DOI 10.51885/1561-4212_2025_3_169
IRSTI 28.23.29

A CNN-LSTM HYBRID MODEL FOR PREDICTING AIR QUALITY AND DETECTING ANOMALIES WITH GAUSSIAN APPROXIMATION

ГИБРИДНАЯ МОДЕЛЬ CNN-LSTM ДЛЯ ПРОГНОЗИРОВАНИЯ КАЧЕСТВА ВОЗДУХА И ОБНАРУЖЕНИЯ АНОМАЛИЙ С ПОМОЩЬЮ ГАУССОВОЙ АППРОКСИМАЦИИ

ГАУСС ЖУЫҚТАУЫН ҚОЛДАНУ АРҚЫЛЫ АУА САПАСЫН БОЛЖАУ ЖӘНЕ АНОМАЛИЯЛАРДЫ АНЫҚТАУҒА АРНАЛҒАН ГИБРИДТІ CNN-LSTM МОДЕЛІ

Y. Vays [1], **A. F. Omojola** [1], **S. Rakhmetullina** [1], **A. Urkumbayeva** [1]*
[1]D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
*Corresponding author: Full name, e-mail: Urkumbayeva Aliya, *aurkumbaeva@edu.ektu.kz*

Keywords:
Air quality, Machine learning, Atmospheric pollution, Environmental monitoring, Anomaly detection.

ABSTRACT

Air pollution is a global issue affecting the health of people, the sustainability of the environment, and the planning of urban areas. The present work utilises a smart air quality data monitoring analysis system that uses machine learning algorithms in forecasting and studying atmospheric pollution concentration. Multi-pollutant forecasting in Ust-Kamenogorsk utilises the collaborative use of LSTM and CNN. The Gaussian approximation is used in detecting outliers and meteorological input is added in order to support predictive precision. The LSTM-CNN blended model was utilized in predicting the concentration of different contaminants, including PM2.5, PM10, NO_2, SO_2, CO, and O_3. The predictive accuracy of the model was average, considering its Root Mean Squared Error (RMSE) of 0.3297. Mean absolute error (MAE) was 0.2741, indicating differences in prediction ability among contaminants. However, R^2 score at -0.3210 suggests that the model needs to be tuned for greater predictability. Identification of outliers was done through residual analysis, which provided a 1.0 recall but poor precision of 0.0676, indicating high false positive rate. Despite its limitations, the model has the capacity to anticipate air quality in real time and detect anomalies. Future enhancements will include hyperparameter optimization, the addition of new data sources, and the refining of the anomaly detection method for greater accuracy and dependability. This contribution goes toward the development of intelligent air quality monitoring technologies to support data-driven environmental management and policy.

Түйінді сөздер:
ауа сапасы, машиналық оқыту, атмосфераның ластануы, қоршаған ортаны бақылау, ауытқуларды анықтау.

ТҮЙІНДЕМЕ

Ауаның ластануы – адам денсаулығына, қоршаған ортаның тұрақтылығына және қалалық аумақтарды жоспарлауға әсер ететін жаһандық мәселе. Бұл жұмыста Өскемен атмосферасындағы ластаушы заттардың шоғырлануын болжау және зерттеу үшін машиналық оқыту алгоритмдерін (LSTM және CNN) пайдаланатын

ауа сапасының мониторингі деректерін талдаудың зияткерлік жүйесі қарастырылады. Гаусс жуықтауы шығарындыларды анықтау үшін қолданылады және болжамның дәлдігін қолдау үшін метеорологиялық деректер қосылады. LSTM-CNN аралас моделі PM2.5, PM10, NO_2, SO_2, CO және O_3 сияқты әртүрлі ластаушы заттардың шоғырлануын болжайды. Модельді болжау дәлдігі орташа, оның орташа квадраттық қателігі (RMSE) 0,3297, орташа абсолютті қателік (MAE) 0,2741, бұл болжамды деректердегі дәлсіздікті көрсетеді. Дегенмен, R^2 -0.3210 көрсеткіші болжам дәлдігін жақсарту үшін модельді одан әрі реттеу қажет екенін көрсетеді. Шығарындыларды анықтау қалдықтарды талдау арқылы жүргізілді, ол 1,0 қайтарып алуды қамтамасыз етті, бірақ 0,0676 дәлдігі төмен, бұл жалған позитивтердің жоғары пайызын көрсетеді. Шектеулерге қарамастан, модель нақты уақыт режимінде ауа сапасын болжауға және ауытқуларды анықтауға қабілетті. Әрі қарай жақсартулар гиперпараметрлерді оңтайландыруды, жаңа деректер көздерін қосуды және дәлдік пен сенімділікті арттыру үшін ауытқуларды анықтау әдісін нақтылауды қамтиды. Бұл мақала деректерге негізделген экологиялық менеджмент пен саясатты қолдау үшін ауа сапасын бақылаудың интеллектуалды технологияларын әзірлеуге бағытталған.

