Air Quality Index Forecasting in the HRcity Smart City System Based on an LSTM Prediction Model

Bruno Trstenjak
Medimurje University of Applied Sciences in Cakovec, Croatia
btrstenjak@mev.hr (corresponding author)

Sanja Brekalo
Medimurje University of Applied Sciences in Cakovec, Croatia
sbrekalo@mev.hr

Jurica Trstenjak
Medimurje University of Applied Sciences in Cakovec, Croatia
jtrstenjak@mev.hr

Received: 22 April 2025 | Revised: 19 May 2025 | Accepted: 1 June 2025

 DOI: https://doi.org/10.48084/etasr.11665

ABSTRACT

The HRcity project aims to implement various digital technologies for the needs of cities in the Republic of Croatia. The system itself consists of several different components specialized for the needs of citizens and various institutions operating in the city areas. One of these components monitors air quality and measures the Air Quality Index (AQI). This paper presents the structure of the AQI component and the principles for measuring and monitoring pollution values and determining the AQI. Air quality prediction models, based on LSTM, were implemented based on pollutant measurements collected during three years of system use. RMSE, MAE, and MAPE metrics were used to evaluate the performance of the LSTM models. The experimental evaluation shows that the LSTM models implemented with other elements of the AQI achieved very good accuracy.

Keywords-smart city; LSTM prediction; air quality index; HRcity system

I. INTRODUCTION

As society develops around the world, people's awareness and aspiration for a healthy life are growing. As air is a basic element for every human being, the quality of the air a person breathes directly affects his or her health. The European Union recognizes the importance of preserving air quality to improve human health and environmental protection. Some of the key air quality initiatives and strategies in the EU include the Clean Air for All Europeans strategy, which focuses on reducing greenhouse gas emissions and air pollution by 2030 [1]. The term air quality is closely related to the Air Quality Index (AQI) [2]. EU standards determine the parameters that are monitored in the AQI calculation process. The calculation of the AQI is based on the values of the following six air pollutants: measuring the concentration of carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), and particulate matter (PM2.5 and PM10) [3].

The HRcity project aims to bring Croatian municipalities and cities into the digital era and enable them to become "smart cities" [4]. The system consists of several modules (web and mobile applications) that allow citizens to access information and directly communicate with various institutions that work to meet the needs of the city in various domains. One such module is the E-ecology module, which monitors air quality in individual cities, informs citizens about the AQI, and alerts them if the detected index is dangerous for their health.

This paper presents an AQI prediction model, based on LSTM, which is embedded in the HRcity system. With increasing awareness of the importance of air quality in which people live, especially in urban centers, many scientific studies have been published on the subject. In [5], an analysis and comparison of machine learning methods for air quality forecasting was presented. A qualitative analysis showed that the methods used can be classified into three basic groups: regression algorithms, deep learning algorithms, and hybrid algorithms. In [6], an SVR model was used to predict PM2.5 and PM10 particle concentrations in London. This SVR-based model used Gaussian kernel functions to determine the Gaussian distribution. In [7], an ARIMA model was developed and several simulations, using data from the city of Dakar,

showed that it was better suited to predict PM10 pollution. In contrast to the regression approach, the model in [8] used multitask methods and RNN to predict AQI in China. This study proposed a novel Spatial-Temporal Deep Multitask Learning (ST-DMTL) framework for air quality forecasting based on dynamic spatial panels of multiple data sources. In [9], a Deep Neural Network (DNN)-based approach was proposed, consisting of a spatial transformation component and a deep distributed fusion network. In [10], a nested LSTM network was developed to predict AQI in Beijing, combining multiple nested LSTM networks (MTMC-NLSTM). In [11], a multivariate regression model was presented to predict AQI, based on the use of independent variables and their values measured a few days earlier than the prediction day, calculating the correlation coefficients between the predicted and measured values. The third group includes hybrid models. In [12], a hybrid model merged three methods, LSTM, GRU, and CNN, to predict daily PM2.5 in Taiwan. In [13], a slightly different approach was presented, using GA and an encoder-decoder model with LSTM to predict PM2.5 concentrations. In [14], a modified approach was presented to collect data on pollutant concentrations. This study focused on modeling the air quality pattern in a given region by adopting both fixed and moving IoT sensors, placed on vehicles patrolling the region. This study demonstrated the feasibility of this approach to effectively measure and predict air quality using different machine learning algorithms with real-world data.

