International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A Hybrid Bi-Directional LSTM and Genetic Algorithm-Optimized Extreme Learning Machine for Air Quality Forecasting

Dr.P.Janaki Ramulu[1], Kamarajugadda Manaswini[2], Kommoju Naga Sowmya[3], J Divya[4], Dadigari Sai Manisha[5]

[1] Principal, Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5]BTech Students ,Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*—This study presents a robust and efficient framework for air quality forecasting that combines a Genetic Algorithm–optimized Kernel Extreme Learning Machine with a Bidirectional Long Short-Term Memory network to improve the prediction of PM2.5 concentrations. The Genetic Algorithm is utilized to fine-tune kernel parameters and optimize the hidden layer structure of the Extreme Learning Machine, which helps in selecting relevant features and reducing prediction errors. To better capture temporal dynamics, the Bidirectional LSTM processes time-series data in both forward and backward directions, allowing the model to understand dependencies from past as well as future observations. This dual-direction learning enhances the model's ability to represent complex pollution trends and long-term patterns more effectively. The proposed hybrid approach is tested on real-world PM2.5 datasets and compared with traditional models such as standard Extreme Learning Machine and Support Vector Machine. The evaluation, based on metrics like Mean Squared Error and Root Mean Squared Error, shows a clear improvement in prediction accuracy and consistency. Overall, the results demonstrate that the proposed model offers a reliable and scalable solution for air quality prediction, making it suitable for applications such as real-time monitoring, early warning systems, and informed environmental planning.**

***Keywords*—Air Quality Forecasting, PM2.5 Prediction, Bidirectional LSTM, Kernel Extreme Learning Machine, Genetic Algorithm Optimization, Hybrid Deep Learning Model, Time-Series Analysis, Environmental Monitoring, Machine Learning, Pollution Prediction, Feature Optimization, Predictive Modeling**

I. INTRODUCTION

Air pollution has become a major global concern, affecting both human health and the environment. With rapid industrial growth, urbanization, and increasing vehicle usage, the level of harmful pollutants in the air has risen significantly. Pollutants such as PM2.5, PM10, sulfur dioxide, nitrogen dioxide, carbon monoxide, and ozone are known to cause serious health problems, including respiratory and heart diseases. Research shows that air pollution contributes to millions of premature

deaths every year, highlighting the importance of reliable monitoring and forecasting systems. As discussed in *Air pollution: A global problem needs local fixes* [1], tackling this issue requires not only global awareness but also region-specific solutions. Accurate air quality prediction can play a crucial role in helping policymakers and healthcare systems take preventive actions and improve environmental planning.

Over the past few years, different statistical and machine learning methods have been developed to predict air quality. However, traditional models often fail to handle the complex and nonlinear nature of atmospheric data. More advanced approaches, including deep learning and Bayesian models, have shown better performance in analyzing time-series data. For example, Bayesian LSTM models have been used to study the impact of pollution control measures [2], while several review studies summarize the strengths and limitations of existing forecasting techniques [3]. In addition, computational intelligence methods such as neural networks and hybrid models have been applied to estimate pollutant levels [5]. Despite these improvements, issues like data inconsistency, feature selection, and lack of generalization still limit prediction accuracy.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

To overcome these challenges, researchers have started combining deep learning models with optimization techniques. Recurrent neural networks, especially LSTM models, are effective in learning patterns from sequential data and capturing long-term dependencies [11]. Hybrid models like CNN-LSTM and other deep learning frameworks have further improved forecasting performance in environmental applications [13], [14]. At the same time, optimization methods such as Genetic Algorithms are widely used to fine-tune model parameters and enhance prediction accuracy. These techniques help in selecting important features, adjusting hyperparameters, and improving the learning process. Previous studies have also shown that combining multiple approaches, such as spatio-temporal deep learning models, can lead to better PM2.5 prediction results [7].

Moreover, the rise of Internet of Things (IoT) technologies and smart monitoring systems has made it possible to collect large amounts of real-time air quality data [6], [12]. This has increased the demand for more efficient and scalable forecasting models. Although models like Extreme Learning Machines are known for their fast training speed, they often face stability issues due to random parameter initialization. To address this problem, optimization-based techniques have been introduced to improve their reliability and performance. In this context, the present study proposes a hybrid framework that combines Genetic Algorithm optimization, Kernel Extreme Learning Machine, and Bidirectional LSTM. The goal is to achieve more accurate, stable, and reliable air quality predictions under different environmental conditions.

