**Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine**

**Abstract:**

Air quality has always been one of the most important environmental concerns for the general public and society. Using machine learning algorithms for Air Quality Index (AQI) prediction is helpful for the analysis of future air quality trends from a macro perspective. When conventionally using a single machine learning model to predict air quality, it is challenging to achieve a good prediction outcome under various AQI fluctuation trends. In order to effectively address this problem, a genetic algorithm-based improved extreme learning machine (GA-KELM) prediction method is enhanced. First, a kernel method is introduced to produce the kernel matrix which replaces the output matrix of the hidden layer. To address the issue of the conventional limit learning machine where the number of hidden nodes and the random generation of thresholds and weights lead to the degradation of the network learning ability, a genetic algorithm is then used to optimize the number of hidden nodes and layers of the kernel limit learning machine. The thresholds, the weights, and the root mean square error are used to define the fitness function. Finally, the least squares method is applied to compute the output weights of the model. Genetic algorithms are able to find the optimal solution in the search space and gradually improve the performance of the model through an iterative optimization process. In order to verify the predictive ability of GA-KELM, based on the collected basic data of long-term air quality forecast at a monitoring point in a city in China, the optimized kernel extreme learning machine is applied to predict air quality (SO2, NO2, PM10, CO, O3, PM2.5 concentration and AQI), with comparative experiments based CMAQ (Community Multiscale Air Quality), SVM (Support Vector Machines) and DBN-BP (Deep Belief Networks with Back-Propagation). The results show that the proposed model trains faster and makes more accurate predictions.

**Keywords - Machine Learning, Regression, Prediction**

**INTRODUCTION**

Air pollution is a prevalent environmental problem in the twenty-first century. In light of the rapid industrialization and urbanization, air pollution is getting worse, which greatly affects our living environment and health [1]. Li et al. came to the conclusion that outdoor physical activity poses numerous health risks due to ambient air pollution in China. [2], [3]. According to the Chinese Ambient Air Quality Standards (GB3095-2012), there are six conventional air pollutants used to measure air quality: sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter with a particle size less than 10 microns (PM10), particulate matter with a particle size less than 2.5 microns (PM2.5), ozone (O3), and carbon monoxide (CO) [4], [5], [6]. These pollutants have adverse effects on human health. The International Energy Agency estimates that air pollution causes 6.5 million premature deaths per year, while long-term exposure to pollutants, such as fine particles (e.g.,PM2.5) or traffic-related pollutants, is linked to higher rates of lung cancer, coronary heart disease, and other illnesses [7], [8]. Therefore, studies on air quality prediction are particularly important and are considered a key factor for environmental protection. In order to more comprehensively assess the health effects of air pollution, numerous air quality monitoring stations have been set up in major cities. Air quality predictions can be made based on the data collected from these stations. Air quality monitoring, modeling, and accurate predictions are important for having a clear understanding of future pollution levels and their associated health risks. Recently, the inherent property of machine learning algorithms to automatically learn features at multiple levels of abstraction has become increasingly important in providing solutions to this challenging task [9], [10]. However, the model only forecasts PM10 and SO2 levels, and it is also challenging to obtain measurement values needed to construct the dataset [11]. Wu Q. et al. proposed an optimal-hybrid model for daily AQI prediction considering air pollutant factors, with the model’s inputs being the six atmospheric pollutants. However, neural networks typically struggle with slow learning, a tendency to fall into local minima, and a complex network training process. Based on the generalized inverse matrix theory, Huang et al. proposed an extreme learning machine (ELM) algorithm with a feedforward neural network that includes a single hidden layer, such that the problems of conventional neural network algorithms are circumvented. The ELM algorithm used to predict the AQI outperformed neural networks in terms of parameter selection, training speed, and prediction accuracy [12]. However, the parameters of the hidden layer nodes and the number of nodes in the test hidden layer are selected at random, which puts the prediction accuracy to a great test. In order to solve the aforementioned problems, we propose to optimize the number of ELM hidden layer nodes, thresholds, and weights, along with an improved genetic algorithm (GA) that uses root mean square error (RMSE) as the fitness function, to obtain the optimal network structure for air quality prediction [14]. The number of hidden layer nodes is updated by continuous coding discretization, the input weights and hidden layer thresholds are updated by continuous coding, and the update thresholds and weights are selected with the number of updated layers to form a hierarchical control structure [15]. The proposed GA-based improved extreme learning machine (GA-KELM) algorithm is applied to air quality prediction, and its performance is compared with that of community multiscale air quality modeling system (CMAQ), support vector regression (SVR), and deep belief network-back propagation (DBN-BP). The results show that the accuracy of the proposed GA-KELM algorithm is reliable for air quality prediction [16]. In this study, an improved extreme learning machine model based on a genetic algorithm is designed and applied to AQI prediction. To verify the effectiveness of the model, we conducted tests on three real-world datasets. The results confirmed that the proposed method has superior performance and outperforms some advanced methods currently in use. The main contributions of this paper are: (1) modifying the ELM activation function or using the kernel function to improve the prediction accuracy, (2) optimizing the ELM using GA to improve the stability of the results and further enhance the prediction accuracy, and (3) obtaining the correlation analysis results of atmospheric environmental quality prediction parameters by comprehensively considering each relevant factor in line with the actual situation. The remainder of this paper is organized as follows. Section II presents related work. Section III describes ELM and the proposed GA-KELM, and illustrates the improvements using the model. Section IV discusses experimental results where GA-KELM is compared with several other methods in terms of prediction results. The last section concludes the entire work and presents directions for future research.