II. SYSTEM ARCHITECTURE AND RESEARCH METHOD

A. AQI Module Architecture

The HRcity AQI monitoring system component is deployed in a cloud environment, which allows it to be used anywhere. The system is flexible and allows users to choose which pollutants they want to monitor. Figure 1 shows the architecture of the AQI module. Air pollutant concentration sensors and IoT devices installed in urban areas typically measure the following parameters: NO_2, SO_2, O_3, CO, PM2.5, and PM10. Measured values of pollutant concentrations are sent to the system every hour using an API. The data received are recorded in a cloud database. The interface divides the received data into three groups: Air pollutant data, comprising the concentrations of each measured pollutant, meteo data, which are meteorological variables, and station data consisting of geographical data and some details that make it recognizable throughout the system. Based on the air pollutant data received, the system immediately calculates the AQI. The calculated AQI is recorded in the central database and displayed to users on mobile devices in real time. Data from the central database are used in two ways. At point A, the data are sent to the prediction model, where forecasting is performed for 6 pollutants. The prediction model consists of 6 autonomous prediction submodels. In this way, all cities can use the system, regardless of whether they monitor the concentration of all pollutants, some of them, or only a specific one (regardless of the type of measuring devices). Before the prediction process, the data go through the preprocessing phase. The prediction results of each submodel together form the final prediction of the AQI. The prediction result obtained can be displayed on a mobile device via the visualization API (point B).

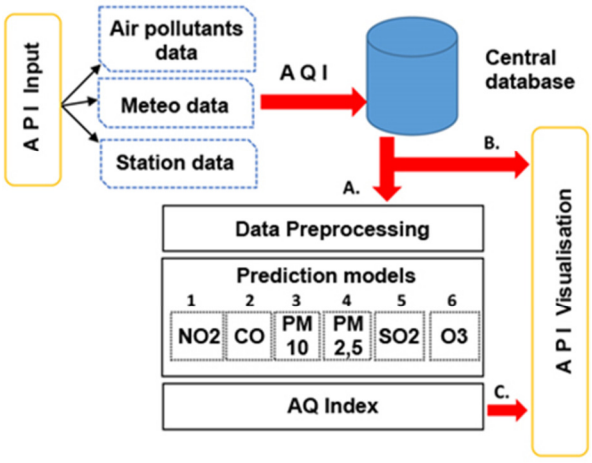

Fig. 1. AQI module architecture.

The prediction is calculated on a daily basis, with an offset of 5 days ahead. The application records the prediction results in a database to analyze the prediction accuracy of the existing model and obtain feedback to fine-tune the future prediction model.

B. Dataset

A dataset from data collected during three years of operation of the HRcity system was used. Data were collected for each city or municipality separately through measuring devices installed in their locations. Each measuring station, depending on the sensors it uses, sends air pollutant data on an hourly basis. Approximately 26,000 measurement records were collected for each device in the central database. Since the HRcity system includes cities and municipalities from different locations in Croatia, each of them using devices from different manufacturers to collect measurements, a private dataset was created for each of them. In some municipalities and cities, only some pollutants were measured, such as PM2.5 and PM10 particles, while in others the entire spectrum of air pollutants was measured. The maximum number of pollutants supported by the system is listed in Table 1. Due to the possibility of further analyses, meteo and station data were added to the existing air pollutant data. Meteorological data were recorded for all cities and municipalities, regardless of the type and characteristics of the measuring station. Table I shows the meteorological and pollutant variables recorded from the monitoring stations or the IoT devices.

Due to the variety of measuring devices, a strategy was determined to form a prediction model for each pollutant, i.e. six submodels, depending on the pollutant being monitored. Data collected on an hourly basis from measurement stations are recorded in the system's central database and used to update individual prediction models. The correction, update and training of the models are carried out every six (6) months for each city or municipality individually. This is an attempt to refresh the dataset and retrain the existing prediction models.

TABLE I. POLLUTANTS AND METEOROLOGICAL VARIABLES COLLECTED IN MONITORING STATIONS

Air pollutant data	Meteo data	Station data
NO_2 (ppm)	Temperature	Station ID
CO (ppm)	Humidity	GEO location
PM10 ($\mu g/m^3$)	Pressure	Date & Time
PM2.5 ($\mu g/m^3$)		
SO_2(ppm)		
O_3 (ppm)		

C. AQI Determination Method

In the process of determining the AQI, the HRcity system uses EU standards and values. Currently, there are several different standards for determining AQI used in different countries. The standards differ in the set AQI range as well as in the specification of major pollutants whose values are taken into account during the determination of AQI. The European Environment Agency devised the European Air Quality Index (EAQI). This standard measures five basic pollutants: PM10, PM2.5, ground-level ozone (O_3), nitrogen dioxide (NO_2), and sulphur dioxide (SO_2). The AQI is divided into six basic categories. Each category is determined with a range of numerical values. Usually, each category is associated with a specific color to provide visualization for users who monitor the connection between air quality and human health [15]. The HRcity system has integrated the EAQI standard. Table II shows the ranges for individual pollutants. All values are expressed in $\mu g/m^3$.

