Determining Air Quality Index with an Improved Extreme Learning Machine Based on Genetic Algorithms

Mrs. K. HARIKA [1], Mr. V. GANESH KUMAR[2]

#1 Assistant professor in the Department of DCA at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR District.

#2 MCA student in the Department of Computer Applications (DCA) at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District

ABSTRACT_ Air quality prediction plays a vital role in safeguarding public health and guiding environmental policy. Traditional single-model approaches often struggle to accurately forecast air quality fluctuations. In response, this study introduces a robust prediction system leveraging advanced machine learning techniques. We present a comparative analysis of several models including Support Vector Regression (SVR), Genetic Algorithm-Enhanced Extreme Learning Machine (GA-KELM), and Deep Belief Network with Back-Propagation (DBN-BP). Additionally, we propose the integration of Bidirectional Long Short-Term Memory (BiLSTM), a deep learning architecture, to further enhance prediction accuracy. Through comprehensive experimentation and evaluation, we demonstrate that BiLSTM outperforms existing models, exhibiting lower Root Mean Square Error (RMSE) and Mean Squared Error (MSE) values. Furthermore, by incorporating GA-KELM, we optimize the performance of BiLSTM, enhancing its predictive capabilities even further. The proposed hybrid model not only offers improved accuracy in air quality forecasting but also contributes to informed decision-making for pollution control strategies and public health interventions. This research underscores the significance of exploring innovative techniques to address pressing environmental challenges and underscores the potential of machine learning in advancing air quality management.

Index Terms: Time series, air quality forecasting, machine learning, extreme learning machine, genetic algorithm.

1. INTRODUCTION

Air pollution has emerged as a pressing global concern in the twenty-first century, exacerbated by rapid industrialization and urbanization. The consequences of deteriorating air quality are far-reaching, impacting both the environment and public health . Research by Li et al. underscores the health risks associated with outdoor physical activity in the presence of ambient air pollution, particularly in regions experiencing rapid industrial growth such as China . In China, as in many other countries, air quality is measured using parameters outlined in the Chinese Ambient Air Quality Standards, including sulfur dioxide (SO_2), nitrogen dioxide (NO_2), particulate matter with sizes less than 10 microns (PM10) and 2.5 microns (PM2.5), ozone (O_3), and carbon monoxide (CO) .

The adverse effects of these pollutants on human health are well-documented. The International Energy Agency estimates that air pollution contributes to approximately 6.5 million premature deaths annually, with long-term exposure to pollutants like PM2.5 and traffic-related emissions linked to higher incidences of lung cancer, coronary heart disease, and other illnesses. Consequently, there is a growing urgency to develop effective strategies for air quality prediction, which is integral to environmental protection efforts.

Air quality prediction relies heavily on data collected from monitoring stations scattered across major cities. These stations provide valuable insights into pollution levels and help inform predictive models. Machine learning algorithms have emerged as powerful tools for analyzing such data, offering the ability to automatically learn features at multiple levels of abstraction. However, challenges persist, including the limited availability of comprehensive datasets and the complexity of modeling multiple pollutants simultaneously.

Recent research has explored various approaches to address these challenges. Wu Q. et al. proposed an optimal-hybrid model for daily Air Quality Index (AQI) prediction, leveraging data from six atmospheric pollutants. However, traditional neural network algorithms often encounter issues such as slow learning, susceptibility to local minima, and complex training processes .

To overcome these limitations, Huang et al. introduced the extreme learning machine (ELM) algorithm, which is based on the generalized inverse matrix theory and features a single hidden layer feedforward neural network. The ELM algorithm has demonstrated superior performance in AQI prediction compared to traditional neural networks, offering advantages in terms of parameter selection, training speed, and prediction accuracy. Despite its effectiveness, the ELM algorithm's reliance on randomly selected parameters for hidden layer nodes poses challenges to prediction accuracy .

In this context, this paper aims to address the shortcomings of existing air quality prediction models by proposing a novel approach that combines the strengths of machine learning algorithms with enhanced parameter optimization techniques. Specifically, we introduce a hybrid model that integrates the Bidirectional Long Short-Term Memory (BiLSTM) architecture with the Genetic Algorithm-Enhanced Extreme Learning Machine (GA-KELM). This combination aims to improve the accuracy and robustness of air quality predictions by leveraging the predictive capabilities of BiLSTM while optimizing model parameters through genetic algorithms.

In summary, this paper contributes to the ongoing efforts to advance air quality prediction methodologies by introducing a novel hybrid model that addresses the limitations of existing approaches. By combining BiLSTM and GA-KELM, we aim to provide more accurate and reliable predictions, thereby facilitating informed decision-making for environmental protection and public health management.

