A K-MEDOID CENTERED EFFICIENT KNOWLEDGE RETRIEVAL SYSTEM FOR AGRICULTURE

Pankaj Kumar Sahu* , Dr. Mukesh Kumar**
*Department of Computer Science & Engineering
Rabindranath Tagore University, India
Email- pksmdb94@gmail.com
**Department of Computer Science & Engineering
Rabindranath Tagore University, Raisen, Madhya Pradesh, India
Email- goutam.mukesh@gmail.com

Abstract- Air constitutes a vital element for life on our planet, and the air quality in a given area significantly impacts the well-being of its residents. The issue of pollutions of air poses a substantial threat to both person health and the surrounding. If this ongoing trend of air pollution persists, it can lead to several severe consequences for the Earth's atmosphere, including an increase in global temperatures that can have detrimental effects on life. The ramifications of air contamination encompass global warming, the greenhouse effect, and a host of respiratory ailments, including bronchitis and lung cancer. To evaluate quality of the air in a specific region, the Air Quality Index (AQI) serves as a key indicator. The proposed approach seeks to predict AQI values using Artificial Neural Networks (ANN) and regression models. Additionally, it aims to forecast the concentration of the primary pollutant responsible for influencing AQI in a given area. The ability to predict future air quality, as indicated by AQI, is crucial for effective air quality monitoring, incorporating factors such as pollutant levels, temperature, relative humidity, and more. Forecasting future air quality and determining the primary pollutant's concentration aids in preventing the deterioration of air quality by addressing pollutant sources. Multiple neural network algorithms, including feedforward neural networks, cascade forward neural networks, and Elman recurrent neural networks, are employed. A machine learning ensemble approach combines these neural network models to enhance accuracy compared to the existing models, which predominantly rely on feedforward neural networks. This proposed machine learning (ML) technique has demonstrated superior accuracy compared to existing methods, as measured by various metrics such as Correlation Coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), and Index of Agreement (IA).

Keywords: Elman recurrent neural network, Cascade forward neural network,

1. INTRODUCTION

Air quality monitoring and prediction are essential for understanding and managing the impact of pollutants on the environment and human health. Monitoring provides direct measurements of air quality, while prediction models offer valuable insights into the temporal and spatial distribution of pollutants. Air quality models are cost-effective tools used to estimate air quality standards, making them crucial for regulatory officials to evaluate emissions' effects on ambient air quality and develop strategies to meet quality requirements .

Air quality models are mathematical representations of pollutant transport, diffusion, and chemical reactions originating from pollution sources. They incorporate various parameters affecting pollutant concentrations at different distances downwind of emission sources. These models use input data to describe emissions, meteorological conditions, and topography to generate air quality forecasts. Depending on the complexity of input variables, models can be simple or advanced, with advanced models suited for scenarios involving photochemical air pollution, dispersion in complex terrain, and long-range pollutant transport. Simpler models are better suited for predicting particulate matter pollutants downwind of sources.

Air quality predictions inherently carry a higher degree of uncertainty compared to other forecasts because they must account for meteorological variables as well as pollutant concentrations. Forecasting models reduce uncertainties by utilizing prior information and adjusting the model with additional data, such as past measurement values and meteorological parameters. Among the pollutants, carbon monoxide (CO) and fine particulate matter with aerodynamic diameters less than 10 µm (PM10) and 2.5 µm (PM2.5) are commonly predicted.

Monitoring air quality through measurements provides valuable information about the air we breathe. Long-term monitoring is particularly useful for identifying patterns that inform air quality control policies, including spatial pollution disparities and temporal variations. While monitoring itself doesn't reduce pollution, it helps identify its sources and levels and assess the effectiveness of pollution control efforts. Monitoring units are established in areas with significant air pollution issues, collecting large volumes of data over time. The challenge is to interpret and validate this data effectively, and this is where Machine Learning (ML) comes into play.

