Indoor Air Pollution Forecasting using Deep Neural Networks

Jorge Altamirano-Astorga, Ita-Andehui Santiago-Castillejos, Luz Hernández-Martínez, and Edgar Roman-Rangel

Instituto Tecnológico Autónomo de México, Mexico City, Mexico
https://philwebsurfer.github.io/dlfinal/ {jaltami9,isantia2,lhern123,edgar.roman}@itam.mx

Abstract. Atmospheric pollution components have negative effects in the health and life of people. Outdoor pollution has been extensively studied, but a large portion of people stay indoors. Our research focuses on indoor pollution forecasting using deep learning techniques coupled with the large processing capabilities of the cloud computing. This paper also shares the implementation using an open source approach of the code for modeling time-series of different data sources. Further research can leverage the outcomes of our research.

1 Introduction

Monitoring the air quality and forecasting air pollution are both paramount for many daily activities [4], as air quality has direct impact on human health, agriculture, and the environment in general. Moreover, it is directly related to global warming and climate change [11].

Typically, air quality systems focus on monitoring specific components of outdoors atmospheric pollutants, including NO_2, NO_3, NO_x, O_2, CO_2, PM_{10}, $PM_{2.5}$, and IAQ index, using public monitoring stations that report their concentrations either daily, hourly, or by the minute [15]. However, some times these reports can be inaccurate, noisy, or simply not provided by the official sources during specific periods [12]. Outdoor air pollution has been demonstrated to have negative consequences in human health, specially in large metropolitan areas. Furthermore, tracking indoors air quality has been a neglected research area.

To face the potential lack of public information, as well as to contribute to the monitoring of indoors air quality, we propose the use of affordable indoors devices, that report air quality every few seconds, and that can be installed inside apartments. For the task of predicting air quality IAQ, we evaluate the predictive performance of different deep neural network-based approaches. Namely, we compare the performance of the multi-layer perceptron (MLP), convolutional neural networks (CNN), and the long short-term memory (LSTM). Our results show that it is possible to obtain an air quality monitoring system that is affordable and reliable for domestic usage.

Concretely, our contribution consists in presenting an exploration of different deep learning models and architectures with full open code, which allows replicability of time-series forecasting using cloud techniques.

The rest of this paper is organized as follows. Section 2 comments on previous work using deep learning techniques for air quality prediction. Section 3 provides details on the data acquired with the indoors device. Section 4 explains the different architectures evaluated for the task of air quality forecast. Section 5 presents our results, and section 6 our conclusions.

2 Related Work

There are some underlying reproductions that use models of neural systems from the perspective of their understanding of how neural networks work. Among other things, we find that some authors

work with artificial neural nets based on MLP-based architecture [3], unlike this research which tries different architectures, mainly based on LSTM. In addition to this, we try with different time windows and other hyperparameter tuning in order to find the one that best predicts the air quality index.

Some researchers have proposed a systems to monitor individual pollution components (pollutants) through wireless networks connecting arrays of sensors. Other studies have incorporated indexes (such IAQ) or score-based systems to determine and forecast IAQ level. We adopted the latter. Additionally, we compared for data accuracy from different sources such as OpenWeatherMap [18] and Mexico City Government Air Quality Monitoring [19,17].

Based on the scope of our investigations, it is common that IAQ monitoring Artificial Neural Networks (ANNs) are used to predict or forecast the value of air quality or the value of air pollution. This project differs from previous research since we use a wide variety of architectures and expanded processing capacity with Google Vertex. The functionality of this system and the methodology of this project are explained in the next section.

No outstanding results were found in weather or pollution forecasting in the papers cited. So, it is natural to build and improve upon these weaknesses and address them.

In the paper of Abdullah [1] the authors make special emphasis on the complexity and non-linear associations between air quality, meteorological, and traffic variables. This is often a limitation of traditional machine learning methods. They also had the idea of updating ANN weights with genetic algorithms, calling it Optimized ANN (OANN). The data includes 16 hours of daily information from 2014 to 2016, it came from the Ministry of Works, Malaysia, while the air pollution and meteorological datasets are collected from the Department of Environment (DOE), Malaysia. These data were used to predict the concentration of CO, NO, NO_2, and NO_x. The models included are ANN, RF, Decision Trees, and the metrics used to measure their performance were: MAE and MSE for all four pollutants.

