Can Machine Learning Models be Used to Predict Pollutants based on Measured Other Pollutants?

Steven Poore[1] and Cristinel Ababei[2]
[1]Electrical and Computer Engineering, University of Kentucky
[2]Electrical and Computer Engineering, Marquette University
Email: stevenpoore@uky.edu, cristinel.ababei@marquette.edu

***Abstract*—In this paper, we investigate the use of machine learning (ML) models to estimate or predict concentrations of pollutants based on measured concentrations of other pollutants. Such models could be used in air quality index (AQI) detection systems to decrease the number of physical sensors in order to reduce overall and maintenance costs. Five different long-short term memory (LSTM) models were explored in the preliminary investigation. The most accurate model was then selected for further refinement via simple hyperparameter search. The final refined model was trained and tested on four different air quality datasets from four different countries. Simulation results indicate that prediction of pollutant concentrations based solely on measured concentrations of other pollutants is not accurate enough to warrant total sensor replacement with ML models. However, when the same ML models are provided as input past measurements of the predicted pollutant rather than previously predicted values, the prediction accuracy is excellent. We conclude that while ML models are not yet accurate enough to completely replace physical sensors, such models could be very helpful to provide predictions in situations of sensor failure and thus to guarantee continuous sensor fusion processes.**

***Index Terms*—air quality index (AQI), machine learning, long-short term memory, LSTM**

I. Introduction and Related Work

Monitoring of air quality is extremely important because air pollution can negatively impact human health; in extreme cases premature mortality are side-effects of common indoor pollutants. Particularly, monitoring indoor air quality is very important as people in the US and Canada spent on average 87 percent of their time in buildings, according to a National Human Activity Pattern Survey [1]. Causes of indoor pollutant concentration include cooking with heat and gas, furnaces, smoking, and vehicle emission [1], [2]. Due to the major health risks, it is important that air pollution is monitored to ensure a safe indoor environment, as the first step in improving air quality in an environment is knowing when significant pollution is present.

An effective way to measure air pollution is the air quality index (AQI) scale, which uses measurements of concentrations of multiple pollutants to estimate the AQI value described later in the paper. The majority of previous work studied methods of predicting future pollutant levels based on present pollutant measurements using machine-learning algorithms for mostly outdoor environments. For example, the study from citeMa2020 proposed a multivariate linear regression model with outdoor NO2, SO2, O3, CO, PM2.5, PM10 concentrations and temperature as inputs to predict future AQI levels. The dataset used to develop the model was from the 2018 data for China's online air quality and analysis platform. The study reported that the model predictions were within 10% to the real values. The work in [3] tested multiple machine-learning algorithms for air-quality prediction. The study found that gradient boosting regression (GBR) and random forest regression (RFR) algorithms offered the best prediction accuracies. Similarly, the study in [4] tested multiple machine learning (ML) models including decision tree, random forest, support vector machine (SVM), and artificial neural network (ANN) to predict AQI. The data used to train and test the models came from a Kaggle online dataset which featured hourly levels of AQI and pollutants measured at different stations across various cities in India over three years. The study reported that the random forest model was the best with a maximum of 74% accuracy. The authors of [5] compared results from various studies on outdoor AQI prediction and concluded that neural networks (NNs) and boosting models were superior in providing accurate predictions for outdoor AQI. We only found one study in [6] that focused on an indoor environment. It presented a regression model for real-time prediction of PM10 in an indoor environment using CO, VOC and humidity as inputs to train the model. The dataset used for this study was gathered from an indoor air quality (IAQ) monitoring system installed inside the National Institute of Technical Teachers Training and Research in Chandigarh, India. The IAQ system recorded pollutant concentration measurements every fifteen minutes for six months to gather the data used for the model.