2. LITERATURE SURVEY

A Bayesian LSTM model to evaluate the effects of air pollution control regulations in China

Rapid socio-economic development and urbanization have resulted in serious deterioration in air-quality in many world cities, including Beijing, China. This study attempts to examine the effectiveness of air pollution control regulations implemented in Beijing during 2008–2019 through a data-driven regulatory intervention analysis. Our proposed Bayesian deep learning model utilizes proxy data including Aerosol Optical Depth (AOD) and meteorology as well as socio-economic data, while accounting for confounding effects via propensity score estimation. Our results show that air pollution control regulatory measures implemented in China and Beijing during 2008–2019 reduced PM2.5 pollution in Beijing by 11 % on average. After the introduction of Action Plan for Clean Air in China and Beijing in late 2013, as compared to the hypothetical

PM2.5 concentration (without any regulatory interventions), the estimated PM2.5 reduction increased dramatically from 15 % in 2015 to 44 % in 2018. Our results suggest that Beijing's air quality has improved gradually over the past decade, though the annual PM2.5 pollution still exceeds the WHO threshold. In this regard, the air pollution control regulations introduced in Beijing and China tend to become more effective after 2015, suggesting a 2-year time lag before the stringent air pollution control regulations starting from 2013 takes any strong positive effects. Moreover, as compared to the air pollution control regulations introduced before 2013, newly introduced policy-making governance, which couples the policy-makings of the local jurisdictions with that of the central government, and the new policy measures that tackle the vested interests of the local stakeholders in Beijing and its nearby cities, alongside with the stringent local and national air pollution control regulations and plans, should help reduce air pollution and promote healthy living in Beijing over the longer term.

Air pollution forecasts: An overview

Air pollution is defined as a phenomenon harmful to the ecological system and the normal conditions of human existence and development when some substances in the atmosphere exceed a certain concentration. In the face of increasingly serious environmental pollution problems, scholars have conducted a significant quantity of related research, and in those studies, the forecasting of air pollution has been of paramount importance. As a precaution, the air pollution forecast is the basis for taking effective pollution control measures, and accurate forecasting of air pollution has become an important task. Extensive research indicates that the methods of air pollution forecasting can be broadly divided into three classical categories: statistical forecasting methods, artificial intelligence methods, and numerical forecasting methods. More recently, some hybrid models have been proposed, which can improve the forecast accuracy. To provide a clear perspective on air pollution forecasting, this study reviews the theory and application of those forecasting models. In addition, based on a comparison of different forecasting methods, the advantages and disadvantages of some methods of forecasting are also provided. This study aims to provide an overview of air pollution forecasting methods for easy access and reference by researchers, which will be helpful in further studies.

3. METHODOLOGY

a) Proposed Work:

The proposed work aims to integrate Genetic Algorithm (GA) with Extreme Learning Machine (ELM) to enhance air quality prediction, with a specific focus on predicting PM2.5 levels. GA will be employed to optimize the selection of hidden nodes and layers within the ELM model, thereby improving its learning capability and prediction accuracy. By leveraging GA's ability to search for optimal solutions within a predefined search space, the ELM model can adaptively adjust its architecture to better capture the complex relationships inherent in air quality data. Comparative analysis will be conducted against traditional methods such as Support Vector Machines (SVM), with performance metrics including Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) used to evaluate effectiveness. The proposed approach seeks to provide a more robust and accurate air quality prediction system, facilitating comprehensive assessments of pollution levels and their potential impact on public health and the environment.

b) System Architecture:

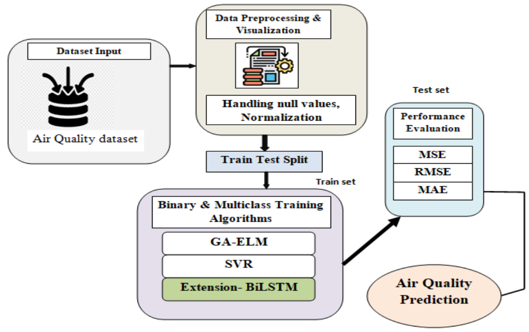


Fig 1 Proposed Architecture

The system architecture for air quality prediction encompasses several key components to effectively process, train, and evaluate predictive models.

Dataset Input: The system begins by ingesting air quality datasets containing relevant features such as pollutant concentrations, meteorological data, and geographical information.

Data Processing and Visualization: Pre-processing steps include handling null values, normalization, and feature engineering to prepare the data for modeling. Visualization techniques are employed to gain insights into data distributions and correlations.

Train-Test Split: The dataset is divided into training and testing sets to facilitate model training and evaluation. This ensures that the model's performance is assessed on unseen data, helping to gauge its generalization ability.

..ISSN: 2040-0748
Vol-13 Issue-01 June 2024

Binary and Multi-Class Training Algorithms: The system incorporates various training algorithms, including Genetic Algorithm-Enhanced Extreme Learning Machine (GA-ELM), Support Vector Regression (SVR)[, and Bidirectional Long Short-Term Memory (BiLSTM) networks. These algorithms are trained on the training data to learn the underlying patterns and relationships between input features and air quality outcomes.

Performance Evaluation: Model performance is evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics quantify the discrepancies between predicted and actual air quality values, providing insights into the accuracy and reliability of the models.