ML is a powerful tool for interpreting complex, high-dimensional datasets and is particularly suitable for air quality forecasting. ML algorithms can create precise predictive models from training data, making them adaptable to various tasks (Zhu D, 2018). The research aims to investigate the applicability of ML techniques in the operational conditions of air quality monitoring, specifically for predicting the daily peak concentration of a major pollutant using data from monitoring stations.

These forecasting models, incorporating ML techniques, can serve as efficient decision support systems in air quality monitoring centers, enabling proactive regulation enforcement when Air Quality Index (AQI) values exceed acceptable levels, thus preventing potential health risks to urban residents. ML offers researchers and air quality practitioners a powerful tool for management and prediction in this critical field.

2. NEURAL NETWORKS METHODS

The Air pollution is an escalating global concern, notably prevalent in urban centers, stemming from diverse sources like household heating, increased vehicular traffic, energy generation, and industrial operations. The deleterious impacts of pollution on human well-being have prompted intensified scrutiny, thus emphasizing the importance of monitoring and assessing air quality parameters. In response to this challenge, the adoption of Artificial Intelligence (AI) techniques has emerged as an effective approach for modeling intricate, non-linear phenomena.

The environmental dilemma of ambient air quality, especially in metropolitan regions of developing nations, has reached critical proportions. The detrimental health ramifications of pollution have necessitated a shift towards AI-driven methodologies, displacing traditional statistical approaches. Notably, Artificial Neural Networks (ANNs) have emerged as pivotal players in this transformation. ANNs are mathematical constructs, inspired by neural networks in the human brain, capable of learning from experiences and discerning intricate associations and dependencies. Their versatility and utility have been well-established for tasks involving modeling and prediction.

Within the realm of atmospheric sciences, ANNs find a significant application in the prediction of the Air Quality Index (AQI), a widely adopted metric for conveying air pollution severity to the general populace. Numerous methodologies exist for computing the AQI, yet there is no universally applicable approach that accommodates all scenarios. These methods diverge in terms of pollutant considerations, sampling durations, air quality classifications, and breakpoint definitions. Consequently, the application of various AQI techniques to a common case study frequently results in substantial disparities in air quality assessments.

Recent research in the realm of ANNs underscores their ability to discern patterns and grasp intricate relationships, rendering them a potent tool for predictive tasks. ANNs offer particular appeal owing to their prowess in addressing problems that hinge on extensive but not rigidly defined knowledge and vast datasets. Their proficiency in generalization, signifying their capability to extrapolate predictions beyond the training data, even amidst noise in the dataset, is a notable asset. Furthermore, ANNs, functioning as universal function approximators, can model continuous functions with remarkable precision, rendering them suitable for a diverse array of forecasting applications.

Numerous investigations have underscored the promise of employing Artificial Neural Network (ANN)-based models in the realm of air quality prediction. For instance, Huang (2021) devised a model for urban atmospheric air quality forecasting that amalgamated data mining with Backpropagation Neural Networks (BPN). This innovative hybrid approach yielded heightened predictive precision, thereby offering robust decision support for environmental protection agencies. Tamas et al. (2019) harnessed a hybrid model that combined Multilayer Perceptron (MLP) and clustering to predict pollutant concentrations, delivering enhanced peak detection capabilities. Osowski and Garanty (2007) introduced a hybrid model integrating ANN and regression techniques for daily air pollution prediction, with neural networks of SVM lineage enhancing forecasting accuracy. Agirre-Basurko et al. (2006) harnessed MLP-based models for real-time forecasts of O3 and NO2, showcasing superior performance compared to regression models. Aishwarya & Babu (2017) accentuated the potential of ANN models, particularly when coupled with methodologies like Genetic Algorithms (GA), to bolster predictive accuracy for both univariate and multivariate datasets. Ghazali & Hakim Ismail (2012) utilized a straightforward feed-forward neural network to predict air quality, achieving a reasonable level of performance. Rahman et al. (2013) conducted a comparative analysis of ANN, ARIMA, and Fuzzy Time Series (FTS) for Air Pollution Index (API) prediction, with ANNs registering the smallest forecasting error. Al-Alawi et al. (2008) showcased a hybrid model that integrated Multiple Linear Regression (MLR), Principal Component Analysis (PCA), and ANN to predict lower atmosphere ozone concentrations, culminating in enhanced accuracy.

