ISSN 2007-9737

Bio-Inspired Optimization of Fuzzy Inference Rules for Air Quality Prediction in an Ensemble Framework

Francisco Javier Moreno Vazquez[1], Felipe Trujillo Romero[2,*], Amanda Enriqueta Violante Gavira[1]

[1] University of Guanajuato, Departamento de Ingeniería Mecánica, Mexico

[2] University of Guanajuato, Departamento de Ingeniería Electrónica, Mexico

fdj.trujillo@ugto.mx

Abstract. This study presents a novel ensemble learning approach designed to improve a fuzzy inference system (FIS) for forecasting PM2.5 pollution levels. The suggested model integrates the ensemble approach with a FIS to enhance predictive accuracy. By developing a collection of FIS, each trained on distinct subsets of the data, this method utilizes model diversity to enhance overall performance. Optimization algorithms are utilized to refine the FIS parameters, thereby improving the model's predictive performance. The performance of the optimized ensemble FIS is assessed through the analysis of a real-world dataset concerning PM2.5 pollution levels. The findings demonstrate that the suggested approach surpasses the conventional ensemble algorithm, such as the commonly utilized Random Forest, in terms of accuracy and robustness. The optimized ensemble FIS presents a compelling approach for accurate air quality forecasting, highlighting its significance as an essential instrument for environmental assessment and safeguarding public health.

Keywords. Ensemble models, fuzzy inference systems, particulate matter, optimization.

1 Introduction

In recent years, air pollution has emerged as a pressing environmental and public health concern, with particulate matter (PM), particularly PM2.5, posing significant risks due to its ability to penetrate the respiratory system [1, 2, 3]. While traditional reference monitors offer precise measurements, their limited distribution and substantial maintenance costs constrain both spatial and temporal coverage.

Consequently, there has been an increasing use of machine learning models to predict air quality in a manner that is both cost-effective and efficient [4, 5].

The integration of the outputs from multiple models has been shown to enhance prediction accuracy and robustness, particularly in the context of complex and ambiguous datasets [6]. Bagging techniques, including Random Forests, are a means of mitigating overfitting by training several models on bootstrapped data subsets and aggregating their predictions through averaging or majority voting [7].

Despite the documented success of ensemble models across multiple domains, their application in the context of fuzzy inference systems (FIS) remains largely unexplored.

Fuzzy logic provides a structured approach to managing uncertainty and imprecision through the use of human-like linguistic expressions [8].

FIS employs fuzzy logic to model intricate, nonlinear systems effectively by utilizing processes such as fuzzification, inference, and defuzzification. Nonetheless, traditional neuro-fuzzy systems, such as ANFIS, despite their effectiveness, may exhibit a tendency to overfit when confronted with intricate models [9].

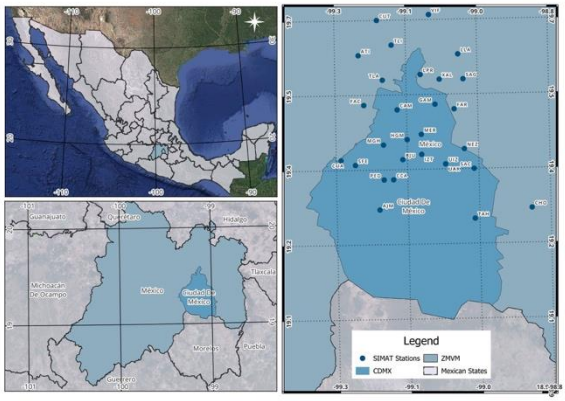

Fig. 1. Location of the SIMAT stations

A brief literature review of related topics yielded the following works, which may be of particular relevance to this study. For instance, in [10], the authors present a PM2.5 prediction method that utilizes image contrast-sensitive features and a weighted bagging-based neural network (WBBNN). This method involves the extraction of three features, their fuzzification, and the implementation of a weighted bagging strategy. Santana et al. [11] investigated the modeling and prediction of respiratory diseases, hospitalizations, and associated costs using an adaptive neuro-fuzzy inference system (ANFIS) based on air pollutant data. In another study, Zhang et al. [12] proposed a novel fuzzy forecasting system that enhances forecasting accuracy and certainty by leveraging fuzzy time series data, a data preprocessing technique, a multi-objective bat optimization algorithm, and forecasting algorithms. This system is designed to improve air quality supervision. In [13], Bhanja et al. propose a novel air quality forecasting method that utilizes type-2 fuzzy time series (FTS) and the FTSBO algorithm, a variant of the butterfly optimization algorithm.

The FTSBO algorithm demonstrates a commendable degree of efficacy in comparison to prevailing methodologies, thereby empowering administrators to exercise control over air pollution. The Takagi-Sugeno fuzzy inference system is also employed to formulate IF-THEN rules for urban air quality modeling. This system is optimized to minimize approximation error, thereby facilitating a rule base that describes the impact of individual input variables on the overall output [14]. In [15], a fuzzy-based methodology is proposed for the identification of air pollutant hotspots and critical urban areas during heatwaves. The model utilizes spatial interpolation and fuzzification to determine pollutant concentrations in assessing high-risk areas. Finally, Lima et al. [16] explore the application of fuzzy logic in air quality control systems, focusing on membership functions, decision-making engines, and defuzzification methods. The findings indicate that defuzzification exerts a more substantial influence on these systems.

In this study, we propose a novel ensemble approach that integrates bagging with FIS models, further optimized using a genetic algorithm. The methodology under consideration employs the distinct strengths of disparate fuzzy inference systems while concurrently addressing their limitations through a diversified ensemble approach. This approach is designed to deliver reliable and precise PM2.5 predictions, adeptly addressing the spatial and temporal complexities inherent in air quality data. The performance of our optimized bagging FIS is evaluated by employing a real-world dataset of PM2.5 pollution levels and contrasting it with the leading ensemble algorithm, Random Forest (RF).

