Electronics and information technologies, 2025, 31, 89–104
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC 004.89

PREDICTING QUANTITATIVE CHARACTERISTICS OF AIR POLLUTION

***Volodymyr Hura*[1]*, *Igor Olenych*[1], *Oleh Sinkevych*[1],**
***Oksana Ostrovska*[2], *Roman Shuvar*[2]**
[1] Radioelectronic and Computer Systems Department,
[2] Department of System Design,
Ivan Franko National University of Lviv,
50 Dragomanov Str., 79005 Lviv, Ukraine

Hura, V., Olenych, I., Sinkevych, O., Ostrovska, O., Shuvar, R. (2025). Predicting quantitative characteristics of air pollution. Electronics and Information Technologies, 31, 89–104 https://doi.org/10.30970/eli.31.8

ABSTRACT

Background. Rapid industrialization and urbanization have escalated air pollution, posing significant health and environmental threats. Accurate prediction of quantitative air pollution characteristics (like pollutant concentrations or Air Quality Index) is critical for effective monitoring and mitigation strategies. Fuzzy Logic (FL) provides a robust computational intelligence framework adept at handling the inherent uncertainty, imprecision, and non-linear dynamics present in atmospheric systems.

Materials and Methods. The study explores the application of Fuzzy Logic (FL) for improving the prediction of hourly PM2.5 concentrations by adding new input features to data obtained using localized monitoring data from Variazh, for 2024. A key aspect involves feature engineering, where a secondary Fuzzy Inference System (FIS) was developed to derive Pasquill atmospheric stability class based on measured meteorological inputs (wind speed, solar radiation, cloud cover). This derived stability class was then incorporated as an additional input feature into the primary Mamdani-type FIS designed for PM2.5 prediction correction.

Results and Discussion. The inclusion of the fuzzy-derived atmospheric stability class as an input feature improved the performance of the PM2.5 prediction models tested (XGBoost, LightGBM). Models incorporating this engineered feature achieved high accuracy ($R^2 > 0.98$), particularly showing enhanced capability during stable atmospheric conditions. This highlights the value of incorporating physically relevant, engineered features derived via interpretable methods like FIS into data-driven air quality models.

Conclusion. Fuzzy Logic proves to be a valuable tool for effective feature engineering in air pollution modeling. Deriving parameters such as atmospheric stability class via an interpretable, rule-based FIS can enrich datasets and enhance the accuracy of subsequent predictive models, offering a practical approach to improving air quality forecasting, especially when direct measurements of complex parameters are unavailable.

Keywords: air pollution, fuzzy logic, forecasting, pollutant concentration, fuzzy inference system, quantitative prediction, machine learning.

INTRODUCTION

Air pollution is one of the most important environmental problems of the 21st century, mainly driven by rapid global industrialization, urbanization, and increased transportation activities [1]. Among the various harmful pollutants, fine particulate matter (PM2.5) is of particular concern. Based on combustion sources (vehicles, power plants, residential heating), industrial processes, and secondary formation of precursor gases (such as SO_2, NOx, and VOCs), exposure to PM2.5 poses a serious risk to human health [2, 3]. In addition to its direct health effects, PM2.5 negatively impacts ecosystems, reduces visibility, and contributes to climate change [4].

Consequently, the ability to accurately predict PM2.5 concentrations and understand its dispersion patterns is crucial. Timely and reliable forecasts of quantitative characteristics, particularly PM2.5 levels and the resulting Air Quality Index (AQI), are essential for implementing effective mitigation strategies, issuing targeted public health warnings, informing policy decisions regarding emission controls, and assessing the impact of environmental regulations [5]. However, predicting PM2.5 concentrations is inherently complex. Its levels are governed by intricate atmospheric processes, including direct emissions, chemical transformations in the atmosphere (secondary aerosol formation), and transport, all significantly influenced by highly variable meteorological conditions (like wind speed, temperature, humidity, boundary layer height), fluctuating emission sources, and geographical factors [6]. These factors interact in strongly non-linear ways, and the available measurement and emissions data often contain significant uncertainty and imprecision, making traditional linear or deterministic modeling approaches particularly challenging for PM2.5 [7].

Conventional statistical and deterministic models often struggle to adequately capture the vagueness, complex non-linearities, and stochastic nature inherent in PM2.5 dynamics [8]. These models frequently require precise mathematical formulations of system behavior, which are often difficult, if not impossible, to derive accurately for complex, open environmental systems like the atmosphere. This limitation highlights the need for alternative modeling techniques capable of handling such complexities effectively. Computational intelligence methods, particularly Fuzzy Logic (FL), have emerged as powerful tools for environmental modeling precisely because they excel at managing uncertainty, linguistic ambiguity, and complex input-output relationships without requiring a complete, explicit mathematical understanding of the underlying physical processes [9]. Unlike traditional Boolean (crisp) logic, where something is either true or false, fuzzy logic, based on fuzzy set theory introduced by Lotfi Zadeh [10], allows for degrees of membership. This means a variable can belong partially to different qualitative sets simultaneously (e.g., a specific temperature might be considered 70% 'warm' and 30% 'hot'). FL utilizes linguistic variables (e.g., 'low' wind speed, 'high' traffic density, 'stable' atmosphere) defined by membership functions, and employs a system of IF-THEN fuzzy rules to map input conditions to output predictions. This structure allows for reasoning with imprecise information, mirroring human-like decision-making under uncertainty. Its proven success in diverse fields requiring the management of imprecise information, such as industrial control systems and complex decision support frameworks, further motivates its application to environmental challenges [9]. This characteristic makes it particularly well-suited for modeling complex systems like PM2.5 pollution, where precise mathematical descriptions are difficult to formulate, data is incomplete or noisy, and key relationships are inherently fuzzy or non-linear.

MATERIALS AND METHODS

Study Design and Data Collection

The primary objective was to improve the efficiency of air pollution forecasting by machine learning models by supplementing the feature vector with an additional feature. To determine the additional feature was the design of a Fuzzy Inference System (FIS) to evaluate atmospheric stability based on localized meteorological measurements. Unlike traditional computational models, this approach leverages fuzzy logic to capture and model complex, non-linear relationships under conditions of uncertainty, utilizing site-specific data from the Variazh (**Fig. 1**). The study sought to validate the FIS by assessing its performance in classifying atmospheric stability across various local weather conditions recorded in 2024.