In contrast to previous work that focused on forecasting future AQI values, in this paper, our goal is to develop machine learning models to predict or estimate concentrations of selected pollutants based on measurements of other pollutants. We are interested in such models because they could be used as substitutes for actual sensor units to reduce system and maintenance costs or as a technique to fill in missing measurement data in realtime during sensor failures. This approach would enable accurate AQI estimations while decreasing system and maintenance costs by reducing the amount of sensors used to a minimal number.

II. Background on Air Quality Index

In this paper, we assume that the air quality is estimated using the classic United States Environmental Protection Agency (US EPA) AQI model, which is one of the most common AQI

Table I
POLLUTANTS RANGE VALUES AND HEALTH CATEGORIES.

AQI Less Than	Concentration ($\mu g/m^3$)						(Index j) Health Risk Category
	$SO2$	$NO2$	CO	$O3$	$PM2.5$	$PM10$	
50	50	40	2	100	35	50	(1) Good
100	150	80	4	160	75	150	(2) Moderate
150	475	180	14	215	115	250	(3) Unhealthy for sensitive groups
200	800	280	24	265	150	350	(4) Unhealthy
300	1600	565	36	800	250	420	(4) Very unhealthy
400	2100	750	48	1000	350	500	(6) Hazardous
500	2620	940	60	1200	500	600	(7) Severe

models [7]. The index is calculated using the concentration values of six pollutants. The numerical scale used is 0-500 and the overall index is calculated by taking the maximum value among all individual sub-AQI indices of all considered pollutants.

Given the measured concentrations of the six, $n = 6$, pollutants SO2, NO2, PM2.5 PM10, CO, and O3, the individual sub-AQI index AQI_i of each pollutant $i = 1..n$ is calculated with the following expressions:

$$AQI_i = \frac{(AQI_{i,j} - AQI_{i,j-1})}{(C_{i,j} - C_{i,j-1})} \times (C_{i,m} - C_{i,j-1}) + AQI_{i,j-1}, j > 1 \quad (1)$$

$$AQI_i = AQI_{i,1} \times \frac{C_{i,m}}{C_{i,1}}, j = 1 \quad (2)$$

Where j is the health category index. The seven health categories are shown in Table I. $C_{i,m}$ is the measured concentration of pollutant i, which falls into the range defined by $C_{i,j-1}$ and $C_{i,j}$, as the upper limit concentrations of the health category indices $j-1$ and j. $AQI_{i,j}$, $AQI_{i,j-1}$ are the AQI values corresponding to $C_{i,j-1}$, $C_{i,j}$ for pollutant i. Once the individual sub-AQI values are calculated as above, the total AQI is calculated as the maximum value among all of them:

$$AQI = \max_{i=1..n} AQI_i \quad (3)$$

III. DATASETS WITH POLLUTANT CONCENTRATIONS

A. Datasets Summary

We use four different datasets, which are all publicly available on the Internet. These datasets and their characteristics are listed in Table II. All datasets have pollutant concentrations measured hourly, i.e., 24 values per day. Dataset 1 has data from multiple sites in India, but we only use the measurements from Amaravati, Andhra Pradesh. Dataset 2 is from Hong Kong, dataset 3 is from Beijing, and dataset 4 is from Seoul. All datasets have hourly measurements of NO2, SO2, O3, CO, PM2.5, and PM10.

All datasets required a limited amount of pre-processing, which we describe next. For each dataset, we pruned the first entries until until we reached the first 24 h with complete data. That is because we then filled missing data values with the values of 24 h earlier, i.e., the concentration at the same hour in the previous day. Also, the last entries in the dataset were pruned such that data for the last day has all 24 values. In this way, each dataset starts with a datapoint at 0 h and ends with a datapoint at 23 h. Our final pre-processed datasets will be made available in the github repository we have created for this project.

