IRSTI 28.23.37 https://doi.org/10.26577/jpcsit20253101

Meyir Yedilkhan[1], Azamat Berdyshev[2], Maksat Galiyev[3], Timur Merembayev[4,*]

[1] National Research Nuclear University MEPhI, Moscow, Russia
[2] International University of Information Technology, Almaty, Kazakhstan
[3] Suleyman Demirel University, Kaskelen, Kazakhstan
[4] Institute of Information and Computational Technologies CS MSHE RK, Almaty, Kazakhstan
*e-mail: timur.merembayev@gmail.com

AIR QUALITY PREDICTION BASED ON THE LSTM WITH ATTENTION USING METEOROLOGICAL DATA IN URBAN AREA IN KAZAKHSTAN

Abstract. This study investigates air pollution prediction in urban Kazakhstan, specifically focusing on Almaty, utilizing machine learning models, LightGBM, and Long Short-Term Memory (LSTM) with an attention mechanism. The research addresses the limitations of current air quality monitoring systems and aims to improve the accuracy of predicting PM2.5 and PM10 concentrations using meteorological data. Results demonstrated that while LightGBM efficiently handled tabular data, LSTM with attention exhibited predictive accuracy by capturing temporal dependencies and handling data variability more effectively. LSTM with attention achieved RMSE values of 5.54 and 5.69 for PM2.5 and PM10, respectively, compared to LightGBM's 4.75 and 5.76. The findings also highlight correlations between pollution levels and environmental conditions such as time of day, wind direction, and temperature. The results suggest that early-morning hours tend to have lower PM2.5 concentrations, while pollution levels generally decline as air temperature rises. These insights reinforce the potential of deep learning models to improve urban air quality management by enabling more precise and adaptive forecasting. We conclude that LSTM with attention is better suited for air quality predictions in complex urban environments, especially under dynamic meteorological conditions.

Key words: Air pollution, LSTM with attention, LightGBM, PM2.5, PM10, Kazakhstan.

1. Introduction

Environmental pollution and climate change are among the global community's most serious challenges. Due to climate-related water shortages, agricultural and industrial development, rapid growth of motor transport, cities, and insufficient environmental control, Kazakhstan is among the most difficult [1]. For example, in 2011, concentrations of key pollutants (PM10, NO2, and SO2) exceeded the annual limit values of the European Union (EU) in ten out of eleven selected cities [2]. In 2022, Kazakhstan ranked 33rd out of 115 countries in the world ranking of urban pollution levels (the higher the rank, the higher the pollution) [3]. The persistent nature of urban pollution is associated with the geographical features of some cities in foothill basins, such as Almaty (Figure 1), and industrial facilities emitting hazardous substances (for example, Oskemen).

In Kazakhstan, systems inform the population about air pollution, such as the airkaz.kz portal. However, despite all its undoubted advantages, such a portal is limited to measuring air pollution in certain places, giving preference to certain types of pollution (micro dust particles PM2.5 in the ground layer of the atmosphere). The lack of available data limits modern analysis and forecasting methods using machine learning methods (ML).

2. Materials and Methods

2.1. Data collection

Data collection is an important step before data analysis. Historical data analysis helps to identify useful and necessary patterns in a large data set. With the help of data analysis, various methods can be applied to extract useful information, make predictions, or detect anomalies [10-12].

Air pollution occurs for several reasons; for example, the air is mixed with harmful gases such as NO2, SO2, ozone, chlorofluorocarbons, CO2, and carbon monoxide. NO2 is released into the environment due to coal, fuel, oil, and gas combustion. It is a dangerous and noticeable air pollutant. Automobile exhaust is the main source of carbon monoxide. Carbon monoxide enters our body by binding to heme proteins, which affects the brain and cardiovascular system. Air pollution also occurs due to particulate matter such as PM2.5 and PM10. When PM particles are inhaled, they are retained in the tissues, causing harm to both the respiratory and respiratory systems. Air pollutants are harmful to human and health [13, 14, 18]. Air Quality Index (AQI) is a numerical value defined or set to inform the public about the air quality. The public can be advised to take appropriate precautions, and the responsible government agency can take necessary actions. The higher the AQI value, the worse the air quality. Each country has its standards using standard AQI ranges. Specific colors are used to indicate air quality and corresponding public health recommendations. This calculation used different values of air pollutants collected at meteorological stations [15]. The study aims to compare open APIs for collecting and analyzing environmental indicators, focusing on air quality in Almaty, Kazakhstan. We assessed the strengths, limitations, and usability of different open APIs (such as OpenWeather, Stormglass, Weatherbit, Yandex Weather, and AQICN) in terms of providing accurate and timely data for environmental monitoring.

