INTELLIGENT MACHINE LEARNING MODELLING FOR AIR QUALITY INDEX PREDICTION

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

by

Roba Zayed

Department of Electronic and Electrical Engineering, College of Engineering, Design and Physical Sciences, Brunel University London

March 2024

Abstract

Air quality research affects global warming, climate change, health effects, and urban planning. Predicting air quality status is complex, with increasingly sophisticated monitoring devices for various different gases and other components. Air quality measurements contribute to broader socio-economic development factors, in addition to direct environmental and healthcare impacts. Many methods have been used by researchers to present air-quality levels, reflecting different disciplines and different national standards. This work aims to develop a model to predict the air-quality index, which measures air pollution levels, in order to support healthcare in congested areas.

This research presents machine learning models and techniques to predict air quality levels in cities, and to provide 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 sustainable urban development.

This study presents the experimental multivariate Deep Neural Network model and Markov switching model as part of research to develop a hybrid (DNN and Markov) air quality prediction model, with appropriate accuracy attainment, in order to support decisions with timely air-quality measurements by representing the output or air quality levels using neuro-fuzzy logic.

DNN-Markov modelling techniques are used to predict air quality, with comparative analysis of locations in Jordan and the UK. Multivariate time series analysis of Big Data from traffic-heavy locations was used, based on environmental conditions at peak hours, giving a highly accurate prediction of the air-quality index for the next hour at a given location, under specific environmental conditions. The air quality index was represented using Neuro Fuzzy Logic as a method to contribute in air quality index predictions within blurry (boundary) values. The selected DNN-Markov hybrid model could predict air quality with accuracy of around (RMSE 7.86) for the location in England, and around (RMSE 15.27) for the one in Jordan.

List of Publications

[1] 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

[2] Zayed, R. and Abbod, M. (2022) 'Hybrid intelligent modelling for air quality prediction deep learning and Markov chain unconventional framework', *International Journal of Simulation – Systems, Science and Technology*, 23, pp. 3.1-3.6. Available at: https://doi.org/10.5013/IJSSST.a.23.01.03

[3] Zayed, R., & Abbod, M. (2024). Air Quality Index Prediction Using DNN-Markov Modeling. Applied Artificial Intelligence, 38(1). https://doi.org/10.1080/08839514.2024.2371540

[4] Zayed, R., & Abbod, M. (2024). Breathable cities: Dynamic machine learning modelling approaches for advanced air pollution control. Applied Sciences, 14(13), 5581. https://doi.org/10.3390/app14135581

Declaration

I declare that the research in this thesis is the author's work and submitted for the first time to the Post Graduate Research Office at Brunel University London. The study was originated, composed and reviewed by the mentioned author in the Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, UK. All information derived from other works has been referenced and acknowledged.

Roba Zayed

March 2024

London, UK

Dedication

I dedicate this thesis to my mother Rula, for encouraging me for taking the step for the PhD. I also dedicate this work to my father Professor Faheem, for supporting me to stay directed towards my goal in every possible way. I am also grateful for the support from my sisters, Shaima, Nada, Ayah and Rasha. I dedicate this achievement to my family, who gave me the confidence to take my decision in doing my PhD, and to complete this milestone

Their encouragement and support have been enlightening, as they value what I do to grow at the professional and personal levels, and they did everything I needed to help me flourish.

I appreciate every experience that enriched my intellectual knowledge, and I will hold the wisdom of this journey with me in sharing the knowledge gained in every possible way.

This thesis stands as a collective investment of those who have supported me in my academic endeavours. The support has been remarkable in achieving this milestone, and I am extremely grateful for the wisdom, encouragement, and for always inspiring me to do more.

With heartfelt gratitude,

Roba Zayed

Acknowledgements

I am extremely grateful for the PhD journey, which was an absolute rich learning experience from all perspectives, in terms of theoretical knowledge, practical experience, and gaining in-depth skills in the field of machine learning. It was a true self-reflection journey.

I am extremely grateful for the support provided from my supervisor, Professor Maysam Abbod, for his great guidance and advices throughout my PhD journey. His mentorship has been of extreme importance in developing and shaping this thesis and has significantly contributed to my growth as a researcher. His knowledge in the field has enriched the intellectual landscape of this work. Building and growing my career alongside doctorate was crucial part of the journey and it wouldn't have been possible without Professor Abbod's support. I am truly fortunate to have had a supervisor who not only imparted valuable academic insights, but also demonstrated a genuine commitment to my success. His caring approach has made a significant and positive impact on my entire PhD experience.

Thank you to my family for the support. I am grateful for all that they have done to help me pursue my ambition in every possible way. I would also like to thank everyone who has supported me along the way. Thank you to my friends, research fellows and people who I worked with and shared experiences with on daily basis.

I also want to extend my gratitude to the reviewers whose feedback has improved the quality of this thesis.

I am excited to carry forward all of the knowledge I have gained to the next chapter of my life, and I am looking forward to what is coming next. This unforgettable experience will remain in my mind forever.

Roba Zayed

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AdaGrad	Adaptive gradient algorithm
ADAM	Adaptive moment estimation
ADLs	Activities of daily living
AH	Absolute humidity
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference system
ANN	Artificial Neural Network
AQI	Air Quality Index
AQP	Air quality prediction
AQS	Air quality system
AR	Auto regression coefficient
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with Explanatory Variable
Bi-LSTM	Bi-Directional Long Short-Term Memory
BME	Bayesian Maximum Entropy
CAQI	Common AQI
CEP	Complex Event Processing
CEP	Complex Event Processing
CNN	Convolution Neural Network
CO	Carbon monoxide
DBM	Deep Boltzmann Method
DBN	Deep Belief Network
DNN	Deep Neural Network
DNNM	Deep Neural Network Markov
DTMC	Discrete-Time Markov Chain
DTN	Deep Transformer Network
EDNN	Encoder-Decoder NN
ELM	Extreme Learning Machines
ELU	Exponential Linear Unit
EPA	Environmental Protection Agency
FAQI	Fuzzy-Based Air Quality Index
FAQLP	Fuzzy Air Quality Levels Prediction
FC-LSTM	Fully Connected LSTM
GAM	Greater Amman Municipality

GD	Gradient descent
GHG	Greenhouse Gas
GNN	Graph Neural Network
GP-LSTM	Gaussian Process LSTM
GRU	Gated Recurrent Unit
IGP-LSTM	Input GP-LSTM
IPCC	Intergovernmental Panel on Climate Change
KHG	King Hussein Gardens
KNN	K-Nearest Neighbours Algorithm
LSTM	Long Short-Term Memory
LUR	Land Use Regression
LVAQ	Limit Values for Air Quality
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MBGD	Mini-Batch Gradient Descent
ML	Machine Learning
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MSAR	Markov-Switching Autoregressive
MSE	Mean Square Error
MS-VAR	Markov-Switching Vector Autoregressive
NAAQS	U.S. National Ambient Air Quality Standards
NaN	Not a Number
NN	Neural Network
NO_2	Nitrogen dioxide
O_3	Ozone
PAN	Peroxyacetyl nitrate
Pb	Lead
PM	Particulate matter
POPs	Persistent organic pollutants
PSI	Pollutant Standards Index
PVC	Polyvinyl chloride
R2	Pearson Correlation Coefficient R-squared
RAQI	Revised AQI
ReLU	Rectified Linear Units
RMSE	Root Mean Square Error

RMSProp	Root Mean Square Propagation
RNN	Recurrent NN
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SDGs	Sustainable Development Goals
SO_2	Sulphur dioxide
SP-LSTM	Shared-Private LSTM
SPM	Suspended particulate matter
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
THC	Total hydrocarbons
TL-BLSTM	Transferred Bi-Directional Long Short-Term Memory
UCI	UC Irvine
UNEP	UN Environment Programme
UNFCCC	UN Framework Convention on Climate Change
VARMA	Vector Autoregressive Moving Average
VSL	Value of statistical life
WHO	World Health Organization
WMO	World Meteorological Organization

Chapter 1
Introduction

This chapter introduces about the air quality problem addressed by this research, and related issues and impacts addressed by this study. It gives an overview of machine learning (ML) prediction approaches in the air quality domain. Furthermore, the research aim and objectives are highlighted, in alignment with the identified research contributions and impacts.

1.1 Background

Air pollution is a universal cause for people to band together to solve and prevent its detrimental impacts on human health. As the number of adults increases in the global population, increasing standards of living, including automobile and electricity use, entail increasing volumes of industrial emissions. Keeping in mind the short- and long-term effects, developed and developing countries are directing their environmental efforts for monitoring air quality. However, developing countries still focus primarily on conventional economic growth, based on fossil fuel combustion and legacy technologies, and there are still many leagues to go in order to develop roadmaps to fundamentally improve air quality. As discussed in the following chapter, many air quality evaluations focus on air quality in relation to certain high-impact gases, particularly carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter (PM), and sulphur dioxide (SO_2).

Existing literature also posits that there can be potential reductions in CO_2, SO_2, nitrogen oxide, and CO emissions by changing energy consumption trends. However, Asia is facing particular issues, notably with regard to PM exceeding acceptable limits. Furthermore, ground level O_3 demonstrates average values but exceeds the limits values in all analysis by several factors. In low income countries, economic development is associated with high expectations of increased air pollution (Coelho *et al.*, 2021). Many economic analysts hypothesize that an initial period of increased air pollution (and other forms of negative environmental impacts) during the first phase of socio-economic development is followed by a gradual decrease in such impacts, and ultimately a reduction, as populations become more affluent an environmentally responsible (Leal and Marques, 2022). However, it is clearly preferable for developed and developing countries to seek to minimize negative environmental impacts as much as possible.

While there are debates in the literature on the most instrumental forms of transportation affecting certain aspects of the environment, because of the many diverse factors that have direct and indirect impacts on environmental dimensions, there is general agreement that road transport predominates as the most egregious contributor, and road and air transportation are considered to be the principal contributors to greenhouse gas (GHG) emissions from the transport sector perspective (Font *et al.*, 2019). Most countries consider energy consumption and the transport sector's contributions to be the major constituents of overall to GHG emissions assessment. However, transportation is a secondary phenomenon, reflecting socio-economic activities that require people to move around, including for employment, education, leisure, and other uses. The availability of various forms of transportation left people with many options and expanded life opportunities in the modern world.

With the availability of various means of travel, demand has increased, and the transportation sector is a major macroeconomic component of national economies, as well as a source of indicators for measuring the sustainability of cities and quality of life. Travel networks have been growing in complexity, and so has commensurate travel-related data. However, the majority of transportation is undertaken by road, and road transport remains the dominant mode of travelling and the main cause of travel-related pollution, with growing emissions. A systematic monitoring of PM_{10} by Grivas and Chaloulakou (2006) raised concerns about many areas exceeding EU-legalized limit values. The reason for severity of issue was highlighted due to large and rapidly growing size of vehicles vehicle fleet, diesel exhaust emissions, and the topography of some cities. More recent studies have reiterated that most developing and developed countries do not meet recommended air quality standards for NO_2 near roads, and the decline in near-road NO_2 has been much less than expected, which is largely related to more usage of diesel vehicles in many areas (Font *et al.*, 2019).

The transportation sector plays a major part in air quality indicators worldwide, and accordingly affects the overall global Air Quality Index (AQI), 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. Air quality is a trending issue which varies in complexity at national and international levels. There are numerous standards with which to measure air quality, the most widely used of which are the US Environmental Protection Agency's (EPA) Pollutant Standards Index (PSI), and the subsequently evolved version, the AQI (Cheng *et al.*, 2007). The Revised AQI (RAQI) was also developed as an alternative to both the previously mentioned

standards (van den Elshout, Léger and Nussio, 2008; Tealab, 2018). Air quality indices have many disadvantages (such as a lack of standardization), which can make it difficult to compare air quality at the national level, because different countries and institutions use different methods of measuring air quality, while varying factors affect the hourly pollution concentration (Monteiro *et al.*, 2017).

As explained in Chapter 2, the 'AQI' standard has become so pervasive that 'AQI' is commonly used as the default description to express the level of pollutants' concentration over a period of time, and this is the common way in which researchers refer to work on air quality indicators (Plaia and Ruggieri, 2011). There is still a gap in the AQI system, even though it has been extensively implemented in the US and elsewhere since 1999. Many countries are unable to adopt it, due to the high cost of $PM_{2.5}$ and other necessary monitoring systems, which can entail a severe financial burden for some countries. Due to such instrumental costs, Cheng *et al.* (2007) claimed that full AQI implementation would not be likely in the near future, and noted the need for a reliable and comparable air-quality index standard to understand the situation in different countries. More recently, Tripathi and Pathak (2021) noted that it will not be possible to develop a universal air-quality index which covers all situations and types of pollution, and argued that the focus should instead be on particularly vulnerable (i.e., highly polluted) zones.

A universal technique needs to be developed, as we lack a reliable methodology with which to respond to human exposure to pollution. This is needed to ensure quality of life, especially given the differences between air pollutants in different locations. No universal air-quality index exists, in particular for vulnerable (highly polluted) areas, and a method for identifying zones with high air pollution is also needed, as there is no international air-quality index (Tripathi and Pathak, 2021). The current air-quality index methodologies are limited, as they do not consider pollutant numbers and variations, and do not measure the health implications of exposure to pollutants in the environment. While the EPA's (2023) AQI standards are used in several countries, and currently comprise the closest thing to a universal tool, the existing literature has shown that it manifests many gaps and shortfalls as an international standard, and the prevailing milieu is one of international incoherence in standards and measurements (van den Elshout, Léger and Nussio, 2008; Bishoi, Prakash and Jain, 2009; Monteiro *et al.*, 2017).

RAQI does not consider pollutants with maximum concentrations, unlike the original AQI, and instead calculates the concentrations of other pollutants. This could produce a more accurate assessment of air quality than the AQI, as it also considers the contribution of other gases to pollution. Furthermore, one pollution air-quality index mentioned in the literature

provides a simple method of calculating the weighted mean values of sub-indices of the most critical pollutants (Sowlat *et al.*, 2011). Because of these fundamental differences between tools, there is a common consensus that necessary global standards are not currently in place. It is also understood that a more dynamic system is required which can accommodate a mixture of different pollutants and which is sufficiently sensitive to boundary-level pollutant predictions. Consequently, researchers have been evaluating various methods of producing a universal tool to measure the health implications of pollution (Mandal and Gorai, 2014).

Some algorithms have been applied to solve issues with current domain systems. Fuzzy logic, a decision-based model representing uncertainties, is substituted for other available methods when a blurry or otherwise boundary air-quality level is presented (Baatarchuluun, Sung and Lee, 2020). Fuzzy-logic air-quality index has been used to produce a logical, reliable and dynamic way to present the health effects of pollutants to the public (Niharika and Rao, 2014). Fuzzy logic has the ability to map different categories with uncertain values, which can be described as 'fuzziness' (Sowlat *et al.*, 2011; Kang *et al.*, 2018). This study proposes a comprehensive framework to address the gaps observed in the literature concerning the prediction of air quality and problems identified with air quality assessment using AQI.

1.2 Research Motivations

Traditional methods to predict air quality suffer from disadvantages such as their limited accuracy (e.g., inability to predict extreme measurement points), not being able to determine cut-offs, inefficient approaches for better output prediction, and equal treatment for old and new data. The uses of Big Data and ML have been proposed as advancements on the traditional methods, and they have been widely used in air quality prediction (AQP). Several studies have evaluated air quality using ML algorithms, exploring variations in ML models to predict air quality (Rybarczyk and Zalakeviciute, 2018). Big Data has enabled modelling more dynamic air quality systems (AQSs), which are behaviourally heterogeneous; such models take data from various resources. AQP 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 have been adopted for air pollution forecasting, including more recently available techniques such as ANN, Fuzzy Logic and Genetic Algorithms were used in air pollution modelling (Alkasassbeh *et al.*, 2013). Several ML approaches have been used by

experts, researchers and others, with different parameters combined for AQP. 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 (Rybarczyk and Zalakeviciute, 2018).

There are two fundamental paradigms for air pollution modelling: traditional chemistry dispersion analysis (of chemically inert species), and ML. 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 factors instrumental in the complex system of air pollutants and other features are highly non-linear. Consequently, the simplest statistical approaches of regression analysis and Autoregressive Integrated Moving Average (ARIMA) models would not be commensurate with the pertinent complexity. Generally, more advanced statistical ML methods such as Support Vector Machine (SVM), ANN, 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. Experimental research has concluded that there is a close relationship between the concentration of air pollutants and meteorological variables and pollution (Zhang and Ding, 2017).

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 Network (ANN) analysis. Statistical techniques do not consider dynamic chemical and physical processes, and instead rely on historical data to predict future concentrations of air pollution (Zhang and Ding, 2017).

1.3 Aims and Objectives

The fundamental aim 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 (DNN) through a niche experimental framework. The experiment is part of applied Big Data Artificial Intelligence (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 national contexts: England and Jordan.

The following are the main objectives necessary to achieve the stated aim:

- Reduce data complexity processing through selecting best ML methods to support air quality domain
- Produce a reliable and accurate model to predict air quality
- Produce an AQI model for policy and regulations supporting health and climate change issues
- Produce a prediction model considering transportation/traffic factor

The following are the research questions guiding this research:

- What parameters have been used for air quality modelling?
- What are the gaps that need to be researched in the air quality field?
- What are the independent and dependent (input and output) variables to be used building air quality model to predict AQI?
- What are ML models for best accuracy prediction?
- What data pre-processing techniques can achieve the best results?
- What air-quality indices are commonly used nowadays?
- What improvements can be applied to current air-quality indices?

1.4 Contributions to Knowledge

The study conducted several methods in an effort to contribute with a new and efficient approach to predict air quality for the next hour. DNN-Markov approach was selected, as model validation indicated its promising potential for efficiency, and it enables using a simple linear model for backup, given the complexity and losses that could occur with DNN models. Its characteristics boost performance. This research represents a set of contributions, including but not limited to the following:

- Proposing a novel hourly prediction model.
- Testing multivariate input and output models that support the complexity of AQP.
- Hybrid modelling methods (combining Markov and DNN).
- Access and analysis for hourly regional data, with added value due to the increased accuracy of data results (particularly for Jordan Data).
- An AQI model generated based on hourly data, to produce better results and accuracy.
- Extending research on transportation factors (pertaining to transportation emissions).
- Addressing data refinement and model accuracy by generating a model to cover such challenges (such as missing data and reducing noise).
- Proposing the best combination of models to cover complex gases that are currently creating challenges in prediction (such as PM).
- A hybrid model considering static and dynamic variables, for more accurate results.
- AQI representation.

Hourly data in this thesis refers to the frequency at which air quality data is collected and predictions are made. In this study ‘hourly’ refers to the collection and analysis of air quality data at one-hour intervals. Further, predictions are generated for the subsequent hour using machine learning models trained on historical data.

1.5 Research Impact

While conducting the literature review, it was discovered that there has been a salient shift in academic discourse from ‘climate change’ to ‘climate crisis’, and world leaders have expressed increased fears that global warming will cross the safety threshold of 2°C. This trend is reflected in newspaper headlines such as ‘“Untold human suffering”: 11,000 scientists from across world unite to declare global climate emergency’. This headline was designed to emphasize 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’ (Weston, 2019).

Most indicators, however, are not very promising for humanity, given the severe increase in global CO_2 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 actions to enhance the natural environment, such as protecting ecosystems and promoting sustainable practices.

Furthermore, 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 (Ripple *et al.*, 2020). As a result of pressure from human activity since the presentation of the UN’s 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 (Fiore, Naik and Leibensperger, 2015).

1.6 Overview of Thesis Structure

This thesis constitutes seven chapters, structured to fulfil the research aims and objectives stated in this introductory chapter, and developed around the technical aspects of the comparative study between air quality predictive index in UK and Jordan.

Chapter 2 reviews literature concerning air quality definitions, impacts, and regulation; machine learning (ML) approaches to AQP; and related studies and applications. It identifies research gaps and the need to develop improved tools for decision makers concerning air pollution. Previous work by researchers in air quality domains and different ML approaches provide a robust basis upon which this study seeks to extend. The chapter reviews optimization methods and hyper-parameter tuning strategies that are crucial in model developments. This chapter then summarizes the findings and highlighted the identified gaps in current literature.

Chapter 3 discusses the experiment design framework for modelling and the methodologies used in the construction of the proposed models for this study. Data sources, collection, data imputation, and pre-processing are explained. The proposed modelling approaches for the experiments are discussed, and a summary for the design framework is included as a flow chart as illustration for the methodology framework.

Chapter 4 covers the stand-alone models implementation and results to predict air quality, the approaches and techniques used to build the stand-alone ANN model, DNN model, and Markov Model.

Chapter 5 presents hybrid modelling based on some selection and methodologies to create hybrid models from the stand-alone models discussed in previous chapter, with the aim of producing a reliable model for air quality modelling with a suitable accuracy, as a proposed solution to the identified problem in this research.

Chapter 6 presents neuro-fuzzy logic representation for the predicted output of the selected hybrid model; it introduces the several air quality standards available, especially those which are commonly used globally, and then represents the outputs based on AQI.

Chapter 7 presents critical conclusions of this study and discusses the future work arising from this research; it lists the gaps discovered in this research and directions for further inquiry. Also, limitations are identified within the scope of this study.

Chapter 2
Literature Review

2.1 Introduction

This chapter provides an analysis of background literature pertaining to this research area, with a thorough critical literature review on AQP methods. It identifies literature gaps addressed by this study, and its contributions unpacked in later chapters. This is a thorough literature review which aims at integrative research to build an AQI model using the best selection of ML methods in the domain, to support decision makers with more effective tools for emissions and pollution regulation. The ultimate rationale for such research is to improve air quality, and reduce related negative issues such health and climate change, by reducing emissions. The study focuses on data from traffic areas, to evaluate the impacts of transportation on air quality and in efforts to translate available data to emissions share from transportation in selected areas.

2.2 Fundamentals of Air Quality

2.2.1 Definition and Guidelines

While the term ‘air pollution’ reflects the presence of pollutants in the air, the broader concept of ‘air quality’ alludes to the general quality of the air we breathe. Clean air is a very basic need for humanity, for health and other life aspects. The ways in which to define and measure air quality, and the commensurate data required, are in fact very complex (Plaia and Ruggieri, 2011). ‘AQI’ has come to be used over the years to express the level of pollutants concentration over a period of time, in a way that is understandable by the public and decision makers.

2.2.2 Health, Environment, Social, and Economic Impacts

Liu *et al.* (2023) conducted a bibliometric review of 100 studies published over 20 years in an effort to study the economic impacts of air pollution, specifically in terms of a cost-benefit analysis. They concluded that air quality, heath, climate, and economic growth in all countries’ scales are all interconnected, and would need long-term strategies and sustainable policies that support holistic sustainable economic development in order to address all pertinent dimensions. Air quality impacts are interconnected, spanning health, social, and environmental systems, all of which in turn have reciprocal economic impacts on cities. For instance, reduced environmental (i.e., air) quality increases hospital admissions

from those affected by pollution, such as asthma, stroke, and respiratory illnesses, which entails economic costs of healthcare and reduced labour capacity. The interconnected dimensions pertaining to air quality are displayed in Figures 2-1 and 2-2.

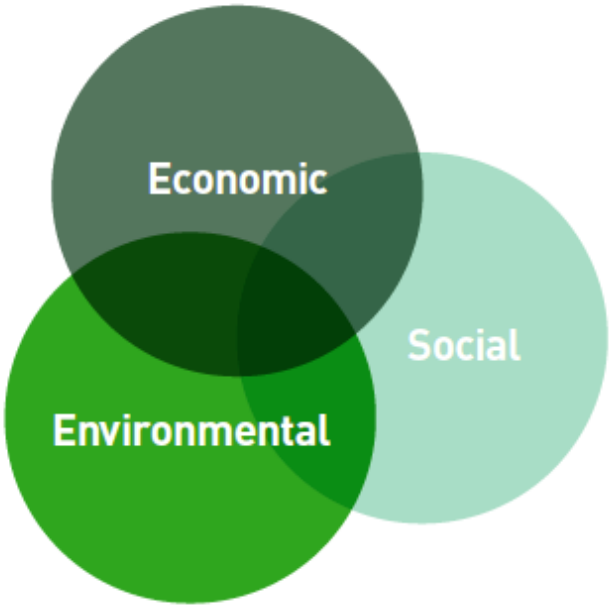

Figure 2-1 Intersection of economic, environmental, and social air pollution impacts

Source: Su et al. (2021)

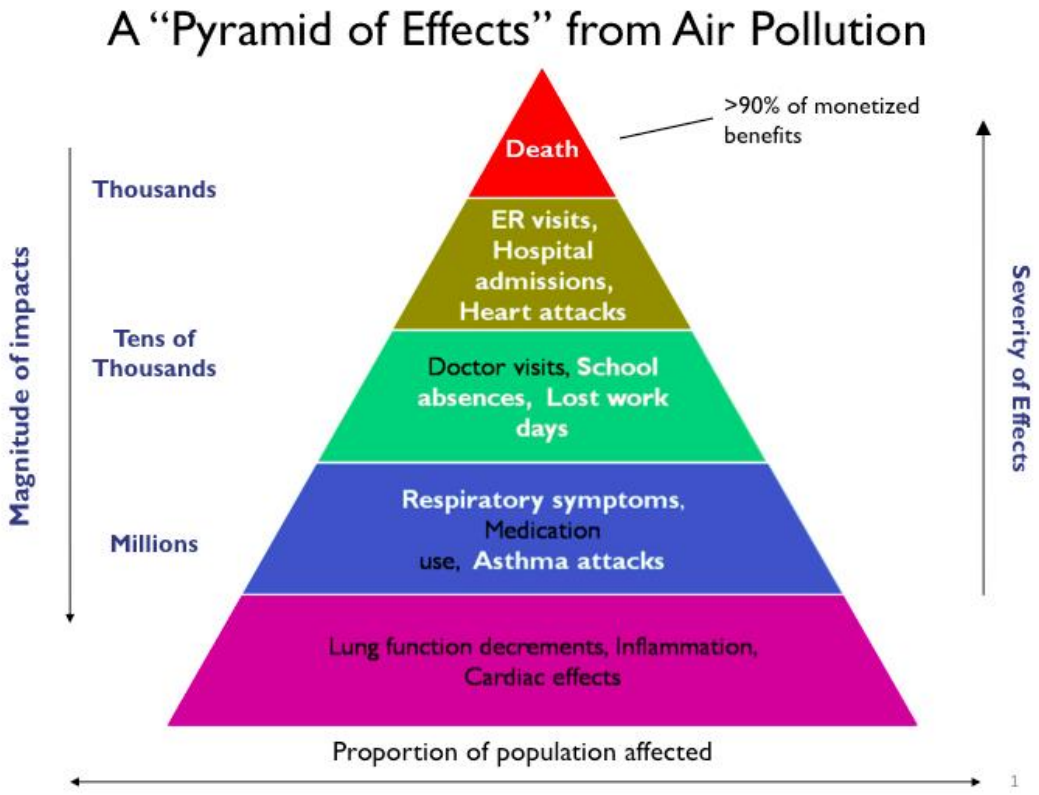

Figure 2-2 Air pollution impacts

Source: EPA (2023)

2.2.2.1 Health Impacts

The health implications of exposure to air pollutants are the most intensively studied, and researchers are constantly studying the linkages between many diseases and air pollution (see Figure 2-3). The World Health Organization (WHO) refers to air pollution as 'the silent killer', because of its apparently innocuous short-term effects, but its profound long-term impacts, including the ever-increasing number of related deaths every year. Air pollution is considered as a risk factor for major diseases relating to changes in lung functions, including asthma and cardiovascular illnesses, and it is particularly dangerous in relation to adverse pregnancy outcomes (e.g., heart failure, atherosclerosis, cardiac arrest, and arrhythmias), all of which could decrease life expectancy and hence leads to death (Méndez, Merayo and Núñez, 2023).

Emerging research suggests correlations between diabetes among women aged under 50 years old and exposure to O_3 air pollutants, while SO_2 appears to have a high correlation with psychiatric illnesses. Toxic pollutants are known to cause disorders such as Alzheimer's and Parkinson's disease, and are instrumental in developing psychological distress. Some research indicates a correlation between seasons and exposure to air pollution impacts on

health, with fluctuations between warm and cold weather being potentially associated with many forms of mental illness (Tripathi and Pathak, 2021).

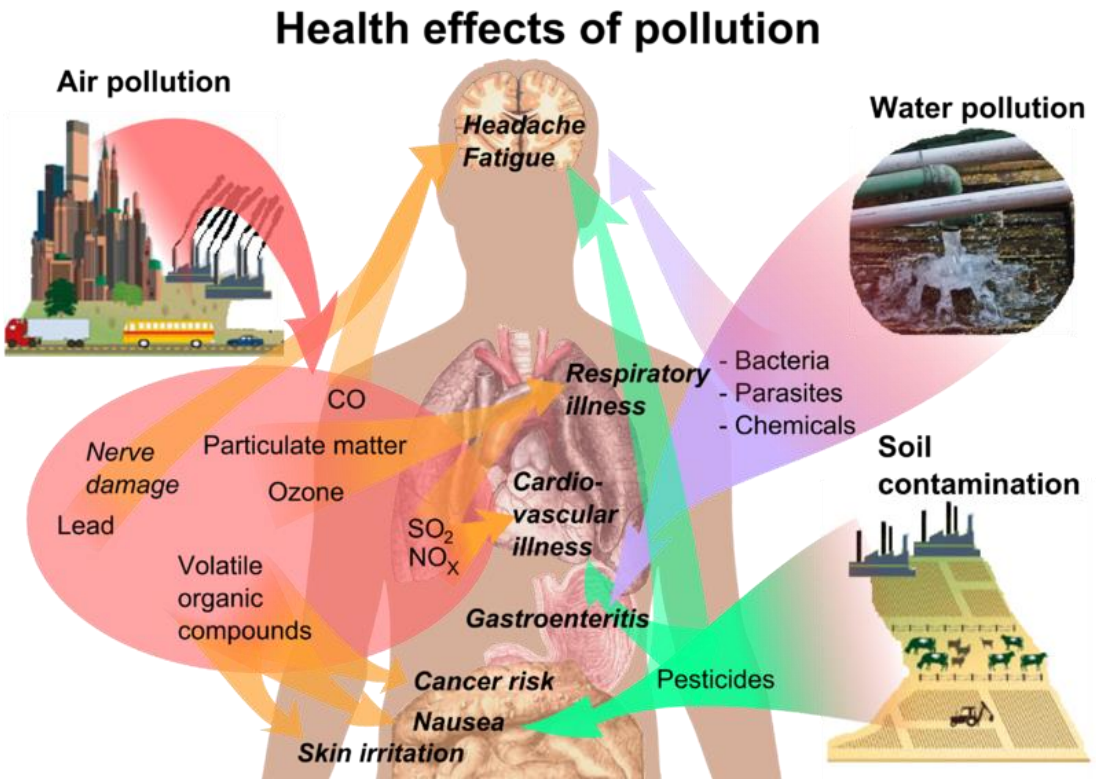

Figure 2-3 Health effects of forms of pollution

Source: FCCMG (2016)

2.2.2.2 Environmental Impacts

Researchers have considered the impact that weather conditions, such as wind direction, can affect air quality, depending on neighbourhood areas. Strong wind speeds generally promote the rapid travel of pollutants to other places and different distances, while high temperatures contribute to photochemical reactions. Conversely, rain generally cleans the air, although it can also cause acid rain and soil pollution due to exposure to airborne toxins (see Figure 2-4). There is a need for more quantitative studies of air quality to help in climate change reduction, based on scientific understanding of air quality and climate change ties, specifically with regard to various environmental impacts (Liu *et al.*, 2017).

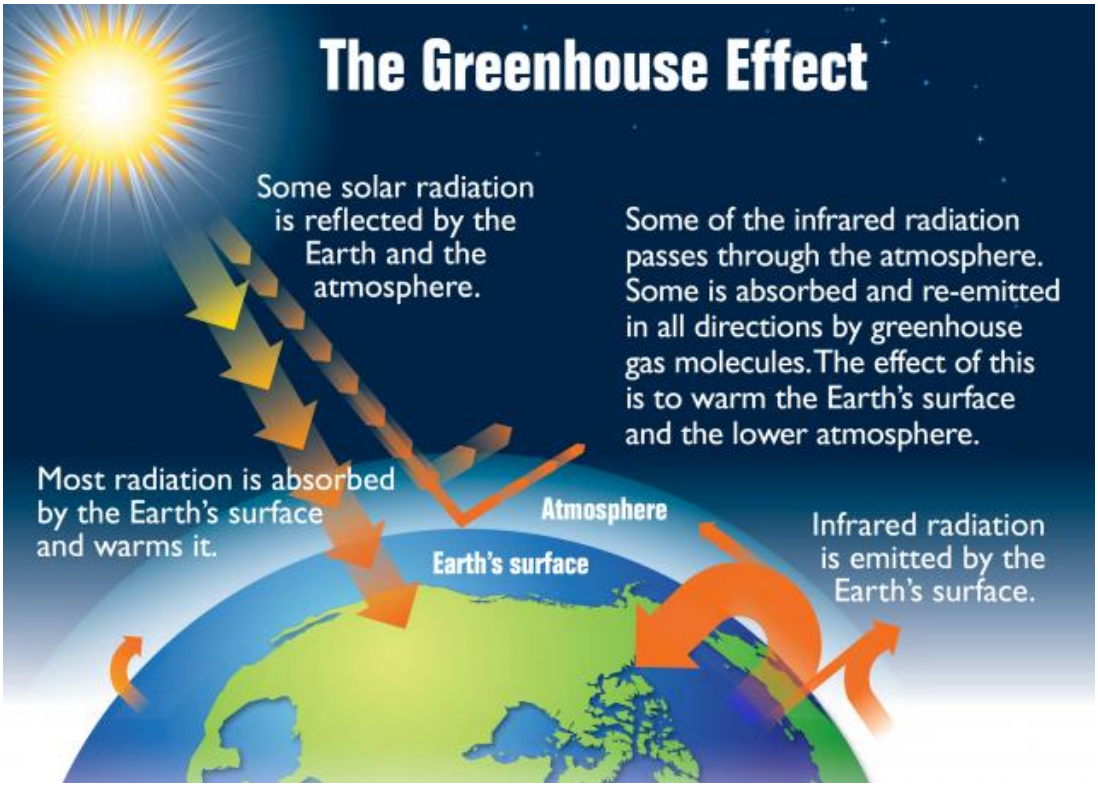

Figure 2-4 Emissions and climate change linkage

Source: EPA (2023)

2.2.2.3 Social Impacts

Cities worldwide are affected by various consequences of air pollution, and in the worst cases there may be scenarios where people are unable to undertaken normal activities of daily living (ADLs). For instance, if pollution levels are high, especially for gases that are classified by WHO as having high and direct impacts on human health, such as PM. The consequences of such scenarios are on the rise, as some cities tend to close some areas where pollution is at a high level, as it could also reduce or prevent visibility, which causes a lack of clarity on the roads, and which could become very dangerous for drivers in the case of smog and fine PM.

2.2.2.4 Economic Impacts

The link between air pollution or the air quality to health is evident, as discussed previously. Elderly people and those with respiratory problems (e.g., lung diseases or asthma) are particularly vulnerable to immediate impacts from outdoor exposure to pollutants, and are more susceptible to require hospitalization, thereby increasing healthcare and macroeconomic costs associated with poor air quality. Improved understanding of air pollution throughout the years has led to increasing demand to evaluate the broader socio-

economic impacts related to pollution-linked health issues. Economic evaluation of air pollution impacts consider the most egregious impacts to be premature death in terms of the value of statistical life (VSL), but work absenteeism and direct healthcare costs are more common and pervasive pollution impacts that can be used to evaluate pollution-related risk (Hall, Brajer and Lurmann, 2010).

2.2.3 Regulatory Framework

2.2.3.1 Air Quality Standards

Air pollution and climate change are controlled by several standards, agreements, and measures ranging between mandatory, voluntary and integrated initiatives (Hall, Brajer and Lurmann, 2010). The Paris Agreement was signed on Earth Day, 2016, at the UN headquarters in New York. Its main objective is to keep the global temperature rise below 2°C, and ideally to limit the temperature increase even further to 1.5°C. Additionally, the Intergovernmental Panel on Climate Change (IPCC) was formed by World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP) 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 GHG guidelines have a detailed method for estimating GHG emissions by source and removals by sinks.

Furthermore, the Kyoto Protocol is an international agreement aimed at reducing CO_2 and GHG emissions in the atmosphere; it was linked to the UN Framework Convention on Climate Change (UNFCCC), adopted in Japan in 1997. As a result of excessive human activity and since the presentation of SDGs, related bodies have aimed to reduce social, economic, and environmental imbalances at several scales, as per the 'Our Common Future' policy. 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 AQS and resources.

