Hybrid Intelligent Modelling for Air Quality Prediction Deep Learning and Markov Chain Unconventional Framework

Roba Zayed, Maysam Abbod

Email: roba.zayed@brunel.ac.uk; maysam.abbod@brunel.ac.uk

Department of Electronic and Electrical Engineering
Brunel University London, Uxbridge, United Kingdom.

***Abstract* - The purpose of this research is to build an innovative prediction model, define measurable (quantifiable) data and use them to measure air quality in selected cities. This study presents a multivariate hybrid Markov-switching dynamic model using a multi-state transition method for multiple outputs and a Deep Neural Network through a niche experimental framework. The experiment is part of applied big-data AI research which aims to predict air quality and present a reliable system which will provide an air-quality index using hybrid model. This will become a tool for decision-makers concerned with related air-quality issues. This research presents a multi-input multi-output hybrid model with reliable accuracy of hourly time-series data, and provides the large dataset in this study. This aims to cover the gap in high big-data prediction accuracy for the domain (hourly frequency) and to form a more standardized air-quality index by comparing results in two selected cities: London and Jordan.**

Keywords - Markov Chain, DNN, Prediction, AI, Hybrid Modelling, Air Quality

I. INTRODUCTION

While conducting the literature review, a major shift has been noticed from 'climate change' to 'climate crisis', and world leaders have expressed increased fears that global warming will cross the safety threshold of 2º Celsius. A recent newspaper headline illustrates this: "'Untold human suffering': 11,000 scientists from across world unite to declare global climate emergency". This headline was designed to emphasise the level of emergency and danger caused by climate change, as was the comment that "'Despite 40 years of major global negotiations, we conduct business as usual and have failed to address this crisis'" [1]. Most indicators, however, are not very promising for humanity, given the severe increase in global carbon dioxide emissions. It has been claimed that up to a third of the reduction in emissions needed by 2030 to satisfy the Paris Agreement could be achieved by natural actions. Further, reductions in fuel consumption could be implemented using effective policies. In terms of the economy and population, we should work on reducing the impact of population growth on GHG emissions, and also have active regulatory policies that can ensure social integrity and maintain the long-term sustainability of the biosphere [2]. As a result of pressure from human activity since the presentation of the Sustainable Development Goals (SDGs), bodies including the UN have aimed to reduce social, economic, environmental imbalances at several scales. Air quality and climate change influence each other and air pollutants also contribute to atmospheric changes [3].

The review by Rybarczyk and Zalakeviciute (2018) highlighted the increasing trend of using machine-learning approaches to monitor air quality in the period since 2010, while the use of big data and machine learning has been proposed as advances on traditional methods. Big-data and machine-learning approaches have been used widely to predict air quality [4]. Nevertheless, examples of highly accurate air-quality prediction methods for big-data considering temporal resolution are limited in the existing literature [5]. There are several examples of research into air-quality evaluation which use machine-learning algorithms with various ML models to predict air quality. Big data has formed a way to model more dynamic air-quality systems which are behaviorally heterogeneous and take data from various sources. Many algorithms, methods and techniques have been used in air-pollution modeling [6] [7]. It is noted that prediction methods sometimes do not support the aimed accuracy; and there are inefficient approaches to better output prediction. Therefore, the existing literature has suggested using hybrid models to overcome several of these limitations and taking advantage of using different methods with more than one model [8] [9]. There are many ways of combining DNN (LSTM) and Markov models (hybrid models) to improve predictions. Some approaches are presented, such as Markov trained on LSTM states, a hybrid model in which Markov is trained first to predict states which are then passed to LSTM to predict outputs. In another method, a jointly trained hybrid model combines LSTM outputs with Markov states. The aim of utilizing hybrid models is to use the advantages of LSTM but make it more interpretable [10]. The Markov chain is one of the classical statistical (stochastic) models which represent a linear method of data analysis and are used in time-series predictions with high interpretability [11]. The purpose of this study is to present the experimental multivariate Markov switching model as part of research to develop hybrid air-quality prediction model (Markov and DNN) which aims for an appropriate level of accuracy, in order to support leaders' decisions by providing timely air-quality measurements.