Studies on air quality prediction have been conducted using various methods, including machine learning methods, statistical models, and deep learning approaches. Some of the most commonly used methods include convolutional neural networks [7], regression neural networks, fuzzy prediction models with empirical mode decomposition [8], and deep spatiotemporal graph networks [8].

Machine learning methods have been widely used to predict air quality, especially when using deep learning models. For example, a deep learning-based approach was proposed to predict air quality [17] using landscape photographs taken by mobile cameras. The model captured the complex relationships between air quality and environmental factors, such as temperature and humidity, and achieved accurate air quality predictions compared to traditional methods.

Another study used a regression neural network to predict air quality [5]. The model improved the prediction accuracy and reduced the time complexity of traditional methods. In addition, a deep spatiotemporal graph network with optimization was proposed to study the changing pattern of time series and the effect of spatial spreading [7]. The model achieved high accuracy in predicting PM2.5 concentration.

Statistical models, including linear regression [5] and fuzzy reasoning, have also been used to predict air quality. Linear regression has been used to predict the values of air pollutants such as PM2.5, PM10, SO2, NO2, NO, and CO [5]. Fuzzy reasoning has been used to evaluate the importance of pollutants in air quality assessments.

Empirical mode decomposition models of fuzzy prediction have also been proposed for air quality prediction [6]. These models use fuzzy logic to decompose data into different modes and predict air quality based on the decomposed modes.

Overall, air quality prediction research has focused on developing accurate and efficient air quality prediction methods using various data sources, including monitoring data, landscape photographs, and environmental factors.

2.2. Regression model – LightGBM

This research paper used two machine learning approaches, LightGBM and LSTM, to predict PM2.5 and PM10. LightGBM is a relatively new algorithm based on the gradient descent algorithm with extensive machine learning and data science applications. The main problem with gradient boosting algorithms is that they evaluate all the data to determine potential split points, which affects performance. LightGBM was adapted to improve the efficiency of the optimal search process [10,11].

Another reason LightGBM was chosen is that ensemble boosting algorithms have proven to solve practical classification and regression problems. This algorithm can provide high performance with multi-class imbalanced data and has shown efficiency in solving regression problems for machine learning competitions.

A feature of using LightGBM for data regression problems with time series data is modifying the input data into a tabular form. In particular, datetime needs to be transformed from a

time format to a numeric one. In this study, we generated 4 additional features from datetime: 'year', 'month', 'day', 'hour'. The listed features are integers, and we can use them in the model and the model will take into account the main patterns in the data.

The goal of the LightGBM algorithm is to obtain an estimate $\widehat{F(X)}$ of the function $F(X)$, mapping X (input features) to Y (target variable) while minimizing the loss function $L(Y, F(X))$.

In gradient boosting, each new algorithm b_i tree) is added to the already constructed composition:

$$a_i(x) = a_{i-1}(x) + b_i(x) \quad (1)$$

Where x represents the input features for predictions, such an algorithm adjusts the answers of the algorithm $a_i(x)$to correct the answers on the training set. If we consider several algorithms, the algorithm will be as follows:

$$a_n(x) = \sum_{i=1}^{N} b_i(x) \quad (2)$$

Where t iterations are the iterative process of adding models (trees) to the ensemble, for the classification problem, the loss function has several variants, one of which is:

$$L(Y, F(x)) = log(1 + exp(-YF(x))) \quad (3)$$

where $F(x) = a_n(x) + s_i, s = (s_1, \ldots\ldots, s_1)$ is the shift (correction) vector. The loss function:

$$\sum_{i=1}^{I} log(1 + exp(-y_i * (a_n(x_i) + s_i))) \rightarrow min \quad (4)$$

2.3. Regression model – LSTM and LSTM with attention

The second model we used to predict PM2.5 and PM10 is LSTM, a type of RNN. A recurrent neural network has a chain of recurrent neural network modules. RNNs have problems with short-term memory, gradient explosion, gradient vanishing, and long-term temporal dependencies. Hochreiter and Schmidhuber [16] developed LSTM to overcome the gradient vanishing problem.

LSTM has an internal structure known as multiplicative gates through which the flow of information is regulated. It has a more complex block called a memory cell, and with the help of this memory cell, the gates decide which data should be remembered or forgotten. This block can allow relevant information to be passed along a long sequence chain, allowing it to make accurate predictions. In a standard LSTM network, three main gates are available: a forget gate, an input gate, and an output gate. The forget gate controls whether information about the previous block's state should be retained or discarded in the block's current state. The input gate controls how much current information is added to the block's state. The output gate controls whether the current value in the cell contributes to the output. The LSTM neural network with all three gates is shown in Figure 1.