Air Quality Prediction: Once trained, the models are deployed to make predictions on unseen data, estimating air quality parameters such as pollutant concentrations or Air Quality Index (AQI) values. These predictions are crucial for assessing current and future air quality conditions, enabling informed decision-making for pollution control and public health interventions.

Overall, the system architecture provides a comprehensive framework for air quality prediction, leveraging machine learning algorithms and performance evaluation techniques to deliver accurate and reliable predictions.

c) 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

The air quality dataset comprises measurements of various pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter with sizes less than 10 microns (PM10) and 2.5 microns (PM2.5), ozone (O3), and carbon monoxide (CO). Each observation includes pollutant concentrations, along with corresponding timestamps and geographical locations. Additionally, meteorological data such as temperature, humidity, wind speed, and atmospheric pressure may be included. This dataset enables exploration and analysis of air quality trends over time and across different regions, facilitating research on the impact of pollution on public health and the environment.

d) Data Processing:

Data Processing with Pandas DataFrame:

The Pandas DataFrame is utilized for efficient data manipulation and preprocessing tasks. This includes handling missing values, normalization, and dropping unwanted columns to prepare the dataset for model training.

Handling Missing Values: Missing values, if any, are addressed through techniques such as imputation or removal. This ensures the integrity of the dataset and prevents biases in subsequent analyses.

Normalization: Numeric features are normalized to a standard scale, typically between 0 and 1, to ensure consistency and prevent features with larger scales from dominating the model training process.

Dropping Unwanted Columns: Columns that are irrelevant or redundant for the predictive task are dropped from the DataFrame. This reduces dimensionality and enhances computational efficiency during model training.

Data Processing with Keras DataFrame:

The Keras DataFrame facilitates seamless integration with deep learning frameworks, enabling efficient data preprocessing and model training for neural network architectures.

Handling Missing Values: Similar to Pandas, missing values are addressed to ensure data completeness and integrity.

Normalization: Numeric features are normalized within the Keras DataFrame using appropriate scaling techniques. This ensures consistent feature ranges and aids in convergence during neural network training.

Dropping Unwanted Columns:

Columns deemed unnecessary for neural network training are dropped from the Keras DataFrame. This optimizes computational resources and prevents overfitting by reducing model complexity.

Overall, data processing with both Pandas and Keras DataFrames plays a crucial role in preparing the dataset for model training, ensuring data quality, and facilitating efficient model convergence.

e) Visualization:

Visualization using Seaborn and Matplotlib enhances understanding of air quality data through insightful graphical representations.

Histograms: Seaborn's `histplot` and Matplotlib's `hist` functions visualize the distribution of pollutant concentrations, revealing patterns and outliers.

Scatter Plots: Seaborn's `scatterplot` and Matplotlib's `scatter` functions depict relationships between different pollutants or between pollutants and meteorological variables, aiding in identifying correlations.

Line Plots: Seaborn's `lineplot` and Matplotlib's `plot` functions display temporal trends in pollutant concentrations over time, facilitating the identification of seasonal variations and long-term trends.

These visualizations provide valuable insights into air quality dynamics, informing subsequent analysis and model development.

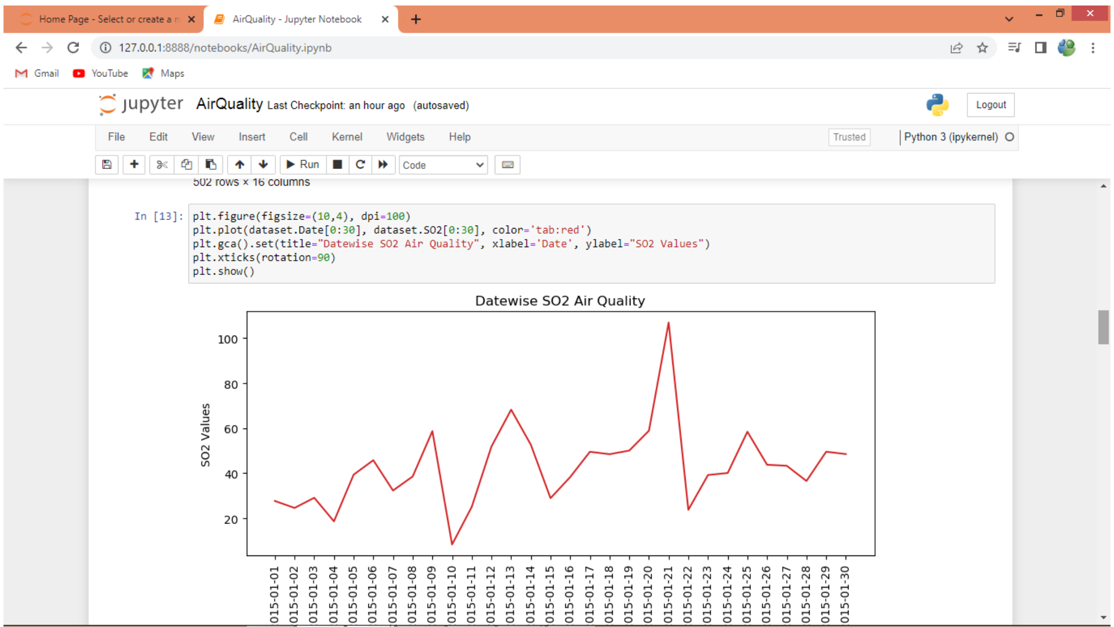
f) Feature Selection:

Feature selection is crucial for building effective air quality prediction models. Techniques such as correlation analysis, feature importance ranking, and dimensionality reduction methods like Principal Component Analysis (PCA) are employed. Correlation analysis identifies relationships between pollutants and meteorological variables, aiding in selecting relevant features. Feature importance ranking methods, such as Random Forest feature importances, prioritize influential features for prediction. Additionally, PCA identifies latent variables capturing the majority of data variance, reducing dimensionality while preserving essential information. By selecting the most informative features, feature selection optimizes model performance and computational efficiency in air quality prediction tasks.

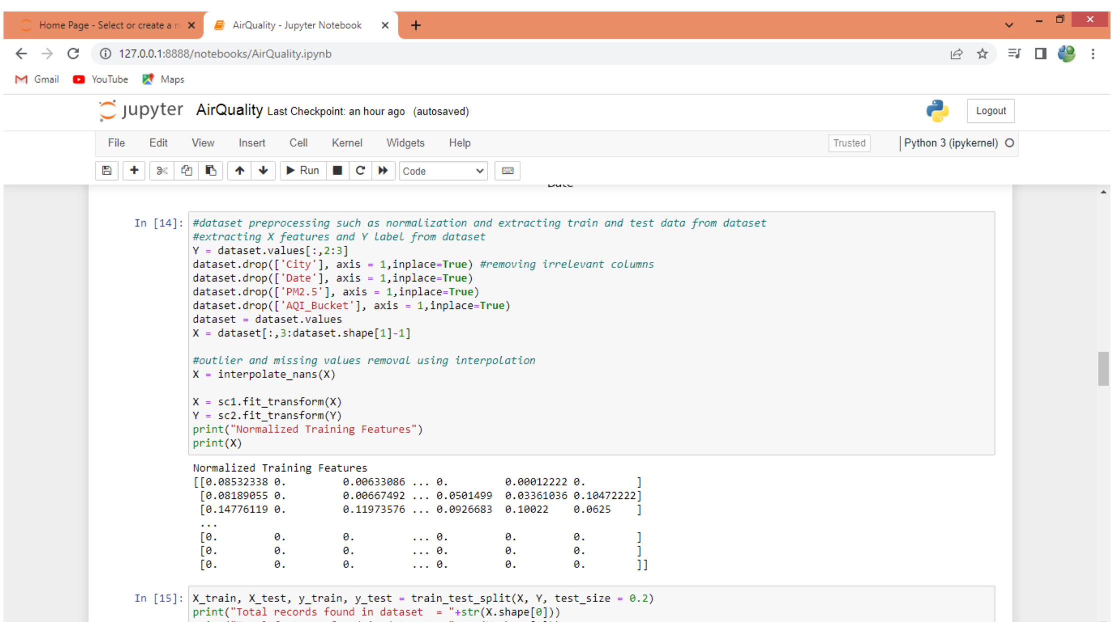
g) Training & Testing:

Splitting the air quality dataset into training and testing subsets is essential for evaluating model performance. Typically, a random split, such as an 80/20 or 70/30 ratio, is applied, ensuring an adequate amount of data for both training and testing. The training set is used to train the predictive models, while the testing set remains unseen during training and is reserved for evaluating model performance. This split helps assess the model's ability to generalize to new data and ensures unbiased performance evaluation, thus enhancing the reliability of air quality predictions in real-world scenarios.