II. RELATED WORK

Li et al., [2019] [1]examined air pollution as a serious global concern that requires region-specific solutions rather than a one-size-fits-all approach. Their study highlighted how increasing industrial activities and urban growth have intensified pollution levels, directly affecting human health and environmental sustainability. The authors emphasized that pollutants such as particulate matter and harmful gases contribute to respiratory and cardiovascular diseases, leading to a rise in premature deaths. They argued that effective air quality management depends on combining global awareness with local-level interventions. The paper also discussed the importance of data-driven strategies in understanding pollution patterns and supporting policy decisions. By focusing on both scientific analysis and practical solutions, the study provides valuable insights into managing air pollution. It serves as a strong foundation for further research in air quality prediction and highlights the importance of accurate forecasting systems in reducing health risks and improving environmental planning.

Han et al.,[2018][2]proposed a Bayesian Long Short-Term Memory (LSTM) model to analyze the impact of air pollution control policies in China. Their approach combined deep learning with probabilistic modeling to improve prediction reliability and handle uncertainty in environmental data. The study focused on capturing temporal patterns in air quality data, showing that LSTM networks are effective for time-series forecasting. By incorporating Bayesian techniques, the model was able to provide more stable and interpretable predictions compared to standard methods. The authors demonstrated that their model could evaluate the effectiveness of pollution control measures over time, offering useful insights for policymakers. The results indicated that advanced deep learning models can better understand complex environmental trends. This work highlights the potential of combining statistical and deep learning approaches to enhance air quality prediction and decision-making processes, especially in regions facing severe pollution challenges.

Bai et al., [2018] [3]presented a detailed review of various air pollution forecasting techniques, covering both traditional statistical methods and modern machine learning approaches. The study compared different models based on their ability to handle complex and dynamic environmental data. The authors pointed out that conventional models often struggle with nonlinear relationships and large datasets, which limits their prediction accuracy. In contrast, machine learning and deep learning models offer better performance due to their ability to automatically learn patterns from data. The paper also discussed challenges such as data quality, feature selection, and model generalization, which continue to affect forecasting results. By analyzing multiple approaches, the study provides a clear understanding of the strengths and weaknesses of existing methods. This work serves as a useful reference for researchers aiming to develop improved air quality prediction models and highlights the need for hybrid and optimized techniques.

Du et al., [2019] [14]developed a hybrid deep learning framework for air quality forecasting that combines multiple neural network models to improve prediction performance. Their approach

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

focused on capturing both spatial and temporal features of air pollution data, which are often difficult to model using single techniques. The study demonstrated that integrating different deep learning methods can significantly enhance prediction accuracy compared to traditional models. The authors also highlighted the importance of handling large-scale datasets and extracting meaningful features to improve forecasting results. Their model was tested on real-world data and showed consistent improvements in performance metrics. The findings suggest that hybrid deep learning frameworks are highly effective for environmental prediction tasks. This research contributes to the growing interest in combining multiple models to overcome individual limitations and achieve more reliable air quality forecasting.

Feng et al., [2019] [15] introduced a hybrid approach that combines recurrent neural networks (RNN) with random forest algorithms for predicting atmospheric pollutant levels. Their study aimed to improve both accuracy and stability by leveraging the strengths of different machine learning techniques. The RNN component was used to capture temporal dependencies in time-series data, while the random forest helped in handling nonlinear relationships and reducing overfitting. The authors applied their model to real-world data from Hangzhou, China, and observed improved prediction performance compared to standalone models. They also emphasized the importance of feature selection and data preprocessing in achieving better results. The study demonstrated that combining deep learning with ensemble methods can effectively address the challenges of air quality prediction. This work provides a practical approach for developing robust forecasting systems and highlights the benefits of hybrid modeling in environmental applications.