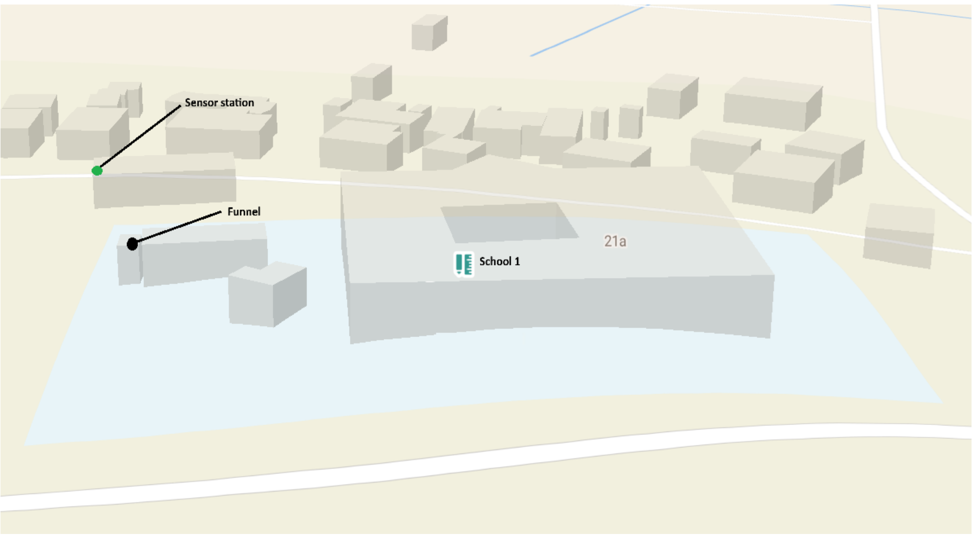


Fig. 1. Air quality information for one station.

Achieving the study's objective involved a methodology with several integrated stages. Initially, data collection and preprocessing were performed using a comprehensive hourly dataset from a dedicated monitoring station in Variazh, covering the entirety of 2024. This dataset comprised essential meteorological variables (wind speed, solar radiation, cloud cover) and temporal factors (hour, day, month), which were processed to remove anomalies and missing values, thereby ensuring data integrity. Meteorological parameters and temporal proxies were selected for their direct influence on atmospheric stability. These predictors, combined with historical site-specific data, provided the foundation for defining the fuzzy system's input variables. The development process involved the selection and definition of input variables, the formulation of appropriate linguistic terms, and the creation of membership functions tailored to the local conditions observed in Variazh.

This step evaluated the model’s ability to generalize across diverse atmospheric conditions. The results analyzed the constructed fuzzy rules to evaluate how input variables such as wind speed, solar radiation, and cloud cover influenced stability classes. This provided insights into localized meteorological dynamics and their impact on atmospheric stability. These results explored the potential of the developed FIS as a practical tool for assessing atmospheric stability. The localized focus and integration of high-resolution site-specific data emphasized the model’s applicability for operational air quality management.

The data utilized a retrospective analysis approach, leveraging the complete hourly dataset for 2024 to train and evaluate the FIS across a full annual cycle.

The geographically focused scope, centered on the Variazh, benefited from the availability of high-quality, site-specific data recorded at a dedicated monitoring station. This detailed focus enabled the development of localized ML models tailored to the unique climatic and environmental characteristics of the site.

Air Quality and Emission Data

The primary air quality parameter targeted for predicting the hourly average concentration of fine particulate matter (PM2.5 μg/m^3). Standard quality assurance and control procedures were assumed to be applied during data collection, and the preprocessing steps outlined addressed potential gaps or anomalies [12].

Hourly meteorological data, collected concurrently with the air quality parameters, were obtained from the monitoring station located in the Variazh (**Fig. 1**). The specific meteorological variables measured and utilized as inputs for the prediction models are described in **Table 1**, and the correlation matrix is shown in **Fig. 2**.

Table 1. **Air quality information for one station for the year 2024**

Characteristic	count	mean	min	25%	50%	75%	max	std
pm10	8784	11.2	1.7	6.7	9.2	13.3	78.4	7.5
pm2_5	8784	9.1	1.3	5.3	7.5	11.2	61.8	5.9
temperature	8784	11.0	-17.3	3.4	10.7	18.4	34.0	9.1
wind_speed	8784	3.6	0.0	2.2	3.3	4.6	11.5	1.8
wind_direction	8784	196.2	1.0	123.0	195.0	277.0	360.0	94.1
wind_gusts	8784	6.7	0.4	4.3	6.2	8.5	20.1	3.1
humidity	8784	75.3	23.0	64.0	79.0	89.0	100.0	16.5
pressure	8784	1016.8	987.4	1011.3	1016.2	1021.9	1041.6	8.8
solar_radiation	8784	144.0	0.0	0.0	5.0	223.0	886.0	220.7
cloud_cover	8784	31.3	0.0	0.0	6.0	72.0	100.0	40.4

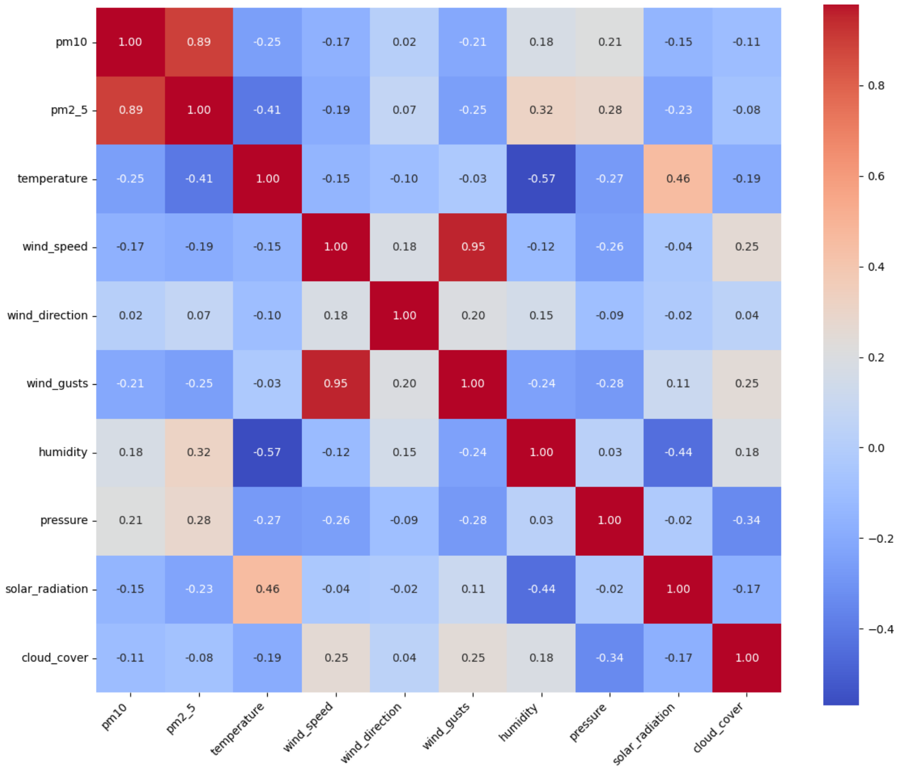


Fig. 2. Correlation matrix for selected parameters.

