Accurate Air Quality Index Forecasting Using Bi-LSTM Neural Network

Sachu Joshna[1], Ravva Harshitha[2], Addagunta Avinash[3], Dr. Digavinti Sreenivasulu[4]

[1]B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[2] B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[3] B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[4]Associate Professor, Dept. of CSE(DS), Institute of Aeronautical Engineering, Telangana, India

---

Abstract - *An Assurance of human wellbeing and bearing of ecological approach rely basically upon air quality prediction. Changes in air quality are hard to predict for most standard single-model frameworks. This paper provides areas of strength for a framework that makes use of cutting-edge machine learning techniques. From a close viewpoint, we investigate various models such as Support Vector Regression (SVR), Deep Belief Network with Back-Propagation (DBN-BP), and Genetic Algorithm-Enhanced Extreme Learning Machine (GA-KELM). In addition, we suggest integrating a deep learning architecture known as bidirectional long short-term memory (BiLSTM) to further improve prediction accuracy even more. After extensive evaluation and testing, we demonstrate that BiLSTM exhibits lower Mean Squared Error (MSE) and Root Mean Square Error (RMSE) values, outperforming existing models. Additionally, we enhance BiLSTM's display by using GA-KELM, hence significantly enhancing its predictive capabilities. In addition to providing. Aside from giving better precision in air quality prediction, the recommended hybrid model assists with directing general wellbeing efforts and contamination control arrangements through informed choices. This study underscores the need of researching imaginative ways to deal with handle earnest ecological issues and the conceivable outcomes of ML in further developing air quality control.*

Key Words: genetic algorithms, time series, machine learning, extreme learning machines, and air quality forecasts.

1.INTRODUCTION

Ascending as a significant overall issue in the twenty-first 100 years, air pollution is exasperated by quick industrialization and urbanization [1]. Declining air quality influences general wellbeing as well as the climate [2]. Li et al's. concentrates on feature the wellbeing dangers associated with open air actual practice within the sight of encompassing air contamination, particularly in regions like China that are quick seeing modern advancement [3]. As in numerous different countries, China estimates air quality utilizing rules characterized in the Chinese Surrounding Air Quality Norms.

These pollutants have plainly unfortunate results for human wellbeing [5 Long-term exposure to air pollutants such as PM2.5 and emissions from moving vehicles has been linked to a higher chance of developing lung cancer, heart disease, and other illnesses. According to estimates from the International Energy Agency, air pollution results in almost 6.5 million premature deaths annually. [6]. Subsequently, the advancement of effective frameworks for air quality expectation turns out to be increasingly more significant as ecological insurance drives rely upon this [7]. Prediction of air quality for the most part relies upon data accumulated from checking stations spread all through significant urban communities [8]. These locales guide estimate models and deal savvy investigation of contamination levels. [9] ML calculations have become more powerful devices for assessing such information. In any case, there are challenges like the shortage of careful datasets and the trouble demonstrating numerous foreign substances simultaneously [10].

New investigations have taken a gander at numerous ways of meeting these challenges. Utilizing information from six air impurities [11]. Conventional neural network calculations do, nonetheless, oftentimes go against issues like languid learning, aversion to neighborhood minima, and troublesome training strategies [12].

In view of the lengthy converse lattice hypothesis and with a single hidden layer FNN, Huang et al. introduced ELM way to deal with beat these limitations [13]. As for boundary determination, preparing time, and prediction accuracy the ELM calculation has shown preferred execution in AQI prediction over conventional neural networks [14]. The ELM calculation's dependence on arbitrarily picked boundaries for buried layer hubs presents hardships to prediction accuracy regardless of whether its proficiency.

In this regard, the goal of this work is to improve upon the current models for predicting air quality by proposing a new strategy combining with the benefits of ML computations further developed boundary augmentation techniques. We present explicitly a crossover model consolidating the GA-KELM architecture with the BiLSTM engineering. Utilizing the prescient powers of BiLSTM and hereditary calculation improvement of model boundaries, this blend looks to expand the accuracy and versatility of air quality projections [16].

This work presents another hybrid model that settles the requirements of current strategies, thusly supporting the constant endeavors to further develop air quality forecast methods. Consolidating BiLSTM with GA-KELM will assist us with giving more exact and reliable conjectures, along these lines supporting wise decision-production for general wellbeing the executives and natural security.