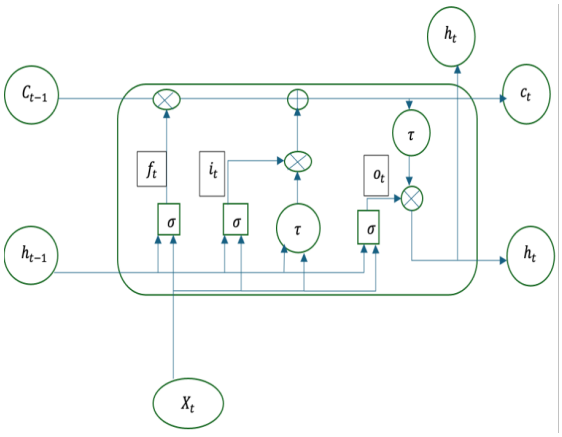

Figure 1 – The architecture of LSTM.

The input gate i_t decides whether new infor–mation will be added to the LSTM memory/cell state. It consists of two parts: first, the sigmoid layer decides how much current input information should be kept in the state or which values should be updated. Then, the tanh layer creates a vector of new values C_t that are added to the memory. The following equations describe both parts.

$$\begin{aligned} i_t &= \sigma(X_t U_t + h_{t-1} W_i + b_i \\ C_t &= tanh(X_t U_c + h_{t-1} W_c + b_c) \end{aligned} \quad (5)$$

In equations (5), W_i and U_i represent the weight matrix of the input gate i_t, tanh is the activation function, U_c and W_c the weight matrix of the new memory cell C_t, b_i, b_c represent the bias of the input layer and the bias vector of the new memory cell C_t, respectively.

From the combination of the above two layers, i.e., equations (2) the update for the cell state C_t is achieved by forgetting the current value of the cell state using the forget gate layer and then updating it by multiplying the old value (i.e., C_{t-1}) and then adding the new value of C_t. In equation (4) C_{t-1} represents the cell state at the last instant and the value of the cell state at the current instant.

$$C_t = f_t \otimes C_{t-1} \oplus i_t \otimes C_t \quad (6)$$

The output gate is needed to determine what part of the memory contributes to the output. It first analyzes the information to be processed as output using the sigmoid function calculated by equation (6). Taking the nonlinear value of the tanh function of the current state of one, it multiplies the output from the output gate and then calculates the value using equation (7).

$$\begin{aligned} o_t &= \sigma(X_t U_o + h_{t-1} W_o + b_o \\ h_t &= o_t \otimes tanh(C_t) \end{aligned} \quad (7)$$

In this study, we propose to use the attention approach in LSTM. LSTM network has related RNN, LSTM architecture can process up to 20 sequences but faces difficulties when processing more than 20 sequences. In addition, in some cases, ordinary LSTMs cannot make a more realistic and accurate prediction. Therefore, we propose to combine the attention mechanism with the LSTM network to solve the above problems and improve its prediction accuracy. The purpose of combining the attention network in the LSTM algorithm is to correct the shortcomings of RNN and optimize the LSTM neural network. The architecture of LSTM with attention and the internal structure of the attention block and LSTM layers are shown in Figure 2.