2 Study Area

The research locale is the Metropolitan Area of the Valley of Mexico (ZMVM, according to its Spanish abbreviation). This region is among the most densely inhabited and urbanized in Latin America. The ZMVM encompasses Mexico City and its adjacent municipalities within the State of Mexico, collectively constituting a substantial megacity with a population exceeding 20 million [17]. The significance of this zone is attributable to Mexico City's status as the fifth largest urban area globally and its role as a prominent economic center in Latin America [18]. Figure 1 illustrates the geographical distribution of the Atmospheric Monitoring System (SIMAT) monitoring stations.

The dataset encompasses exhaustive records of particulate matter (PM10, PM2.5), nitrogen dioxide (NO2), ozone (O3), sulfur dioxide (SO2), and carbon monoxide (CO) levels, which were obtained from the Automatic Atmospheric

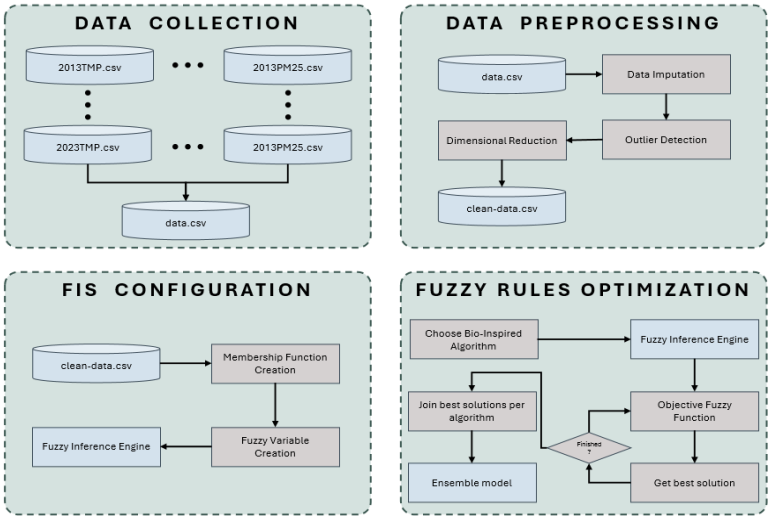

Fig. 2. Paper workflow diagram

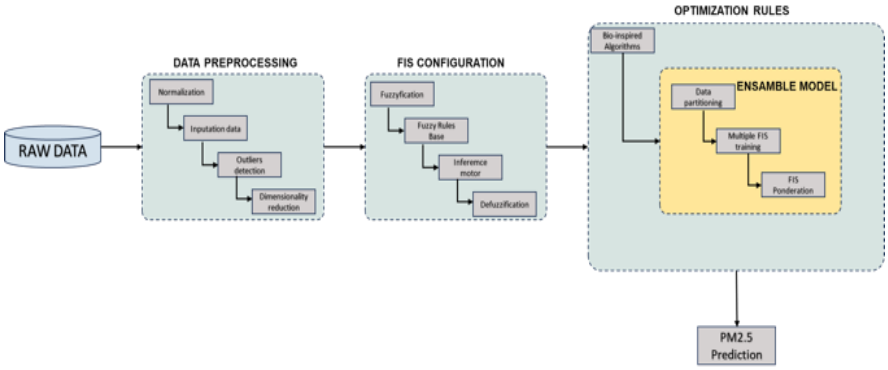

Fig. 3. Detail workflow diagram

Monitoring Network (Red Automática de Monitoreo Atmosférico, RAMA). Moreover, data concerning relative humidity (HR) and temperature (TMP) were obtained from the Meteorology and Solar Radiation Network (Red Meteorológica y de Radiación Solar, REDMET) [19].

3 Methodology

The present study focuses on the development of a comprehensive fuzzy inference system ensemble (FISE) for predicting PM2.5 levels, incorporating optimization methods to enhance the accuracy of forecasts. The methodology is comprised of four critical stages, each of which is vital for ensuring data reliability, model effectiveness, and computational efficiency. Figures 2 and 3 offer a visual representation of the workflow.

1. The process of data collection entails the procurement of data from all monitoring stations within the SIMAT system.
2. Data preprocessing entails the resolution of issues such as missing data and outliers, in addition to the utilization of dimensionality reduction techniques to identify principal components within the dataset.
3. The configuration of the FIS (Fuzzy Information System) involves the establishment of the variables and functions that are essential for the development of the fuzzy inference engine. Following the configuration, the engine must be

exported for utilization in subsequent optimization processes.

4. The utilization of bio-inspired algorithms to enhance the fuzzy rules employed by the fuzzy inference engine is referred to as "Fuzzy Rules Optimization."

3.1 Data Collection

The initial phase of this research, as illustrated in Figure 2, involves the collection of data. The files were collected from January 1st, 2013, to December 31st, 2023. The selected time intervals indicate a significant increase in mortality associated with particulate matter pollution in Mexico City, exceeding the 30-year average [20].

The data collected by SIMAT monitoring stations is structured as follows:

- The file naming convention adheres to a specific structure, whereby each file is named according to the year, followed by the relevant variable code or abbreviation. For instance, the relative humidity in 2014 is denoted as 2014RH, whereas the CO in 2019 is represented as 2019CO, and so forth.
- The subsequent step entailed the extraction of data from each file, given that the content comprised multiple stations arranged in columns. However, the necessity for automation arises from the periodic renaming or discontinuation of certain stations. Consequently, an identification process was conducted to ascertain the monitoring stations documented in the files within the designated time frame. This analysis yielded a total of 26 stations, which are listed below: The following are the acronyms: CHO, TLI, CUA, UAX, UIZ, MON, ACO, SJA, FAC, LLA, TLA, PED, NEZ, VIF, SFE, MER, COY, IZT, HGM, LPR, XAL, SAG, TAH, CAM, CUT, ATI.
- The data have been meticulously organized on a monthly basis for each year, culminating in a total of 3,432 samples. This comprehensive set encompasses 26 stations over a span of 11 years, with 12 months accounted for in each year.