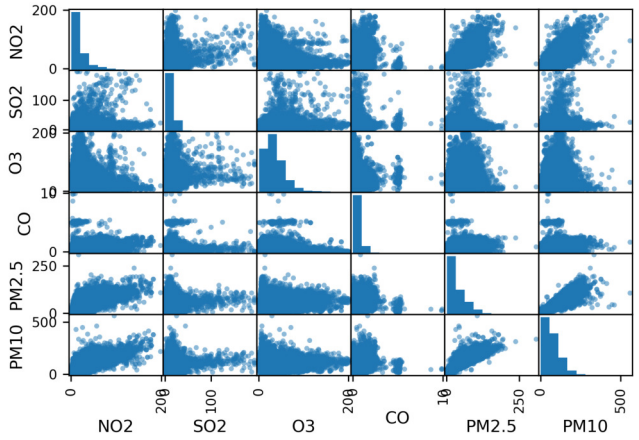

Figure 1. Scatter matrix for dataset 1.

B. Correlations and Trends

To investigate correlations between the six pollutants, a scatter matrix was created for each dataset. For example, the scatter matrix plot for dataset 1 is shown in Fig. 1. Because the scatter plots for the other datasets are similar, we do not show them here in the interest of space. Looking at the these scatter matrices, some key correlations can be noticed. NO2 appears to have good positive correlation with particulate matter pollutants PM2.5 and PM10, negative correlation with CO, and weak correlations to SO2 and O3. Similarly, PM2.5 shows a good correlation with NO2 and PM10, somewhat good correlation with SO2, negative correlation with CO, and weak correlation to O3. To a large extent, these observations were confirmed by previous literature. For example, the study in [8] found that the pollutants in the group formed by SO2, NO, NO2, CO are more correlated among themselves. The other group of correlated pollutants includes PM10, PM2.5, and O3. The study also reported a strong negative correlation between O3 and NO and NO2: when concentration of NOx decreases the O3 concentration increases. Similar negative correlation was observed for O3 with SO2 and particulate matter in [9]. The study in [9] reported also a positive correlation between particulate matter and NO2, SO2, and CO. A strong correlation between PM10 and SO2 was found in [10]. Therefore, informed by the above observations and previous literature, in this paper, we decided to investigate NO2 and PM2.5 as the candidate pollutants whose concentrations to predict.

IV. A CASE FOR AQI SYSTEMS WITH FEWER SENSORS

A. General Approach

Our idea is to build an AQI system that uses the EPA model from section II, but, in which one or more pollutant concentrations are not directly measured with costly sensor units but, indirectly estimated using machine learning models.

Table II
SUMMARY OF DATASETS WITH OUTDOOR POLLUTANT CONCENTRATIONS.

			Measured Pollutant Concentrations					
Dataset	**Location**	**Length**	$NO2$	$SO2$	$O3$	CO	$PM2.5$	$PM10$
1	Amaravati, Andhra Pradesh, India	2017/12/02 - 2020/06/30	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$
2	Beijing, China	2013/03/01 - 2017/02/28	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$
3	Seoul, Korea	2017/01/01 - 2019/01/23	(ppb)	(ppb)	(ppb)	(ppb)	$(\mu g/m^3)$	$(\mu g/m^3)$
4	Madrid, Spain	2009/12/31 - 2022/03/31	$(\mu g/m^3)$	$(\mu g/m^3)$	$(\mu g/m^3)$	(mg/m^3)	$(\mu g/m^3)$	$(\mu g/m^3)$

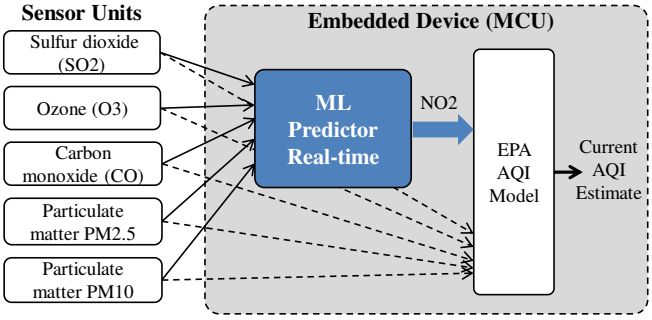

Figure 2. System level block diagram illustrates the replacement of the NO2 sensor unit with an ML model.