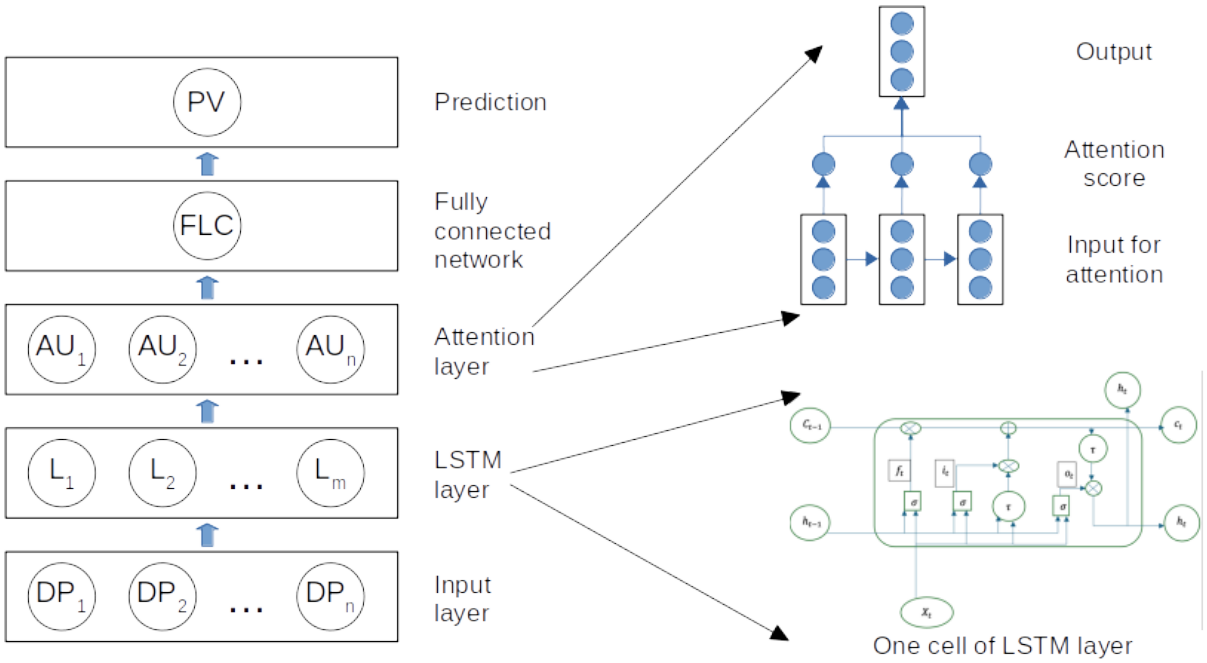

Figure 2 – The architecture of LSTM with attention.

As shown in Figure 2, the five layers of the proposed LSTM network model are the input layer, LSTM layer, attention layer, dense layer, and output layer. All these five layers should be in the correct order. The input layer (or data point layer) takes the data points into the model. The LSTM layer uses the hidden short-term memory layer to infer the input data into rich features. The attention layer calculates a weight vector in which the most important weights are extracted from the hidden state information and weights all the hidden states of the subsequent time steps. In the dense layer, each neuron of layer N is connected to each neuron of layer N+1. Therefore, the dense layer is often called a fully connected layer. The output layer uses the feature vectors to analyze and predict the time series data.

Integrating the attention mechanism into the LSTM network helps us to exploit the long sequence data retention problem as a flexible solution. It digs out a robust and explicit connection by intelligently assigning more weights to each critical part of the input data, which allows us to make accurate predictions using production data. LSTM works well with time series data, but when combined with an attention mechanism, it can predict production more accurately and accurately, even with noisy data. This study has made a significant contribution by adding the attention mechanism, its placement (e.g. where to add an attention layer, either before the dense layer or after the dropout layer), and the number of attention layers in the LSTM network. This fusion allows the AI-driven algorithm to work efficiently.

3. Results

Figure 3 shows the stages of predicting the concentration of air pollutants PM2.5 and PM10. The process begins with collecting data, which is then processed to clean and prepare it for modeling. The following features act as input parameters for the model: wind_dir; wind_speed; temp; humidity; bar; year; day; month, based on these features, we predict PM2.5, PM10. The main justification for using input features is the relatively easy organization of collecting this data using stationary or portable sensors. In addition, we used open sources (Open API) to collect weather data in cases where sensors could not collect data.

At the preprocessing stage, outliers are removed, and data is normalized to reduce the impact of abnormal values and bring the data to a single scale. Then, the data is divided into training and testing sets to evaluate the model's accuracy. Two techniques are used to build models: LightGBM and LSTM. The LightGBM model, suitable for tabular data, enables rapid predictions and is evaluated using the RMSE (root mean square error) metric [9]. The model is then applied to predict PM2.5 and PM10 concentrations. The LSTM recurrent neural network, which is particularly effective for time series analysis, is also used [12]. After training, the LSTM model is also evaluated using the RMSE metric and applied to predict PM2.5 and PM10. Thus, both models (LightGBM and LSTM) are applied in parallel, allowing for a comparative analysis of their performance and selecting the optimal approach for predicting air pollutants.

Since the collected data set is relatively small (902 records and 8 features) and quite variable, the data was divided into training and test data multiple times to evaluate the modeling results. The division into training and test samples was performed randomly. In a proportion of 80% to 20%, i.e., 721 records were included in the training sample, and 181 were included in the test sample.