3.2 Data Preprocessing

Data preprocessing constitutes a pivotal phase in data analysis, as it ensures the quality, consistency, and reliability of the input data prior to modeling [21]. Data obtained from monitoring stations often demonstrate gaps, anomalies, and discrepancies, which can negatively impact the efficacy of the FIS. This stage enhances the dataset's integrity through the implementation of normalization, cleaning, and transformation techniques, thereby enabling the model to generate more accurate and meaningful predictions of PM2.5 concentrations.

The initial aspect of our preprocessing that we do not explicitly categorize as a stage is normalization, which is a straightforward procedure where we adjust the range of values to fall between 0 and 1.

3.2.1 Data Imputation

The normalized data facilitates improved interval management; however, we encounter an issue with missing data. To address this, an imputation algorithm based on K nearest neighbors is employed, ensuring that the filled missing data exhibits characteristics akin to those of nearby data points [22].

3.2.2 Outlier Detection

Despite the completion of missing data, anomalies persist in the data, attributable to either sensor errors or data exceeding standard limits due to extraordinary events. This complicates data analysis further. In order to address this issue, there exist specific techniques for identifying outliers within and outside the data. The Isolation Forest algorithm is one such technique that utilizes an ensemble of trees to randomly partition the data. Outliers, which are few and distinct, are isolated in fewer splits compared to normal points [23, 24].

Furthermore, this algorithm exhibits linear time complexity with a low constant and minimal memory usage. Additionally, the data undergoes a process of winsorization, a technique employed to reduce the impact of outliers stemming from random sources and deviating from standard settings [25].

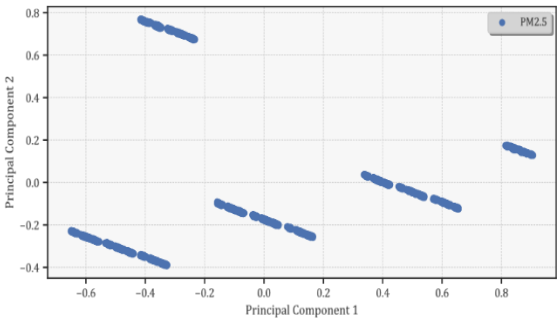

Fig. 4. Preprocess data comparison

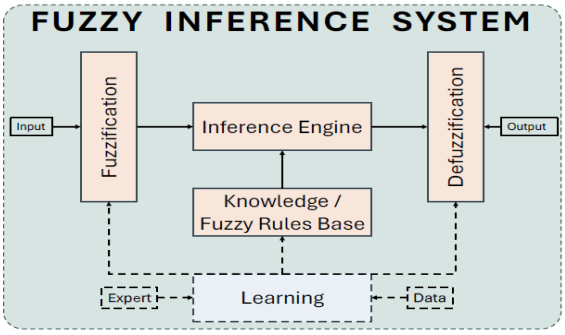

Fig. 5. Components of a typical FIS setup

3.2.3 Dimensional Reduction

As demonstrated in Section 2, the dataset under consideration encompasses environmental variables (TMP, RH) and pollutants (O3, CO, SO2, NO2, PM10, and PM2.5). Given the potential for substantial multicollinearity among the variables, particularly with PM2.5 as the target variable, Principal Component Analysis (PCA) is employed to address this issue [26, 27]. Principal component analysis (PCA) is a data transformation technique that transforms variables into a collection of orthogonal principle components, thereby encapsulating the majority of the variation present in the original data. This approach not only facilitates linear analysis but also enhances the resilience and efficiency of AI model training by mitigating issues associated with multicollinearity [28].

Upon completion of the data processing, a more refined and visually appealing dataset is obtained. This enhancement is distinctly demonstrated in Figure 4, which presents the state of the data following preprocessing.

3.3 FIS Configuration

Fuzzy logic is a mathematical framework that enables the representation of uncertainty and imprecision in decision-making processes. The field of fuzzy logic has seen extensive application across a variety of disciplines, including control systems, pattern recognition, and optimization [29]. This approach is particularly useful for modeling systems that are ambiguous and complex, where traditional binary logic methods may not be suitable. It allows for a transition from discrete 0 and 1 values to continuous values between 0 and 1.

The primary application of this framework is in a FIS, also known as an Expert System. An Expert System consists of four main components: fuzzification, the rule base, the inference engine, and defuzzification. The components of a FIS are illustrated in Figure 5.

The initial phase in the FIS is the fuzzification process, wherein crisp input data is converted into fuzzy sets employing membership functions (MF). These functions map the input data to linguistic variables, facilitating the representation of uncertainty and imprecision. The second component is the rule base, which contains a set of if-then rules that define the relationship between the input and output variables. Each rule consists of an antecedent (IF) and a consequent (THEN) part, linking the input variables to the output variable.

The inference engine constitutes the third component of the FIS, wherein the rules are applied to the input data to generate the output. This process entails the amalgamation of the fuzzy sets from the antecedents of the rules, thereby facilitating the determination of the degree of membership of the output variable. The final step in the process is the defuzzification stage, which transforms the output that is characterized by fuzziness into a set of crisp values, thereby yielding a result that is both clear and interpretable [30].