II. MARKOV CHAIN THEORY

Following the generalized Hamilton model (1988, 1989), the Markov-switching model is represented by a

general autoregressive component. It is a state-dependent model which has received much attention in dependent data modeling. As an extension of Hamilton's Markov-switching model and others, different approaches are used to satisfy the different capabilities of Markov Chain theory [12].

A Markov system can be described as a set of N states: $S_1,S_2,S_3\ldots\ldots S_N$. A change in state (state transition) according to a set of probabilities (a chance that any state can be reached from any other states) can be expressed in the equation and illustrations below in Eq. 1 and Eq. 2 [13]:

$$P[q_t=S_i|\ q_{t-1}=S_i,q_{t-2}=S_k,\ldots] \quad (1)$$
$$= P[q_t=S_i|\ q_{t-1}=S_i] \quad (2)$$

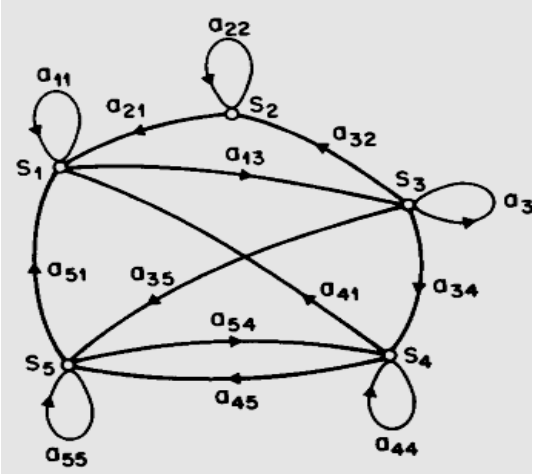

Fig. 1. A Markov chain with 5 states (labeled S, to S,), with selected state transitions [13].

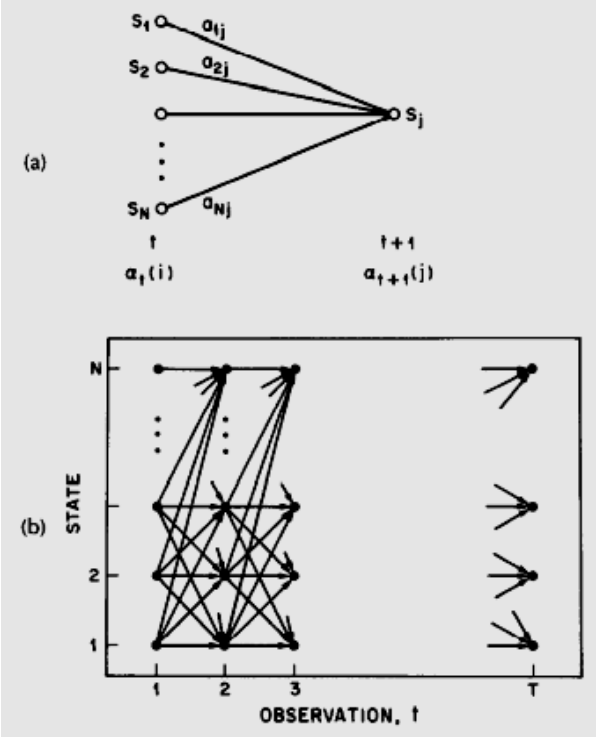

Fig. 2. Illustration of observations t, and states i [13].

III. DATA COLLECTION AND PREPARATION

A. Data Sets

After studying multiple factors impacting air quality predictions, and researching data availability and completeness in this field, the researchers aimed to achieve satisfactory results. Data used in this experiment were collected from two locations: London, UK and Amman, Jordan.

1) First location: Marylebone Road, data between 2014 and 2018 (hourly data)

Data points (data size): 43824
Source:
https://www.londonair.org.uk/LondonAir/Default.aspx

2) Second location: GAM (Greater Amman Municipality), data between 2014 and 2018 (hourly data)
Data points (data size): 26268
Source: collected from Jordanian Ministry of Environment-traffic locations.

B. Data Preprocessing

Parts of the data, in particular the Marylebone Road-London meteorological data, were not complete, so data from the nearest location with complete data-sets were selected to complete empty rows within the same column. Other empty rows from columns concerning gases were completed using the value of the previous row, as the nearest reading for the next hour of missing data. Further, normalization was then applied to data.