The U.S. National Ambient Air Quality Standards (NAAQS) are limits on atmospheric pollutions concentrations that impact health and environment by the EPA under authority of the Clean Air Act (42 U.S.C. 7401 et seq.), It consist of six criteria pollutants: O_3, atmospheric PM, lead, CO, SOx and NOx. They are emitted from industry, mining, transportation, electricity generation and agriculture (Wang *et al.*, 2023). 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 (Polezer *et al.*, 2023), and the EU's Limit Values for Air Quality (LVAQ), in addition to NAAQS. The latter is chiefly concerned with criteria for the following common air pollutants that harm the environment and human health: CO, lead (Pb), NO_2, O_3, PM, and SO_2. Outdoor air pollution is mainly created mainly by automobiles and various industries which are responsible for the climatic change (Niharika and Rao, 2014). There are two main categories of air pollution in terms of their functional source:

Primary Pollutants: results from combustion of fuel and industrial operations

Secondary Pollutants: results from the reaction of primary pollutants.

The AQS is used in assessing air quality contains ambient air pollution data by EPA, 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 NAAQS contains two types of standards, primary and secondary. As different pollutions have different effects, the standards are built to accommodate the protection against it (short and long term). AQI quantifies air quality in a region, and it is used in government agencies to communicate the air pollution status (Kang *et al.*, 2018).

Tables 2-1, 2-2, and 2-3 present more details on the regulatory framework pertaining to air pollution.

Table 2-1 Air pollutants and their sources

Pollutants	Sources
SO_2/ oxides of sulphur	Power plants, sulphuric acid manufacture, boilers, ore refining, petroleum refining.
Suspended particulate matter (SPM) (from sulphates and nitrates)	Fine particles (synthetic or natural). Automobiles, power plants, boilers. Industries requiring crushing and grinding (e.g., quarrying, 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, aluminium refining, nuclear industry, steel industry, oil refineries/
Oxides of nitrogen (NO, NO_2, NOx)	Automobiles, power plants, nitric acid manufacture.
Peroxyacetyl nitrate (PAN)	Secondary pollutant.
Persistent organic pollutants (POPs)	Produced through industrial processes and waste incineration.
Formaldehyde	Secondary pollutant.
O_3	Secondary pollutant, formed from chemical reaction during sunlight.
CO	Automobiles, from combustion 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; emitted from agricultural processes and farm animals.
CO_2	From volcanic activity and hot springs, combustion processes, cars and plants.

Source: Niharika and Rao (2014)

Table 2-2 NAAQS criteria for pollutants and standards

Pollutant	Primary/ Secondary	Averaging Time	Level	Form
CO	Primary	8 hours	9 ppm	Not to be exceeded more than once per Year
		1 hour	35 ppm	
Pb	Primary and secondary	Rolling 3 month average	0.15 µg/m3	Not to be exceeded
NO_2	Primary	1 hour	100ppb	98th percentile of 1-hour daily maximum concentrations, averaged over 3 years
		1 year	53 ppb	Annual mean
O_3	Primary and secondary	8 hours	0.07 ppm	Annual fourth-highest daily maximum 8-hour concentration, averaged over 3 years

Source: Kang *et al.* (2018)

Table 2-3 AQI classification

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

Source: Kang *et al.* (2018)

2.2.3.2 Air Quality Systems

There are discussions in the existing literature about several AQI standards or systems that have been used by researches for air quality levels (Bishoi, Prakash and Jain, 2009). Table 2-4 summarizes gaps identified in existing research, and many works have cited the need for further work in this area, to address known deficiencies in some standards (Plaia and Ruggieri, 2011; Mandal and Gorai, 2014). Also, some novel researches discovered the benefits of applying fuzzy logic for the use in AQI (Sowlat *et al.*, 2011), especially in relation to the need of a more comprehensive system to model the complexity of the existence several pollutants in the atmosphere in a period of time, which could need careful consideration of how some gases could become dominant in certain setups (Lokys, Junk and Krein, 2015).

Table 2-4 Reviewed AQI systems' highlights

Article	Remarks
Comparison of the Revised Air Quality Index with the PSI and AQI indices (Cheng *et al.*, 2007)	The Pollution Standards Index (PSI) was established due the increasing number of people suffering from respiratory problems, and subsequently developed into AQI. RAQI was developed as an alternative to PSI and AQI, achieving more significant outcomes as it covers a wider range of pollutants and concentration levels. RAQI gave more accurate results than PSI and AQI, with certain abilities to distinguish certain pollutants. The cost of establishing a monitoring system covering PM2.5 is particularly high for many countries to implement, but this is necessary, especially in light of global O_3 problems
Novel, fuzzy-based air quality index (FAQI) for air quality assessment (Sowlat *et al.*, 2011)	In this study, FAQI using fuzzy logic was proposed using different pollutants, based on different weighting factors. The FAQI was suggested as of the limitations provided using AQI, so as FAQI is seen as a more sensitive tool. Results of FAQI were compared to AQI USEPA, and the authors flagged FAQI as a comprehensive, reliable method for decision makers.
Towards an improved air quality index (Monteiro *et al.*, 2017)	The authors claimed that there are no universally significant methods that covers for all specifics situations of air quality, and pointed out that methods of AQI could differentiate based the number of pollutants, air quality levels categories and boundaries points, and sampling period. The current AQI has limitations, as it is difficult to compare air quality levels across countries using it. The authors suggested adding PM2.5 specific standards and indexing for specific sources of pollutants ('traffic areas', 'industries', and 'others'), and adding a 'natural events' factor, as well as producing data when monitoring is not working. The authors acknowledged the complexity of developing more significant AQI, and pointed out the need to study long-term factors for air pollutants, along with health related descriptions
A comparative study of air quality index based on factor analysis and EPA methods for an urban environment (Bishoi, Prakash and Jain, 2009)	Factor analysis of the National Air Quality Index (NAQI) was suggested to be used to cover the gaps in the EPA's system. The authors claimed that NAQI could be used for comparing daily and seasonal pollution levels in different areas, to allow monitoring of seasonal trends.
Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative (van den Elshout, Léger and Nussio, 2008)	The authors proposed a new Common AQI (CAQI), to be able to compare air quality levels across Europe. It consists of two indices, one for roadside sites and the other for average city background conditions. The structure is assumed to bring consistency when comparing parameters.

2.3 Machine Learning Approaches to Air Quality Prediction

2.3.1 ML Background and Theory

This section explains the ML foundation of this research, including ML theory, and its relation to AI as a key future technology. Alzubi, Nayyar and Kumar (2018) outlined the history of the formation of the ML and AI fields. They noted that 'machines' in the context of ML comprise systems or computers, while 'learning' refers to the process of acquiring new knowledge, skills, behaviours, and techniques The term 'machine learning' was defined by Arthur Samuel in 1959 as 'the learning capability of computers that provides learning capability to computers without being explicitly programmed'. Mitchelle developed the definition to be more applicable to engineering applications: 'A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E'.

While the research on ML field has been developing over more than six decades, it has always been guided by the 'Turning Test', developed by Alan Turning in the 1950s, in alignment with which Sameul developed a learning algorithm with high capabilities in 1952. Martin Minsky and John McCarty with Claude Shannon and Nathan Rochester popularized the term 'Artificial Intelligence' in a conference in 1956, and in 1958, Frank Rosenblatt initiated the development of ANN through the perceptron concept. After this progress, massive work was undertaken in the ML field, and a major highlight was presented in 2006 with the deep learning presented by Geoffrey Hinton, which has been a tremendous improvement in support for the ANN architecture to cover multiple layers of neurons. Researchers are expecting more development of the DNN concept, with plenty of forthcoming innovative applications (Alzubi, Nayyar and Kumar, 2018).

2.3.2 Modelling Review for Air Quality Prediction

Many factors play a role in the prediction of air quality, as it is not standalone in air *per se*; rather, it is affected by atmospheric conditions, and has time dependencies. Air-quality concentrations have dependencies on (and state-fluctuations between) air pollutants and impacts such as meteorological events. Moreover, the availability of valid air-quality datasets is occasionally a barrier to optimal forecasting in particular domains (Tripathi and Pathak, 2021). A high level of pollutants and the consequences of this have made it necessary for immediate action to be taken to reduce pollution levels. The importance of this action has drawn the attention of researchers into air quality investigations. Because of the importance of this research, work has been done exploring different perspectives, such as parameters,

temporal dimensions, and spatial interactions (Cheng *et al.*, 2007; Alnawaiseh and Hashim, 2014; Masih, 2019).

Emissions are a complex mixture of gases and meteorological conditions in the atmosphere. This combination of factors presents limitations for forecasting modelling, due to non-linearity and a lack of meteorological parameters in some regions. A review by Masih (2019) argued that there are two types of ML, forecasting and estimation, and ensemble learning and linear regression can be used for modelling estimates. However, the author suggested that methods such as NN and SVM may be more useful for forecasting (see Table 2-6).

A systematic review by Tealab (2018) reported that Neural Networks (NNs), SVMs, and ensemble learning algorithm are commonly used, as these are known for their ability to capture non-linearity in modelling. A survey of ML algorithms used to forecast air quality by Méndez, Merayo and Núñez (2023) concluded that the main algorithms applied for pollutant concentration predictions can be classified into three different categories: classical regression based algorithms, ML regression based algorithms, and deep learning algorithms. A number of algorithms fall within the supervised learning classification paradigm, such as decision trees, naïve Bayes, SVM, and regression analysis (the latter of which may be linear, logistic, and polynomial). Despite the importance of traditional methods of predictive modelling, such as Auto-Regressive Integrated Moving Average (ARIMA), Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) and other statistical methods of linear modelling, other methods such as deep learning are needed in order to better model the non-linearity of data between parameters, and to give reliable performance and accuracy in time-series data (Siami-Namini, Tavakoli and Namin, 2019).

There are also other used methods for air pollution, such as Multiple Linear Regression (MLR), which proves efficiency in linear relationships between output and multiple independent inputs. Multivariate linear modelling can be particularly suitable for PM predictions. Moreover, SVM and Bayesian Maximum Entropy (BME) have been considered as emerging methods offering comparable performance to Vector Autoregressive Moving Average (VARMA), ARIMA, and MLP.

The following subsections review literature on relevant ML models that support the accuracy of AQPs, given the optimal aim of considering factors such as non-linearity nature of air quality components, and multi-dimensional parameters relating to the complexity of emissions etc. A comprehensive review of ML models used in air quality studies is given below, studying related techniques and analysing their parameters, data frequency, data

sizes and type of model used. The subsequent discussion is formed based on the review and analysis of the gaps and findings of existing literature.

2.3.2.1 Artificial Neural Networks

ML, also known as ‘predictive analytics’, is simply a collection of instructions (algorithms) that accumulate data and learn from it, to improve over time. Many applications use ML nowadays, such as using of NNs to solve various industrial problems through the use of available data. Part of this science is also reliable on statistics, but that is not the case for all ML implementations. NNs were originally conceptualized as models of the functionality biological brain (Shao and Shen, 2023), and ANNs posit that weights between nodes in data systems are analogous to connected neurons in the brain. The neuron is the initial element of an NN, in which the input and output values are exchanged. There are number of NN types, the more basic of which, such as feed forward networks, have manifest limitations in modelling time prediction tasks. ANN is basically a perceptron, as Figure 2-5 shows. It consists of multiple input external links, one output, and an internal input (bias) (Staudemeyer and Morris, 2019). The perceptron receives a vector of real-values, and the output of the perceptron is Boolean (zero or one). A more developed form of ANN is the feed-forward neural network, which expands from the most basic type of artificial neuron concept to encompass a number of neurons structured in layers, each neuron of which has a computed weighted sum of its inputs. A single layer perceptron network is one layer consisting of a set of neurons. Sets of neurons structured in several layers can be defined as a multilayer feed-forward neural networks (Scabini and Bruno, 2023).

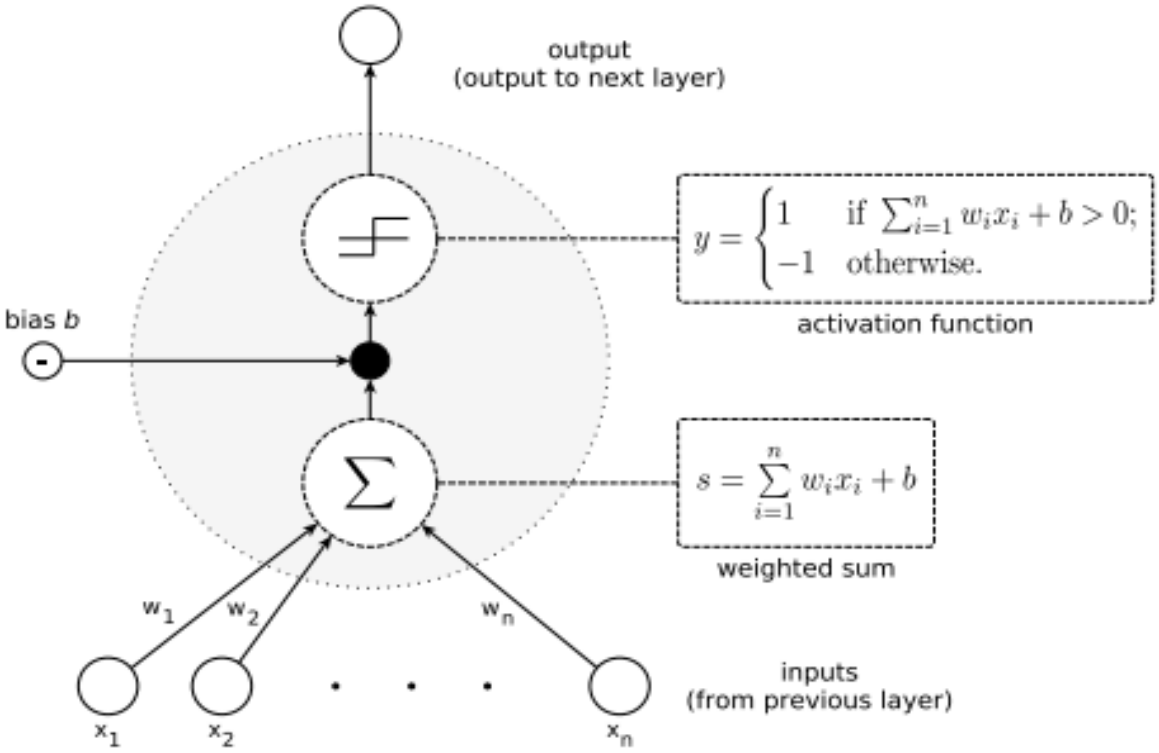

Figure 2-5- The structure of the most basic type of artificial neuron, called a perceptron

Source: Staudemeyer and Morris (2019)

ANN models had already been widely used for air pollution concentration predictions during the early 2000s, as noted by Grivas and Chaloulakou (2006), whose findings confirmed the superior performance of ANNs in comparison to traditional statistical methods such as multiple regression, classification and regression trees, and autoregressive models. ANN models have shown better performance than 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, including a propensity to local minimum and poor generalization, lack of analytical model selection approach, being time consuming (in relation to finding the best architecture), and its trial-and-error weighting mechanism.

ANN models have the ability to capture the highly non-linear character of those processes serving a wide range of gaseous prediction. An analysis of a range of research into AQP has reported that NNs are one of the most reliable, cost-effective machine-learning tools for prediction purposes. Based on the training methodology, there can be supervised NN where inputs and outputs are given to the network, unsupervised NN this is when no output is given and reinforcement NN and reinforcement learning where the NN learns from past decisions

and weights are consequently adjusted, based on related output/response health, which produces relatively high accuracy (Grivas and Chaloulakou, 2006).

Recently, NN models have been used for PM mass concentrations predictions, which is a more complex task compared to the forecasting of gaseous pollutants. This is due to the complexity of the related processes (the formation, transportation, and removal of aerosol in the atmosphere). ANN models have the ability to capture the highly non-linear character of those processes, serving a wide range of gaseous prediction purposes. Time series ANN research has saturated existing literature in this field (Niska *et al.*, 2004; Shrestha and Mahmood, 2019; Baatarchuluun, Sung and Lee, 2020), indicating NNs' efficiency. Nevertheless, the use of ANN to solve problems still depends to a great extent on the skill and experience of modellers, and few systematic procedures exist for it (particularly for time series which require non-linear forecasting).

2.3.2.2 Deep Neural Networks

When the universal approximation theorem slowed the evolution of Deep Neural Networks (DNN), the back propagation learning algorithm took over as a leap forward in DNN, in which the automation of feature extractors' added value compared to traditional machine-learning techniques. However, given the shortcomings of deep learning algorithms, DNN was shaped into different architectures and training techniques, offering revolutionary resolutions for deep-learning models. An increase in layers led to greater capacity for network learning, but not necessarily an improvement in accuracy. Due to the limitations in achieving high accuracy in forecasting model, especially when complex parameters are modelled, many researchers directed their efforts towards deep (rather than shallow) learning architectures (Zaini *et al.*, 2022). Despite the high potential of deep learning models, their implementation for air quality domain would still be seen not fully utilized with many areas for development.

A review of deep learning NN for time series AQP by Alzubi, Nayyar and Kumar (2018) looked at different strategies ways on how deep learning were developed through recent studies including features extraction, data decomposition, and other deep learning components. The study reviewed different elements of deep learning models, such as the model topology, input parameters, output parameters and performance criteria.

2.3.2.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are considered to be an extension of conventional feed-forward NNs, with added ability to process sequential inputs with a memory which is featured to handle previous sequential information. However, this memory is limited to only some steps back (Siami-Namini, Tavakoli and Namin, 2019). The RNN differs from traditional ANN

in having the basic unit of the hidden layer as the memory block which contains memory cells with self-connections for memorizing the temporal state and a pair of adaptive multiplicative gating units to control information flow in the block and the input and output control the activation in the block (Krakovna and Doshi-Velez, 2016). The following is a representation of RNN model updates.

Assuming x represents a sequence of length T:

$$x = (x1, x2, x3, x4, x5, x6, x7.....xT) \quad (2.1)$$

h_t represents RNN memory at time step t:

$$h_t = \sigma\ (W_x^{xt} + W_h^{ht-1} + b_t) \quad (2.2)$$

where W_x and W_h are weight matrices and b_t is a constant bias.

RNNs have various forms, including the following:

- One input to many outputs
- Many inputs to one output
- Many inputs to many outputs

Some drawbacks for RNNs should be mentioned, such as ‘vanishing gradients’, in which case input information passing though many layers vanishes when reaching to the beginning or the end layer, or ‘exploding gradients’, whereby input information passes through many layers that will end up with a large gradient when reaching to the beginning or the end layer. Such issues during the training of RNNs create more problems when long-term dependencies occur (see Table 2-6).

2.3.2.4 Long Short-Term Memory

Long Short-Term Memory (LSTM) models are a type of RNNs. While traditional RNNs are limited to 10-time steps back, LSTM models achieve greater reliability for long-time series data models due to their learning capacity encompassing more than 1,000-time steps. The LSTM architecture helps to accommodate inputs for long-term dependencies (i.e., it captures features and preserves information over a long period) (Ma *et al.*, 2019). LSTM model is based on three gates:

- Forget gate: decisions gate
- Output gate: output results presented
- Input gate: information to be added to the memory

The structure and functionality of an LSTM cell is displayed in Figure 2-6.

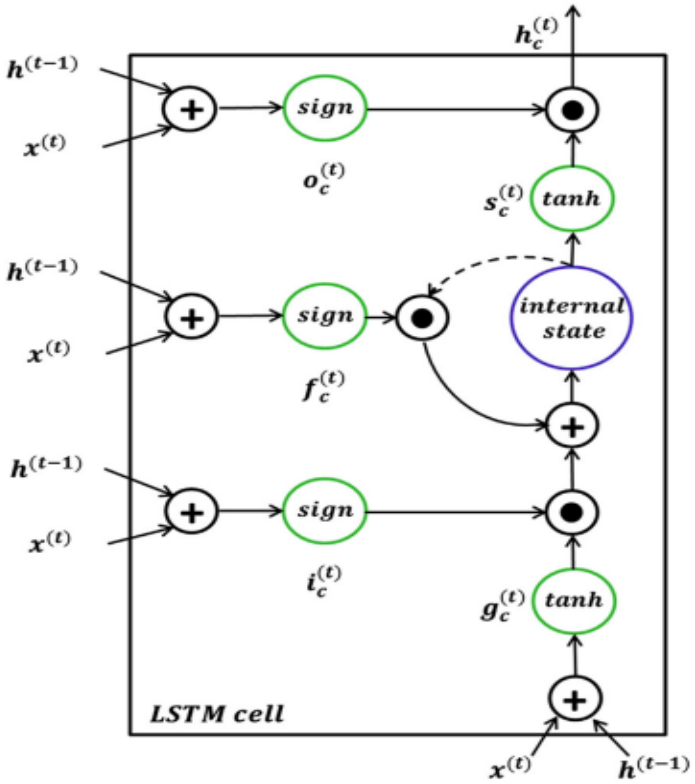

Figure 2-6 Structure of an LSTM cell

Source: Ma et al. (2019)

2.3.2.5 Bi-Directional Long Short-Term Memory

Bi-Directional Long Short-Term Memory (Bi-LSTM) was proposed by Graves and Schmidhuber (2005) as a development of the traditional LSTM. The upgrade consisted of a modelling adjustment to improve learning, whereby LSTM is applied twice to the input data (forward and backward), thereby increasing long-term training dependencies and hence increasing prediction accuracy.

In contrast to legacy LSTM studies, Siami-Namini, Tavakoli and Namin (2019) more recently experimented with Bi-LSTM in their research into forecasting time series, discovering that it outperformed LSTM and ARIMA. They recommended the use of Bi-LSTM, rather than LSTM, for time-series prediction. However, they did not test the validity of the experiment for multivariate time series, pointing out that further experiments would be needed to demonstrate its efficiency with multivariate time series. It was also stated that LSTM is relatively faster in training, because Bi-LSTM trains smaller batches of data, since input data is used for training. Furthermore, the training method for Bi-LSTM is eponymously bi-

directional (i.e., from left to right, and then from right to left), which differentiates it from other comparative algorithms.

Recent forecasting research has compared modern ML with time-series algorithms such as LSTM, stacked-LSTM, bidirectional-LSTM networks, XGBoost, and an ensemble of Gradient Boosting Regressor, Linear Support Vector Regression, and an Extra-Trees Regressor. As complex statistical models have become expensive, time-series algorithms are taking over, and stacked LSTM and bidirectional LSTM have been found to perform better than other modern machine-learning algorithms (Barrera-Animas *et al.*, 2022) (see Figure 2-7). Transferred Bi-Directional Long Short-Term Memory (TL- Bi-LSTM) has also been proposed as a model for AQP (Ma *et al.*, 2019). This method uses Bi-directional LSTM to learn from long-term dependencies, while also learning from small temporal resolutions and passing this learning on to large temporal resolutions. Key researchers into Bi-LSTM encouraged more research into such models, to improve prediction performance (Ma *et al.*, 2019; Siami-Namini, Tavakoli and Namin, 2019). Most of the existing literature in this area reported that machine-learning methods were not reliably accurate when dealing with Big Data. This is where research is especially required, to provide efficient methods which will allow models to produce high accuracy, particularly when dealing with large datasets (Tealab, 2018).

2.3.2.6 Markov Chains

'Markov' is a linear statistical method has been named after the 20th-century Russian mathematician Andrey Markov, who worked on development for the Markov processes after the known Poisson process (the process in continuous time) (Chen and Wu, 2020). Markov chains were already widely used in time series studies concerning the variability of events over time by the late 20th century, as an extension of the generalized linear models. A Markov-switching model was presented in 1988-89, as an extension to work on state-space representation, and a Markov-switching dynamic regression model was used to model the dynamic behaviour of time-series variables, with switching represented by Discrete-Time Markov Chain (DTMC) objects (Kim, 1994; MacDonald and Zucchini, 2016).

Statistical linear methods have previously been used for air-pollution time-series prediction. However, over the past decade, research has been carried out into the use of other ML methods for AQP, whose complex nature and variations in data sources and parameters foments the urgent need for more accurate measurement methods demonstrates, and thus further research (Ameer *et al.*, 2019). Time series create several ways of forming models, while discrete-valued time series are required in many applications (e.g., to show the sequence of events, number of defective items in a particular set-up, or number of cases of

a disease in a given area and time). In discrete models, it is important to rely on the discrete nature of the data when building the distribution, so that normal distribution should not always be chosen (Ameer *et al.*, 2019).

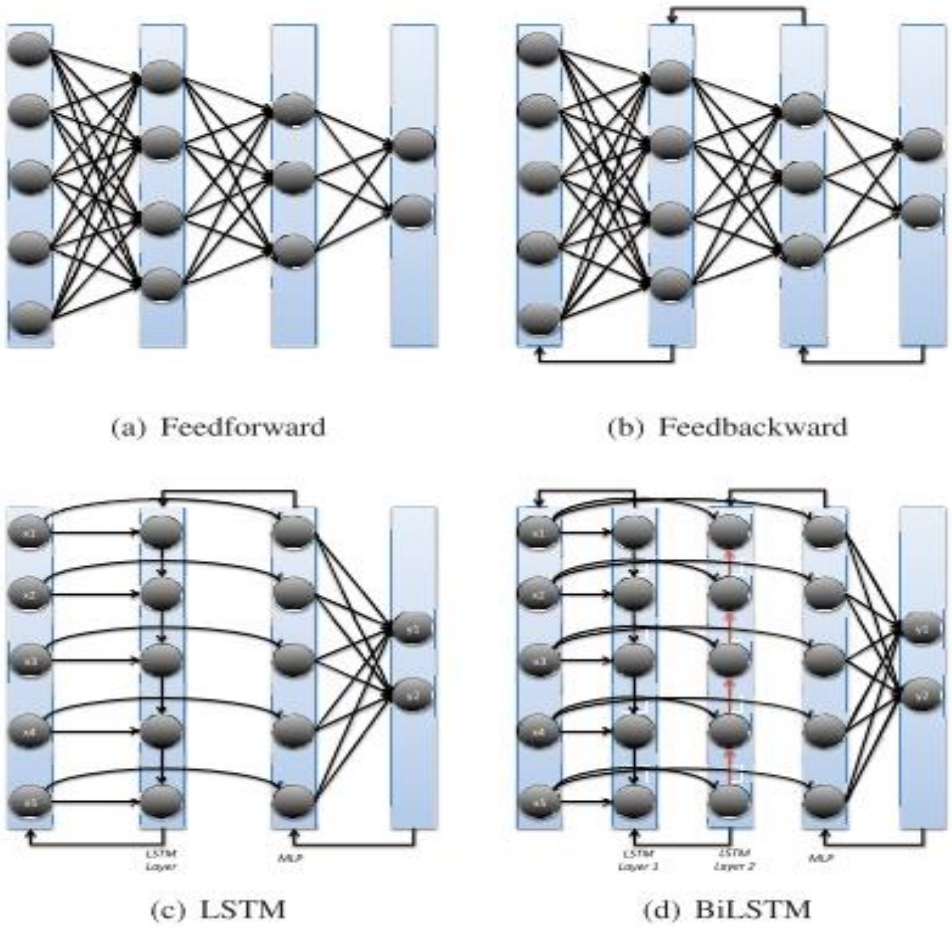

Figure 2-7 LSTM and Bi-LSTM representation

Source: Siami-Namini, Tavakoli and Namin (2019)

DTMC has been adopted for the prediction of air quality in several fields, and it offers advantages in estimating the probability distribution and analysing probabilistic behaviours for some applications. It has been used in many environmental domains and applications. While providing a simple linear methodology for predicting, Markov discrete time (or the stochastic process) is 'memory-less', and depends on the current state only in transitioning to future states. It is also less time-consuming in comparison with other ML methods, improving its performance value (Chen and Wu, 2020).

MacDonald and Zucchini (2016) proposed Markov-switching vector autoregressive (MS-VAR) model as a solution for non-linear time-series models. Finite-state Markov chains form a discrete-time stochastic (independent) process transitioning from state to state; the prediction depends on the immediate past (i.e., the sequence of previous states) (Kemeny

and Snell, 1983). Markov neglects past information but uses the outcome of the most recent experiment to predict the future. This is a process described by moving from state to state with the transition probability. The Markov switching dynamic method starts with defining states or thresholds as the beginning point, then building the transition matrix (state transitions and probability). Figures 2-8, 2-9, and 2-10 represent the observed frequencies of transitions from one state to another. The Markov-switching model is represented by the transition probabilities.

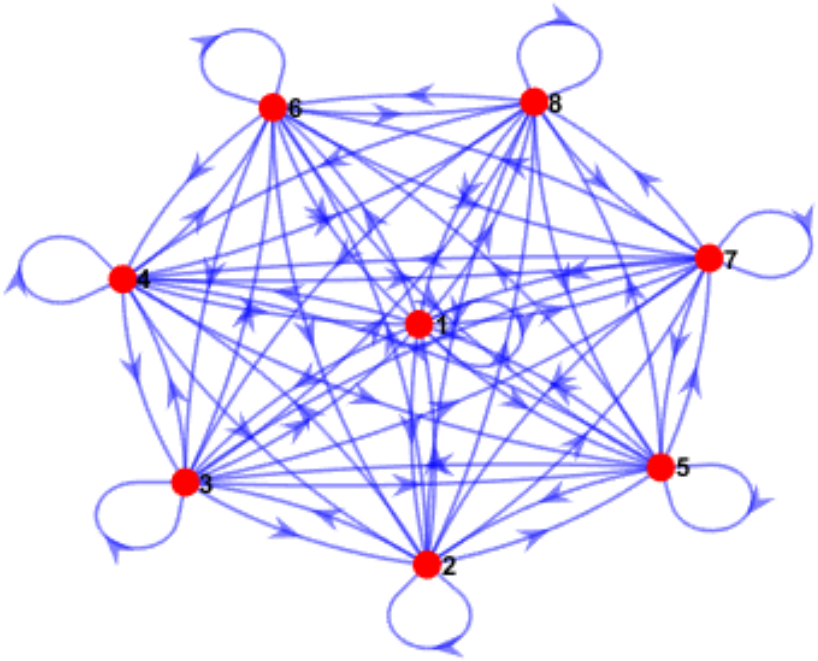

Figure 2-8 Eight (8) states transitions – MATLAB illustration

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 (Wang *et al.*, 2019). There is a gap in the existing literature in using Markov chain model development for forecasting air quality (Zakaria *et al.*, 2019), notwithstanding studies of Markov chains used in several domains to predict long-run behaviour and determine efficiency. Markov is very commonly used in stock prediction, but Markov chain studies in air quality analysis remain very limited in comparison to the more prolific ANN and DNN research on AQP (Zakaria *et al.*, 2019).

Following the generalized model of Hamilton (1989), introducing the Markov-switching model for time series analysis, 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 modelling (Kim, 1994). Markov-Switching Vector Auto Regression (MS-VAR) is a generalized form of the VAR model, consisting of a serially distribution

independent regime. This framework represents the probability of the states (transition between the states) based on unobservable regime variable *st* for the observed vector *yt* (Hamilton, 1994). Lam's approach as used by Kim (1994) is an extension that is generalized from the Hamilton (1994) model, which proposed estimation using the sum of previous states as an additional state variable.

As an extension of Hamilton's Markov-switching model and others, different approaches have been used to satisfy the different capabilities of Markov chain theory (Kim, 1994). A Markov system can be described as a set of N states: S1, S2, S3……SN. 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 (Eq. 2.3, Eq. 2.4) (Rabiner, 1989):

$$P[qt=Si|\ qt-1=Si,qt-2=Sk,\ldots] \quad (2.3)$$

$$P[qt=Si|\ qt-1=Si] \quad (2.4)$$

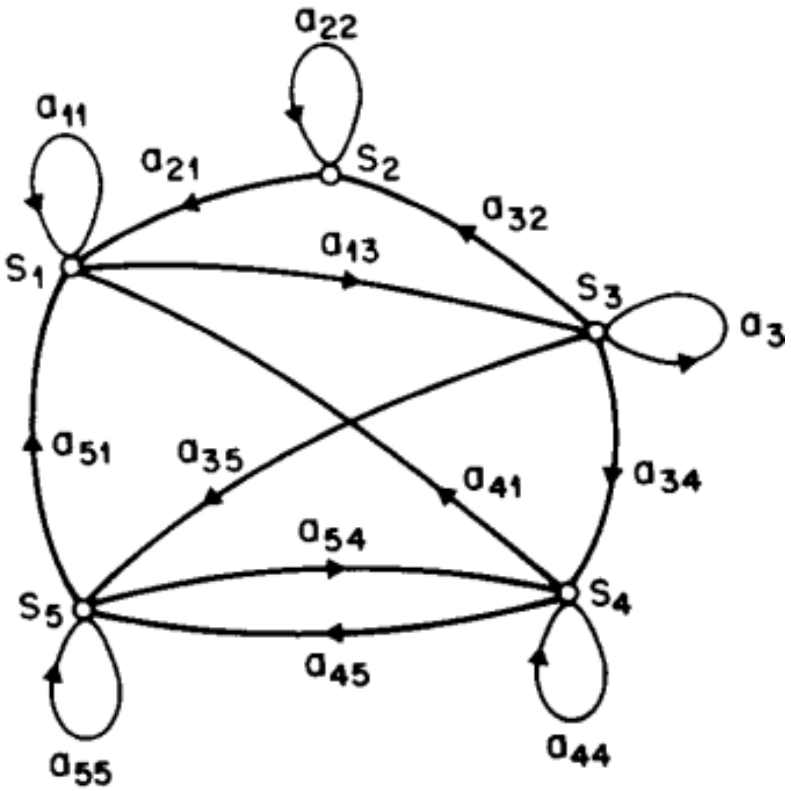

Figure 2-9 A Markov chain with 5 states (labeled S, to S,), with selected state transitions

Source: Rabiner (1989)

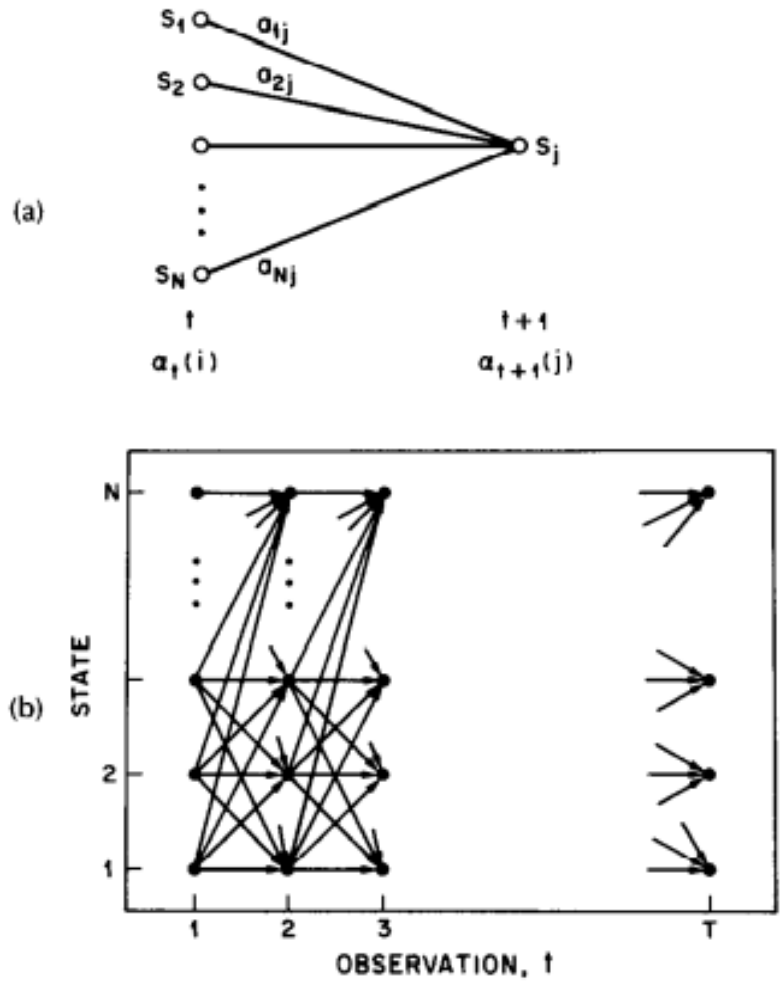

Figure 2-10 Illustration of observations t, and states i

Source: Rabiner (1989)

2.3.2.7 Hybrid Models

Hybrid models were already subject to intensive research by the 2000s, and subsequent innovations in model selection and connections have emerged. The dynamism in hybrid model studies reflects the flexibility and ease of developing such models, with great scope for enhancements and developing new architectures to support different domains and prediction objectives (Liao *et al.*, 2021). So this space of hybrid modelling is very revolutionary with forthcoming developments. A 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 ML have been proposed as advances on traditional methods. Big data and ML approaches have been used widely to predict air quality. Nevertheless, examples of highly accurate AQP methods for Big Data considering temporal resolution are limited in the existing literature (Ma *et al.*, 2019).

