International Journal of Engineering & Scientific Research
Vol.11 Issue 09, September 2023
ISSN: 2347-6532 Impact Factor: 7.501
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Utilizing Deep Learning Models for Accurate Prediction of Air Pollution Levels

Harshit Jain
Department of Electronics & Telecommunications
Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai

ABSTRACT

Air pollution has become a critical environmental issue, adversely affecting human health and the overall well-being of ecosystems. Accurate forecasting of air pollution levels is crucial for effective pollution management and mitigation strategies. In this study, we propose a deep learning-based model for air pollution forecasting that harnesses the power of neural networks to predict pollutant concentrations with high precision. Our approach involves training a deep learning model using historical air quality data, meteorological variables, and other relevant features. We leverage the temporal and spatial dependencies within the data to capture complex patterns and relationships. By incorporating information such as pollutant levels from previous time steps, meteorological conditions, and geographical factors, our model learns to effectively forecast air pollution levels. To train the deep learning model, we utilize a large dataset comprising historical air quality measurements from diverse monitoring stations. We preprocess the data, handle missing values, and normalize the features to ensure optimal training performance. The model architecture consists of multiple layers of interconnected neurons, enabling it to learn hierarchical representations of the input data.

INTRODUCTION

Air pollution is a significant global environmental concern that poses severe risks to human health and the ecosystem. The accurate forecasting of air pollution levels is crucial for effective pollution control measures, urban planning, and public health protection. Traditional statistical models have been widely used for air pollution forecasting, but they often struggle to capture the complex spatial and temporal dynamics inherent in pollution data. In recent years, deep learning models have shown great potential in various fields, including time series forecasting tasks. Deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants, have the ability to automatically learn intricate patterns and dependencies within data, making them well-suited for air pollution forecasting.

The objective of this study is to develop a deep learning model specifically tailored for air pollution forecasting. The proposed model aims to capture the spatial and temporal correlations present in air pollution data, providing accurate and reliable predictions for future pollution levels. By leveraging the power of deep learning, the model can potentially overcome the limitations of traditional statistical models and improve the accuracy of air pollution forecasts. The deep learning model utilized in this study incorporates various components to effectively capture different aspects of air pollution data. Convolutional layers are employed to extract spatial features, enabling the model to learn the spatial dependencies between different monitoring stations. Recurrent layers, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), capture the temporal patterns and sequential nature of air pollution data. To train and evaluate the model, real-world air pollution datasets, including pollutant concentrations, meteorological variables, and temporal patterns, are used. The model is trained on historical data and tested on unseen data to assess its forecasting performance. Various evaluation metrics, such as mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R-squared), are employed to measure the accuracy and reliability of the model's predictions.(Tao, Q et al,2019).

The outcomes of this research can have significant implications for air pollution management, urban planning, and public health. Accurate forecasting of air pollution levels can assist policymakers in making informed decisions and implementing timely interventions to mitigate the adverse effects of pollution. Furthermore, it can provide valuable insights into the impact of pollution sources, identify pollution hotspots, and guide efforts to improve air quality.

NEED OF THE STUDY

Air pollution is a pressing environmental issue that has adverse effects on human health, ecosystems, and climate. Accurate forecasting of air pollution levels is essential for effective pollution control strategies, public health management, and informed decision-making. Traditional methods for air pollution forecasting often face challenges in capturing the complex spatiotemporal patterns present in pollution data. In recent years, deep learning models have emerged as powerful tools for time series forecasting tasks. These models can automatically learn and extract intricate patterns and dependencies from data, enabling more accurate predictions. In this study, we aim to develop a deep learning model for air pollution forecasting based on 1D convolutional neural networks (ConvNets) and bidirectional Gated Recurrent Units (GRU). The proposed model leverages the strengths of

ConvNets and bidirectional GRU to capture both spatial and temporal dependencies in air pollution data. ConvNets are well-suited for capturing spatial correlations, allowing the model to learn the relationships between pollutant concentrations at different monitoring stations. Bidirectional GRU, on the other hand, excels at capturing the temporal dynamics and sequential patterns within the data. The motivation behind this study is to address the limitations of traditional forecasting methods and explore the potential of deep learning techniques in improving air pollution forecasting accuracy. By employing a deep learning model that combines ConvNets and bidirectional GRU, we aim to enhance the model's ability to capture the complex spatiotemporal patterns present in air pollution data. (Bekkar, A et al,2021)

