Big Data AI System for Air Quality Prediction

Roba Zayed[1,*], Maysam Abbod[1]

[1] *Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, UK*

***Abstract*—Air Quality has been a research field for many investigators from varied disciplines in respect to global warming, climate change, health effect theories and others. Predicting air quality status is becoming more complex with time due to different gases and other components. This paper aims at presenting machine learning models and techniques to predict air quality levels in cities and providing accurate measures to support data driven decision making in various sectors aligned with sustainable development, economic growth and social values. It supports air quality policies formulation with a future vision to eliminate global related consequences, save the world from the pollution and to close the gap in air quality index standardization, with an emphasis on cities sustainable development.**

***Keywords*— Big Data, AI, Air Quality, Prediction.**

I. INTRODUCTION

Traditional methods to predict air quality suffer from disadvantages such as their limited accuracy (unable to predict extreme points), cut offs cannot be determined, inefficient approaches for better output prediction and equal treatment for old and new data. The uses of big data and machine learning have been proposed as advancement to the traditional methods. Big data and machine learning approaches have been widely used in air quality prediction. There are several researches on air quality evaluation using machine learning algorithms with variation in ML models to predict air quality [1]. Big data has formed a way to model more dynamic air quality systems which are behaviorally heterogeneous; such models take data from various resources. Air quality prediction helps in various ways, directly impacting the environment. However, it is still complex due to the processes and the strong coupling across many parameters, which affect the modelling process. Many techniques were adopted for air pollution forecasting. Recently, techniques such as ANN, Fuzzy Logic and Genetic Algorithms were used in air pollution modelling [2].

Recently, several machine learning approaches have been used by experts, researchers and others, with different parameters combined for air quality prediction. Consequently, it is difficult to understand the reasons for which algorithms are being selected to solve the different ranges of world challenges. This is further made difficult due to the growing number of studies. The aim of this research is to conduct a literature and approach several algorithms and their performance in determining the air quality domain, taking into consideration multiple factors for comparison [1].

II. LITERATURE REVIEW

The two approaches used for air pollution modelling are chemistry dispersion (chemically inert species) and machine learning. Unlike other models, statistical techniques do not take physical and chemical aspects into consideration. Moreover, they use historical data by training the model and then predicting air pollution concentration according to prediction features such as meteorology, land use, time, planetary boundary layer, elevation, human activity, pollutant covariates, and so on. The relationship between air pollution and other features/factors—which can be classified as a complex system of air pollutant and other factors—are highly non-linear; therefore, the simplest statistical approaches— Regression and Autoregressive Integrated Moving Average (ARIMA) models—would not support the complexity enough. Generally, more advanced statistical machine learning as Support Vector Machines, Artificial Neural Networks, and Ensemble Learning have a higher predictive performance than other traditional approaches. Meteorological conditions/ variables such as wind speed, relative humidity and temperature have significant impact on the levels of air pollution. Also, Zhang and Ding experiments have concluded that there is a close relationship between the concentration of air pollutants and meteorological variables and pollution.

There are two methods to predict air pollution concentration: deterministic and stochastic. The deterministic method models the relationship between the physical and chemical transportation process of air pollutants, in terms of the influences of meteorological variables with mathematical models to predict the level of air pollution. The statistical approach learns from historical data and predicts the future accordingly. Researchers suggest using time series to predict the relation between metrological variables and air pollution without the necessity of presenting/modelling the physical relation using methods as time series analysis, Bayesian filter and artificial neural networks. Statistical techniques do not consider chemical and physical processes and use historical data instead to predict the future concentration of air pollution [3].

A. *Air Pollution*

Air pollution is a cause for people to band together to solve and prevent the detrimental on human health, contribution to the depletion of the Ozone layer, acid rain, photochemical smog. Moreover, it is harmful as global population increases, which means growth of automobile use and industrial emissions. Keeping in mind the short- and long-term effects, developed and developing countries are

Manuscript received July 21, 2021; accepted December 01, 2021. *Corresponding author: roba.zayed@brunel.ac.uk

directing their environmental efforts for monitoring air quality. However, developing countries are still focusing on economic growth and there are still more areas to discover in improving air quality.