There are several examples of research into air-quality evaluation which use machine-learning algorithms with various ML models to predict air quality, as discussed later in this chapter (see Tables 2-6, 2-7, 2-8, and 2-9). Big Data has formed a way to model more dynamic air-quality systems which are behaviourally heterogeneous, taking data from various sources. Many algorithms, methods, and techniques have been used in air-pollution modelling (Alkasassbeh *et al.*, 2013). 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 (Zheng *et al.*, 2015).

Based on several air quality evaluations and after surveying the existing literature (see Table 2-7 and Table 2-8), 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 particularly useful for decision making purposes. 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 (Zheng *et al.*, 2015). A review by Liao *et al.* (2021) highlighted that PM, NO_x, and O_3 have been of particular interest to researchers, and pointed out that ANN was the preferred way to study PM and O_3, while Land Use Regression (LUR) is broadly used in studies of NO_x. Furthermore, the study noted the emergence of more hybrid techniques as a growing trend, with rising challenges for multiple pollution prediction in light of pollution interactions. $PM_{2.5}$ and PM_{10} studies are increasing, and the growing number of fine PM studies have typically used ANN.

Moreover, a systematic literature review by Zaini *et al.* (2022) explained deep learning strategies and possible methods for AQP, considering DNN, RNN, convolution NN (CNN); hybrid and multiple deep learning methods; and ensemble learning. The review pointed out that hybrid models were developed based on different setup of data and other requirements. The review reported that most studies using hybrid models appeared to choose CNN-LSTM, which supports long-term time series data at the expense of a complex process, due to the large datasets required. The study recommended the use of hybrid metaheuristic algorithms and deep learning, and more research directed to this field is expected to contribute to generating new findings and driving theoretical development.

2.3.3 AQP Methods and Techniques

2.3.3.1 Feature Selection and Data Pre-Processing in Machine Learning

There are several methods for features selection and data pre-processing used in the literature, and numerous associated techniques (see Table 2-7, 2-8, and 2-9). The air quality domain has been of interest for many researches across numerous disciplines, and features selection for any proposed prediction for the field is of primary concern. The nature of air quality and gaseous activity in local and global atmospheres is inherently complex, and feature selection should be done in light with the factors that affects emissions concentration in air. The most important step is to determine the factors impacting the concentrations that influence the prediction process, and such factors should be included in the study of the air quality forecasting based on the aims and objectives. Accordingly, pre-processing should be done based on the geography, locations, frequency of data and many others. However, few studies have been concerned with the identification of effective parameter in the prediction based on statistical methods. Baldasano, Valera and Jiménez (2003) performed a study with the aim of selecting the best statistical model and refinement method for air pollution and metrological data, in order to predict missing data, filter noise, and determine the most influencing parameters in air pollution prediction. They compared selected cities in developed and developing countries by evaluating air quality values and comparing these to air quality guidelines (European and WHO limits), and their analysis showed that there is general decrease in the worldwide pollutants concentrations. They claimed that this was mainly due to the restrictions applied by governments, and national and international organizations. However, they claimed that this is not generally the case for developing countries, as pollution remains high, with expected trend of possible increase in ground level concentrations.

The data measured for the two-year period study (2001-2002) by the National Air Pollution Monitoring Network randomly divided the dataset into training, validation, and test sets for modelling purposes (development, evaluation). The training of the NNs was conducted using the bulk of the dataset (3/4) and the remaining cases were equally divided to validation (improve generalization ability) and test (statistical comparison of the performance of the different models) sets (Grivas and Chaloulakou, 2006). Three NN models were developed for each station, the first of which had a full set of input variables; the second used variables selected by a genetic algorithm optimization procedure; and the third did not use meteorological input variables. The NN architecture used was feed forward, which is able to approximate every measurable function. Grivas and Chaloulakou (2006) pointed out that the results were satisfactory, and provided superior results compared to multiple linear

regression models. They claimed that NN is particularly germane to PM_{10} predictions, given its characteristic of being a complex gas to predict (see Table 2-9).

Rybarczyk and Zalakeviciute (2018) studied analysed ML prediction models' parameters, for the scenarios adumbrated below. (1) Estimation models using 'predictive features as contaminant covariates, meteorology, etc.'. (2) Forecasting models using 'historical data to predict pollution concentration)'. (3) Different types of ML algorithm used with main categories (ANNs, SVM, ensemble learning, regressions, and hybrid versions of these algorithms). (4) Methods used by authors/ (5) The nature of the predicted parameters. (6) The geographical location where studies were performed. (7) Dataset details (timespan, quantity of monitoring stations, and number of instances). (8) Specific information about the dataset regarding used predictive attributes such as (pollutant covariates, meteorology, land use, time, human activity, and atmospheric phenomena). (9) Evaluation method. (10) Tested algorithm's performance (compared in terms of models' accuracy and/or prediction of actual values). (11) The computational cost of the method (Rybarczyk and Zalakeviciute, 2018).

2.3.3.2 Air Quality Prediction Architectures and Optimization Techniques

Hybrid deep learning and algorithms optimization are also advancements in deep learning studies, and accuracy improvements are strongly called for (Zaini *et al.*, 2022). Architectures for deep learning are varied, with the most commonly used being DNN, RNN, CNN, and some others, such as Deep Belief Network (DBN) and Deep Boltzmann Method (DBM), with numerous deep learning strategies. Improving accuracy and reducing error for models requires significant tuning for hyper-parameters, which plays a vital role in models' performance. This can entail a prolonged process of experimenting tweaks for different attributes as required for the model design and architecture. A systematic review of time series forecasting by Tealab (2018) was undertaken in order to check the performance of NNs, and it was reported that many of the studies measured used hybrid models (adding the residuals of the linear model as an input to the ANN model) as a technique.

Niharika and Rao (2014) noted that Transfer Learning can be used for further improvements, in case of in inadequate data faced during learning, as this method was found to yield lower error rates across several studies. There are many forms of creating hybrid models for forecasting; however, choosing suitable architecture and strategies for building models is crucial for optimal forecasting. There are many ways of combining DNN (LSTM) and Markov models (hybrid models), for instance to improve predictions. Markov trained on LSTM offers 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.

Rakholia *et al.* (2023) used different deep learning topologies to assess the possibility of accuracies so they could improve predictions. They tested deep learning models using different combinations (i.e., GRU + GRU, LSTM + LSTM, RNN + RNN, GRU + LSTM, GRU + RNN, LSTM + RNN, LSTM + GRU, RNN + GRU, and RNN + LSTM), developed and run for ten times. They concluded that LSTM+LSTM gave the best results in comparison to other explored architectures. Further to the experiment, Rakholia *et al.* (2023) searched the literature for similar work, compared methodologies such as CNN-LSTM and TL-BLSTM for $PM_{2.5}$ predictions, and reported that LSTM-LSTM outperformed these alternatives. A multistep multivariate prediction model was developed, which was posited as a global model for predicting several gases emissions without the need for multiple models and it is related complications. A global forecasting model was suggested; N-BEATS NN based to forecast multi pollutant simultaneously. N-BEATS consist of a deep stack fully connected layers with backwards and forward residual links, the study developed a multivariate multi-step (multiple time series with different covariate features).

LSTMs can be designed in several forms, and it is very important for implementing algorithms with appropriate results to do the configuration exercise. Navares and Aznarte (2020) conducted a comprehensive experiment to determine an efficient and accurate LSTM model. They noted that the number of hidden nodes to be used in LSTM cannot be specified using known techniques, which poses challenges due to designing LSTM models based on trial-and-error. The study ran a Fully Connected LSTM (FC-LSTM), which is the most common type of LSTM, along with Gaussian Process LSTM (GP-LSTM), whereby LSTMs are grouped to link with a specific selected pollutant. Input GP-LSTM (IGP-LSTM) is another suggested architecture, using an input assistant for each group in the network. Inputs are grouped by pollutants with a fully connected output layer. Furthermore, a simple Shared-Private LSTM (SP-LSTM) with all inputs to be used to train the network and outputs one single variable.

Figures 2-11, 2-12, 2-13, and 2-14 illustrate each suggested architecture. Navares and Aznarte (2020) concluded by assessing different topologies for air quality forecasting in the Madrid area, and the statistical results showed that IGP-LSTM and GP-LSTM outperformed the other studied algorithms. It was noted that there were differences between pollutants behaviours' in different circumstances, such as locations, and this was considered in the study dimensions.

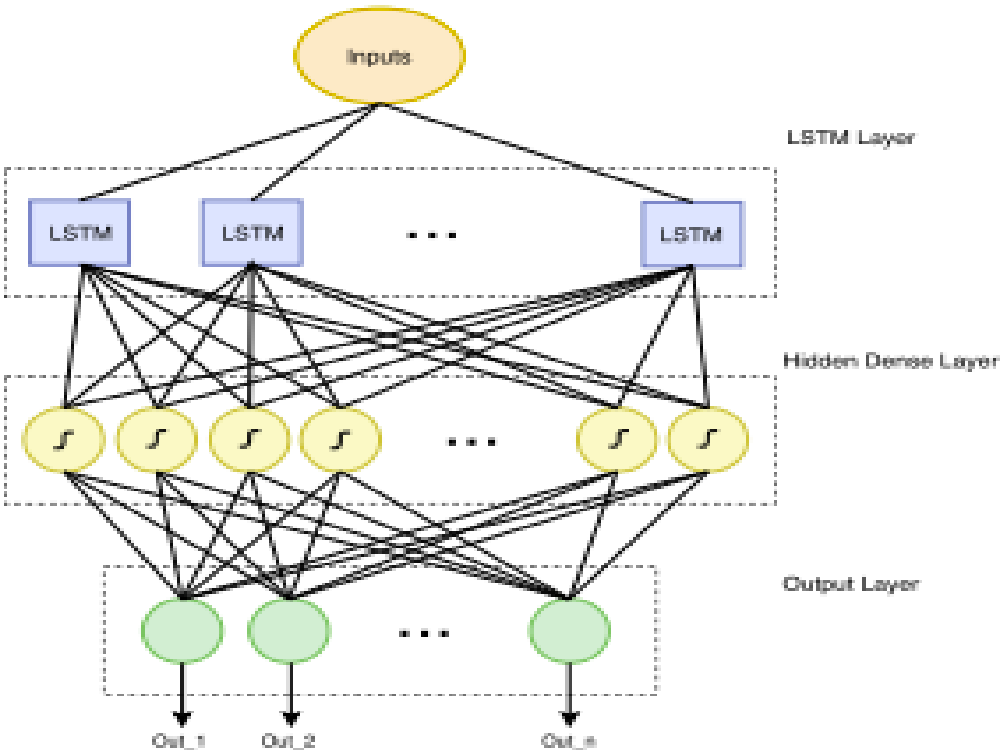

Figure 2-11 FC-LSTM architecture

Source: Navares and Aznarte (2020)

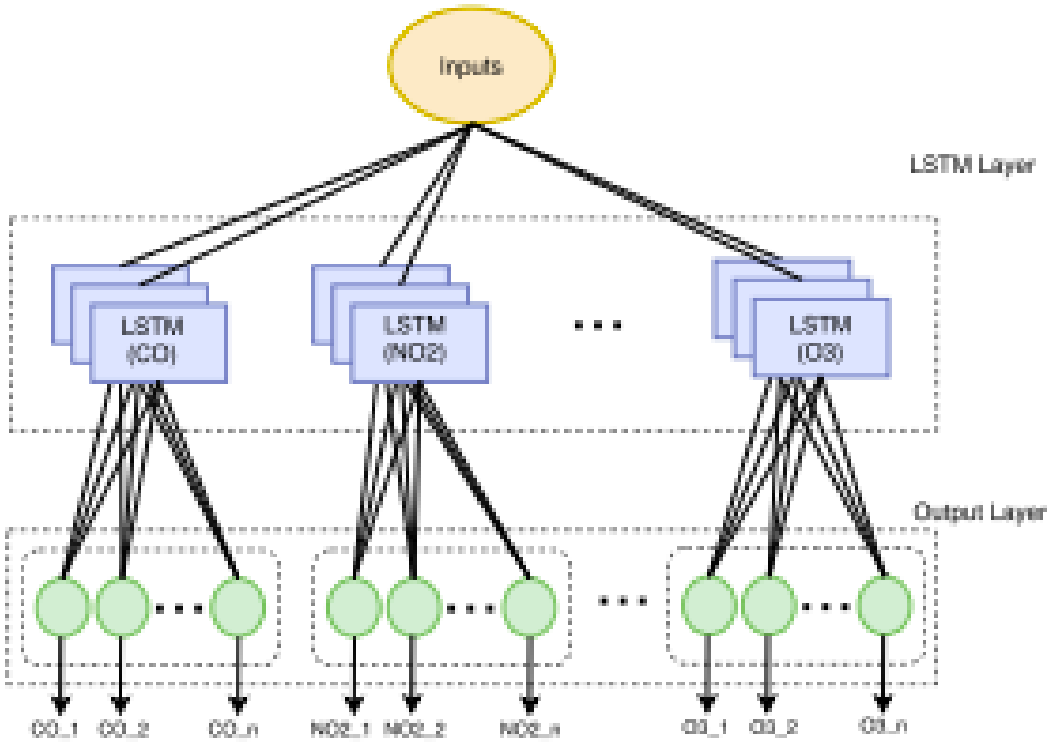

Figure 2-12 GP-LSTM architecture

Source: Navares and Aznarte (2020)

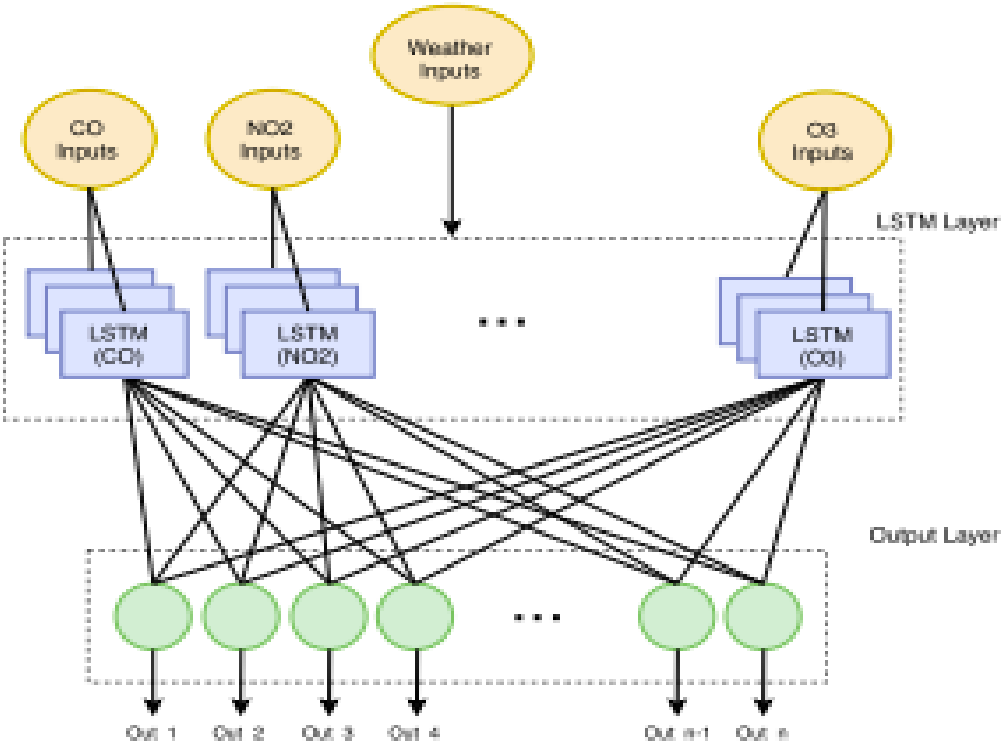

Figure 2-13 IGP-LSTM architecture

Source: Navares and Aznarte (2020)

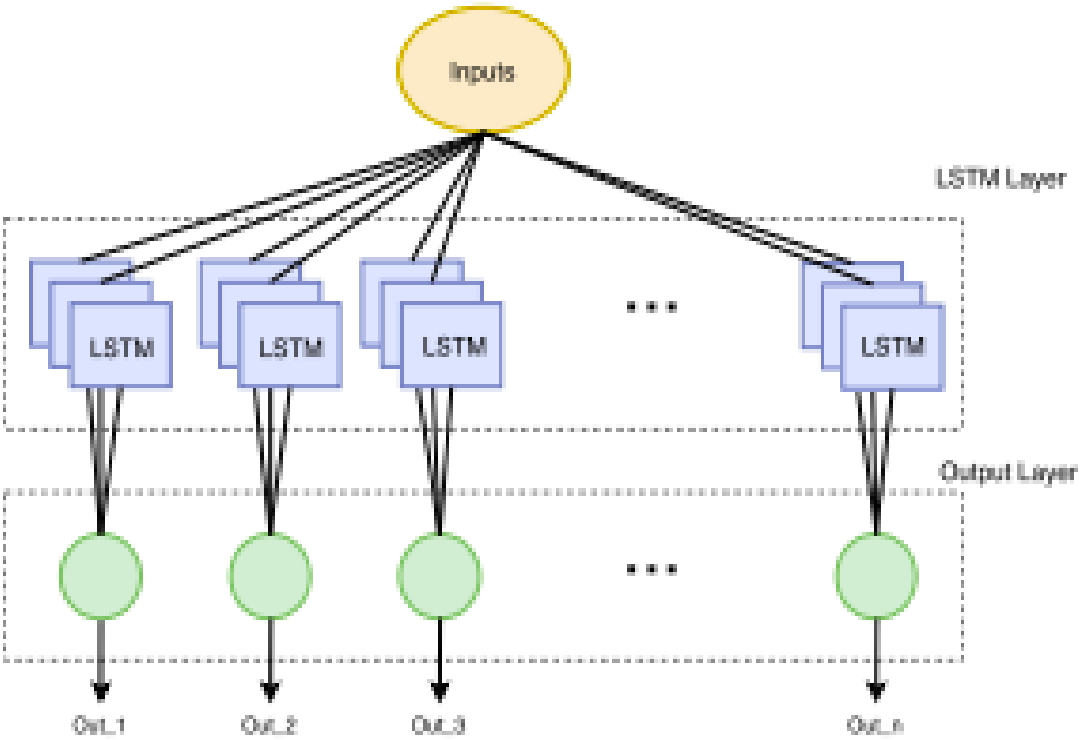

Figure 2-14 SP-LSTM architecture

Source: Navares and Aznarte (2020)

2.3.3.3 Hyper-Parameters Tuning and Optimization Techniques

Hyper-parameters selection and optimization is one of the appealing topics when discussing ML methods. It is necessary to tune hyper-parameters so they optimally fit the data and model requirements, to achieve suitable or desired accuracy (see Table 2-5). Hyper-parameters include batch size, optimizer, loss function, number of hidden layers, number of hidden neurons in each layer, learning rate, models used in the layers, dropout rate, epoch, batch size, activation function, train size, number of hours, seed, input shape, duration (Eren, Aksangür and Erden, 2023). All of these parameters are of great interest in related studies to build the architecture of the algorithm and for optimization (Ma *et al.*, 2019; Hu *et al.*, 2023).

In relation to explorative deep learning techniques, Zhou *et al.* (2019) considered that using L2 regularization, dropout neuron, and Mini-Batch Gradient Descent (MBGD) algorithms would be useful for overcoming over fitting problems. Optimal processing time and avoiding over-fitting is fundamental for model design considerations, and DNN architectures are dependent on the number of hidden layers and neurons. However, other factors are also of significant importance for performance, including learning rate, activation function, optimizer algorithm, and others that can be tuned to optimize model performance and accuracy (Zhou *et al.*, 2019).

Deep residual learning can resolve some of the problems faced when using DNN. Moreover, oddball gradient descent proposes a methodology in which the training examples are proportional to respective errors for training data: the higher the error, the more occasions are needed to train the data. Adaptive learning rates are very important for DNN training: adaptive methods such as Delta-bar Algorithm, AdaGrad, RMSProp, and Adam can alter learning rates and increase training speed (see Table 2-5). Batch normalization also improves the accuracy of DNN, as well as decreasing over-fitting chances. The dropout method is increasingly used alongside regularization techniques, to eliminate issues of over-fitting (Shrestha and Mahmood, 2019).

Table 2-5 Hyper-parameters highlights

Parameter	Functions	Remarks
Activation Function	Rectified Linear Units (ReLU)	Activation function, mostly used in deep learning models ReLU variants: Exponential Linear Unit (ELU), Parametric ReLU, LeakyReLU
	Linear, sigmoid and hyperbolic tangent	Other activation functions used for deep learning models
Optimization Algorithm	Gradient descent (GD) Adaptive gradient algorithm (AdaGrad) Root mean square propagation (RMSProp) Adaptive moment estimation (ADAM)	GD variants: Stochastic Gradient Descent (SGD), Mini-Batch Gradient Descent (MGD) Considering different types of optimizer algorithm yields different forecasting results

2.4 Model Performance Measures

There are several methods used in evaluating models performance in AQP domains. A considerate number of studies considered regressing as main evaluator for data relations, also studies mostly showed RMSE as a main evaluator measure for prediction results as error measure.

Root Mean Squared Error (RMSE) can be represented using the following formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)} \tag{2.5}$$

Other standards methods have been also used by researchers, such as Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) which are represented by the below equations:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.6}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.7}$$

2.5 Literature Review on Air Quality Studies and Applications

2.5.1 Searching Process

The literature review was performed over three primary stages. The first round consisted of collective results from searching academic databases (Google Scholar, IEEE Xplore,

JSTOR, Web of Science, Scopus, and ScienceDirect) to find articles, journals, and workshops, etc. concerning the key search words (e.g., 'Big Data', 'machine learning', 'air quality', 'transportation', 'air pollution', and 'emissions'). Database hits were filtered based on progressively screening their titles and abstracts and then reading full-text versions, until only studies whose content included real representation of ML prediction models from several countries remained.

In the second stage, systematic review journals/articles for ML and air quality were analysed to compare wider ML prediction studies based on several parameters, including accuracy, platform used, data size, method/technique, index of quality, and data span (e.g., hourly/daily).

Third stage of refinement restricted the scope to studies published during 2018-2020, to include the latest trends, and including the keywords 'machine learning' and 'air quality' (to ensure direct relevance to this research).

After these three stages, the remaining high-quality, relevant studies pertaining to the comparison of ML models and methodologies for air quality analysis were read multiple times, and were subsequently reviewed and compared in depth, and the resultant analysis is summarised in tables expounded throughout this chapter. Updates to the literature search were subsequently conducted, to include more recent years, as the subject was developing with time and new publications were checked. Different sets of combinations for keywords were considered as well, using high indexed publications. The search strategy included the main keywords 'air quality', 'prediction', 'forecasting', 'air pollution', 'air pollutant', 'machine learning', 'deep learning', 'neural network', and 'modelling', using different combinations and Boolean operators, to case a broad net.

2.5.2 Discussion

Rybarczyk and Zalakeviciute (2018) conducted a high quality systematic review that systematically compared algorithms using different parameters (see Table 2-9). They selected and analysed recent ML studies (journal articles) in pollution research, to identify ML algorithms for predicting air quality. The review reported that the most commonly used algorithms were (in descending order): ensemble learning methods, ANNs, SVMs, and linear regressions. As can be concluded from analysed studies (as displayed in Table 2-9), air pollution forecasting has been extensively undertaken using MLR, ARIMA, Support Vector Regression (SVR), Random Forest (RF), and K-nearest neighbours (KNN). Different types of ANNs contributed to the field of forecasting, including MLP, CNN, RNN, LSTM, Gated Recurrent Unit (GRU), and Encoder-Decoder NN (EDNN).

CNN was found to improve performance in some AQP studies; for instance, Alzubi, Nayyar and Kumar (2018) compared three different RNN architectures to predict hourly $PM_{2.5}$ concentrations. Various metrics (MAE, MSE, RMSE, R2, and MAPE) were compared, and the experiment showed that CNN offered the best performance, followed by LSTM and then GRU. Nevertheless, as deep learning continues to contribute to the field, DNN proved to have a competitive edge in performance over some other algorithms used in air quality studies, as explained in Table 2-9, especially for complex settings, wherein the DNN architecture supports different capacities.

As explained previously, researchers have cautioned interpretation in relation to the complex nature of emissions in the real atmosphere; this has created dilemmas for the best fit algorithms to be used to ensure coverage of such complexities. While the received wisdom affirms that there is no one ideal solution for accurate AQP, existing literature has encouraged continuous research in the field to identify relatively more accurate options, specifically in light of the fast growing field in ML and the new methods that could be used in the future (see Table 2-9).

Some gases were found to present particular challenges for researchers, such as O_3 and PM, due to their nature and behaviour in the atmosphere. For instance, the quick exchange of O_3 between the upper layer and the surface creates complexities using statistical methods. MLP and ANN can be used to solve such challenges related to some gases characteristics; however, due to the ANN structure, it could be not always helpful to use it, if accurate performance is a priority, and using other additional methods is advisable in order to corroborate ANN outputs (Liao *et al.*, 2021). Even with the many applications NNs have been used for, and the continuing recommendations to use them as very applicable algorithms, consistent negative issues have been discovered with their use, including over-fitting and generalization. While such issues have not been resolved even after several trials and studies, it has also been stated that some of them were solved using some other methods (Siami-Namini, Tavakoli and Namin, 2019).

For instance, a very early work by Grivas and Chaloulakou (2006) screened researches for hourly concentration prediction for PM using ANN, and the results showed that ANN models outperformed linear regression ones, and that the use of meteorological predictors enhanced their performance. However, ANN did not show the same performance in predicting hourly NO_2 concentrations. The study suggested the implementation of several methods to compare results is required; in line with this the current literature (see Table 2-7, 2-8, and 2-9), there is a need to develop several methods for AQP, due the challenging nature of

gases, and the implementation of mixed methods is encouraged to address the existing gaps in AQP.

As linear modelling is presented as important field in the existing literature, there is limited research on such dimension of studies for air quality modelling to solve some challenges in predicting some types of gases. The significance of such forms of modelling has been highlighted in a forecasting study of off-shore wind power fluctuations, which adopted Markov-Switching Autoregressive (MSAR) models, using the regime-switching feature to represent the fluctuating nature of wind power data. Although the field has been presented with several challenges and limitations, it has been thought that setting parameters for such models entails severe gaps, and estimation is fundamentally challenging. However, some solutions to the challenges faced can be considered viable Pinson and Madsen (2012).

In the MSAR model, the coefficient can vary slightly over time, and maximum likelihood estimation is used. Unlike stationary MSAR models, where coefficients do not change over time, non-stationary MSAR is built by maximizing the likelihood of the dataset observations. Chen and Wu (2020) conducted an empirical case study of AQI prediction in Taipei city. They presented a dynamic prediction applying DTMC in a process to predict AQI short-term value and identify primary pollutants in the area. The authors recommended that the outputs of their study be used as a base to form more comprehensive air quality controls, albeit they acknowledged that their findings were limited in scope. They encouraged further studies from different regions, to compare the inconsistencies of air quality states in urban environments in different climate conditions.

The review (Table 2-8) took into consideration several factors when studied, and it showed that air quality research consists of many factors. When building models and choosing algorithms, a careful comparison and evaluation of selected factors should be undertaken, including the following:

- Parameters (inputs and outputs)
- Data size
- Data frequency
- Data structure
- Data relations (linear or non-linear)

For building models of suitable accuracy for the studied domain, some consideration should be mapped such as:

- Model type

- Features selections
- Model architecture
- Optimization techniques
- Hyper-parameters
- Model integrations

Observations arising from the analysis of the literature reveal that several approaches can be followed for AQP modelling, and in most cases there is no single method that fits all air quality domain prediction modelling requirements. There must be a high emphasis on the aims and objectives of the research or application when selecting models, and equally the data related to the study should be of high consideration when building models. Different model topologies should be considered, and more specifically integration methods (hybrid modelling), which is of increasing research interest. Hybrid modelling can create extremely large numbers of possibilities between different model types, architectures, topologies, and the number and ways in which different models can be integrated to offer more effective solutions.

The identified gaps in existing literature arising from the analysis can be summarized as follows, which are addressed by the current research:

- Limited research used hourly data frequency when doing AQP.
- Limited research explored multi-inputs and multi-outputs when studying air quality parameters for AQP, to model the real complexity in the atmosphere. Many factors must be present in the prediction, collectively with other gases, as this is more realistic for scenario setups.
- Limited research used Markov chains, specifically the Markov Switching Dynamic Regression model type, in light with multi-inputs and multi-outputs parameters
- Limited research mapped the complexity of the nature of some gases, and there are very limited suggestions for addressing this limitation in light of the multiple instrumental factors at play (such as the presence of other gases and meteorological conditions).
- Limited comparative studies have compared cities or areas with multi-input and multi-outputs parameters.

The conducted literature review has explored several areas in relation to the problem identified, and the aims and objectives of this study. This research studies the existing literature for the methods that were used by other researchers in air quality domain and the

literature review was developed in several parts, to support the contextualization of the current study in relation to the findings of existing literature:

- Prediction methods:
  - Non-linear ML methods for prediction
  - Linear ML methods for prediction
  - ML prediction architecture
  - ML optimization techniques

The literature review identified several gaps and areas worth covering and discovering. This research addresses some identified areas, and formulates pathways for future advancement and development. It should be highlighted that there is very limited research addressing transportation factors and impacts on air quality (especially in Asia, and specifically for this wok where Jordan was studied). It has been reported that processing air quality data is complicated for several reasons. First of all, in both developed and developing countries the available data are very limited, and only a few countries publish related information (Delavar *et al.*, 2019).

The transportation sector undoubtedly plays a massive part in air quality indicators worldwide, and accordingly affects overall global AQI, as well as local indices for each country. Therefore, there are regional and worldwide efforts to study the bottom line of transportation, to try to come up with solutions in several directions to improve air quality, as well as to reduce traffic and related consequences, including GHG emissions. The literature has clearly stated the lack of meteorological data, specifically for humidity, for many cities worldwide. On the other hand, it has been reported that processing air quality data was not easy for several reasons. In Africa, data are very limited, and only a few countries publish pertinent and reliable information (Baldasano, Valera and Jiménez, 2003). There is limited research using hourly air quality data for prediction, and very few addressing transportations.

Table 2-6 ML methods insights from the literature

Method	Advantages	Drawbacks
ANN	Complex nonlinear relationships between the concentration of air pollutants and the corresponding meteorological variables. ANN models have the ability to capture the highly non-linear character of those processes serving wide range of gaseous prediction NN models has been used for PM mass concentrations predictions (it is known that PM concentration modelling is more complex compared to the forecasting to gaseous pollutants due to the complexity of the processes). Able to follow both linear and non- linear patterns.	-Local minimum and poor generalization, lack of analytical model selection approach, time consuming in finding best architecture and its weight by trial and error. -Computationally expensive
MLR	Complex model for prediction.	Data noise can affect regression based results negatively.
SVM	Robust and reliable prediction results, as well as handling multidimensional dataset with small number of samples for training.	-Sensitivity to noise -Computational complexity
RNN	RNN can process sequential data and it can build sequential structure of the historical data.	-Training instabilities -Vanishing and exploding gradient issues

Table 2-7 Literature review (initial round performed-2019/2020)

Article and Region	Quality Indicator	Data Size	ML Model	Parameters
PM_{10} prediction using genetic programming: A case study in Salt, Jordan (Faris *et al.*, 2014) Jordan	MSE-Training: 222.85 MSE-Testing: 212.76 MAE-Training: 11.88 MAE-Testing: 10.58	5 stations around Al-Fuhais Cement Plant 1-year (26 November 2006 to 25 November 2007)	GP tree model for genetic programming	PM_{10}, temperature, relative humidity and wind speed

Table 2-7 Literature review (initial round performed-2019/2020)

Article and Region	Quality Indicator	Data Size	ML Model	Parameters
A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran (Delavar *et al.*, 2019) Iran	RMSE R2 NARX gives minimum error and the best determination coefficient, minimum time calculation	January 2007 to January 2011	SVR, NARX, ANN, and GWR	O_3, SO_2, NO_x, CH4, total hydrocarbons (THC) and meteorological data (air pressure, temperature, wind speed and direction, and air humidity)
A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study (García Nieto *et al.*, 2013) Spain	Correlation coefficient for gases relationships	January 2006 to December 2008	SVR	CO, NO, NO_2, SO_2,O_3, and PM_{10}
Time series forecasting of air pollutant concentration levels using machine learning (Patra, 2017) Italy	RMSE	From March 2004 to February 2005 (1 year)	ANN, SVM, ARIMA	CO, NO_2 plus temperature and relative humidity
Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing- Tianjin-Shijiazhuang Bing-Chun (Liu *et al.*, 2017) China	MSE RMSE MAE MAPE	Daily data (January 1, 2014 to April 30, 2016)	SVR	$PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3,CO minimum temperature, maximum temperature , weather , win direction and wind power

Table 2-7 Literature review (initial round performed-2019/2020)

Article and Region	Quality Indicator	Data Size	ML Model	Parameters
A deep recurrent neural network for air quality classification (Zhao *et al.*, 2018) China	Notations for CP and IAQI-classification measures (otherwise not mentioned in the paper)	Training data from January 1, 2010 to December 31, 2014 Test data from January 1, 2015 to December 31, 2015	RNN model, SVM and RF	CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10}
Forecasting fine-grained air quality based on Big Data (Zheng *et al.*, 2015) China	Mean of the predicted maximum and minimum values against the mean of the truth AQIs during the interval, also calculating the absolute error of each time interval	Hourly data from 43 cities in China	Hybrid model (Regression, ANN)	Not specific (air quality, meteorological data and weather forecasts) Each station gases (NO_2, SO_2, O_3, CO, $PM_{2.5}$ and PM_{10})
Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong (Zhang and Ding, 2017) Hong Kong	R^2, root mean square error (RMSE)	8 air quality parameters in 2 monitoring stations for 6 years	ELM	Focus on (NO2, O3, PM2.5, NOx, SO2)

Table 2-7 Literature review (initial round performed-2019/2020)

Article and Region	Quality Indicator	Data Size	ML Model	Parameters
Prediction of PM_{10} and TSP air pollution parameters using ANN autoregressive, external input models: A case study in Salt, Jordan (Alkasassbeh *et al.*, 2013) Jordan	Quite promising; capable of achieving acceptable error level after number of iterations using three neurons in the hidden layer MSE, ED, MD, and MMRE Prediction error method based on introduction of measure of closeness, specified in terms of the Mean Square Error (MSE) error criteria	8 monitoring stations in Salt, Jordan, over a 1-year period (25 November 2006-25 November 2007) Sampling frequency of 24 hours	Two ANN-based Auto Regressive with eXternal (ANNARX) input models	Particulate Matters (PM_{10}), Total Suspended Particles (TSP), Temperature (Temp), Relative Humidity (RH), Wind Speed (WS)
Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain (Martínez-España *et al.*, 2018) Spain	RMSE < 11 μ/m^3	4 cities in the Murcia region (Spain) Real data from 4 stations for air quality measurement	RF, Decision Tree, Random Committee, Bagging and KNN	Ozone level (O_3)
ANN models for prediction of PM_{10} hourly concentrations in the Greater Area of Athens, Greece (Grivas and Chaloulakou, 2006) Greece	MAE, RMSE, and some other performance indicators	4 sites in Athens Hourly date for 2-year period (1 January 2001 to 31 December 2002)	For each station, three MLP NN models were developed. The first uses the full set of the input variables (MLPf) the second uses the variables selected by a genetic algorithm optimization procedure (GA-MLP) and the third is developed without meteorological input variables (MLPnomet).	PM_{10}, temperature, relative humidity, wind speed and wind direction

Table 2-8 Literature review comparison and highlights

Study	Study Region	Model	Remarks
Moscoso-López *et al.* (2022)	Spain/Algeciras	ANNs, standard sequence-to-sequence LSTMs and a new LSTM-based approach (LSTMNA)	LSTMNA models provide slightly better performances than ANNs and standard LSTMs. Studied parameters: CO, NO_2, O_3, PM_{10} and SO_2. Performance criteria performed: correlation coefficient (R), the MSE, MAE and the index of agreement (d).
Rakholia *et al.* (2023)	Vietnam	N-BEATS architecture	Studied parameters: NO_2, CO, O_3, and SO. Performance criteria performed: MAPE, MAE, RMSE.
Ameer *et al.* (2019)	Different locations	Regression techniques used: Decision Tree, RF, MLP, Gradient Boosting	Results showed that RF regression overperformed other techniques for pollution prediction. The experiment showed good performance with different sets with different sizes and characteristics and for different locations. Processing time for RF regression was much less than for gradient boosting and MLP algorithms. The error rate was the least for RF regression than other techniques. It performed well in identifying data peaks. Decision tree processing time was lower than for all other techniques, but with the highest error rate, and no ability to identify data peaks. RF regression can be considered as the best technique for air quality pollution for this study. Gradient boosting regression can be considered as the wort technique with high processing time and high error rate.
Tripathi and Pathak (2021)	India	CNN-LSTM	Authors claimed that CNN-LSTM gave the best performance in their study Addressed challenges and limitations of the research in deep learning field. Performance criteria performed: RMSE, MAE, MAPE, MSE, Decision Coefficient, Agreement Index, Nash-Sutcliffe Efficiency Index, and Percentage Bias.
Siami-Namini, Tavakoli and Namin (2019)	Online data from Yahoo Finances	LSTM and Bi-LSTM	Studied algorithms: LSTM and Bi-LSTM. Bi-LSTM slower in training than LSTM. Recommendation to use Bi-LSTM for time series forecasting instead of LSTM. Forecasting problems for multivariate and seasonal time series needs further research

Table 2-9 summarizes the findings of systematic reviews, encompassing a large array of papers analysed by researchers grouped under different thematic concepts. It can be

concluded from the comprehensive reviews included in the table that hybrid models have gained increasing attention in recent years due to superior performance and more accuracy. It was thought generally that there would be no intrinsic restrictions on how models could be combined; researchers have been innovative in the way to combine models together driven by the aims and objectives of the studies. Models as such depend generally on the architecture, parameters, or relevant variables, which could increase complexity, thus specifying parameters is of extreme importance in building models (see Table 2-7 and Table 2-8).

