Vol. 7, No. 1 | January – June 2023

Air quality forecasting based on machine and deeplearning models: an IoT application

[1]Khalid Khan, [1]Affan Alim, [2]Humayun Qureshi, [3]Imran Sabir, and [1]Ibrahim Hassan

Abstract:

Harmful gasoline and particulate objects that exist in the air and above the cut-off values are dangerous for human, animal, and plant health. Essentially, it leads to lung cancer, throat infection, heart attack, and other diseases. The early forecasting of these objects may help for precautions of safety. In this paper, it is proposed to use the regression-based model auto regression integrated moving average (AIRMA) and deep learning-based model long short-term memory (LSTM) for air quality prediction. The air quality forecasting performance also depends on the quality of the available dataset. In this study, real-time data is collected from 10 different locations based on an IoT system, which is developed locally for a funded project of the Higher Education Commission (HEC). The main idea of this study is to validate the real-time collected dataset. Two objects, particulate PM2.5, and gasoline Ammonia are considered for four different locations for forecasting. Due to several issues such that electricity, Wi-Fi, sensor calibration, and collected data are not in their finest position. A number of prepossessing steps are applied to raw data to bring it into a usable form. Regardless of these issues, proposed models based on data collected by IoT system, outperform two air objects PM2.5 and Ammonia. For the case of Ammonia, an RMSE value of 0.562 is obtained which is very low to the mean value of 5.15 which indicates high performance. Similarly, very close values of 0.186 and 0.133 of RMSE and MAE were achieved respectively, and reflect the low variance in error. The LSTM-based experiment for Ammonia prediction, comparable to a very low RMSE value of 1.948 is achieved from the corresponding mean. A very small difference value of 0.287 between RMSE and MAE is obtained indicating a low variance in predicting error and high precision.

Keywords: *Air pollution, Air quality, Machine Learning, human health, Forecasting pm2.5, Ammonia.*

1 Introduction:

Due to adverse effect on biological objects, air pollution and its quality has become the focus of people's daily life. The pollutant air contains carbon monoxide (*CO*), Nitrogen dioxide (NO_2), sulfate dioxide (SO2), Ammonia (NH_3), particulate matter *PM* 2.5, *PM* 10, etc. [1]. Above the cut-off point, any of the pollutant ingredients would cause severe health issues, breathing difficulty, lung cancer, cardiovascular disease, respiratory and metabolic disorders, etc. [2]. The authorized organizations monitor the concentration of pollutant ingredients and share the current situation of air quality. The forecasting of the air quality is a challenging job, it would facilitate people, regarding the safety precaution, don't go out unnecessarily, wearing a specific mask in current situations, etc. The involvement of machine learning makes it possible to analyze past data and forecast the pollution in the air [3], [4], [5] Essentially, the results are dependent on performance. The data via IoT, meteorological data, historical data, etc. However, good machine learning algorithms play a vital role to make the prediction and forecasting reliable and applicable. The most recent work [4], the multiple linear regressive (MLR) is used on time series dependent variables. With the increase in the

[1]College of Computing and Information Sciences, Karachi Institute of Economics and Technology email: khalid.khan@kiet.edu.pk., affanalim@kiet.edu.pk and Ibrahimhassan@live.co.uk.
[2]Datalog, Pakistan. email: humayun.qureshi@datalog.pk.
[3]SEPA, Sindh Pakistan. email: Imransabir@sipa.gov.pk.