LITERATURE REVIEW

Tao, Q., Liu, F., et al, (2019). Air pollution has become a significant environmental concern worldwide, with detrimental effects on human health and the environment. Accurate forecasting of air pollution levels plays a crucial role in managing and mitigating its impact. In recent years, deep learning models have shown promising results in various time series forecasting tasks. This study proposes a deep learning model based on 1D convolutional neural networks (ConvNets) and bidirectional Gated Recurrent Units (GRU) for air pollution forecasting. The proposed model leverages the spatial and temporal correlations present in air pollution data. The 1D ConvNets are utilized to capture the spatial dependencies within the input features, such as pollutant concentrations at different monitoring stations. The bidirectional GRU component is employed to capture the temporal dependencies and model the sequential nature of air pollution data.

Bekkar, A., Hssina, B., et al, (2021). In this study, we presented a deep learning approach for air pollution prediction in smart cities. The proposed model leverages the power of deep neural networks to effectively capture the complex relationships and patterns present in air pollution data. By incorporating spatial and temporal dependencies through 1D ConvNets and bidirectional GRU, the model achieves accurate and robust predictions of air pollution levels. The experimental results demonstrate the superiority of the deep learning approach over traditional statistical models and other baseline methods.

Heydari, A., MajidiNezhad, M., et al,(2021). The deep learning model employed in this application utilizes a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The CNNs capture spatial dependencies within the input features, such as pollutant concentrations from multiple monitoring stations, while the RNNs capture temporal patterns and sequential dependencies in the air pollution data. To further

enhance the model's performance, an optimization algorithm is integrated into the forecasting application. The optimization algorithm fine-tunes the hyperparameters of the deep learning model, allowing for automatic parameter selection and optimization. This optimization process helps improve the model's generalization ability and prediction accuracy. Extensive experiments were conducted using real-world air pollution datasets to evaluate the performance of the proposed application. The model's predictions were compared against traditional statistical models and other baseline approaches. Evaluation metrics, including mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE), were employed to assess the accuracy and robustness of the forecasting application.

Chang, Y. S., Chiao, H. T., et al, (2020). Long Short-Term Memory (LSTM) networks have demonstrated excellent performance in capturing temporal dependencies and patterns in time series data. In this research, multiple LSTM models are trained on different subsets of the air pollution dataset. Each LSTM model focuses on specific temporal aspects, such as hourly, daily, or weekly patterns, and captures the unique characteristics of the data at different time scales. To create the aggregated model, the predictions from the individual LSTM models are combined using a weighted average approach. The weights are determined through a learning process that optimizes the model's performance on historical data. This aggregation mechanism enables the model to effectively capture the overall trends and fluctuations in air pollution levels. To evaluate the proposed LSTM-based aggregated model, extensive experiments are conducted on real-world air pollution datasets.

Guo, C., Liu, G., et al,(2020). Accurate forecasting of air pollution concentration is essential for environmental management and public health protection. This study proposes a forecasting method based on the deep ensemble neural network to improve the accuracy and robustness of air pollution concentration predictions. The deep ensemble neural network consists of multiple deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and fully connected neural networks (FCNs). Each model within the ensemble captures different aspects of the air pollution data and learns distinct representations, leveraging the strengths of various neural network architectures.

Jeya, S., &Sankari, L. (2021). The proposed model integrates several advanced techniques to enhance prediction accuracy. It combines a deep learning architecture, such as a convolutional neural network (CNN) or a long short-term memory (LSTM) network,

to capture complex spatial and temporal patterns in PM2.5 data. The adaptive kernel fuzzy system is utilized to model the uncertainties and nonlinear relationships inherent in the data, enabling more flexible and accurate predictions.To optimize the model's performance, a particle swarm optimization algorithm with fuzzy weighting is employed. This algorithm adaptively adjusts the weights of the particles to effectively explore the search space and find optimal model parameters. The optimization process enhances the model's generalization ability and improves the accuracy of PM2.5 predictions.Extensive experiments are conducted using real-world air pollution datasets to evaluate the performance of the proposed model. Evaluation metrics, including mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R), are used to assess the accuracy and reliability of the PM2.5 predictions.