IV. METHODOLOGY

The experiment conducted multivariate output prediction for several gases in two cities, as shown below:
London, UK-Output: CO, NO, NO2, NOx, O3, PM10, SO2
Amman, Jordan-Output: PM10, NO2, CO, SO2

In the study, the researchers considered the impact of weather conditions; temperature, humidity, wind direction and wind speeds, which generally promote the rapid movement of pollutants to other places and different distances.

All reported results/performance are based on experiments using MATLAB R2020a.

The authors performed the experiment at three levels, as shown below:

- DNN (Deep Neural Network): a previous study by the authors explained the details of the DNN model using this data-set in the section on data collection and preparation [14].
- Markov Chain: the Markov-switching regression model provides a prediction method using MS-VAR (details in Section II (Markov Chain Theory) and section IVA (Parameters Setup)).

- Hybrid (Markov Chain and DNN): a combination of DNN and Markov is presented in this work following the method below:

DNN and Markov were combined, first using HMM outputs (running the algorithm) and then feeding DNN as output and running DNN with the input data and a Markov output. This is a proposed method of achieving better performance.

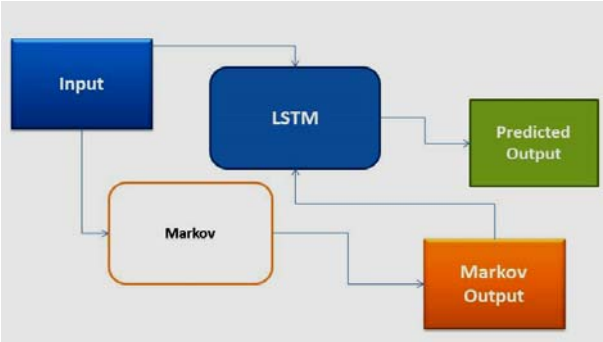

Fig. 3. Proposed Hybrid modelling (Markov Chain and DNN) by authors

First, a DNN trial was performed and then another stand-alone trial of Markov. Then the two models were combined as discussed in Fig. 3.

It is worth mentioning that other methods were used in experiments, in an attempt to increase accuracy. However, the method shown in Fig.3 was selected as the most accurate, compared to the other methods mentioned below. Some of these methods are:

- Method A: Calculation for error was performed for DNN and added to the Markov outputs. Then a Markov run was performed with new outputs. Alternatively, the Markov error was calculated and added to the DNN results. Then a DNN run was performed with new outputs, in an attempt to check for appropriate levels of accuracy.
- Method B: The mean was checked for both results (mean DNN and mean Markov). The mean was used as a hybrid method of predicting outputs using Markov and then predicting outputs using DNN. The mean of both predictions was then taken.
- Method C: Simulated Markov results (input and output) were used as input and output of DNN. A run was then performed using the new simulated states and the predicted output from Markov.
- The LSTM output is predicted using LSTM and then the Markov output is predicted using Markov.

A. Parameters Setup

1) Markov Parameters

The multivariate Markov-switching dynamic system experiment consists of multiple parameters for inputs and outputs. The parameter set-up was focused on the achievement of suitable results and consistency across models. TABLE I and TABLE II show the parameter set-up for inputs and outputs.

TABLE I. MARKOV MODEL PARAMETERS (INPUTS)

Parameters	Models	
	Markov Jordan	***Markov England***
AR	Input1: corr(Input1,Output1) Input2: corr(Input2, Output1) Input3: corr(Input3, Output1) Input4: corr(Input4, Output1)	Input1: corr(Input1,Output1) Input2: corr(Input2, Output1) Input3: corr(Input3, Output1) Input4: corr(Input4, Output1)
Beta	Input1:set to 1 Input2:set to 1 Input3:set to 1 Input4:set to 1	Input1: set to 1 Input2:set to 1 Input3:set to 1 Input4:set to 1
Constant	meanInput1=mean(Input1) meanInput2=mean(Input2) meanInput3=mean(Input3) meanInput4=mean(Input4)	meanInput1=mean(Input1) meanInput2=mean(Input2) meanInput3=mean(Input3) meanInput4=mean(Input4)
Variance	Input1: std(Input1) Input2: std(Input2) Input3: std(Input3) Input4: std(Input4)	Input1: std(Input1) Input2: std(Input2) Input3: std(Input3) Input4: std(Input4)