Ключевые слова:
качество воздуха, машинное обучение, загрязнение атмосферы, мониторинг окружающей среды, обнаружение аномалий.

АННОТАЦИЯ

Загрязнение воздуха – глобальная проблема, влияющая на здоровье людей, устойчивость окружающей среды и планирование городских территорий. В данной работе рассматривается интеллектуальная система анализа данных мониторинга качества воздуха, использующая алгоритмы машинного обучения (LSTM и CNN) для прогнозирования и изучения концентрации загрязняющих веществ в атмосфере Усть-Каменогорска. Гауссова аппроксимация применяется для обнаружения выбросов, а метеорологические данные добавляются для поддержки точности прогноза. Смешанная модель LSTM-CNN прогнозирует концентрацию различных загрязняющих веществ, включая PM2.5, PM10, NO_2, SO_2, CO и O_3. Точность прогнозирования модели средняя, ее среднеквадратичная ошибка (RMSE) 0,3297, средняя абсолютная ошибка (MAE) 0,2741, что указывает на неточность в предсказанных данных. Однако показатель R^2 -0,3210 свидетельствует о том, что модель нуждается в дальнейшей настройке для повышения точности прогноза. Идентификация выбросов проводилась с помощью анализа остатков, который обеспечил 1,0 отзыв, но низкую точность в 0,0676, что указывает на высокий процент ложных срабатываний. Несмотря на ограничения, модель способна прогнозировать качество воздуха в реальном времени и выявлять аномалии. Дальнейшие усовершенования будут включать оптимизацию гиперпараметров, добавление новых источников данных и доработку метода обнаружения аномалий для повышения точности и надежности. Данная статья направлена на разработку интеллектуальных технологий мониторинга качества воздуха для поддержки экологического менеджмента и политики на основе данных.

INTRODUCTION

Air pollution is one of today's most pressing environmental challenges. Air quality should be evaluated and assessed on a regular basis to ensure better living circumstances. The US Environmental Protection Agency (EPA) uses the air quality index (AQI) to standardize air

quality. However, AQI demands precise and reliable sensor data as well as complicated calculations, which portable air quality measuring equipment cannot provide. Air pollution is a major environmental concern that affects millions of people throughout the world (Singh and Singh, 2017; Manisalidis et al. 2020). Increased industrialization and urbanization in the past decades have resulted in excessive emissions of air pollutants such as particulate matter (PM2.5, PM10), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3) (Hoque et al. 2020; Meo et al. 2024; Meo et al. 2024). These air pollutants have serious health impacts, from respiratory and cardiovascular diseases to other chronic diseases. Further, air pollution speeds up environmental deterioration by altering climatic patterns and breaking down ecosystems. The need for reduced air pollution necessitates the development of advanced monitoring and forecasting systems that give policymakers and urban planners timely and accurate information.

Traditional air quality monitoring is based on stationary sensor networks that collect information from various points (Mihaita et al. 2019). While these systems are able to offer location-specific pollution measurements, they are associated with many limitations including the need for extensive maintenance, spatially limited coverage, and the slowness of data processing. Historically, there are traditional forecast models based on statistical and regression-based methods incapable of addressing complex inter-relationships among numerous different contaminants and weather indicators (Zhang et al. 2022). There is an interesting alternative presented by machine learning with the help of big data to discover complex patterns and enhance predictability accuracy (Qolomany et al. 2019; Ahmad et al. 2022).

Development in artificial intelligence and machine learning over the past couple of years has made it possible to develop very complex models that can quite effectively predict air quality (Tien et al. 2022; Ma et al. 2019). Deep learning techniques, such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), have been found to perform well in the process of time-series analysis and spatial data feature extraction of pollution data. The work employs the synergy of sequence learning with LSTM and feature learning with CNN to create a cost-effective and scalable model for air quality monitoring. In addition to pollutant concentration estimation, the model employs anomaly detection techniques for identifying unexpected surges in pollution levels, hence improving air quality estimation accuracy.

As a result of increased industrialization and urbanization, air pollution has become a serious issue necessitating effective and scalable monitoring strategies. Installing and maintaining conventional air quality monitoring stations is excessively expensive in most locations, particularly developing countries. Secondly, existing predictive techniques are rudimentary regression schemes that are devoid of the potential to capture the intricate spatiotemporal relationships of contaminants. With advancements in machine learning, in particular deep learning, it is now possible to use it to create data-driven predictive models to improve air quality monitoring and forecasting.

The motivation behind this work is the requirement for an intelligent system not only to predict pollutant concentration but also to identify anomalies that can indicate spikes in pollution or sensor malfunction. The industrially polluted region of Ust-Kamenogorsk is an ideal case study to experiment with the techniques described above. The use of LSTM networks for learning temporal patterns and CNNs for feature learning improves predictability, whereas Gaussian approximation ensures robust anomaly detection.