3.3.1 Membership Functions

In the context of fuzzy logic, membership functions delineate curves that quantify the degree of an element's affiliation with a specific set. The properties of these sets are characterized by a value range that typically extends from 0 to 1,

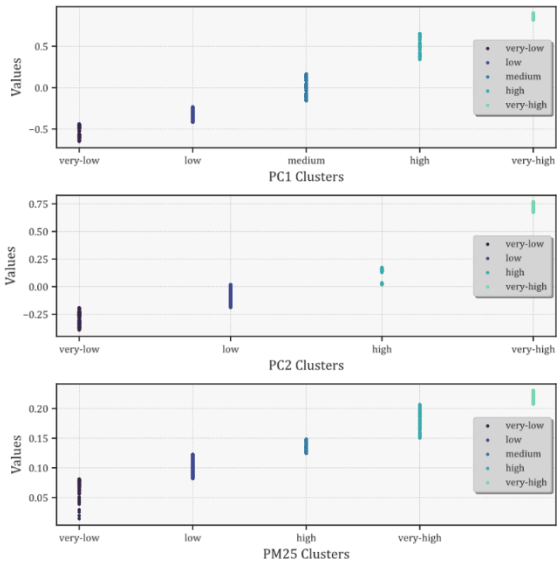

Fig. 6. Clustered variables to define the membership functions

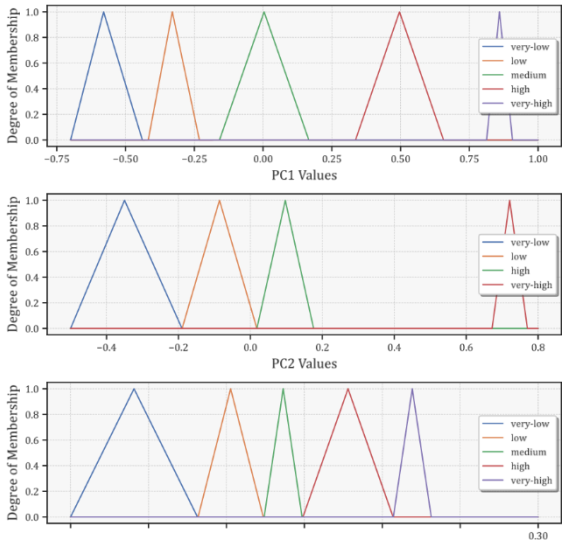

Fig 7. Clustered variables to define the membership functions

where membership is either true or false, with varying degrees of membership to a fuzzy set.

The configuration of membership functions manifests in a variety of forms, with the most prevalent patterns being trapezoidal, triangular, singleton, and S-shaped. For the present study, triangular membership functions were selected. Equation 1 presents the formula for determining the degree of membership of a value within each fuzzy set. In this formula, "a" represents the lower limit of the variable, "b" signifies the upper limit of the variable, and "m" denotes the position of the top vertex of the triangle on the x-axis.

The development of membership functions necessitates an essential preliminary step: the effective partitioning of the data domain. The K-means clustering algorithm is employed to ascertain the optimal number of clusters in accordance with the data distribution [31]. As demonstrated in Figure 6, the resultant clusters are indicative of discrete segments within the dataset:

$$A(x) = \begin{cases} 0 & \text{if } x \leq a \\ \dfrac{x-a}{m-a} & \text{if } x \in (a, m] \\ \dfrac{b-x}{b-m} & \text{if } x \in (m, b] \\ 0 & \text{if } x \geq b \end{cases} \quad (1)$$

After establishing the clusters and determining their value ranges, the subsequent step is to define the fuzzy variables. Each fuzzy variable is augmented with pertinent linguistic labels to facilitate intuitive human comprehension.

3.4 Fuzzy Rules Optimization

Given the intricate nature of atmospheric pollution patterns and the inherent uncertainty in environmental data, the optimization of FIS performance is imperative. Derivative-free optimization (DFO) methods are particularly well-suited for this task, as they eliminate the need for derivative calculations of the objective function. This characteristic enhances their effectiveness in addressing problems characterized by non-smooth, noisy, or costly objective function evaluations [32].

3.4.1 Bio-Inspired Algorithm Selection

The efficacy of bio-inspired algorithms in addressing complex optimization problems has been demonstrated. These biological techniques effectively manage the interplay between exploration and exploitation, making them highly adept at discovering optimal solutions across various application areas [33].

This research is grounded in a framework for the implementation and assessment of bio-

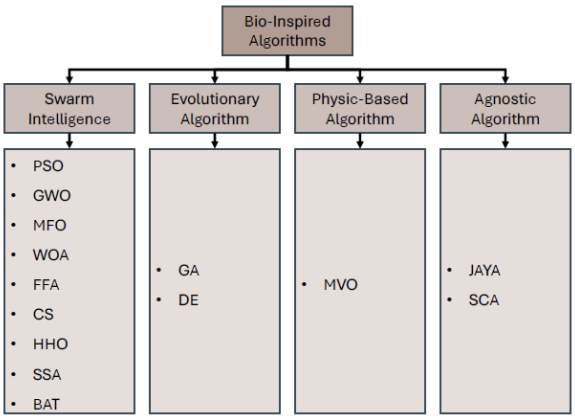

Fig 8. Classification of metaheuristic bio-inspired algorithms

Table 1. Fuzzy rules example

PC1 MF	PC2 MF			
	VL	L	H	VH
VL	H	M	M	VL
L	VH	L	VL	L
M	M	VL	VH	VL
H	M	VL	L	M
VH	H	VH	VL	VL

inspired algorithms, EvoloPy [34, 35, 36]. This framework encompasses a diverse array of algorithms, which can be categorized as illustrated in Figure 8.

The subsequent explanation will elucidate the various categories into which algorithms can be classified:

- The initial category within the bio-inspired domain is swarm intelligence, which emulates the predatory or collective behaviors observed in natural organisms. The following algorithms are included in this study: particle swarm optimization (PSO) [37], gray wolf optimizer (GWO) [38], moth-flame optimization (MFO) [39], whale optimization algorithm (WOA) [40], firefly algorithm (FFA) [41], cuckoo search (CS) [42], harris hawk optimization (HHO) [43], salp swarm algorithm (SSA) [44], and bat algorithm (BAT) [45].
- The second category of bio-inspired algorithms encompasses evolutionary algorithms, which are derived from the natural evolutionary process. The algorithms incorporated in this study are the genetic algorithm (GA) [46] and the differential evolution (DE) [47] algorithm.
- The third category of bio-inspired algorithms draws primarily from phenomena in nature that are rooted in physics or chemistry. The algorithm under consideration incorporates multiverse optimization (MVO) [48].
- In the final category, agnostic algorithms often draw inspiration from seemingly unrelated sources, including ideologies or mathematical processes. The algorithms incorporated in this study are JAYA [49] and the sine cosine algorithm (SCA) [50].