While researchers have alluded to the lack of a systematic process for hybrid models, and the possibility of generating an unlimited number of scenarios in combining different models to maximize potential, this could scale-up AQP research while potentially precluding practical application (Tealab, 2018). Hence, the complexity of selecting one method over another entails numerous academic and practical trade-offs, to manage general and particular difficulties. It is recommended to carefully study the selection of different variables of studies and evaluate existing literature models based on relatively comparable parameters, besides evaluating the proposed methods on their fit for the performed study; studying the data characteristics in exploratory data analysis can drive the study appropriately.

Some reviews (Rybarczyk and Zalakeviciute, 2018; Masih, 2019) indicated that there is limitation in using estimation ML for non-linear modelling. Furthermore, there is a noticeable debate between estimation and forecasting; as estimation methods claimed to give more precision, but forecasting is more dynamic, and there is obvious evidence from the literature that using more complex models such as DNN can address the disadvantages of simpler models, but this generally entails significantly greater computation requirements.

Existing research has clearly observed the preponderance of certain algorithms in the field of AQP, and it is thought that usage and discovery of other algorithms could create more paths for more accuracy. Moreover, the existing literature suggested that there are limitations associated with including certain pollutants in predictions (see Table 2-9).

Table 2-9: Summary of analysed surveys and systematic reviews

Time series forecasting using artificial neural networks methodologies: A systematic review (Tealab, 2018) **SYSTEMATIC REVIEW**
Studied new ANN models (2006-2016). Presented evidence that hybrid model predictions proved more accurate than traditional ANN models (such as back propagation with single hidden layers), despite the lack of a systematic process for hybrid model development. Explored some new models in terms of architecture, complexity, relevant variable selection, parameters estimation, and implementation and evaluation. *Recommended research addressing literature gaps:* Specifying criteria for relevant variable selections (on which basis to be selected); methodology development for the selection of ANN architectures; creation of evaluating models methodology that tests generalization of models
Machine learning approaches for outdoor air quality modelling: A systematic review (Rybarczyk and Zalakeviciute, 2018) **SURVEY**
46 papers systematically reviewed to determine why some algorithms are selected over others in prediction. Addressed the main need for ML-based statistical models to overcome limitations of deterministic techniques, to model non-linear relationships between concentrations with required accuracy. Provided details about algorithms and how they are applied to enhance accuracy (principles of algorithms). *Main findings:* Showed that estimation problems usually apply ensemble learning and regressions. Forecasting problems mostly use NNs and SVMs. Identified challenges in improving peaks prediction and contaminants (such as nanoparticles). Claimed that ML is mainly used in Eurasian and North America continents. *The review presents two types of studies in ML:* Estimation: pollutants concentration estimation (using ensemble learning, regression. Forecasting problems (using NN and SVM) gives priority to accuracy over interpretability. Estimation is more precise than forecasting; hence, forecasting is more variable. More complex methods such as deep learning are needed to accommodate the complexity of predicting air pollution ahead of time (days or hours), although such complex methods have a drawbacks of being very computationally demanding. The study emphasized on the suitability of ML to predict air quality. Traditional deterministic methods showed complexity to model fine PM, while ML approaches (estimation and forecasting) showed high accuracy relatively to other emission gases (however lower precision is noted for peak values). Accuracy is higher for medium and small peaks than high concentration of pollutants (high peaks). Forecasting for some gases such as CO and NOx is limited in terms of performance. Assessed models showed better performance in peak weather conditions. The study suggested future directions, developing models that enhance pollution peaks prediction, and models that improves critical pollutants such as CO and NOx.

Table 2-9: Summary of analysed surveys and systematic reviews

Machine learning algorithms to forecast air quality: A survey (Méndez, Merayo and Núñez, 2023) **SYSTEMATIC REVIEW**
155 publications were studied. Direct correlation between the most polluted and the most studied countries. Increasing trend in number of ML for pollution studies. For the studied pollutant measures, nearly half of studied papers used AQI. For air pollutant concentration, total 54 papers showed that PM2.5 is the most predicted. Pollutant features are the most used; weather variables are used very often. For ML techniques, DL are more used than regression algorithms, and hybrid algorithms include both types. The most used algorithms are LSTM and MLP. Algorithms less frequently used include CNN, RNN, GRU, and auto-encoders. The most used regression algorithms are SVR and RF. Less frequently used ones are DT, ARIMA, KNN, and Boosting. There is increasing trend for the future use of Deep Transformer Networks. Air quality and climate change have been correlated in recent studies, creating a need to develop models for early warning of climate change consequences that could be caused by air pollution (for sustainable cities and societies). There is increasing recent popularity in using Graph NNs for air quality forecasting, which could model dynamic interactions (e.g., different cities, neighbourhoods, and streets) with distance-based weights. There are recent applications for using Temporal Convolutional Networks (TCNs) specifically for PM2.5. There is a recent mention for the use of recent application of Complex Event Processing (CEP) for air quality forecasting.
Statistical approaches for forecasting primary air pollutants: A review (Liao *et al.*, 2021) **SYSTEMATIC REVIEW**
Quantitative analysis of research published between 1990 and 2018, identifying trends. In this study it was found that most papers mainly focused on air pollution and relation to health diseases, urban pollution exposure models and land use regression methods. PM, NOx, and O3 are the most studied pollutants. A preference on using ANN when studying PM and O3. LUR was mostly used in NOx studies. Hybrid methods (a combination of models) become the most used methods between 2010 and 2018. Future expectations of mixed methods of statistical predictions to predict multiple pollutants at the same time. Interactions between pollutants are a challenging part of air pollution prediction future research. There is an increasing trend for studying PM and the influence it has on air pollution. Research papers studied show that PM is the most studied emission, then NOx and O3. The most used methods are ANN, LUR, multiple linear statistical analysis, and multi-method coupling models. The work highlighted the high importance of early warning system studies. The work pointed out the increase of accuracy for AQP studies over years of efforts in the domain and discussed that there still gaps in the domains and work to be done in this regards. The work highlighted the necessity to study the interaction or relation between air pollutants, human health and the urban environment. The interaction between pollutants, in particular PM-NOx relation and PM-O3 (as main combination of interest).

Table 2-9: Summary of analysed surveys and systematic reviews

A systematic literature review of deep learning neural network for time series air quality forecasting (Zaini *et al.*, 2022) **SYSTEMATIC REVIEW**
Reviewed of the recent studies of deep learning applications for time series air quality forecasting. Combinations of multiple components that produced hybrid forecasting models were suggested in this paper for potential superior performances and improving accuracy. Hybrid models may increase the computational complexity and reduce the time efficiency of the models, which can be a downside of using hybrid models. Studied main components for deep learning (features extraction, data decomposition, spatiotemporal dependency). Various combinations of deep learning input parameters were presented for different problems requirements (different applications studied).
Machine learning algorithms in air quality modelling (Masih, 2019) **SYSTEMATIC REVIEW**
Analysed 38 studies applying ML techniques. Studied input predictors and the impact of inputs on prediction accuracy improvements. Considered the geographical locations of studies. Explored techniques applied for pollutant concentration (forecasting /estimation). Analysed algorithms applied (linear regression, NN, SVM, ensemble learning, etc.). *The study concluded the following.* ML techniques are usually used and applied in North America and Europe. Multicomponent analysis (factorial analysis) showed that estimation for pollutions were done using ensemble learning and linear regression, but forecasting commonly used NNs and SVM. The study reported that ensemble learning and regression outperformed NN and SVM for the conducted studies, noting estimation models' low variability and standard deviation. Forecasting is still very limited with NN and SVM, and other models and pollutants should be considered (NOx and SO2; currently there is more focus on PM10 and PM2.5). Suggested considering other models (such as ensemble learning or others) to improve model accuracy.

2.5.3 Summary of Findings

The in-depth review of related AQP literature reveals numerous insights pertinent to this research, including the fundamental problem of the limited amount of air quality data available for the Middle East region (Baldasano, Valera and Jiménez, 2003). This limits the contribution data can make to prediction, and the lack of access to Big Data related to air quality is also an issue. Furthermore, there are many air quality monitors errors reported in the literature that affect reading accuracy, and accordingly there is a chance of faulty data. It appears that there is limited research using Markov chain for AQP, which represents a big gap in this area. Markov chain can be used to support the dynamic nature of the air quality, and represent a simple linear method for prediction as a backup, and compensate for the errors resulting from the complexity of other models such as deep learning when combined

in hybrid structure. The use of Markov chain is considered as a major contribution of this thesis.

Moreover, there is an identified literature gap concerning multi-input and multi-output prediction for emissions, and specifically concerning hourly data, with the time factor taken into consideration. Studies (see Table 2-7) mostly considered one or two outputs to be predicted for the air quality (gases) using the same model as multivariate (mostly studies built model for one output for instance) in light of some input factors, which commensurately affects outputs.

The literature has been conducted to study the features that are used and the models were designed to contribute in a feasible way, considering possible important factors that could affect gaseous concentrations within the complexity of the atmosphere components to be addressed. Also, the existing literature showed limitation in the presented comparative geographical contexts (e.g., comparing countries or cities) in the field of AQP analysis. Another point to consider major is the lack of a unified framework to ultimately represent AQI across countries, which makes comparison for pollution levels almost impossible; hence, there is a need for a methodology to build a global unified AQI framework.

As it can be concluded from the reviewed literature that identifying features and parameters for model is extremely important, as a building block for future model development, and the architecture design or topology of such models. Factors that should be considered when modelling include data size and frequency (hourly, daily, weekly, and monthly or yearly). Further, the number of inputs and outputs is a major consideration in deciding the selecting of the suitable algorithms to be used for predictions. Moreover, when studying the literature for the models used to predict air quality evaluation, commonly used methods (such as RMSE, MSE, and MAE etc.) should be considered to allow feasible comparison across models, in light of previously experimented parameters and achieved results. Hyper-parameters and optimization are of high importance when considering the architecture of the algorithms, and can have a massive impact on the results. This literature review studied different hyper-parameters from different studies, focusing on issues such as batch size, regularization, epoch and other consideration in reference to the importance of the current work experiment, which addresses the literature gaps summarized below.

2.5.4 Identified Literature Gaps

2.5.4.1 Limited Transportation Focus

The research acknowledges a gap in the literature concerning the impact of transportation on air quality, especially in Asia, and specifically in the Jordan area. Limited availability of air quality data, particularly in developed countries, poses a challenge.

2.5.4.2 Data Processing Challenges

Processing air quality data is reported to be difficult, especially in developed countries where data are scarce. The literature notes challenges in obtaining comprehensive information, and this may affect the accuracy of AQPs.

2.5.4.3 Input Data Refinement

The literature points out a lack of published reports regarding input data refinement for network learning, with a specific mention of studies aiming to improve accuracy by selecting the best methods for air pollution prediction. Effective parameter identification is also noted as an area with limited research.

2.5.4.4 Limited Meteorological Data

A clear limitation is the scarcity of meteorological data, specifically humidity, for many cities, particularly in Africa. This shortage of data (which is analogous for Jordan and the Middle East) could impact the precision of AQPs.

2.5.4.5 Regional and Global Efforts

The study emphasizes the global importance of the transportation sector in influencing air quality indicators. While regional and worldwide efforts are underway to study the impact of transportation on air quality, challenges in addressing traffic-related consequences and reducing emissions persist (Chapman, 2007).

2.5.4.6 Incomplete Air Quality Standards Compliance

The literature indicates that many developing and developed countries do not meet air quality standards, particularly for NO_2 near roads. Road transport, especially diesel vehicles, is identified as a dominant contributor to GHG emissions.

2.5.4.7 Focus on Specific Pollutants

The research highlights a predominant focus in the literature on specific pollutants, such as CO, NO_2, O_3, PM, and SO_2. The potential for reducing emissions is discussed, but challenges remain, especially with PM levels in Asia.

2.5.4.8 Air Pollution Concentration Disparities

It is mentioned that air pollution concentration remains high in poor countries with low income, and reversing the impact of air pollution is an on-going challenge, as observed in a study analysing trends from 1990 to 2000 (Fenger, 2009).

Chapter 3

Experimental Design Framework for Air Quality Modelling

3.1 Overview

This chapter describes the data collection process, including the pre-processing performed for collected data. Furthermore, methodologies are discussed to achieve suitable prediction accuracy. The stages of performing the experiments are described, along with a flowchart that identifies main elements of the design framework for this study.

3.2 Data Collection

3.2.1 Overview

Data was collected after completing the comprehensive review of related literature presented in the previous chapter. The literature studied the parameters used by other researchers for AQPs to give insights on the needed for the carried study of this research. This informed the view of the factors affecting gas concentrations in air, and these were selected as parameters for the models in this research. Accordingly, in the light of the aim of undertaking a Big Data comparative study, data was selected in order to compare developed and developing cities, for which sufficient data fulfilling the aim and objectives of the research was available and accessible. The data used in this research was collected from three different sources which are described in the following sections of this chapter.

There were four data phases: collection, processing, modelling and obtaining of outputs. While optimal accuracy was the intended aim, challenges were presented by issues, including data losses and data sparseness (issues of data collection); noisy and incomplete data (issues of data pre-processing); and accuracy and scalability (issues of data modelling). Some solutions to these issues were suggested by the literature, including removing noise (by filtering data such as null).

During the data check undertaken before the modelling phase, major data losses were found for temperature, wind speed, wind direction and humidity. Accordingly, other data sources were provided: emails were sent to representatives of the areas listed above, who suggested using similar data from the nearest available area to the one selected (for instance, London City Airport was said to be ‘the nearest area to Marylebone Road’). The missing data was retrieved using R software, and was replaced (by checking where it was null or zero, and

replacing it with the 'City Airport' dataset values for the corresponding missing hour, where applicable). Table 3-1 displays the monitoring stations, years, and time span for the collected data.

Table 3-1 Main data sources description

	London	**Jordan**
Air Quality Station (Data Capturing)	Marylebone Road	Greater Amman Municipality
Years	2014-2018	2016-2018
Time span	Hourly	Hourly

It should be mentioned that the hourly data used in this work reflects data points for each hour over a 24-hour period. Hourly data enables more accurate modelling, which is essential for immediate responses to public health advisories. This high-resolution data is crucial for short-term responses, in contrast with daily and monthly data, which is more suitable for medium and long-term planning.

3.2.2 Data Sources

3.2.2.1 [England Data] London Air Quality Data Selection

Air quality data were accessed from first location London Air quality data repositories (open data) (http://www.londonair.org.uk/london/asp/datadownload.asp) as first source of data. The accessible repositories include data between 1993 and 2019 (see Table 3-1), spanning several sites in London and offering a humongous amount of open source data, with up to six species selection of gases. Several air quality open data sources were searched, and the selection was based on complete data bases for at least three years of metrological data (such as wind speed, humidity and temperature).

United Kingdom Data Sources Locations:

- Oxford Street
- Westminster – Marylebone Road
- Hillingdon

United Kingdom Data Sources Parameters:

- Years (2014-2018) UK (London) Air Quality Data Selection
- Date
- Time of the day (hourly data-every hour)
- Gases (varied between locations)

- Humidity
- Wind speed
- Wind direction
- Temperature

3.2.2.2 [Jordan Data] Jordanian Ministry of Environment Air Quality Data Selection

The second source of data was collected directly from Jordanian Ministry of Environment for two traffic locations, air quality hourly data from 2016 till 2018 including meteorological related data (temperature, wind speed, wind direction, and humidity) (see Table 3-1). The location of King Hussein Gardens (KHG) had concentrations for NO_2, O_3, PM_{10}, and SO_2; and the Greater Amman Municipality (GAM) location had concentrations for PM_{10}, NO_2, CO, and SO_2. Further meteorological data from 2016 to 2018 was collected from Jordan Meteorological Department to validate humidity values, which showed some odd patterns in the original file retrieved from the Jordanian Ministry of Environment.

Jordan Data Sources Locations:

- Greater Amman Municipality (GAM)
- King Hussein Gardens (KHG)

Jordan Data Sources Parameters:

- Years: 3 years (2016-2018) Attribute Information for Jordanian Ministry of Environment data
- **GAM Location:**
- Date
- Time of the day (hourly data-every hour)
- Gases (PM_{10}, NO_2, CO, SO_2)
- Humidity
- Wind speed
- Wind direction
- Temperature
- **KHG Location:**
- Date
- Time of the day (hourly data-every hour)
- Gases (PM_{10}, O_3, NO_2, SO_2)
- Humidity
- Wind speed
- Wind direction

- Temperature

3.2.2.3 [Italy Data] Attribute Information for UC Irvine (UCI) Machine Learning Repository

The third source of data was accessed from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Air+Quality) (Vito, 2016). The dataset contains 9,358 instances of hourly averaged responses from an array of five metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year).

Italy Data Sources Locations:

- This data was used to test and validate the models for England and Jordan, with different parameters selected (as shown below), depending on the type of test performed, and for which model.

Italy Data Sources Parameters:

- Date (DD/MM/YYYY)
- Time (HH.MM.SS)
- True hourly averaged concentration CO in mg/m^3 (reference analyser)
- PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
- True hourly averaged overall non-metallic hydrocarbons concentration in mg/m^3 (reference analyser)
- True hourly averaged benzene concentration in mg/m^3 (reference analyser)
- PT08.S2 (titanic) hourly averaged sensor response (nominally NMHC targeted)
- True hourly averaged NOx concentration in ppb (reference analyser)
- PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
- True hourly averaged NO_2 concentration in mg/m^3 (reference analyser)
- PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO_2 targeted)
- PT08.S5 (indium oxide) hourly averaged sensor response (nominally O_3 targeted)
- Temperature in °C
- Relative humidity (%)
- Absolute humidity (AH)

3.2.3 Data Points

After studying multiple factors impacting AQPs, and researching data availability and completeness in this field, the researcher and supervision team decided on the strategy to attain the most satisfactory results. Data used in this experiment were collected from two locations (London, UK, and Amman, Jordan) as primary datasets for analysis, with an additional third source (Italy) to validate the model, using external completely new data.

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

Data points (data size): 43824 data point

Source: Open Data (United Kingdom)**:** London Air

https**:**//www.londonair.org.uk/LondonAir/Default.aspx

Inputs: day, month, year, hour, humidity, temperature, wind speed, wind direction

Outputs: CO, NO, NO_2, NO_x, O_3, PM_{10}, SO_2

Second location: GAM (Greater Amman Municipality), data between 2014 and 2018 (hourly data)

Data points (data size): 26268 data point

Data (from traffic locations) was collected from the Jordanian Ministry of Environment.

Source: Closed data (Jordan) - collected from Jordanian Ministry of Environment-traffic locations.

Inputs: day, month, year, hour, humidity, temperature, wind speed, wind direction

Outputs: PM_{10}, NO_2, CO, SO_2

Third location: Italian city

As explained previously, the Italy Data is used for testing and validating England and Jordan models, with inputs and outputs selected depending on the scenario test performed for the data

Source: Open data (Italy)**:** UCI ML Repository: Air Quality Dataset

https://archive.ics.uci.edu/ml/datasets/Air+quality

Variables were selected based on several factors. Firstly, a literature review was conducted to understand previous research in the air quality index domain. Additionally, this research focuses on traffic areas and related pollution at the selected locations, which were provided by the Ministry of Environment for Jordan data and pulled from open-source data for England. Data were selected based on completeness and hourly frequency as needed for the experiment.

3.2.4 Parameters Selection

A thorough literature review was done for input and output selection for studies of how AQP has been undertaken by previous researchers and it was found that a significant number used wind speed, wind direction, humidity, and temperature in different combinations, based on the studies' setups. In several studies conducted in the literature to study the inputs and outputs to be included in this research, 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 has been considered. Furthermore, this study collected available data from the parameters available for selected locations in Jordan and England (hereinafter referred to as the 'Jordan Data' and 'England Data'), as explained below. The selection of input and output was then performed based on the aims and objectives of this study. Most related studies performed prediction in isolation of the other gaseous factors (as explained in detail in Tables 2-7 and 2-8), and this research aims to provide multivariate output predictions by having multiple outputs.

3.2.5 Data Units

The following units were used for the England Data and Jordan Data from the raw data of the source locations. This is a good reference for when the data units needs conversion (as in Chapter 6), to be able to do the air quality levels representations based on USEPA standards.

England Data: CO (µg/ m^3) NO (µg/ m^3), NO_2 (µg/ m^3), NOx (µg/ m^3), O_3 (µg/ m^3), PM_{10} (µg/ m^3), SO_2 (µg/ m^3)

Jordan Data: PM_{10} (µg/m^3), NO_2 (ppb), CO (ppb), SO_2 (ppb)

3.2.6 Data Correlations

Data correlations for data were performed for some datasets to gather the relations between parameters before performing any modelling (see Figure 3-1 and Table 3-2).

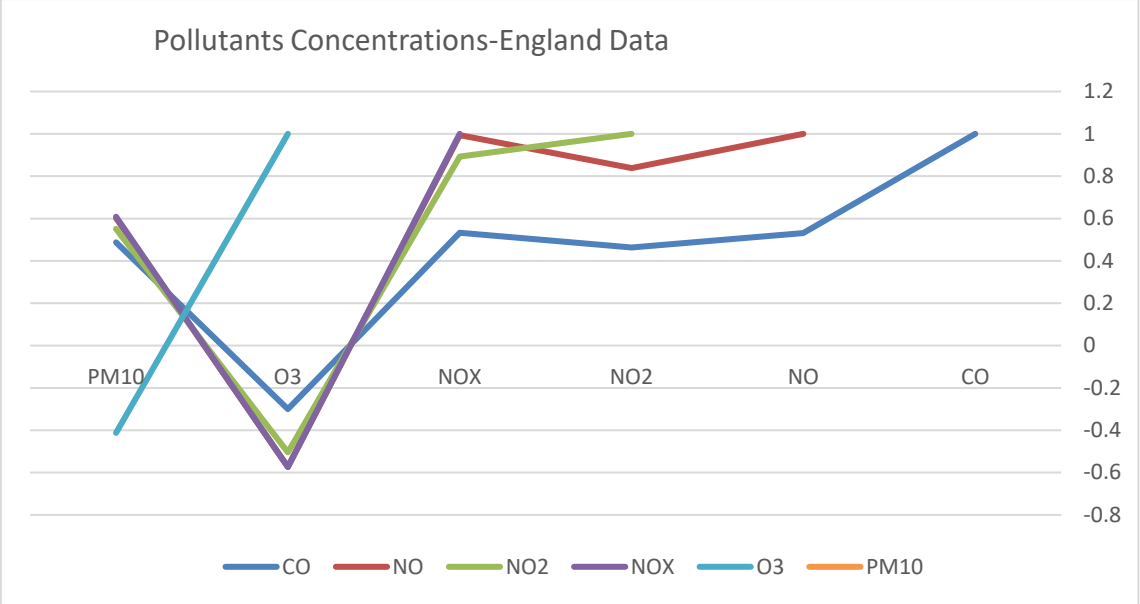

Figure 3-1 England Data correlation

Table 3-2 P-values table (correlation England Data)

	CO	NO	NO_2	NO_x	O_3	PM_{10}
CO	1					
NO	0.531676	1				
NO_2	0.463115	0.838613	1			
NO_x	0.53226	0.994263	0.89206	1		
O_3	-0.30008	-0.57249	-0.50454	-0.57405	1	
PM_{10}	0.487672	0.602425	0.550681	0.608147	-0.41223	1

3.2.7 Data Pre-Processing

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, to represent meteorological data for the studied area. Furthermore, other empty rows from gases were filled with *movemean* and then with previous values within the same column, as it is assumed this is the nearest reading for the next hour's missing data. In addition, normalization was applied to data.

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

Data pre-processing: 43824 points

Data preparation was done using multiple methods, as described below:

Input Data: Day, Month, Year, Hour columns were created by splitting the date time from original raw data (from monitors) to separated columns as in the sample below (see Figure 3-2). The year column was transformed to year 1, 2, 3, etc. (instead of 2014, 2015, 2016 etc.)

ReadingDateTime	Day	Month	Year	Hour
01/09/2014 00:00	01	9	1	0
01/09/2014 01:00	01	9	1	1
01/09/2014 02:00	01	9	1	2
01/09/2014 03:00	01	9	1	3
01/09/2014 04:00	01	9	1	4
01/09/2014 05:00	01	9	1	5
01/09/2014 06:00	01	9	1	6
01/09/2014 07:00	01	9	1	7
01/09/2014 08:00	01	9	1	8

Figure 3-2 Data example accessed using Excel sheet, showing reading date time and split (Day, Month, Year, Hour)

As can be seen from Figure 3-2, column 5. TMP, 6. WDIR, 7. WSPD, 8. RH are empty and 9. ws_C, 10. wd_C, 11. air_temp_C, 12. RH_C are the values of replacement from London City Airport (see Figure 3-3).

N	O	P	Q	R	S	T	U
5. TMP	6. WDIR	7. WSPD	8. RH	9.ws_C	10.wd_C	11.air_temp_C	12.RH_C
				3.4	235.4	15.5	77.4
				3.6	240.0	15.0	82.4
				3.1	230.0	14.5	85.1
				2.4	231.1	14.0	87.9
				2.6	240.0	14.0	87.9
				3.4	230.0	14.0	85.1
				3.1	240.0	14.0	85.1
				3.1	240.0	15.5	82.5
				2.9	245.4	17.0	77.5
				3.9	245.3	17.5	77.5
				4.4	255.3	17.5	91.0
				4.4	250.0	17.5	91.0
				4.6	254.5	18.0	91.1
				4.1	260.0	18.5	85.5
				4.4	250.0	18.5	91.1

Figure 3-3 Wind speed, wind direction, temperature, humidity data accessed from Excel sheet

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 'fill missing command' with *movemean* values.

Following the above data manipulation, a further 'fill missing command' was used, up to the length of the data, to replace any remaining missing data using previous value as the nearest reading for the next hour of missing data. Furthermore, normalization was then applied to data as explained below.

3.2.8 Data Normalization

3.2.8.1 Min-Max Scaling

Data were analysed, and there was a need to normalize some parts of the data, such as the humidity column, which was normalized as per the following equation:

$$X_{\text{normalized}} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

3.2.8.2 Data Partitioning

The model initially pulls data from the identified source, and for this research experiment, Excel sheet data source was primarily provided (as per the procedures for data sources and manipulation discussed previously). The data are divided into training, test, and validation sets, with ratios of 0.8, 0.1, and 0.1 (respectively), and these were further subject to random partitions from the matrix of inputs and outputs.

3.2.9 Data Characteristics and Features Selection

England Data repositories include data from 1993–2019, as explained previously, encompassing several sites in London, 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 three years with concentrations (CO, NO, NO_2, NO_x, O_3, PM_{10}, SO_2) and meteorological data, such as wind speed, humidity and temperature.

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 (temperature, wind speed, wind direction, and humidity) for the GAM location, with concentrations for PM_{10}, NO_2, CO, and SO_2. Further meteorological data from 2016 to 2018 was collected from the Jordan Meteorological Department, to validate humidity values, as some odd patterns were detected in the original file retrieved from Jordanian Ministry of Environment, after consultation with experts in the field from the Department of Statistics in Jordan.

The experiments in this research were conducted for multivariate output prediction for several gases in two cities:

- London, UK Output: CO, NO, NO_2, NO_x, O_3, PM_{10}, SO_2
- Amman, Jordan Output: PM_{10}, NO_2, CO, SO_2

3.3 Modelling Approaches Proposed

In this study several modelling approaches were conducted to fulfil the aim and objectives of this study, and to present a comprehensive modelling approach. Two different datasets were used from two locations (England and Jordan) for the main comparative study, with another dataset from Italy. The modelling was done following several gradual steps, as explained below.

This research purpose is to build prediction model and defining measurable (quantifiable) data and set them to measurable index (AQI). Several ML methods were used to compare results and model results using MATLAB R2020a software in the initial stage, using ANN methods, as listed below:

- Fitting tool app (MATLAB)
- Time series App (MATLAB)
- ANN –*nntool* (MATLAB)

Results were compared and time series app has proved better results in all tested cases and scenarios for the same dataset. After checking with several ANN models, the DNN model was developed first, based on the Jordanian data. Hyper-parameters tuning was performed to the model until a suitable accuracy is achieved. The model was then used for England Data, and gradual modifications were applied to achieve suitable accuracy. A Markov model was then developed independently for Jordan Data and England Data, and it was built to fulfil data structure and the objectives of the experiment and after approaching sensible accuracy. Developing the hybrid model took a huge part of the experiment, to have better accuracy than both developed models independently. In the next stage, wider experiments performed using MATLAB R2020a, conducted at several levels, as listed below:

- Developing stand-alone DNN model
  - *Working on hyper-parameters to optimize the model*
- Developing stand-alone Markov Chain model

- *The Markov-switching regression model provides a prediction method using MS-VAR*
- *Working on parameters setup to optimize model*

- Developing hybrid model connecting the two model (DNN and Markov chain) using multiple methods, as discussed below.

Several experiments conducted for the hybrid model, to test the best possible scenario for the hybrid model accuracy, to achieve superior accuracy to stand-alone models. It is worth mentioning that other methods were used in experiments, in an attempt to increase accuracy. However, Method 5 (described in Figure 3-8) was selected as the most accurate, compared to the other methods mentioned below. Some of these methods scenarios are:

- **Method (Scenario 1)**: 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 (Scenario 2)**: 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 outputs were predicted using DNN. The mean of both predictions was then taken.
- **Method (Scenario 3)**: Simulated Markov results (input and output) were used as input and output data for DNN. A run was then performed using the new simulated states and the predicted output from Markov (Figure 3-4)

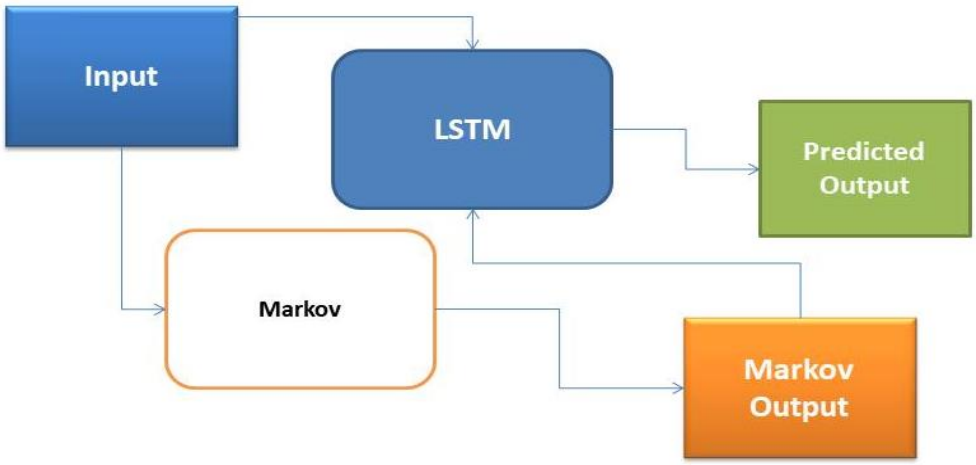

Figure 3-4 Proposed hybrid modelling (Markov Chain and DNN)

The modelling stage required combining the data from various sources, in order to achieve an appropriate level of accuracy. Taking into consideration the data used and the domain (air quality), following these steps to obtain the final results. Deep learning and Markov models were combined in multiple ways (methods), as adumbrated below.

3.3.1 Method 1

In first experiment, calculation for error was performed for DNN and added to Markov outputs and then Markov run was performed with new outputs. In another way, Markov error was calculated and added to DNN results and then DNN run was performed with new outputs in efforts to check suitable accuracy (see Figure 3-5).

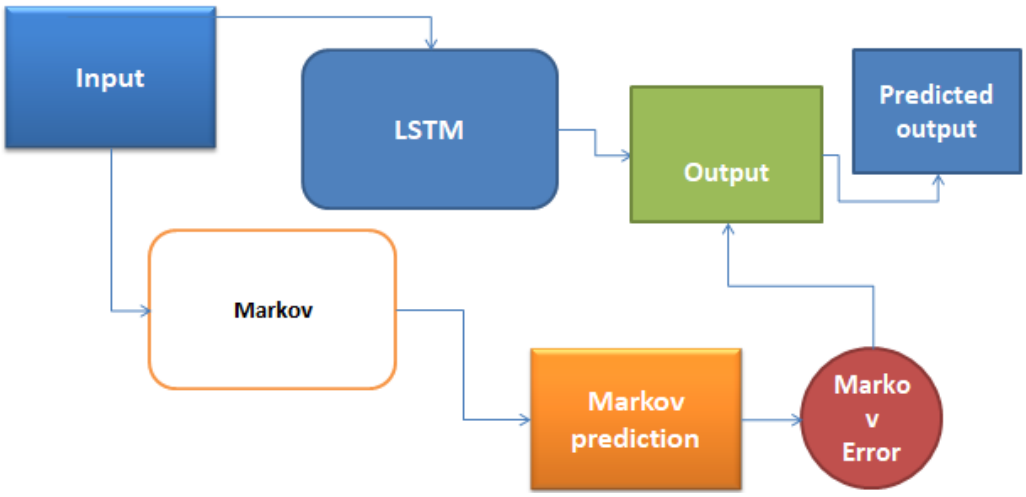

Figure 3-5 Example of prediction error consideration for hybrid model

3.3.2 Method 2

In the second experiment, the mean was checked for both results (mean DNN and mean Markov). Mean values were used as a hybrid method, predicting output using Markov, and then predicting output using DNN, and taking the mean of both predictions.

3.3.3 Method 3

In another trial, DNN and Markov were combined at first using HMM outputs after running the algorithm with outputs for DNN (as new outputs), or using HMM simulated states as inputs for DNN (as new inputs). DNN was run with the new input and original outputs. This is a way to make RNNs more interpretable, and to achieve better performance (see Figure 3-6).