Based on several air quality evaluations and after surveying the literature in order to improve air quality is on the following high-impact gases: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone (O3), Particulate matter (PM) and Sulphur Dioxide (SO2). It has been discussed in the existing literature that there can be potential reduction in Carbon Dioxide, Sulphur Dioxide, Nitrogen Oxide and Carbon Monoxide emissions by changing energy consumption trends.

B. Air Quality Standards

Air pollution and climate change are controlled by several standards, agreements and measures between mandatory, voluntary and integrated initiatives. The Paris agreement was signed on Earth Day, 2016 at United Nations headquarters in New York. Its main objective is to keep the global temperature rise below 2 degrees Celsius and to limit the temperature increase even further to 1.5 degrees Celsius. Additionally, the IPCC has been formed by World Meteorological Organization (WMO) and United Nations Environment as an international body to provide scientific assessment measures on climate change, its implications and potential future risks.

In 2013, the IPCC provided more clarity about the role of human activities in climate change though its fifth assessment report. IPCC greenhouse gases guidelines have a detailed method for estimating GHG emissions by source and removals by sinks. Further, the Kyoto Protocol is an international agreement aimed at reducing CO2 emissions and greenhouse gases (GHG) in the atmosphere; it was linked to the United Nations Framework Convention on Climate Change (UNFCCC) and adopted by Japan in 1997. As a result of excessive human activity and since the presentation of Sustainable Development Goals (SDGs), there were bodies such as the UN whose aim was to reduce social, economic, environmental imbalance at several scales "Our Common Future". The concentrations of air pollutants recorded over a given time period and considering the effect of each pollutant on health and environment forms what is called "Air Quality Standards". The primary source of standards, criteria and policies is at the local level of central organization that monitors and controls air quality system and resources.

The U.S. National Ambient Air Quality Standards (NAAQS) are limits on atmospheric pollutions concentrations that impact health and environment by the United States Environmental Protection Agency (EPA) under authority of the Clean Air Act (42 U.S.C. 7401 et seq.), It consist of six criteria pollutants: ozone (O3), atmospheric particulate matter, lead, Carbon Monoxide (CO), Sulphur Oxides (SOx) and Nitrogen Oxides (NOx). They are emitted from industry, mining, transportation, electricity generation and agriculture. However, combustion of fossil fuels or industrial processes is main contributors.

To control air quality levels, several guidelines and measures were implemented such as guideline values considered by WHO, the EU labels that limit values for air quality (LVAQ), the National Ambient Air Quality Standards (NAAQS). There are criteria air pollutants that are common throughout the United States which can harm the environment and health:

- Carbon Monoxide (CO)
- Lead (Pb)
- Nitrogen Dioxide (NO2)
- Ozone (O3)
- Particulate Matter (PM)
- Sulphur Dioxide (SO2)

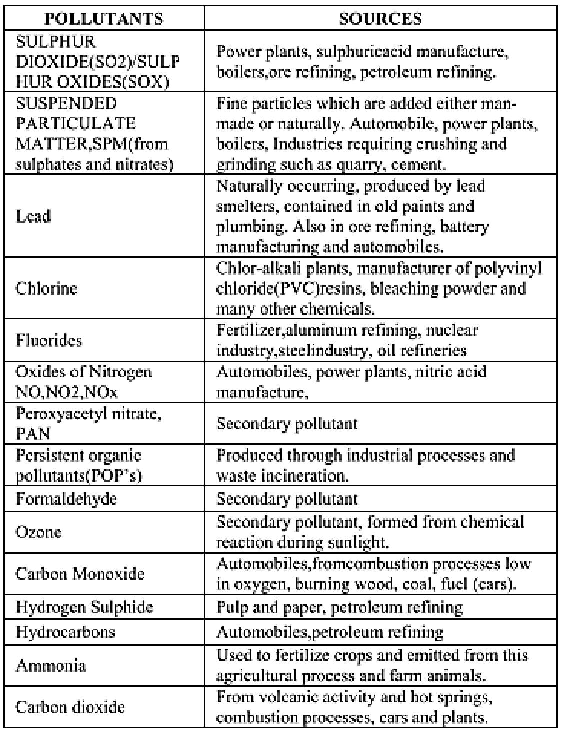
The outdoor air pollution is created mainly by automobiles and various industries which are responsible for the climatic change [4]; there are two categories of air pollution:

Primary Pollutants: results from combustion of fuel and industrial operations.