**LITERATURE REVIEW**

Air quality prediction has been extensively researched in the literature [17]. In recent years, numerous researchers have made significant contributions to the field by leveraging quantitative studies and the latest techniques to identify various air quality patterns and their underlying trends [18]. Existing work in this area relies on statistical methods and shallow machine learning models to address the problem of air quality prediction [19]. Agarwal and Sahu [20] conducted air quality prediction studies by employing statistical models. Lary et al. [21] combined remote sensing and meteorological data with groundbased PM2.5 observations. Zheng et al. [22] proposed a hybrid prediction method that combines a linear regression- based temporal prediction method with an ANN-based spatial prediction method for pollutant concentrations. Zheng et al. [23] used a data-based approach for the next 48 hours of PM2.5 prediction, implementing a prediction model based on linear regression and neural network. They combined meteorological data, weather forecast data, and air quality data from monitoring stations. Rajput and Sharma [24] used a multiple regression model to represent the changes in air quality index (AQI), considering ambient temperature, relative humidity, and barometric pressure as the main parameters in the regression model for AQI calculation [25]. These classical methods and models all have the advantages of simple algorithms, easy processing, and acceptable prediction results. However, obtaining precise and specific air quality prediction values remains challenging [26]. Elbaz et. al. [27] proposed a novel deep learning approach that extracts high-level abstractions to capture the spatiotemporal characteristics of NEOM city in Saudi Arabia at hourly and daily intervals. Campbell et al. [28] described the development of FV3GFSv16 coupled with the ‘‘stateof-the-art’’ CMAQ model version 5.3.1. Jin et al. [29] proposed an interpretable variational Bayesian deep learning model with self-filtering capability for PM2.5 prediction information, which effectively improves prediction accuracy. Zhou et al. [30], [31], [32] proposed a method based on an improved Grasshopper optimization algorithm to classify the color difference of dyed fabrics using kernel extreme learning machine. In this study, the classification of color differences in dyed fabric images is performed using the kernel limit learning machine, and the kernel function parameters are optimized by the improved Grasshopper optimization algorithm to achieve color difference classification of dyed fabric images. Xue et al. [33] proposed a GA-based air quality prediction model to optimize the parameters of the weighted extreme learning machine (WELM). Despite the progress made by the aforementioned methods, they also exhibit limitations; their training efficiency is relatively low, and deep learning algorithms are not yet fully mature. These challenges present greater obstacles for the application of deep learning, necessitating improvements to existing models, the development of new models, and the enhancement of their predictive capabilities [34], [35]. The use of statistical or numerical forecasting techniques is subject to several limitations. Neural networks are widely used because of their unique associative abilities, memory, and distinctive learning [36], [37]. Given the highly nonlinear nature of AQI changes and the strong generalization and nonlinear characterization abilities of neural networks, the nuclear limit learning machine neural network model, also known as kernel extreme learning machine (KELM), is employed to investigate air quality prediction using a real dataset. The weights and threshold values of KELM are optimized using a genetic optimization algorithm [38].