TABLE II. POLLUTANTS AND METEO VARIABLES COLLECTED IN MONITORING STATIONS

AQI	PM2.5	PM10	NO_2	O_3	SO_2
Good	0-10	0-20	0-40	0-50	0-100
Fair	10-20	20-40	40-90	50-100	100-200
Moderate	20-25	40-50	90-120	100-130	200-350
Poor	25-50	50-100	120-230	130-240	350-500
Very poor	50-75	100-150	230-340	240-380	500-750
Extremely poor	75-800	150-200	340-1000	380-800	750-1250

D. LSTM Prediction Model

LSTM was applied as the basic prediction model for AQI forecasting. The decision to choose this model is based on its reliability, the good results achieved in the area of AQI prediction, and the good software support for its implementation [16-17]. The LSTM model, a type of Recurrent Neural Network (RNN) consists of three (3) main layers: the input layer, the hidden (forget) layer, and the output layer. A Python environment based on Tensorflow and Keras was used to develop and implement the model [18].

At the beginning of the model, a sequence layer is used. The LSTM is used for tasks involving sequential data. The hidden layers consisted of two LSTM levels with 50 neurons and the output layer involved two dense layers. In the AQI module, six such LSTM models were implemented, one for each pollutant.

Figure 2 shows the general structure of the prediction model. The structure begins with preprocessing and adapting the input data to improve the accuracy of the model [19]. After preprocessing, the weights for each attribute are determined using the Information Gain method [20]. Feature selection was performed based on the obtained results on weight values and the correlation analysis between features. The data prepared in this way was sent to the LSTM model. The last stage involves the evaluation of the prediction results using four metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) [21].

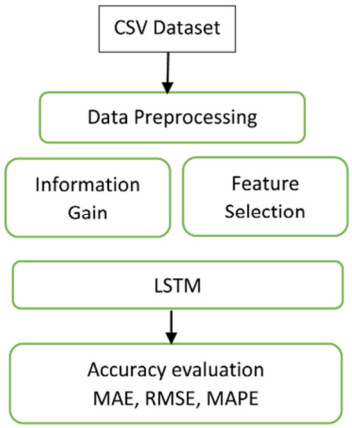

Fig. 2. The general structure of the prediction model.

III. RESULTS AND DISCUSSION

Six autonomous LSTM models were developed based on the measurement data collected in each selected city area. For each Croatian city in the system, initial measurements were carried out for a minimum of 6 months to create a base for shaping a prediction model tailored to its characteristics and location. Three years of monitoring data were used to develop the prediction model. Data from the first two years were used to train the model, and data from the third year were used to test it.

In addition, meteorological data from observation stations were also integrated to improve prediction accuracy. A correlation analysis was performed for all pollutants with the aim of detecting positive correlations between them. The heatmap [22] in Figure 3 shows a correlation between PM10 and other characteristics of air quality. Correlation analysis allows for the selection of features to achieve the highest possible prediction accuracy. The correlation analysis revealed that the concentration of PM10 has a positive correlation with several features. According to this analysis, air pressure has the greatest influence on particle concentration. This is logical because high pressure affects the pressure of the particles and increases their concentration in the lower parts of the air where the measuring devices are placed.

Each city, independently of others, has registered a certain number of pollutants that it wants to measure in its urban area. The system performs AQI forecasting once a day for each registered pollutant separately. The prediction results obtained are merged into the final calculation and AQI predictions.