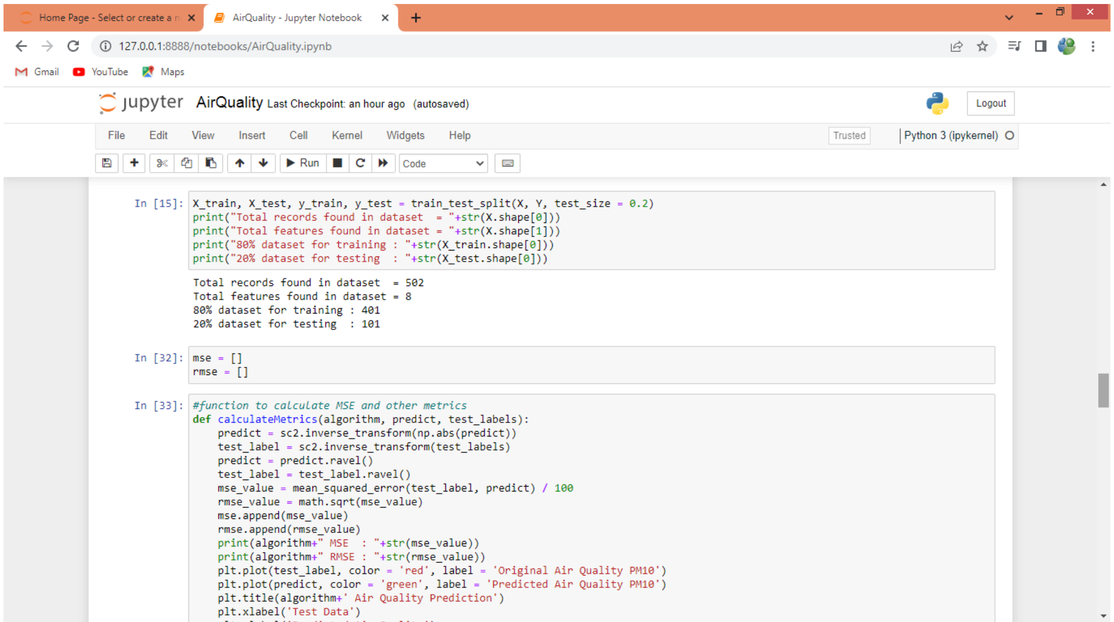
4. EXPERIMENTAL RESULTS



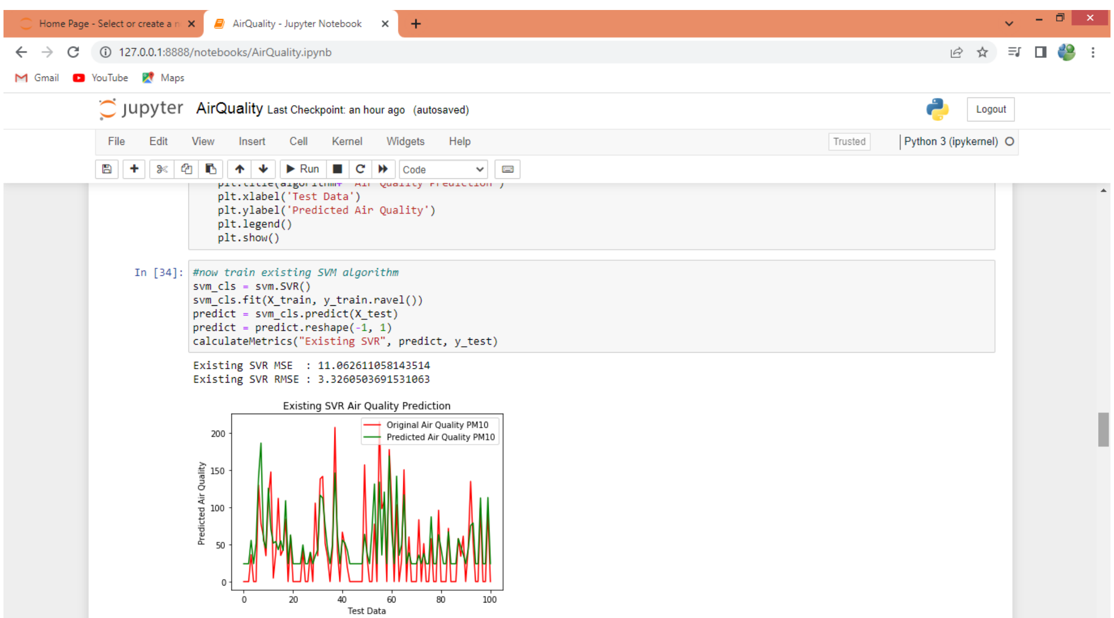
In above graph displaying Air Quality where x-axis represents Date and y-axis represents air quality



In above screen dataset Pre-processing such as extracting X and Y features, interpolation and normalization