Collectively, these investigations underscore the potential of ANNs in the domain of air quality prediction, particularly concerning individual pollutants and the Air Quality Index (AQI). ANNs exhibit an aptitude for modeling intricate relationships and capturing non-linear patterns, rendering them a valuable tool for addressing the multifaceted challenges associated with air quality forecasting .

METHODOLOGY

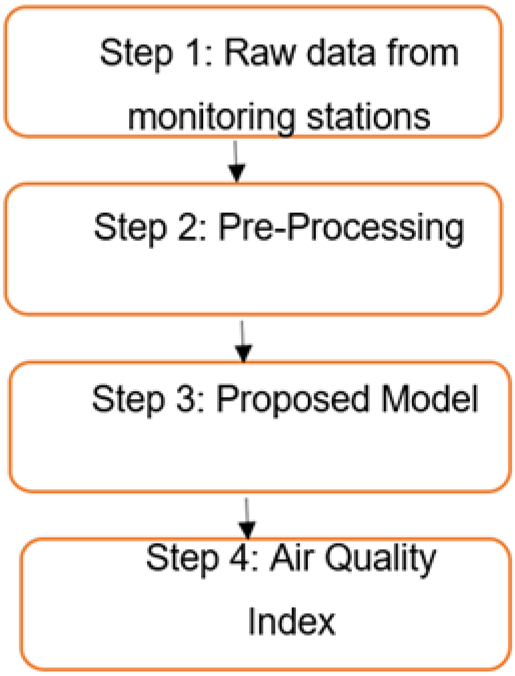
Step 1: Raw data from monitoring stations
Step 2: Pre-Processing
Step 3: Proposed Model
Step 4: Air Quality Index

Fig General methodology for models

STEP 1:

The requisite data is retrieved from the official websites of the pollution control board or environmental agency. This dataset comprises hourly measurements of pollutant concentration, AQI values, and meteorological data. Employing the specified concentration thresholds for each pollutant, outliers are identified and subsequently eliminated from the dataset.

STEP 2:

This phase of the study encompasses the identification of principal pollutants and the initial data processing. The primary pollutants, as designated by the CPCB or EPA for a specific geographical area, were defined and selected for analysis. It is important to note that we encountered a challenge related to data consistency, specifically in terms of the availability of data with a higher number of samples (encompassing concentrations of primary pollutants over a greater number of days) and minimal instances of outliers, ideally none. Upon acquiring the dataset from the CPCB or EPA, we proceeded with data preprocessing. This preprocessing procedure entailed data normalization, which was implemented based on the particular modeling approach under consideration. For the regression algorithms employing Multiple Linear Regression (MLR), Mean normalization was applied, while Max-Min normalization was employed for other models. The Max-Min normalization process adhered to the formulation detailed in Equation 3.11.

$$Xn= \quad (Xi-X\ min/(X\ max-X\ min)$$

STEP 3:

Here, Xi represents the ith data sample, Xmin denotes the minimum value in the dataset, and Xmax signifies the maximum value within the dataset.

Following the pre-processing stage, the pertinent input variables, encompassing pollutant concentrations and historical data, are thoughtfully selected for the predictive modeling. Subsequently, the proposed models are trained using these chosen input predictors and an appropriate number of samples derived from the dataset. The trained models are then put to the test using the remaining dataset samples for evaluation.samples.

STEP 4:

Following the rigorous training and comprehensive testing of the proposed models using a rich dataset, the anticipated outcomes encompass the forecasting of Air Quality Index (AQI) and pollutant concentrations. Subsequently, an exhaustive evaluation of the model's efficacy transpires, leveraging a suite of performance metrics including the coefficient of determination (R), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Index of Agreement (IA). This investigation upholds the standards of scientific inquiry and analysis, thus advancing our understanding in the realm of air quality modeling.