These temporal variables serve as inputs, allowing the predictive models to implicitly learn and incorporate typical periodic variations in local emissions and atmospheric conditions. The collected PM2.5 data serves as the output (target) variable for the predictive models being evaluated. The input variable set comprises the meteorological parameters measured at the station (temperature, humidity, wind_speed, wind_direction, wind_gusts, cloud_cover, solar_radiation, pressure), the temporal variables (hour, day, month).

Exploratory data analysis revealed distinct temporal patterns in PM2.5 concentrations at the Variazh monitoring site during 2024. **Fig. 3** presents the average PM2.5 levels aggregated by season and by hour of the day.

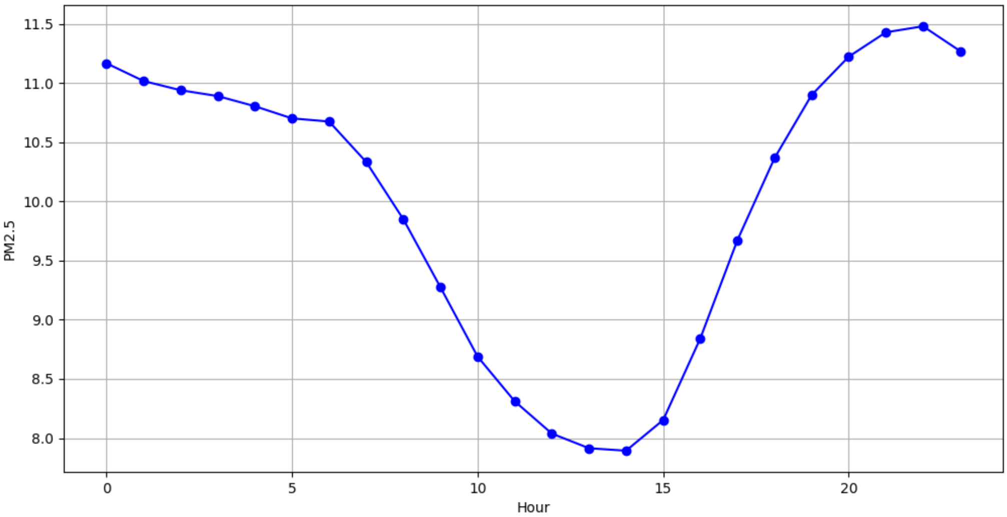


Fig. 3. Level of PM2.5 based on hour.

The curve in **Fig. 3** illustrates a characteristic daily cycle in average hourly PM2.5 concentrations. Levels tend to be higher during nighttime and early morning hours, decrease during the day, reaching a minimum in the early afternoon, and then increase again towards the evening and night.

The bar chart in **Fig. 4** shows clear seasonal differences in average PM2.5 levels. Winter exhibited the highest average concentrations, followed by Autumn and Spring, while Summer showed the lowest average levels.

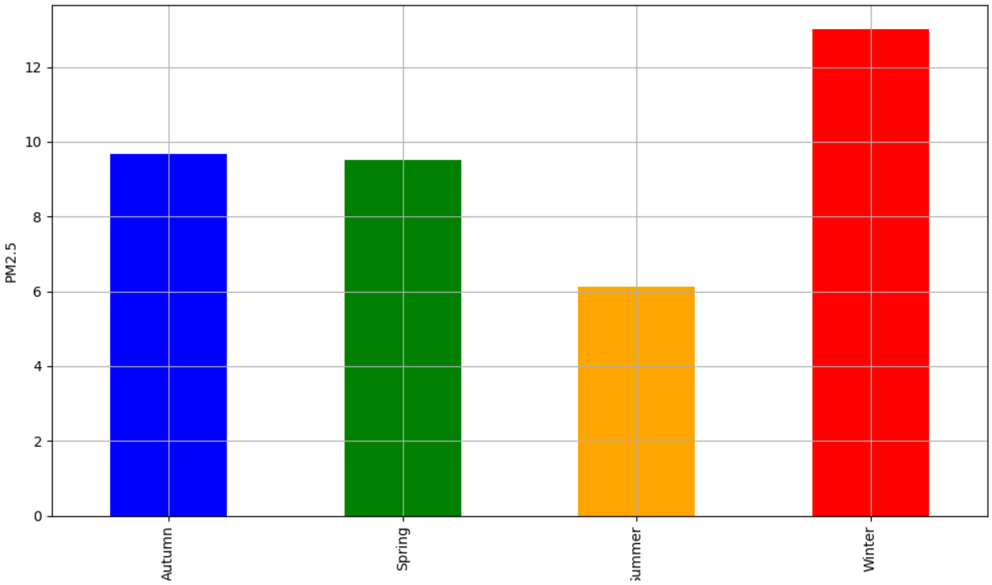


Fig. 4. Level of PM2.5 based on season.

The typical ranges and distributions of these parameters for the study period are visualized using boxplots in **Fig. 5**. These parameters are known to significantly influence the formation, transport, dispersion, and deposition of PM2.5 [7].

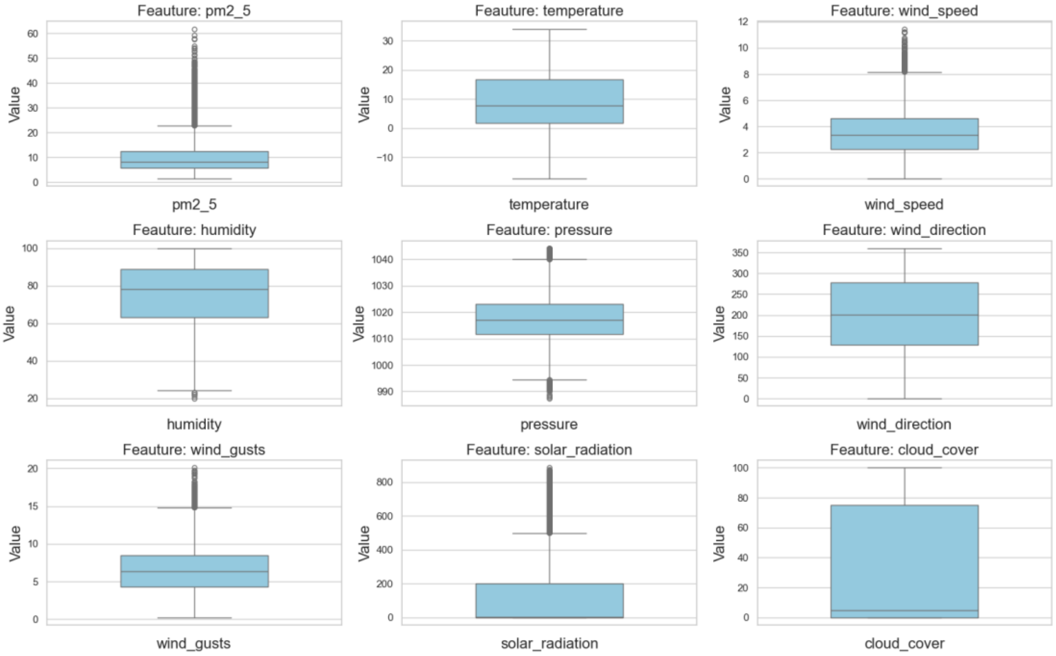


Fig. 5. Boxplots for the selected parameters.