### Existing System and Its Disadvantages

\*\*Existing System:\*\*

- The existing system typically uses a single machine learning model, such as traditional neural networks, for predicting air quality.

- Some of the prominent models referenced include the Community Multiscale Air Quality (CMAQ) modeling system, Support Vector Regression (SVR), and Deep Belief Network with Back-Propagation (DBN-BP).

\*\*Disadvantages:\*\*

1. \*\*Single Model Limitations:\*\* Using a single machine learning model often results in suboptimal performance due to the varied and fluctuating nature of AQI data.

2. \*\*Slow Learning:\*\* Traditional neural networks can suffer from slow learning processes.

3. \*\*Local Minima:\*\* Neural networks tend to get stuck in local minima, which affects prediction accuracy.

4. \*\*Complex Training Process:\*\* Training neural networks can be complex and time-consuming.

5. \*\*Random Parameter Selection:\*\* For methods like Extreme Learning Machine (ELM), the number of hidden nodes and the random generation of thresholds and weights lead to a degradation in the network's learning ability and prediction accuracy.

### Proposed System and Its Advantages

\*\*Proposed System:\*\*

- The proposed system is a Genetic Algorithm-based Improved Extreme Learning Machine (GA-KELM).

- This system introduces a kernel method to generate a kernel matrix that replaces the output matrix of the hidden layer in the ELM.

- A genetic algorithm (GA) is used to optimize the number of hidden nodes, the thresholds, and the weights of the ELM.

- The Root Mean Square Error (RMSE) is employed as the fitness function for the GA.

- The Least Squares Method is applied to compute the output weights of the model.

\*\*Advantages:\*\*

1. \*\*Improved Prediction Accuracy:\*\* The use of a kernel method in ELM and optimization via GA enhances the prediction accuracy by addressing the random parameter selection issue.

2. \*\*Optimized Network Structure:\*\* GA optimizes the number of hidden nodes and layers, as well as the thresholds and weights, leading to a more effective and efficient learning process.

3. \*\*Faster Training:\*\* The proposed GA-KELM trains faster compared to traditional methods.

4. \*\*Enhanced Stability:\*\* The optimization process via GA improves the stability of the results, reducing variability and enhancing reliability.

5. \*\*Iterative Improvement:\*\* GA iteratively finds the optimal solution in the search space, progressively improving model performance.

6. \*\*Comparative Superiority:\*\* Experiments demonstrate that GA-KELM outperforms existing models like CMAQ, SVR, and DBN-BP in terms of accuracy and training speed.

**CONCLUSION :**

The economic development achieved by the country through rapid urbanization is polluting the environment in an alarming way and putting people’s lives in danger. Therefore, a correct analysis and accurate prediction of air quality remains a primary condition to achieve the objective of sustainable development. This paper focuses on the problem of prediction model design, and investigates the problems related to the optimization of the model parameters. A GA-KELM model is designed, implemented, and tested. It is experimentally proven to be more efficient than the classical shallow learning and can effectively explore and learn the interdependence of multivariate air quality correlation time series such as temperature, humidity, wind speed, SO2, and PM10. Therefore, the GA-KELM model developed in this study can be used to provide valuable support to vulnerable groups and trigger early warning of adverse air quality events. However, there are still areas for further investigation and improvement. In recent years, numerous advanced algorithms and optimization methods based on genetic algorithms and population intelligence have emerged. Therefore, future research should explore the underlying significance and value of combinatorial intelligence optimization algorithms such as the Limit Learning Machine. Additionally, we acknowledge the need to address the issue of manually setting the number of hidden layer nodes in the optimal Limit Learning Machine. Although the Dynamic Extreme Learning Machine (DELM) algorithm offers adaptive determination of hidden layer nodes without human intervention, further work should be dedicated to this aspect. Moreover, to enhance the accuracy and validity of air quality measurement and assessment, it is crucial to integrate pollutant emission factors and meteorological factors into the evaluation system. This integration will enable a more precise and comprehensive evaluation of air quality. In conclusion, our study highlights the significance of the GA-KELM model in predicting air quality. We have addressed the optimization challenges and demonstrated its superiority over traditional methods. However, there is still room for improvement and further research. Future studies should delve into advanced optimization algorithms based on genetic algorithms and population intelligence, explore the potential of the Limit Learning Machine, and strive for adaptive determination of hidden layer nodes. Furthermore, the integration of pollutant emission factors and meteorological factors into the evaluation system will advance the accuracy and reliability of air quality measurement and assessment.