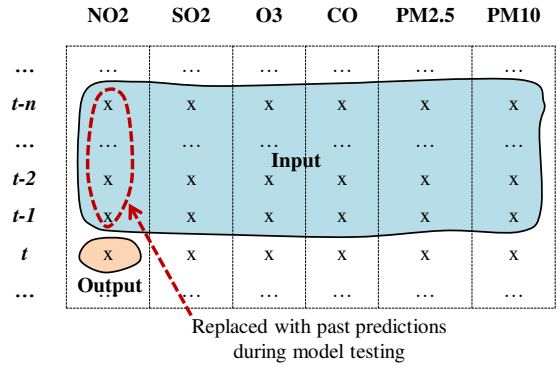

Figure 3. Illustration of what is used as input and output for studied ML models out of a given multivariate dataset.

In other words, recalling the discussion from the previous section, we use a machine learning model to predict for example the concentration of NO2 (or PM2.5) based on measured concentrations of the other pollutants as well as on previous predictions of NO2. Implementing this model into an AQI estimation device would allow for an AQI measurement to be made with only five sensors instead of six, reducing overall and maintenance costs while still providing an accurate AQI estimation. Fig. 2 illustrates how the ML prediction is used in the overall AQI estimation system.

B. Machine Learning Models

We have developed and investigated preliminarily five different machine learning models, which were tested on dataset 4 presented in section III. The models tested include: 1) *Model 1:* Simple long-short term memory (LSTM), 2) *Model 2:* Stacked LSTM, 3) *Model 3:* LSTM Encoder-decoder, 4) *Model 4:* CNN-LSTM Encoder-decoder, and 5) *Model 5:* Deep neural network (DNN). All LSTM models were developed to be single/multi-step multi-variate input with single-step univariate output. That is, the input into the model represented the pollutant concentrations from one or more time steps as illustrated in Fig. 3. In the case of the DNN model, the above inputs were all fed directly as a simple array of values.

Table III
PREDICTION ERRORS OF NO2/PM2.5 FOR DATASET 1.

Model	**RMSE**	**MAE**
LSTM	4.29/7.82	2.474/5.42
Stacked LSTM	4.35/8.02	3.25/5.74
LSTM Encoder-decoder	4.36/7.9	3.25/5.54
CNN-LSTM Encoder-decoder	4.51/9.53	3.19/6.76
DNN	4.33/7.25	2.64/5.06

During this model development phase, the number of epochs (100), batch size (100), optimizer (adam), loss function (mae), instances in the input (2 time steps), and predicted variable (NO2 or PM2.5) were kept the same between the models for consistency. The dataset was split into 70%/30% for training/testing. The dataset was normalized to $(0, 1)$ before training. To evaluate the trained models, we use root-mean-square error (RMSE) and mean-absolute error (MAE). The prediction errors of each of the five models are listed in Table III. We observe that after testing these models, *Model 1* was the most accurate. Therefore, *Model 1* is selected to be refined and investigated further.

C. Model Refinement

After determining the best model (*Model 1:* Simple LSTM) among those investigated, additional investigations were performed to further refine the selected model to increase its prediction accuracy. To that end, we explored multiple different values for various model hyperparameters, including: number of units in the LSTM layer, batch size, loss function, optimizer, and number of time-steps used as input. In all experiments we used EarlyStopping with the argument *patience*=10. A summary of the investigated hyperparameters is presented in Table IV.

During this investigation, only the hyperparameter being refined was changed while others were kept constant. Each such experiment would be run 10 times and the average values for RMSE were calculated. For example, the results of refining batch size are shown using a box and whisker plot in Fig. 4. It shows that a $batch\_size = 30$ provided best results in this experiment. After this investigation, we found the most accurate predictions were obtained with a refined model that:

Table IV
SUMMARY OF INVESTIGATED HYPERPARAMETERS.