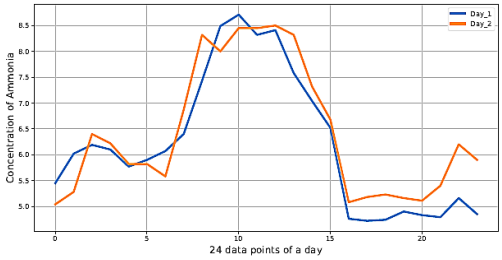

Fig. 1: Two days of Ammonia concentration at the same location

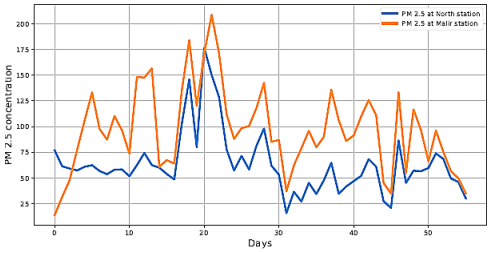

Fig. 2: PM 2.5 pollutant concentration at two stations in the same time period

availability of pollution (particulate matter and gasoline) in the air, an automated and sophisticated forecasting and prediction system for the air quality index is in dire need. IoT and Machine learning has been overwhelming and many researchers have been using multiple machine learning algorithms supporting vector machine (SVM), LSTM, autoregression integrated moving average (ARIMA), multiple linear regressive, etc. for forecasting and prediction. The present study proposes forecasting using neural networks long short-term memory LSTM, and autoregression integrated moving average ARIMA on time-series dependent features. The main focus is to validate the real-time collected data via IoT from 10 different locations. With this method, historical air quality data from 10 different locations in Karachi are used to construct the relevant feature. IoT-based systems are installed at these 10 locations for measuring the concentration of particulate matter, gasoline, and meteorological objects like temperature, humidity, PM2.5, ammonia, methane, Nitrogen dioxide, and carbon mono oxide. Further, a dataset is constructed using values of selected features over time. However, there are several limitations and challenges for forecasting and prediction, as we have a small size of the dataset, and IoT-based air quality collected data are dependent on many factors like fire, traffic, etc [8]. As shown in figure-1, a 5 days activity of Ammonia at the same location and, in Figure - 2, Ammonia activity at three different locations on the same day.

The following are the main contribution of the proposed method:

i. The data collection from IoT-based systems, these IoT systems are installed at ten different locations in Karachi city. Due to some external hurdles, sometime IoT-based systems are not able to send data to the cloud. Time and date synchronization issues exist in available data. These are handled and addressed in this research.

ii. Preparation and cleaning of the dataset. The collected data has missing values, un-synchronized with respect to date. According to the contribution, combining the different location datasets and cleaning them such that they can be used for prediction and forecasting.

iii. We proposed to use the two state-of-the-art machine learning algorithms LSTM with specific parameters and ARIMA with the best ordering.

The proposed approach yielded excellent performance in all four experiments in terms of MAE, RMSE, and MSE. For the case of Ammonia, a value of 0.562 RMSE is obtained which is very close to the mean value. Similarly, very close values of 0.186 and 0.133 of RMSE and MAE were achieved respectively. The LSTM-based experiment for Ammonia prediction, comparable to a very low value of 1.948 RMSE was achieved from the corresponding mean. A very small difference value of 0.287 between RMSE and MAE indicates a low variance in predicting error and high precision. The rest of the paper is organized as follows: In Section- II, the literature review is dissuaded, whereas Mathematical tools and proposed methodology are discussed in Section III, and experiment details and results are discussed in section-IV. The conclusions of the paper are provided in section V.

2 Related Work:

The industries, forest fire, traffic, and Ozone are the primary sources of air pollution. It is a dire need to develop a system which intimate about the pollution index of PM and

gasoline in air for corrective measures. A wide range of studies have already been done by researchers for forecasting and prediction. In this section, a number of research works are discussed in terms of its effectiveness and influencing factors of air pollution. Sivakumar el. al. [4] propose air quality index prediction using a machine learning algorithm. The author has proposed a multivariate regressive function using multiple linear regression MLR for AIQ prediction with consideration of two-time series features from the dataset. The MLR outperform compared with other developed model of AQI prediction. A boosting-based algorithm is proposed by Ying Zhang. al. [6] for air quality (AQ) prediction. The author has used the meteorological and historical data prepared after merging. In [6], the author proposed to use the LightGBM model to predict the concentration of PM2.5 in different locations of Beijing for a day. The results of the LightGBM sliding window are superior to other schemes. Zhendong at. al. [9] has studied PM2.5-based air quality forecasting. The author has proposed to use a hybrid deep learning model combine with variational decomposition (VMD) and bidirectional long-short memory BiLSTM for the prediction of PM2.5. The experimental results of VMD-BiLSTM outperform compared with individual EMDbased and VMD-based models. The dataset was collected from different cities of China. Another deep learning-based air quality forecasting has been proposed in [10]. The author targeted PM2.5 forecasting which correlated features. A hybrid deep learning architecture has been proposed by constitutional neural networks (1D-CNN) and Bi-directional long short-term memory. The experiment was conducted on a real-world dataset. Substantial work has been done for air pollution prediction using deep learning and ARIMA based model with promising results. In [11], an ARIMA model has be used for air pollution prediction. The author also used the artificial neural network (ANN) model and compared the results of two models. ANN results are more promising compared with ARIMA. In [12], Author has used machine learning algorithms Decision tree regression, Multilayer perceptron, gradient boosting regressor and random forest regression. The author use the 5 years