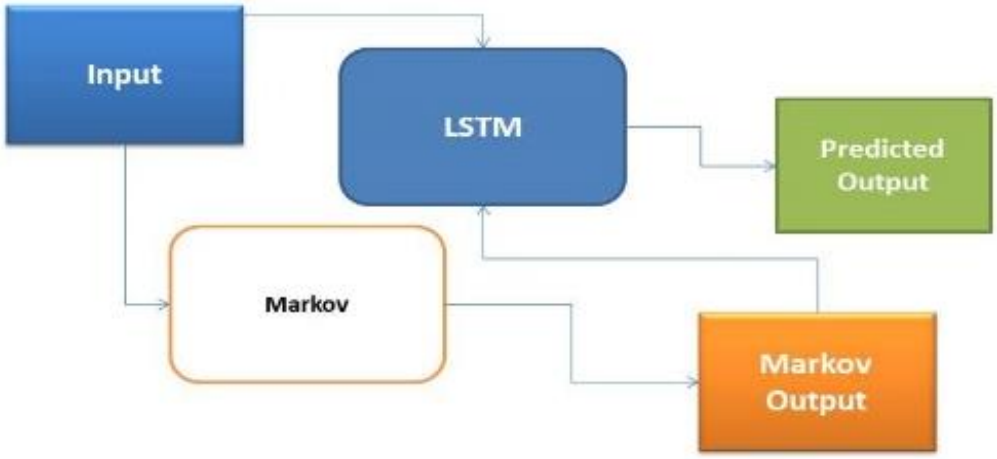

Figure 3-6 Example of Markov output consideration for hybrid mode

3.3.4 Method 4

Furthermore, a trial was performed by having both simulated Markov results (input and output) as input and output to DNN, which was then run using the new simulated states and predicted output from Markov (see Figure 3-7).

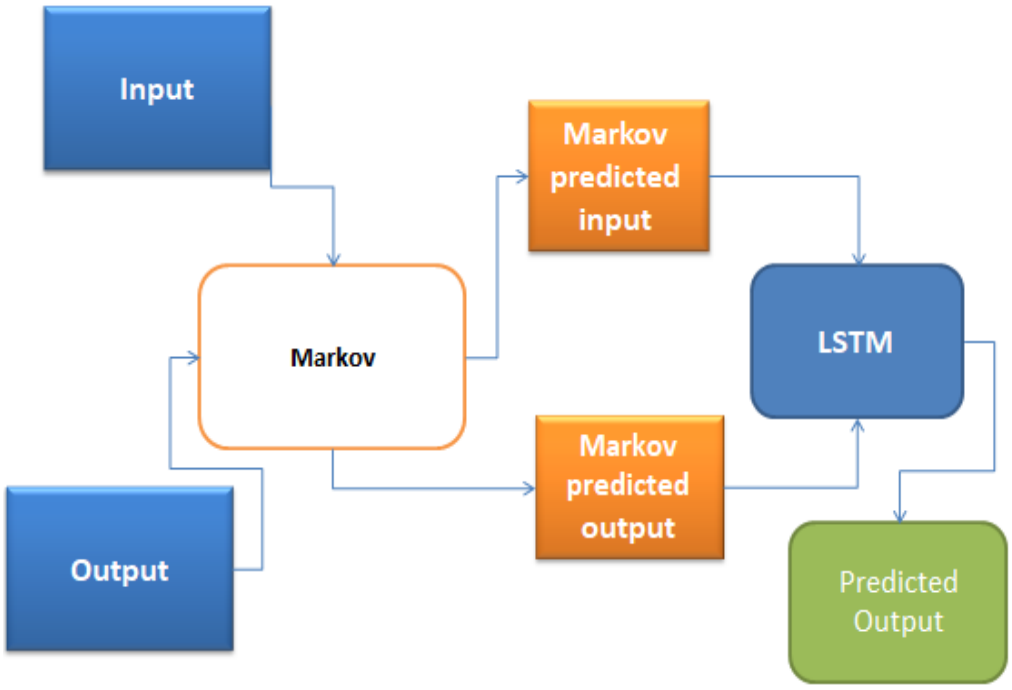

Figure 3-7 Example of simulate input and outputs using Markov as input and output to DNN

3.3.5 Method 5

LSTM output is predicted using LSTM, and then output is predicted using Markov.

The data output (gas concentration only) was used as output data for the Markov model. (The missing data was processed first with a moving mean for each gas and then as a

previous value, to back up any missing values for which the data was completed by the mean).

The gas concentration results from the DNN model were used as a feed to the Markov model (acting as the output). The inputs (i.e. the original inputs for the DNN model) were: day, month, year, hour, wind speed, wind direction, temperature and humidity. The final results were the outputs predicted after the running of the Markov model (see Figure 3-8). More details explained in Chapter 5 and Chapter 6 (see Figure 6-3) for full process details.

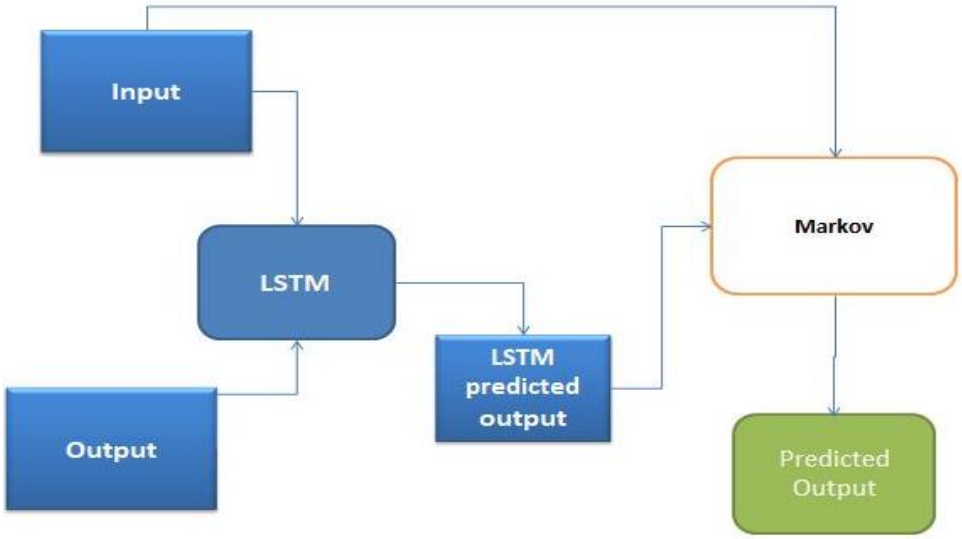

Figure 3-8 Example of LSTM output consideration for hybrid mode

3.4 Experimental Design Framework

The purpose of this research is to build a prediction model (next hour forecasting) and define measurable (quantifiable) data and compare different models for AQP recommendation. Several ML methods were used to compare results and model results using MATLAB R2020a software. The study analysed related literature to discover the researched algorithms in the domain, and then experimented with the top niche of them in an effort to build a novel, more efficient model, to support AQP with improved accuracy. The experimental stages are summarized below.

- **Experimental Stages**

Stage 1: Data collection

Stage 2: Data pre-processing

Sage 3: Data preparation

Stage 4: Model development

Model development phase 1

- NN (feed forward backdrop).
- NN fitting.
- NN time series (NARX).

Model development phase 2

- DNN
- Markov chain
- Hybrid model (DNN and Markov)

Stage 5: Model performance evaluation

3.5 Summary

As a summary for Chapter 3, the flow of the experiment design stages explains the major stages of the experimental work conducted in this thesis (Figure 3-9). As the flow chart shows, data were collected from three sources (data collection process), as presented in section 3.2 of this chapter. Subsequently, data preparation and pre-processing were performed – an important step to ensure data quality. Models were developed based on the aim and objectives of this study in phases, as shown in section 3.4. Following the model setup, data splitting, and model training, testing, and validation were performed. Details that summarize the stages of the experimental design are presented in Figure 3-9.

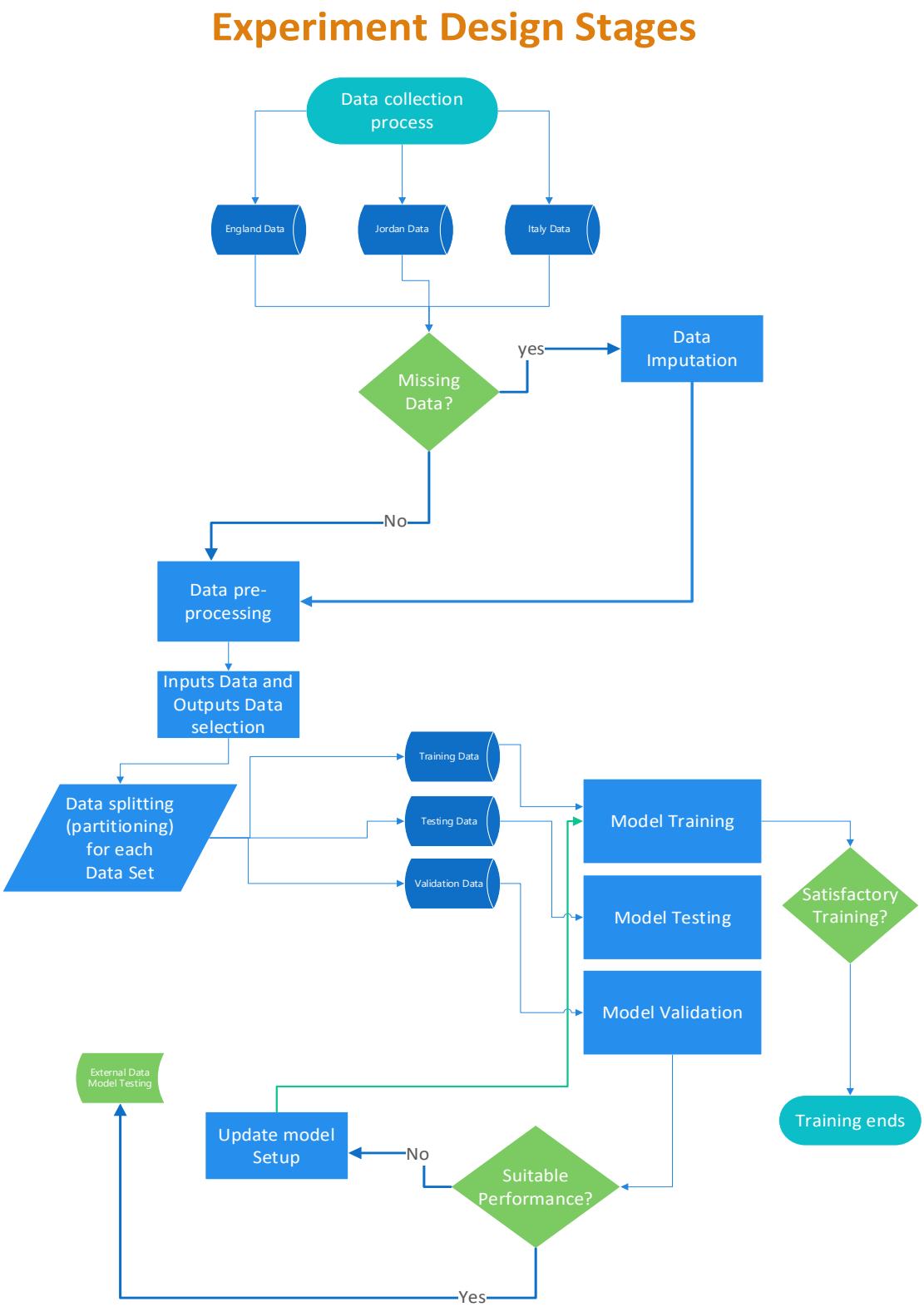

Figure 3-9 Experiment design stages flow chart

Chapter 4
Air Quality Models Using Stand-Alone Models

4.1 Introduction

This chapter presents the methodologies for AQP. The focus of this chapter is on ANN and deep learning models and their architectures. This chapter highlights major contributions of the research concerning ANN and deep learning, and the use of multi-input multi-output. Firstly, simple NN's were conducted with different types (NN-FFB, NN-Fitting, NN-NARX) to assess feasibility for the aim and objectives of this research. DNN was then used with different architectures, to test multiple possible scenarios and select the topology that provides suitable results for the study in terms of accuracy and reliability. This chapter presents and discusses a Markov chain model, and proposes a method of applying Markov through ARIMA representing multi-input and multi-output for simulation. It analyses the Markov chain model and the experiment for which the model was built, to determine if it fulfils the accuracy requirements for the stand-alone model. In the following chapter it is combined with DNN to make hybrid model. The experimental setup explanation explains the data preparation, model testing, validation, and testing the performance of the model with selected external data.

4.2 Theoretical Basis for Architectures

4.2.1 Artificial Neural Network and Deep Neural Network Evolution

ANN has obvious footprints in AQP domain, as is clear from its prolific mentions concerning NN models in existing literature throughout the years (see Tables 2-6, 2-7, and 2-8). As complexity increases in many domains, especially in fields that intersect with various disciplines, such as air quality, many factors affect and are affected. Consequently, there has been a notable increase and a leap in using more complex setup of layers by considering the evolution of DNNs, which offer more capacity to produce more accurate results. The systematic literature review summarized in Table 2-8 revealed that great attention has been paid to the high potential and already achieved results for DNN in air quality studies, but as there is already a rich history of the use of ANN, the research started on the basis of simple prediction using ANN types (see Chapter 3 for background on ANN methods used in this research). DNN was then used to achieve more reliable and accurate results, due to the size and dimensions of data used. This chapter comprehensively

discusses the experiments, parameters, implementations, and results of the described methods used.

4.2.2 Markov Chain Insights

Markov chains are considered as a form of stochastic or statistical modelling; put simply, they concern the probability for an event to happen based on the current state of previous event (with that a sequence of possible events could be represented). Statistical methods which provide estimations provide interpretability and competing performance, unlike ML forecasting, which does not provide the same balance between interpretation for data and performance for models (Rybarczyk and Zalakeviciute, 2018). Although Markov chains are of good use in stock modelling and other domains, they are underused for air pollution forecasting, and the current literature revealed very limited research in the field for Markov chains, especially for multivariate studies. Hence, this research studied Markov chain due to the scale of benefits for interpretability, as it offers a simple linear type of models that could be used to backup complex systems, and compensate for errors.

The theoretical basis for Markov chains used in this study in order to build the Markov chain experiment pertains to the hidden Markov model, modelled using time series regression model (Markov Switching Dynamic Regression Model). The ‘switching’ term represents the mechanism of switching of the regression model coefficients. It has been discovered that this type is best suited for the research experiment and the nature of air quality within this study framework (i.e., having multi-inputs and multi-outputs).

The proposed model in this research is adopted from Hamilton (1989), who proposed an algorithm for predicting future time series values, and suggested that probabilistic inferences should be in place, drawn from the observed time series in which shifts in regime could happen. Hamilton (1989) represented shifts in dynamic behaviour explicitly, and noted that each regime is defined by parameters which are subject to estimation.

4.3 DNN Stand-Alone Model Development

4.3.1 Data Preparation for Training, Testing, and Validation

X is indexed from Column A to Column H: 'A1:H43824' and the number 43824 (England Data) refers to the number of rows in the excel sheet to include the whole dataset. The data partitions are then split into three with 80 percent for training data, 10 percent for testing data, 10 percent for validation data (all randomly) using *dividerand* function. Pre-processing for partitions of the data is performed as discussed in data preparation and processing

methods (see Chapter 3). The method selected and used for filling missing data for this study is to be performed on two stages during run time, as the data will be first filled using '*movmean*', as this will calculate mean over specified length and based on the experiment performed this helps in raising the performance of the model.

The second stage is filling any remaining missing values by using *fillmissing* 'previous' as a backup to fill in what is missing with the previous non-missing value. Following this order for *fillmissing* values ensures that the values are filled first with *movemean*, and when and if values need to go to the next stage. If any missing values remain, the window of fill previous will be smaller, to give a more realistic filling method, as in first step most of the values are filled using *movemean* this will minimize gaps. The previous value picked in this case will be nearer to the hour of the missing value. The data will be in form of matrix partitions in this case for the first instance, and to convert the data into the processing form for the DNN model, num2cell is used to convert numeric arrays to cell arrays (for all training, testing, and validation partition data).

The network ‘net’ is then trained using *trainNetwork* to train the network with parameters input and output as cell entries. Also, layers and options parameters are used, as previously mentioned. After network training prediction is performed for training, testing, and validation data. This method produces the forecasting values of chief concern for this study. The identified parameters are changed based on experimental needs to fulfil this study aim of approaching suitable accuracy for the prediction. The same steps are followed for all of the studied datasets.

4.3.2 DNN Parameters

The *numReponses* variable is set to the size of output training data, and this variable is used during DNN run time for training purposes as a parameter for the training network for *fullyConnectedLayer*. On the other hand, *featureDimension* is set to the size of input training data, and is used as a parameter for *sequenceInputLayer*. *numHiddenUnits* is set to represent number of neurons (hidden units) for the *lstmLayer*, and this parameter has been tweaked through several experiments to have number of hidden units to reach satisfactory performance for the model in relation to the *numReponses* and *featureDimension*.

Tables 4-1 and 4-2 display the particular characteristics of the England and Jordan Data (respectively. Tables 4-3 and 4-4 show the DNN layers architecture and training options applied to both datasets (respectively).

Table 4-1 DNN training options – England Data

England Data	
numHiddenUnits	800
maxepochs	1000
miniBatchSize	900

Table 4-2 DNN training options – Jordan Data

Jordan Data	
numHiddenUnits	300
maxepochs	500
miniBatchSize	256

Table 4-3 DNN layers architecture – England Data

Layer	Parameter
sequenceInputLayer	featureDimension
lstmLayer	numHiddenUnits
dropoutLayer	0.3
lstmLayer	numHiddenUnits
dropoutLayer	0.3
fullyConnectedLayer	numResponses
regressionLayer	

Table 4-4 DNN training options –England Data

Option/ Parameters	Description
'adam'	Optimization algorithm, offering dynamic adjustment for parameters learning rates with bias correction.
'ExecutionEnvironment' 'auto'	Parameter to specify the mode of run for the mode (CPU, GPU, etc.), which impacts the computational resources used and training efficiency.
'MaxEpochs' Maxepochs (defined above in parameters)	The maximum number of passes through the whole dataset during training.
'GradientThreshold' 1	A hyper-parameter assessing in the stability of the training by using a threshold as a maximum limit. If the gradient of parameters exceeds the threshold limit, exploding gradients can be avoided to retain stability.
'InitialLearnRate' 1e-2	A primary hyper-parameter, which initializes the value of the learning rate for the model by defining the step size for weight updates during training. This hyper-parameter impacts model generalization.
'LearnRateSchedule' piecewise	A method for learning rate adjustment during training for model performance improvement.
'LearnRateDropPeriod' 125	A hyper-parameter to specify training iteration after the learning rate reduction (while moving towards a minimum of a loss function), mainly used for better model convergence by controlling the reduction of the learning rate.
'LearnRateDropFactor' 1	A hyper-parameter that sets the factor of learning rate drop during training. It assists in training stability for the model and convergence. A larger drop factor leads to more reduction in the learning rate.
'L2Regularization' 1e-10	A technique used for over fitting control for the model, helps in mode generalization.
'Verbose' 0	Specifies the amount of out displayed when training the model. Zero determines the level of verbose, which means no information is displayed in this case.
'MiniBatchSize' miniBatchSize	The size of mini batch –number of training data batches (subset) used to update model parameters (gradient, weights, etc.) for each iteration.
'Plots' 'training-progress', 'ValidationData', {XVal,YVal}	The training screen that shows model's performance and metrics (losses, accuracy, etc.) as visualization during run time for the model training.
'training-progress'	This gives the status about training performance metrics during model learning from training data, which is essential for monitoring training progress and taking necessary actions to improve model data training.
'ValidationData' {XVal,YVal}	Part of the data to validate the model performance during training. Effectively provides performance evaluation metrics such as validation accuracy, loss and helps in fine-tuning the model to avoid over-fitting.

4.3.3 Hyper-Parameters Tuning and Optimization Techniques

At first, the experiments for DNN model started with one LSTM layer, to see the behaviour of the model. Subsequently, with the same base, parameters were tuned –such as mini batch and epoch, and the DNN Jordan model was doing better than the DNN England one. To close the gap in performance the number of LSTM layers was increased to two, and the performance and reliability increased for both DNN England and DNN Jordan. After this tweaking for parameters, the models were finalized as shown in Table 4-5. According to Zhou *et al.* (2019), adding L2 regularization algorithm and dropout layer could increase the stability of the model, and this research found this to be impactful for the multivariate models in terms of accuracy. After experimenting with L2 regularization there was an immediate improvement, which made a consequential difference in results. Also, other considerations of hyper-parameters and optimization techniques, as shown in Tables 4-3 and 4-4, collectively contributed to form the optimized DNN architecture, after extensive trials to provide the combination that worked best for the studied models and data.

Table 4-5 DNN models comparison England and Jordan Data

Model type	LSTM Layers	L2 Regularization	Mini batch size	Epoch
DNN-Jordan Data	2	1E-10	256	**500**
DNN-England Data	2	1E-10	900	**1000**

4.4 Markov Chain Stand-Alone Model Development

4.4.1 Markov Chain Parameters: Setup, Inputs, and Outputs

The Markov-switching dynamic regression model consists of four states (humidity, wind speed, wind direction, and temperature). Each state was formed using ARIMA with the parameters in Table 4-6. The output model was also formed using ARIMA with the parameters presented in Table 4-7.

Table 4-6 and Table 4-7 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 (AR, beta, constant, and variance) in Table 4-7 (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. 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.

Table 4-6 Markov model output parameters for England and Jordan Data

Parameters	**Models**	
	Markov Jordan	**Markov England**
AR (auto regression coefficient)	Mean (corr(Inputs,Output)	Mean (corr(Inputs,Output)
Beta (regression coefficient)	Set to 1	Set to 1
Constant (mean)	Set to 0	Set to 0
Variance (standard deviation)	Set to 1	Set to 1

*Inputs (represent all four inputs)

*corr (correlation)

*std (standard deviation)

Table 4-7 Markov model input parameters for England and Jordan Data

Parameters	Models	
	Markov Jordan	**Markov England**
AR (auto regression coefficient)	Input1: mean (corr(Input,Output)) Input2: mean (corr(Input,Output)) Input3: mean (corr(Input,Output)) Input4: mean (corr(Input,Output))	Input1: mean (corr(Input,Output)) Input2: mean (corr(Input,Output)) Input3: mean (corr(Input,Output)) Input4: mean (corr(Input,Output))
Beta (regression coefficient)	Input1: set to 1 Input2: set to 1 Input3: set to 1 Input4: set to 1 Input5: set to 1 Input6: set to 1 Input7: set to 1 Input8: set to 1	Input1: set to 1 Input2: set to 1 Input3: set to 1 Input4: set to 1 Input5: set to 1 Input6: set to 1 Input7: set to 1 Input8: set to 1
Constant (mean)	Set to 0	Set to 0
Variance (standard deviation)	Input1: std(Input1) Input2: std(Input2) Input3: std(Input3) Input4: std(Input4) Input5: std(Input5) Input6: std(Input6) Input7: std(Input7) Input8: std(Input8)	Input1: std(Input1) Input2: std(Input2) Input3: std(Input3) Input4: std(Input4) Input5: std(Input5) Input6: std(Input6) Input7: std(Input7) Input8: std(Input8)

*corr (correlation)

*std (standard deviation)

4.4.2 Markov Model Architecture

4.4.2.1 Autoregressive Integrated Moving Average Models

The states of the Markov-switching regression model were identified based on the number of Inputs, each of which was converted to the Autoregressive Integrated Moving Average (ARIMA) model with the experimental number of parameters. The following inputs were used as parameters for the ARIMA model:

- **Inputs**

- Hour: The hour of the day (for 24 hours)
- Day: The day of the month

- Month: The month of the year
- Year: The year
- Wind speed
- Wind direction
- Temperature
- Humidity

4.4.2.2 Markov-Switching Vector Autoregressive

Markov-Switching Vector Autoregressive (MS-VAR) is a type of finite order VAR model of *p*th order and K-dimension time series vector. The assumption of Markov-switching is the regime of a discrete time, discrete state. MS-VAR model supports non-linear predictions, which provides flexibility for discrete shifts (Krolzig, 1997; Chauvet and Hamilton, 2005). MS-VAR makes it possible to have regime dependent parameters, or separate regimes for each shifting parameter. The formulation of MS-VAR model with a finite number of states *st* is inferred from data ('observables') (Hamilton, 1990; Robinson, 2009).

4.4.2.3 Random Walks

Random walks are generally based on probabilities, with no trends or patterns from previous steps, and in times series. They are generated based on mathematical models to form a random process. In the context of Markov switching dynamic modelling, random walk evolves according to the parameters associated during each regime. (Kemeny and Snell, 1983).

4.4.2.4 Probability

To create probability matrix, a state vector is initiated with zeros for 8 instances. The first instance of the vector is then assigned to one, as well as a blank probability matrix is created with zeros. A nested for loop from 1 to 8 used to create Markov probability matrix using random function to generate a random number from a uniform distribution in range (0,1). For this stochastic random matrix, the sum of all elements along the row should be equal to one.

$$\boldsymbol{p} = \begin{bmatrix} \boldsymbol{p_{11}} & \cdots & \boldsymbol{p_{12}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p_{21}} & \cdots & \boldsymbol{p_{22}} \end{bmatrix} \tag{4.1}$$

4.4.2.5 Discrete Time Markov Chain

Discrete Time Markov Chain (DTMC) or stochastic process, represents the sequence of random variables (whereby the next variable value depends only on the value of the current variable), and there is no consideration for past variables. The sequence of states is Markov

chain, the sequence of transitions from state to another which can be described as stochastic matrix (the probability of states transitioning).

4.4.2.6 State Transitioning

State transitioning or the regime-dependent covariance which is represented as described below.

The states are defined using eight variables (MdlX1, MdlX2, MdlX3, MdlX4, MdlX5, MdlX6, MdlX7, and MdlX8), each of which represents inputs as states for the Markov model. Each variable is assigned to ARIMA model. ARIMA models' parameters are 'AR', 'beta', 'Constant', 'Variance'.

The first state definition is shown below; the other states follow the same logic:

```
MdlX1 = arima('AR',ARRR,'beta',1,'Constant',0,'Variance',std(Input1));
```

The ARIMA model consists of AR parameter, beta parameter, constant parameter and variance. AR equals to ARRR (a defined variable for this research used to save discussed values), a value defined before initiating ARIMA models for the states, by finding the correlation between all inputs and outputs using the formula *corr(Input,TrainingData)*. This which returns the matrix of correlation coefficient between x and y (in this case, inputs and outputs), and then the mean of ARRR is calculated. The resultant value is assigned to the ARIMA model's AR parameter, which describes the response process within the regime-auto regression coefficients. The second parameter in the ARIMA model is beta, which is set to 1 for all states. The third parameter is constant, which is set to 0 for all states, and the variance parameter equals the standard deviation for each input. Consequently, there are eight difference variances for each ARIMA model. All eight ARIMA models representing the eight inputs are then stored in one matrix variable MdlX (see Figure 4-1).

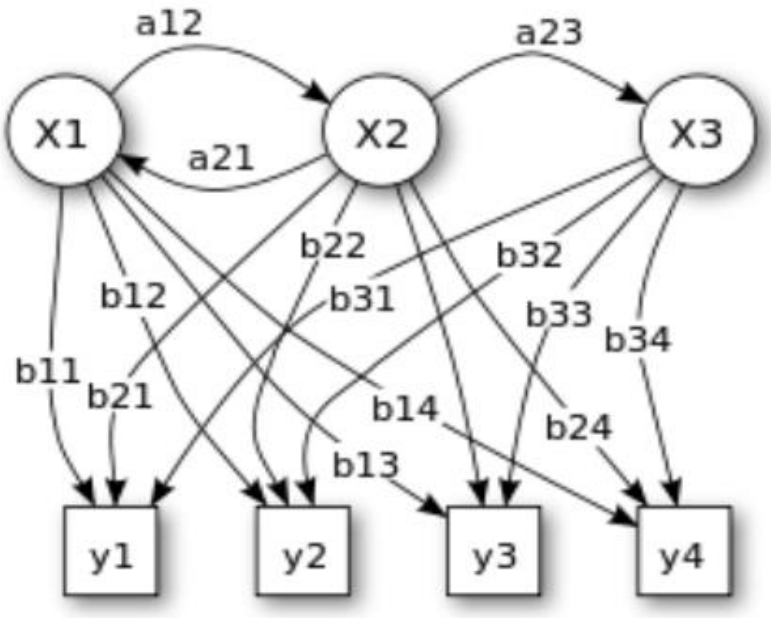

Figure 4-1 Probabilistic parameters of hidden Markov model

(X) represents states, (y) represents possible observations, (a) represents state transition probabilities, (b) represents output probabilities

Source: Tdunning (2012)

4.4.2.7 Input Simulation

Input1, Input 2, Input 3, Input 4, Input 5, Input 6, Input 7, and Input 8 were simulated using the simulation function based on equal probability for each of the four states (transition probability).

4.4.2.8 Outputs Simulation

After determining state transitioning, probabilities matrix and the DTMC object *mc* using state transition matrix *p* and state transition models, outputs were simulated using simulate function with Mdl parameter, with number of observations represented by number of rows of outputs and the observed output data. Each output is represented by simulate function with the specified output parameter.

Each simulation function represents one of the outputs, to form simulated values for all the outputs. All simulated outputs are stored in one defined variable named TrainingDatay.

4.5 Models' Implementation Results

Figures 4-2 and 4-3 show flow charts for the DNN and Markov models. The results are discussed below.

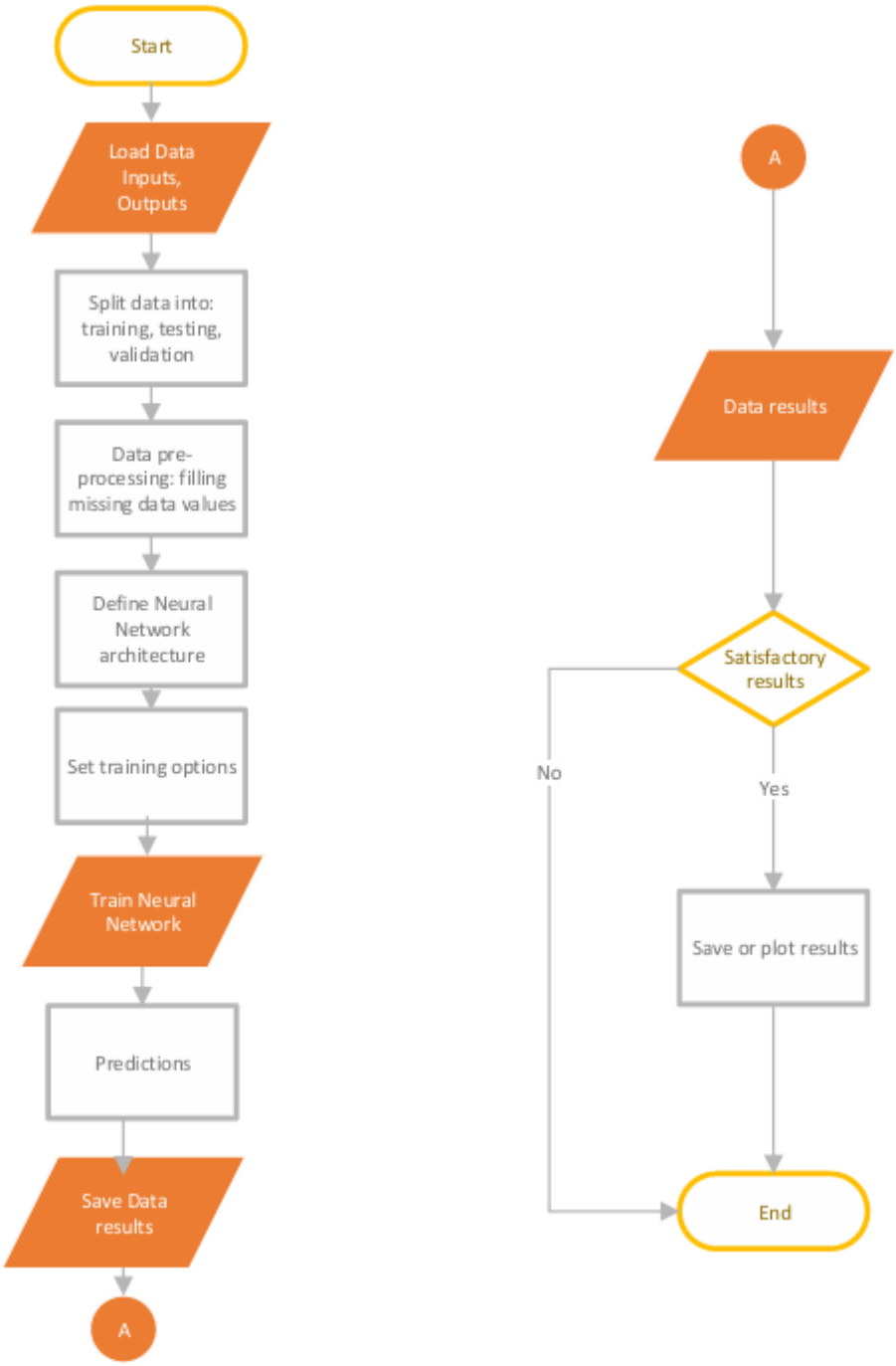

Figure 4-2 Deep learning model flowchart

The flowchart (Figure 4-2) presents the process of performing a deep neural network (DNN) experiment. First, the data is loaded and then randomly split into training, testing, and validation sets. The data is pre-processed at runtime to ensure any remaining missing data is filled. Next, the architecture is defined as discussed thoroughly in Section 4.3 (DNN parameters and hyper-parameters tuning and optimization techniques). The prediction is then performed after training the DNN model. The obtained results are saved and checked

for satisfactory performance by calculating the error. If the results are satisfactory, they are plotted.

Figure 4-3 shows the flowchart for the Markov chain experiment. The process starts with data preparation to fit the data to the model, given the unique nature of the Markov model. A multivariate regression is performed to obtain the parameter setup for the Markov model (beta, sigma, etc.) as specified in detail in Section 4.4. Then, statistical calculations are performed for input/state representation as they are modeled using the ARIMA model for each input/state. A Markov chain simulation for the output is performed to obtain the predicted results. The results are evaluated by calculating the error, and if satisfactory results are achieved, they are plotted.

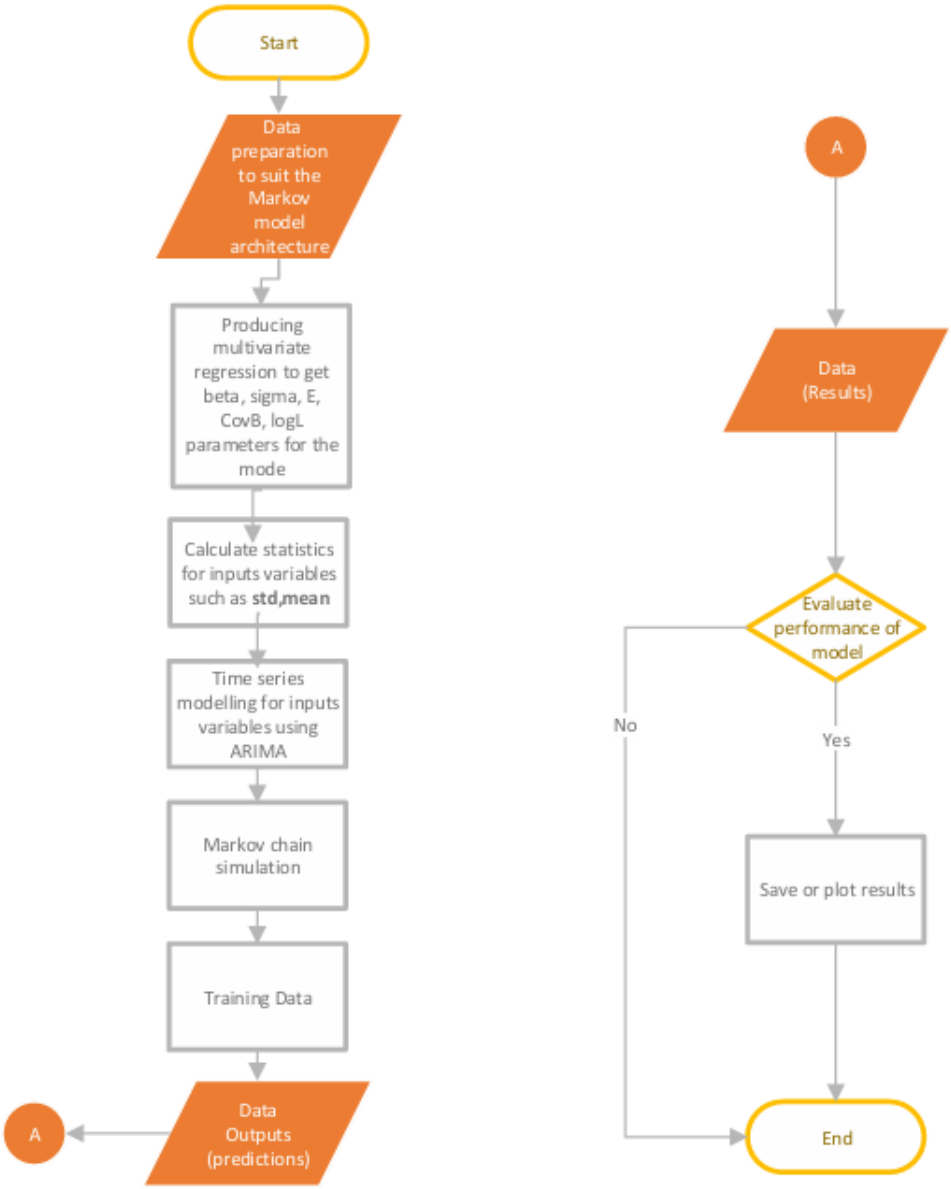

Figure 4-3 Markov chain model flowchart

4.5.1 ANN Results

4.5.1.1 NN-FFB, NN-Fitting, NN-NARX

As described in Chapter 3 (concerning the experimental design), the first experiment was done using ANN (NN-FFB, NN-Fitting, NN-NARX). Table 4-8 shows the comparison results. NARX (dynamic NN) model yielded better results than ANN for the tested cases and scenarios for the same dataset. This is because the structure of this algorithm provides support for time series data. All results/performance are reported based on experiments using MATLAB R2020a (see Table 4-8). Figures 4-4 to 4-9 display the results for the models described in Table 4-8.