The research contribution presents a new hybrid model that maximizes multi-pollutant prediction by combining CNN for feature learning and LSTM for sequence modeling of time. The model improves its predictive accuracy and robustness by including climatic factors like temperature, humidity, wind speed, and atmospheric pressure. The research suggests an anomaly detection system that detects anomalies in the pollution pattern and sends an early

warning for unexpected air quality incidents. It is learned and tested on air quality records from Ust-Kamenogorsk, depicting its usability in real-world environmental monitoring. Performance of the model is compared by RMSE (0.3297), MAE (0.2741), and R^2 (-0.3210) to know where improvement would be needed.

LITERATURE SURVEY

Unnikrishnan and Rajeswari (2024) developed an ambient air pollution early warning system and a daily Air Quality Index (AQI) forecasting system over road networks. Their approach integrates a Gaussian dispersion model and a deep learning algorithm for modeling pollutant dispersion and predicting AQI values more accurately. The model effectively captures pollution dynamics caused by vehicle emissions and environmental factors. However, a key gap is the model's use of static emission rates and limited real-time traffic variability, which can undermine prediction robustness for dynamic traffic regimes. In addition, the model's scalability to larger and more complicated urban domains was not explored to the greatest degree. Future research should investigate adaptive models that use real-time traffic data, feature dynamic emission inventories, and test the system across a range of urban domains to provide enhanced reliability and generalization.

Borah (2024) implemented a Deep Learning-Based Anomaly Detection Approach for Air Pollution Assessment. A deep learning approach was proposed for detecting anomalies in air quality monitoring. To detect spatiotemporal anomalies, we used a combination of CNNs and LSTM networks. The model has a 96% recall rate but has moderate false positive rates. Deep learning algorithms are capable of detecting air pollution anomalies. Real-time implementation is a significant computing difficulty.

Wei et al. (2023): LSTM-Autoencoder-Based Anomaly Detection for Indoor Air Quality Time Series Data. I proposed an LSTM-Autoencoder model for detecting abnormalities in indoor air quality time-series data. Deep learning-based LSTM-Autoencoders were used for time series prediction. For air quality anomaly detection, we achieved an F1-score of 0.87. Deep learning approaches were demonstrated to be beneficial for detecting unanticipated pollution incidents. The anomalies identified were not interpretable, and they could not be explained.

Nguyen et al. (2023) introduced an air pollution forecast model based on a Long Short-Term Memory Bayesian Neural Network (LSTM-BNN). The method combines the temporal sequential learning capability of LSTM networks and the Bayesian inference to produce point forecasts and uncertainty estimation. The method was implemented on actual air quality data and outperformed regular LSTM and deterministic models in forecasting accuracy and uncertainty estimation. Performance showed that LSTM-BNN performed better than benchmarks for pollutant concentration estimation such as PM2.5 and NO_2. Its most notable limitation stated was the tremendous computational cost using Bayesian neural networks, limiting real-time deployment. Future work must involve alleviating computational cost and exploration of light-weight Bayesian architecture to enable rapid and scalable deployment in real air quality sensing systems.

El-Shafeiy et al. (2023) investigated real-time anomaly detection for water quality sensor monitoring using a multivariate deep learning technique. We proposed a deep learning approach for detecting anomalies in water quality sensors in real time. Used a multivariate LSTM-based anomaly detection framework. Accuracy in detecting anomalies exceeded 93%. Deep learning models were shown to be excellent at detecting anomalies in environmental monitoring. Applied to water quality monitoring, with little direct relevance to air pollution data.

Gilik, Ogrenci, and Ozmen (2022) proposes a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for air quality prediction. Spatial features are extracted using CNN, and temporal dependencies in the data are modeled using LSTM, achieving greater prediction accuracy compared to classical

machine learning and individual deep learning models. The method is tested on real datasets and performs better. However, one of the largest gaps in the research is the absence of detailed exploration of external influencing variables such as meteorological variables and sudden environmental events, which can significantly impact air quality but were not meaningfully integrated into the model. The applicability of the model across geographical locations is also untested. Future research must focus on incorporating larger-scale contextual information, enhancing the robustness of models to varying environmental conditions, and developing lightweight architectures that can be deployed in real time.

Zhang, Han, Li, and Lam (2022) proposed Deep-AIR as a hybrid CNN-LSTM model for accurate forecasting and estimation of metro cities for air pollution and metro cities' air quality. The CNN identifies spatial features from heterogeneous city urban data, and the LSTM depicts temporal changes in air quality. The model was tested and trained on large data sets from a range of monitoring stations in urban areas, with both higher spatial resolution and accuracy than current models. Results showed that Deep-AIR was able to detect both local gradients of pollution and larger temporal trends. A significant limitation was that the model relied on dense sensor networks, so it was less well-suited for use in areas with low data coverage. Future studies need to concentrate on improving the generalization model using sparse data and applying transfer learning techniques in a bid to tailor the model to other diverse urban environments with limited retraining.