TABLE I: Location-Wise Pollution Sensors

Sno.	Location	# of dif. Particulates	# of diff. Gasoline	# of diff. Meteorological
1	NN	5	1	2
2	CC	0	2	2
3	MC	5	1	2
4	MY	0	4	2
5	FBA	0	4	2
6	LA	0	4	2
7	ML	5	1	2
8	SEPA	5	1	2
9	SD	5	1	2
10	UNI	0	4	2

Information

of data collected from different cities in China. An interesting discussion on the result section that different models performed a good prediction on different cities dataset. A 4 day ahead forecast of EU regional is provided by Copernicus atmosphere monitoring Service (CAMS). Papa at. al. [13], evaluate the CAMS AM forecast at urban coastal city in Greece. The author compared the performance results of the analog ensemble (AnEn) technique and the deep learning long short-term memory for the particulate matter of PM2.5 and PM10 during the winter season. The AnEn outperforms compared with LSTM.

3 Designing Tool and Methodology

A. Dataset, IoT system, and Study Area:
Data is a primary requirement of machine learning prediction can be collected from primary or secondary sources for implementation of machine learning algorithms. According to literature review, most of the research works addressed the real world data of cities or countries [14] [15]. In our case, data is collected from IoT system, IoT based systems with variety of sensors have been installed at 10 different locations of city Karachi, detail of sensors are described in Table I. The IoT system are connected with cloud via WiFi, and minute-minute meteorological, particulate, and gasoline data are forwarded to cloud. For this work, currently 3-month data is available (but it vary from station to station). A wrangling steps were applied on raw collected data, and

remove the structuring, and outlier issues to prepare the data in such a way to get the maximum efficacy. The following pollution sensors are used as described in Table I.

1. Particular Matter: PM 0.3, PM 0.5, PM 01, PM 2.5 and PM 10
2. Gasoline: Nitrogen Dioxide, Ammonia, Carbon monoxide, and Methane
3. Meteorological: Temperature, and Humidity extensive experiments have been performed at individual locations, individual attributes, and a combination of datasets.

B. Modeling with LSTM

Recurrent Neural Networks (RNN) extract the contextual information, between input and output sequence. Usually, RNN has a vanishing gradient problem, addressed in [16], [17]. In mid of the 1990s, researchers proposed several solutions for vanishing gradient for RNN [18]. The most popular sophistic approach proposed by [19] is LSTM. LSTM has the capability to handle the gradient vanishing problem in the back-propagation step such that it learns the input sequence for a longer time. Due to its efficacy, researchers have commonly used LSTM for time series issues [20]. A generalized closer look at the LSTM memory blocks diagram is shown in Figure 3, and a detailed diagram of the LSTM memory block with the single cell is shown in Figure 8. The output of each gate is a linear combination of weight and bias values shown in the following equations:

The output of the forget gate can be calculated by the following equations:

$$f_i = \sigma[(W_{fh}.h_{t-1}) + (W_{fx}.x_t) + b_f] \quad \ldots (1)$$

$$C_t^f = C_{t-1}.f \text{ same dimensionality} \quad \ldots (2)$$

f_i is an output of forget gate and
C_i^f is the system's current state.

$[W_{fh}, W_{fx}]$ are weight and b_f is bias value

The output of the input values can be achieved by following equations:

$$i_t = \sigma[(W_{ih}.h_{t-1}) + (W_{ix}.x_t) + b_i] \quad \ldots (3)$$

$$g_t = \tanh[(W_{gh}.h_{t-1}) + (W_{gx}.g_t) + b_g] \quad \ldots (4)$$

The output of the input node and gate is

$$C_t^i = i_t.g_t \quad \ldots (5)$$

The final cell sttatus can be calculated by:

$$C_t = C_t^f + C_t^i \quad \ldots (6)$$

The i_t and g_i are the two output of the input gate and node respectively. For equation (3 and 4) the weight are used [W_{ih}, W_{gh}, W_{ix}, W_{gx}] and [b_i, b_g] are biase values. The output of the output gate can be calculated as:

$$O_t = \sigma[(W_{oh}.h_{t-1}) + (W_{ox}.x_t) + b_0] \quad \ldots (7)$$

$$h_t = \tanh(C_t).O_t \quad \ldots (8)$$

C. Modeling with ARIMA

The autoregression-integrated moving average (ARIMA) [21] has 3 terms AR, I, and MA. The AR term corresponds to the lags of the stationaries series, the MA term corresponds to the lags of the forecast error, and the term corresponds to the order of differencing of the series to make it stationary. The AR model has the capability to forecast the future based on its immediate prior value within the time series and the MA model is equal to past error, multiplied by the coefficient. Essentially, ARIMA uses lags of the original series and lags of the forecast errors as regressors.
The auto-regressive model can be defined as the linear regression:

$$x_t = \mu + \xi_1 x_{t-1} + e_t \quad \ldots (9)$$

x_i is a predicted value against the finest value of gradient of ζ. x_{i-1} generated by the lag of time series. A constant μ, which is mean of xi, it is because for the amplitude the line and control the height of regression line. ei is uncorrelated random errors. Likewise, a generalized *qth* order autoregressive model can be define as:

$$x_t = \mu + \xi_1 x_{t-1} + \xi_1 x_{t-2} + \ldots + \xi_1 x_{t-q} + e_t \ldots (10)$$

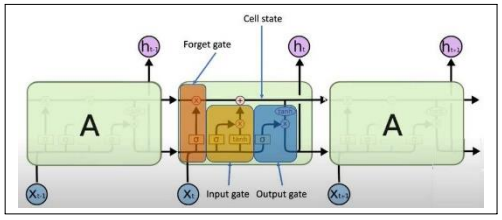

Fig. 3: A generalized multi-cell LSTM

The value of x depends on its previous q-time periods. Here, μ is the mean value.

Moving Average (MA) is a term, to get the mean of previous specific time periods, and calculate the errors. The value of the moving average can be calculated by the linear combination of stochastic white noise error terms as:

$$x_t = \sigma + e_t - \phi e_{t-1} \qquad (11)$$

Where x_t is a moving average of current and past selected periods. σ and φ are the constant and e is the error term. A generalized representation of order p is

$$x_t = \mu + e_t - \phi e_{t-1} - \phi e_{t-2} - \ldots - \phi e_{t-p} \qquad (12)$$

D. Proposed Methodology

1) Dataset: Usually, two steps are involved for air quality prediction (i) data gathering & cleaning, and predictive model. In most cases, the data is collected in real-time. In this study, data has been collected from an IoT system containing sensors. Initially, the minute-minute data of each day of 10 different locations have been sent and stored in the cloud. Three different sensors are installed in the IoT system meteorological, gasoline, and particulate, detailed in Table I.

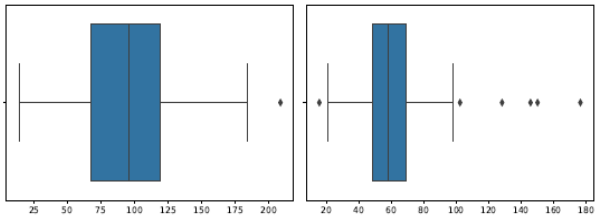

Fig. 4: Outlier presentation of PM 2.5 of two different stations

2) Data Wrangling: Due to electricity failure, Internet (WiFi) problems, and IoT system errors data is not forwarded to the cloud on a specific day, or from a specific sensor which leads to missing values in the dataset. The data has been analyzed for removing the structuring, missing values, and outliers issues. Usually, time series data fill by interpolation [6], in the proposed model, interpolation is used for filling the missing values. In this study, the structuring issue has been handled by regex. Outlier data is harmful to regression models, due to out of calibration, it is a chance to receive a high or low value from the sensor. In our case, a few outliers have been found as shown in Figure 4. For the outlier handling, we preferred to bring these outliers into the dataset by using the NaN technique.

E. Model Development

In this study, real-time data (Ammonia and PM2.5) collected from 10 locations in Karachi is used for prediction based on ARIMA and LSTM models. The collected data is non-seasonal as shown in Figure 6, in different order (p, d, q) of ARIMA has been used for different model buildings. The implementation of ARIMA has already been discussed in III-C. The minute-minute data, average has been calculated for preparing the day-wise sequence. The ARIMA model is sensitive to outliers, the existence of outliers may mislead the model followed by low performance. Outliers have been handled before the ARIMA model based on the 'NaN' method. A scalar normalization is used before the implementation of LSTM as it is the primary requirement of the model.