The objective of the work of Cakir [2], 2020 was to compare the performance of the ANN against multiple linear regression; associating the weather condition with some ambient air measures. The data corresponds to the average hourly concentrations of the particles during the years 2012-2015. The authors make predictions of PM_{10}, NO_2, and O_3. The performance was measured with MAE, RMSE and R^2. They found a correlation between in and out variables, also found that shallow ANNs seem to work better than deeper ones, but MLR seems to work better than ANN. This seems to be more competitive against their previous work (2017), where the MLR equations obtained were used to predict the concentrations for the period of 2012–2013. The same data sets used to simulate both ANN and MLR, but predictors used in ANN are not the same with the independent variables of regression equations. During the model development in the work of 2020, independent variables of the MLR equations were used as a predictor in different ANN configurations; however, the obtained performances were not better than ANN trained with predictors given in this paper.

The dataset used in the work of Singh et al [10] came from Indira Gandhi International Airport at New Delhi, India, which includes 9 months of hourly information about meteorological factors, pollutant concentrations, and traffic information. The target variable is $PM_{2.5}$. Authors present an exploratory model using an LSTM with 75 units; however, they came with no strong conclusion, only that RMSE is lower for deeper LSTM (hence the 75 units).

The work of Saad [5] plays with information of continuous monitoring of indoor air quality among 22 days between 9.00 a.m. and 5.00 p.m. The authors try to identify source influence among 9 input variables (to known: CO_2, CO, O_3, NO_2, VOCs, O_2 and PM_{10}, temperature and humidity) and 5 conditions: ambient air, chemical presence, fragrance presence, foods and beverages presence, and human activity. Within the written work they talk about the engineering application system which is made up of three parts: sensor module cloud, base station, and service-oriented client. They used

2-layers ANN to learn to predict 5 outputs: Ambient environment, Chemical presence, Fragrance presence, Human activity, and Food and beverages. The performance measure used was the mean accuracy, which was found between 45 to 99%. With this measure they concluded that the ANN was a good tool to identify the most relevant predictors.

Bekkar et al [6] in their research published the negative impact on air in human health triggering cardiovascular diseases related to mortality, and therefore having an impact on the domestic economy. It also hints a probable relationship of pollutants and COVID-19 propagation. Their research also highlights the complexity of modelling the $PM_{2.5}$ pollution with different traditional statistical methods and machine learning methods. Bekkar focuses on deep learning with different architectures, comparing them using performance metrics. The dataset used in this research also contained less than 4% of missing values, and the spline linear interpolation method was used to fill the gaps in the missing data. The results presented in the paper show that the combined convolutional and LSTM network combination offered the best results. We found that this research is noteworthy in their ability to present the information and compare models; but lacks information regarding time-series preprocessing methods, such as the window sizes and fall short on explaining why the least history has the best results, which is contrasting with our own research.

Sotomayor-Olmedo et al [7] research focused on using Support Vector Machines with Mexico City data. They explored different kernels for SVM but did not provide further details on preprocessing on how they dealt with the missing values. The research explained briefly, but clearly, the general weather and pollution conditions of Mexico City. The authors of this paper highlight the computational overhead that may impose in forecasting applications, therefore suggesting lower Support Vector Machines with high accuracy as the best option. We tried focusing on covering those areas that could improve replicability of air quality research for other researchers' work.

Ramos-Ibarra et al [8] focused their paper research on the trends of atmospheric pollution in the greater Mexico City area using Kalman filters as an smoothing technique for several pollutants in the region. They use Mean absolute deviation, mean square error and mean absolute percentage error as their metrics for evaluation of their techniques. Their research handled the missing variables through the Kalman filters. Their research focused on the non compliance at the time of the local and global environmental regulations, and forecasting of the pollution concentration on a 7 day horizon from 2008 until 2018. The most remarkable outcome of this research is that the Kalman filters and the techniques used in the paper may be used by decision makers to tackle the air pollution problem from an integral perspective using statistical tools based on data.