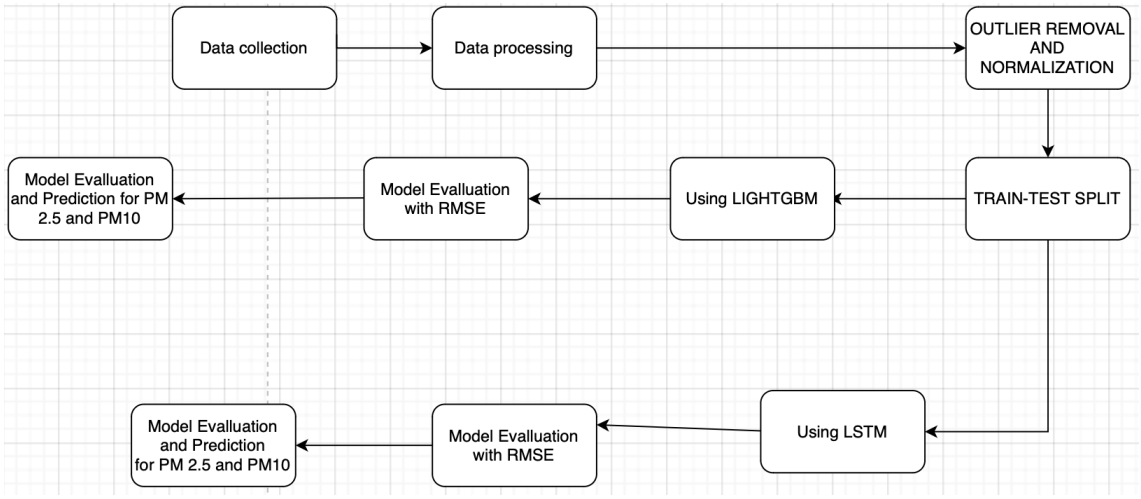

Figure 3 – Steps of data preparation and building LightGBM and LSTM models for predicting PM2.5 and PM10

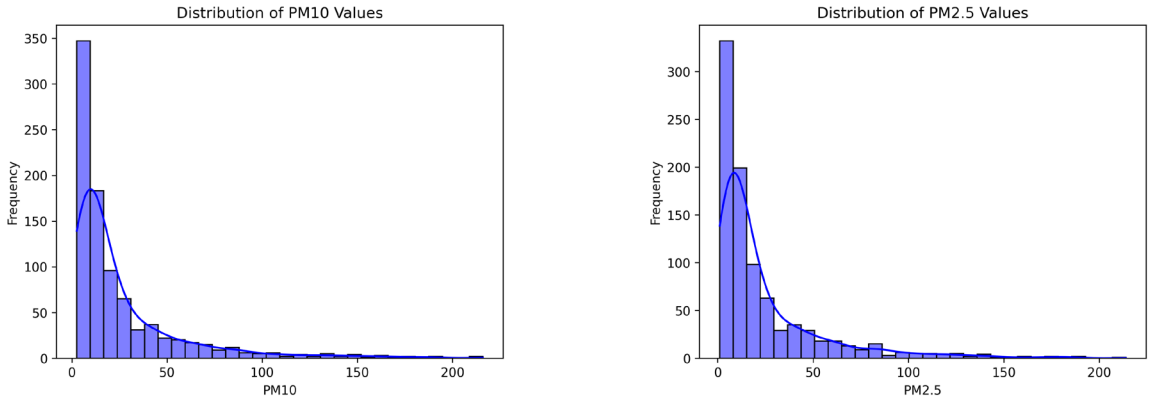

Figure 4 – The distribution of the PM2.5 and PM10.

In Figure 4, we plotted the distribution for PM10 and PM2.5, showing a left-skewed distribution with a predominance of low values and rare high concentrations. Both plots demonstrate that PM10 and PM2.5 have a similar distribution structure, with a high concentration of low and rare high values.

Figure 5 shows the results of predicting PM2.5 and PM10 concentrations using the LSTM model. The RMSE value obtained for PM2.5 is 5.94, and for PM10, it is 5.71, indicating acceptable model accuracy. The actual values are shown in red, the predicted values are shown in blue, and the 95% confidence interval is shown in light blue, reflecting the uncertainty of the predictions. The model accurately reproduces the trends in the data, although there are some noticeable deviations. The confidence interval covers a significant part of the real values, confirming the confidence in the predictions of the LSTM model.

We used the LSTM with Attention algorithm to improve my results on the simple LSTM. After preparing and training the model, we got the following result, which can be seen in Figure 6.