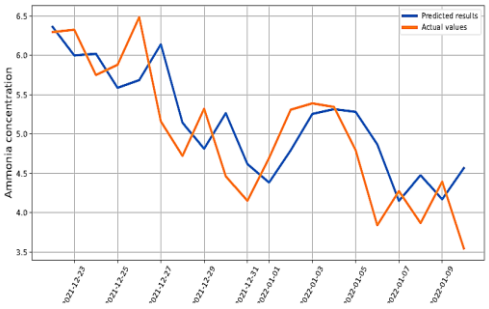

Fig. 5: ARIMA-based: Ammonia prediction of the Maymmar Dataset

4. Result and Discussion

In this paper, we proposed to use ARIMA & LSTM for Ammonia and PM2.5 forecasting. The data was collected from a real-station IoT system installed at 10 different locations in Karachi, detailed in Table I. In this study, four experiments have been performed Ammonia prediction based on ARIMA and LSTM and PM2.5 prediction based on ARIMA and LSTM. In this research, we used data from 4 locations from available 10 locations as others have incomplete data. State-of-the-art performance methods were used to evaluate the experiment results, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

A. Performance Evaluation Tools

Model performance evaluation is one of the key parts of machine learning algorithms. It helps to compare models, testing the weaknesses of the model viz overfitting and underfitting. In this study, we used the most recommended tools for forecasting MAE, MSE, and RMSE. These have different types of indicators with the minimum value towards 0 of MAE indicating an excellent result, the less difference between MAE and RMSE indicates the low variance in prediction error. The mathematical representation of these equations is 13 − 15:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}| y_i - \lambda(x_i) | \quad (13)$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \lambda(x_i))^2 \quad (14)$$

$$RSME = \frac{1}{m}\sqrt{\sum_{i=1}^{m}(y_i - \lambda(x_i))^2} \quad (15)$$

Where y deals with the actual test results and $\lambda(x)$ is the predicted results of the m test sample.

B. Experiment-1

The experiment was conducted based on ARIMA model for forecasting the Ammonia at "Maymmar" and "Federal B. area" locations. Currently, 70 days of data were available which covers three months November, December, and January 2021 − 2022. In this section, the experiment was conducted with consideration of the first 50 days as training, and the remaining 20 days were used for testing for Maymmar station and for F.B. Area station, first 65 days were considered for training and remaining 20 for testing. The parameters of ARIMA order for training was selected (p, d, q) is (1, 0, 0) and (0,1,0) for Maymmar and F.B. area locations dataset, respectively.

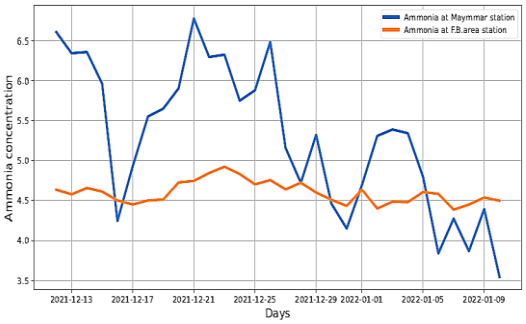

Fig. 6: Non-seasonal behavior of Ammonia at FB Area and Maymmar

TABLE II: ARIMA-based Model forecast of PM2.5 at Malir and North locations

Sno.	Station	Mean	RMSE	MAE	MSE
1	Maymmar	57.96	19.244	13.752	370.359
2	F.B. Area	90.35	28.881	25.755	834.112

Figure 6 shows a non-seasonality behavior of data at both stations. Detailed results are presented in Table III. The experiment results are outperformed as MAE values are << of corresponding means. The model was designed for Maymmar and F.B. area datasets, both have a small MAE of 0.471 and 0.1141, receptively. The minimum values of MAE results prove the excellency of the model and the originality of the dataset and these help for future accuracy of forecasting. The less difference of 0.091 and 0.038 between MAE and RMSE of the Maymmar and F.B. area receptively, reflect the low variance in prediction and high precision. Figures 5 and 7 show the comparison of predicted and actual values of test data.

C. Experiment-2

In this section, ARIMA-based experiments were conducted for PM2.5 prediction of the "Malir" and "North" stations datasets. Malir station has 116 days of data covering 4 months from November 2021 to February 2022 and North station has 70 days of data from November 2021 to January 2022. The IoT system is locally developed for a funded project titled TDF approved by the Higher Education Commission, due to the calibration of the IoT system, electricity issues, and WiFi connectivity lead to irregularity and less data collection of data. Regardless of all issues, the results of the implemented algorithms on the dataset outperform. A Non-seasonality diagram of PM2.5 of Malir and North are shown in Figure 9. The parameters of ARIMA order for training was selected (p,d,q) is (2, 0, 0) and (1,0,0) for Malir and North locations dataset, respectively. Detailed results are shown in Table II.