Bing et al [9] research focused on predicting ozone air pollution in Mexico City forecasting using multiple linear regressions, neural networks, support vector machines, random forests and ensemble techniques. Their research clearly show the difficulties of having a greater prediction performance than 50%. They used all these techniques to evaluate contradictory outcomes of previous research on whether the Support Vector Machines versus Multi-layer Perceptron offer better performance, but the researchers published better performance with ensemble techniques that combine neural networks and support vector machines. This research used the performance metric: RMSE and MAE; but the paper lacked information regarding if those errors were scaled, as is usual in machine learning techniques with different variables in different scales, and it didn't explained how the data preprocessing was handled. Therefore, it is very difficult to replicate their findings. Their conclusions highlighted some of the limitations found using a single station.

None the previously described research presented public code published or it lacked the required technical detail for replicability. A lack of baseline models as suggested in Keras documentation regarding time-series processing was also addressed.

3 Data.

We use the sensor Bosch BME680 [15] that has a published precision of ±5% and a resolution of 1 IAQ. Our exploratory data analysis confirmed this. This sensor collects data every 3 seconds [16].

Additionally, we use two more data sources: OpenWeatherMap and Mexican Federal Government Pollution data (SINAICA), which have new observations every hour.

Preprocessing. We explored the data sets and we had 6,285,103 records of our sensors from February 2nd, 2021 until September 27th, 2021. Our average IAQ readings were 161.23 with an standard deviation of 72.85.

We resampled our sensor data every 1, 2 and 5 minutes. Getting the mean values for 1, 2 and 5 minute windows, respectively. We created a linear interpolation for the missing data points. The 5 minute resampled data was found to be a good balance between dataset size performance wise. Our data was split into training (70% of the data) and validation datasets (30%), without reshuffling, i.e. keeping the oldest records for the training dataset, and the validation dataset with the most recent data.

We had missing data on all datasets. Nonetheless, this was not a big concern on our experiments by using the above mentioned interpolation methods.

Variables. The 5 minute resampled data has 62,724 observations with the following variables:

- Sensor data:
  1. Temperature: continuous variable in Celsius degrees.
  2. Pressure: continuous variable in hectopascals (hPa).
  3. Humidity: continuous variable in relative humidity percentage (% rh).
  4. IAQ: discrete variable in EPA Indoor Air Quality Index.
- SINAICA Government Pollution Data:
  1. NO: continuous variable for Nitric Oxide parts per billion (ppb).
  2. NO_2: continuous variable for Nitrogen Dioxide parts per billion (ppb).
  3. NO_x: continuous variable for Nitrogen Oxide parts per billion (ppb). This is the sum of NO and NO_2 pollutants.
  4. CO: continuous variable for Carbon Monoxide parts per million (ppm).
  5. O_3: continuous variable for Carbon Monoxide parts per million (ppm).
  6. PM_{10}: continuous variable for Particle Matter with diameters of less than 10 microns measured in micrograms per cubic meter (μg / cm^3).
  7. $PM_{2.5}$: continuous variable for Particle Matter with diameters of less than 2.5 microns measured in micrograms per cubic meter (μg / cm^3).
  8. SO_2: continuous variable for Sulfur Dioxide parts per billion (ppb).
- OpenWeatherMap Data:
  1. Outdoor Temperature: continuous variable in Celsius Degrees.
  2. Outdoor Pressure: continuous variable in hectopascals (hPa).
  3. Outdoor Humidity: continuous variable in relative humidity percentage (% rh).

We use the previous list of variables as independent variables to forecast IAQ. More precisely, our models are fed with a vector of length 15, and predict an scalar output.

Postprocessing. All variables described in Variables subsection are numeric but on different scales, therefore we used the `MinMaxScaler()` transformation of Scikit-Learn for training our models. This transformation is inverted for the reporting of the Mean Absolute Error, to have the error metric in an interpretable scale.