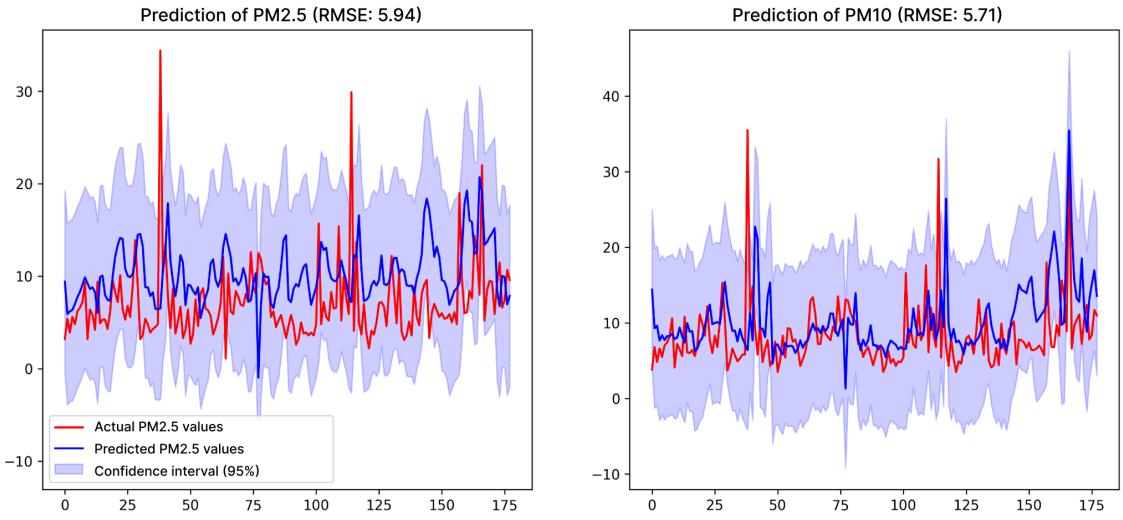

Figure 5 – Prediction of PM2.5 and PM10 concentrations using LSTM model and RMSE quality assessment.

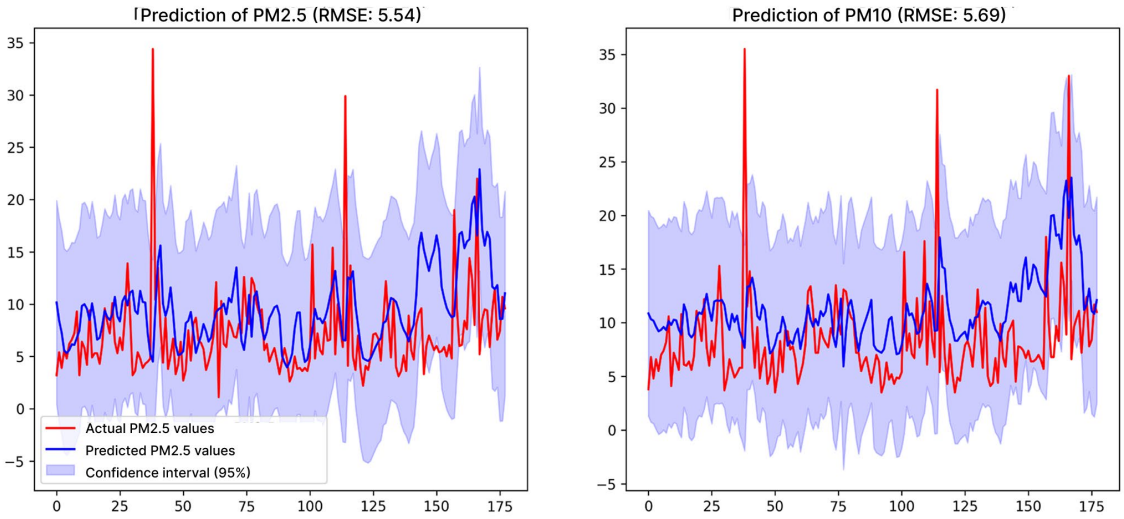

Figure 6 – Prediction of PM2.5 and PM10 concentrations using LSTM with attention and RMSE quality assessment.

Based on Figure 6, the model has slightly improved compared to the simple LSTM regarding the RMSE metric. The prediction of PM2.5 in terms of the RMSE metric improved by 0.40, and the prediction of PM10 in terms of the RMSE metric improved by 0.02, which is a positive result.

We applied the LightGBM machine learning model to predict the PM2.5 and PM10 air pollutant levels. The results are presented in the form of two graphs, where the comparison of the predicted and real values is shown for each particle. For PM2.5, the model achieved an RMSE value of 4.75, which indicates a fairly good prediction quality but with some deviations from the real data. Similarly, for PM10, the model showed an RMSE of 5.76, indicating a similar accuracy level with small errors. In Figure 7, the red dots represent the predicted and actual value pairs, and the dashed line represents the ideal accuracy (coincidence between the predicted and actual values). The distribution of dots around this line allows us to visually assess the model biases, which, although present, remain within a moderate error level. Overall, the LightGBM model accurately predicted air pollution for PM2.5 and PM10.