The experiment results are outperformed as MAE values are << of corresponding means. Both models have a small MAE of 13.752 and 25.785 on the Malir and North dataset, receptively. A close value of MAE and RMSE 13.752 and 19.244 at Malir station shows that the variance of prediction error is minimum having high precision. Similar behavior is shown at North station i.e. MAE and RMSE of 25.755 and 28.881, respectively. The minimum values of MAE results also prove the excellency of the model and originality of the dataset and these help for future accuracy of forecasting. Figures 10 and 10 show the comparison of predicted and actual values of test data.

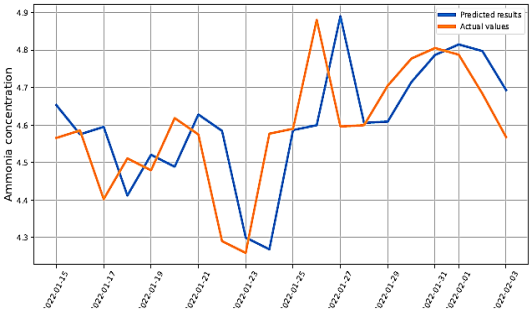

Fig. 7: ARIMA-based: Ammonia prediction of F.B. Area Dataset

TABLE III: ARIMA-based Model forecast of Ammonia at Maymmar and F.B. area locations

Sno.	Station	Mean	RMSE	MAE	MSE
1	Maymmar	5.15	0.562	0.471	0.316
2	F.B. Area	4.80	0.186	0.133	0.034

D. Experiment-3

Another LSTM-based experiment was conducted for "Ammonia" gasoline detection of the "Maymmar" and "F.B. Area" stations datasets. Maymmar station has 70 days of sequential data covering 3 months from November 2021 to January 2022 and F. B. Area station has 85 days without break data from November 2021 to February 2022. Regardless of the small size of the data set, the results of the implemented algorithm on the dataset outperformed. The data of both locations have Nonseasonality behavior as shown in Figure 6. A cumulative 50 and 65 data were selected for the training of the Maymmar and F.B. area respectively and the remaining 20 data for training for both stations. The experiment was run with 100 internal neurons along with 135 and 125 epochs for Maymmar and F.B. area respectively. Detailed results are shown in Table IV. The experiment results are outperforming as RMSE values are << of corresponding means. Both models have small MAE of 1.661 and 1.948 on the Maymmar and F. B area datasets, receptively. A close value of MAE and RMSE 1.661 and 1.948 at Maymmar station shows that the variance of prediction error is minimum having high precision. Similar behavior is shown at F. B Area station of MAE and RMSE of 0.121 and 0.151, respectively. Both algorithms have minimum MAE results that reflect the originality of the dataset which was collected by a real-time IoT system. Figures 13 and 12 show the comparison of predicted and actual values of test data.

E. Experiment-4

The experiment was conducted based on the LSTM model for forecasting the pm2.5 of the "Maymmar" and "Federal B. area" locations datasets. In this experiment, 100 days of sequential data was available which covers 4 months in early October 2021 to

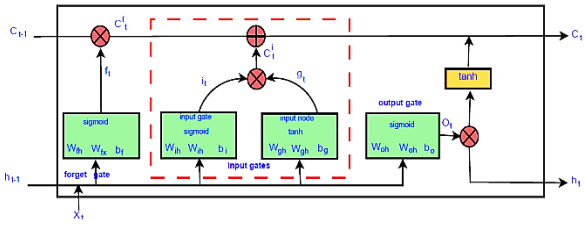

Fig. 8: A detailed view of LSTM cell

early February 2022. In this section, the experiment was conducted with consideration of the first 70 days as training, and the remaining 20 days were used for testing for Malir station and for F.B. Area station, the first 40 days were considered for training, and the remaining 19 for testing. A detailed result is presented in Table V. The experiment results are outperformed as RMSE values are << of the corresponding means of both stations. The model was designed for Malir and North datasets, both have a small MAE of 21.694 and 30.001, receptively. The minimum values of MAE results prove the excellency of the model and the originality of the dataset and these help for the future accuracy of forecasting. A small difference of 3.632 and 6.910 between MAE and RMSE of Malir and North stations receptively, reflect the low variance in prediction error and high precision. Figures 5 and 7 show the comparison of predicted and actual values of test data.