[a.] AR (auto regression coefficient)
[b.] Beta (regression coefficient)
[c.] Constant (mean)
[d.] Variance (standard deviation)
[e.] corr (correlation)
[f.] std (standard deviation)

Table I presents the set of parameters used to build a Markov-switching dynamic system. The Markov model was built using the switching dynamic regression method; the states were represented by a set of multiple ARIMA (moving average) models and each model presented one of the states (temperature, humidity, wind direction and wind speed). The parameters (a,b,c,d) in Table I (input model) and the inputs were accordingly simulated using MSVAR. The output model consisted of the same parameters as the input model but simulated using the output data.

B. Algorithm Structure

The Markov-switching dynamic regression model consists of four states (humidity, wind speed, wind direction and temperature). Each state was formed using ARIMA (autoregressive integrated moving average) with the parameters in TABLE I. The output model was also formed using ARIMA with the parameters presented in TABLE II.

C. Input and Output Simulation

Input1, Input 2, Input 3 and Input 4 were simulated using the simulation function based on equal probability

for each of the four states (transition probability). This is theoretically represented in :

$$P = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25] \quad (3)$$

TABLE II. MARKOV MODEL PARAMETERS (OUTPUTS)

Parameters	Models	
	Markov Jordan	***Markov England***
AR	Mean(corr(Inputs,Output)	Mean(corr(Inputs,Output)
Beta	Output:set to 1	Output:set to 1
Constant	Mean (Output)	Mean (Output)
Variance	Std (Output)	Std (Output)

[g] AR (auto regression coefficient)
[h] Beta (regression coefficient)
[i] Constant (mean)
[j] Variance (standard deviation)
[k] Inputs (represent all four inputs)
[l] corr (correlation)
[m] std (standard deviation)

Data were simulated using observed outputs based on the transition probability and then random walks were performed on the simulated data to obtain predictions using the simulation function.

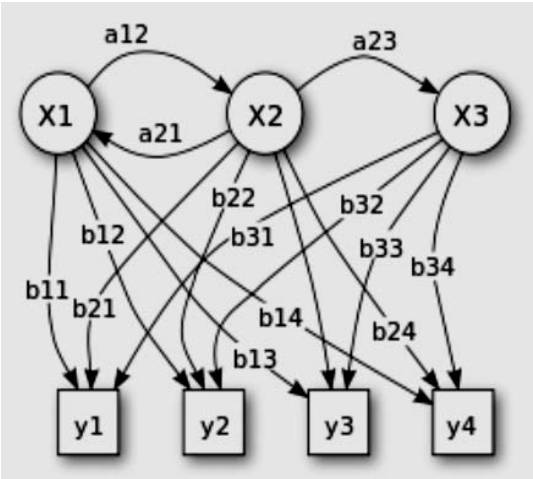

Fig. 4. Illustration of probabilistic parameters of a hidden Markov model, (X) represents states, (y) represents possible observations, (a) represents state transition probabilities, (b) represents output probabilities [15].

V. EXPERIMENT RESULTS AND ANALYSIS

A. Experiment Results

Results are shown in Fig. 4 and Fig. 5 below, which show the correlation between inputs and outputs as a measure of accuracy. It should be noted that both models of the Jordan and England results shown in these figures use multiple inputs (meteorological data) and multiple outputs (gases).

Results show that the Markov model gives an appropriate performance, especially when it is used, in a hybrid model in the current work in progress, to represent a linear method of simulating data.

1) Markov results

Fig. 5 and Fig. 6 present the results of the Markov model, representing the linear part of the hybrid model by simulating the data, as discussed in Section IV (Methodology). This model produced good results as a stand-alone. However, the authors proposed a hybrid model in the following parts of this section, in an attempt to improve accuracy.