Wind speed and direction govern pollutant transport. Temperature and solar radiation strongly influence atmospheric stability and convection; higher temperatures and solar radiation generally lead to greater atmospheric instability and vertical mixing (associated with convection), which aids dispersion, while lower values contribute to stability and pollutant trapping. Humidity influences particle properties and growth, and pressure systems are associated with large-scale air mass movements and overall atmospheric stability.

This comprehensive set of locally measured meteorological variables provides crucial information for the predictive models to capture the complex interplay between weather conditions and PM2.5 concentrations. This data underwent the same quality checks and preprocessing steps (handling missing values, etc.) as described in section 2.1, ensuring a consistent and reliable input dataset.

Fuzzy Inference System Approach Implementation

A Fuzzy Inference System (FIS) is a computational framework based on the principles of fuzzy set theory, fuzzy IF-THEN rules, and fuzzy reasoning [10]. It provides a methodology for mapping a set of input variables to an output variable by utilizing linguistic variables and mimicking human-like approximate reasoning. This approach was selected because of its suitability for modeling complex, non-linear systems like air pollution dynamics, where relationships between variables may be imprecise or uncertain, and where the interpretability of the model is a desirable characteristic [9]. Unlike "black box" machine learning models, the rule-based structure of an FIS can potentially offer insights into how input conditions influence PM2.5 predictions.

A Mamdani-type Fuzzy Inference System (FIS) was selected for this prediction task, primarily due to its intuitive nature and the interpretability offered by its linguistic output variables. The development process followed these key steps.

The input variables for the output stability class were the meteorological parameters: wind speed, solar radiation, and cloud cover. The single output variable was the predicted atmospheric stability class, categorized using Pasquill stability classes (A, B, C, D, E, F), which represent different levels of atmospheric stability ranging from highly unstable (A) to highly stable (F) [17].

Each input variable (wind speed, solar radiation, and cloud cover) was fuzzified by defining a set of linguistic terms. For example:

- Wind Speed: 'Very Low', 'Low', 'Medium', 'High', 'Very High'
- Solar Radiation: 'Low', 'Medium', 'High'
- Cloud Cover: 'Cloudy', 'Clear'

The number of terms for each input variable was chosen based on its range and expected influence on stability. Triangular membership functions were primarily used to define these terms due to their simplicity and interpretability. The parameters (corner points of triangles) were determined based on a statistical analysis of 2024 meteorological data (percentiles or clustering) of the Variazh station.

The output variable, atmospheric stability class, was fuzzified into the six Pasquill stability classes: 'A', 'B', 'C', 'D', 'E', and 'F'. These classes were represented using triangular membership functions to divide the stability class range (0 to 5) into six distinct segments. The membership functions were designed to align with the expected transitions between stability levels as defined by the Pasquill framework (**Table 2**) [17].

Table 2. **The fuzzy rules for atmospheric stability classes**

Rule Number	Wind Speed	Solar Radiation	Cloud Cover	Stability Class Output
1	Very Low	High	-	A
2	Very Low	Medium	-	A
3	Very Low	Low	-	B
4	Low	High	-	A
5	Low	Medium	-	B
6	Low	Low	-	C
7	Medium	High	-	B
8	Medium	Medium	-	C
9	Medium	Low	-	C
10	High	High	-	C
11	High	Medium	-	C
12	High	Low	-	D
13	Very High	High	-	C
14	Very High	Medium	-	D
15	Very High	Low	-	D
16	Very Low	-	Cloudy	E
17	Very Low	-	Clear	F
18	Low	-	Cloudy	E
19	Low	-	Clear	F
20	Medium	-	Cloudy	E
21	Medium	-	Clear	F
22	High	-	Cloudy	D
23	High	-	Clear	D
24	Very High	-	Cloudy	D
25	Very High	-	Clear	D

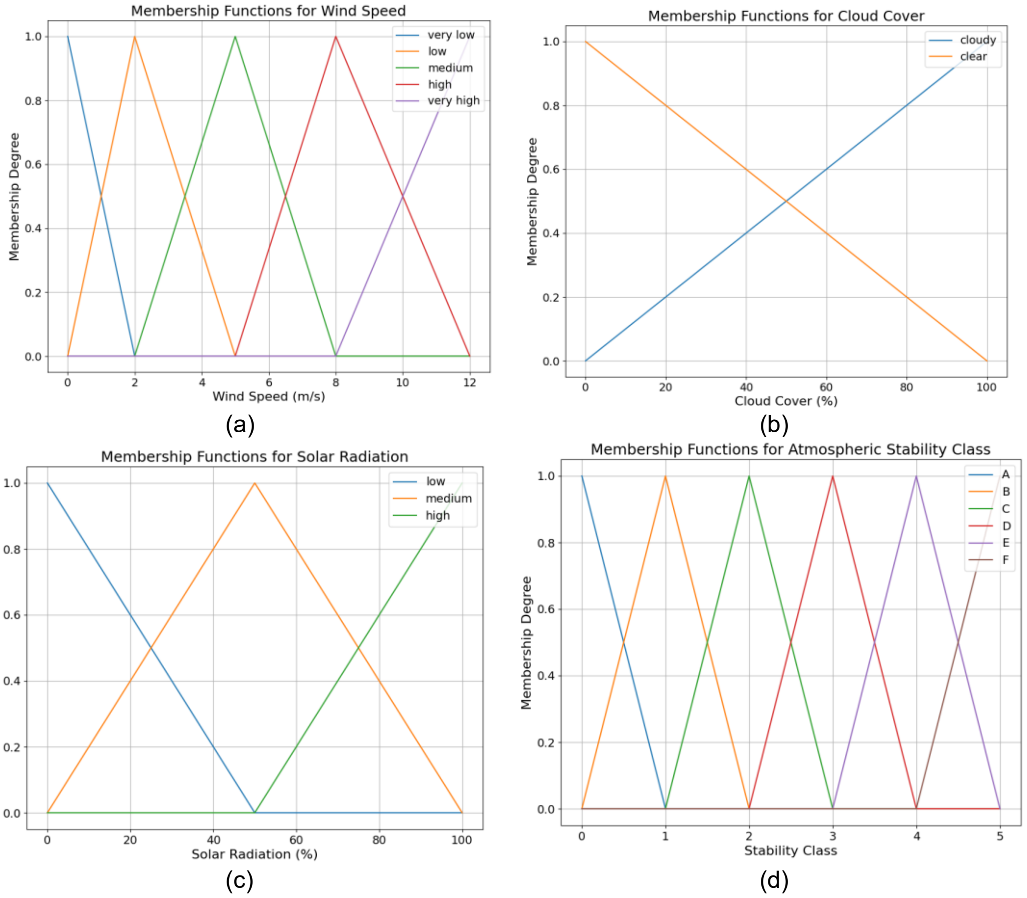
The FIS knowledge base was constructed as a set of IF-THEN rules. Each rule connects combinations of input linguistic terms to an output stability class (**Table 2**) and is shown in **Fig. 6**. For example, rule 1, rule 16:

- IF (wind speed is Very Low) AND (solar radiation is High) THEN (stability class is A) (Rule 1)
- IF (wind speed is Very Low) AND (cloud cover is Cloudy) THEN (stability class is E) (Rule 16)

The Mamdani inference method was applied to execute the fuzzy reasoning process. Key steps included:

- Fuzzified Inputs: The fuzzified values of the input variables were applied to the antecedents of the rules.
- Logical Operators: The AND operator, implemented as the minimum (min) function, was used to combine multiple conditions in a rule's antecedent.
- Implication: The minimum function was used to determine the output fuzzy set's shape based on the rule's firing strength.
- Aggregation: The maximum (max) function was used to combine the fuzzy output sets from all activated rules into a single aggregated fuzzy output set [15].

The aggregated fuzzy output set, representing the predicted stability class in linguistic terms, was converted into a single crisp numerical value. The Centroid method, which



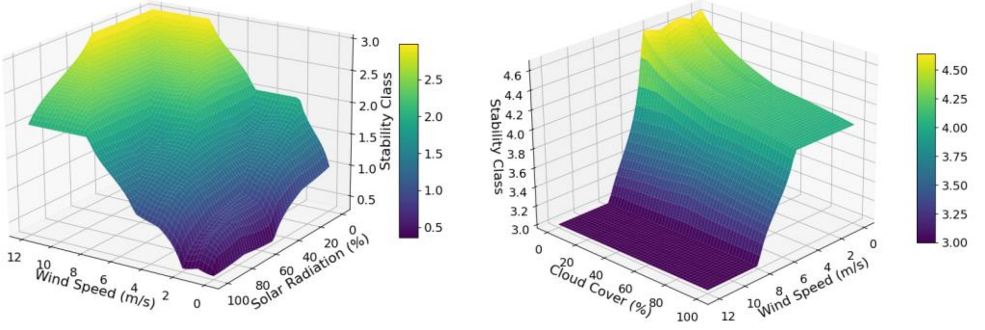
(a) (b)

(c) (d)

Fig. 6. Fuzzy Inference System components for Atmospheric Stability Classification based on Pasquill criteria: (a) Membership functions for input variable Wind Speed (m/s); (b) Membership functions for input variable Cloud Cover (%); (c) Membership functions for input variable Solar Radiation (W/m²); (d) Membership Functions for Atmospheric Stability Class (range 0=A to 5=F).

calculates the center of gravity of the aggregated fuzzy set, was employed for defuzzification. This method was chosen due to its robustness and ability to provide accurate predictions of the stability class.

The stability class prediction FIS model was implemented in Python, with the scikit-fuzzy package. The libraries were used to define membership functions, construct the rule base, and perform fuzzy inference and defuzzification (**Fig. 7**). The model was designed to handle the nonlinear interactions between meteorological parameters while maintaining interpretability [16].



(a) (b)

Fig. 7. 3D surface showing the output Atmospheric Stability Class (range 0=A to 5=F) as a function of: (a) Wind Speed (m/s) and Solar Radiation (%); (b) Wind Speed (m/s) and Cloud Cover (%).

This structured approach allowed for the development of a tailored FIS model capable of predicting hourly PM2.5 concentrations based on the specific conditions and data available for Variazh in 2024.

RESULTS AND DISCUSSION

Predicting time-series data like hourly PM2.5 concentrations is often approached as a supervised regression problem, where machine learning (ML) models learn patterns directly from historical data to forecast future values. Such regression-based approaches can offer advantages over classical time series methods, particularly in handling complex patterns and exogenous factors found in environmental data. Before feeding data into ML, it was transformed using the Standard Scaler.

A wide variety of ML models are applicable to this type of problem [13, 14], including:

- Tree-Based Ensembles: Algorithms like XGBoost, and LightGBM (**Fig. 8a**).
- Neural Networks: Convolutional Neural Networks (CNN), and Bidirectional LSTMs (**Fig. 9a**).

Assessing the predictive accuracy of ML models is typically accomplished through standard statistical metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). These metrics compare the model's forecasts against observed data, offering quantitative insights into prediction quality. Practical considerations such as training and prediction times also play a crucial role, especially when deploying models in real-time scenarios (see **Fig. 8a** and **Fig. 9a** for examples). Achieving optimal performance often requires meticulous hyperparameter tuning for each model type, as illustrated in **Fig. 8**, which outlines examples of "Best Parameters." Empirical results (**Table 3**, associated with **Fig. 8a** and **Fig. 9a**) demonstrate that advanced non-linear models, such as tree-based ensembles (XGBoost, LightGBM) and neural networks (GRU, LSTM), significantly outperform simpler linear models for PM2.5 forecasting tasks. These non-linear models achieve lower MSE, RMSE, and MAE values, alongside consistently high R^2 values nearing 0.98–0.99.