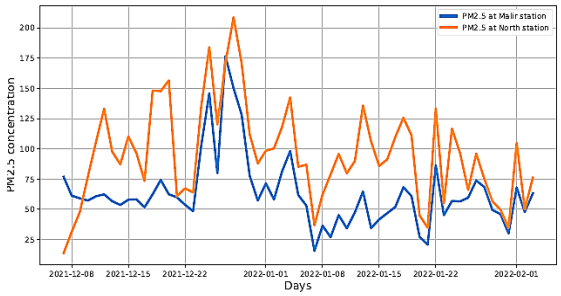

Fig. 9: Non-seasonal behavior of PM2.5 at Malir and North

TABLE IV: LSTM-based Model for the forecast of Ammonia at Maymmar and F. B. Area locations

Sno.	Station	Mean	RMSE	MAE	MSE
1	Maymmar	5.00	1.948	1.661	3.795
2	F.B. Area	4.59	0.151	0.121	0.0229

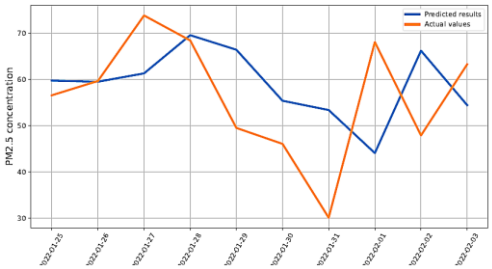

Fig. 10: ARIMA-based: PM2.5 prediction of North Dataset

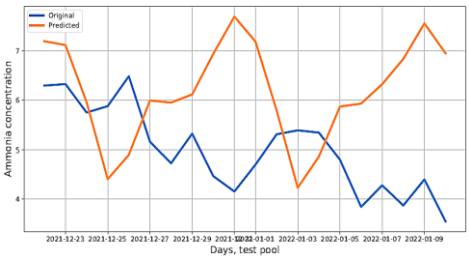

Fig. 12: LSTM-based: Ammonia prediction of Maymmar Dataset

5. Conclusion

In this research air quality prediction is proposed based on a locally developed IoT system for a funded project of the Higher Education Commission (HEC). These IoT systems are installed at 10 different locations of city Karachi with gasoline, particulate, and meteorological sensors. Data is collected from the stations for further analysis and forecasting. In this study, a regression ARIMA and deep learning LSTM algorithms are proposed to use for forecasting of 4 stations data and validate it regarding the IoT-based collection. The experiments were specially designed for PM2.5 and Ammonia prediction at four locations using a primitive IoT-based data collection.

The proposed approach yielded excellent performance in all four experiments in terms of MAE, RMSE, AND MSE. For the case of Ammonia, a value of 0.562 RMSE is obtained which is very close to the mean value. Similarly, very close values of 0.186 and 0.133 of RMSE and MAE were achieved respectively. In the LSTM-based experiment for Ammonia prediction, comparable a very low value of 1.948 RMSE was achieved from the corresponding mean. A very small difference value of 0.287 between RMSE and MAE indicates a low variance in predicting error and high precision.

Fig. 11: ARIMA-based: PM2.5 prediction of Malir Dataset

6. Acknowledgment

This work is supported by the Higher Education Commission (HEC) of Pakistan under the Technology Development Fund (TDF) grant. The project id is TDF 03 − 028.

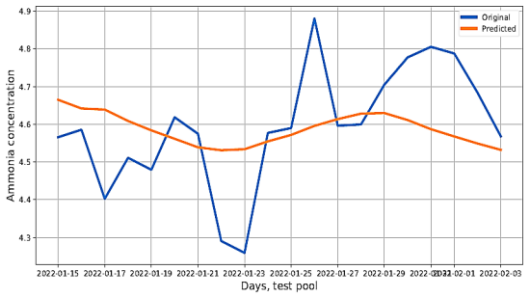

Fig. 13: LSTM-based: Ammonia prediction of F. B. Area Dataset

TABLE V: LSTM-based Model forecast of PM2.5 at Malir and North locations

Sno.	Station	Mean	RMSE	MAE	MSE
1	Malir	49.68	25.317	21.694	640.951
2	North	80.48	36.911	30.001	1362.441