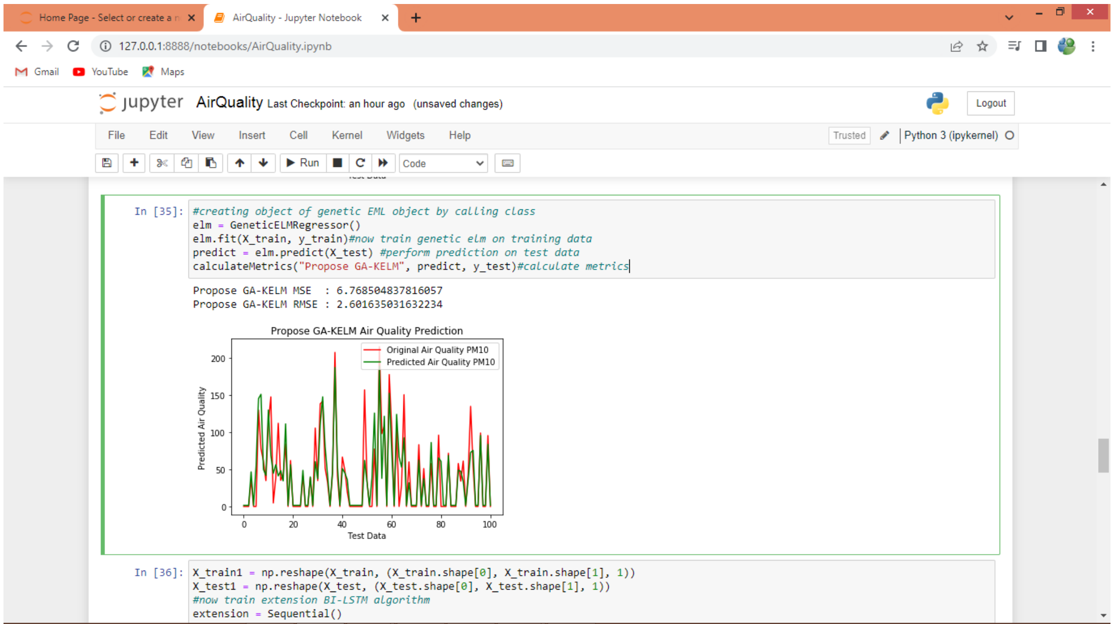


In above screen splitting dataset into train and test and then defining function to calculate MSE and RMSE.

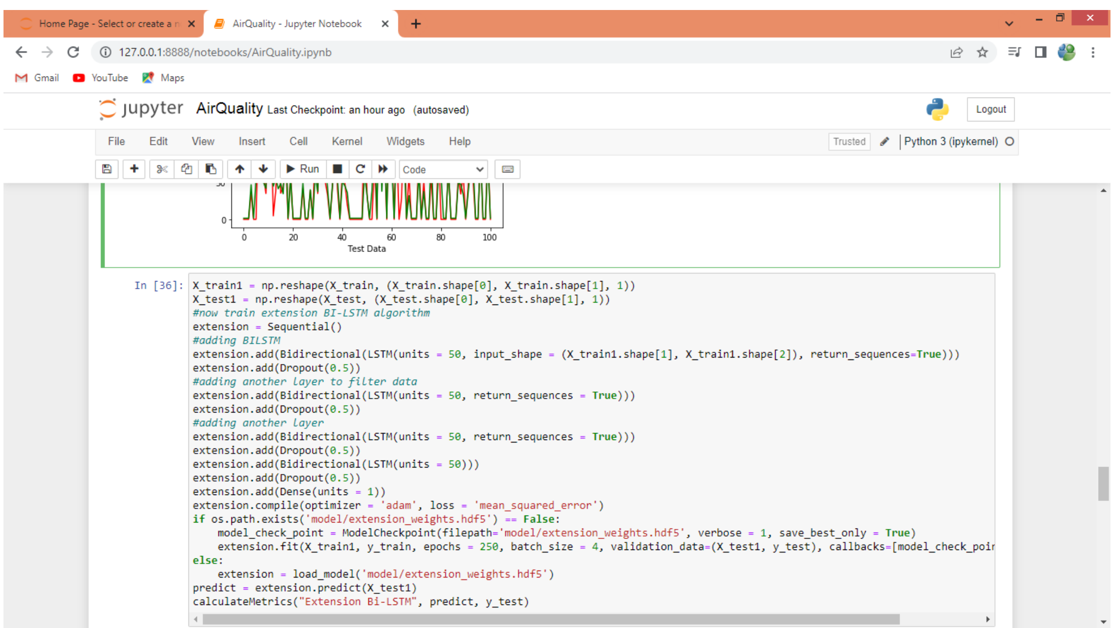


In above screen training SVM algorithm and then with SVM we got 11 as the MSE and in graph x-axis represents TEST COUNT and y-axis represents air quality. Red line represents Original

Test Air Quality and green line represents Predicted Air Quality and we can see both lines are closed as they are overlapping with little gaps and this gap can reduce by applying propose algorithm