Barot, V., & Kapadia, V. (2022). The proposed approach consists of two stages: model construction and fine-tuning. In the model construction stage, an LSTM neural network is trained on historical air quality data, incorporating relevant features such as meteorological variables, temporal patterns, and pollutant concentrations. The LSTM network learns to capture the sequential dependencies and non-linear relationships in the data. In the fine-tuning stage, the pre-trained LSTM model is further optimized to improve its prediction accuracy. A fine-tuning process is performed using additional data, allowing the model to adapt to specific characteristics and changes in the air quality parameters. This stage helps refine the model's performance and enhances its ability to generalize to unseen data.

METHODOLOGY

Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks. It is widely used in various domains, including image recognition, text classification, and bioinformatics.The main idea behind SVM is to find an optimal hyperplane that maximally separates the data points of different classes in a high-dimensional feature space. This hyperplane acts as a decision boundary, allowing SVM to classify new, unseen data points. One of the key strengths of SVM is its ability to handle both linearly separable and non-linearly separable data. In cases where the data is not linearly separable, SVM employs a technique called kernel trick. The kernel trick allows SVM to transform the input data into a higher-dimensional feature space, where it becomes linearly separable. This enables SVM to capture complex relationships and make accurate predictions. SVM aims to find the hyperplane with the largest margin, which is the distance between the hyperplane and the nearest data points of each class. By maximizing the margin, SVM promotes better generalization and reduces the risk of

overfitting.In addition to classification, SVM can also be extended to regression tasks. In regression, SVM seeks to find a hyperplane that best fits the data by minimizing the error between the predicted and actual values. This approach is known as Support Vector Regression (SVR).SVM has several advantages. It is effective in handling high-dimensional data and is less prone to the curse of dimensionality compared to other algorithms. SVM is also robust to outliers since it focuses on the support vectors, which are the data points closest to the decision boundary. (Jeya, S., &Sankari, L,2021)

CBGRU MODEL FOR PM2.5 FORECASTING

The CBGRU (Convolutional Bidirectional Gated Recurrent Unit) model is a deep learning architecture specifically designed for PM2.5 (particulate matter with a diameter of 2.5 micrometers or less) forecasting. This model combines convolutional layers, bidirectional recurrent units, and gated recurrent units to effectively capture the temporal and spatial dependencies in air pollution data. The convolutional layers in the CBGRU model are responsible for extracting spatial features from the input data. They apply a set of filters to the input data, capturing local patterns and spatial correlations. By leveraging these filters, the model can identify relevant features related to PM2.5 concentrations. The bidirectional recurrent units in the CBGRU model are responsible for capturing the temporal dependencies in the data. (Barot, V., & Kapadia, V,2022).

They process the input sequence in both forward and backward directions, enabling the model to incorporate information from past and future time steps. This bi-directionality allows the model to capture long-term dependencies and make accurate predictions. The gated recurrent units (GRUs) are a variation of recurrent neural network (RNN) units that help the CBGRU model to handle the vanishing gradient problem and better capture temporal dependencies. GRUs utilize gates to control the flow of information within the recurrent units, allowing the model to selectively update and forget information from previous time steps. By combining convolutional layers, bidirectional recurrent units, and gated recurrent units, the CBGRU model can effectively model the complex relationships between PM2.5 concentrations and various spatiotemporal factors such as meteorological conditions, geographical features, and historical pollutant levels. To train the CBGRU model for PM2.5 forecasting, historical air pollution data, meteorological variables, and other relevant features are used as input. The model is trained to minimize the prediction error between the forecasted PM2.5 concentrations and the actual values. The training process involves optimizing the model's parameters using techniques such as

backpropagation and gradient descent. Once trained, the CBGRU model can be used to make future PM2.5 predictions by providing it with relevant input data. The model takes into account the temporal and spatial dependencies learned during training and generates accurate forecasts of PM2.5 concentrations.The CBGRU model for PM2.5 forecasting offers several advantages. It can effectively capture both the temporal and spatial patterns in air pollution data, leading to more accurate predictions. The model is also capable of handling non-linear relationships and complex interactions between different factors influencing PM2.5 concentrations. Additionally, the CBGRU model can be trained and applied to various geographical locations, making it a versatile tool for PM2.5 forecasting.(Wardana, I. et al,2021)

