for the air quality classification on the granular computing rule extraction technique, *Engineering Applications of Artificial Intelligence* 133 (2024) 108096. DOI:10.1016/j.engappai.2024.108096.
- [16] M. J. Jiménez-Navarro, M. Lovrić, S. Kecorius, E. K. Nyarko, M. Martínez-Ballesteros, Explainable deep learning on multi-target time series forecasting: An air pollution use case, *Results in Engineering* 24 (2024) 103290. DOI:10.1016/j.rineng.2024.103290.

- [17] Ayesha, M. Noor-ul-Amin, O. Albalawi, N. Mushtaq, E. E. Mahmoud, U. Yasmeen, M. Nabi, Modeling health outcomes of air pollution in the Middle East by using support vector machines and neural networks, *Scientific Reports* 14 (2024) 21517. DOI:10.1038/s41598-024-71694-8.
- [18] S. Ahn, A. Kim, Y. Chung, C. Kang, S. Kim, D. Kwon, J. Park, J. Oh, J. Park, J. Moon, I. Song, J. Min, H. J. Lee, H. Kim, W. Lee, Nationwide Machine Learning-Ensemble $PM_{2.5}$ Mapping Prediction and Forecasting Models in South Korea with High Spatiotemporal Resolution and Health Risk Estimation-Based Evaluations, *Environment & Health* 3 (2025) 878–887. DOI:10.1021/envhealth.4c00201.
- [19] Air Pollution Image Dataset from India and Nepal. URL: <https://www.kaggle.com/datasets/adarshrouniyar/air-pollution-image-dataset-from-india-and-nepal>
- [20] Air Quality System (AQS). URL: <https://www.epa.gov/aqs>
- [21] Air Pollution Data for Seoul. URL: https://figshare.com/articles/dataset/CO_csv/12805643