2. LITERATURE SURVEY

Globally, air pollution has become a serious threat to the environment and public health concern requiring intensive examination to distinguish its sources, results, and moderating procedures. With an eye on the utilization of ML approaches for air quality prediction, we assess significant exploration on air pollution monitoring, forecasting, and control in this writing study.

Accentuating the need of handling air quality issues at the territorial level, Li et al. (2019) underlined air contamination as an overall concern requiring neighborhood arrangements [1]. From this vantage point, restricted air quality observing and gauge frameworks become significantly more significant in directing centered measures. To evaluate the effect of air contamination the board approaches in China, Han et al. (2018) introduced a Bayesian LSTM model, consequently featuring the worth of modern factual techniques for information examination of air quality [2]. Their examination underlines how well LSTM models might gauge what strategy changes will mean for air quality outcomes.

Bai et al. (2018) examined many displaying procedures and information sources applied in air quality prediction [3] along with a synopsis of air contamination gauges. Their research highlights how difficult it is to predict air quality and how important it is to have a few sources of information. including meteorological information, satellite perceptions, and ground-level checking information. Utilizing inertial sensor information of air penmanship, Ding and Xue (2019) recommended a DL technique for essayist ID, thusly demonstrating the versatility of DL strategies in sensor information examination for some purposes [4].

Exploring the change of outside air proportion in cooling frameworks for arriving at wanted inside air quality and greatest energy reserve funds, Cheng et al. (2019) [5] found Their examination underscores the need of boosting ventilation methods to safeguard improve indoor air quality and use less energy. An LSTM model based on period series was developed by Chaudhary et al. (2018) to forecast air pollution foci in notable Indian metropolitan areas, demonstrating the value of Their work contributes to the growing body of research on information-driven methods for predicting air quality.

In view of publicly supported and cloud-based air quality markers, Chen et al. (2018) proposed a metropolitan medical care big data system featuring the potential outcomes of publicly supporting information for checking metropolitan air quality [7]. Their examination underlines how new innovation could assist with expanding the degree of general wellbeing checking and air quality control. Joining CNNs and LSTM networks [8], their examinations show that hybrid DL calculations catch multifaceted spatiotemporal examples in air quality data actually.

The writing audit calls attention to by and large the rising interest in utilizing ML strategies for the board, forecasting, and air quality monitoring. To expand the accuracy and reliability of air quality figures, research has taken a gander at a different range of strategies having LSTM models, DL structures, hybrid ML designs. These improvements could assist with directing proof-based medicines intended to minimize the damaging effects of air pollution on the environment and public health.

3. METHODOLOGY

3.1 Proposed Work

The study coordinates GA & ELM to further develop air quality forecasting, explicitly PM2.5 levels. GA will upgrade the ELM model's secret hubs and layers to further improve learning and prediction. GA's [14] ability to look for ideal arrangements inside a given hunt region permits the ELM model to adaptively modify its engineering to more readily reflect air quality information's convoluted linkages. MSE and RMSE will be utilized to analyze adequacy versus exemplary methodologies like Support Vector Machines (SVM) [16].

3.2 System Architecture

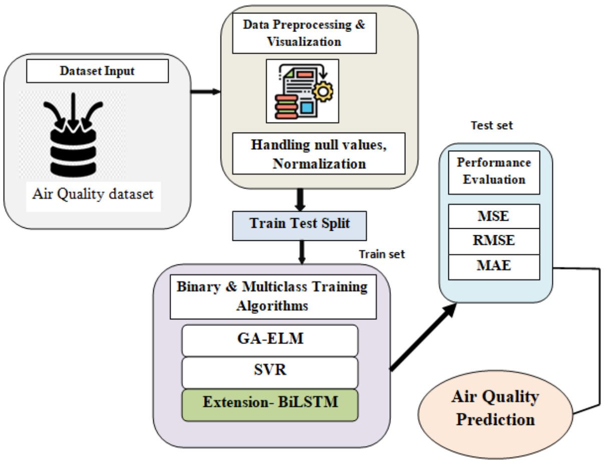

Fig -1: Proposed Architecture

There are several main components to the suggested design for an air quality forecasting system. It starts by entering a dataset on air quality including geographic information, meteorological data, and pollution concentrations. While visualization helps to discover data trends, data processing techniques include normalizing, feature engineering, and handling missing values ready the data for modeling.