Goh et al. (2021) created a real-time in-vehicle air quality monitoring system using a machine learning prediction algorithm. Created a real-time air quality monitoring system for vehicle applications. Machine learning regression techniques, such as Random Forest and Gradient Boosting, were used to estimate indoor vehicle air pollution concentrations. With an accuracy of 85%, the system correctly determined and predicted air pollution concentrations in automobiles. The researchers emphasized the importance of in-vehicle pollution monitoring, particularly for urban passengers. The data was limited to controlled studies in specific vehicle types, reducing its applicability to real-world driving settings.

Jesus, G., Casimiro, A., & Oliveira (2021) Machine Learning for Reliable Outlier Detection in Environmental Monitoring Systems. We used machine learning to investigate outlier detection in air sensor pollution data. For detecting abnormal levels of pollution, Isolation Forest and One-Class SVM were used. Outliers were detected in 95% of cases with few false positives. Machine learning was shown to be a solid method to distinguish sensor errors from true pollution anomalies. Multi-modal sensor data fusion and ensemble techniques were not investigated.

Dai, Huang, Wang, Zeng, and Zhou (2021) introduced an air pollutant concentration predictive model, which is a combination of multi-scale one-dimensional Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Temporal and spatial patterns of air quality data are expected to be learned by the model. Spatial patterns of various scales are learned by multi-scale CNN, and temporal relationships are learned by LSTM. The method was applied to Xi'an, China, air quality observations with the primary pollutants like PM2.5 and PM10. The result indicated that the model introduced in this paper performed better than typical machine learning methods and isolated deep learning models. The study acknowledged limitations like the model's reliance on historical data without taking external dynamic variables like weather conditions and urban activities into consideration. Future studies should focus on incorporating real-time socio-economic and environmental data for further improvement in the performance of predictions and resilience.

Zhang et al. (2019), Predictive Data Feature Exploration-Based Air Quality Prediction Method. To improve the precision of air quality prediction, we proposed a new feature selection technique. Combination of feature engineering and ensemble learning approaches. Improved air quality forecast precision by 12% compared to baseline models. Feature selection considerably

enhances the precision of predicted values in machine learning models. Computationally expensive, needing plenty of resources to train.

SUMMARY OF LITERATURE

Current machine learning-based air quality monitoring models are promising but have limitations, including the inability to handle intricate pollutant interactions, high computational cost, a lack of real-time deployment, and the inability to differentiate between real anomalies and sensor faults. Although deep learning models such as CNN and LSTM enhance accuracy, they are plagued by explainability and scalability problems. Generalizability is limited by the fact that many research concentrate on particular contaminants or indoor air quality. Furthermore, it can be challenging for anomaly detection methods to distinguish between real pollution incidents and sensor malfunctions. This study develops a scalable and interpretable CNN-LSTM-based model with Gaussian approximation for improved real-time anomaly detection and air quality prediction in order to get over these restrictions.

MATERIAL AND METHODS

The study makes use of meteorological factors including temperature, humidity, wind speed, and atmospheric pressure as well as air quality measurements for pollutants like PM2.5, PM10, NO2, SO2, CO, and O3 from monitoring stations in Ust-Kamenogorsk. The following is how the data is preprocessed. The CNN+LSTM Deep Learning workflow is shown in figure 1.

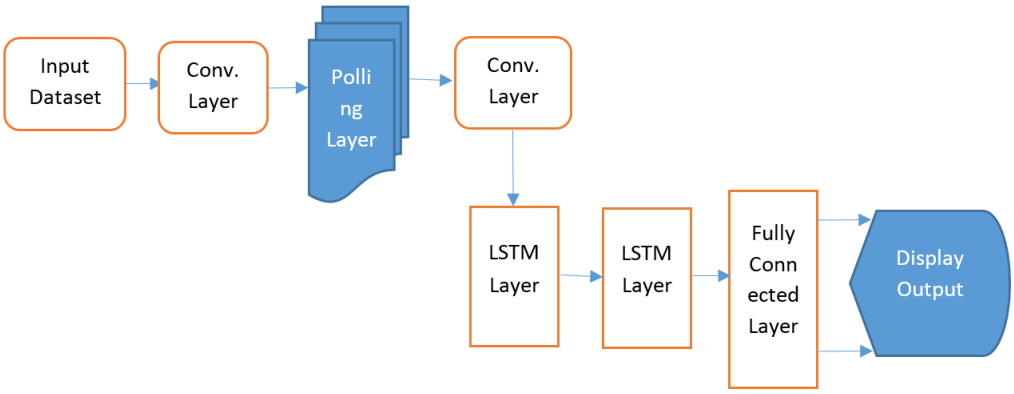

Figure 1. CNN+LSTM workflow

Note – compiled by the authors

Normalization: Pollutant values between 0 and 1 are normalized using min-max scaling.