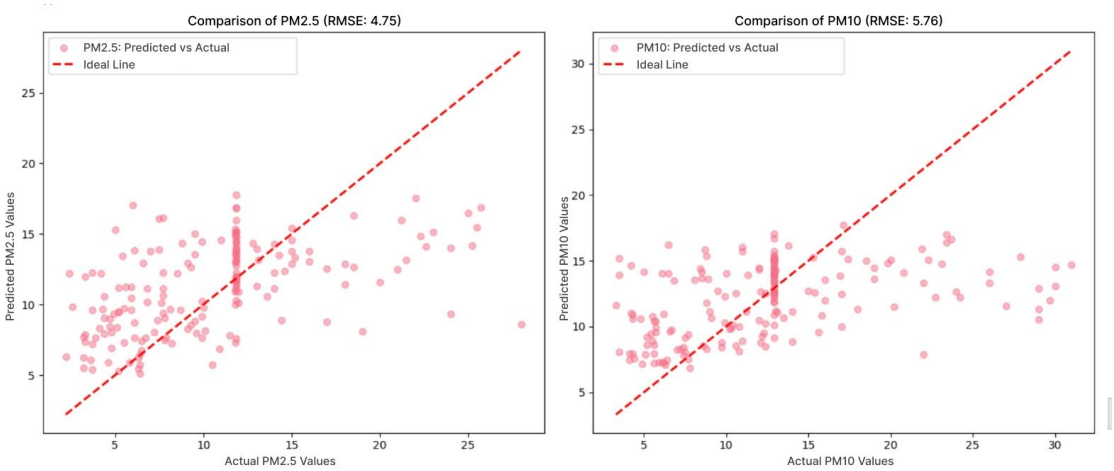

Figure 7 – Comparison of predicted and actual PM2.5 and PM10 values for LightGBM model with RMSE metrics.

We compared two algorithms that we used for air pollution predictions. Comparing LightGBM and LSTM with Attention models for predicting PM2.5 and PM10 concentrations shows that LSTM with Attention outperforms LightGBM regarding prediction accuracy and stability. The RMSE metric for LSTM with Attention is 5.54 for PM2.5 and 5.69 for PM10, lower than that of LightGBM (4.75 and 5.76, respectively), indicating a more minor standard deviation and, therefore, more accurate predictions. The LightGBM plots show significant deviations from the ideal line, especially for high PM10 values, where the scatter of the dots shows the errors in the predictions. At the same time, the LSTM with Attention plots looks smoother, and the confidence intervals (blue shaded background) show the range of possible errors, which allows us to assess the stability of the model better. LSTM with Attention copes better with fluctuations in data and accounts for temporal dependencies more accurately, making it more suitable for time series forecasting tasks like PM2.5 and PM10. Thus, LSTM with Attention is a preferable model for predicting PM2.5 and PM10 concentrations, providing higher accuracy and reliability of predictions.

5. Conclusions

The obtained results confirm some intuitive expectations about air circulation in the foothills. For example, when interpreting the results, we can state that the earlier the data is received (morning), the lower the PM2.5 pollution. The higher the air temperature, the lower the pollution. The lower the wind direction (to the east), the lower the pollution.

The quality of air pollution calculations can increase significantly, especially if the calculations use data from a nearby sensor. For example, if you predict the readings at point 0 using the data from point 1, the R2 value increases by 25%. At the same time, for point 6, using the data from point 1, the value increases more than 2 times – from 0.27 to 0.64. This is expected since point 6 is approximately twice as close to point 1. However, the calculation quality may decrease if the corresponding pair is selected incorrectly (for example, points 0 and 6).

Additionally, the study showed that LightGBM and LSTM with Attention models have different accuracy in predicting air pollutant concentrations. LSTM with Attention demonstrated higher accuracy with RMSE of 5.54 for PM2.5 and 5.69 for PM10, indicating its better ability to model time dependencies and capture seasonal variations. LightGBM, on the contrary, showed slightly higher errors, especially in predicting PM10, which may be due to the limitations of working with time series requiring sequential data analysis. Thus, using LSTM with attention to predict pollutant concentrations in urban agglomeration conditions is more justified, as it provides more accurate and stable results, especially in mountain-valley air circulation conditions.

In addition, we would like to note that LSTM with attention did not show significant

improvement compared to simple LSTM. We assume this is due to the small dataset; LSTM with attention improves its results as the data gets bigger. Monitoring the quality of weather data takes significant time, and the data will only get bigger. Due to this fact, LSTM with attention will be an effective approach for predicting PM2.5 and PM10.

Funding

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan grant number BR24992852.