3 METHODOLOGY

In order to predict the Air Quality Index (AQI) for a particular city, we introduced Artificial Neural Network (ANN) predictors, trained utilizing the Conjugate Gradient Descent method, as pioneered by Hestenes in 1952. Our approach encompasses the implementation of multiple neural network models, including Multilayer Perceptron (MLP), Elman, Radial Basis Function Network (RBFN), and Nonlinear Auto-Regressive with Exogenous Inputs (NARX). These models were trained using data spanning the years 2014 to 2016 in Delhi, India, incorporating four pivotal pollutant levels: NO2, CO, O3, and PM10. To provide a visual representation of the forecasting learning methods.

4 CONJUGATE GRADIENT DESCENT

Within the realm of neural networks, the fundamental backpropagation algorithm is harnessed to facilitate the adjustment of weights. It ensures that the weights align with the steepest gradient descent direction, effectively negating the gradient. This direction, marked by a rapid reduction in the objective function, offers the most direct route for determining a new direction vector. However, it's imperative to recognize that, even though the function exhibits rapid decrease in the negative gradient direction, convergence can be rather sluggish. To counteract this challenge and expedite the convergence process, a line search technique, originally introduced by Box in 1969 and subsequently refined by the work of Arora and Snyman in 2006, is applied. This technique operates along conjugate directions, surpassing the efficacy of steepest gradient descent.

Conjugate Gradient Descent (CGD), as introduced by Fletcher and Reeves in 1964, stands as a well-regarded iterative technique that incorporates adaptive learning rates or step sizes, which are determined based on the conjugate direction following each iteration of the training process. This approach ensures consistent progress toward the solution for each step, aligning with the new direction indicated by gradients. Remarkably, CGD can function as a direct method, providing exact solutions after a series of iterations. Additionally, it operates as an iterative method, delivering progressively improving approximations to the exact solution. This adaptability can lead to the attainment of the desired tolerance level within a relatively small number of iterations, especially when compared to the overall problem size.

5. NEURAL NETWORKS

Artificial Neural Networks (ANN) stand out as the predominant machine learning algorithms employed for time series forecasting, as substantiated by the work of Zhang et al. in 1998. These models comprise an assembly of neurons interconnected within the neural network, fostering intricate connectivity.

Within these models, distinct layers come into play, featuring input and output layers as the visible interfaces. Nestled between these two prominent layers is the hidden layer, an integral component. It's noteworthy that the input layer invariably assumes the role of the first layer, with the output layer culminating the sequence. An intriguing facet of neural network predictors lies in the direct relationship between their efficiency and complexity, particularly concerning the quantity of neurons in the hidden layer. An increase in the number of neurons in the hidden layer invariably enhances both the efficacy and intricacy of the neural network predictor.

5.1 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) stands as the foundational artificial neural network, serving as a critical component of many neural network architectures. It operates as a feedforward model, effectively mapping inputs to appropriate outputs. In the MLP structure, layers are interlinked, and the connections between them are regulated by parameters or weights, as depicted in the fundamental architecture presented in Figure 5.2. In the context of this study, the Conjugate Gradient Descent method is applied to determine the weights governing the connections between neurons within each layer. The learning process within each neuron involves the adjustment of weights following the processing of each input sample. This adjustment is contingent on the magnitude of the error generated in the output layer concerning the anticipated outcome, aligning with the work of Xu et al. in 2002. Each set of pollutant levels, encompassing NO2, CO, O3, and PM10, constitutes an input to the MLP, with activation values from each input layer neuron being propagated to all neurons within the hidden layer. In turn, the neuron within the output layer draws upon the combined activations of the hidden layer neurons to calculate the Air Quality Index (AQI). Notably, the activation function employed at each neuron adheres to a logistic function, guided by the governing equation:

The errors cost function is:

where the variables k t and k y represent observed and predicted value of AQI respectively