Secondary Pollutants: results from the reaction of primary pollutants.

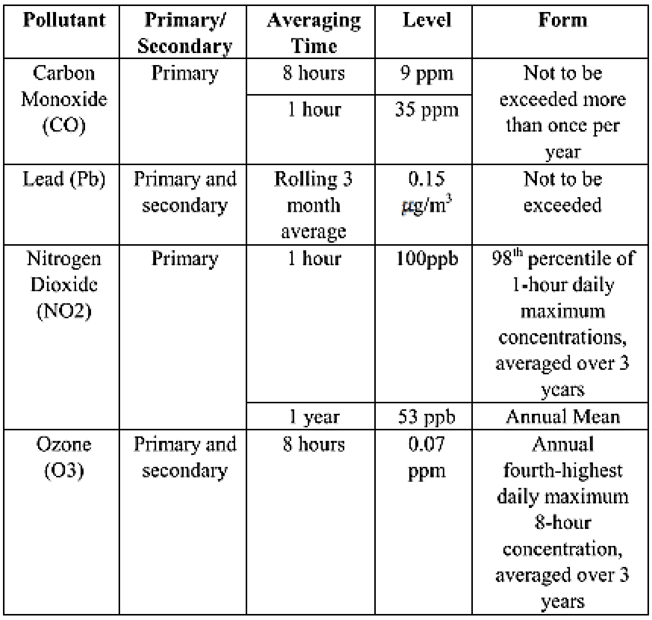
The air quality system (AQS) is used in assessing air quality contains ambient air pollution data by EPA (U.S. Environmental Protection Agency), state, local, meteorological data, descriptive information about each monitoring station (including its geographic location and its operator), data quality assurance/quality control information and tribal air pollution control agencies from thousands of monitors. The National Ambient Air Quality Standard (NAAQS) contains two types of standards—primary and secondary standards. As different pollutions have different effect, the standards are built to accommodate the protection against it (short and long term). Air Quality Index (AQI) quantifies air quality in a region, and it is used in government agencies to communicate the air pollution status [5]. Air pollutants and their sources, NAAQS criteria for pollutants and standards and AQI classification are denoted in Table I, Table II and Table III, respectively.

TABLE I. AIR POLLUTANTS AND THEIR SOURCES [4]



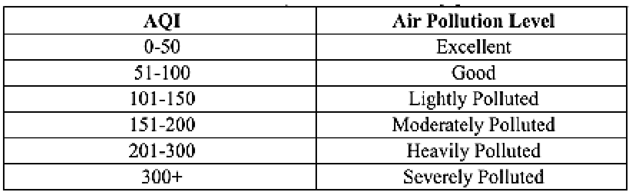
POLLUTANTS	SOURCES
SULPHUR DIOXIDE(SO2)/SULPHUR OXIDES(SOX)	Power plants, sulphuricacid manufacture, boilers,ore refining, petroleum refining.
SUSPENDED PARTICULATE MATTER,SPM(from sulphates and nitrates)	Fine particles which are added either man-made or naturally. Automobile, power plants, boilers, Industries requiring crushing and grinding such as quarry, cement.
Lead	Naturally occurring, produced by lead smelters, contained in old paints and plumbing. Also in ore refining, battery manufacturing and automobiles.
Chlorine	Chlor-alkali plants, manufacturer of polyvinyl chloride(PVC)resins, bleaching powder and many other chemicals.
Fluorides	Fertilizer,aluminum refining, nuclear industry,steelindustry, oil refineries
Oxides of Nitrogen NO,NO2,NOx	Automobiles, power plants, nitric acid manufacture,
Peroxyacetyl nitrate, PAN	Secondary pollutant
Persistent organic pollutants(POP's)	Produced through industrial processes and waste incineration.
Formaldehyde	Secondary pollutant
Ozone	Secondary pollutant, formed from chemical reaction during sunlight.
Carbon Monoxide	Automobiles,fromcombustion processes low in oxygen, burning wood, coal, fuel (cars).
Hydrogen Sulphide	Pulp and paper, petroleum refining
Hydrocarbons	Automobiles,petroleum refining
Ammonia	Used to fertilize crops and emitted from this agricultural process and farm animals.
Carbon dioxide	From volcanic activity and hot springs, combustion processes, cars and plants.