III. DATASET DETAILS

The dataset used in this study consists of real-world air quality measurements collected from different monitoring stations. It includes multiple environmental attributes such as PM2.5, PM10, nitrogen dioxide (NO_2), nitrogen oxides (NOx), ammonia (NH_3), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and other related pollutants. Along with these parameters, additional columns such as city name, date, and air quality index (AQI) categories are also present. These features help in understanding both the chemical composition of air pollution and its variation over time. Since the dataset contains daily observations, it is suitable for time-series analysis and forecasting tasks. During initial exploration, it was observed that some values were missing or recorded as zero, which could affect model performance. Therefore, preprocessing steps were necessary to ensure data quality and consistency before feeding it into machine learning models.

To make the dataset ready for model training, several preprocessing steps were carried out. First, unnecessary columns such as city name and date were removed so that only the important numerical features were retained for prediction. The dataset also contained missing values and unusual data points, which were handled using interpolation techniques. This method estimates missing values based on nearby data, helping to maintain continuity and improve data quality. After cleaning, all features were scaled using Min-Max normalization so that their values fall within a common range, which helps the model learn more effectively. The dataset was then separated into input variables (X) and the target variable (Y). Finally, it was divided into training and testing sets to properly evaluate the model's performance. These preprocessing steps ensure that the data is clean, consistent, and suitable for building accurate and reliable air quality prediction models.

IV. PROPOSED METHODOLOGY

The proposed system follows a step-by-step process to predict air quality using different learning techniques. First, the prepared dataset is taken as input after completing all preprocessing steps such as cleaning, handling missing values, and scaling the data. Only the important pollution-related features are selected for model training. The dataset is then divided into two parts: one for training the models and the other for testing their performance. Interpolation is used to handle missing or abnormal values so that the data becomes more consistent and suitable for analysis. This well-prepared dataset helps the models learn patterns more effectively.

After preparing the data, different models such as Support Vector Machine, GA-KELM, and Bidirectional LSTM are applied. Each model is trained using the training data and then tested using the test data. The GA-KELM model improves performance by optimizing its internal parameters, while the Bi-LSTM model captures patterns over time by learning from both past and future values. The performance of all models is compared using error measures like MSE and RMSE. Based on the

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

results, the Bi-LSTM model performs best and is used for final air quality prediction.

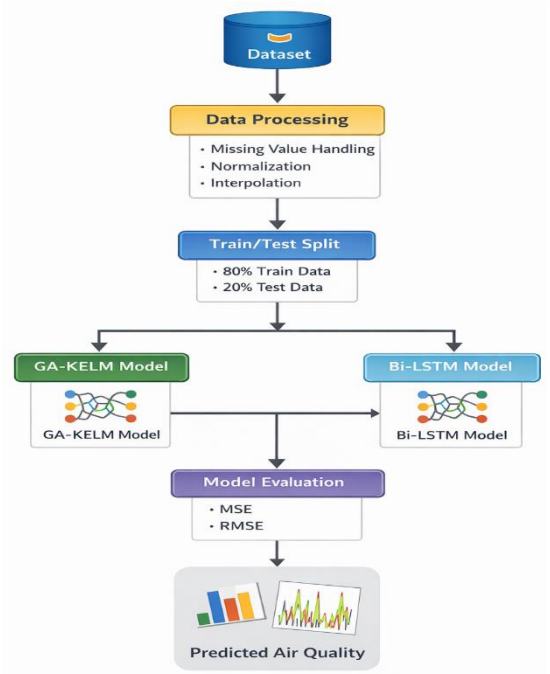


Figure 1: Air Quality Prediction System Using GA-KELM and Bi-LSTM Models

Figure [1] illustrates the overall flow of the proposed air quality prediction system. The process starts with the dataset, which is first cleaned and prepared through steps like handling missing values, applying interpolation, and normalizing the data. After preprocessing, the dataset is divided into training and testing portions to evaluate the model properly. The processed data is then used to train different models such as GA-KELM and Bi-LSTM, which learn patterns from the air pollution data. These models are tested and compared using performance measures like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Based on the comparison, the best model is selected to produce the final air quality predictions.