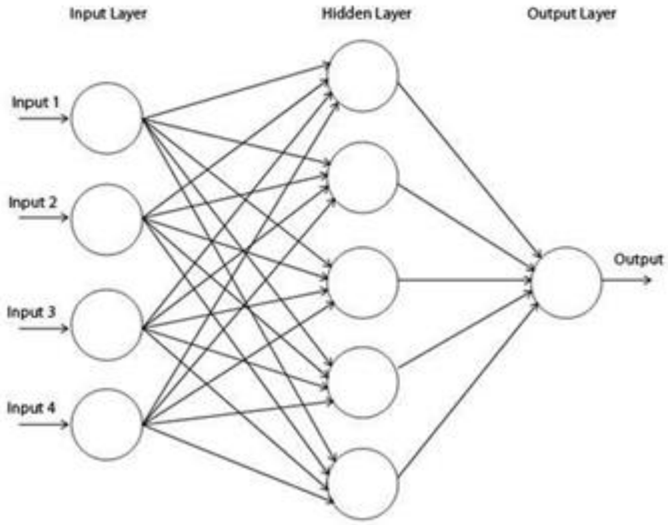

Fig. 5.2 Multilayer perceptron neural network

5.2 Performance Evaluation

In order to assess the effectiveness of each neural predictor, a set of statistical criteria has been selected, encompassing Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), the Correlation Coefficient (R), Root Mean Square Error (RMSE), and the Index of Agreement (IA). These evaluation metrics are defined by the following equations:

$$MAE = \frac{\sum_{k=1}^{n} | t_k - y_k |}{n}$$

where n , "n" signifies the total count of data points, "k_y" represents the predicted value, and "k_t" designates the observed value.

$$MAPE = \frac{\sum_{k=1}^{n} | \frac{t_k - y_k}{t_k} |}{n} \times 100\%$$

Within the framework of these equations, "n" corresponds to the total number of data points, while "k_y" denotes the predicted value, and "k_t" signifies the observed value

In this context, "n" represents the total count of data points, "k_y" designates the predicted value, and "k_t" signifies the observed value.

$$R = \frac{\sum_{k=1}^{n}(t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(t_k - \bar{t})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

where n , In this context, "n" refers to the total count of data points, "k_y" denotes the predicted value, and "k_t" signifies the observed value. Furthermore, "t" represents the average of the observed data, while "y" corresponds to the average of the predicted data.

$$IA = 1 - \frac{\sum_{k=1}^{n}(t_k - y_k)^2}{\sum_{k=1}^{n}(| t_k - \bar{t} | + | y_k - \bar{t} |)^2}$$

In this context, "n" represents the total count of data points, "k_y" designates the predicted value, and "k_t" signifies the observed value. Additionally, "t" corresponds to the average of the observed data.

The assessment of the testing data holds significance as it serves as a gauge of accuracy for each neural network predictor. In pursuit of the optimal model, the most favourable outcomes are characterized by the lowest MAE, MAPE, RMSE values, and the highest R and IA values.

6. RESULTS & COMPARISON

6.1 Performance of MLP

The initial neural network model utilized in this study for forecasting the Air Quality Index (AQI) is the Multilayer Perceptron (MLP). The specific architecture adopted for the MLP in this research is configured as 4-4-1, encompassing the input, hidden, and output layers. The network is subjected to 500 epochs of training, facilitated by the Conjugate Gradient Descent method, leading to the determination of optimal weights.

The training outcomes of the MLP neural network model concerning AQI prediction in Delhi are illustrated in Figure 6.1. Meanwhile, the test results for the MLP neural network model's performance in forecasting Delhi's AQI are presented in Figure 6.2. The regression plots, showcasing the data samples for both the training and testing phases, are depicted in Figure 6.3 and Figure 6.4, respectively.