TABLE II. NAAQS CRITERIA FOR POLLUTANTS AND STANDARDS [5]



Pollutant	Primary/ Secondary	Averaging Time	Level	Form
Carbon Monoxide (CO)	Primary	8 hours	9 ppm	Not to be exceeded more than once per year
		1 hour	35 ppm	
Lead (Pb)	Primary and secondary	Rolling 3 month average	0.15 µg/m^3	Not to be exceeded
Nitrogen Dioxide (NO2)	Primary	1 hour	100ppb	98th percentile of 1-hour daily maximum concentrations, averaged over 3 years
		1 year	53 ppb	Annual Mean
Ozone (O3)	Primary and secondary	8 hours	0.07 ppm	Annual fourth-highest daily maximum 8-hour concentration, averaged over 3 years

TABLE III. AQI CLASSIFICATION [5]



AQI	Air Pollution Level
0-50	Excellent
51-100	Good
101-150	Lightly Polluted
151-200	Moderately Polluted
201-300	Heavily Polluted
300+	Severely Polluted

C. *Air Pollution Due to Transportation Sector*

The Transport Sector's Contribution to Greenhouse Gas Emissions and energy consumption has been a major part of the overall assessment of greenhouse gas emissions indicators for countries. Many forces such as economic, social, education, leisure and others have been major players in the need for people to move from place to another, with technological advancement in transportation, In order to present a more connected world. The availability of various forms of transportation left people with many options and to some extent, presented unpredictability and uncertainly for traveler patterns. With the availability of various means of travel, the demand has increased, and the trend has become as one of the main sectors for measuring the sustainability of cities and quality of life. The travel networks have been growing in complexity and so has traveling data. In many reports, road transport was the dominant mode of travelling and it is the main cause of pollution. Moreover, it is considered as one of the sectors with growing emissions. Road travel and aviation are considered the principal contributors to greenhouse gas emissions from the transport sector perspective [6]. Many developing and developed countries do not meet air quality standards for NO2 near near road transport and the decline in near road NO2 is much less than expected, which can be as a result of more usage of diesel vehicles.

The transportation sector plays a major part in air quality indicators worldwide and accordingly affects the overall global air quality index as well as the local index for each country. Therefore, there are regional and worldwide efforts to study the bottom line of the transportation layer and try to come up with solutions to improve air quality by reducing traffic.

D. *Machine Learning Methods*

In fact, a systematic review by Rybarczyk, Y. and Zalakeviciute is providing a systematic comparison between algorithms using different parameters. In [1] systematic review, recent machine learning studies (journal articles) in pollution research were selected and scanned, concluding with the use of machine learning algorithms for predicting air quality. The review presents the most used algorithms in descending order: Ensemble Learning Methods, Artificial Neural Networks, Support Vector Machines and Linear Regressions. The advantage of using the Ensemble Learning Method (ELM) is that it provides better accuracy.

As emphasized in the literature [7]. ANN models have been widely used for air pollution concentration predictions. The findings of several papers confirm the superior performance of ANNs in comparison to traditional statistical methods—multiple regression, classification and regression trees and autoregressive models. Artificial Neural Networks (ANN) models have shown better performance than Multiple Regression Models (MLR), incorporating complex nonlinear relationships between the concentration of air pollutants and the corresponding meteorological variables and are widely used for the prediction of air pollution concentration. However, ANN has a few drawbacks—potentially falling into the trap of local minimum and poor generalization, lack of analytical model selection approach, time consuming (in finding the best architecture) and its weight by trial and error. ANN models have the ability to capture the highly non-linear character of those processes serving a wide range of gaseous prediction.

Based on several air quality evaluations and after surveying the existing literature, the focus on improving air quality by reducing the emission of these high-impact gases: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone (O3), Particulate matter (PM) and Sulphur Dioxide (SO2). It has been discussed in the literature that there can be potential reduction in carbon dioxide, sulphur dioxide, nitrogen oxide and carbon monoxide emissions by changing energy consumption trends. However, Asia is facing issues by exceeding PM emission. Ground level Ozone demonstrates average values but exceeds the limits values in all analysis by several factors. In low-income countries and the process of their development, there is a high possibility in the increase in air pollution.