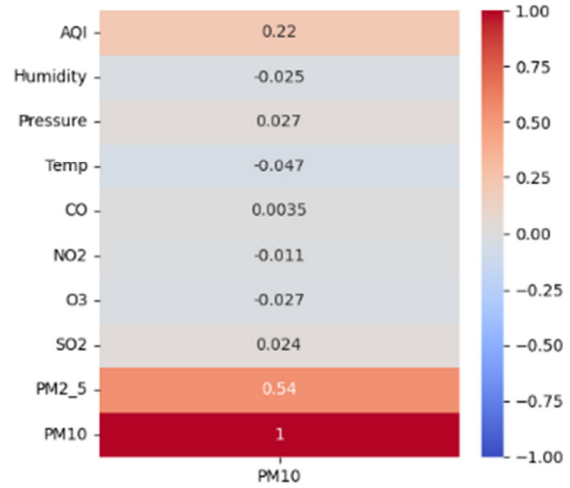

Fig. 3. Correlation heat map between PM10 and air quality features.

Testing the prediction accuracy of the LSTM model for a single pollutant showed high accuracy. The evaluation results show that the proposed model performed well, with satisfactory prediction accuracy. Prediction models were tested for different time series, both short- and long-time. Based on research of citizen interests, short-time forecasting was implemented in the system. During the testing and development of the LSTM prediction model, a large number of different combinations of feature (pollutant) selections were used.

The selection of features to form a prediction model is mainly based on the correlation results obtained between the features and the calculated weight values. The results varied depending on the measurement devices used, the sensors used, and the city locations themselves. For this reason, several LSTM prediction models with minor differences between them were formed. In general, the evaluation metrics showed that the LSTM models achieved good prediction accuracy. Since different LSTM models were used in development and testing, depending on the specifics of each city, the prediction accuracy was measured at the level of the entire system. The mean values for RMSE, MAE, and MAPE were 0.1077, 0.0891, and 0.1325, respectively.

The HRcity system has implemented an automatic AQI forecasting mechanism, deployed in a cloud environment. For each city included in the system, pollutant concentration measurements are collected every hour. The system runs the AQI prediction process once a day for a period of 7 days.

IV. CONCLUSION

This study introduced a system and cloud framework for forecasting AQI in cities. Forecasting is based on LSTM prediction models, which are an integral part of the AQI module implemented in the HRcity system. The LSTM model was chosen based on the good results achieved and presented in previous studies [5] because a proven and adaptable model is needed for the specific diversity of stakeholders. This study presents the concept of measuring the concentration of pollutants with the possibility of dynamic adaptation, depending on the specific devices used in cities. Taking into account the operational strategy of the system, a separate LSTM prediction model was developed for each pollutant in each city. This paper presents the general structure of the prediction model and explains the individual phases of the model's operation.

Before developing the prediction model, a feature correlation analysis was performed, determining weight values for different combinations of pollutants depending on the character of the urban measurement area. Evaluation metrics showed that the models achieved good results. Due to the specifics of the system and the large number of LSTM models that used similar but not the same datasets, an estimate of the average value of the model's performance quality was given. The results showed that the LSTM models achieved results that are within the scope of similar scientific research in the AQI domain [5].

Based on the results of the LSTM prediction models, a cloud AQI module was implemented. The AQI module is responsible for carrying out the process of forecasting the AQI. The system runs the AQI-level prediction process once a day for a period of 7 days. The future period during which the LSTM prediction model will be used and additional analysis of the results obtained will provide guidelines for its further development and possible improvements to the prediction model.

REFERENCES

[1] Decision (EU) 2022/591 of the European Parliament and of the Council of 6 April 2022 on a General Union Environment Action Programme to 2030. European Union, 2022.

[2] A. Alwabli, "Federated Learning for Privacy-Preserving Air Quality Forecasting using IoT Sensors," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 16069–16076, Aug. 2024, https://doi.org/10.48084/etasr.7820.

[3] Z. Karavas, V. Karayannis, and K. Moustakas, "Comparative study of air quality indices in the European Union towards adopting a common air quality index," *Energy & Environment*, vol. 32, no. 6, pp. 959–980, Sep. 2021, https://doi.org/10.1177/0958305X20921846.

[4] B. Trstenjak, L. Butkovic, and S. Brekalo, "HRcity smart city as support in Business decision making for waste disposal," presented at the 98th International Scientific Conference on Economic and Social Development, 2023.

[5] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 10031–10066, Sep. 2023, https://doi.org/10.1007/s10462-023-10424-4.

[6] A. Sotomayor-Olmedo *et al.*, "Evaluating trends of airborne contaminants by using support vector regression techniques," in *CONIELECOMP 2011, 21st International Conference on Electrical Communications and Computers*, San Andres Cholula, Puebla, Mexico, Feb. 2011, pp. 137–141, https://doi.org/10.1109/CONIELECOMP.2011.5749350.