The selection criterion for the algorithms employed was the duration required to execute a specified number of iterations. An overarching experiment was conducted utilizing all algorithms inside the EvoloPy framework. From this experiment, one method from each category was chosen based on optimal performance in the least amount of time. The experimental design comprised 10 individuals, spanning 30 generations with 50 samples. Moreover, to enhance the statistical analysis, this experiment was conducted 30 times, yielding the selected algorithms with optimal results in their respective categories: The following acronyms are of particular relevance in this context: WOA, DE, MVO, and SCA.

3.4.2 Ensembled Model

Upon completion of each optimization iteration of the method, a vector of outcomes is obtained. This vector can be reformulated as a square matrix, representing the values of the respective membership functions for PM2.5.

An illustration of this outcome is presented in Table 1, where the entries correspond to the values obtained from the optimization process.

The methodology employed is of particular relevance due to the division of the data into batches of 150 samples, resulting in a total of 13 batches. The culmination of each batch is an optimization, which results in 13 distinct fuzzy rules for the inference engine.

The resultant rules are assessed separately, and by an aggregation procedure, such as the arithmetic mean, we obtain the final value of our ensemble fuzzy rule set.

Table 2. Parameters used in the experiment

Algorithm	Parameter	Value
DE	mutation factor	0.5
	crossover factor	0.7
MVO	wormhole existence probability (WEP)	(0.2,1)
RF	estimators(trees to create)	13

Table 3. Error metrics for each algorithm

Algorithm	MSE	MAE	R2
WOA	0.0005	0.0182	-0.1371
DE	0.0007	0.0191	-0.7152
MVO	0.0005	0.0195	-0.2376
SCA	0.0006	0.0209	-0.4407
RF	0.0008	0.0214	-0.8877

4 Results

The performance of the ensemble system was evaluated using a set of comprehensive metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared coefficient (R2). The mean-squared error (MSE) is a metric that quantifies the average of the squared differences between the predicted and actual values, thereby providing insight into the degree of discrepancy between the two values:

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

The R-squared (R2) metric is a measure of the proportion of variance in the target variable that is explained by the model. The values range from 0 to 1, with higher values indicating a stronger correspondence between the predicted and actual values.

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

The Mean Absolute Error (MAE) is a quantitative metric that calculates the average absolute discrepancy between the predicted and actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (4)$$

Within the context of the equations, y_i denotes the actual value, $\hat{y}_i$ represents the predicted value, $\bar{y}_i$ is defined as the mean of the actual values, and n is the number of observations.

The experiments were conducted by partitioning the original dataset into training and test sets with a 70:30 split, yielding 2009 training samples and 861 test samples. The optimization process employed a standard configuration comprising 10 individuals across 30 generations. The parameter settings for each optimizer are specified in Table 2.

Table 2 presents the explicit parameter settings for DE, MVO, and RF. However, it should be noted that WOA and SCA are not included in this table. WOA and SCA employ adaptive strategies that entail the dynamic adjustment of pivotal control parameters throughout the optimization process. This objective is typically accomplished by leveraging methodologies such as linear decrement schemes or random vector-based updates, as opposed to the utilization of predefined parameters. Consequently, their performance is governed by these inherent adaptive mechanisms, obviating the need for explicit parameter tuning.

The errors obtained for the test samples are presented in Table 3. The analysis indicates that the machine learning algorithm (RF) yields the highest value, whereas the evolutionary algorithm (DE) among the bio-inspired optimizers results in the lowest value.

A salient finding in Table 3 is the negative coefficient of determination (R2) across all methods. This phenomenon occurs when the predicted values exhibit a substantial deviation from the mean of the observed data, as delineated in Equation 3. While a negative R2 generally indicates that a model's performance is less effective than the use of the mean as a predictor, it is crucial to note that optimized methods result in R2 values that approach zero. In this context, an R2 of 0 indicates optimal performance, suggesting that the predictions are at least as accurate as the mean predictor. These findings underscore that the enhanced, optimized methods more effectively