Training and testing sets separate the dataset so as to assess the generalizing power of the model. Using algorithms such GA-ELM [14], SVR, and BiLSTM, trends between input characteristics and air quality results are found. Prediction accuracy is measured by means of MSE, RMSE, and MAE, thereby evaluating performance. By means of pragmatic insights, the system projects pollution concentrations and AQI values, therefore supporting public health initiatives.

3.3 Dataset

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	0.0	0.0	0.92	18.22	17.15	0.0	0.92	27.64	133.36	0.00	0.02	0.00	0.0	0
1	Ahmedabad	2015-01-02	0.0	0.0	0.97	15.69	16.46	0.0	0.97	24.55	34.06	3.68	5.50	3.77	0.0	0
2	Ahmedabad	2015-01-03	0.0	0.0	17.40	19.30	29.70	0.0	17.40	29.07	30.70	6.80	16.40	2.25	0.0	0
3	Ahmedabad	2015-01-04	0.0	0.0	1.70	18.48	17.97	0.0	1.70	18.59	36.08	4.43	10.14	1.00	0.0	0
4	Ahmedabad	2015-01-05	0.0	0.0	22.10	21.42	37.76	0.0	22.10	39.33	39.31	7.01	18.89	2.78	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
497	Ahmedabad	2016-05-12	0.0	0.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0
498	Ahmedabad	2016-05-13	0.0	0.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0
499	Ahmedabad	2016-05-14	0.0	0.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0
500	Ahmedabad	2016-05-15	0.0	0.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0
501	Ahmedabad	2016-05-16	0.0	0.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0

Fig -2: Dataset

Estimates of a few toxins and particulate matter with a width of less than 10 millimeters (PM10) are included in the air quality dataset. Toxin fixations, related timestamps, and geographic areas are completely remembered for every perception. Meteorological data may likewise be given, for example, temperature, mugginess, wind speed, and gaseous tension. Research on the impacts of contamination on the climate and general wellbeing is made more straightforward by this dataset, which considers the examination and investigation of changes in air quality over the long run and in different geological regions.

3.4 Data Processing

Preparing datasets for model training depends on data processing leveraging Pandas and Keras Data Frames. Pandas effectively manages missing values by imputing or deleting them, therefore guaranteeing dataset integrity and lowering bias. Usually between 0 and 1, normalizing numerical features to a common scale helps to avoid bigger features controlling the model training process. To reduce dimensionality and improve computing speed, unnecessary columns are deleted. Keras Data Frames provide effective data preparation for neural network designs by elegantly interacting with deep learning frameworks. They also preserve data integrity by handling missing values just like Pandas does. Keras Data Frames maximize resource use and minimize overfitting by normalizing numerical characteristics and eliminating useless columns, therefore guaranteeing improved model convergence and performance.

3.5 Visualization

Seaborn and Matplotlib data visualization improves knowledge of air quality statistics. While scatter plots show relationships between pollutants and meteorological data, histograms emphasize patterns and outliers in pollution concentrations. Showcasing seasonal and long-term trends, line graphs demonstrate pollution levels over time. Analysis of air quality dynamics and the assistance of the creation of efficient prediction models depend on these visuals.

3.6 Feature Selection

Viable air quality prediction models need feature selection. We use relationship examination, include importance positioning, and PCA to diminish dimensionality. By recognizing connections amongst foreign substances and meteorological variables, relationship investigation helps pick significant attributes. Positioned include importance approaches like Random Forest favor persuasive highlights for prediction. PCA additionally finds idle factors that make sense of most information variety, bringing down dimensionality while keeping key data. Include determination further develops air quality forecast model execution and computational proficiency by picking the most enlightening attributes.

3.7 Training & Testing

To evaluate the performance of the model it is necessary to divide the air quality dataset into training and testing subsets. Typically, an 80/20 or 70/30 random split is used to ensure enough information to support testing and training. The instruction set trains the predictive model, and the testing set evaluates the model's execution. This split assesses the model's adaptability to fresh data, facilitating the creation of credible air quality forecasts and an objective performance evaluation.

3.8 Algorithms

Genetic Algorithm (GA) is used by Genetic Algorithm with Extreme Learning Machine (GA-ELM). for parameter optimization to improve Extreme Learning Machine (ELM). GA develops solutions to improve the predictive ability of ELM through fitness-based selection, breeding, and mutation. ELM is effective for demanding environmental data such as air quality forecasting by using in order to compute output weights and map input features into a high-dimensional space, a random activation function is used.