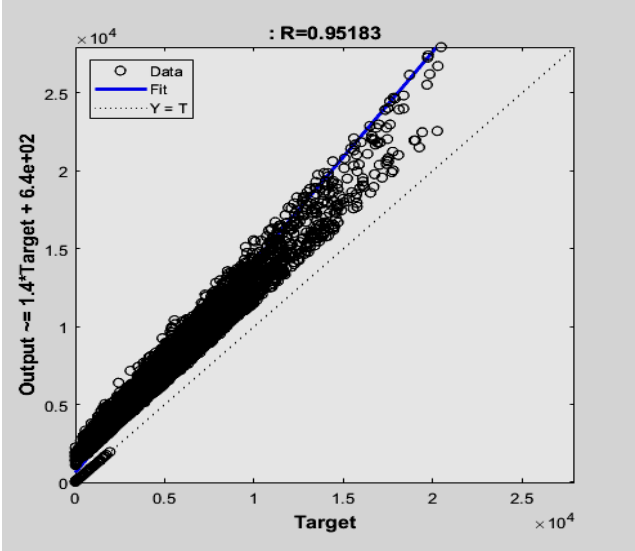

Fig. 5. Markov model results for the Jordan data, with four inputs (wind speed, wind direction, temperature, humidity) and four outputs (PM10, NO2, CO, SO2).

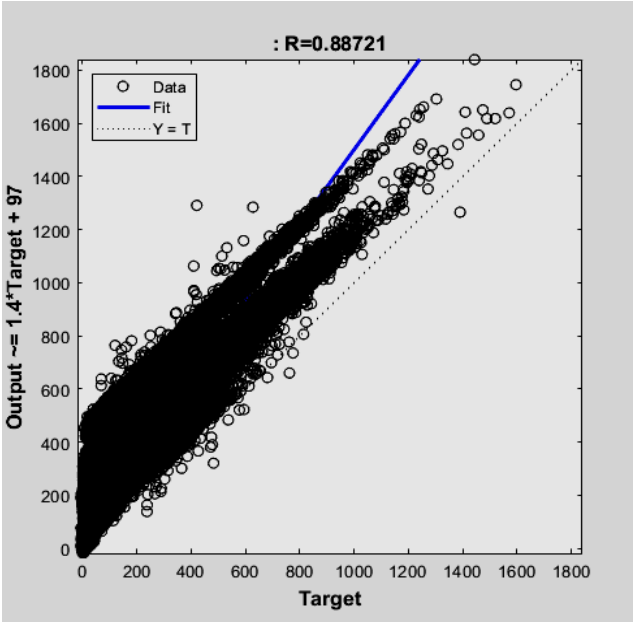

Fig. 6. Markov model results for the England data, with four inputs (wind speed, wind direction, temperature, humidity) and seven outputs (CO, NO, NO2, NOx, O3, PM10, SO2).

2) Hybrid model (Markov and DNN)

Fig. 7 and Fig. 8 show the hybrid model results (Markov and DNN), which proved to be satisfactory results for this study. The combination of models provided a solution to the Markov shortage in big-data prediction, and utilized the advantages of both models to produce better results, satisfying the aim of this study.