**REFERENCES**:

[1] X. Li, L. Jin, and H. Kan, ‘‘Air pollution: A global problem needs local fixes,’’ Nature, vol. 570, no. 7762, pp. 437–439, Jun. 2019.

[2] Y. Han, J. C. K. Lam, and V. O. K. Li, ‘‘A Bayesian LSTM model to evaluate the effects of air pollution control regulations in China,’’ in Proc. IEEE Big Data Workshop (Big Data), Dec. 2018, pp. 4465–4468.

[3] L. Bai, J. Wang, X. Ma, and H. Lu, ‘‘Air pollution forecasts: An overview,’’ Int. J. Environ. Res. Public Health, vol. 15, no. 4, p. 780, 2018.

[4] Y. Ding and Y. Xue, ‘‘A deep learning approach to writer identification using inertial sensor data of air-handwriting,’’ IEICE Trans. Inf. Syst., vol. E102-D, no. 10, pp. 2059–2063, 2019. [5] S.-Q. Dotse, M. I. Petra, L. Dagar, and L. C. De Silva, ‘‘Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam,’’ Atmos. Pollut. Res., vol. 9, no. 2, pp. 358–368, Mar. 2018. [6] M. Jia, A. Komeily, Y. Wang, and R. S. Srinivasan, ‘‘Adopting Internet of Things for the development of smart buildings: A review of enabling technologies and applications,’’ Automat. Construct., vol. 101, pp. 111–126, May 2019. [7] S. Abirami, P. Chitra, R. Madhumitha, and S. R. Kesavan, ‘‘Hybrid spatio-temporal deep learning framework for particulate matter (PM2.5) concentration forecasting,’’ in Proc. Int. Conf. Innov. Trends Inf. Technol. (ICITIIT), Feb. 2020, pp. 1–6. [8] Y. Cheng, S. Zhang, C. Huan, M. O. Oladokun, and Z. Lin, ‘‘Optimization on fresh outdoor air ratio of air conditioning system with stratum ventilation for both targeted indoor air quality and maximal energy saving,’’ Building Environ., vol. 147, pp. 11–22, Jan. 2019. [9] A. C. Cosma and R. Simha, ‘‘Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions,’’ Building Environ., vol. 148, pp. 372–383, Jan. 2019. [10] M. Bhowmik, K. Deb, A. Debnath, and B. Saha, ‘‘Mixed phase Fe2O3/Mn3O4 magnetic nanocomposite for enhanced adsorption of methyl orange dye: Neural network modeling and response surface methodology optimization,’’ Appl. Organometallic Chem., vol. 32, no. 3, p. e4186, Mar. 2018. [11] V. Chaudhary, A. Deshbhratar, V. Kumar, and D. Paul, ‘‘Time series based LSTM model to predict air pollutant’s concentration for prominent cities in India,’’ in Proc. Int. Workshop Utility-Driven Mining (UDM), Aug. 2018, pp. 1–9. [12] M. Chen, J. Yang, L. Hu, M. S. Hossain, and G. Muhammad, ‘‘Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators,’’ IEEE Commun. Mag., vol. 56, no. 11, pp. 14–20, Nov. 2018. [13] R. Chen, X. Wang, W. Zhang, X. Zhu, A. Li, and C. Yang, ‘‘A hybrid CNN-LSTM model for typhoon formation forecasting,’’ GeoInformatica, vol. 23, no. 3, pp. 375–396, Jul. 2019. [14] S. Du, T. Li, Y. Yang, and S. Horng, ‘‘Deep air quality forecasting using hybrid deep learning framework,’’ IEEE Trans. Knowl. Data Eng., vol. 33, no. 6, pp. 2412–2424, Jun. 2021. [15] R. Feng, H.-J. Zheng, H. Gao, A.-R. Zhang, C. Huang, J.