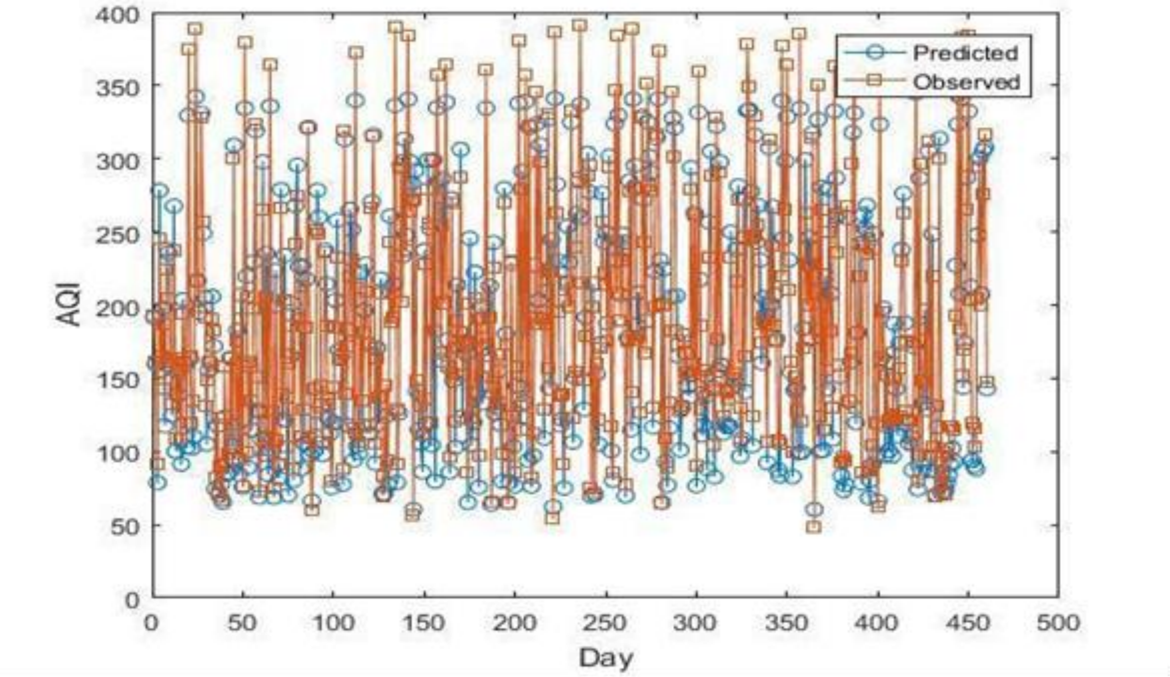

Fig. 6.1 Observed and predicted AQI training samples of MLP

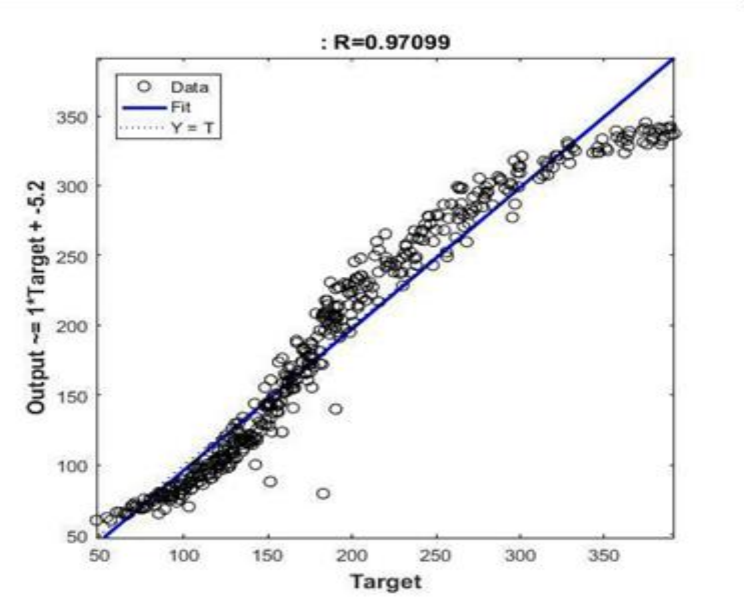

Fig. 6.3 Regression plot of MLP training samples

The regression plots, as depicted in Figure 6.5, exemplify the connection between the observed and predicted values as generated by the MLP neural network model. These plots vividly demonstrate that both the training and testing datasets exhibit a commendable alignment. The performance metrics associated with the MLP model are computed as follows: MAE = 17.72, MAPE = 9.73%, R = 0.955, RMSE = 25.08, and IA = 0.97.