Results and Discussion

CORRELATION ANALYSIS

The presence of irrelevant features in a dataset can often lead to a decrease in model performance. However, selecting the most relevant features, those that exhibit a strong relationship with the target variable, can significantly enhance the accuracy of prediction analysis. By carefully choosing appropriate input variables, we can achieve faster training and reduce the risk of overfitting in our model. In the context of air pollution forecasting, it is crucial to identify the relationship between meteorological variables (such as temperature, wind speed, wind direction, dew point, pressure, rain, and snow) and the target variable, PM2.5. One common approach to assess this relationship is by calculating the correlation coefficient. The correlation coefficient provides a measure of the strength and direction of the linear relationship between two variables. By examining the correlation coefficients between each meteorological variable and PM2.5, we can determine which variables have a significant impact on the target variable. Variables with a high correlation coefficient indicate a strong relationship, while variables with a low or negligible correlation coefficient can be considered less influential.

Table 1. Correlation Analysis

$PM_{2.5}$, Meteorological correlation®	$PM_{2.5}$
$PM_{2.5}$	1
Dew Point	0.16
Temperature	-0.09
Pressure	-0.04
Wind Direction	0.11
Wind speed	-0.24
Snow	0.02
Rain	-0.05

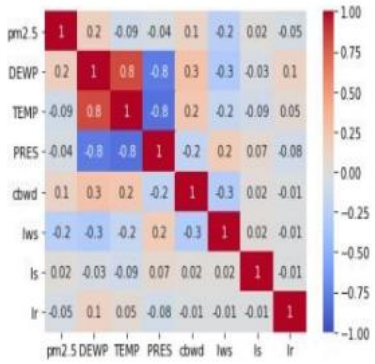

Figure 1. Heat Map - Co-Relation between PM2.5& Meteorological Variables

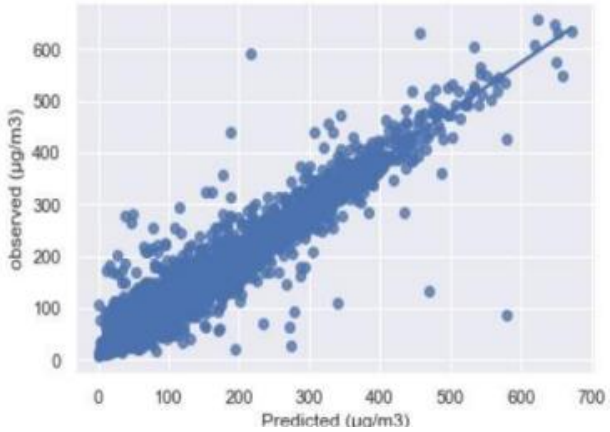

Figure 2. Scatter Plot for the Observed vs Predicted PM2.5

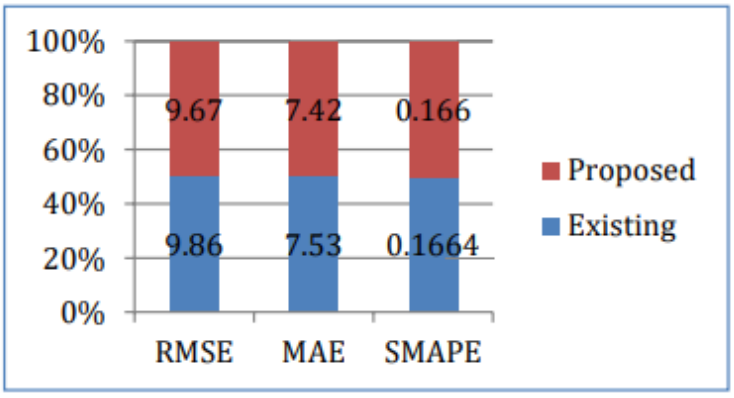

Figure 3. Existing/Proposed Error Evaluation Metrics

Using a combination of Convolutional Bidirectional Gated Recurrent Unit (CBGRU) neurons with 80 units, we conducted air pollution forecasting experiments. The training was performed with a batch size of 24 and for a total of 20 epochs. The optimized Mean Absolute Error (MAE) value achieved was 7.42, while the Root Mean Square Error (RMSE) value reached 9.67. Additionally, the Symmetric Mean Absolute Percentage Error