-X. Zhang, K. Luo, and J.-R. Fan, ‘‘Recurrent neural network and random forest for analysis and accurate forecast of atmospheric pollutants: A case study in Hangzhou, China,’’ J. Cleaner Prod., vol. 231, pp. 1005–1015, Sep. 2019. [16] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, ‘‘Forecasting air quality time series using deep learning,’’ J. Air Waste Manage. Assoc., vol. 68, no. 8, pp. 866–886, Aug. 2018. [17] S. Mahajan, H.-M. Liu, T.-C. Tsai, and L.-J. Chen, ‘‘Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model,’’ IEEE Access, vol. 6, pp. 19193–19204, 2018. [18] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, ‘‘An information framework for creating a smart city through Internet of Things,’’ IEEE Internet Things J., vol. 1, no. 2, pp. 112–121, Apr. 2014. [19] A. Grover, A. Kapoor, and E. Horvitz, ‘‘A deep hybrid model for weather forecasting,’’ in Proc. 21st ACM SIGKDD Int. Conf. Knowl. DiscoveryData Mining, Aug. 2015, p. 379–386. [20] A. Agarwal and M. Sahu, ‘‘Forecasting PM2.5 concentrations using statistical modeling for Bengaluru and Delhi regions,’’ Environ. Monit. Assessment, vol. 195, p. 502, Mar. 2023. [21] D. J. Lary, T. Lary, and B. Sattler, ‘‘Using machine learning to estimate global PM2.5 for environmental health studies,’’ Environ. Health Insights, vol. 9, no. 1, pp. 41–52, 2015. [22] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, ‘‘Forecasting fine-grained air quality based on big data,’’ in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Aug. 2015, pp. 2267–2276. [23] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, ‘‘Urban computing: Concepts, methodologies, and applications,’’ ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, p. 38, Sep. 2014. [24] T. S. Rajput and N. Sharma, ‘‘Multivariate regression analysis of air quality index for Hyderabad city: Forecasting model with hourly frequency,’’ Int. J. Appl. Res., vol. 3, no. 8, pp. 443–447, 2017. [25] Z. Kang and Z. Qu, ‘‘Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou,’’ in Proc. IEEE Comput. Intell. Appl. (ICCIA), Sep. 2017, pp. 155–160, doi: 10.1109/CIAPP.2017.8167199. [26] B. Liu, S. Yan, J. Li, G. Qu, Y. Li, J. Lang, and R. Gu, ‘‘A sequence-tosequence air quality predictor based on the n-step recurrent prediction,’’ IEEE Access, vol. 7, pp. 43331–43345, 2019. [27] K. Elbaz, I. Hoteit, W. M. Shaban, and S.-L. Shen, ‘‘Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM,’’ Chemosphere, vol. 313, Feb. 2022, Art. no. 137636. [28] P. C. Campbell, Y. Tang, P. Lee, B. Baker, D. Tong, R. Saylor, A. Stein, J. Huang, H.-C. Huang, E. Strobach, J. McQueen, L. Pan, I. Stajner, J. Sims, J. Tirado-Delgado, Y. Jung, F. Yang, T. L. Spero, and R. C. Gilliam, ‘‘Development and evaluation of an advanced national air quality forecasting capability using the NOAA Global Forecast System version 16,’’ Geosci. Model Develop., vol. 15, no. 8, pp. 3281–3313, Apr. 2022.