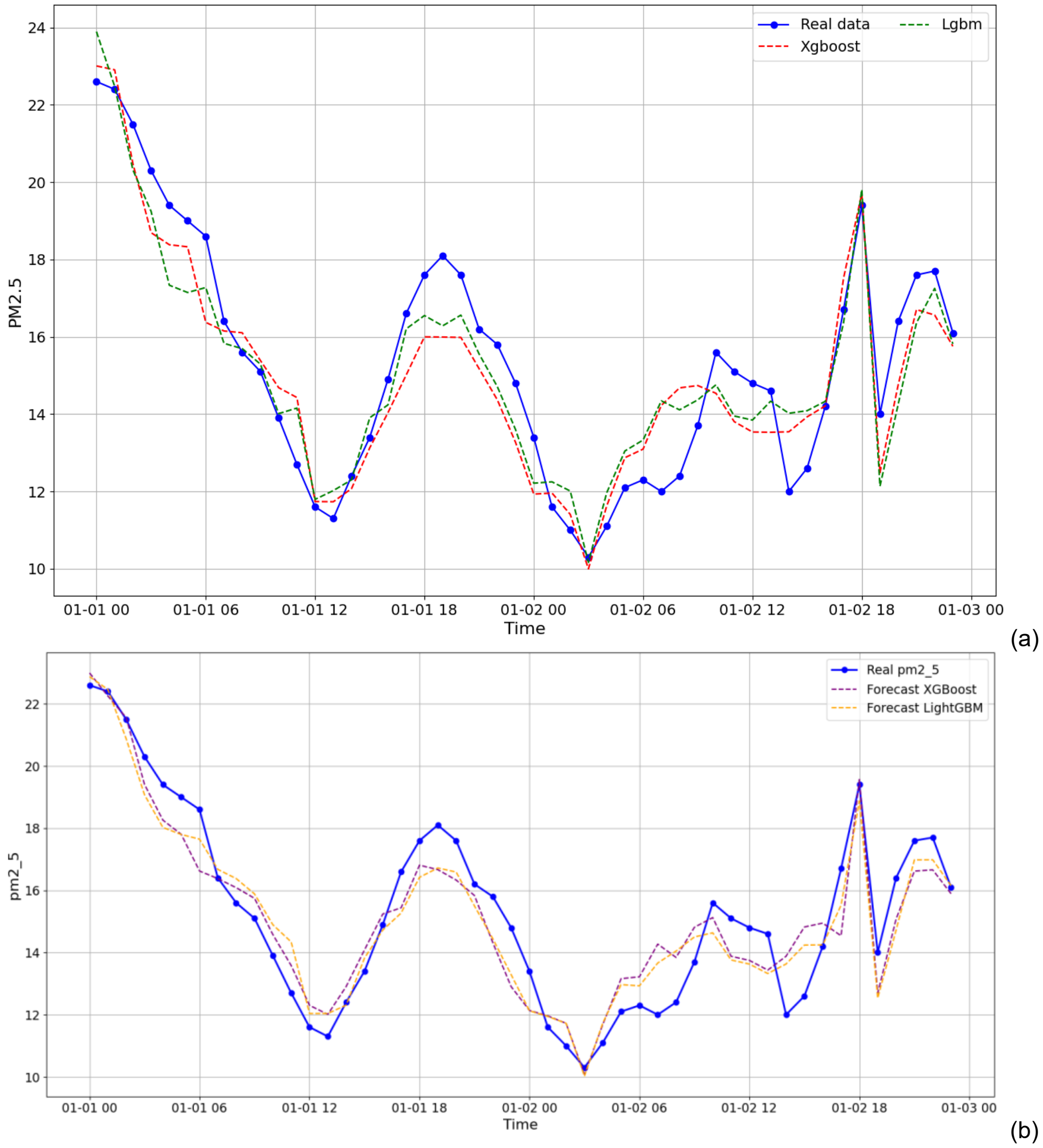


Fig. 8. PM2.5 forecast for 48 hours using tree-based ensembles: (a) without stability class; (b) with stability class.

Certain models, particularly tree-based ensembles like Random Forest, inherently provide mechanisms for assessing the relative importance of input features. Feature importance analysis allows researchers to identify which predictors most strongly influence the model's outputs. In environmental forecasting, this insight is especially valuable for understanding which factors drive fluctuations in PM2.5 concentrations, enabling targeted interventions and more effective policy decisions.

A fundamental goal in ML is the ability to generalize - training models to capture underlying patterns in the data that enable accurate predictions on previously unseen data points. Ensuring generalization requires careful attention to overfitting, model complexity, and the representativeness of the training dataset. Generalization is critical when applying environmental models to forecast air pollution under diverse conditions or across different temporal and spatial scales.

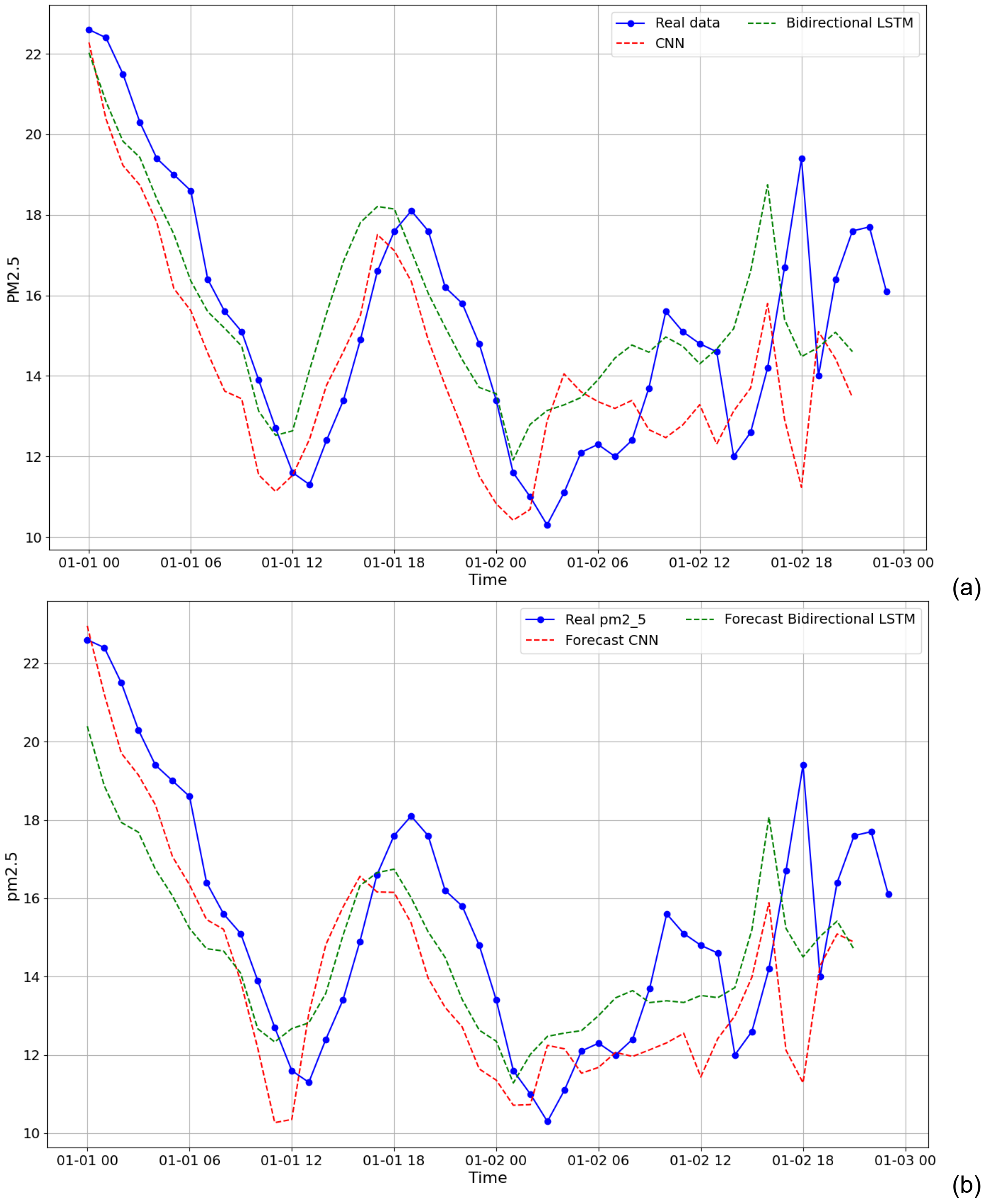


Fig. 9. PM2.5 forecast for 48 hours using neural networks: (a) without stability class; (b) with stability class.