Table 6.1 AQI Performance comparison of neural network models

Model	MAE	MAPE	R	RMSE	IA
MLP	17.72	9.73	0.955	25.08	0.99
Elman	14.26	6.79	0.975	18.85	0.98
RBF	7.33	4.05	0.993	9.69	0.99
NARX	16.57	9.47	0.956	23.80	0.97

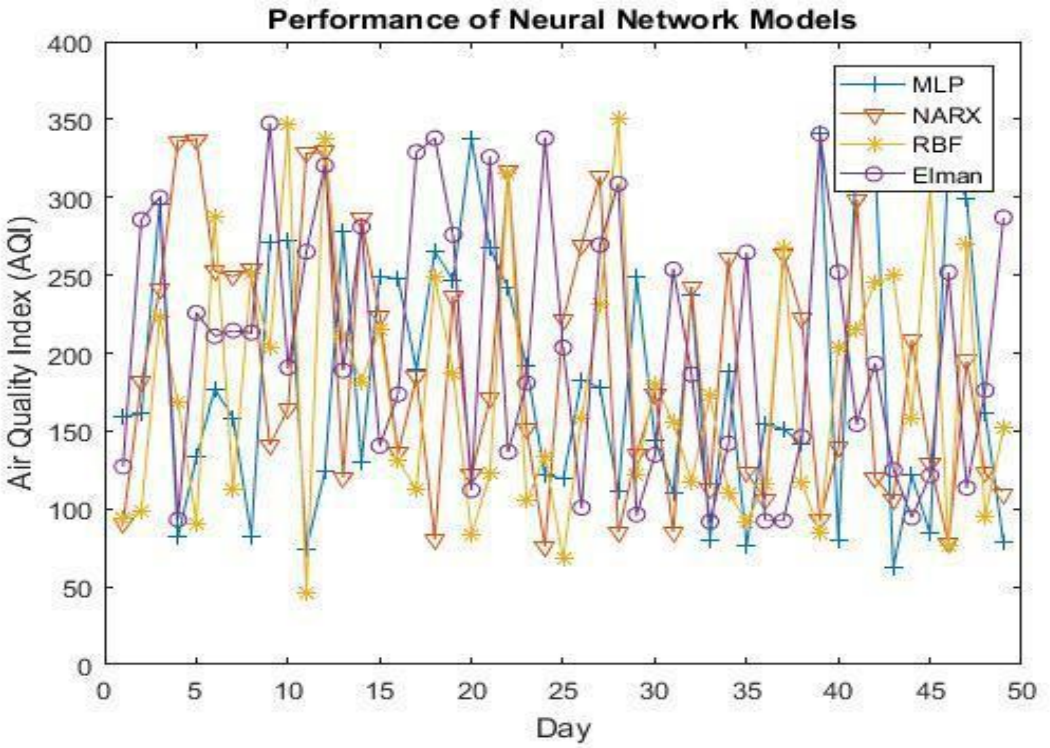

Fig. 6.5 Performance comparison of various network models

7. CONCLUSION

Air Quality Index (AQI) forecasting plays a crucial role in raising awareness among individuals and environmental agencies regarding the potential harm associated with poor air quality. In the course of this study, several neural network models, including Multilayer Perceptron (MLP), Elman's Recurrent Neural Network, Radial Basis Function Network (RBFN), and Nonlinear AutoRegressive with eXogenous (NARX) models, were employed to predict AQI. The comparative efficiency of these models is summarized in Table 6.1.

Notably, the RBF neural network model outperformed the others, delivering more accurate AQI predictions. It achieved impressive performance metrics, with MAE = 7.33, MAPE = 4.05%, R = 0.993, RMSE = 9.69, and IA = 0.99. These findings are illustrated in Figure 5.22, showcasing the test results for all the neural network models considered.