Table 4-8 NN results for England and Jordan Data

Model type	Location	Accuracy	No. Hidden Units	Training Function
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**

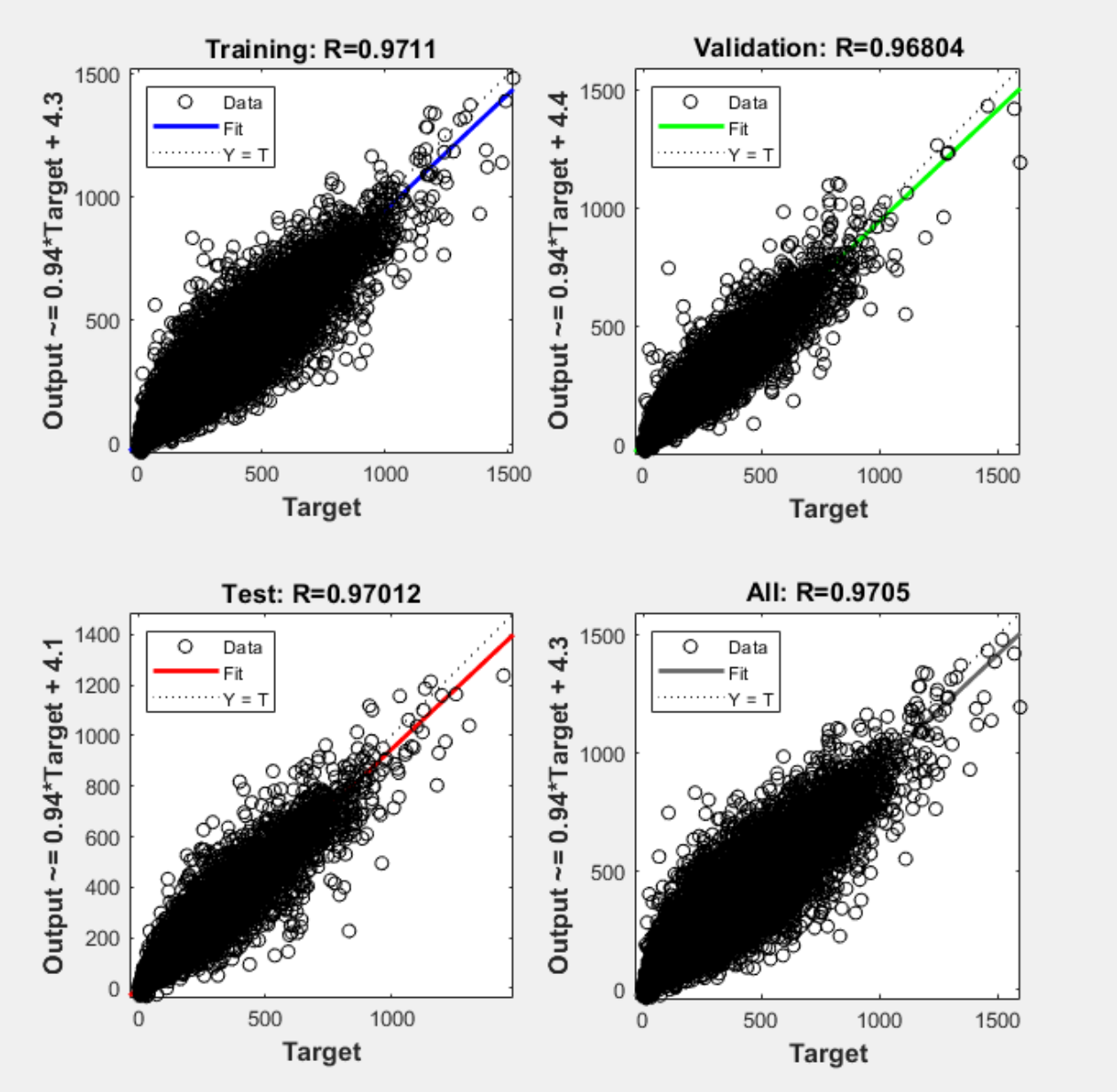

Figure 4-4 Westminster – Marylebone Rd results (Central London) – NARX

	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

Figure 4-5 Westminster – Marylebone Rd (Central London) – NARX results/errors

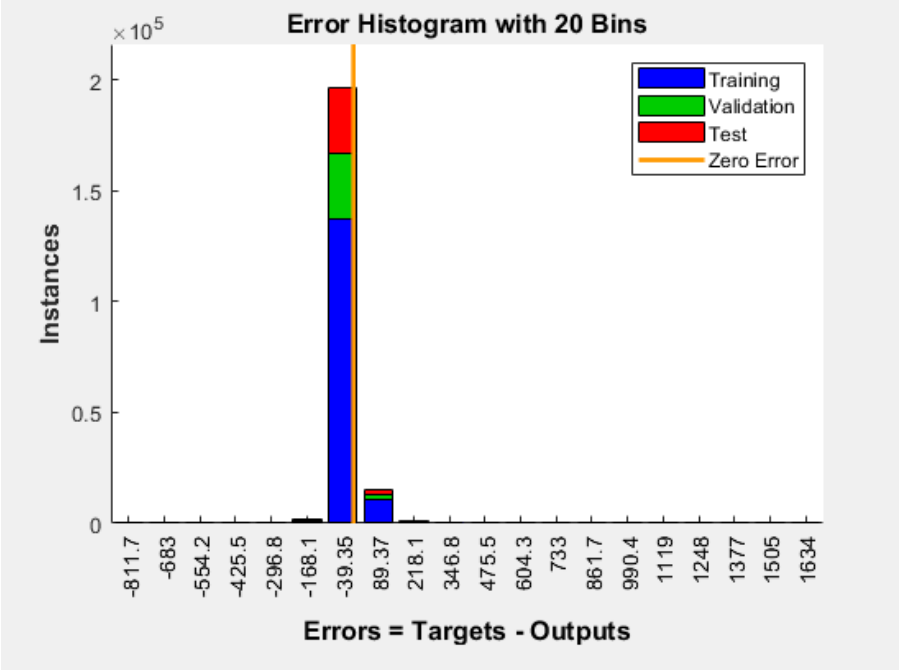

Figure 4-6 Westminster – Marylebone Rd (Central London) – NARX error histogram

Figure 4-4 shows the accuracy of the NARX model for England data, presenting the results for the training, validation, and testing data. In addition, Figure 4-5 clarifies the MSE and R values, which precisely measure the accuracy of the NARX model. Additionally, the error histogram (Figure 4-6) visualizes the prediction errors (the difference between predicted and actual values).

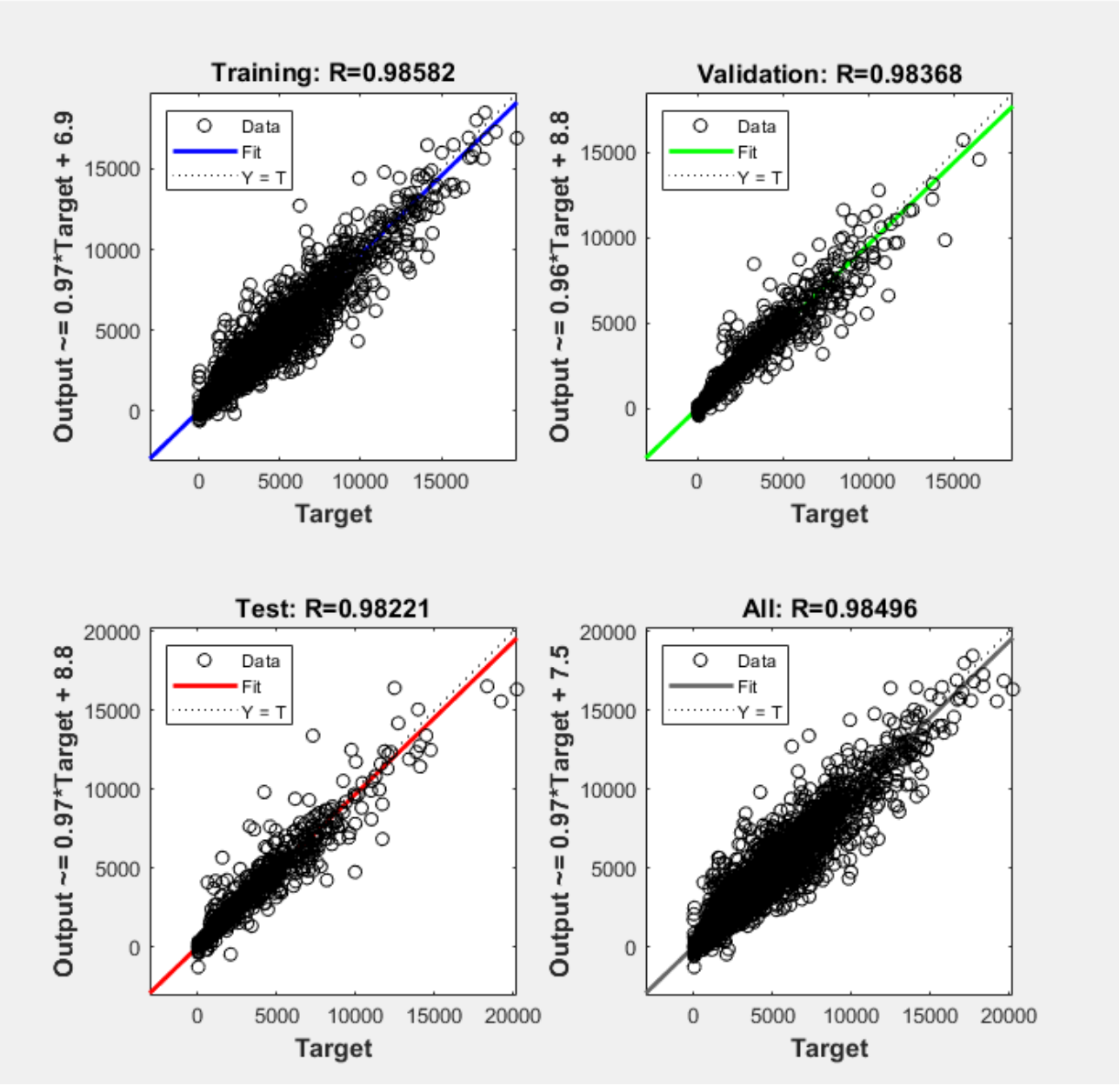

Figure 4-7 GAM location results (Jordan) – NARX

	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 4-8 GAM location (Jordan) – NARX results/errors

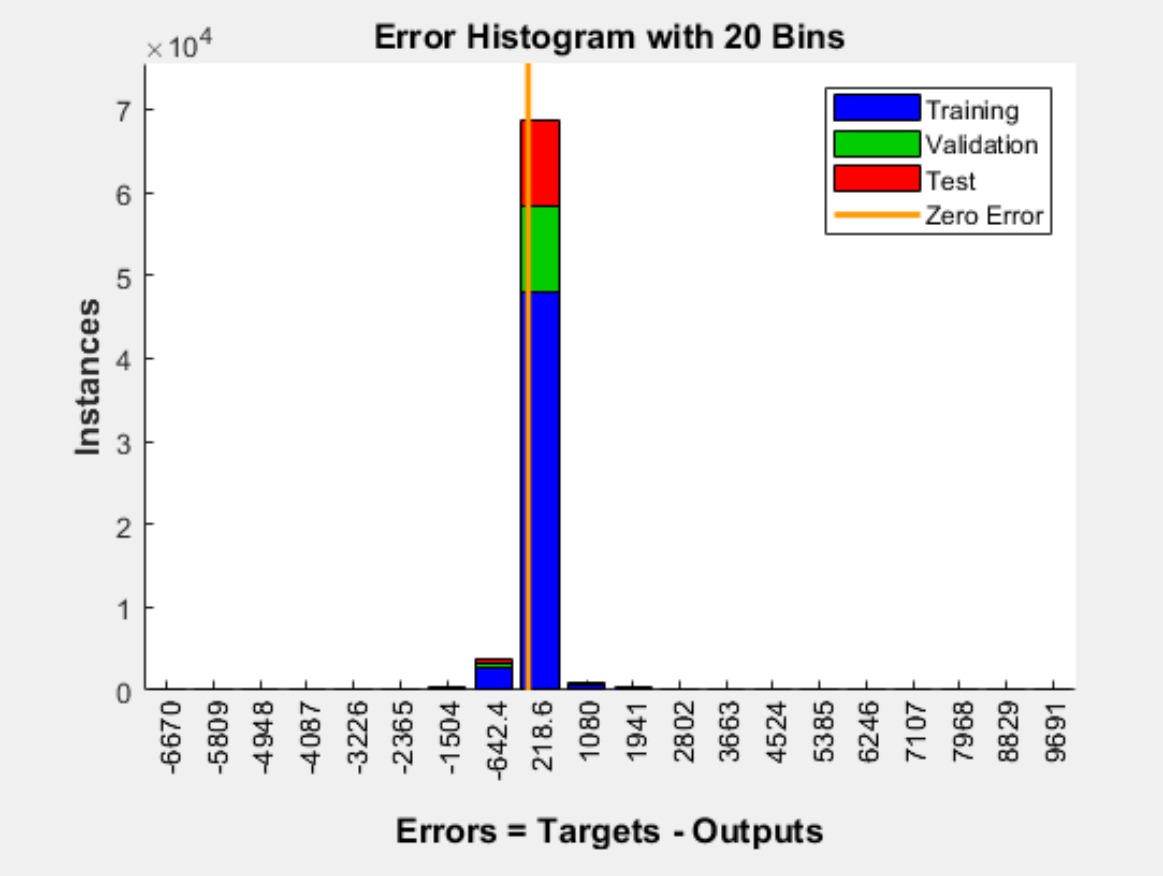

Figure 4-9 GAM location results (Jordan) – NARX error histogram

Figure 4-7 shows the accuracy of the NARX model for Jordan data, presenting the results for the training, validation, and testing data. In addition, Figure 4-8 clarifies the MSE and R values, which precisely measure the accuracy of the NARX model. Additionally, the error histogram (Figure 4-9) visualizes the prediction errors (the difference between predicted and actual values).

In comparison, it can be seen that the NARX model achieved good results. The accuracy is higher for the Jordan data (around 0.98) compared to the England data (around 0.97). This difference is due to the variation in the number of parameters and the size of the data used. This comparison provides valuable insights into the model's behaviour under different parameter setups and data sizes.

4.5.1.2 DNN LSTM Models

Table 4-9 and 4-10 present results for DNN England and DNN Jordan

Table 4-9 DNN (LSTM) results – England Data

Model	Data Part	RMSE Value
DNN (LSTM)- England	Training Data	**51.521**
DNN (LSTM)- England	Testing Data	**53.371**
DNN (LSTM)- England	Validation Data	**52.121**

Table 4-10 DNN (LSTM) results – Jordan Data

Model	Data Part	RMSE Value
DNN (LSTM)- Jordan	Training Data	**38.345**
DNN (LSTM)- Jordan	Testing Data	**79.814**
DNN (LSTM)- Jordan	Validation Data	**74.190**

4.5.2 Markov Chain Results

Table 4-11 and 4-12 present results for Markov England and Markov Jordan models. Table 4-13 compares the Markov and DNN models for both England and Jordan Data.

Table 4-11 Markov results – England Data

Model	RMSE Value
Markov- England	**11.1347**

Table 4-12 Markov results – Jordan Data

Model	RMSE Value
Markov- Jordan	**15.6624**

Table 4-13 DNN and Markov comparison table for England and Jordan Data (test data)

Model	England RMSE	Jordan RMSE
DNN	53.3712	**77.7665**
Markov	11.1347	**15.6624**

4.6 Discussion of Outcomes

DNN models were developed for the England Data and Jordan Data, with some variations between the parameter values in both models, depending on the needs, data size and parameters (as explained in Chapter 3). Two LSTM layers were used for each model, one sequence input layer, two drop out layers (0.3), a fully connected layer and a regression layer. Training options for the models were execution environment (CPU), L2Regularization, Mini Batch Size, and Max Epochs. The details of the structure of the DNN models can be found in Table 4-5. Parameters' values were specified following Zhou *et al.* (2019), based on systematic experiments and observations. Parameters were optimized in iterative reviews through several trials.

4.7 Summary

Predicting air quality is challenging because of the complexity of its processes and the strong coupling across all parameters, which is more complex in some gases than in others, such as with PM. Limited data access in some regions is a problem, and missing data from monitors is a common occurrence; further methods and validations for data replacement/removal accuracy are necessitated by such conditions. Furthermore, generating accurate results in light of data factors is inherently more challenging in dynamic systems.

Adding L2 Regularization layer to DNN model improved the model results. Tweaking the mini batch value to suitable value depending on the mode (between 32 and 1024) improved the model results. Epoch value plays a role in enhancing results, which depends on the model as well (Zhou *et al.*, 2019). Hyper-parameters tuning for the models proved increase in overall model performance, with consideration to data points, input and output numbers (multivariate input and output). Experiments were performed with hyper-parameters tuning, with a view to achieving suitable accuracy to fulfil the objectives of this study. The following chapter presents the results for hybrid air quality modelling.

Chapter 5
Hybrid Air Quality Modelling

5.1 Introduction

This chapter presents the proposed hybrid model, combining the stand-alone models discussed in Chapter 4, with further details about the structure. Model selection is presented and the results between algorithms are presented and discussed in different ways, with introduction of the selection based on the best accuracy. Comparison tables with error and regression data are presented. This chapter discusses model selection and lists optimization for already developed models using different algorithms, and summarizes tweaking and refinements of parameters that can help in improving accuracy.

5.2 Overview

Modelling for accurate results needs further identification for hybrid methods, to produce suitable results in support of the aims and objective of the study. This chapter presents experimental outcomes of using non-conventional hybrid modelling. The model was built using DNN and Markov chain model (Markov-switching dynamic). The initial assumption for the experiment is on the basis of developing a deep learning model which could provide capacity to the complexity of the multivariate inputs and outputs for the AQP model, and then simulate results using a simple linear method. Markov chain was selected for this purpose, to backup data and compensate for any losses, errors, redundancy and external factors.

5.3 Model Architecture

The deep-learning modelling architecture was built using methods and techniques discussed in the previous chapter. Due to the advantages of the DNN model in Big Data prediction, many trials were performed, using tailored parameters to fit the data requirements, to produce appropriate accuracy for the predictions. A DNN model with two LSTM layers was used as the first method of predicting air quality for the selected data. A separate experiment was performed using Markov-chain modelling, and then hybrid modelling was developed, so that the test data was fed to the Markov model. This produced the required outputs and gave an indication of appropriate levels of accuracy.

By its nature, the Markov model requires data to be prepared in a certain way, so the Markov-switching regression (Kim, 1994) was tailored to this particular research, and was

treated in a special way to fulfil the specific aims of the model. The initial input consisted of multiple inputs of eight parameters. When preparing to feed the Markov model with data, the dataset was split randomly in a ratio of 0.8:0.1:0.1 – 0.8 for training, 0.1 for testing, and 0.1 for validation. The data was then filled with move-mean method, and then previous value method, to back up any missing values. Indexing for both the input and output data was done in such a way as to treat each input and output as separated parameters. This procedure was based on previous trials, in which it was attempted to replace missing values on the run time using different methods.

The method described gave good results when it was applied to the England Data and Jordan Data. This method of replacing the data was also suitable for hourly data, as the first replacement of mean values through the move-mean method was aligned with the data frequency. Because of the nature of the data selected, there was a realized pattern of values, which were relatively close to each other's readings on many occasions. This was due to the impact of weather conditions as a collective atmospheric effect, and the steep increase or decrease in associated values. Move-mean was chosen as the primary method of replacing missing cover values, as it can give near-average replacements. It should be noted here that the data contains negative temperature values, which were not manipulated, as they represent the reality of weather conditions and sub-zero temperatures, particularly in winter.

The Markov model was based on several input models. Each input was represented by an ARIMA model, which was built using a number of variables: AR (auto regression coefficient), beta (regression coefficient); constant (mean); and variance (standard deviation). The AR variable for each input was calculated by using the (corr) function for each input and output, and then the mean was taken as the result of the correlation (the value being used for each input ARIMA model). The beta was a fixed value of 1, and the constant was a fixed value of zero. The standard deviation changed according to each input value. An std function was used for each input: std (Input1), std (Input2), std (Input3), std (Input4), std (Input5), std (Input6), std (Input7) and std (Input8).

A DTMC object was used for the switching-technique Markov-switching dynamic regression model msVAR object, which stored the parameter values of the model. The DTMC object took the *P* parameter as referred to for the probability of the transition. When the output values stored in the Mdl variable were then simulated using the simulation function in MATLAB, which took the saved Mdl representing the input side. Subsequently, a number of observations (referring to the number of data rows used in the experiment) and the output were produced.

A simulation object was used for each output, and it should be mentioned here that the output was named Training Data in the code, so that TrainingData1 represented the value of Output1, TrainingData2 represented Output2, TrainingData3 represented Output3 and TrainingData4 represented Output4. The probability transition was created in eight different states, based on the eight input values. The assumption for the probability matrix was to produce an 8-by-8 matrix between zero and one, by generating a random number from a uniform distribution in range (0,1). All new outputs were represented using Training Data, as all simulated outputs are stored in this variable. The transition probabilities linked each state to the next one; the earlier described model created a Markov-switching dynamic regression model, which supported the dynamic behaviour of the time series through the set of state transition probabilities. ARIMA and msVAR were used to create the dynamic regression model.

The DNN and Markov model were trained using the methods described above. Trained models were saved appropriately, and the resultant output of the DNN was fed to the Markov model as a new output, and the values were predicted using the previously trained model parameters. The new output represented the predicted values for the hybrid model (both DNN and Markov). Data manipulation was performed in order to execute the hybrid model. This was done by using the output data of the model as the output for the first (i.e., DNN) model. The resulting values were then used as the output of the previously trained Markov model, as test data was used alongside other data in this experiment. A third source of data was considered to be external to the other data. This was used to predict the output, in order to validate the model and show how well it would perform with new data.

The modelling results are presented in section 5.4. To perform the hybrid modelling, DNN model that was presented in Chapter 4's saved results was called to the workspace using the import feature, and the already saved Markov model results were called. The test partition of the data that was predicted using DNN model was fed to the Markov model as new output for the latter, and the already saved setup for Markov was used to simulate outputs. The same methods were used for both the England Data and Jordan Data.

5.4 Experimental Results

5.4.1 DNN and Markov

As Tables 5-1 and 5-2 demonstrate, the accuracy of the hybrid models in the selected locations in England and Jordan is better than that of the DNN and Markov models. The hybrid models provided good accuracy in both experiments. Moreover, the performance of

the models was validated using the new data. Improved accuracy was also noticed when the same hybrid models were used. This shows that they are preferable to the standalone models, in the light of the multivariate data from both England and Jordan. This study shows that combining two models supporting the time-series nature of air quality data has enhanced the experimental results. The first experiment was performed to obtain appropriate results for each individual model. The hybrid model was then applied to the experiment to achieve the required level of prediction accuracy. In comparison, the overall performance improved using hybrid modelling. Experiments of this kind are recommended when using Big Data for prediction, especially when modelling limitations arise (Zaini *et al.*, 2022).

Table 5-1 Modelling results: England

Model	RMSE
DNN	53.371
Markov	11.134
Hybrid (DNN and Markov)	9.889

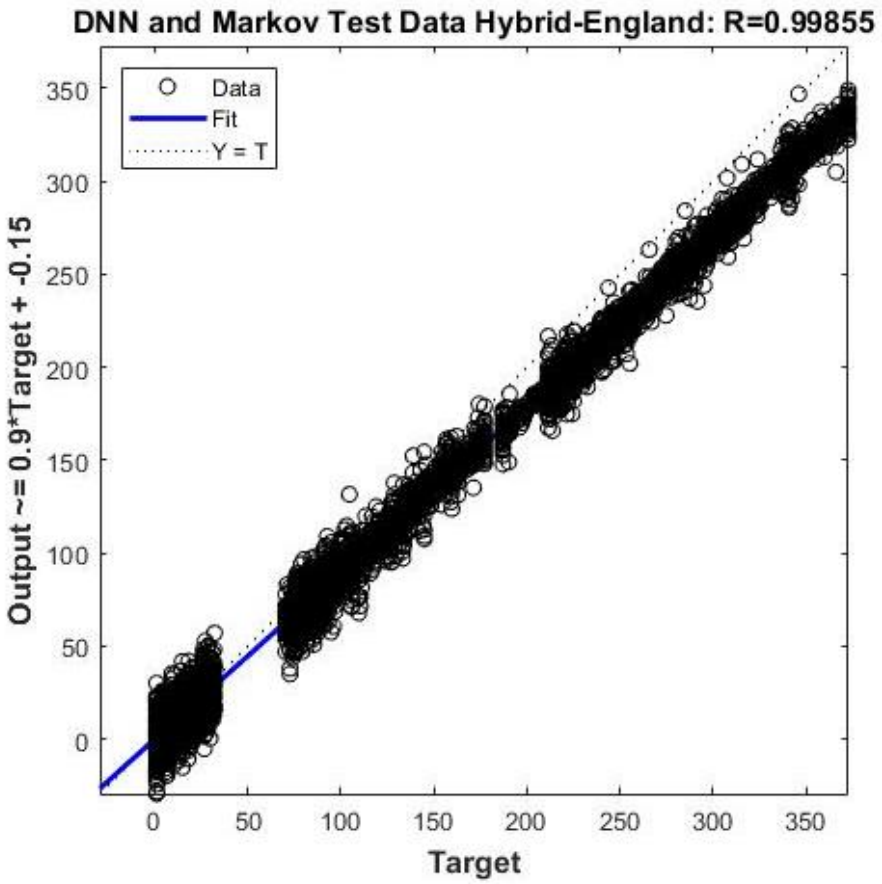

Figure 5-1 England air quality index prediction process

Table 5-2 Modelling results: Jordan

Model	RMSE
DNN	77.7665
Markov	15.662
Hybrid (DNN and Markov)	14.877

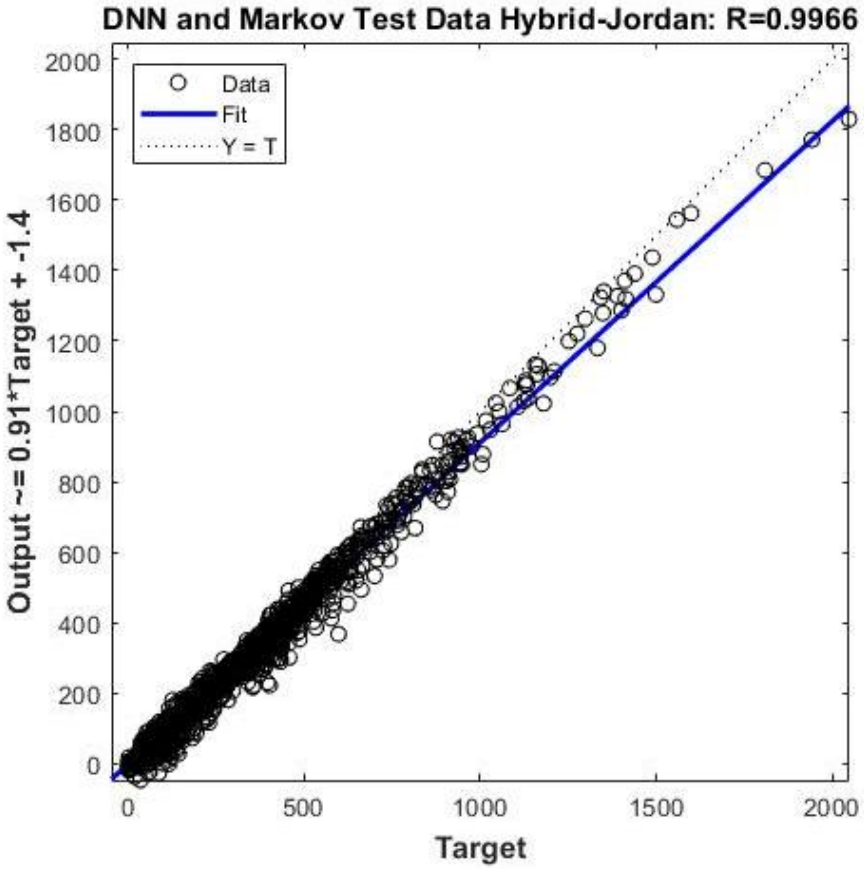

Figure 5-2 Jordan air quality index prediction process

Table 5-1 presents the RMSE for the standalone models DNN, Markov, and the hybrid model (DNN and Markov) for England data. Table 5-2 presents the RMSE for the standalone models DNN, Markov, and the hybrid model (DNN and Markov) for Jordan data. Both tables illustrate that the hybrid model outperformed the standalone models, indicating that using the hybrid model is recommended due to its superior performance with different datasets. Figure 5-1 and Figure 5-2 show regression R for England and Jordan data as listed in Table 5-5.

To validate the performance of the models, new data was selected from a data source that was not used in the experiment, in order to validate and evaluate the models, and calculate the error for the models, using RMSE. The first stage of the validation was data preparation

from the new source. Data was selected with similarities to the original data specifically to fulfil the requirements of the study, to ensure that the model would perform well with similar studies and data, and show that it was reliable. Data was selected from the Italian Data source (as explained in Chapter 3), which resembled the data used to build the models mentioned above.

After preparation of the input data, the data was partitioned to fit the number of rows selected for the test data. The new input data was then fed to the DNN model (after the previously saved DNN model results were loaded to the MATLAB workspace). A run of the prediction was then performed with the current settings, but without retraining the DNN (the pre-saved model set-up was used). Finally, the new predicted result was fed to the Markov model (after the previously saved Markov model results were loaded to the MATLAB workspace), and the results from this run were considered for the validation of the hybrid modelling results.

A new source of data was used to validate the Jordan modelling: KHG location data, provided by the Ministry of the Environment. Firstly, the DNN and Markov models were trained and each set of results saved separately. The externally sourced test data was predicted (fed) first to the DNN and then to the Markov. The results were validated using the new Italy Data. The output was used to perform a new DNN run, using this data source.

The accuracy rates of the hybrid models for both England and Jordan Data were better than those of the stand-alone DNN and the stand-alone Markov models. The hybrid model showed good accuracy in both experiments. Its performance was validated using new data, and it was found to achieve greater accuracy compared to the standalone models (see Table 5-3 and 5-4). Furthermore, Table 5-5 shows hybrid model performance evaluations (R and MAE).

Table 5-3 Modelling validation (new data source from Italy): Jordan modelling

Model	**RMSE**
DNN	57.494
Markov	10.486
Hybrid (DNN and Markov)	7.866

Table 5-4 Modelling validation (new data source from Italy): England modelling

Model	RMSE
DNN	113.389
Markov	18.702
Hybrid (DNN and Markov)	15.277

Table 5-3 and 5-4 show the model's generalization ability by achieving good accuracy using a completely new source of data (Italy data). The results displayed in Tables 5-3 and 5-4 demonstrate the efficient use of the DNN–Markov model in larger datasets, as DNN–Markov outperformed the standalone models in the case of England and Jordan data. Hence, within the scope of this study, the size of the data for a multivariate setup must be considered. There must be a good balance of trade-offs (performance, training time, accuracy, etc.) when selecting the algorithm for air quality prediction.

Table 5-5 Hybrid models' performance evaluation

Model	R	MAE
Hybrid (DNN and Markov) – Jordan	0.9966	8.2426
Hybrid (DNN and Markov) – England	0.9986	14.6901

Table 5-5 shows a comparison between Jordan and England models using performance evaluation metrics R and MAE.

5.4.2 Experimental Scenario Results

This section highlights some of the experimental results produced throughout the study. Several factors were instrumental in obtaining different results, such as the architecture (topology) of the model (e.g., the number of DNN model layers and type affected results). Furthermore, hyper-parameters tuning is of major impact to the overall model performance. In multi-input and multi-output deep learning (Zhou *et al.*, 2019), L2 regularization can be used to avoid over-fitting issues by optimizing weight parameters. In line with this, this research adopted the implementation of such an algorithm within the DNN architecture, which achieved a leap forward in enhancing the results.

Moreover, mini-batch gradient descent and dropout neurons were of equal importance to the developed model in this study. Tweaking the mini batch value to be something in between 32-1024 is recommended (Zhou *et al.*, 2019), depending on the model architecture and number of layers needed based on the datasets to be modelled. Chapter 4 explained the set

of parameters selected for this study, based on several experiments, which produced different results for the scenarios expounded below. Section 3.3 explained the methods or scenarios followed. The results obtained from the experimental of some scenarios are discussed below.

There is obvious evidence from previous studies that there is a need for models that provide reliable high accuracy prediction while dealing with large datasets (Tealab, 2018; Ma *et al.*, 2019; Siami-Namini, Tavakoli and Namin, 2019). Additionally, there is a lack of ML models giving reliably accurate forecasting when predicting using Big Data, and put simply there is an absence of reliable methods. The use of Bi-LSTM as a method to improve prediction accuracy in the presence of Big Data has been suggested (Ma *et al.*, 2019; Siami-Namini, Tavakoli and Namin, 2019). However, Siami-Namini, Tavakoli and Namin (2019) claimed that there are still gaps in the area of BiLSTM itself, especially for multivariate time series, and encouraged further specific research to address this. Following the mapped needs claimed in different research studies expounded in Chapter 2, this research conducted multivariate series with Big Data, using the Bi-LSTM method to test for the same setup as presented previously for DNN (LSTM).

The results displayed in Table 5-6 and 5-7 show that there might be efficient use for Bi-LSTM in larger datasets, as Bi-LSTM out performed LSTM in the case of England Data, but it appeared that LSTM outperformed Bi-LSTM for Jordan Data. Hence, in the scope of this study, there must be consideration for the size of data when using Bi-LSTM for multivariate setup, especially as training takes a lot of time for BiLSTM. There must be a good balance for trade-offs (performance, training time, accuracy, etc.) between using BiLSTM and LSTM.

Table 5-6 BiLSTM England Data

Model	Data Part	RMSE Value
DNN (BiLSTM)- England	Training Data	51.338
DNN (BiLSTM)- England	Testing Data	52.698
DNN (BiLSTM)- England	Validation Data	53.4368

Table 5-7 BiLSTM Jordan Data

Model	Data Part	RMSE Value
DNN (BiLSTM)- Jordan	Training Data	39.223
DNN (BiLSTM)- Jordan	Testing Data	80.0189
DNN (BiLSTM)- Jordan	Validation Data	75.5875

As it can be seen from the different experimental scenarios presented in this thesis, after several trials, DNN-Markov proved to offer the best results. Even the order of the experiment made a difference; for instance, trials for the Markov-DNN model were performed, meaning the Markov outputs were used to feed DNN, but in fact DNN-Markov proved performance. For example, in other scenarios, the mean of the two models (DNN and Markov) was used, and some other trials such as residual error and others (more scenarios are explained in section 3.3).

5.5 Summary

The results of the Markov model represent the linear part of the hybrid model by simulating the data, as discussed in Chapter 3. This model produced good results as a stand-alone. However, a hybrid model was used in this chapter, in an attempt to improve accuracy. Hybrid (Markov and DNN) model results proved to be satisfactory for the objectives of 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 research. There was a marked 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 AQP.