Handling Missing Data: Linear interpolation is used to fill in missing numbers.

Time-Series Structuring: To create sequences for model training, a sliding window procedure is used.

CNN-LSTM with Gaussian Approximation Model Architecture

The following are included in the hybrid model in the suggested work:

CNN Layers: Use 1D convolutional layers to extract spatial information from air quality data.

Long-term temporal relationships in pollution trends are captured by LSTM layers.

Fully Connected Layers: Predict pollution with multiple outputs.

Gaussian Approximation: Used to identify anomalies by looking at residual model prediction errors.

Algorithm 1: The model Used in the Study
Input: Time series data on air quality is taken in by the input layer.
Output: Performance Evaluation.

Step 1: Start

Step 2: Conv1D Layer: Acquires knowledge of spatial dependencies (64 filters, kernel size 3, ReLU activation).

Step 3: Additional spatial patterns are extracted by the Conv1D Layer (32 filters, kernel size 3, ReLU activation).

Step 4: Long-term dependencies in pollutant concentrations are modeled by the LSTM Layer (64 units).

Step 5: Pollutant concentration forecasts are produced by the fully connected dense layer (6 units).

Step 6: The Gaussian Approximation Layer calculates residuals in order to identify anomalies.

Step 7: Stop

Training and Evaluation of the Model. Mean Squared Error (MSE) is the training loss function.

Adam is an optimization algorithm with a 0.001 learning rate.

Model evaluation metrics include R2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Evaluation of Anomaly Detection: F1-score, precision, and recall are used to gauge detection accuracy.

RESULT AND DISCUSSION

The CNN-LSTM hybrid model for air quality prediction is demonstrated in algorithm 1. An input layer that takes in a time-series data set of pollutant levels over time (24,6) is proposed. Short oscillations are identified with the assistance of the Conv1D layers, where spatial correlations between the air quality are learned from the data. The LSTM layer is best for predictions as it can capture long dependence on pollution rates. For various contaminants, the thick output layer produces multi-class predictions as displayed in Figure 2.

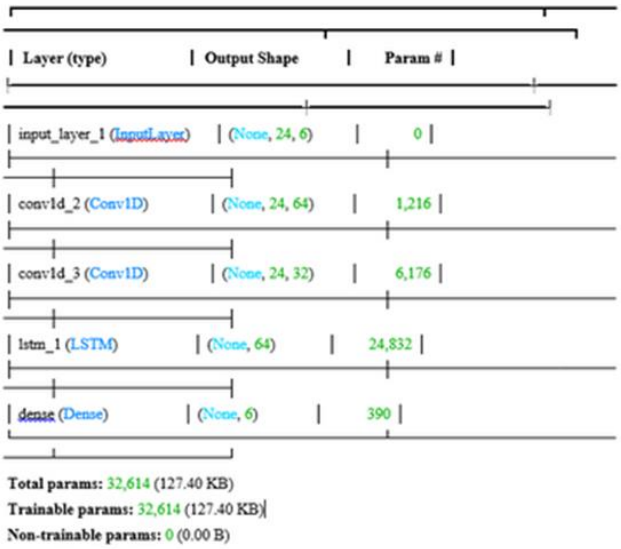

Figure 2. Multi Pollutant Prediction

Note – compiled by the authors

The 32,614 total parameters demonstrate the model's complexity without making it computationally expensive. By removing surprise deviations, the Gaussian Approximation is integrated into anomaly detection, improving reliability. This model efficiently captures geographical and temporal correlations in air quality data, a major improvement over conventional machine learning methods. Figure 3 displays the Training versus the Validation Loss Plot.

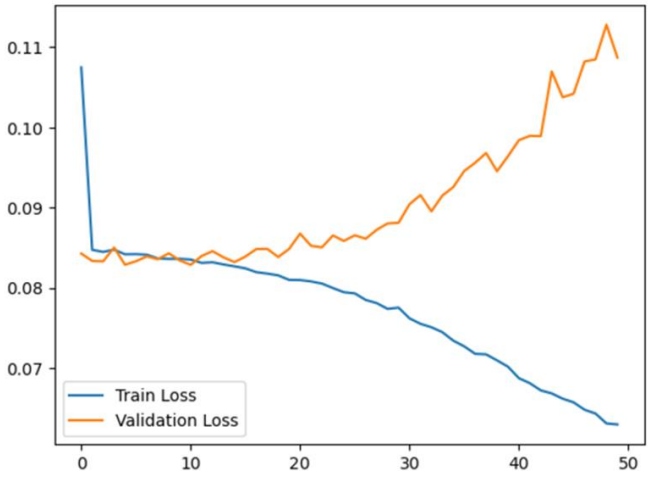

Figure 3. Training and Validation Loss Plot

Note – compiled by the authors

Train loss (blue) and validation loss (orange) are shown against 50 epochs in the second graph, which displays the model's learning curve. The initial drop of the two lines indicates good learning. However, after about 15 epochs, there are indications of overfitting as the validation loss begins to rise while the train loss continues to fall. When a model overfits, it performs well on training data but poorly on fresh, unseen test data. Regularization strategies including early halting, dropout layers, and hyperparameter adjustment can help to lessen this problem. The model may be memorizing patterns instead of generalizing, which is a problem that requires more fine-tuning, as evidenced by the steady increase in validation loss. The Actual and predicted PM2.5 Level figure is shown in figure 4.