V. RESULT AND DISCUSSION

The experimental results clearly show the effectiveness of the proposed models in predicting air quality with improved accuracy. After completing data preprocessing and dividing the dataset into training and testing sets, multiple models including Support Vector Machine (SVM), GA-KELM, and Bidirectional LSTM were trained and evaluated. The SVM model produced a Mean Squared Error (MSE) of around 11, indicating moderate prediction performance with some noticeable gaps between actual and predicted values. The GA-KELM model showed better results, reducing the MSE to approximately 6 by optimizing model parameters, which helped in improving prediction accuracy. The best performance was achieved by the Bidirectional LSTM model, which further reduced the error to nearly 5, showing its strong ability to capture time-based patterns in the data. Graphical results demonstrated that the predicted values closely followed the actual air quality values, especially in the Bi-LSTM model where the lines almost overlapped. Comparative analysis using MSE and RMSE confirmed that the Bi-LSTM model outperformed the other approaches. Overall, the results indicate that the proposed hybrid approach is effective, reliable, and suitable for accurate air quality forecasting.

The dataset is organized in a tabular format where each row represents daily air pollution data for a specific city. It includes several important features such as PM2.5, PM10, NO, NO_2, NOx, NH_3, CO, SO_2, O_3, and other pollutant-related values, along with additional columns like date, AQI, and AQI category. These features provide detailed information about air conditions and pollution levels over time. The dataset is well-structured and prepared for further processing and analysis. Such organized data is important for training machine learning models, identifying patterns, and improving the accuracy of air quality prediction.

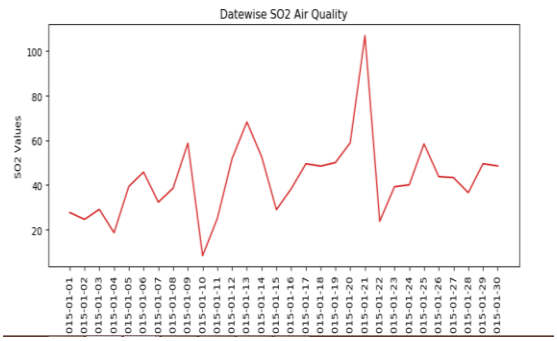


Figure 2: Date-wise SO_2 Air Quality Variation

Figure [2] shows the change in SO_2 levels across different dates. The x-axis represents the dates, while the y-axis shows the SO_2 values. The graph highlights how air quality varies over time, with noticeable rises and drops in pollution levels. This visualization helps in understanding daily trends and patterns in air pollution.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

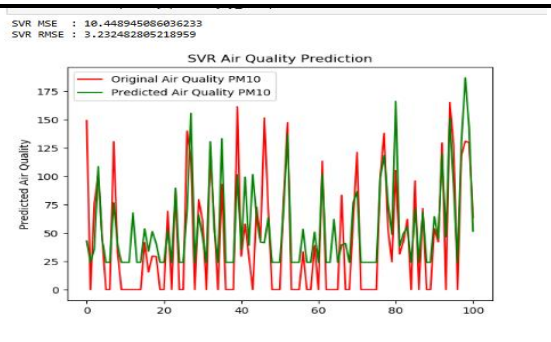


Figure 3: SVM-Based Air Quality Prediction Results

Figure [3] shows the prediction results obtained using the Support Vector Machine (SVM) model. The graph compares the original air quality values (red line) with the predicted values (green line). The x-axis represents the test data points, while the y-axis shows the predicted air quality levels. It can be observed that the predicted values follow the actual values to some extent, although there are noticeable gaps between the two lines. This indicates moderate prediction accuracy of the SVM model.

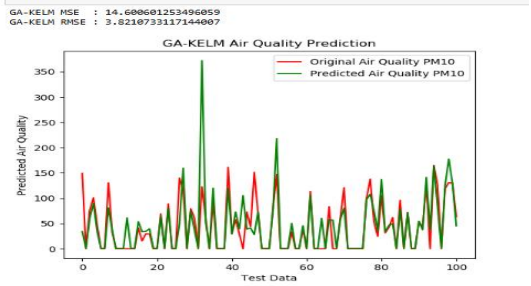


Figure 4: GA-KELM-Based Air Quality Prediction Results

Figure [4] shows the prediction results using the GA-KELM model. The graph compares the original air quality values (red line) with the predicted values (green line). The x-axis represents the test data, and the y-axis shows air quality levels. It can be observed that the predicted values are closer to the actual values compared to the SVM model, with fewer gaps between the lines. This indicates improved prediction accuracy and better performance of the GA-KELM model.