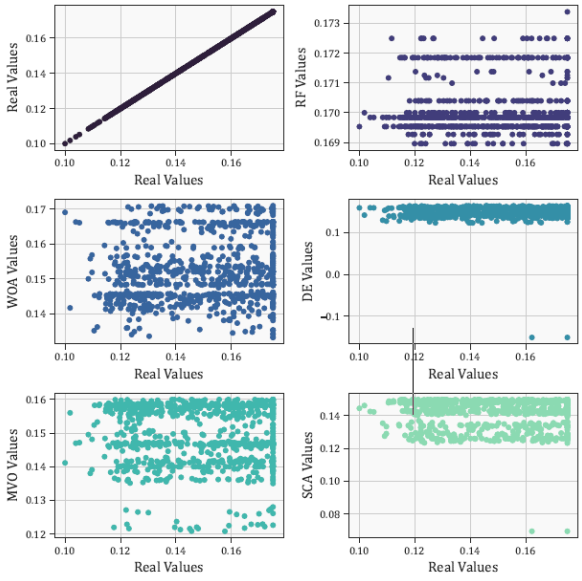

Fig. 9. Regression plots for the predicted values

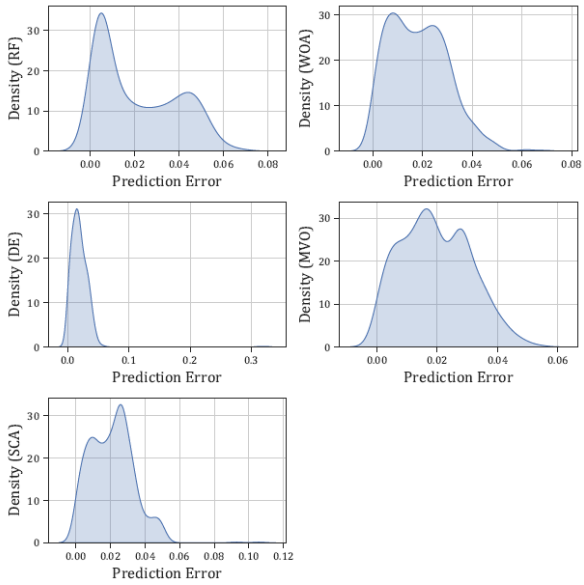

Fig. 10. Error density for the predicted values

minimize prediction errors, with the swarm algorithms (WOA) achieving the lowest error values and the machine learning approach (RF) yielding the highest error among the compared techniques.

Figure 9 presents the regression plot, which compares the predicted values with the actual measurements. Despite the regression line's deviation from the theoretical 1:1 line, as evidenced by the R2 values in Table 3, this should not be construed as an indication of the model's inadequacy.

Due to the imprecise nature of the model, its efficacy lies not in achieving exact precision for each prediction, but rather in its ability to minimize overall error and optimize outcomes under uncertainty.

To further elucidate this point, Figure 10 presents the distribution of the MSE across the sample. The findings suggest that the DE and SCA algorithms result in MSE values that are nearly zero, thereby demonstrating a high degree of accuracy in prediction. Conversely, algorithms such as RF, MVO, and WOA demonstrate broader MSE distributions, indicating comparatively larger prediction errors. These findings underscore the efficacy of the proposed approach in reducing error variance, even in cases where conventional regression metrics are inadequate in fully capturing the model's optimization capabilities.

5 Discussion

The findings of this study demonstrate that the implementation of ensemble models results in a substantial enhancement of the effectiveness of FIS in the forecasting of particulate matter pollution. The integration of numerous FIS has been demonstrated to enhance resilience and accuracy, thereby ensuring more reliable projections. The findings of the present study indicate a substantial correlation between the number of FIS in the ensemble and forecast accuracy, with larger ensembles producing improved performance. A collection of several weaker FIS can outperform a single, strong FIS, illustrating the advantage of employing different models to tackle intricate, nonlinear, and unpredictable environmental data [51].

However, the inherent algorithmic complexity of fuzzy logic systems poses significant challenges for training and experimental reproduction [29]. The subsequent undertakings will center on enhancing computational efficiency through the utilization of batch processing and parallelization techniques, which are imperative for minimizing

execution durations in complex experimental configurations [52].

In contrast to the use of spatial techniques [15] or Takagi-Sugeno systems [14], this work integrates multidimensional variables (pollutants and meteorological) using PCA, demonstrating robustness to noise and complex correlations. However, future improvements could include process parallelization [52] and the adoption of type-2 membership functions [16] for greater flexibility. While Random Forest [7] maintains competitiveness in the context of generalization, the proposal demonstrates particular aptitude in applications necessitating interpretability, such as in the domain of air quality public policy.

In summary, the study presented in this paper advances PM2.5 prediction by balancing accuracy and transparency. However, it requires adjustments to scale efficiently in operational environments.

6 Conclusion and Future Work

The findings of this study demonstrate that the implementation of ensemble models results in a substantial enhancement of the effectiveness of FIS in the forecasting of particulate matter pollution. The integration of numerous FIS has been demonstrated to enhance resilience and accuracy, thereby ensuring more reliable projections. The findings of the present study indicate a substantial correlation between the number of FIS in the ensemble and forecast accuracy, with larger ensembles producing improved performance. A collection of several weaker FIS has been shown to outperform a single, strong FIS, illustrating the advantage of employing different models to tackle intricate, nonlinear, and unpredictable environmental data [51].

However, the inherent algorithmic complexity of fuzzy logic systems poses significant challenges for training and experimental reproduction [29]. The subsequent undertakings will center on enhancing computational efficiency through the utilization of batch processing and parallelization techniques, which are imperative for minimizing execution durations in complex experimental configurations [52].

References

1. **WHO (2021).** Air pollution. https://www.who.int/health- topics/ air-pollution.
2. **Cao, R., Li, B., Wang, Z., Peng, Z.-R., Tao, S., Lou, S. (2020).** Using a distributed air sensor network to investigate the spatiotemporal patterns of PM2.5 concentrations. Environmental Pollution, 264, 114549. Doi: 10.1016/j.envpol.2020.114549.
3. **Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., Samet, J. M. (2006).** Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA, Vol. 295, pp. 1127–1134. Doi: 10.1001/jama.295. 10.1127.
4. **Tanzer, R., Malings, C., Hauryliuk, A., Subramanian, R., Presto, A. (2019).** Demonstration of a Low-Cost Multi-Pollutant Network to Quantify Intra-Urban Spatial Variations in Air Pollutant Source Impacts and to Evaluate Environmental Justice. International Journal of Environmental Research and Public Health, Vol. 16, 2523. Doi: 10.3390/ijerph16142523.
5. **Woltmann, L., Deepe, J., Hartmann, C., Lehner, W. (2023).** evalPM: a framework for evaluating machine learning models for particulate matter prediction. Environmental Monitoring and Assessment, Vol. 195. Doi: 10.1007/s10661-023-11996-y.
6. **Fouillen, E., Boyer, C., Sangnier, M. (2023).** Proximal boosting: Aggregating weak learners to minimize non-differentiable losses. Neurocomputing, 520. Doi: 10.1016/j.neucom. 2022. 11.065.
7. **Bourel, M., Cugliari, J., Goude, Y., Poggi, J. M. (2024).** Boosting diversity in regression ensembles. Statistical Analysis and Data Mining, 17. Doi: 10.1002/sam.11654.
8. **Laufer, E. (2023).** Personal Statistics-Based Heart Rate Evaluation Using Interval Type-2 Fuzzy Sets. Computacion y Sistemas, Vol. 27. Doi: 10.13053/ CyS-27-4-4784.
9. **Termeh, S. V. R., Kornejady, A., Pourghasemi, H. R., Keesstra, S. (2018).** Flood susceptibility mapping using novel

ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. Science of The Total Environment, Vol. 615, pp. 438–451. Doi: 10.1016/j.scitotenv.2017. 09.262.