Support Vector Regressor (SVR) generates a regression model by using a hyperplane that maximizes the distance between data points, thereby reducing the error function. SVR can record complex interactions between inputs and outputs, making it very effective in predicting nonlinear trends in air quality data. SVR faithfully reflects complex interactions by optimizing parameters such as kernel type and regularization, improving the overall forecast quality.

BiLSTM: By analyzing time the extended bidirectional long short-term memory (BiLSTM), when given series data in both forward and backward directions, generates: It is especially helpful for comprehending how air quality data changes over time as bidirectional processing allows it to capture both past and future environments. Since long-term dependencies and trends are well handled by BiLSTM, it also allows for more accurate prediction of air quality changes over time.

4. EXPERIMENT RESULTS

Mean squared error (MSE): MSE in statistics checks factual model error. It assesses the normal and noticed values' typical squared distinction. The MSE equivalent 0 when a model is sans mistake. It is worth ascents as model error rises. Furthermore, alluded to as the standard deviation of the mean (MSD).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

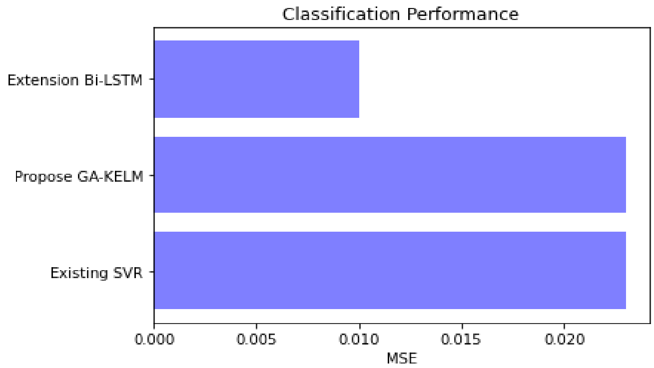

Fig -3: MSE Comparison Graph

Root Mean Square Error (RMSE): This measure assesses the standard deviation between the genuine information and the extended upsides of a measurable model. It is, numerically, the residuals' standard deviation. Residuals are the information point distance from the regression line.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\|y(i) - \hat{y}(i)\|^2}{N}},$$

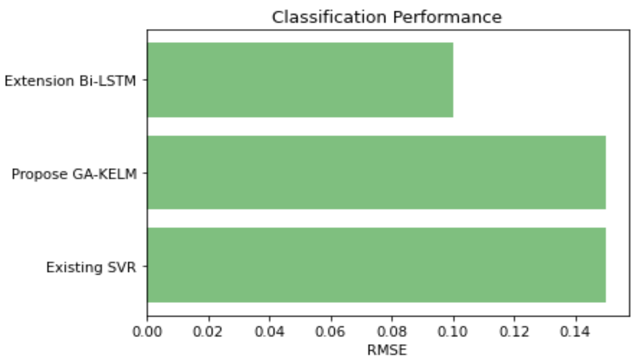

Fig -4: RMSE Comparison Graph

Mean Absolute Error (MAE): The inaccuracy in your estimations is the level of misstep. It is the variety between the "true" and recorded values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

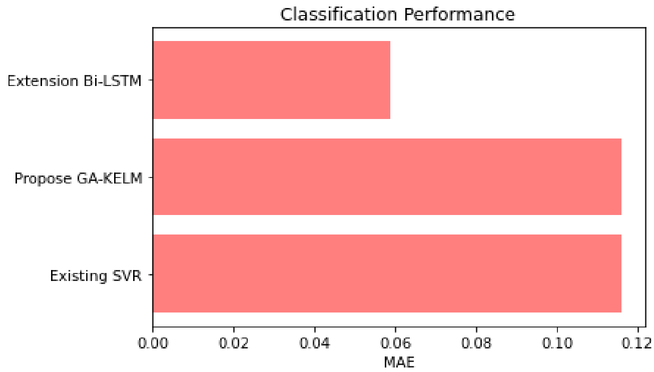

Fig -5: MAE Comparison Graph

	ML Model	MSE	RMSE	MAE
0	Existing SVR	0.023	0.15	0.116
1	Propose GA-KELM	0.023	0.15	0.116
2	Extension Bi-LSTM	0.010	0.10	0.059