Furthermore, there are significant variations in air pollutants, causing air changes over location and time. A generic prediction of the overall air quality in a city is not that useful for decision making. Moreover, there are some sudden changes which can be caused by unusual weather conditions (inflection points). To tackle such challenges and shortages that can present in a general statistics model, researchers proposed models to stress the need to consider hybrid models to predict air quality and cover some gaps and shortages by some modelling methods [8].

III. DATA COLLECTION

Air quality data was accessed from London Air quality data repositories-open data (see links section) as the primary source of data. The repositories include data (1993–2019). It includes several sites in London and there is a considerable data that can be used, with up to six species of gases selection. Several air quality open data sources have been searched and the selection was based on complete data bases for at least 3 years with concentrations (CO, NO, NO2, NOX, O3, PM10, SO2) and meteorological data, such as wind speed, humidity and temperature.

TABLE IV. DATA COLLECTION SPECIFICATION

	London	Jordan
Air Quality Station (Data Capturing)	Marylebone Road	GAM
Years	2014-2018	2016-2018
Time span	Hourly Data (every hour)	Hourly Data (every hour)

The secondary source of data was the Jordanian Ministry of Environment for traffic locations, air quality hourly data from 2016 till 2018, including meteorological related such as (temperature, wind speed, wind direction, humidity); GAM (Greater Amman Municipality) location with concentration for (PM10, NO2, CO, SO2). Further meteorological data from 2016 to 2018 was collected from the Jordan Meteorological Department to validate humidity values as it showed some odd patterns in the original file retrieved from Jordanian Ministry of Environment after consultation with experts in the field from Department of Statistics in Jordan (dos.gov.jo).

The retrieved data required pre-processing for missing values and normalization of selected columns, such as humidity. London city airport data was used to fill in some parts to complete missing meteorological data. Filling data in Marylebone road-London was based on the most complete dataset from nearest location as London city airport is the nearest location with the most complete dataset. Only empty rows were filled by city airport data as it should represent meteorological data for the studied area. Further, other empty rows from gases were filled with previous values within the same column, as it is assumed this is the nearest reading for the next hour missing data. In addition, normalization was applied to data.

London-UK Data [10]:

Input: Day, Month, Year, Hour, Humidity, Temperature, Wind Speed, Wind Direction

Output: CO, NO, NO2, NOx, O3, PM10, SO2

Size: 43824 data point

Amman-Jordan Data [9]:

Input: Day, Month, Year, Hour, Humidity, Temperature, Wind Speed, Wind Direction

Output: PM10, NO2, CO, SO2

Size: 26268 data point

IV. Modelling And Results

The purpose of this research is to build a prediction model (next hour forecasting) and define measurable (quantifiable) data and compare different models for air quality prediction recommendation. Several machine learning regression methods have been used to compare results and model results using MATLAB R2020a software.

- Neural Network (feed forward backdrop).
- Neural Network Fitting.
- Neural Network Time Series (NARX).

NARX (dynamic neural network) model has proved better results in tested cases and scenarios for the same data set (as seen in Fig. 1, Fig. 2 and Fig. 3) as of the structure nature of this algorithm in supporting time series data All results/performance are reported based on experiments using MATLAB R2020a. In the second phase, DNN (Deep Neural Network) models were applied to the same datasets.

In Fig 1, Fig 2 and Fig. 3, results, results/errors and error histograms are denoted for Westminster–Marylebone Road (central London) NARX, respectively.

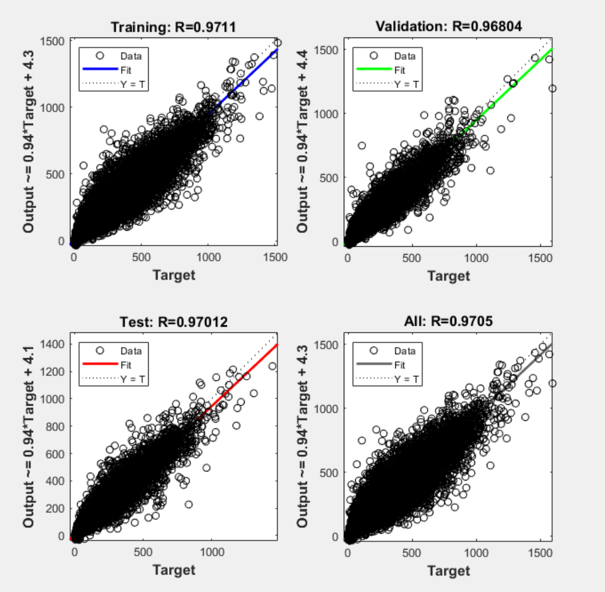
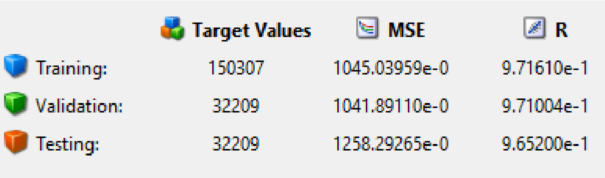