Hyperparameter	Values investigted
units on LSTM layer	5, 10, 30, 60
batch size	10, 30, 50, 70, 100, 200
loss function	mae, mse, binary_crossentropy
optimizer	Adam, Sgd, Adagrad, Adadelta, RMSprop
n_input_steps	1, 2, 4, 5, 6, 12, 24

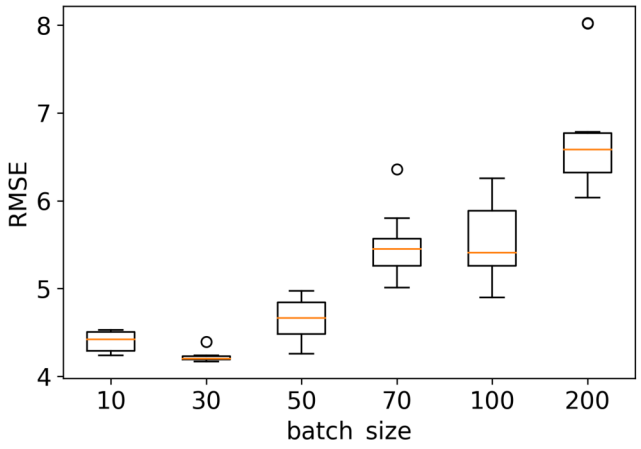

Figure 4. RMSE value results of NO2 prediction with *Model 1* tested on dataset 1 during model refinement in terms of *batch_size*.

used 30 units on the LSTM layer, batch size of 30, loss function *mae*, and optimizer *Adam*, and problem being framed with 2 lag time-steps used as input.

V. RESULTS

In this section, we conduct two types of experiments. In *Experiment 1*, the refined model is tested on the test portion of the datasets, where the past values of the predicted pollutants (NO2 or PM2.5) are their actual real measurements (available in the original datasets). In *Experiment 2*, we eliminate completely the actual real measurements of the target pollutants from the test portion of the dataset and replace them with the actual predictions of either NO2 or PM2.5. In other words, for any prediction at any time during the evaluation of the test dataset, the past values of NO2 and PM2.5 are previous predictions of NO2 and PM2.5 obtained with the refined model - during the step-by-step forward evaluation. For the first n_input_steps evaluations, the past values of NO2 and PM2.5 are set to 0.5 (middle value of the normalization interval) in order to mimic their unavailability in the practical implementation of the proposed model - where the proposed model replaces actual NO2 and SO2 sensors. In that case, at the beginning of AQI monitoring no prior estimations of NO2 and PM2.5 are available, and therefore, they must be initialized to some default values. In all experiments, each dataset was split into 70%/30% for training/testing of the final refined model. Also, datasets were normalized to $(0, 1)$ before training.

Table V
PREDICTION ERRORS OF NO2/PM2.5 FOR ALL FOUR DATASETS OBTAINED WITH REFINED MODEL IN *Experiment 1*.

Dataset	RMSE	MAE
Dataset 1	3.33/6.7	2.02/4.55
Dataset 2	14.52/19.1	9.3/10.41
Dataset 3	0.006/18.8	0.004/3.64
Dataset 4	9.33/5.17	6.61/3.56