This research affirms the effectiveness of Artificial Neural Networks (ANN) as a powerful predictive methodology. Furthermore, it suggests that by harnessing ensemble techniques, despite the associated computational complexity, the efficiency of these neural network models can be substantially enhanced.

REFERENCES

[1] Agirre-Basurko, E., Ibarra-Berastegi, G., and Madariaga, I. (2018), 'Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area', Environmental Modelling and Software 21(4), 430-446.

[2] Aishwarya, D. C., and Babu, C. N. (2017), Prediction of Time Series Data Using GA-BPNN Based Hybrid ANN Model, 'IEEE 7th International Advance Computing Conference (IACC)' pp. 848-853.

[3] Al-Alawi, S. M., Abdul-Wahab, S. A., and Bakheit, C. S. (2018), 'Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone', Environmental Modelling and Software 23(4), 396-403.

[4] Antanasijević, D. Z., Pocajt, V. V., Povrenović, D. S., Ristić, M. Đ., and Perić-Grujić, A. A., (2013), 'PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization', Science of the Total Environment 443, 511-519.

[5] Arora, J. S., Snyman J.A. (2019), 'Practical Mathematical Optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms', Structural and Multidisciplinary Optimization 31(3), 249-249.

[6] Ashish, M., and Rashmi, B. (2018), 'Prediction of daily air pollution using wavelet decomposition and adaptive-network-based fuzzy inference system', International Journal of Environmental Sciences 2(1), 185.

[7] Athanasiadis, I. N., Kaburlasos, V. G., Mitkas, P. A., and Petridis, V. (2023), Applying machine learning techniques on air quality data for real-time decision support, 'First international NAISO symposium on information technologies in environmental engineering (ITEE'2003)'.

[8] Bai, L., Wang, J., Ma, X., and Lu, H. (2018), 'Air pollution forecasts: An overview', International journal of environmental research and public health 15(4), 780.

[9] Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J. and Sabde, Y. (2019), 'The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017'. The Lancet Planetary Health 3(1), e26-e39.

[10] Bellinger, C., Jabbar, M. S. M., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17(1), 907.

[11] Bělohlávek, R., Klir, G. J., Lewis, H. W., and Way, E. (2012), 'On the capability of fuzzy set theory to represent concepts', International Journal of General Systems 31(6), 569-585.

[12] Bottou, L. (1998), ‘Online learning and stochastic approximations’, On-line learning in neural networks 17(9), 142.

[13] Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., and Vitabile, S. (2007), ‘Two-day ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo’, Atmospheric Environment 41(14), 2967-2995.

[14] Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., and Coggins, J. (2018), ‘Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter’, Proceedings of the National Academy of Sciences 115(38), 9592-9597.

[15] Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. (2012), ‘Sample size selection in optimization methods for machine learning’, Mathematical programming 134(1), 127-155.

[16] Cai, S., Wang, Y., Zhao, B., Wang, S., Chang, X., and Hao, J. (2017), ‘The impact of the “air pollution prevention and control action plan” on PM2. 5 concentrations in Jing-Jin-Ji region during 2012–2020’, Science of the Total Environment 580, 197-209.

[17] Cascio, W. E., and Long, T. C. (2018), ‘Ambient Air Quality and Cardiovascular Health Translation of Environmental Research for Public Health and Clinical Care’, North Carolina medical journal 79(5), 306-312.

[18] Chaloulakou, A., Saisana, M., and Spyrellis, N. (2003), ‘Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens’, Science of the Total Environment 313(1-3), 1-13.

[19] Chang, K. W., and Roth, D. (2011), Selective block minimization for faster convergence of limited memory large-scale linear models, ‘Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining’, pp. 699-707.

[20] Chong, E. K., and Zak, S. H. (2013), ‘An introduction to optimization’, Vol. 76 John Wiley and Sons.

[21] Cobourn, W. G. (2010), ‘An enhanced PM 2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations’, Atmospheric Environment 44(25), 3015-3023.