Fig. 1. Westminster-Marylebone Road results (central London)-NARX



	Target Values	MSE	R
Training:	150307	1045.03959e-0	9.71610e-1
Validation:	32209	1041.89110e-0	9.71004e-1
Testing:	32209	1258.29265e-0	9.65200e-1

Fig. 2. Westminster – Marylebone Road (central London)-NARX results/errors

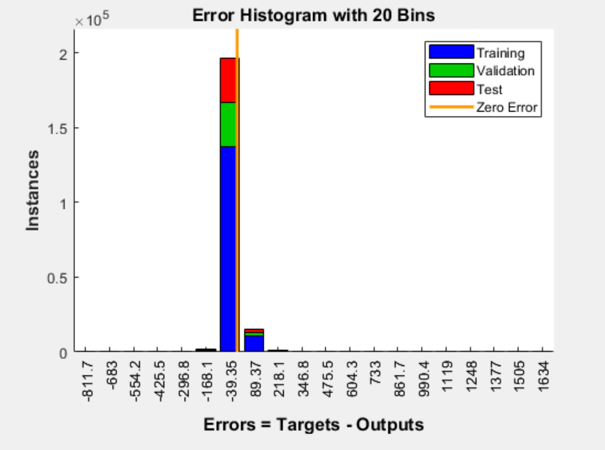


Fig. 3. Westminster–Marylebone Road (central London)-NARX error histogram

In Fig 4, Fig 5 and Fig. 6, results, results/errors and error histograms are denoted for (Jordan)-NARX, respectively.

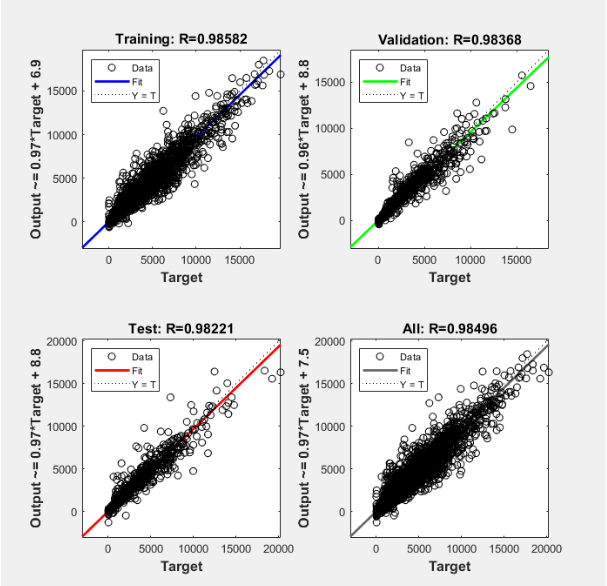
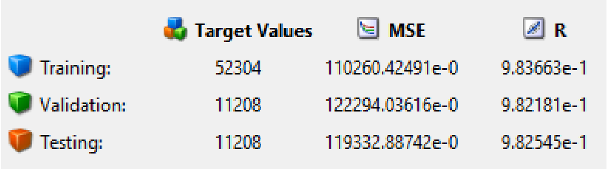


Figure 4. GAM-location results (Jordan)-NARX



	Target Values	MSE	R
Training:	52304	110260.42491e-0	9.83663e-1
Validation:	11208	122294.03616e-0	9.82181e-1
Testing:	11208	119332.88742e-0	9.82545e-1