[7] B. Ngom, M. Diallo, M. R. Seyc, M. S. Drame, C. Cambier, and N. Marilleau, "PM10 Data Assimilation on Real-time Agent-based Simulation using Machine Learning Models: case of Dakar Urban Air Pollution Study," in *2021 IEEE/ACM 25th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, Valencia, Spain, Sep. 2021, pp. 1–4, https://doi.org/10.1109/DS-RT52167.2021.9576143.

[8] X. Sun, W. Xu, H. Jiang, and Q. Wang, "A deep multitask learning approach for air quality prediction," *Annals of Operations Research*, vol. 303, no. 1, pp. 51–79, Aug. 2021, https://doi.org/10.1007/s10479-020-03734-1.

[9] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep Distributed Fusion Network for Air Quality Prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Apr. 2018, pp. 965–973, https://doi.org/10.1145/3219819.3219822.

[10] N. Jin, Y. Zeng, K. Yan, and Z. Ji, "Multivariate Air Quality Forecasting With Nested Long Short Term Memory Neural Network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8514–8522, Sep. 2021, https://doi.org/10.1109/TII.2021.3065425.

[11] S. M. Choi, H. Choi, and W. Paik, "Multivariate Regression Modeling for Coastal Urban Air Quality Estimates," *Applied Sciences*, vol. 13, no. 19, Sep. 2023, Art. no. 10556, https://doi.org/10.3390/app131910556.

[12] P. W. Chiang and S. J. Horng, "Hybrid Time-Series Framework for Daily-Based PM2.5 Forecasting," *IEEE Access*, vol. 9, pp. 104162–104176, 2021, https://doi.org/10.1109/ACCESS.2021.3099111.

[13] M. H. Nguyen, P. L. Nguyen, K. Nguyen, V. A. Le, T. H. Nguyen, and Y. Ji, "PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model," *IEEE Access*, vol. 9, pp. 57338–57350, 2021, https://doi.org/10.1109/ACCESS.2021.3072280.

[14] D. Zhang and S. S. Woo, "Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network," *IEEE Access*, vol. 8, pp. 89584–89594, 2020, https://doi.org/10.1109/ACCESS.2020.2993547.

[15] "Standards for Air Quality Indices in Different Countries (AQI)," *Atmotech*. https://atmotube.com/blog/standards-for-air-quality-indices-in-different-countries-aqi.

[16] Isam Drewil and R. J. Al-Bahadili, "Forecast Air Pollution in Smart City Using Deep Learning Techniques: A Review," *Multicultural Education*, vol. 7, no. 5, pp. 38–47, May 2021, https://doi.org/10.5281/ZENODO.4737746.

[17] H. He and F. Luo, "Study of LSTM Air Quality Index Prediction Based on Forecasting Timeliness," *IOP Conference Series: Earth and Environmental Science*, vol. 446, no. 3, Oct. 2020, Art. no. 032113, https://doi.org/10.1088/1755-1315/446/3/032113.

[18] F. Aljuaydi, M. Zidan, and A. M. Elshewey, "A Deep Learning CNN-GRU-RNN Model for Sustainable Development Prediction in Al-Kharj City," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20321–20327, Feb. 2025, https://doi.org/10.48084/etasr.9247.

[19] G. Y. V. Tang, K. G. Pillay, and A. Mustapha, "The Impact of Data Preprocessing Order on LASSO and Elastic Net Capabilities," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20264–20270, Feb. 2025, https://doi.org/10.48084/etasr.9611.

[20] K. Qu, J. Xu, Q. Hou, K. Qu, and Y. Sun, "Feature selection using Information Gain and decision information in neighborhood decision system," *Applied Soft Computing*, vol. 136, Mar. 2023, Art. no. 110100, https://doi.org/10.1016/j.asoc.2023.110100.

[21] Y. Liu, P. Wang, Y. Li, L. Wen, and X. Deng, "Air quality prediction models based on meteorological factors and real-time data of industrial waste gas," *Scientific Reports*, vol. 12, no. 1, Jun. 2022, Art. no. 9253, https://doi.org/10.1038/s41598-022-13579-2.

[22] M. Niu, Y. Zhang, and Z. Ren, "Deep Learning-Based PM2.5 Long Time-Series Prediction by Fusing Multisource Data—A Case Study of Beijing," *Atmosphere*, vol. 14, no. 2, Feb. 2023, Art. no. 340, https://doi.org/10.3390/atmos14020340.