10. **Qiao, J., He, Z., Du, S. (2020).** Prediction of PM2.5 concentration based on weighted bagging and image contrast-sensitive features. Stochastic Environmental Research and Risk Assessment, Vol. 34, No. 3–4, pp. 561-573. Doi: 10.1007/ s00477-020-01787-z.
11. **Santana, J. C. C., Miranda, A. C., Rosa, J. M., Berssaneti, F. T., Ho, L. L., da Silva Ferreira, L. L., Gomes, R. A., de Araújo, S. A., Belan, P. A. (2024).** A neuro-fuzzy model to predict respiratory disease hospitalizations arising from the effects of traffic-related air pollution in São Paulo. Clean Technologies and Environmental Policy. Doi: 10.1007/s10098-024-02877-0.
12. **Zhang, H., Wang, J., Nie, Y. (2024).** A novel optimization model based on fuzzy time series for short-term Air Quality Index forecasting. Knowledge-Based Systems, Vol. 296, 111905. Doi: 10.1016/j. knosys.2024.111905.
13. **Bhanja, S., Das, A. (2024).** An air quality forecasting method using fuzzy time series with butterfly optimization algorithm. Microsystem Technologies, Vol. 30, No. 5, pp. 613–623. Doi: 10.1007/s00542-023- 05591-x.
14. **Michalíková, A. (2025).** Explanation of Air Quality Data Using Takagi–Sugeno Fuzzy Inference System. Applied Sciences, Vol. 15, No. 7. Doi: 10.3390/app15073461.
15. **Cardone, B., Di Martino, F., Mauriello, C., Miraglia, V. (2025).** A Fuzzy-Based Model to Detect Hotspots of Air Pollutants During Heatwaves in Urban Settlements. Sensors, Vol. 25, No. 7. Doi: 10.3390/s25072160.
16. **Lima, J. F., Patiño-León, A., Orellana, M., Zambrano-Martinez, J. L. (2025).** Evaluating the Impact of Membership Functions and Defuzzification Methods in a Fuzzy System: Case of Air Quality Levels. Applied Sciences, Vol. 15, No. 4. Doi: 10.3390/app15041934.
17. **INEGI (2020).** Información de México para Niños. https://cuentame.inegi.org.mx/ monografias/informa cion/df/poblacion/.
18. **Bell, M. L., Davis, D. L., Gouveia, N., Borja-Aburto, V. H., Cifuentes, L. A. (2006).** The avoidable health effects of air pollution in three Latin American cities: Santiago, São Paulo, and Mexico City. Environmental Research, Vol. 100. Doi: 10.1016/j.envres.2005.08.002.
19. **SIMAT (2024).** Dirección de Monitoreo Atmosférico. http://www.aire.cdmx.gob.mx/entorno s/entorno%5C_detalle.php?est=eXl9.
20. **IHME, G. B. o. D. (2021).** Deaths from outdoor air pollution. https://ourworldindata.org/ outdoor-air-pollution.
21. **Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., Engel, T. (2024).** Survey: Time-series data preprocessing: A survey and an empirical analysis. Journal of Engineering Research (Kuwait). Doi: 10.1016/j.jer.2024.02.018.
22. **Zhang, S. (2012).** Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, Vol. 85. Doi: 10.1016/j.jss. 2012.05.073.
23. **Liu, F. T., Ting, K. M., Zhou, Z. H. (2012).** Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, Vol. 6. Doi: 10.1145/2133360.2133363.
24. **Abri, K. A., Sidhu, M. S. (2024).** Machine Learning Approaches to Advanced Outlier Detection in Psychological Datasets. International Journal of Electrical and Computer Engineering Systems, Vol. 15. Doi: 10.32985/ijeces.15.1.2.
25. **Lien, D., Balakrishnan, N. (2023).** Some results on multiple regression analysis with data cleaned by trimming and winsorization. Communications in Statistics: Simulation and Computation, Vol. 52. Doi: 10. 1080/03610918.2021.1982974.
26. **Jollife, I. T., Cadima, J. (2016).** Principal component analysis: A review and recent developments. Doi: 10.1098/rsta.2015.0202.
27. **Zateroglu, M. T. (2024).** Forecasting particulate matter concentrations by combining statistical models. Journal of King Saud University - Science, 36. Doi: 10.1016/j.jksus.2024.103090.