We also applied time-series processing using Keras `timeseries_dataset_from_array()` function with hyperparameter tuning to find the "sweet spot" of performance and accuracy. This function creates a tensor, i.e., a vector of arrays with the history of previous observations of certain length in a sliding window fashion. This is "learned" during the training process of the models we applied in this research.

Data Access We are planning to publish the sensor data on the paper repository hosted on GitHub for replicability and further research purposes. SINAICA Mexican Government data will be published in this repository. Due to the rights of OpenWeatherMap, we cannot publish them, but they are easily afforded in their website [18].

4 Methods

The main purpose of this research focuses on Deep Learning techniques using neural networks machine learning. We tested several architectures, hyperparameters and different types of neurons to minimize the mean square error (MSE).

Procedure and General Workflow. Our research used Google coLaboratory [20] as our interactive experiment environment using ADAM as our optimizer, though we used Stochastic Gradient Descent on some experiments. We minimized the MSE as our loss function. All our experiments were performed with Tensorflow in Python [22].

For scalability purposes we used the novel platform Google Vertex AI [21] platform, as it offered more performance and automated hyperparameter tuning successfully using a Python script. This code and the notebook experiments code will be published on GitHub.

Experimental protocol We compare different types of artificial neurons:

- **DNN**: Dense neural networks (Multilayer perceptron, MLP) are deep neural networks that are the foundation for artificial neural networks (ANNs) with multiple layers between input and output layers.
- **RNN**: Recurrent neural networks are the most classical architecture used for time series prediction problems.
- **CNN**: Convolutional Neural Network is very popular in image processing applying 2-D convolutions. However, it is also useful for one-dimensional data using 1-D convolutions.
- **LSTM**: "Long-Short-Term Memory" (LSTM) neural networks that are an evolution of RNNs developed to overcome the disappearing gradient problem;
- A mix of the best models, in this case, **CNN + LSTM**.

5 Results.

Our best results were consistently as shown in Figure 1:

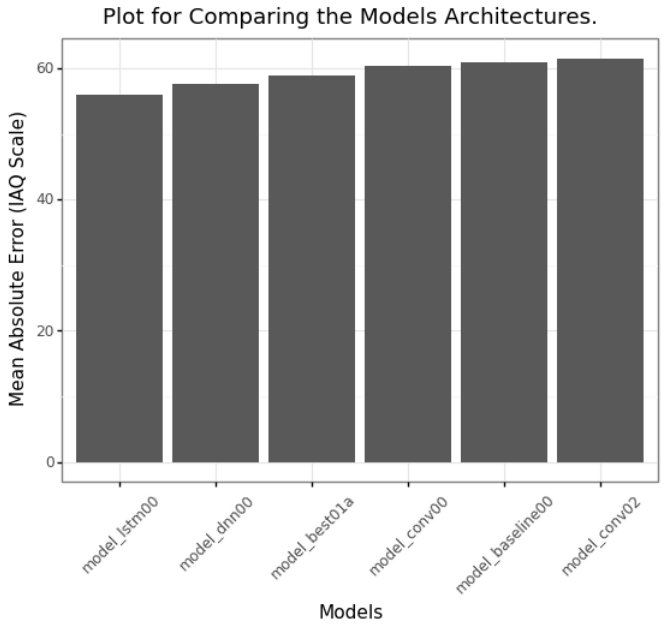

Fig. 1. We tested 10 different neural network types, including: Dense (MLP), LSTM, Convolutional, Recursive networks and combinations of them.