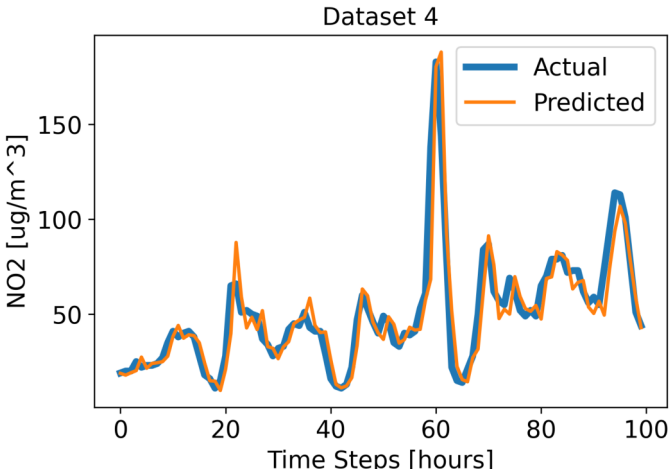

Figure 5. Prediction results of NO2 with refined model on 100 points from the test portion of dataset 4 in *Experiment 1*.

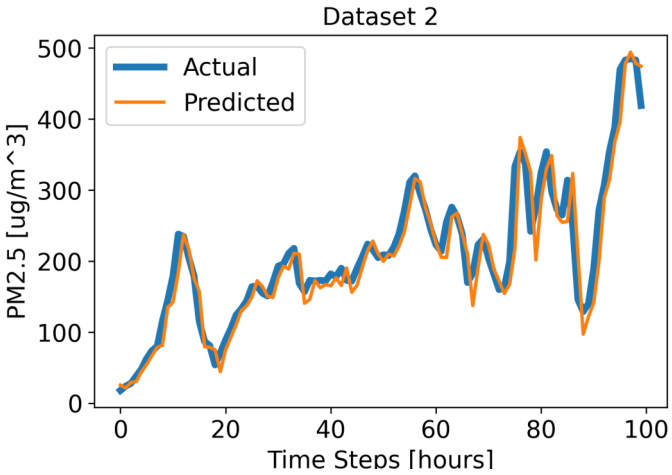

Figure 6. Prediction results of PM2.5 with refined model on 100 points from the test portion of dataset 2 in *Experiment 1*.

The results from *Experiment 1* are summarized in Table V, where prediction errors of NO2 and PM2.5 are listed. As examples of predictions, Fig. 5 shows the actual and predicted first 100 values from the test portion of the dataset 4, while Fig. 6 shows the actual and predicted first 100 values from the test portion of the dataset 2. Similar high accuracy predictions are observed in all cases for all datasets.

The results from *Experiment 2* are summarized in Table VI, where prediction errors of NO2 and PM2.5 are listed.

Table VI
PREDICTION ERRORS OF NO2/PM2.5 IN *Experiment 2*, WHEN PAST VALUES OF NO2 AND PM2.5 ARE AVAILABLE ONLY AS PRIOR PREDICTIONS DURING TESTING.

Dataset	RMSE	MAE
Dataset 1	9.86/16.11	7.44/10.73
Dataset 2	44.27/54.67	38.34/37.27
Dataset 3	0.019/17.82	0.016/9.7
Dataset 4	33.55/7.45	20.58/4.91

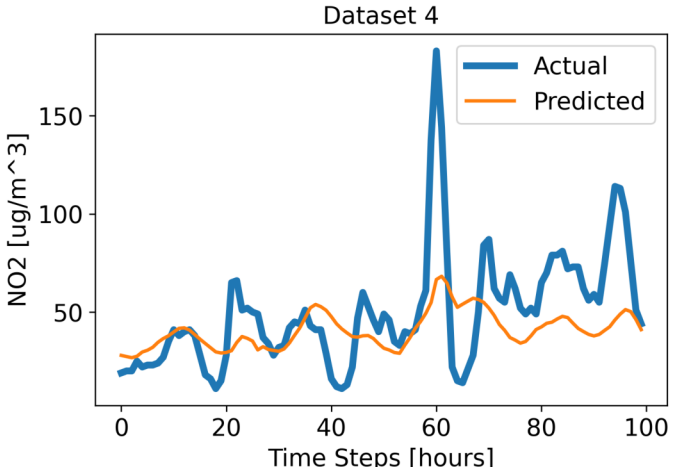

Figure 7. Prediction results of NO2 with refined model on 100 points from the test portion of dataset 4 in *Experiment 2*; past values of NO2 are available only as prior predictions.