Fig -6: Performance Evaluation

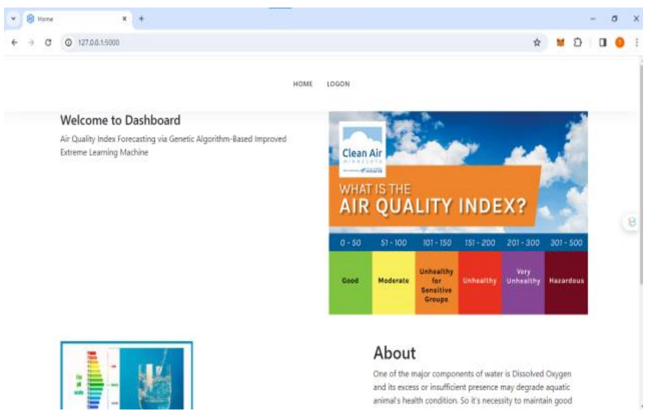

Fig -7: Home Page

Fig -8: Registration Page

Fig -9: Login Page

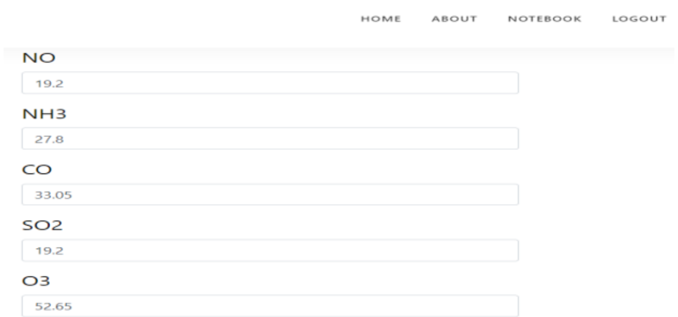

Fig -10: Upload Input Data

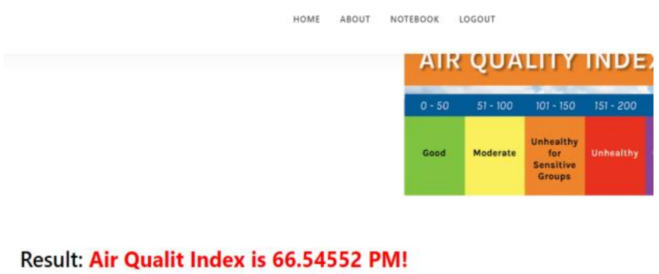

Fig -11: Final Outcome

5. CONCLUSIONS

In summary, the integration of the "Genetic Algorithm with Extreme Learning Machine" (GA-KELM) and the extension of the "Bidirectional Long Short-Term Memory" (BiLSTM) [14] address vital improvements in air quality prediction, giving expanded accuracy and further developing natural administration navigation. The task's impact is additionally expanded by sending the BiLSTM model inside an instinctive Flask framework, which gives general society and scientists the same valuable admittance to air quality forecasts. This not just empowers individuals to go with decisions that are best for their wellbeing and prosperity, however it likewise makes it simpler to make a deterrent move against the harming influence that the environment is affected by air pollution.

6. FUTURE SCOPE

There are numerous forthcoming innovative work potential open doors. Further turn of events and enhancement of the BiLSTM and GA-KELM models should focus on anticipated accuracy and versatility. Incorporation of constant information streams and more enhanced components might work on model capacities. Beyond air quality prediction, applying these models to environment demonstrating or ecological effect evaluations might give critical bits of knowledge. At long last, making air quality forecast advancements, for example, portable applications and online stages, more open and usable will assist with tackling ecological issues. High level computational techniques and their utilization to reduce natural issues and increment worldwide personal satisfaction are what's in store.

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.

[29] X.-B. Jin, Z.-Y. Wang, W.-T. Gong, J.-L. Kong, Y.-T. Bai, T.-L. Su, H.-J. Ma, and P. Chakrabarti, "Variational Bayesian network with information interpretability filtering for air quality forecasting," Mathematics, vol. 11, no. 4, p. 837, Feb. 2023.

[30] J. Li, W. Shi, and D. Yang, "Color difference classification of dyed fabrics via a kernel extreme learning machine based on an improved grasshopper optimization algorithm," Color Res. Appl., vol. 46, no. 2, pp. 388–401, Apr. 2021.