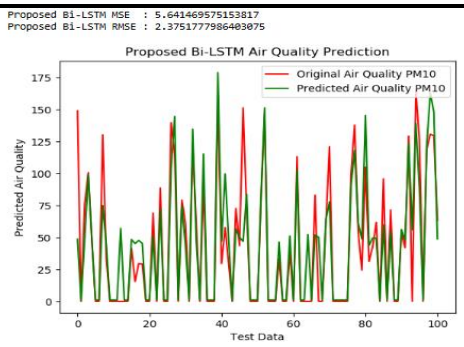


Figure 5: Bi-LSTM-Based Air Quality Prediction Results

Figure [5] shows the prediction results obtained using the Bi-LSTM model. The graph compares the original air quality values (red line) with the predicted values (green line). The x-axis represents the test data points, while the y-axis shows air quality levels. It can be seen that the predicted values closely match the actual values, with minimal differences between the two lines. This indicates that the Bi-LSTM model provides higher prediction accuracy and better performance compared to other models.

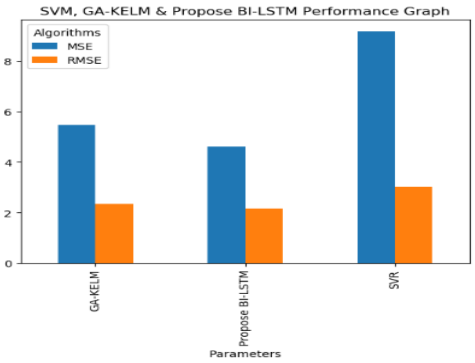


Figure 6: Performance Comparison Graph

Figure [6] shows a comparison of different models based on their performance metrics. The graph presents Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for SVM, GA-KELM, and Bi-LSTM using bar charts. The x-axis represents the models, while the y-axis shows the error values. It can be observed that the Bi-LSTM model has the lowest MSE and RMSE, indicating better prediction accuracy. GA-KELM performs better than SVM, while SVM shows the highest error among the three models. This comparison clearly highlights the superiority of the Bi-LSTM model.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

```
Test Data = [37.73  0.   7.29 80.09 44.76  0.   0.   0. ]=====> Predicted Air Quality : 72.341194
Test Data = [33.42  0.   8.92 54.28 47.42  0.   0.   0.27]=====> Predicted Air Quality : 45.065
Test Data = [26.86  0.   4.32 48.73 39.94  0.02 0.   3.55]=====> Predicted Air Quality : 54.90878
Test Data = [18.75  0.   1.41 35.91 56.15  0.   0.   3.31]=====> Predicted Air Quality : 55.852264
Test Data = [28.24  0.   6.34 66.58 53.14  9.7  9.63 16.49]=====> Predicted Air Quality : 87.42933
Test Data = [44.21  0.  14.19 65.91 31.88  7.72 17.46  2.51]=====> Predicted Air Quality : 82.72617
Test Data = [57.33  0.  18.18 80.43 40.11 11.29 24.35  3.35]=====> Predicted Air Quality : 118.35026
Test Data = [77.85  0.  37.49 99.72 36.47 17.   52.66  6.98]=====> Predicted Air Quality : 167.8433
Test Data = [44.87  0.  15.55 69.71 33.78 10.47 28.53  3.22]=====> Predicted Air Quality : 102.91522
Test Data = [47.31  0.  19.85 73.23 30.57 10.28 31.25  4.  ]=====> Predicted Air Quality : 99.804115
```

Figure 6: Prediction Output for Test Data

Figure [6] shows the final prediction output generated by the Bi-LSTM model on test data. The input values are displayed alongside their corresponding predicted air quality values. This output demonstrates how the trained model can take new data and provide accurate predictions. It also confirms the effectiveness of the model in handling unseen data for air quality forecasting.

DISCUSSION

The results of this study show that selecting the right model plays an important role in achieving accurate air quality predictions. Among all the models used, the Bi-LSTM model performed the best because it can learn patterns from both past and future data, making it suitable for time-series analysis. The GA-KELM model also gave better results than SVM by improving parameter selection and reducing errors. In comparison, the SVM model showed lower performance due to its limited ability to handle complex patterns in the data. Data preprocessing steps such as handling missing values, interpolation, and normalization helped improve the overall performance of the models. Splitting the dataset into training and testing ensured fair evaluation, while graphical comparisons made it easier to understand the results. Overall, the study shows that combining optimization methods with deep learning can provide more accurate and reliable air quality predictions.