28. **Feng, X., Park, D. S., Liang, Y., Pandey, R., Papeş, M. (2019).** Collinearity in ecological niche modeling: Confusions and challenges. Ecology and Evolution, Vol. 9. Doi: 10.1002/ece3.5555.
29. **Fuentes-Penna, A., Díaz-Parra, O., de D. González-Ibarra, J., Flores, P. E., Simancas-Acevedo, E., Aguilar-Ortiz, J., Ruiz-Vanoye, J. A. (2023).** Complexity on Fuzzy set and Fuzzy Logic for Air Quality. International Journal of Combinatorial Optimization Problems and Informatics, Vol. 14, pp. 43–48. Doi: 10.61467/ 2007.1558.2023. v14i2.367.
30. **Warner, J., Sexauer, J., scikit-fuzzy, twmeggs, alexsavio, Unnikrishnan, A., Castelão, G., Pontes, F. A., Uelwer, T., pd2f, laurazh, Batista, F., alexbuy, den Broeck, W. V., Song, W., Badger, T. G., Pérez, R. A. M., Power, J. F., Mishra, H., (2019).** JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2. Doi: 10.5281/zenodo.3541386.
31. **Jin, X., Han, J. (2011).** K-Means Clustering. Encyclopedia of Machine Learning, pp. 563–564. Springer US. Doi: 10.1007/978-0-387-30164-8_425.
32. **Audet, C., Hare, W. (2017).** Derivative-free and blackbox optimization. Springer, pp. 3–6.
33. **Zhong, C., Li, G., Meng, Z., Li, H., Yildiz, A. R., Mirjalili, S. (2025).** Starfish optimization algorithm (SFOA): A bio-inspired metaheuristic algorithm for global optimization compared with 100 optimizers. Neural Computing and Applications, Vol. 37, pp. 3641–3683. Doi: 10.1007/s00521-024-10694-1.
34. **Faris, H., Aljarah, I., Mirjalili, S., Castillo, P. A., Merelo, J. J. (2016).** EvoloPy: An Open-source Nature-inspired Optimization Framework in Python. Doi: 10.5220/0006048201710177.
35. **Khurma, R. A., Aljarah, I., Sharieh, A., Mirjalili, S. (2020).** EvoloPy-FS: An Open-Source Nature-Inspired Optimization Framework in Python for Feature Selection. Evolutionary Machine Learning Techniques, Springer, Singapore, pp. 131–173, Vol. 1. Doi: 10.1007/ 978-981-32-9990-0_8.
36. **Qaddoura, R., Faris, H., Aljarah, I., Castillo, P. A. (2021).** EvoCluster: An Open-Source Nature-Inspired Optimization Clustering Framework. SN Computer Science, Vol. 2, pp. 185. Doi: 10.1007/s42979-021-00511-0.
37. **Kumar, T., Bhargava, A. K., Sharma, M. K., Dhiman, N., Nain, N. (2024).** Hybrid approach of type-2 fuzzy inference system and PSO in asthma disease. Clinical eHealth, Vol. 7. Doi: 10.1016/j. ceh.2024.01.001.
38. **Mirjalili, S., Mirjalili, S. M., Lewis, A. (2014).** Grey Wolf Optimizer. Advances in Engineering Software, Vol. 69, pp. 46–61. Doi: 10.1016/j.advengsoft. 2013.12.007.
39. **Mirjalili, S. (2015).** Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems, Vol. 89. Doi: 10.1016/j.knosys.2015.07.006.
40. **Mirjalili, S., Lewis, A. (2016).** The Whale Optimization Algorithm. Advances in Engineering Software, Vol. 95. Doi: 10.1016/j.advengsoft.2016. 01.008.
41. **Yang, X. S. (2009).** Firefly algorithms for multimodal optimization. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5792 LNCS, pp. 169–178. Doi: 10.1007/978-3-642-04944-6_14.
42. **Tappiti, C., Lin, T. K. (2024).** Optimization of hybrid platform for high-tech equipment and building vibration mitigation using evolutionary algorithms. Structures, Vol. 60. Doi: 10.1016/j.istruc.2023. 105818.
43. **Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., Chen, H. (2019).** Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems, Vol. 97. Doi: 10.1016/ j.future.2019.02.028.
44. **Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., Mirjalili, S. M. (2017).** Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software, Vol. 114, pp. 163–191. Doi: 10.1016/ j.advengsoft.2017.07.002.
45. **Yang, X. S., Gandomi, A. H. (2012).** Bat algorithm: A novel approach for global engineering optimization. Engineering

Computations (Swansea, Wales), Vol. 29. Doi: 10.1108/02644401211235834

46. **Jayaram, M. A., Chandana, M. (2024).** Design of flexible pavements through fuzzy inference system with genetic algorithm optimized rule base. International Journal of Transportation Science and Technology, Vol. 13. Doi: 10.1016/j.ijtst.2023. 03.001.
47. **Meziane, K. B., Dib, F., Benaya, N., Boumhidi, I. (2023).** Optimized fuzzy PI controller for variable speed wind turbine using DE algorithm. International Journal of Power Electronics and Drive Systems, Vol. 14. Doi: 10.11591/ijpeds.v14.i3.pp1684-1693.
48. **Mirjalili, S., Mirjalili, S. M., Hatamlou, A. (2016).** Multi-Verse Optimizer: A nature-inspired algorithm for global optimization. Neural Computing and Applications, Vol. 27. Doi: 10.1007/s00521-015-1870-7.
49. **Rao, R. V. (2016).** Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. International Journal of Industrial Engineering Computations, Vol. 7, pp. 19–34. Doi: 10.5267/j.ijiec.2015.8.004.
50. **Mirjalili, S. (2016).** SCA: A Sine Cosine Algorithm for solving optimization problems. Knowledge-Based Systems, Vol. 96. Doi: 10.1016/j.knosys. 2015.12.022.
51. **Zaman, A., Nassar, R. U. D., Alyami, M., Shah, S., Rehman, M. F., Hakeem, I. Y., Farooq, F. (2023).** Forecasting the strength of micro/nano silica in cementitious matrix by machine learning approaches. Materials Today Communications, Vol. 37. Doi: 10.1016/j.mtcomm.2023.107066.
52. **Miliauskaite, J., Kalibatiene, D. (2020).** Complexity Issues in Data-Driven Fuzzy Inference Systems: Systematic Literature Review. Doi: 1007/978-3-030-57672-1_15.

Article received on 05/06/2025; accepted on 04/10/2025.
**Corresponding author is Felipe Trujillo Romero.*