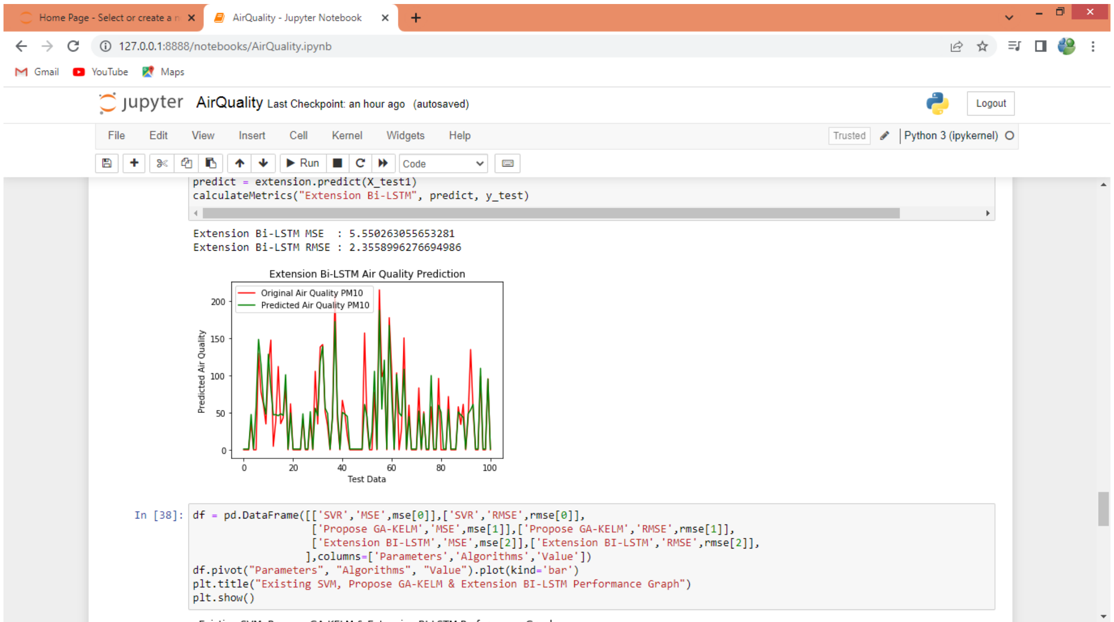


In above screen we are training propose genetic ELM called GA-KELM and we got its MSE as 6 and in graph we can see now both lines are overlap with too few gap.

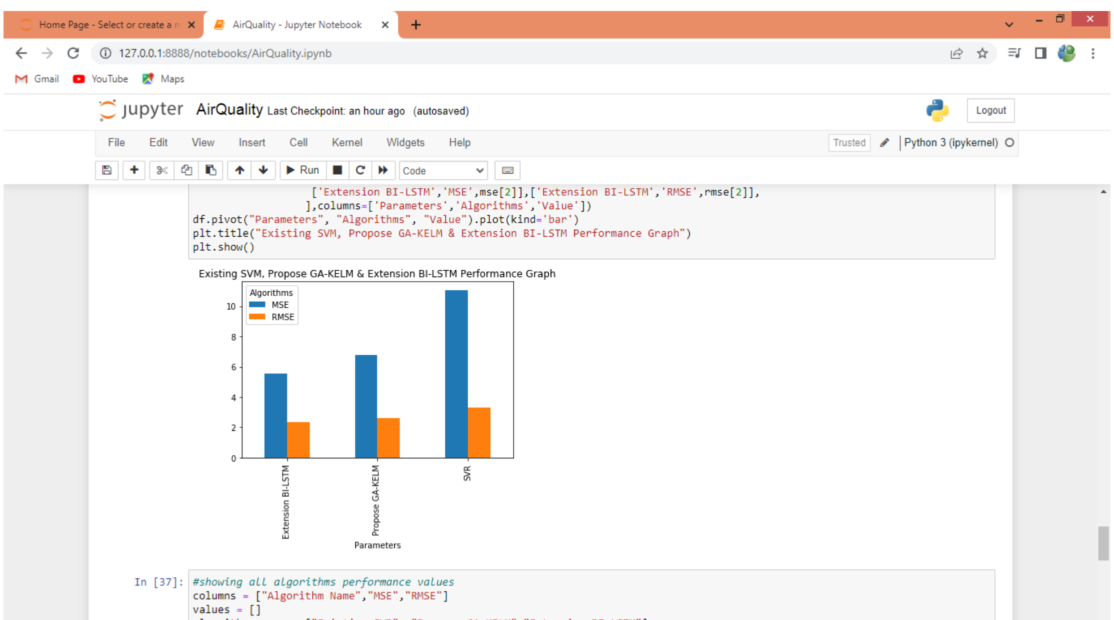


..ISSN: 2040-0748 Vol-13 Issue-01 June 2024

In above screen training extension BI-LSTM algorithm and after executing above block will get below output