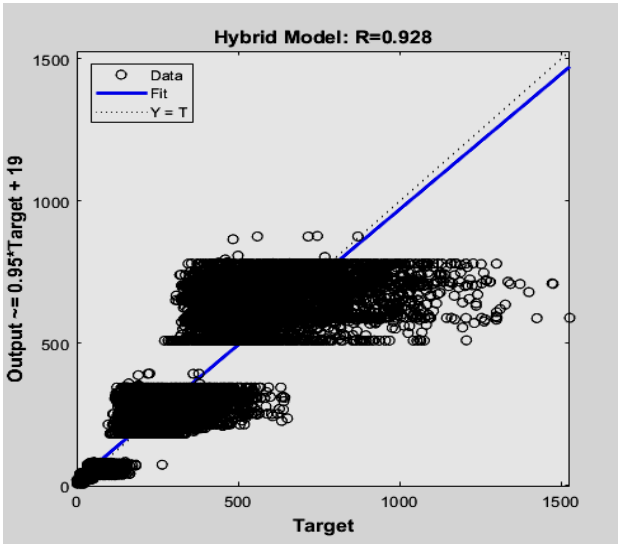

Fig. 7. Hybrid model: Markov Chain and DNN-England

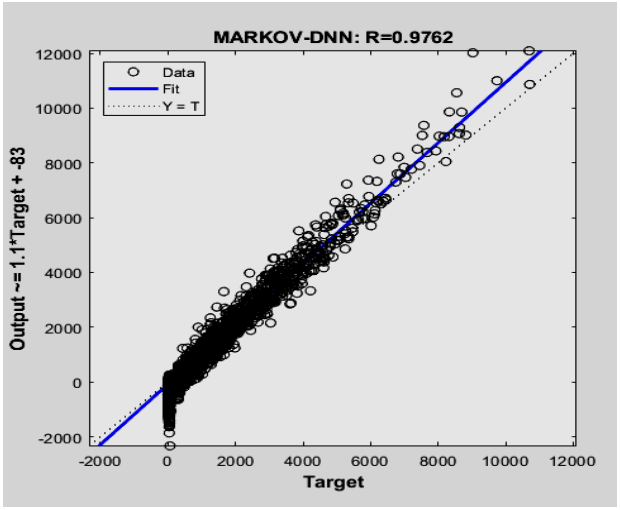

Fig. 8. Hybrid model: Markov Chain and DNN-Jordan

VI. DISCUSSION

Previous work on air-quality prediction using deep learning (DNN) by the authors of this paper [14] is presented here for the purpose of comparison. The DNN models represent non-linear methods of data prediction using MATLAB R2020a software. This is shown in Table III.

TABLE III. DNN MODELLING RESULTS [14]

Model Type	Location	Accuracy
DNN	Jordan	0.97
DNN	England	0.83

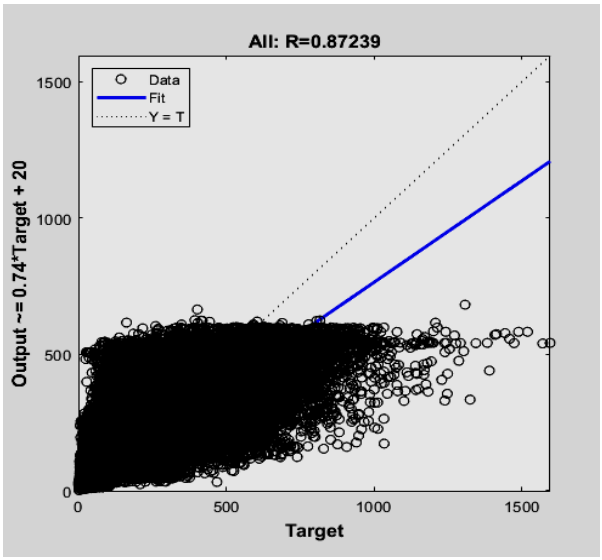

Fig. 9. Westminster-Marylebone Road location (central London): DNN results.

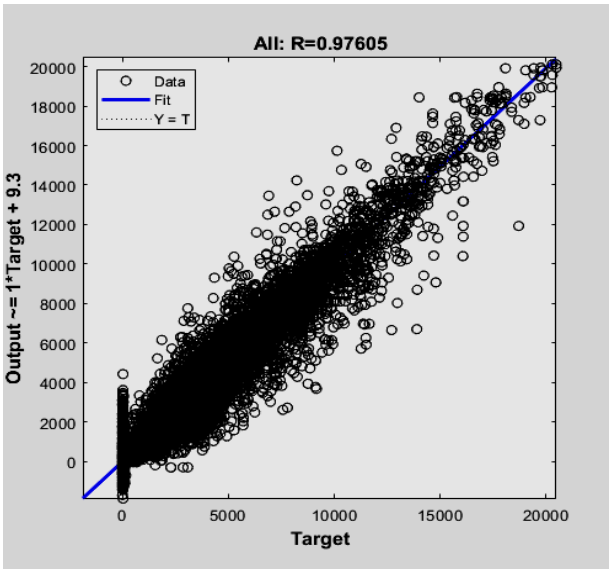

Fig. 10. GAM-location (Jordan): DNN results: Conclusion and Future Direction.

As can be seen from Fig. 7 and Fig. 8, there is an improvement in the results when using hybrid methods (Markov and DNN). This supports the aim of this study by providing the level of accuracy required for air-quality prediction.