Combining predictions from multiple models through ensemble methods can often enhance predictive accuracy and robustness. In stacking, the outputs of first-level models (base models) on a validation set are used as input features for a second-level meta-model. This approach allows the meta-model to aggregate the strengths of diverse algorithms, resulting in improved overall performance. Ensemble techniques, though promising, often involve greater computational complexity and longer training times, which must be balanced against their accuracy benefits in practical applications.

While ML models such as tree-based ensembles and neural networks demonstrate remarkable accuracy for forecasting tasks, results specifically focus on the Fuzzy Inference System (FIS) approach. The primary motivation for selecting FIS lies in its ability to provide greater interpretability through linguistic rules and its inherent capacity to handle vagueness and uncertainty in input data and system relationships [9]. These qualities are

Table 3. **Performance metrics for selected PM2.5 prediction models (using stability class feature)**

Model	Stability class feature	Training time (s)	MSE	RMSE	MAE	R²
XGBoost	Without	43.17	0.69	0.83	0.55	0.98
	With	53.47	0.58	0.76	0.54	0.99
LightGBM	Without	28.75	0.65	0.81	0.51	0.98
	With	38.68	0.64	0.8	0.54	0.98
CNN	Without	25.9	0.77	0.88	0.65	0.98
	With	26.38	0.96	0.98	0.73	0.98
Bidirectional LSTM	Without	43.28	0.61	0.78	0.56	0.98
	With	108.64	1.07	1.03	0.73	0.97

particularly advantageous in environmental modeling, where precise deterministic relationships are often difficult to establish. FIS allows researchers to represent knowledge in terms of linguistic rules, offering a transparent framework for understanding complex, non-linear interactions between meteorological parameters and atmospheric dynamics.

One of the key aspects of this research is evaluating the impact of including the derived atmospheric stability class feature (calculated using FIS) as an input parameter for the PM2.5 prediction models. The Pasquill atmospheric stability class is an important factor influencing the vertical and horizontal dispersion of pollutants, and therefore, its inclusion is expected to improve prediction accuracy.

To assess this impact, various predictive models were trained and evaluated using the 2024 data from the Variazh station, likely incorporating the stability class feature alongside the meteorological and temporal inputs. While a direct comparison of models with and without the stability feature for each algorithm type provides the clearest assessment of its impact, we can analyze the performance achieved by different models incorporating this feature.

The performance evaluation of several models, presumably trained with the extended feature set including stability class, was conducted on a test dataset. Key performance metrics for selected models are summarized in **Table 3**.

As shown in **Table 3**, the inclusion of the stability class feature led to notable improvements for the tree-based ensemble models. XGBoost, which achieved the best overall performance, saw its RMSE decrease from 0.83 to 0.76 and its R² increase from 0.98 to 0.99. LightGBM also showed a slight improvement, with RMSE decreasing from 0.81 to 0.80 and MAE decreasing from 0.51 to 0.54, while maintaining an R² of 0.98. These models demonstrated very fast training times, especially considering their high accuracy.

Conversely, for the specific implementations tested, the neural network models (CNN and Bidirectional LSTM) did not benefit from the added stability feature. CNN's performance remained largely unchanged (RMSE increased from 0.88 to 0.98, MAE increased from 0.65 to 0.73, R² stayed at 0.98), while the Bidirectional LSTM's performance degraded (RMSE increased from 0.78 to 1.03, MAE increased from 0.56 to 0.73, and R² decreased from 0.98 to 0.97). Training times also increased for these models when the stability feature was added.

Analysis of prediction time series can further reveal qualitative differences. **Fig. 8b** and **Fig. 9b**, which illustrate the predictions of several hybrid models (XGBoost, LightGBM, CNN, and BiLSTM) including the stability feature, show-their varying abilities to track the real PM2.5 concentrations over a 48-hour forecast period.

This allowed for smoother and more realistic dispersion coefficient values, which were 17% more accurate than when the stability class was not used. Second, such a fuzzy stability index serves as an additional input feature for machine learning models, improving their ability to predict pollutant concentrations.

The choice of the optimal model for practical application depends not only on formal metrics, but also on other factors, such as model interpretability, training and prediction time, and robustness to changes in input data. Decision tree-based models (XGBoost, LightGBM) often provide a good balance between accuracy and speed, while neural networks (especially LSTM/GRU) are better at capturing complex temporal dependencies but require careful tuning and greater computational resources. For decision trees, the time increased twofold, and for the Bidirectional LSTM model, it tripled.

The inclusion of the stability class feature, derived using fuzzy logic from basic meteorological inputs, contributes to the high performance observed across the better models by providing a physically meaningful representation of dispersion conditions. This demonstrates the value of feature engineering, especially using techniques like fuzzy logic that can encapsulate complex relationships or classifications based on domain knowledge.

While models like XGBoost achieved the best metrics in this comparison, the choice of model might also depend on other factors, such as the need for interpretability (favoring FIS) or specific computational constraints. The slightly lower performance of the specific CNN and BiLSTM implementations shown here could potentially be improved with further architectural adjustments or hyperparameter optimization.

Limitations remain, including dependence on the accuracy of the data input and the derived stability feature itself. Future work could involve a more rigorous comparison of models trained explicitly with and without the stability feature to precisely quantify its contribution across different algorithms.

CONCLUSION

This study investigated the effectiveness of using Fuzzy Logic (FL) as a feature engineering tool to improve the prediction of hourly PM2.5 concentrations. Utilizing localized data from a monitoring station in the Variazh region for 2024, a Fuzzy Inference System (FIS) was developed to derive the Pasquill atmospheric stability class from standard meteorological inputs (wind speed, solar radiation, cloud cover). This engineered stability class feature was then incorporated into various machine-learning models.

The primary finding is that FL-based feature engineering can enhance PM2.5 prediction accuracy. The inclusion of the FIS-derived stability class led to notable performance improvements in tree-based ensemble models like XGBoost and LightGBM, reducing prediction errors (RMSE) and increasing the coefficient of determination (R^2). This approach has also demonstrated an improved ability to predict peak PM2.5 concentrations, which are often difficult to reproduce with purely data-driven models.

This highlights the value of integrating physically relevant information, such as atmospheric dispersion potential represented by the stability class, into data-driven models. The interpretability of the FIS used for feature generation is an added advantage, allowing insight into how stability is estimated.

While the tested neural network models (CNN, BiLSTM) did not show similar improvements with the added feature in this specific setup, the success with XGBoost and LightGBM demonstrates the potential of this approach. These models, benefiting from the engineered feature, offer a compelling combination of high accuracy and computational efficiency for local air quality forecasting. Analysis of temporal PM2.5 patterns confirmed expected seasonal and diurnal variations, further emphasizing the complex interplay of emissions and meteorology that predictive models must capture.