Figure 5. GAM-location (Jordan)-NARX results/errors

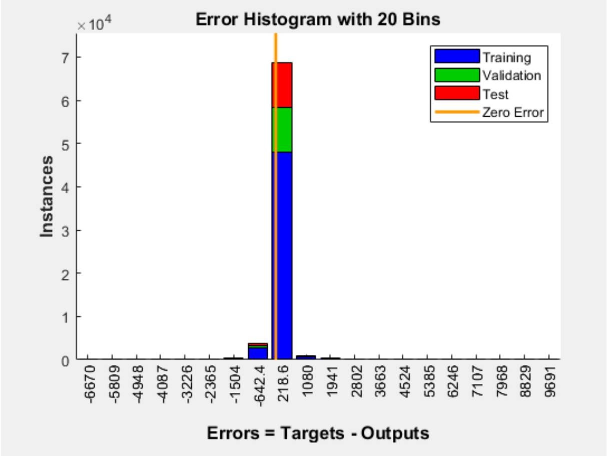


Figure 6. GAM-location results (Jordan)-NARX error histogram

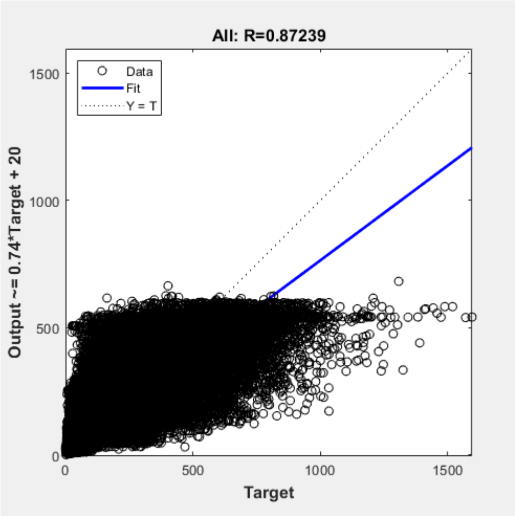


Figure 7. Westminster-Marylebone Road location (central London)–DNN results

In Fig. 7, DNN results are shown for Westminster-Marylebone Road location (central London). In addition, in Fig. 8, DNN results are shown for GAM-location (Jordan)

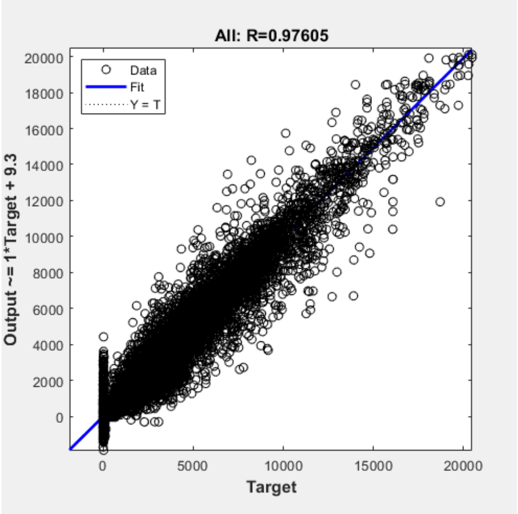


Figure 8. GAM-location (Jordan)-DNN results

A DNN was used to model both the English and Jordanian data, as seen in Table IV. There were some variations between the parameter values in both models, depending on the needs, data size and parameters (check data details in the Data section). Two LSTM layers were used for each model: a sequence-input layer, two dropout layers (0.3), a fully connected layer and a regression layer. Training options for the models were: the execution environment (CPU), L2 Regularization, Mini Batch Size and Max Epochs. The details of the structure of the DNN models can be found in Table V. The parameter values were specified following scientific papers [9] and based on systematic experiments. Parameters were optimized in iterative reviews through several trials (see Table IV for more details of the DNN models structure).

TABLE IV. STRUCTURE OF DNN MODELS

Model type	**LSTM Layers**	**L2 Regularization**	**Mini batch size**	**Epoch**
DNN-Jordan Data	2	1E-10	1024	1500
DNN-England Data	2	1E-10	900	1000

TABLE V. RESULTS OF DNN MODELS

Model type	**Location**	**Accuracy**	**No. hidden units**	**Training Fun**
NN -FFB	Jordan	0.94	20	trainlm
NN -FFB	England	0.89	25	trainlm
NN-Fitting	Jordan	0.93	20	trainlm
NN-Fitting	England	0.88	25	trainlm
NN-NARX	Jordan	0.98	20	trainlm
NN-NARX	England	0.97	25	trainlm
DNN	Jordan	0.97	1400	trainNetwork
DNN	England	0.83	800	trainNetwork

The different methods shown in Table V are discussed in Section IV (Modelling and Results). The table shows accuracy, represented by correlation between inputs and outputs using the following methods: Neural Network (feed forward backdrop), Neural Network Fitting, Neural Network

Time series (NARX) and DNN.