TABLE IV. HYBRID MODELLING RESULTS

Hybrid Model	*Input*	*Output*	*Hybrid Method*	*Accuracy*
HMM and DNN-England	Input	Predicted from Markov (simulated output)	Markov (output prediction) feeding to DNN	0.92
HMM and DNN-Jordan	Input	Predicted from Markov (simulated output)	Markov (output prediction) feeding to DNN	0.97

In comparison, the overall performance improved using hybrid modeling. The researchers would recommend experiments of this kind when using big data for prediction, especially when modeling limitations arise.

VII. CONCLUSION AND FUTURE DIRECTIONS

The study presents satisfactory performance of the hybrid Markov and DNN model. Due to the limitations of the Markov Chain in predicting long-term time-series data [11], the direction of this study suggested that the hybrid (Markov-LSTM) model would produce improvements, and the experiment demonstrated this. A forward-looking Air Quality Index (AQI) will be developed further, when appropriate levels of accuracy in reference to air quality prediction have been achieved. Further methods of implementing the Markov and DNN hybrid are being explored to fulfil the aims and objectives of this study, while other models are also being investigated to see if they also improve accuracy. As has been discussed in the experimental summary, some modeling methods outperformed others, especially when Markov and DNN were combined. However, not all combinations of methods will give good results. Further validation of the best performing models will be conducted using case studies (with data from another source) to test the models for further developments.

REFERENCES

[1] Global climate emergency: 11,000 scientists from across world unite to issue unprecedented declaration | The Independent. Available at: https://www.independent.co.uk/environment/climate-emergency-scientists-emissions-letter-climate-change-a9185786.html?fbclid=IwAR1WNs5HLQaGxIac50scgXXayAhJTE42nofsKwrTIyT0IIyJZulCokXlzF0

[2] Ripple, W. J., Wolf, C., Newsome, T. M., Barnard, P., & Moomaw, W. R. (2020). World Scientists' Warning of a Climate Emergency Sintific Sagnatories form 153 Countries (lit in supliment file S1), (Vol. 70, Issue 1). https://academic.oup.com/bioscience

[3] Fiore, A. M., Naik, V., & Leibensperger, E. M. (2015). Air quality and climate connections. Journal of the Air and Waste Management Association, 65(6), 645–685. https://doi.org/10.1080/10962247.2015.1040526

[4] 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.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73

[5] Ma, J., Cheng, J. C. P., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmospheric Environment, 214(April), 116885. https://doi.org/10.1016/j.atmosenv.2019.116885

[6] Alkasassbeh, M., Alkasassbeh, M., Sheta, A. F., Faris, H., & Turabieh, H. (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), 999–1009. https://doi.org/10.5829/idosi.mejsr.2013.14.7.2171

[7] Faris, H., Alkasassbeh, M., Ghatasheh, N., & Harfoushi, O. (2014). PM10 prediction using genetic programming: A case study in Salt, Jordan. Life Science Journal, 11(2), 86–92.

[8] Zheng, Y. et al. (2015) 'Forecasting Fine-Grained Air Quality Based on Big Data'. doi: 10.1145/2783258.2788573Predicting.

[9] SRao, P. (2014) A survey on Air Quality forecasting Techniques. Available at: www.ijcsit.com

[10] Krakovna, V., & Doshi-Velez, F. (2016). Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models. Whi. http://arxiv.org/abs/1611.05934

[11] Wang, P., Wang, H., Zhang, H., Lu, F., & Wu, S. (2019). A hybrid markov and LSTM model for indoor location prediction. IEEE Access, 7, 185928–185940. https://doi.org/10.1109/ACCESS.2019.2961559

[12] Kim, C. J. (1994). Dynamic linear models with Markov-switching. Journal of Econometrics, 60(1–2), 1–22. https://doi.org/10.1016/0304-4076(94)90036-1

[13] Rabiner, L. R. (1989). Tutorial on Hmm and Applications.Pdf. In Proceedings of the IEEE (Vol. 77, pp. 257–286). http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=18626

[14] Zayed, R. and Abbod, M. (2022) "Big Data AI System for Air Quality Prediction", Data Science and Applications, 4(2), pp. 5-10. Available at: http://www.jdatasci.com/index.php/jdatasci/article/view/63

[15] https://en.wikipedia.org/wiki/Hidden_Markov_model#/media/File:HiddenMarkovModel.svg