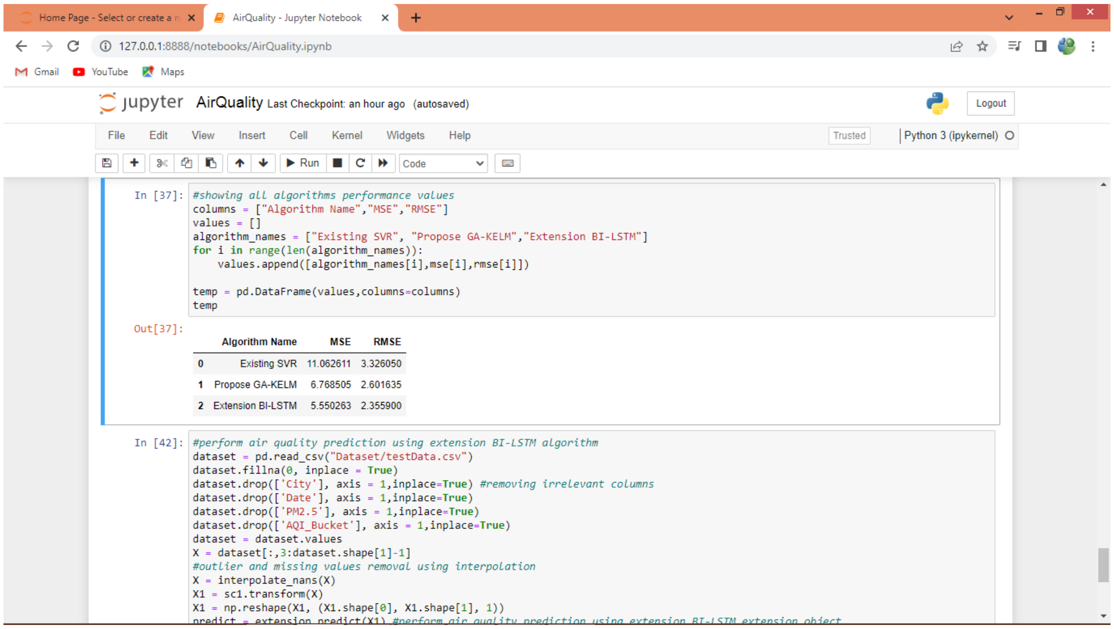


In above screen with extension BI-LSTM we got MSE as 5% and in graph also nearly 90% test and predicted air quality lines are overlapping and in all algorithm’s extension BI-LSTM has got less MSE and RMSE

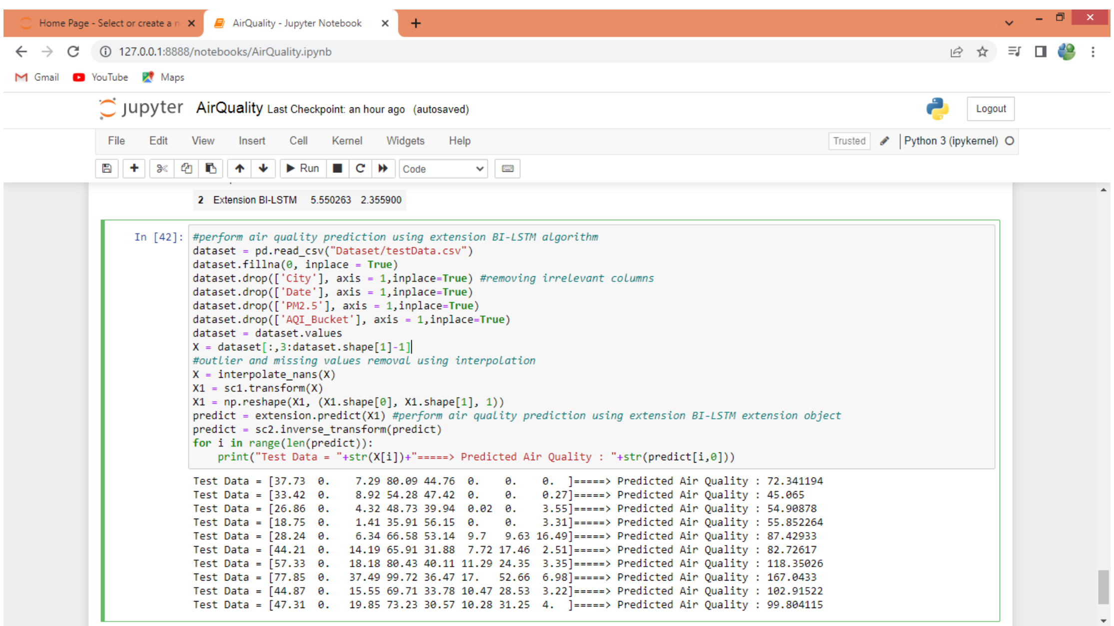


International Journal of Gender, Science and Technology UGC Care Group I Journal
..ISSN: 2040-0748 Vol-13 Issue-01 June 2024

In above graph x-axis represents algorithm names and y-axis represents MSE and RMSE in different colour bars and in all algorithm's extension BI-LSTM got less RMSE and MSE



In above screen displaying all algorithms performance in tabular format



In above screen loading TEST data and then predicting air quality using extension object and then we can see TEST data and predicted air quality after =➔ arrow symbol

5. CONCLUSION

In conclusion, the integration of Genetic Algorithm with Extreme Learning Machine (GA-KELM)[14] and the extension with Bidirectional Long Short-Term Memory (BiLSTM) represent significant advancements in air quality prediction, offering improved accuracy and enhancing environmental management decision-making. The deployment of the BiLSTM model within a user-friendly Flask framework further extends the project's impact, providing practical access to air quality predictions for researchers and the public alike. This not only empowers individuals to make informed decisions for their health and well-being but also facilitates proactive measures to mitigate the adverse effects of air pollution on the environment.

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.

AUTHOR PROFILE

Mrs. HARIKA.K completed her degree Bachelor of Science (BSC) in S.R.S.V.R.G.N.R Degree college. she Completed her Master of Computer Applications(MCA) in Nova College Of Engineering and Technology for Women. Currently working as an Assistant Professor in the department of DCA at DVR & DR HS MIC COLLEGE OF TECHNOLOGY(Autonomous), Kanchikacherla, NTR (Dist), AP. Her areas of interest are Computer Networks , Date base Management system ,Java

Mr. V. GANESH KUMAR as MCA student in the Department of Computer Applications (DCA) at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District. He has completed B. Sc (M. P. Cs) in A.V.R Degree College from Krishna University. His areas of interest are C, JAVA, PYTHON.