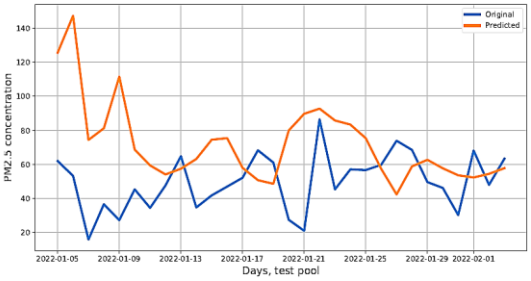

Fig. 14: LSTM-based: PM2.5 prediction of Malir Dataset

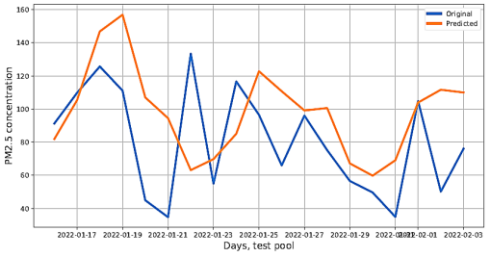

Fig. 15: LSTM-based: PM2.5 prediction of North Dataset

References

[1] U. Nation, "UN Environment Program 2022," https://www.unep.org/explore-topics/air/what-we-do/monitoring airquality/, 2022, [Online; accessed 1972-2022].

[2] D. Iskandaryan, F. Ramos, and S. Trilles, "Air quality prediction in smart cities using machine learning technologies based on sensor data: a review," *Applied Sciences*, vol. 10, no. 7, p. 2401, 2020.

[3] W. Mao, W. Wang, L. Jiao, S. Zhao, and A. Liu, "Modeling air quality prediction using a deep learning approach: Method optimization and evaluation," *Sustainable Cities and Society*, vol. 65, p. 102567, 2021.

[4] S. Sigamani and R. Venkatesan, "Air quality index prediction with the influence of meteorological parameters using a machine learning model for iot application," *Arabian Journal of Geosciences*, vol. 15, no. 4, pp. 1–12, 2022.

[5] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," *Big data and cognitive computing*, vol. 2, no. 1, p. 5, 2018.

[6] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, and L. Huang, "A predictive data feature exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30 732–30 743, 2019.

[7] J. K. Sethi and M. Mittal, "A new feature selection method based on machine learning technique for air quality dataset," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 697–705, 2019.

[8] Y. Cheng, S. Zhang, C. Huan, M. O. Oladokun, and Z. Lin, "Optimization on fresh outdoor air ratio of air conditioning system with stratum ventilation for both targeted indoor air quality and maximal energy saving," *Building and Environment*, vol. 147, pp. 11–22, 2019.

[9] Z. Zhang, Y. Zeng, and K. Yan, "A hybrid deep learning technology for pm2. 5 air quality forecasting," *Environmental Science and Pollution Research*, vol. 28, no. 29, pp. 39 409–39 422, 2021.

[10] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Deep air quality forecasting using hybrid deep learning framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2412–2424, 2019.

[11] B. A. A. Abdulali and N. Masseran, "Artificial neural network (ann) and arima models for better forecast of the air pollution data in malaysia," *Sch J Phys Math Stat*, vol. 10, pp. 184–196, 2021.

[12] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128 325– 128 338, 2019.

[13] A. Pappa and I. Kioutsioukis, "Forecasting particulate pollution in an urban area: From copernicus to sub-km scale," *Atmosphere*, vol. 12, no. 7, p. 881, 2021.

[14] Q. Wu and H. Lin, "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors," *Science of the Total Environment*, vol. 683, pp. 808–821, 2019.

[15] A. Barthwal and D. Acharya, "An iot based sensing system for modeling and forecasting urban air quality," *Wireless Personal Communications*, vol. 116, no. 4, pp. 3503–3526, 2021.

[16] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[17] Y. Hu, A. Huber, J. Anumula, and S.-C. Liu, "Overcoming the vanishing gradient problem in plain recurrent networks," *arXiv preprint arXiv:1801.06105*, 2018.

[18] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in narx recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 7, no. 6, pp. 1329–1338, 1996.

[19] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," Advances in neural information processing systems, vol. 9, 1996.

[20] C.-J. Huang and P.-H. Kuo, "A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities," Sensors, vol. 18, no. 7, p. 2220, 2018

[21] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid arima–ann model for forecasting time series data," Applied Soft Computing, vol. 23, pp. 27–38, 2014.