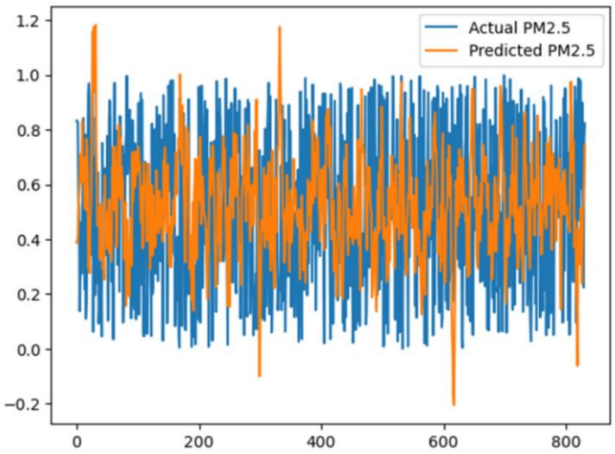

Figure 4. Actual and Predict PM2.5 Levels

Note – compiled by the authors

A comparison of the projected and actual PM2.5 levels is shown by this plot. The blue line represents the model's prediction, while the orange line represents the actual PM2.5 measurement. Although the model typically tracks the trend of the real data, there are occasional peaks and troughs where it deviates. Moderate accuracy is indicated by the RMSE value of 0.3297 and the MAE value of 0.2741. Conversely, negative R2 values (-0.3210) indicate that the model is poorly fitted and that adjustment is necessary. The variance in PM2.5 concentration is influenced by external environmental conditions, and accuracy can be improved by include more robust feature selection, such as wind speed or vehicle density. The DBSCAN clustering anomalies detection is shown in figure 5.

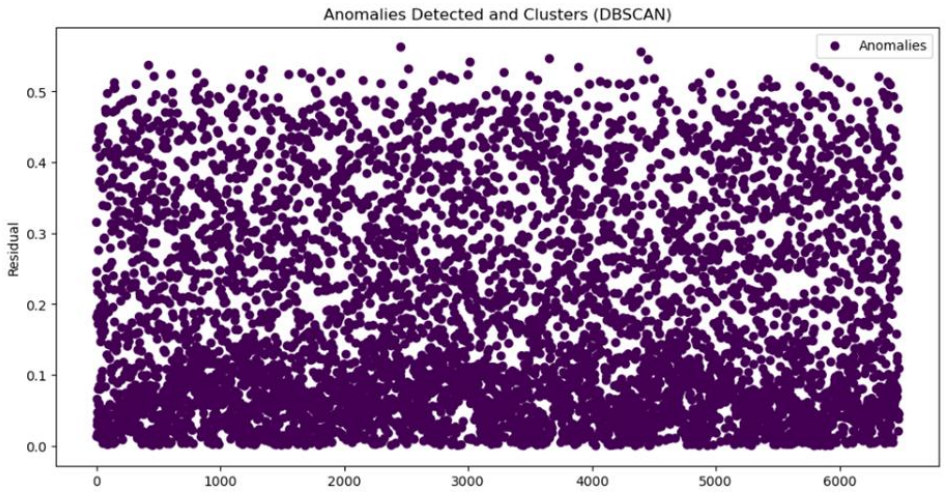

Figure 5. Anomalies detected using DBSCAN Clustering

Note – compiled by the authors

The plot uses the DBSCAN clustering algorithm to exhibit the anomalies (purple dots) that were found. Extreme deviations were identified using residual analysis, which is calculated by subtracting the observed value from the anticipated value. Anomalies were three standard deviations from the mean residual value. To detect noise and group similar abnormalities, DBSCAN was employed. However, DBSCAN's efficacy was limited since either only one cluster was found or the outliers were categorized as noise. This suggests that either the data did not contain distinct anomaly patterns or the DBSCAN hyperparameters (min_samples and eps) were not properly configured. The model had a very high recall in properly identifying anomalies (recall = 1.0), but it also produced a large number of false positives, as evidenced by the poor F1-score value of 0.0676 and low accuracy of 0.1342. More sophisticated techniques like autoencoders or isolation forests could be used to detect anomalies more effectively.

DISCUSSION

Pollutant concentrations such as PM2.5, PM10, NO2, SO2, CO, and O3 were predicted using the hybrid LSTM-CNN model. The following performance indicators were noted throughout the model's training and testing:

1. Root Mean Squared Error (RMSE): The model's RMSE of 0.3297 indicated a moderate level of prediction error. Although this suggests that the model can accurately depict the pollutant concentration trend, it could be improved to lower the inaccuracy.