VI. CONCLUSION

This project successfully demonstrates the use of machine learning and deep learning techniques for accurate air quality prediction. By applying essential preprocessing steps such as handling missing values, interpolation, and normalization, the dataset was effectively prepared for model training. Among the models used, the Bi-LSTM model achieved the best performance with lower error values compared to SVM and GA-KELM. The GA-KELM model also showed improved results due to optimization, while SVM provided comparatively moderate accuracy. The results confirm that the system can predict air quality values reliably for new data inputs. Visualization through graphs helped in clearly analyzing model performance and comparing predictions with actual values. Overall, the proposed approach provides an effective solution for air quality forecasting and can support better environmental monitoring. This study also highlights the importance of combining optimization techniques with deep learning models and offers a strong foundation for future improvements and real-time applications.

REFERENCES

[1] X. Li, L. Jinn, and H. Kan, “Air pollution: A global problem needs local fixes,” Nature, vol.570, pp.437–439, 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 AI Air Qual. EstimationForecasting Health Advice, 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, 2018, Art. no. 780.

[4] Y. Ding and Y. Xu, “A Deep Learning Approach to Writer Identification Using Inertial Sensor Data of Air-Handwriting,” IEICE Transient. & Syst., vol. E102-D, no.10, pp.2059–2063, 2019.

[5] S.-Q. Dotes, M. I. Petra, L. Dagger, and L. C. De Silva, “Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam,” Atmos. Pollute. Res., vol. 9, no. 2, pp. 358368,2018.

[6] M. Jia, A. Comelily, Y. Wang, and R. S. Srinivasan, “Adopting Internet of Things for the development of smart buildings: A review of enablingtechnologygives and applications,” Automat. Construct., vol. 101, pp. 111126, May 2019.

[7] Abram, S., Chitra, P., Mahometh, R., Keserwan, S.R., 2020. Hybrid Spatio-temporal Deep Learning Framework for Particulate Matter (PM2.5) Concentration Forecasting.2020 International Conference on Innovative Trends in Information Technology (ICITIIT) 1e6.

[8] Y. Cheng, S. Zhang, C. Huan, M. O. Loudoun, and Z. Lin, “Optimization on fresh outdoor air ratio

of air conditioning system with stratum venti-lotion for both targeted indoor air quality and maximal energy saving, "Building Environ., vol. 147, pp. 1122, Jan. 2019.

[9] A. C. Cosmas and R. Simcha, "Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions," Building Environ., vol. 148, pp. 372383, Jan. 2019.

[10] Bhowmik, M., Deb, K., Debnath, A., Asha, B., 2018. Mixed phase Fe 2 O 3/Mn 3 O 4 magnetic nanocomposite for enhanced adsorption of methyl orange dye: neural network modeling and response surface methodology optimization. Appl.Organomet. Chem. 32, e4186 https://doi.org/10.1002/aoc.4186.

[11] Chaudhary, V., Ditchwater, A., Kumar, V., Paul, D., 2018. Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in Indium, Aug.

[12] Chen, M., Yang, J., Hu, L., Shamim Hossain, M., Muhammad, G., 2018. Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators. IEEE Commune. Mag. 56, 14e20. https://doi.org/10.1109/MCOM.2018.1700571

[13] Chen, R., Wang, X., Zhang, W., Zhu, X., Li, A., Yang, C., 2019. A hybrid CNN-LSTM model for typhoon formation forecasting. Geoinformatics 23, 375e396. https://doi.org/10.1007/s10707-019-00355-0.

[14] Du, S., Li, T., Yang, Y., Hong, S.-J., 2019. Deep air quality forecasting using hybrid deep learning framework. IEEE Trans. Knowle. Data Eng. https://doi.org/10.1109/ tkde.2019.2954510, 1-1..

[15] Feng, R., Zheng, H., Gao, H., Zhang, A., Huang, C., Zhang, J., Luo, K., Fan, J., 2019.Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. J. Clean. Prod.