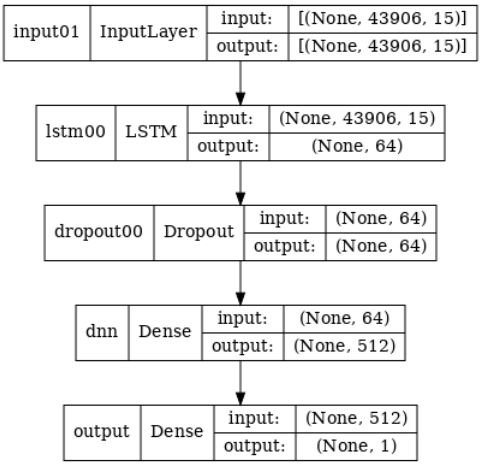

Fig. 2. lstm00 model is a 3 layer combining LSTM, dropout and dense layers in the following fashion. LSTM layer was implemented as 64 LSTM neurons, a second dropout layer with a rate of 0.5 to avoid overfitting and lastly a 512 MLP RELU activated layer. This model has a total of 576 neurons.

Model	Time	Epochs	W. Sz.	Stride	Samp. Rt.	MSE	MAE	resample
lstm00	47.62	11	30	2	2	0.0249	55.97	10 Min
dnn00	29.58	14	15	1	2	0.0309	57.58	10 Min
best01a	890.75	28	15	1	2	0.0244	59.03	05 Min
conv00	79.88	14	15	1	2	0.0301	60.34	10 Min
baseline00	24.97	14	15	1	2	0.0325	60.90	10 Min
conv02	799.29	28	15	1	2	0.0207	61.58	05 Min

Table 1. Comparison of different architectures. Stride and for Sample Rate were handled as hyperparameters of Tensorflow time-series window generator [13,14]. W. Sz. means Window size in days.

Performance of the Models. These architectures were tuned with different time-series hyperparameters to check for consistency of performance on our architectures with our data and to optimize our computational budget. We found that the re sampling size of our dataset, i.e.: instead of processing all our sensor records every 3 seconds it is useful to get the means of these records every 1, 2 and 5 minutes.

Our budget did not allow us to fully test the performance of the 1 minute re sampling data, as it was very expensive. And the 2 and 5 minutes resamples are comparable performance-wise. We tested an even larger resample of 10 minutes with great results.

6 Conclusions.

This paper faced the problem of not having accurate and up-to-date information of air quality while living in a city with known pollution problems. This lack of information led us to ask ourselves what could we contribute as Data Scientists, so we have developed some air quality prediction models based on neural network systems. We proposed a variety of methods with both the scopes of concerned: to perform multi-layer perceptron models for predicting IAQ concentration in an indoor sensor, and to compare the performances between different time-windows and a baseline model.

Data for model classification training was collected using an IAQ Bosch monitoring system. In addition to information from our sensor, we collated data with information from external sources that included particulates such as carbon dioxide, carbon monoxide, ozone, nitrogen dioxide, oxygen, volatile organic compounds, and particulates, temperature, and humidity. Based on the results of the network models, the LSTM model was the best with a MAE of 55.97, nevertheless, all our models presented a similar performance.

In general, it can be concluded that the system delivered a high classification rate based on LSTM. We attribute this success to the appropriate use of time windows and the advantage that Vertex gave us to be able to perform tests with large amounts of data.

The developed models can help environmental agencies and governments monitor air pollution levels more efficiently. Moreover, the model can help to correct missing information in order to protect the health of the citizens who are inhabiting in metropolitan areas.

With high hopes, in the future we would like to work, at least, with data from a full year, to deal with seasonality and its effects on the measurement. We will probably extend this work to a project where the effects of air quality in a pandemic environment are considered.

7 Acknowledgments

We thank the Asociación Mexicana de Cultura A.C., for its support. We also thank Montserrat Altamirano-Astorga for her valuable help proofreading this manuscript.