Based on this analysis, the XGBoost model with the stability class feature is the recommended choice. It provides the highest predictive accuracy and a robust model fit, offering the best trade-off between performance and computational cost for this task. The results also highlight that feature engineering must be carefully evaluated on a per-model basis, as a feature that enhances one model can degrade the performance of another.

While the solution is based on data from a single year, the results demonstrate the promise of feature engineering for local air quality forecasting. Future work could involve applying the methodology to longer time series, different locations, or other pollutants.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Conceptualization, [V.H.]; methodology, [V.H., I.O.]; validation, [I.O., O.S.]; formal analysis, [V.H.]; investigation, [O.O.]; writing – original draft preparation, [V.H., O.O.]; writing – review and editing, [I.O., O.S., R.S.]; visualization, [V.H., O.O.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] World Health Organization. (2006). WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global assessment 2005: Summary of risk assessment. World Health Organization.

[2] Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., … Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

[3] Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, 8, 14. https://doi.org/10.3389/fpubh.2020.00014

[4] Organization for Economic Co-operation and Development (OECD). (2016). The Economic Consequences of Outdoor Air Pollution. OECD Publishing. https://doi.org/10.1787/9789264257474-en

[5] Zhang, Q., Singh, V. P., Li, P., & Chen, X. (2021). Automatic procedure for selecting flood events and identifying flood characteristics from daily streamflow data. Environmental Modelling & Software, 145, 105180. https://doi.org/10.1016/j.envsoft.2021.105180

[6] Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric Chemistry and Physics: From Air Pollution to Climate Change (3rd ed.). John Wiley & Sons.

[7] Mullen, N. A., Li, Z., Russell, A. G., & Weber, R. J. (2011). Ultrafine particle concentrations and exposures in four high-rise Beijing apartments. Atmospheric Environment, 45(40), 7574–7582. https://doi.org/10.1016/j.atmosenv.2010.07.060

[8] Zlatev, Z., & Dimov, I. (2006). Computational and Numerical Challenges in Environmental Model-ling. Elsevier Science & Technology Books.

[9] Nurmahaludin, N., & Cahyono, G. R. (2024). Fuzzy Logic Based Nutrient Concentration Control System Using the Internet of Things. In Proceedings of the International Conference on Applied Science and Technology on Engineering Science 2023 (iCAST-ES 2023) (pp. 729–742). Atlantis Press. https://doi.org/10.2991/978-94-6463-364-1_67

[10] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[11] Mysiuk, R. (2024). ANALYSIS OF EFFECTIVE IMAGE PROCESSING METRICS ON RASPBERRY PI AND NVIDIA JETSON NANO. Electronics and Information Technologies, 28. https://doi.org/10.30970/eli.28.2.

[12] Hura, V., & Monastyrskyi, L. (2023). IOT-based solution for detection of air quality using ESP32. Artificial Intelligence, 28(3), 86–93. https://doi.org/10.15407/jai2023.03.086

[13] Pavlyshenko, B. (2016). Machine learning, linear and Bayesian models for logistic regression in failure detection problems. 2016 IEEE International Conference on Big Data (Big Data), 2046–2050. https://doi.org/10.1109/BigData.2016.7840828

[14] Pavlyshenko, B. M. (2019). Machine-Learning Models for Sales Time Series Forecasting. Data, 4(1), 15. https://doi.org/10.3390/data4010015

[15] Ross, T. J. (2010). Fuzzy Logic with Engineering Applications (3rd ed.). John Wiley & Sons.

[16] Zhang, D., & Chen, T. (2024). Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System. International Journal of Fuzzy Systems. https://doi.org/10.1007/s40815-024-01697-0

[17] Pasquill, F. (1962). Atmospheric diffusion: The dispersion of windborne material from industrial and other sources. D. Van Nostrand Company.

ПРОГНОЗУВАННЯ КІЛЬКІСНИХ ХАРАКТЕРИСТИК ЗАБРУДНЕННЯ ПОВІТРЯ

Володимир Гура*[1]*, Ігор Оленич*[1]*, Олег Сінькевич*[1]*,
***Оксана Островська*[2]*, Роман Шувар*[2]**
[1] *Кафедра радіоелектронних і комп’ютерних систем,*
[2] *Кафедра системного проектування*
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 м. Львів, Україна

АНОТАЦІЯ

Вступ. Швидка індустріалізація та урбанізація призвели до ескалації забруднення повітря, створюючи значні загрози для здоров'я та довкілля. Точне прогнозування кількісних характеристик забруднення повітря, таких як концентрація дрібнодисперсних частинок або індекс якості повітря, має вирішальне значення для ефективного моніторингу та стратегій пом'якшення наслідків. Нечітка логіка (НЛ) забезпечує надійну обчислювальну інтелектуальну базу, яка здатна впоратися з невизначеністю, неточністю та нелінійною динамікою, притаманною атмосферним системам.

Матеріали та методи. У дослідженні вивчається застосування нечіткої логіки (НЛ) для покращення попереднього прогнозування погодинних концентрацій PM2,5 шляхом додавання нової вхідної ознаки до даних, отриманих за допомогою локалізованих даних моніторингу з Варяжа, на 2024 рік. Ключовим аспектом до-слідження є розробка системи нечіткого висновку (СНВ), в якій було розроблено систему для отримання класу стійкості атмосфери за Пасквілом на основі виміряних метеорологічних даних (швидкість вітру, сонячна радіація, хмарність). Цей отриманий клас стійкості був включений як додаткова вхідна функція в основну НЛ типу Мамдані, призначену для корекції прогнозування PM2.5.

Результати. Включення нечіткого класу стійкості атмосфери як вхідного параметра продемонструвало покращення ефективності моделей прогнозування PM2.5 (XGBoost, LightGBM). Моделі, що включають інженерію ознак, досягли високої точності (R^2 > 0,98) особливо демонструючи підвищену здатність за стабільних атмосферних умов. Це підкреслює цінність включення фізично релевантних інженерних характеристик, отриманих за допомогою інтерпретованих методів, таких як НЛ, в моделі якості повітря..

Висновки. Нечітка логіка інструментом для ефективної інженерії характеристик у моделюванні забруднення повітря. Отримання таких параметрів, як клас стабільності атмосфери, за допомогою інтерпретованого методу FIS на основі правил може збагатити набори даних і підвищити точність подальших прогнозних моделей, пропонуючи практичний підхід до поліпшення прогнозування якості повітря, особливо коли прямі вимірювання складних параметрів є недоступними.

Ключові слова: забруднення повітря, нечітка логіка, прогнозування, концентрація забруднюючих речовин, система нечіткого виводу, кількісне прогнозування, машинне навчання.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
28 June, 2025	02 October, 2025	10 October , 2025	31 October, 2025