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 (Wang *et al.*, 2019), the direction of this study suggested that the hybrid (Markov-LSTM) model would produce improvements, and the experiment demonstrated this. A forward-looking AQI was developed further, as presented in Chapter 6, when appropriate levels of accuracy in reference to AQP have been achieved. Further methods of implementing the Markov and DNN hybrid are explored to fulfil the aims and objectives of this study, while other models are also investigated to see if they also improve accuracy. As discussed in the experimental summary, some modelling 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 was conducted using case studies (with data from another source) to test the models for further developments.

The study conducted several methods in an effort to contribute with a new and efficient approach to predict AQI for the next hour; DNN-Markov approach was selected as validation proved the model's performance in terms of promising potential for efficiency. It is also efficacious to deploy a simple linear model, enabling backup in the fact of complexity and potential losses that could occur with the DNN model, which thereby boosts performance. Among its main contributions to knowledge, this work proposes an hourly prediction model, with multivariate input and output models supporting the complexity of AQP. It proposes a hybrid model, combining Markov and DNN models considering static and dynamic variables for accurate results and AQI representation.

The developed solution offers hourly generation of the AQI model, to produce more accurate results, and improved access for added value for decision makers for the selected regions (especially concerning the data for Jordan). The research considers the transportation factor (share of transportation emissions), and addresses data refinement and model accuracy by generating a model to cover such challenges (such as missing data and reducing noise). It proposes the best combination of tested models to cover complex gases that are currently creating challenges in prediction, such as PM.

The proposed multi-input multi-output hybrid Deep Neural Network Markov (DNNM) model achieves 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 AQI by comparing results in two selected areas: England and Jordan (i.e., London and Amman). The following are the main objectives of the proposed solution:

- Reduced data complexity processing through selecting the best ML methods to support air quality analysis.
- Increased reliability and accurate modelling to predict air quality.
- An effective AQI model for policy and regulation, supporting health and climate change issues.
- Considering transportation/traffic factors.

Chapter 6
AQI Framework Using Neuro-Fuzzy Logic

6.1 Introduction

This chapter builds the AQI levels representations using Adaptive Neuro Fuzzy Logic (ANFIS). First, the results of the selected model from hybrid modelling with most suitable results (as explained in the previous chapter) are used, and then output of air quality levels is represented using Fuzzy Logic, following the air quality standards which discussed in the next section. The EPA (2023) criteria were selected as the standard to follow in designing the fuzzy logic.

6.2 EPA Air Quality Standards (2023)

Air pollution has affected many aspects of life, such as health, where reported respiratory irritation issues are increasing (Coelho *et al.*, 2021). The first AQI was developed by the EPA as a response to the major economic, health and environmental consequences of this (Bishoi, Prakash and Jain, 2009). In comparison to the PSI and AQI, the RAQI has been found to give good results. There has been a limitation to existing research, due to the cost of developing an AQI system for $PM_{2.5}$ (which refers to particulates – tiny particles or droplets in the air) from the base PSI. This required further developmental research into systems to cover other literature gaps in the field. There are many standards available; however, this research proposed the development of a neuro-fuzzy-logic to support the boundary areas of the AQI as a further enhancement for representing air quality levels based on EPA standard as the most suitable one for this study.

This research has industrial significance, as air pollution has affected many aspects of life, the most egregious of which is health, with increased prevalence of respiratory irritation issues (Coelho *et al.*, 2021). The first AQI was developed by the EPA as a response to the major economic, health, and environmental consequences of industrial activities (Bishoi, Prakash and Jain, 2009). The EPA standard is adopted in this study to represent air quality levels as it has been thought of as the most standard that could be represented across cities from the available standards.

6.3 Fuzzy Logic

Fuzzy logic can be considered as decision making tool, and it is a subset of the intelligent system field; it is used in the simulation of non-linear behaviour using the fuzzy logic framework. Despite its name, it is actually more of a *precise* logic for rational decisions in light of uncertainty (Singpurwalla and Booker, 2004). As explained previously, fuzzy logic was originally proposed as a solution to handle uncertainty by approximation (Baatarchuluun, Sung and Lee, 2020). A fuzzy system includes a membership function, which can be in different curve shapes (trapezoidal, triangular, or Gaussian); the curve shows the connection between each input point and value between 0 and 1 (Sowlat *et al.*, 2011).

This study shows the importance of using fuzzy logic, whereby the latter can give approximations, which is an added benefit to the already developed models. This is because when data is collected there is a need for data replacement, and data filling following different strategies makes data more accurate in terms of values. The subsequent deployment of fuzzy logic as the last stage is useful for boundary areas, and also gives approximation for the values within the specified rules, which makes the whole system more reliable for use in AQI.

6.4 AQI Experiment

6.4.1 AQI (England)

The AQI for England was produced using the following method:

- Firstly, the most accurate output was selected from the predictive modelling.
- The units were converted for some gases to comply with the requirements of the EPA (see Table 6-1).
- The maximum value of each gas was determined using loop and max functions (showing which gas had the highest value at the specified point in time).
- The AQI was found for each gas concentration at the specified point of time, according to the EPA standards (representing the AQI levels).

Table 6-1 Conversion of units for emissions

Unit of emission	EPA Unit	Conversion
CO (mg/m^3)	mg/m^3	24.45 × CO concentration /28.01
NO (µg/m^3)	µg/m^3	24.45 × NO concentration /30
NO_2 (µg/m^3)	µg/m^3	24.45 × NO_2 concentration /46.006
NO_x (µg/m^3)	µg/m^3	24.45 × NO_x concentration /46.006
O_3 (µg/m^3)	µg/m^3	24.45 × O_3 concentration /48.0
SO_2 (ug/m^3)	µg/m^3	24.45 × SO_2 concentration /64.06

6.4.2 AQI (Jordan)

The air-quality index for Jordan was produced using the following method:

- Firstly, outputs were selected from the predictive modelling.
- The units were converted for some gases to comply with the requirements of the selected EPA's AQI standard.
- CO was the only gas reading in the Jordan Data that needed unit conversion (from ppb to ppm, dividing the values by 1000).
- The maximum value of the gases was found using the loop and max functions (to determine which gas had the highest value at that point in time).
- The AQI could then be found for each gas concentration at the specified point in time, following the EPA standards (representing the AQI levels) (EPA, 2023) (see Figure 6-1). PM did not require any conversion, as all the units for the collected data matched the relevant EPA unit.

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*...air quality conditions are:*	*...as symbolized by this color:*
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Figure 6-1 EPA Health AQI

Source: EPA (2023)

6.4.3 AQI Calculation Flow for England and Jordan

After finalizing DNN-Markov models, the final results selected for the models for both England and Jordan were saved, and were then loaded as preparation for AQI calculations. As the selected AQI standard is EPA, there is a need to convert the units of the results to the matching unit in EPA standard, which will make the categorizing the AQI level feasible, based on the value range of index (Figure 6-1 shows the EPA standard specification). After the correct conversion and based on the needed steps, as detailed in sections 6.4.1 and 6.4.2 the data was assigned to the relevant category level, marked from 0 to 7 in the MATLAB code, as shown in the following flow chart (Figure 6-2) (e.g., 0 when it is less than zero, and 7 when it is more than 500).

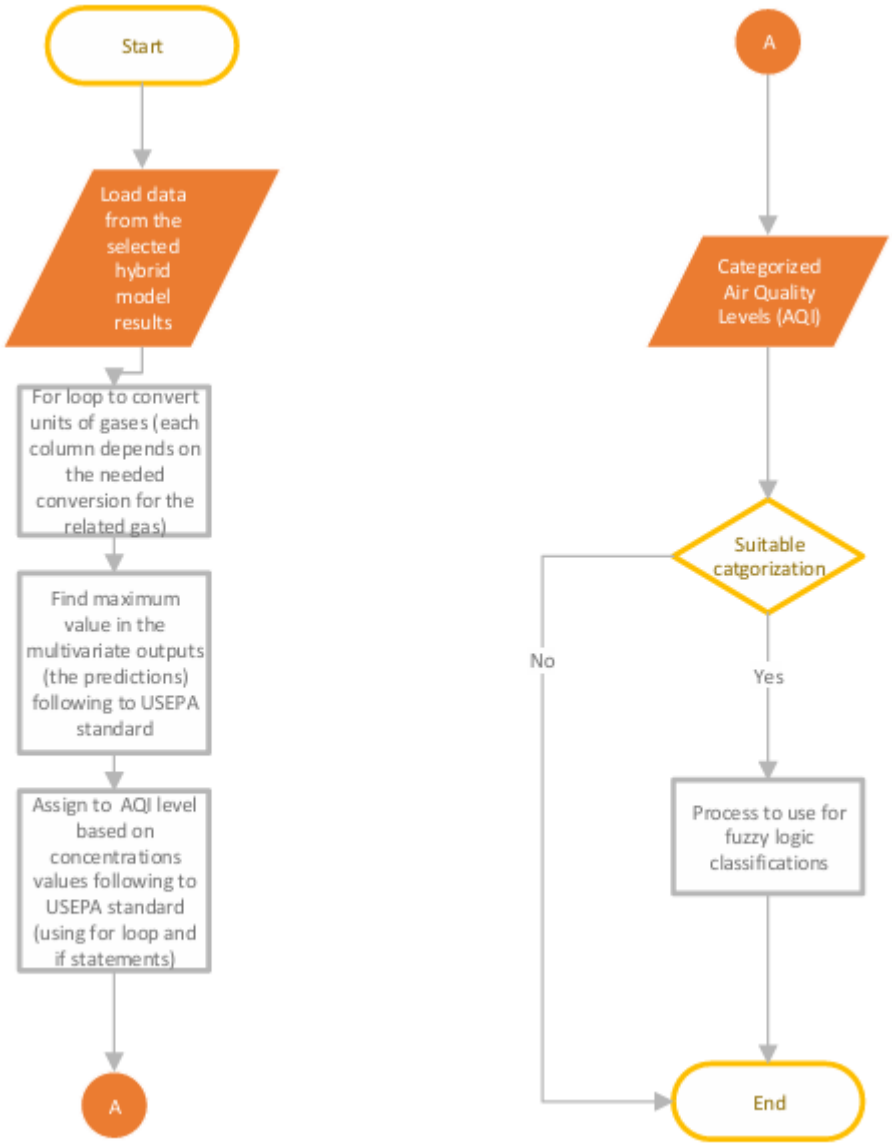

Figure 6-2 AQI – Levels calculations flow chart

First of all, the outputs were prepared in two parts: one for fuzzy logic (ANFIS) training and the other for fuzzy logic model testing. Training data consist of the raw data that were collected from monitors after pre-processing and preparation, and then each row was assigned to a specific air quality level following the United States Environmental Protection Agency (USEPA) standard, using the flow logic shown in Figure 6-2. Testing data consist of the predicted outputs explained previously, with the hybrid model and each row then being assigned to a specific air quality level following the USEPA standard, using the flow logic in Figure 6-2.

The training data and test data were used to train the fuzzy logic model in iterations, and then the fuzzy logic was evaluated using the predicted output data to obtain the resultant

levels from the fuzzy logic evaluation. Afterward, the actual (air quality levels based on the USEPA assignment) and predicted (air quality level from the fuzzy logic) classification were represented using ROC and confusion matrices (as reported in the following section). Figure 6-3 shows the whole AQI prediction process presented in this paper as part of continuous research performed in the air quality prediction field in this work.

After finalizing the DNN–Markov models, the final results selected for the models for both England and Jordan were saved and were then loaded in preparation for AQI calculations. As the selected AQI standard is USEPA, it was necessary to convert the units of the results to the matching unit in the USEPA standard to make categorizing the AQI level feasible based on the value range of the index. After the correct conversion and based on the needed steps, the data were assigned to the relevant category level, marked from 1 to 7 in the MATLAB code (e.g., 1 when it is less than zero and 7 when it is more than 500).

The air quality index prediction process (Figure 6-3) summarizes the collective experimental framework of this work, consisting of three main parts: the DNN, Markov, and fuzzy logic models. The experiment of this framework started by using the selected results from DNN standalone model-test data, as described earlier in this section. These data were then used as output data for the Markov model, and a run was executed for the already saved results for the Markov standalone model using the test output data from DNN. Afterward, the data from the hybrid (DNN–Markov) model were fed to the fuzzy logic model, and the model results were evaluated.

6.5 Neuro-Fuzzy Logic AQI Prediction Framework

This section summarizes the collective experimental framework of this thesis, which consists of three main parts: the DNN, Markov, and fuzzy logic models (see Figure 6-3). The experiment of this framework starts by using the selected results from DNN standalone model-test data (see Chapter 4). This data is then used as output data for Markov model, and a run is executed for the already saved results for Markov standalone model (see Chapter 4), but using the test output data from DNN. Afterwards, the data from the hybrid (DNN-Markov) model (presented in Chapter 5) is fed to the fuzzy logic model, as described in this chapter (see section 6.4.3) and the model results are evaluated. Figure 6-3 presents a comprehensive overview of the whole framework of this study.

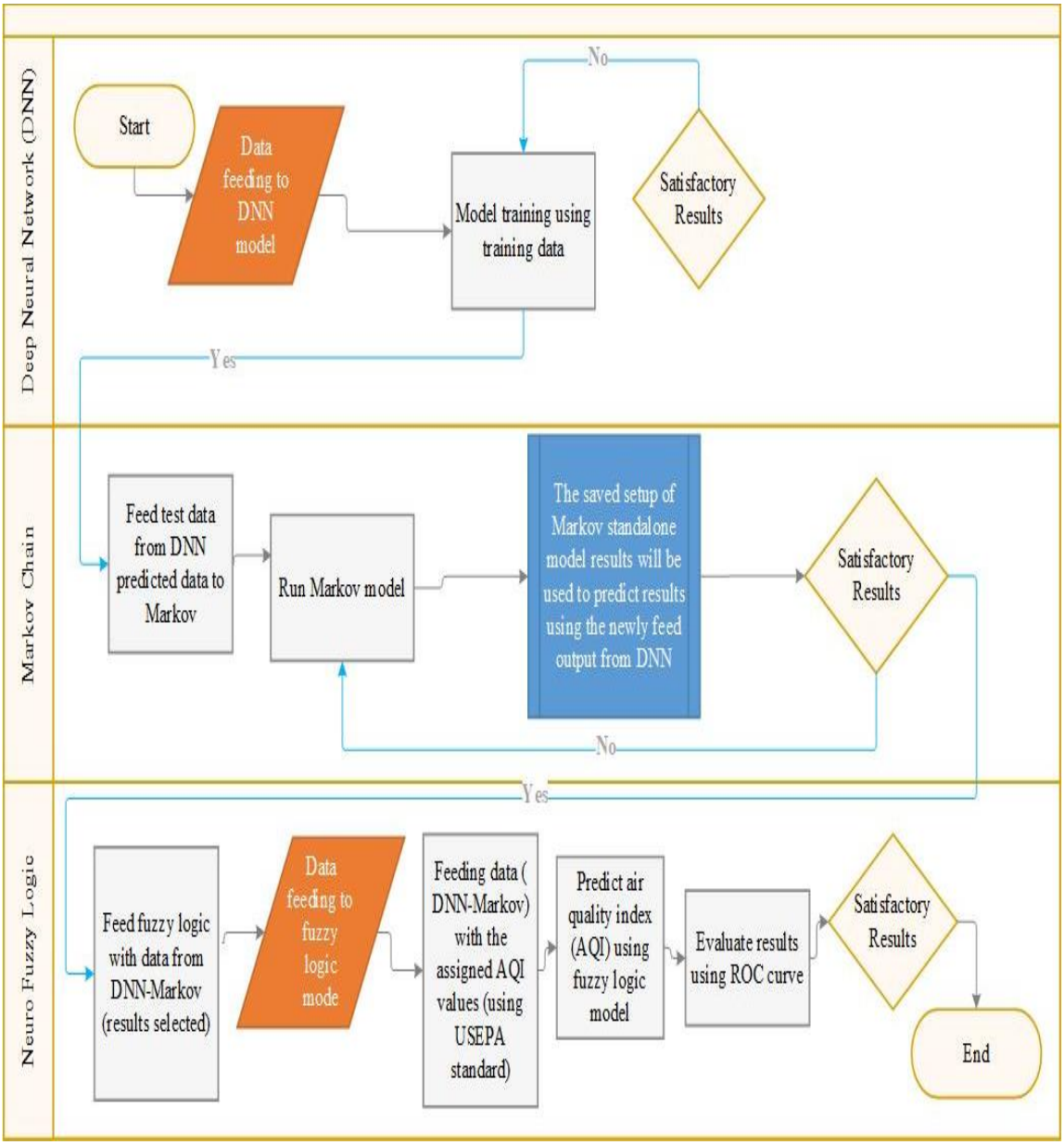

Figure 6-3 AQI prediction process

6.5.1 Neuro-Fuzzy Logic Representation

Neuro-fuzzy logic was used in this study to represent the AQI for the predicted measurements. The data inputs (for fuzzy logic) for England were: Input1 (CO), Input2 (NO), Input3 (NO_2), Input4 (NO_x), Input5 (O_3), Input6 (PM_{10}) and Input7 (SO_2). The data inputs (for fuzzy logic) for Jordan were: Input1 (PM_{10}), Input2 (NO_2), Input3 (CO), and Input4 (SO_2).

In this neuro-fuzzy logic model the outputs were considered to be inputs for the model. Firstly, for the training data, the initial outputs represented the inputs, while the outputs represented the AQI assigned to each value, based on the EPA levels. Secondly, for the testing data, the predicted outputs represented the inputs, while the outputs represented the

AQI assigned to each value, based on the EPA levels. For the model setup, a Gaussian (*gaussmf*) was used as the input MF (membership function) type, and a linear function was selected for the output MF (membership function) type.

For the purposes of illustration, CO was selected to check the AQI representation against all other inputs. The control surface in Figure 6-4 (representing England Data) shows the overall mapping between Inputs and Outputs. It can be seen that the output in this case (CO) is at highest value relatively when Input4 (NO_x) and Input5 (O_3) are high, which is almost AQI (4) on a scale from 1 to 7 for the AQI. In addition, it is clear that Input6 (PM_{10}) and Input7 (SO_2) are influencing the AQI levels. It can be concluded that different gases with different concentrations affect the output level of gases in light of weather conditions such as wind speed, wind direction, temperature, and humidity.

As can be seen from Figure 6-5 (representing Jordan Data), Input4 (SO_2) is greatly impacting the pollution level, with the highest value for AQI (5) on a scale from 1 to 7. Input 2 (NO_2) has a moderate influence on the AQI levels. Figure 6-8 represents neuro-fuzzy rules for the Jordan Data as an example, and it is used to evaluate the created rules to validate the fuzzy model. Two membership functions were used for each variable (see Figures 6-6 and 6-7).

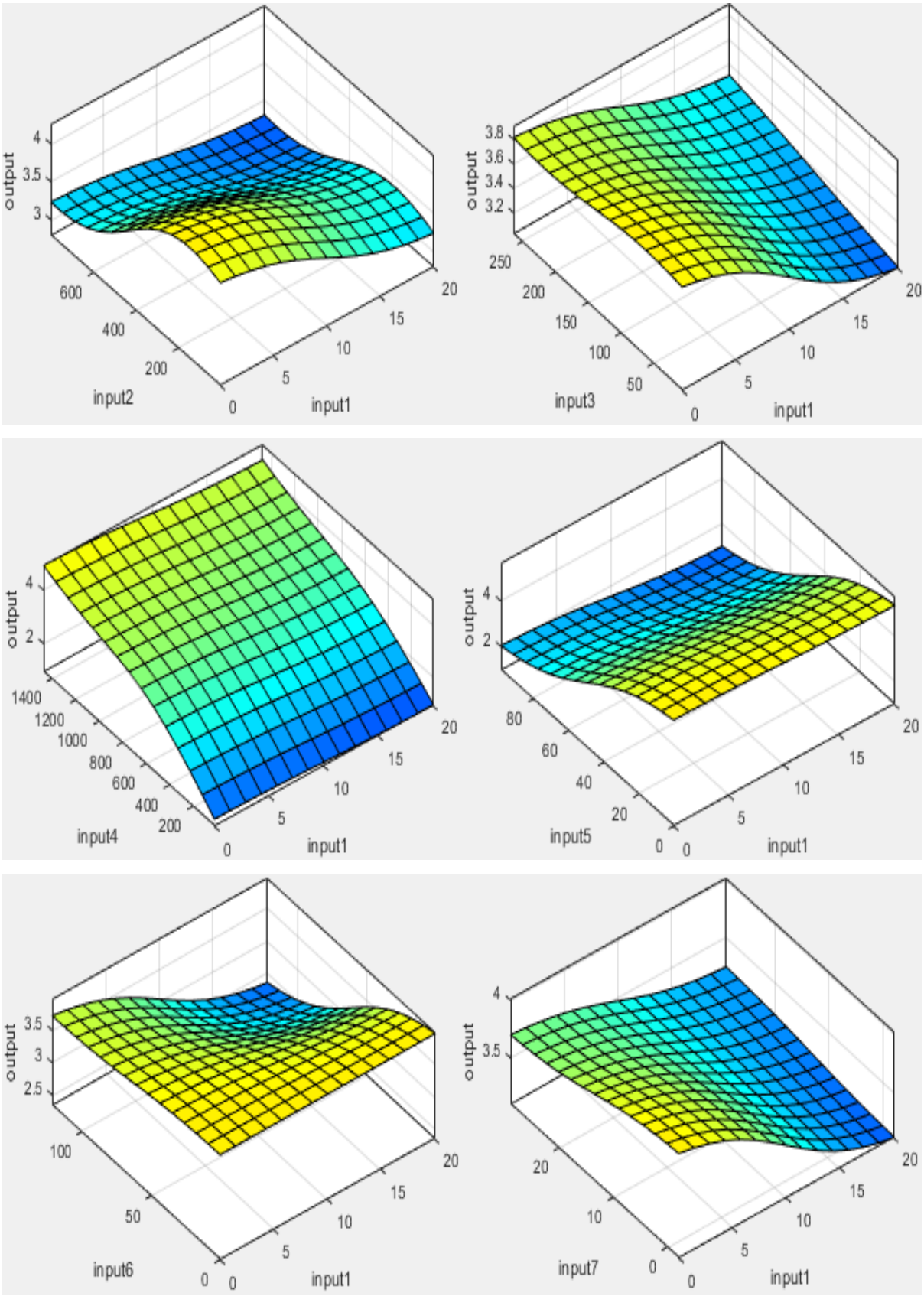

Figure 6-4 Neuro-fuzzy logic representing England AQI data

128 rules (sample representation of the first input (CO) and all other inputs)

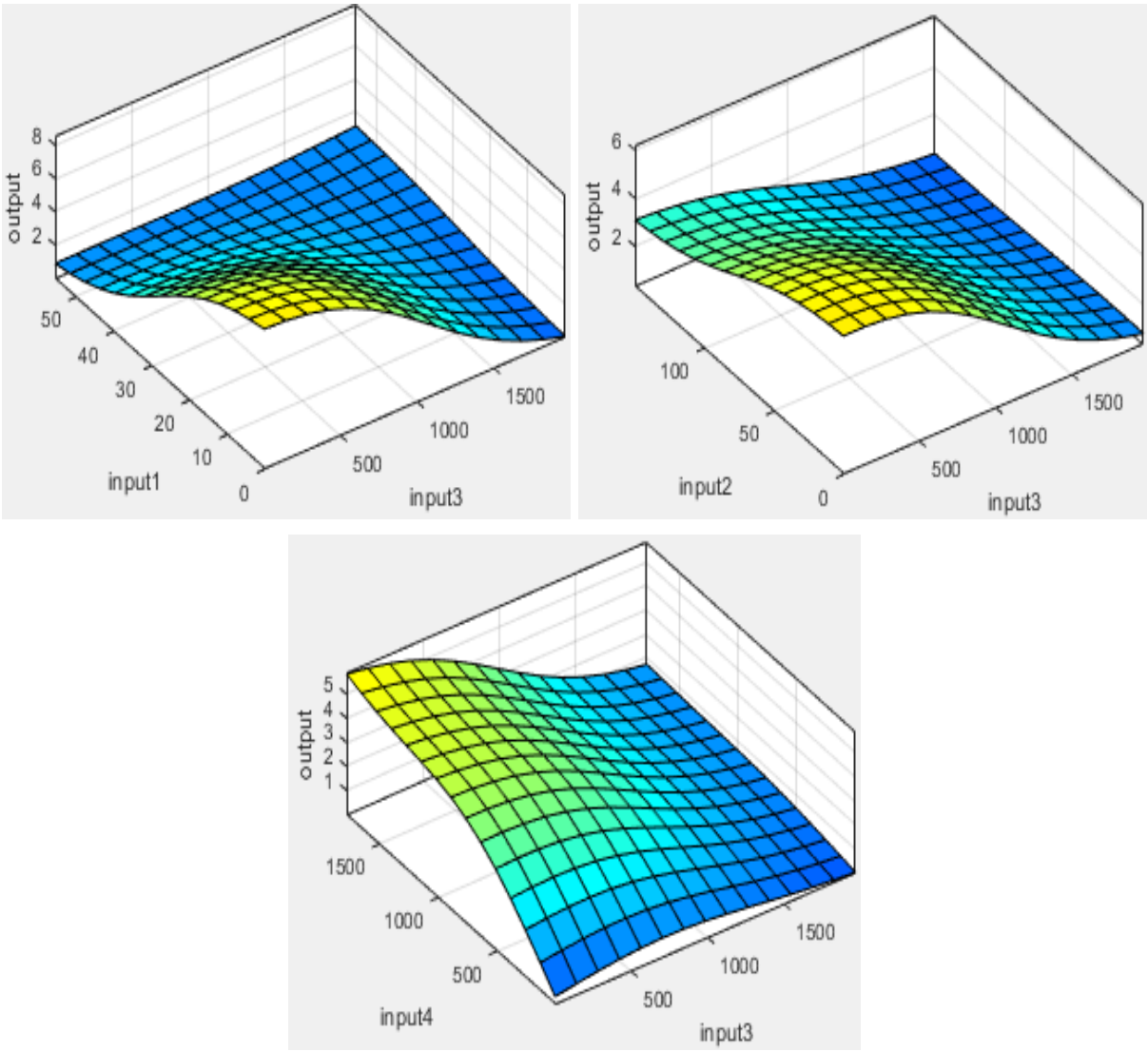

Figure 6-5 Neuro-fuzzy logic representing Jordan AQI data

16 rules (sample representation of the first input3 (CO) and all other inputs)

Neuro Fuzzy Designer (Fuzzy Logic Toolbox 2.7) (ANFIS) was used from MATLAB R2020a to model AQI. Data were split into AQI training data and AQI testing data. The training data consisted of the initial/raw data that was collected and processed, and then a new column created for the calculated AQI index using EPA USA standard method. AQI testing data consists of the predicted output from the hybrid models (as discussed in Chapter 5), and a new column created for the calculated AQI index using EPA standard method.

The first step to build the Neuro-Fuzzy logic for AQI is to load the training data (as per the above description for training data) and then to choose the above mentioned parameters (membership function, epochs, etc.). Data was trained using the loaded training data, and then test data was loaded (as per the above description for testing data), and the model was tested. Figure 6-6 shows the screen of the Neuro-Fuzzy Designer. Using the Jordan Data as

an example, the Neuro-Fuzzy logic structure is shown in Figure 6-7, and the Neuro-Fuzzy logic rules representation is shown in Figure 6-8.

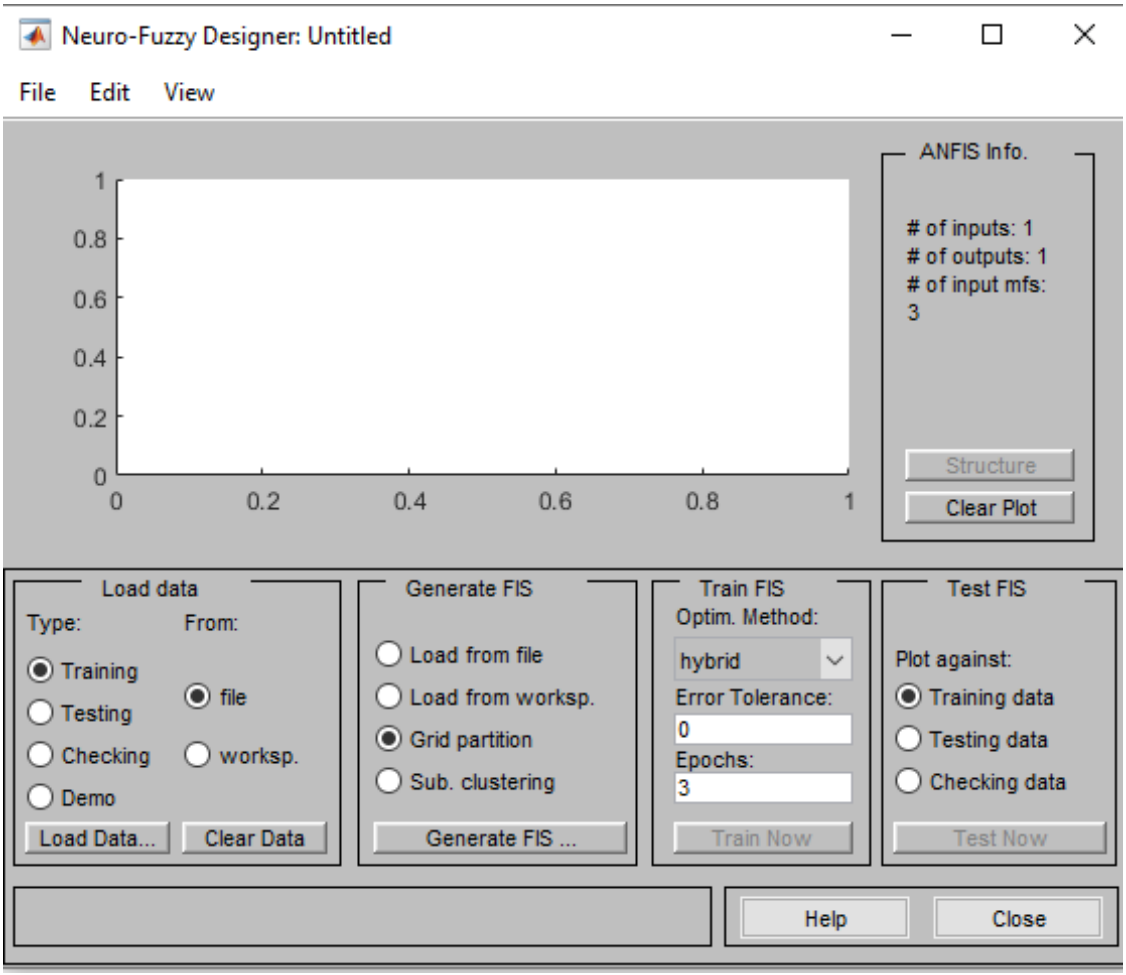

Figure 6-6 Neuro-Fuzzy Logic Designer Tool

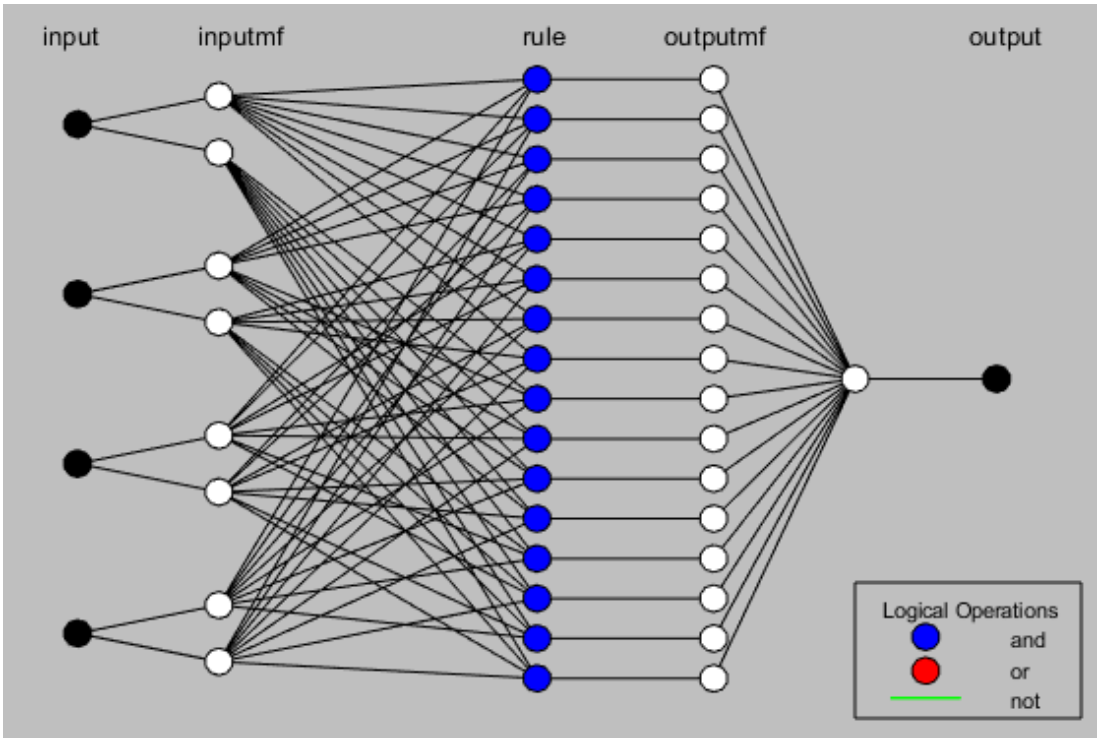

Figure 6-7 Representation of neuro-fuzzy logic structure (data from Jordan)

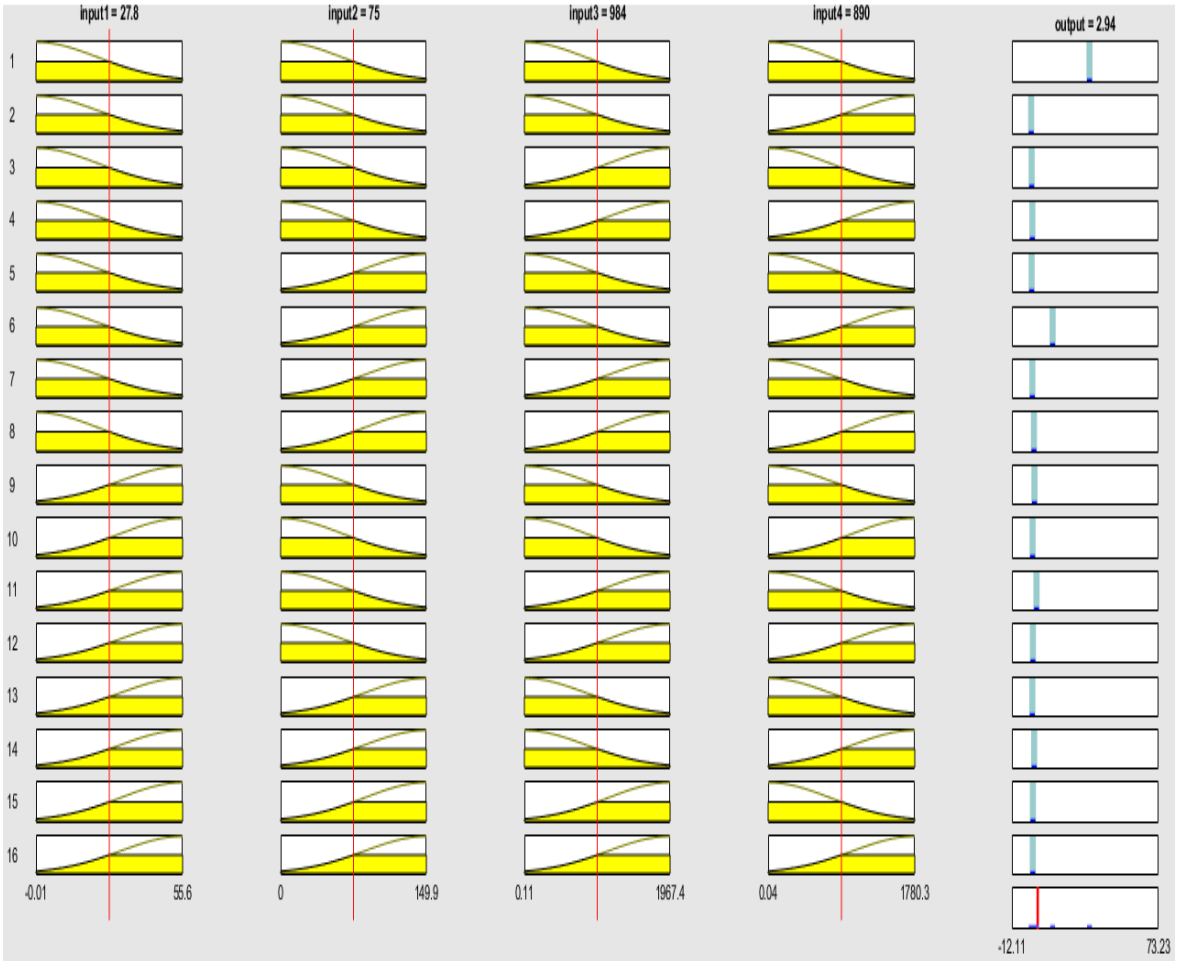

Figure 6-8 Representation of neuro-fuzzy logic rules (data from Jordan)