V. Discussion

Predicting air quality is challenging because of the complexity of its processes and the strong coupling across all parameters, which can be more complex in one gas than another, such as PM. Data access in some regions can be considered as a limitation in other regions, as some data from monitors is missing. (This means that further methods and validations are required for data replacement and/or removal accuracy). Furthermore, generating accurate results in the light of these data factors is becoming more challenging in dynamic systems.

Adding an L2 Regularization layer to the DNN model and tweaking the mini batch value to a suitable value, depending on the mode (between 32 and 1024), both improve the model results. Epoch value, dependent on the model, also plays a role in enhancing results [11].

VI. Conclusion And Future Work

It should be highlighted that there is very little research that focuses on transport and its impact on air quality, especially in Asia. Processing air-quality data was not easy for several reasons — first of all, data are very limited in developed countries, and only a few countries publish this information [12]. There is limited research which uses hourly air-quality data for predictions and also addresses transport. Furthermore, it has been claimed that there is no published report concerning the refinement of input data for network learning. Few existing studies focus on the identification of effective parameters. This study aims to use prediction, based on statistical methods, to select the best statistical model and refinement method for air pollution and meteorological data, in order to predict and classify air quality. This study is part of wider research into designing and modelling the Air Quality Index, determining the best machine-learning methods for monitoring air quality. After completing the prediction models using DNN, hybrid models can be used to improve accuracy. The models generated are able to deal with missing data problems, complex gas predictions and accuracy issues. In the future, this study recommends building an Air Quality Index model using deep learning and classification. It should be mentioned that other prediction methods are subject to consideration in support of the study directions.

Acknowledgment

Part of this work was presented orally at the IV. International Conference on Data Science and Applications 2021 (ICONDATA'21).

References

[1] Rybarczyk, Y. and Zalakeviciute, R. (2018) 'Machine learning Approaches for outdoor air quality modelling: A systematic review', Applied Sciences. MDPI AG, 8(12), p. 2570. doi: 10.3390/app8122570.

[2] Alkasassbeh, M et al. (2013) 'Prediction of PM10 and TSP Air Pollution Parameters Using Artificial Neural Network Autoregressive, External Input Models: A Case Study in Salt, Jordan', Middle-East Journal of Scientific Research, 14(7), pp. 999–1009. doi: 10.5829/idosi.mejsr.2013.14.7.2171.

[3] Zhang, J., & Ding, W. (2017). Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. International Journal of Environmental Research and Public Health, 14(2), 114. https://doi.org/10.3390/ijerph14020114

[4] Rao, P. (2014) A survey on Air Quality forecasting Techniques. Available at: www.ijcsit.com (Accessed: 1 February 2020).

[5] Gao, J. (2018) 'Air Quality Prediction: Big Data and Machine Learning Approaches'. doi: 10.18178/ijesd.2018.9.1.1066.

[6] Chapman, L. (2007). Transport and climate change: a review. Journal of Transport Geography, 15(5), 354–367. https://doi.org/10.1016/j.jtrangeo.2006.11.008

[7] Grivas, G. and Chaloulakou, A. (2006) 'Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece', Atmospheric Environment. Pergamon, 40(7), pp. 1216–1229. doi: 10.1016/j.atmosenv.2005.10.036.

[8] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting Fine-Grained Air Quality Based on Big Data. https://doi.org/10.1145/2783258.2788573

[9] dos.gov.jo (Accessing date: 21 July 2021)

[10] https://www.londonair.org.uk/LondonAir/Default.aspx

[11] Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., & Wang, Y. S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. Journal of Cleaner Production, 209, 134–145. https://doi.org/10.1016/j.jclepro.2018.10.243

[12] Baldasano, J. M., Valera, E., & Jiménez, P. (2003). Air quality data from large cities. Science of the Total Environment, 307(1–3), 141–165. https://doi.org/10.1016/S0048-9697(02)00537-5.