References

1. Abdullah, A., Raja S., Thulasyammal R., Mohsen M, Ibrahim A.: An Optimized Artificial Neural Network Model Using Genetic Algorithm for Prediction of Traffic Emission Concentrations. International Journal of Advanced Computer Science and Applications, vol. 12. The Science and Information Organization. https://doi.org/10.14569/IJACSA.2021.0120693. (2021)
2. Cakir, S., Moro S.: Evaluating the Performance of Ann in Predicting the Concentrations of Ambient Air Pollutants in Nicosia. Atmospheric Pollution Research, Vol. 11, Issue 12, pp. 2327-2334. https://doi.org/10.1016/j.apr.2020.06.011. (2020)
3. Ghazali S., Lokman H.: Air Quality Prediction Using Artificial Neural Network. Universiti Tun Hussein Onn Malaysia. The International Conference on Civil and Environmental Engineering Sustainability. http://eprints.uthm.edu.my/2528/1/Air_Quality_Prediction_Using_Artificial_Neural_Network.pdf. (2012)
4. Patni J., Sharma H.: Air Quality Prediction using Artificial Neural Networks. International Conference on Automation, Computational and Technology Management. https://doi.org/10.1109/icactm.2019.8776774. (2019)
5. Saad S., Andrew A, Shakaff A., Saad A., Yuzof A., Zakaria A.: Classifying Sources Influencing Indoor Air Quality (IAQ) Using Artificial Neural Network (ANN). Sensors. **15**(5). (2015)
6. Bekkar, A., Hssina, B., Douzi, S. et al. Air-pollution prediction in smart city, deep learning approach. Journal of Big Data **8**(161). (2021)
7. Sotomayor-Olmedo, A., Aceves-Fernández, M., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arreguín, J., and Vargas-Soto, J.: Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach, Int. J. Intell. Science, **3**(3), pp. 126-135. (2013)
8. Ramos-Ibarra, E., Silva, E.: Trend estimation and forecasting of atmospheric pollutants in the Mexico City Metropolitan Area through a non-parametric perspective, Atmósfera: **33**(4), p.401-420. https://doi.org/10.20937/ATM.52757. (2020)
9. Bing, G., Ordieres–Meré, J., and Cabrera, C.: Prediction models for ozone in metropolitan area of Mexico City based on artificial intelligence techniques. International Journal of Information and Decision Sciences, **7**(2), 115-139. (2015)
10. S. K. Singh, R. Yang, A. Behjat, R. Rai, S. Chowdhury and I. Matei, PI-LSTM: Physics-Infused Long Short-Term Memory Network. 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019, pp. 34-41. (2019)
11. Kim, Y., Knowles, S., Manley, J., Radoias, V. Long-run health consequences of air pollution: Evidence from Indonesia's forest fires of 1997. Economics & Human Biology **26**, pp. 186-198, https://doi.org/10.1016/j.ehb.2017.03.006. (2017)
12. NORMA Oficial Mexicana NOM-156-SEMARNAT-2012, Establecimiento y operación de sistemas de monitoreo de la calidad del aire. §10.4.2, pp 8, 14. (2012)
13. Timeseries forecasting for weather prediction, https://keras.io/examples/timeseries/timeseries_weather_forecasting/.
14. Tensorflow: Tutorial on Time series forecasting Time series forecasting, https://www.tensorflow.org/tutorials/structured_data/time_series.
15. Bosch BME680 Datasheet, https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme680-ds001.pdf.
16. Mancuso, D. Indoor Air Quality Monitor | Hackster.io, https://www.hackster.io/damancuso/indoor-air-quality-monitor-b181e9. Last accessed 3 Oct 2021.
17. Dirección de Monitoreo Atmosférico de la Secretaría del Medio Ambiente del Gobierno de la Ciudad de México, http://www.aire.cdmx.gob.mx/.
18. OpenWeatherMap: History weather bulk for Camarones (19.48,-99.18) from January 01, 1979 to September 27, 2021.
19. Sistema Nacional de Información de la Calidad del Aire del Gobierno Federal México] https://sinaica.inecc.gob.mx/. Last accessed 3 Oct 2021.
20. Google AI Blog: Doing Data Science with coLaboratory. https://ai.googleblog.com/2014/08/doing-data-science-with-colaboratory.html.

21. Google Cloud launches Vertex AI, unified platform for MLOps | Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-launches-vertex-ai-unified-platform-for-mlops.
22. Abadi, M., Agarwal, A., Barham, P., et al: TensorFlow: Large-scale machine learning on heterogeneous systems. https://arxiv.org/abs/1603.04467. (2015)