Author Contributions

Conceptualization, T.M. and M.Y.; Methodology, T.M. and A.B.; Software, M.Y. and A.B.; Validation, M.Y. and M.G..; Formal Analysis, T.M., M.G.; Investigation, T.M. and M.Y.; Resources, T.M.; Data Curation, T.M.; Writing – Original Draft Preparation, T.M. and M.Y.; Writing – Review & Editing, A.B., M.G. and M.Y.; Visualization, T.M. and M.Y.; Supervision, T.M.; Project Administration, T.M.; Funding Acquisition, T.M.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. A. Russell et al. "A spatial survey of environmental indicators for Kazakhstan: an examination of current conditions and future needs." *International journal of environmental research* 12 (2018): 735-748.
2. Atmospheric air quality in urban areas. – URL: https://www.unece.org/fileadmin/DAM/env/europe/monitoring/Indicators/A-2-ru-final.pdf (date of access 20.07.2023).
3. Kazakhstan is in the top position in terms of pollution. – URL: https://dknews.kz/ru/eksklyuziv-dk/221987-kazahstan-v-top-poziciyah-po-urovnyu-zagryazneniya (date of access 20.07.2023).
4. Karatayev, M., Rivotti, P., Mourão, Z. S., Konadu, D. D., Shah, N., & Clarke, M. "The water-energy-food nexus in Kazakhstan: challenges and opportunities." *Energy Procedia* 125 (2017): 63-70.
5. Current Pollution Index by City. – URL: https://www.numbeo.com/pollution/rankings_current.jsp
6. Kerimray, A., Azbanbayev, E., Kenessov, B., Plotitsyn, P., Alimbayeva, D., & Karaca, F. "Spatiotemporal variations and contributing factors of air pollutants in Almaty, Kazakhstan." *Aerosol and Air Quality Research* 20.6 (2020): 1340-1352.
7. Nugmanova, D., Feshchenko, Y., Iashyna, L., Gyrina, O., Malynovska, K., Mammadbayov, E., ... & Vasylyev, A. "The prevalence, burden and risk factors associated with chronic obstructive pulmonary disease in Commonwealth of Independent States (Ukraine, Kazakhstan and Azerbaijan): results of the CORE study." BMC pulmonary medicine 18 (2018): 1-14.
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y.. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30 (2017).
9. How to use a 4s 40A BMS Module to build Battery Packs? – URL: https://circuitdigest.com/electronic-circuits/how-to-use-a-4s-40a-bms-module-to-build-battery-packs.
10. Gore, Ranjana Waman, and Deepa S. Deshpande. "Voting method for AQI prediction and monitoring air pollution using real-time data." 2020 international conference on smart innovations in design, environment, management, planning and computing (ICSIDEMPC). IEEE, 2020.
11. Rudakov, V., Timur, M., & Yedilkhan, A. "Comparison of Time Series Databases." 2023 17th International Conference on Electronics Computer and Computation (ICECCO). IEEE, 2023.
12. Rudakov, V., Timur, M., Yedilkhan, A., & Perizat, O. "Time Series Analysis of Biogas Monitoring with Deep Learning Approaches." 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI). IEEE, 2023.
13. Shabanpour, R., Golshani, N., Auld, J., & Mohammadian, A. "Dynamics of activity time-of-day choice." *Transportation Research Record* 2665.1 (2017): 51-59.
14. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Neftissov, A., Vatskel, V., Yedilkhan, D., & Herych, M. (2022). BUILDING A MODEL FOR CHOOSING A STRATEGY FOR REDUCING AIR POLLUTION BASED ON DATA PREDICTIVE ANALYSIS. *Eastern-European Journal of Enterprise Technologies*, 117(4).
15. Yandex.Weather API, https://yandex.ru/dev/weather/ last accessed 2024/09/20.
16. Hochreiter, S. "Long Short-term Memory." *Neural Computation MIT-Press* (1997).
17. Biloshchytskyi, A., Kuchanskyi, O., Andrashko, Y., Neftissov, A., Yedilkhan, D., & Vatskel, V. (2023, March). Models and methods for monitoring, air purification, and forecasting environmental pollution. In 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 107-112). IEEE.
18. Sarsenova, Z., Yedilkhan, D., Yermekov, A., Saleshova, S., & Amirgaliyev, B. (2024). ANALYSIS AND ASSESSMENT OF AIR QUALITY IN ASTANA: COMPARISON OF POLLUTANT LEVELS AND THEIR IMPACT ON HEALTH. *Scientific Journal of Astana IT University*, 98-117.

Information About Authors:

Meyir Yedilkhan is a Master's student in the Almaty Branch of National Research Nuclear University MEPhI (Moscow, Russia, meir.yedilkhan@gmail.com). His research interests include the development of computer vision and data mining. ORCID iD: 0009-0003-6513-4982

Azamat Berdyshev is a PhD student in the Information Systems department at the International University of Information Technology (Almaty, Kazakhstan, Aberdysh@gmail.com). His research interests include the development of LLM algorithms. ORCID iD: 0000-0003-0574-1580

Maksat Galiyev is a PhD student at Suleyman Demirel University (Kaskelen, Kazakhstan, galiev.maksat@gmail.com). His research interests include the development of software engineering and data mining. ORCID iD: 0009-0006-2045-7907

Timur Merembayev, PhD is a Research Assistant at the Institute of Information and Computational Technologies (Almaty, Kazakhstan, timur.merembayev@gmail.com). His current research covers various topics in a mathematical simulation of physical processes, machine learning, and geoscience problems. ORCID iD: 0000-0001-8185-235X.