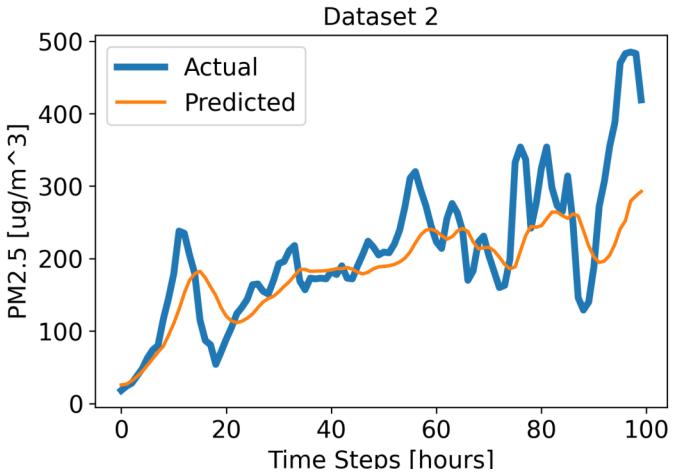

Figure 8. Prediction results of PM2.5 with refined model on 100 points from the test portion of dataset 2 in *Experiment 2*; past values of PM2.5 are available only as prior predictions.

Again, as examples of predictions, Fig. 7 shows the actual and predicted first 100 values from the test portion of the dataset 4, while Fig. 8 shows the actual and predicted first 100 values from the test portion of the dataset 2.

A. Discussion

We observe that in the first set of experiments (*Experiment 1*), the prediction quality is excellent. This can be seen in Table V, where RMSE and MAE values are very small for all investigated datasets, as well as in figures Fig. 5 and Fig. 6, where it can be seen that the predicted values follow very closely the ground truth values, which are available from the original datasets. In other words, the refined model can predict very well either NO2 or PM2.5 in *Experiment 1*. This is expected because in *Experiment 1*, the true real measurements of NO2 and PM2.5 are available as n_input_steps past values during each evaluation data point from the test portion of the datasets. In this case, each prediction of NO2/PM2.5 is done using as input to the model actual real measurements of past values of NO2/PM2.5 - the same way as it was done during the model training. While the results from *Experiment 1* are excellent, *Experiment 1* is not the real scenario in which we wanted to employ such prediction models (as described in Fig. 2). That is because, in the scenario described in Fig. 2, the ML models are intended to completely replace sensors for NO2 or PM2.5, case in which we do not have available the *ground truth values* of the concentrations of these pollutants, from the previous time steps, to feed as part of the input to the models used during inference. We only have available past *predicted values* from the previous time steps. This scenario is what *Experiment 2* mimics.

However, in *Experiment 2*, the prediction quality degrades - as observed in Table VI and figures Fig. 7 and Fig. 8. In this experiment, the actual real measurements (i.e., ground truth) of past values of NO2/PM2.5 are not available anymore, and each prediction of NO2/PM2.5 is now done using as input to the model *previous predictions obtained with the same model*. In this case, the prediction quality depends more on how correlated NO2/PM2.5 were to the other pollutants, as discussed earlier (Fig. 1). Such correlations are not perfect and therefore the refined model is bound to introduce prediction errors. In addition, because during testing past actual measurement values of NO2/PM2.5 are not available, the refined model will miss any variation or impact over NO2/PM2.5 caused by independent and random events that may have affected the true measurements recorded during the test interval.