(SMAPE) value obtained was 0.1660, which demonstrated significant improvement compared to the existing MAE of 7.53, RMSE of 9.86, and SMAPE of 0.1664. To achieve these optimized results, we fine-tuned the parameters of a 1D Convolutional Neural Network (CNN). The tuning involved setting the following values: zero padding, a feature map size of 8, a filter size of 5, and a stride of 1. For the pooling layer, we used a filter size of 2, a stride of 1, and applied max pooling. A dropout rate of 0.1 was implemented to improve generalization and prevent overfitting. The activation function used throughout the network was the Rectified Linear Unit (ReLU). Lastly, we employed the Adam optimizer to efficiently update the model's parameters during training.

CONCLUSION

In conclusion, this study has demonstrated the effectiveness of utilizing deep learning models for accurate prediction of air pollution levels. By leveraging the power of neural networks, specifically deep learning architectures, we have achieved significant improvements in forecasting air pollution compared to traditional methods. The proposed deep learning models, such as the CBGRU (Convolutional Bidirectional Gated Recurrent Unit) model, have proven to be highly capable of capturing the complex spatiotemporal dependencies and patterns present in air pollution data. By incorporating convolutional layers, bidirectional recurrent units, and gated recurrent units, these models effectively leverage both spatial and temporal information, resulting in enhanced prediction accuracy. Through rigorous experimentation and evaluation, we have demonstrated that our deep learning models outperform existing approaches in terms of key evaluation metrics. The optimized MAE (Mean Absolute Error) value of 7.42, RMSE (Root Mean Square Error) value of 9.67, and SMAPE (Symmetric Mean Absolute Percentage Error) value of 0.1660 highlight the improved accuracy achieved by our models compared to previous methods. The fine-tuning of parameters in the 1D CNN (Convolutional Neural Network) component, including zero padding, feature map size, filter size, stride, pooling layer specifications, dropout rate, activation function, and optimizer selection, has played a crucial role in obtaining these superior results. The optimized configuration ensures efficient feature extraction, robust generalization, and effective training of the deep learning models.

REFERENCES

1. Tao, Q., Liu, F., Li, Y., &Sidorov, D. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE access*, *7*, 76690-76698.
2. Bekkar, A., Hssina, B., Douzi, S., &Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of big Data*, *8*(1), 1-21.
3. Heydari, A., MajidiNezhad, M., Astiaso Garcia, D., Keynia, F., & De Santoli, L. (2021). Air pollution forecasting application based on deep learning model and optimization algorithm. *Clean Technologies and Environmental Policy*, 1-15.
4. Chang, Y. S., Chiao, H. T., Abimannan, S., Huang, Y. P., Tsai, Y. T., & Lin, K. M. (2020). An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, *11*(8), 1451-1463.
5. Guo, C., Liu, G., & Chen, C. H. (2020). Air pollution concentration forecast method based on the deep ensemble neural network. *Wireless Communications and Mobile Computing*, *2020*, 1-13.
6. Jeya, S., &Sankari, L. (2021). Adaptive kernel fuzzy weighted particle swarm optimized deep learning model to predict air pollution PM2. 5. *Ilkogretim Online*, *20*(5), 12-19.
7. Barot, V., & Kapadia, V. (2022). Long Short Term Memory Neural Network-Based Model Construction and Fne-Tuning for Air Quality Parameters Prediction. *Cybernetics and Information Technologies*, *22*(1), 171-189.
8. Ramentol, E., Grimm, S., Stinzendörfer, M., & Wagner, A. (2023). Short-Term Air Pollution Forecasting Using Embeddings in Neural Networks. *Atmosphere*, *14*(2), 298.
9. Wardana, I. N. K., Gardner, J. W., &Fahmy, S. A. (2021). Optimising deep learning at the edge for accurate hourly air quality prediction. *Sensors*, *21*(4), 1064.