2. MAE (Mean Absolute Error): The model's average prediction error is approximately 27.4%, with an MAE of 0.2741. This shows that the model can estimate pollutant quantities, but it also shows that prediction accuracy varies throughout contaminants.

3. R^2 (R-Squared): A score of -0.3210 for R^2 suggests that the model does not match the test data adequately. R^2 negative values typically indicate that the model's prediction is worse than the average of the actual values. This suggests that the model still need fine-tuning to improve prediction accuracy.

4. Test Loss: With a test loss of 0.1087, the overall model loss is rather small, but it does highlight areas for performance improvement.

KEY FINDINGS

The growing validation loss indicates that to avoid overfitting, regularization techniques like dropout layers and early stopping must be used.

Although it needs more accurate peak value projections, the CNN-LSTM model can identify pollution patterns. To reduce prediction errors, feature engineering may entail including more environmental variables.

Reasonable anomaly segmentation could not be produced using DBSCAN clustering. Anomaly classification would be enhanced by additional techniques like hybrid models or dynamic thresholding.

The model shows promise for monitoring air quality in real-time, but it needs to be optimized for application in different metropolitan locations.

THREAT TO VALIDITY

The validity of this study is threatened by issues with model development, statistical results, generalizability, and data dependability. Missing values and sensor errors compromise internal validity, while the data's spatial specificity restricts outward validity. Due to the underrepresentation of significant environmental variables, feature selection compromises construct validity. Overfitting and low precision (0.0676) in identifying anomalies put statistical validity at risk, increasing the likelihood of false positives. Validation is further limited by the absence of labeled anomaly data. Future research should focus on enhancing robust validation methods, feature selection, hyperparameter tweaking, and diverse datasets.

CONCLUSION

The study recommended developing a CNN-LSTM hybrid model for anomaly detection and air quality prediction that incorporates an in-painted Gaussian approximation. With a reasonable RMSE of 0.3297 and MAE of 0.2741, the model demonstrated efficacy in learning the trend of pollutants and may be used as a real-time air quality forecasting model. The model functioned well overall, despite the negative R2 (-0.3210) value suggesting room for improvement to boost prediction accuracy. The anomaly detection process based on DBSCAN clustering has a 1.0 recall rate, a poor precision of 0.0676, and a significant rate of false positives. Overfitting was shown by the validation loss curve, indicating that while the model did well on training data, its generalization ability is still lacking. Despite these obstacles, this study presents data-driven strategies for reducing environmental pollution and aids in the development of machine learning-enhanced air quality monitoring systems. Future research should focus on using transfer learning and integrating additional environmental elements (such as traffic congestion and industrial toxins) to raise forecast accuracy to improve the system's efficiency and scalability as suggested.

To strengthen the model and avoid overfitting, dropout layers, early halting, and hyperparameter adjustment are used.

REFERENCES

Ahmad, T., Madonski, R., Zhang, D., Huang, C., & Mujeeb, A. (2022). Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. Renewable and Sustainable Energy Reviews, 160, 112128. https://doi.org/10.1016/j.rser.2022.112128

Borah, A. (2024). Deep Learning based Anomaly Detection Approach for Air Pollution Assessment. IEEE Transactions on Big Data.

Dai, H., Huang, G., Wang, J., Zeng, H., & Zhou, F. (2021). Prediction of air pollutant concentration based on one-dimensional multi-scale CNN-LSTM considering spatial-temporal characteristics: A case study of Xi'an, China. Atmosphere, 12(12), 1626. https://doi.org/10.3390/atmos12121626

El-Shafeiy, E., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique. Sensors, 23(20), 8613. https://doi.org/10.3390/s23208613

Gilik, A., Ogrenci, A. S., & Ozmen, A. (2022). Air quality prediction using CNN+ LSTM-based hybrid deep learning architecture. Environmental science and pollution research, 1-19. DOI: 10.1007/s11356-021-16227-w

Goh, C. C., Kamarudin, L. M., Zakaria, A., Nishizaki, H., Ramli, N., Mao, X., ... & Elham, M. F. (2021). Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. Sensors, 21(15), 4956.