We note that the accuracy of the refined LSTM models as substitutes of actual hardware sensors is not satisfactory to warrant their use as complete replacements of actual sensor units. Nevertheless, given the excellent performance observed in *Experiment 1*, these models still have practical value: they could be incorporated in the AQI system, but used as a back-up in the event of sensor units failure. In other words, between the moment when a sensor failure occurs and its repair/replacement, the proposed models could be triggered to generate estimations in real-time and thus to fill in measurement values that would otherwise be missing from the failed sensors. Furthermore, in the context of recently proposed *machine learning sensors* paradigm [11], the refined LSTM models can be integrated with the sensor units themselves and be used as estimation techniques in cases of missing measurements due to sensing elements failure.

VI. Conclusion

We investigated the use of LSTM models to estimate or predict concentrations of NO2 and PM2.5 pollutants based on measured concentrations of other pollutants. The objective was to answer the question of whether such models could be used as substitutes for actual sensor units in practical air quality index (AQI) detection systems - in order to reduce the overall cost of such systems. Results from *Experiment 2* (Table VI and Figs. 7, 8) indicated that prediction is not accurate enough to warrant total sensor replacement with ML models. However, when the ground truth values of the predicted pollutant concentrations at times previous to the prediction time are available in *Experiment 1* - and such values are included as input to the ML models - then, the prediction quality is excellent (Table V and Figs. 5, 6). Therefore, we concluded that the proposed ML models could still be very valuable in providing predictions in situations of temporary sensors failure - situations that are emulated by *Experiment 1*. The complete implementation in Python of the model development presented in this paper, together with the cleaned datasets, will be made publicly available at [12].

Acknowledgment

This work was supported by the National Science Foundation (NSF) grant REU Site 1950082. Any findings and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] R. Perez-Padilla, A. Schilmann, and H. Riojas-Rodriguez, "Respiratory health effects of indoor air pollution," *The International Journal of Tuberculosis and Lung Disease, vol. 14, no. 9, pp. 1079-1086*, 2010.

[2] J.A. Hoskins, "Health effects due to indoor air pollution," *Sage Publications, Indoor and Built Environment, vol. 12, no. 6, pp. 427-433*, 2003.

[3] C. Li, Y. Li, and Y. Bao, "Research on air quality prediction based on machine learning," *Int. Conf. on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2021.

[4] K.M.O.V.K. Kekulanadara, B.T.G.S. Kumara, and B. Kuhaneswaran, "Machine learning approach for predicting air quality index," *Int. Conf. on Decision Aid Sciences and Application (DASA)*, 2021.

[5] T. Madan, S. Sagar, and D. Virmani, "Air quality prediction using machine learning algorithms - a review," *Int. Conf. on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020.

[6] J. Saini, M. Dutta, and G. Marques, "Indoor air quality monitoring with IoT: predicting PM10 for enhanced decision support," *Int. Conf. on Decision Aid Sciences and Application (DASA)*, 2020.

[7] U.S. Environmental Protection Agency, "Technical assistance document for the reporting of daily air quality - The air quality index (AQI)," *EPA*, 2018.

[8] P.J. Garcia Nieto, E.F. Combarro, J.J. del Coz Diaz, and E. Montanes, "A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study," *Applied Mathematics and Computation*, vol. 219, pp. 8923-8937, 2013.

[9] R. Nigam, K. Pandya, A.J. Luis, R. Sengupta, and M. Kotha, "Positive effects of COVID-19 lockdown on air quality of industrial cities (Ankleshwar and Vapi) of Western India," *Nature Reports*, 2021.

[10] O.F. Althuwaynee, A.L. Balogun, and W. Al Madhoun, "Air pollution hazard assessment using decision tree algorithms and bivariate probability cluster polar function: evaluating inter-correlation clusters of PM10 and other air pollutants," *GIScience and Remote Sensing*, vol. 57, no. 2, pp. 207-226, 2020.

[11] P. Warden et al., "Machine learning sensors," *arXiv:2206.03266 [cs.LG]*, 2022.

[12] ML models for prediction of pollutant concentrations, GitHub Repository, https://github.com/eigenpi/ml-pollutant-prediction, 2023.