Hoque, M. M., Ashraf, Z., Kabir, H., Sarker, E., & Nasrin, S. (2020). Meteorological influences on seasonal variations of air pollutants (SO2, NO2, O3, CO, PM2. 5 and PM10) in the Dhaka megacity. Am. J. Pure Appl. Biosci, 2(2), 15-23. DOI:10.34104/ajpab.020.15023

Jesus, G., Casimiro, A., & Oliveira, A. (2021). Using machine learning for dependable outlier detection in environmental monitoring systems. ACM Transactions on Cyber-Physical Systems, 5(3), 1-30. https://doi.org/10.1145/3445812

Ma, J., Cheng, J. C., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmospheric Environment, 214, 116885.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. Frontiers in public health, 8, 14. doi: 10.3389/fpubh.2020.00014

Meo, S. A., Salih, M. A., Alkhalifah, J. M., Alsomali, A. H., & Almushawah, A. A. (2024). Effect of air pollutants particulate matter pm2. 5, pm10, carbon monoxide (co), nitrogen dioxide (no2), sulfur dioxide (so2), and ozone (o3) on fractional exhaled nitric oxide (feno). Pakistan Journal of Medical Sciences, 40(8), 1719. doi: 10.12669/pjms.40.8.9630

Meo, S. A., Salih, M. A., Alkhalifah, J. M., Alsomali, A. H., & Almushawah, A. A. (2024). Environmental pollutants particulate matter (PM2. 5, PM10), Carbon Monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), and Ozone (O3) impact on lung functions. Journal of King Saud University-Science, 103280. DOI: 10.26355/eurrev_202401_35079

Mihăiţă, A. S., Dupont, L., Chery, O., Camargo, M., & Cai, C. (2019). Evaluating air quality by combining stationary, smart mobile pollution monitoring and data-driven modelling. Journal of cleaner production, 221, 398-418. https://doi.org/10.1016/j.jclepro.2019.02.179

Nguyen, H. A., Ha, Q. P., Duc, H., Azzi, M., Jiang, N., Barthelemy, X., & Riley, M. (2023). Long short-term memory Bayesian neural network for air pollution forecast. IEEE Access, 11, 35710-35725.

Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J., & Fong, A. C. (2019). Leveraging machine learning and big data for smart buildings: A comprehensive

survey. IEEE Access, 7, 90316-90356.
Singh, R. L., & Singh, P. K. (2017). Global environmental problems. Principles and applications of environmental biotechnology for a sustainable future, 13-41.
Tien, P. W., Wei, S., Darkwa, J., Wood, C., & Calautit, J. K. (2022). Machine learning and deep learning methods for enhancing building energy efficiency and indoor environmental quality–a review. Energy and AI, 10, 100198. https://doi.org/10.1016/j.egyai.2022.100198
Unnikrishnan, A., & Rajeswari, S. (2024). Forecasting Daily Air Quality Index and Early Warning System for Estimating Ambient Air Pollution on Road Networks Using Gaussian Dispersion Model with Deep Learning Algorithm. Tehnički glasnik, 18(4), 549-559.
Wei, Y., Jang-Jaccard, J., Xu, W., Sabrina, F., Camtepe, S., & Boulic, M. (2023). LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. IEEE Sensors Journal, 23(4), 3787-3800.
Zhang, Q., Han, Y., Li, V. O., & Lam, J. C. (2022). Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE Access, 10, 55818-55841.
Zhang, Q., Han, Y., Li, V. O., & Lam, J. C. (2022). Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE Access, 10, 55818-55841. Doi:10.1109/ACCESS.2022.3174853
Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., ... & Huang, L. (2019). A predictive data feature exploration-based air quality prediction approach. IEEE Access, 7, 30732-30743.

Авторлар туралы мәліметтер
Информация об авторах
Information about authors

Вайс Юрий Андреевич – техника ғылымдарының кандидаты, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
Вайс Юрий Андреевич – кандидат технических наук, Восточно-Казахстанский технический университет им. Д. Серикбаева, г. Усть-Каменогорск, Казахстан
Vays Yuriy Andreevich – Candidate of Technical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan,
e-mail: YuVais@edu.ektu.kz,
ORCID: 0000-0002-2964-8260,

Омоджола Айогоке Феликс – Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен, Қазақстан
Омоджола Айогоке Феликс – Восточно-Казахстанский технический университет им. Д. Серикбаева, Усть-Каменогорск, Казахстан
Omojola Ayogoke Felix – D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan,
e-mail: felixayogoke1@gmail.com
ORCID: https://orcid.org/0009-0005-0297-6044

Рахметуллина Сәуле Жәдігерқызы – техника ғылымдарының кандидаты, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
Рахметуллина Сауле Жадыгеровна – кандидат технических наук, Восточно-Казахстанский технический университет им. Д. Серикбаева, Усть-Каменогорск, Казахстан
Rakhmetullina Saule Zhadygerovna – Candidate of Technical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan,
e-mail: SRakhmetullina@edu.ektu.kz,
ORCID: 0000-0002-1729-3343

Уркумбаева Алия Муратовна – техника ғылымдарының кандидаты, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан
Уркумбаева Алия Муратовна – кандидат технических наук, Восточно-Казахстанский технический университет им. Д. Серикбаева, г. Усть-Каменогорск, Казахстан
Urkumbayeva Aliya Muratovna – Candidate of Technical Sciences, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan,
e-mail: aurkumbaeva@edu.ektu.kz,
ORCID: 0009-0008-9630-6448