6.5.2 ANFIS Jordan Model

Figures 6-9 to 6-12 show the fuzzy logic results (ROC and confusion matrices) for the Jordan model, subject to the following data:

- Number of nodes: 193
- Number of linear parameters: 405
- Number of nonlinear parameters: 24
- Total number of parameters: 429
- Number of training data pairs: 2627
- Number of checking data pairs: 0
- Number of fuzzy rules: 81

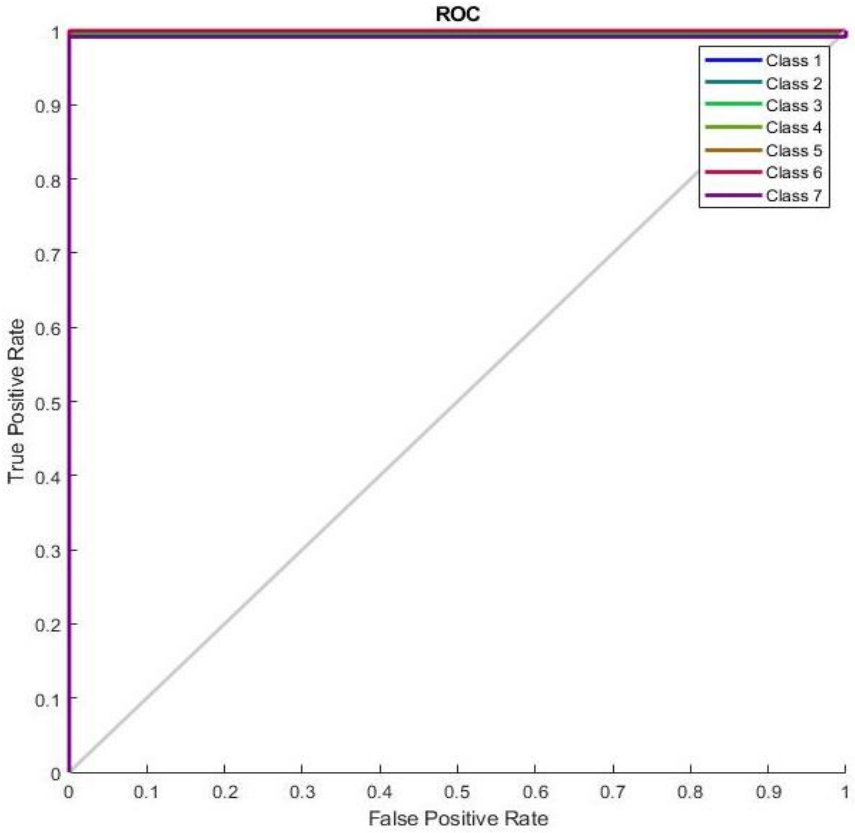

Figure 6-9 ROC curve for AQI – Jordan test data

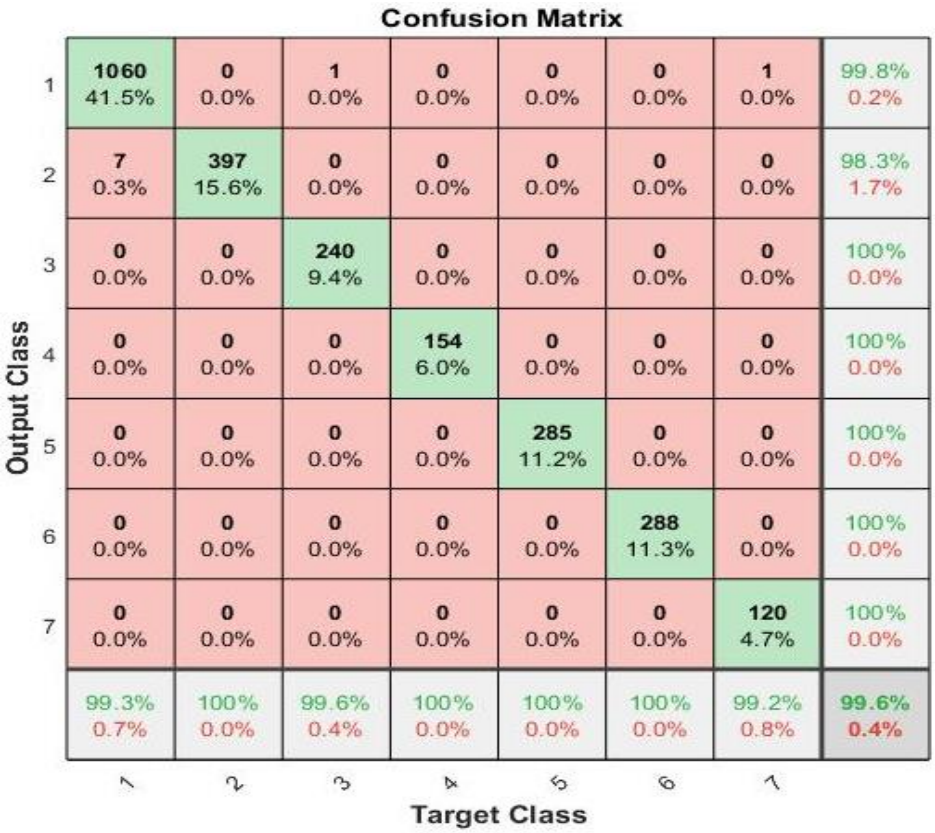

Figure 6-10 Confusion matrix for AQI – Jordan test data

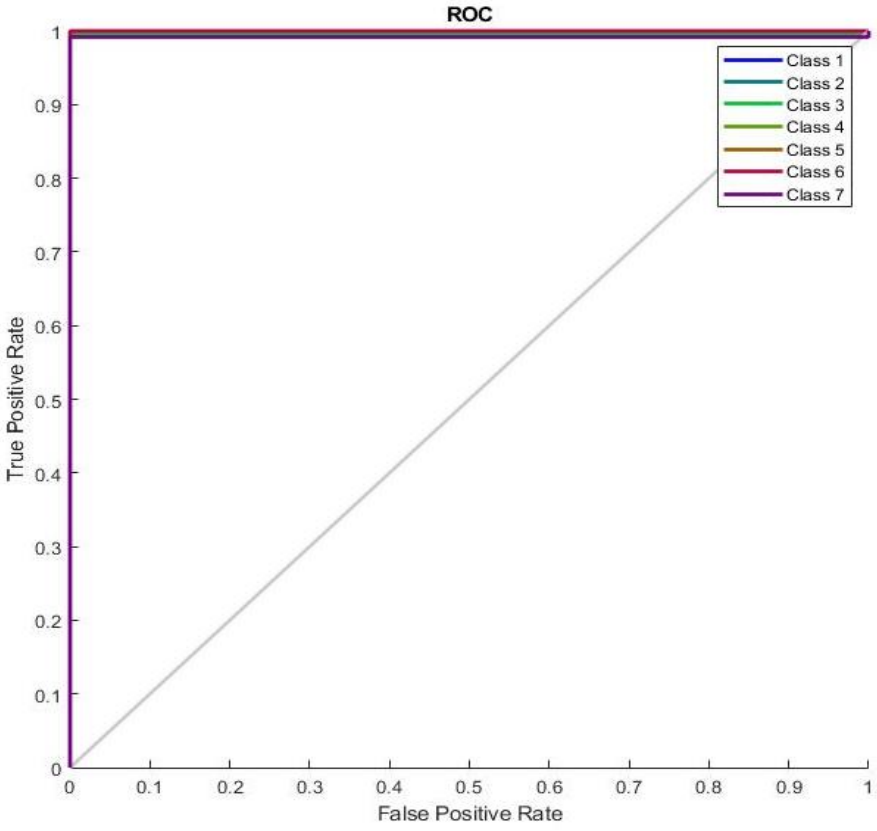

Figure 6-11 ROC curve for AQI – Jordan training data

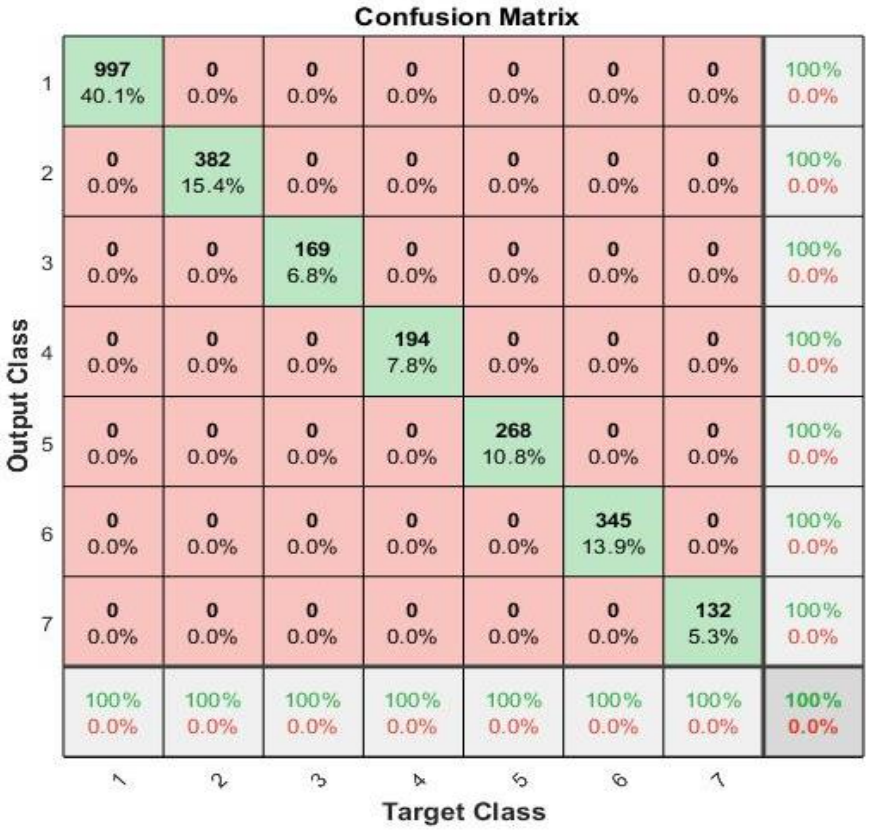

Figure 6-12 Confusion matrix for AQI – Jordan training data

6.5.3 ANFIS England Model

Figures 6-13 to 6-16 show the fuzzy logic results (ROC and confusion matrices) for the England model, subject to the following data:

- Number of nodes: 294
- Number of linear parameters: 1024
- Number of nonlinear parameters: 28
- Total number of parameters: 1052
- Number of training data pairs: 4382
- Number of checking data pairs: 0
- Number of fuzzy rules: 128

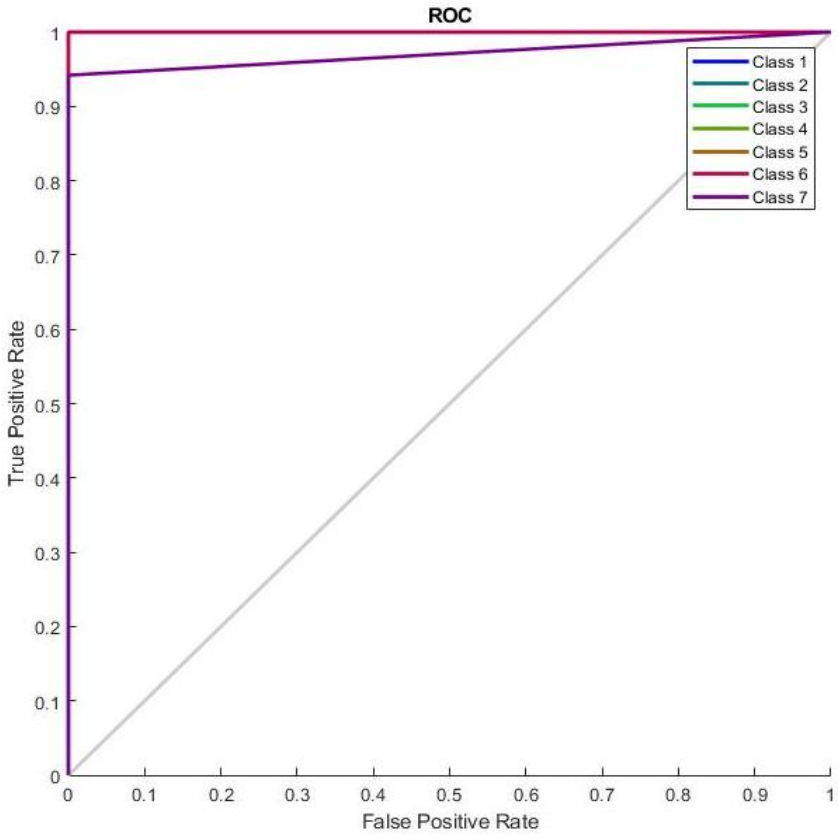

Figure 6-13 ROC curve for AQI – England training data

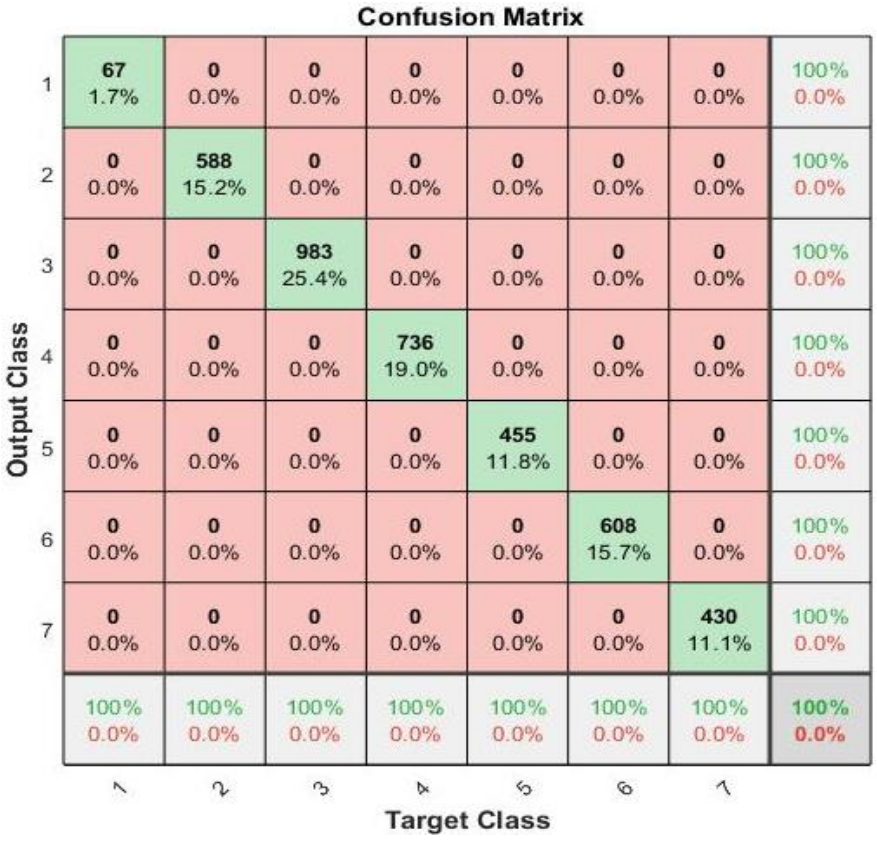

Figure 6-14 Confusion matrix for AQI – England training data

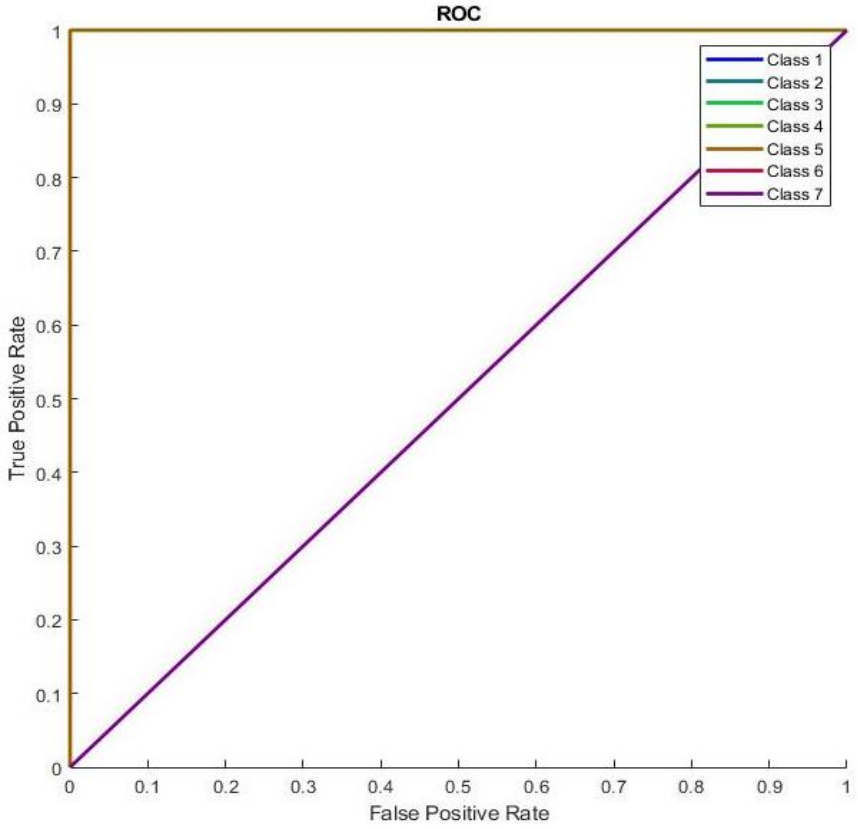

Figure 6-15 ROC curve for AQI – England test data

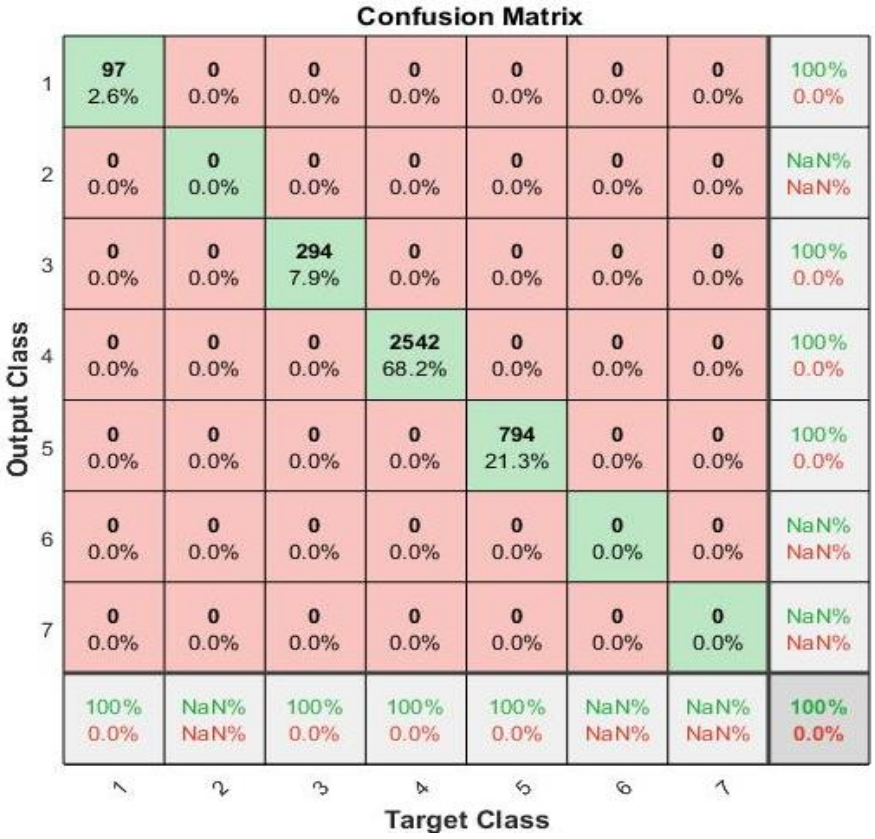

Figure 6-16 Confusion matrix for AQI – England test data

6.5.4 ANFIS Italy Model

Figures 6-17 and 6-18 show the fuzzy logic results (ROC and confusion matrices) for the Italy model as validation step.

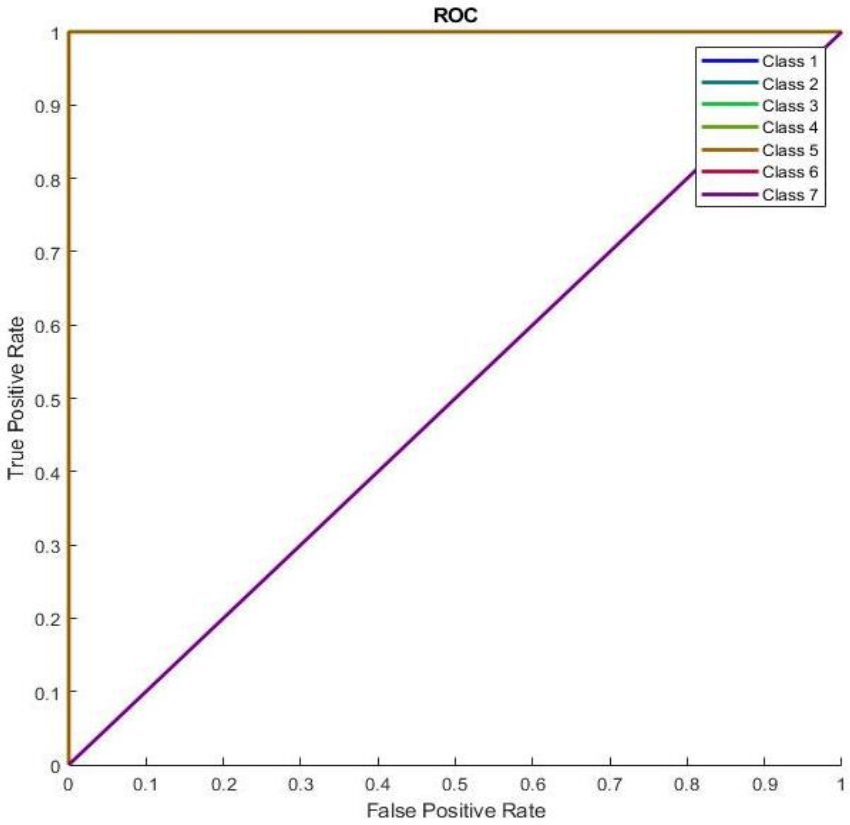

Figure 6-17 ROC curve for AQI – Italy test data

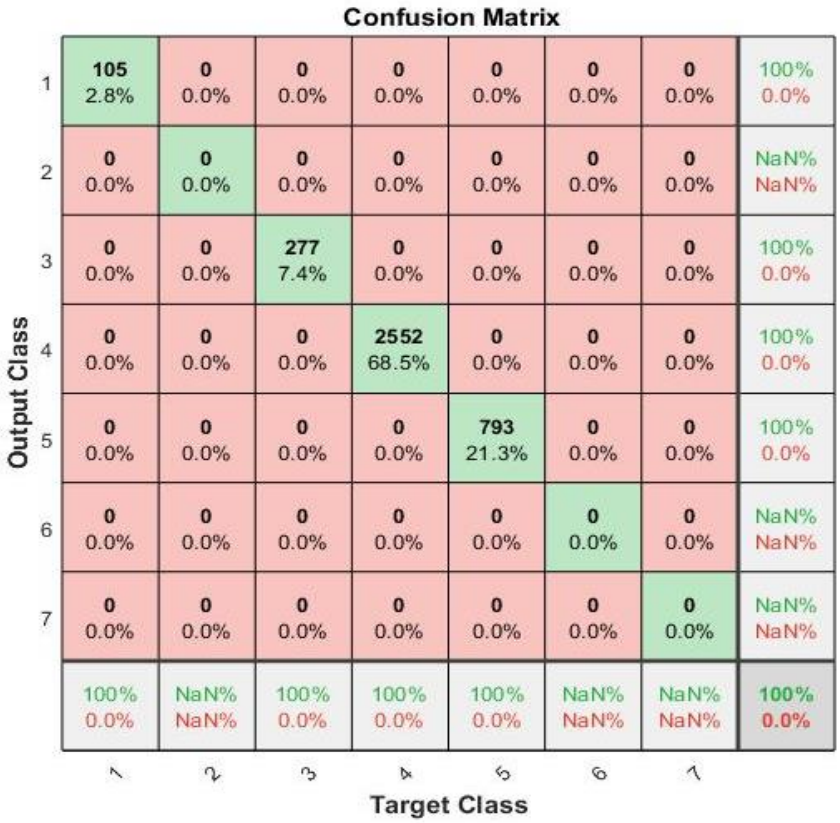

Figure 6-18 Confusion matrix for AQI – Italy test data

6.6 Summary

This chapter discussed AQI representation using fuzzy logic as a framework to be used to cover the blurry areas of AQI where indices are in between range of values. After studying several standards for AQP, this research suggested the use of fuzzy logic as an extended method to cover some limitations found in several standards, in which the fuzzy logic represent a more dynamic way to support cross countries comparisons as well. This research developed the EPA standards to address their acknowledged limitations by constructing a Fuzzy Air Quality Levels Prediction (FAQLP). The novel solution categorizes air quality to corresponding ranges (actual levels), and classifies new fuzzy levels (predicted levels), using a fuzzy logic model to enforce more realistic predictions. This model can solve the issue of values at or near boundaries, when there is uncertainty about air quality levels.

Chapter 7
Conclusions and Future Work

7.1 Main Outcomes

This concluding chapter revisits the identified research gaps in existing literature addressed by this research. It reiterates the covered topics, and highlights the contributions made to this research area. It acknowledges the limitations of this study, and identifies future research directions.

It should be highlighted from the outset that there is very limited research that focuses on transportation and its impacts on air quality, especially in Asia. There is limited research using hourly air quality data for prediction, and which specifically addresses transportation. Furthermore, processing air quality data is not easy, for several reasons, and there have been few studies that focused on the identification of effective parameters (check chapter 2).

The experimental investigations presented in this thesis used statistical methods for prediction, seeking to select the best statistical model and the best refinement method for air pollution and meteorological data modelling, in order to predict and categorize to air quality levels. This study is part of emerging research to design and model the AQI, seeking to discover the best ML methods to monitor the air quality domain. After completing the prediction models using DNN, hybrid models were found to achieve improved accuracy. The generated models are able to cover missing data problems and complex gases predictions and accuracy issues. In the future, this study recommends building a comprehensive AQI model that allows better comparisons across cities; however, this would entail a massive research effort, as many factors needs to be studied that affect air quality in different geographical locations. It should be mentioned that other prediction methods can be considered to support the identified directions for future research.

After completing both the Markov Chain and DNN modelling, the results were assessed and reported. As the aim of this research was to increase accuracy and obtain more reliable results, a hybrid model was proposed by the researchers, to ensure better interpretation of the data and more appropriate results. Many experimental trials were conducted in order to achieve the best scenario possible with the data available for this study. The DNN-Markov model was determined to be the best hybrid model, for data from both England and Jordan.

This research argues that the predictive framework for air quality indices presented in this study offers an effective method of measuring healthy levels of the air we breathe every day. At times of high toxic pollutant exposure, the researcher has introduced methods of predicting hourly emissions concentrations and producing related AQI levels as a control system for vulnerable areas.

Chapter 2 reviewed existing literature, mapping the gaps found in AQP domain and the methods that were developed by researchers. Modelling techniques and optimization were discovered from the current literature, along with studies and applications from researches in the domain to support the argument. The findings of existing literature were summarized, to set the scene for the subsequent chapters, seeking to contribute to air quality research and address the gaps identified from reviewed studies.

Chapter 3 explained the setup framework underlying the experimental work presented in subsequent chapters, from the data to experimental design stages. The chapter presented the data collection process, parameters, pre-processing, and modelling stages, and the experimental design framework that summarizes the flow from data to modelling.

Chapter 4 presented all viable proposed modelling solutions and scenarios for the identified scope of this research, with a detailed description of the algorithm used, discussing the architectures and parameters of the stand-alone models. The results for each selected model of the study and other experimented scenarios were also presented in this chapter. ANN, DNN (LSTM) and Markov chain models were included, along with some other studied scenarios, such as DNN (Bi-LSTM).

Chapter 5 concluded the experimental scenarios of the studied methodologies, architectures, and optimization of standalone models, and then hybridization was performed. The chapter presented the best reached scenario of the hybrid model and the selection along with results and performance of models. It concluded that the DNN-Markov model yielded the best performance, with the greatest improvement as indicated by data validation (using new data). The proposed solution to the identified problem presented a new and niche methodology to predict air quality, and covered relevant listed gaps in previous studies, which adds to the contributions of this research.

Chapter 6 presented the AQI framework using neuro-fuzzy logic. It built on the discussion expounded in Chapter 5 to discuss the conversions needed for the results, and connected the dots in proposing neuro-fuzzy logic to illustrate the refined outputs.

7.2 Challenges and Limitations

7.2.1 Data Challenges

The issue of data posed primary and fundamental challenges to this research, as this is the building block for all other elements. Challenges faced during this study pertained to data availability and quality. The models consist of multivariate data (i.e., with multi-inputs and multi-outputs), which made data pre-processing and training time more challenging. In particular, the DNN model took approximately three days to run for each trial for England Data, and two days for Jordan Data).

7.2.1.1 Incomplete or Sparse Data

Air pollutant concentrations are measured using monitors in several locations, and related devices, as with any other system, can have some downtimes during which data recording is suspended. Consequently, missing data renders datasets incomplete and inconsistent. There was a large portion of missing data in the collected datasets due to downtime issues, which created challenges in the efforts to find reasonable solutions for the problem (data could appear as zeros, NANs, or even empty records).

7.2.1.2 Data Quality Issues

In some cases data may appear to be complete, while measures are not accurate, which can be explored at the beginning by auditing and analysing data, and undertaking exploratory analysis pertinent to the domain. For instance, in this research, exploring the data revealed rows which showed humidity as 100 for Jordan Data, which could not be possible in the real scenario, so there was an immediate need to normalize, replace, or delete such data.

7.2.2 Model Challenges (Complexity)

Challenges were faced in selecting suitable algorithms for study, requiring a thorough literature review, studying many elements and factors. Besides the overhead of DNN and the computational resources required, hyper-tuning models is a significant part of the research for optimization, which takes quite a long time for trials to run and find the optimal combination of suitable parameters for the study. Integrating two models to create hybrid model consisted of several steps, scenarios, and configurations to attain suitable results, which are also effectively aligned with the aims and objectives of the study.

7.2.3 Temporal and Spatial Variability

There can be differences in pollutant concentrations across time and location; for instance, based on seasonal fluctuations or traffic areas patterns, or due to differences between cities or regions. Such variability creates difficulties in analysing data patterns to ensure data quality and imputation strategies needed to deal with missing data.

7.3 Future Improvements and Research Directions

This research presented several contributions to the field of AQP through studying the literature, identifying gaps and proposing solutions to existential challenges. This work discussed several methods to cover the gaps in accuracy and reliability when predicting with Big Data and further identified a method for AQI interpretation from the predicted data. It discussed the representation of pollution levels, noting the gap identified from the reviewed literature in terms of the need for a global framework for a unified method to measure air quality indicators in the presence of varied standards in countries and regions with no availability of any supportive global standards for pollution level comparisons.

Therefore, the researcher suggests building a global AQI framework as a future improvement for the work presented in this thesis. As the presented literature reviewed in Chapter 2 indicates the lack of availability of a global universal AQI system that could compare air pollution levels between countries, there is clearly an existential standardization issue. Future research is needed to drive progress towards a more global unified framework for AQI prediction, which would create rich potential for project funding as an extension to this research. This can be adopted in universities or research institutions with real implementation projects to help produce a unified global AQI prediction system.

7.4 Developments in AQP field

Looking further beyond the outcomes of the current study, recent literature indicates the emergence and evolution of new methods of air quality assessment that can be used in experiments to test the suggested unified framework for AQI. Deep Transformer Networks (DTNs) are a possible future use in the area of air quality forecasting, besides the increasing popularity of Graph Neural Networks (GNNs) to model dynamic interactions, whereby air quality factors can be studied, and relational factors can be mapped. Furthermore, Temporal Convolutional Networks (TCNs) could be considered, specifically for complex gases such $PM_{2.5}$, to cover the complexity in modelling such gases. Moreover, Complex Event Processing (CEP) has recently been used in some applications, as evident from the

literature. It worth mentioning there has been clear identification in the literature for the need to find the relation and linkage between air quality and climate change, and developing models for early warning climate change systems will be needed to support sustainable cities and societies.

This is a rich research area where there could be many areas that could be identified as holding potential for future directions extending beyond this work.

References

[1] AirNow (2018) *Air Quality Index (AQI) – A guide to air quality and your health*. Available at: https://web.archive.org/web/20180618144741/https:/airnow.gov/index.cfm?action=aqibasics.aqi (Accessed: 10 February 2024).

[2] Alkasassbeh, M., Sheta, I. F., Faris, H. and 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), pp. 999-1009. Available at: http://dx.doi.org/10.5829/idosi.mejsr.2013.14.7.2171

[3] Alnawaiseh, N. A. and Hashim, J. H. (2014) 'Respiratory symptoms from particulate air pollution related to vehicle traffic in Amman, Jordan', *European Journal of Scientific Research*, 120(4), pp. 550-563.

[4] Alzubi, J., Nayyar, A. and Kumar, A. (2018) 'Machine learning from theory to algorithms: An overview', *Journal of Physics: Conference Series*, 1142(1), pp. 0-15. Available at: https://doi.org/10.1088/1742-6596/1142/1/012012

[5] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Ul Islam, S., Asghar, M. N. (2019) 'Comparative analysis of machine learning techniques for predicting air quality in smart cities', *IEEE Access*, 7, pp. 128325-128338. Available at: https://doi.org/10.1109/ACCESS.2019.2925082

[6] Baatarchuluun, K., Sung, Y.-S. and Lee, M. (2020) 'Air pollution prediction model using artificial neural network and fuzzy theory', *International Journal of Internet, Broadcasting and Communication*, 12(3), pp. 149-155.

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

[8] Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D. and Akanbi, A. L. (2022) 'Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting', *Machine Learning with Applications*, 7, p. 100204. Available at: https://doi.org/10.1016/j.mlwa.2021.100204

[9] Bishoi, B., Prakash, A. and Jain, V. K. (2009) 'A comparative study of air quality index based on factor analysis and EPA methods for an urban environment', *Aerosol and Air Quality Research*, 9(1), pp. 1-17. Available at: https://doi.org/10.4209/aaqr.2008.02.0007

[10] Chapman, L. (2007) ‘Transport and climate change: a review’, *Journal of Transport Geography*, 15(5), pp. 354-367. Available at: https://doi.org/10.1016/j.jtrangeo.2006.11.008

[11] Chauvet, M. and Hamilton, J. (2005) ‘Dating business cycle turning points’, National Bureau of Economic Research, Working Paper No. 11422. Available at: https://doi.org/10.3386/w11422

[12] Chen, J. C. and Wu, Y. J. (2020) ‘Discrete-time Markov chain for prediction of air quality index’, *Journal of Ambient Intelligence and Humanized Computing*, p. 0123456789. Available at: https://doi.org/10.1007/s12652-020-02036-5

[13] Cheng, W.-L., Chen, Y.-S., Zhang, J., Lyons, T. J., Pai, J. L. and Chang, S.-H. (2007) ‘Comparison of the Revised Air Quality Index with the PSI and AQI indices’, *Science of the Total Environment*, 382(2-3), pp. 191-198. Available at: https://doi.org/10.1016/j.scitotenv.2007.04.036.

[14] Coelho, S., Rafael, S., Lopes, D., Miranda, A. I. and Ferreira, J. (2021) ‘How changing climate may influence air pollution control strategies for 2030?’, *Science of the Total Environment*, 758, p. 143911. Available at: https://doi.org/10.1016/j.scitotenv.2020.143911

[15] Delavar, M. R., Gholami, A., Shiran, G. R., Rashidi, Y., Nakhaeizadeh, G. R., Fedra, K. and Hatefi Afshar, S. (2019) ‘A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran’, *International Journal of Geo-Information*, 8(2), p. 99. Available at: https://doi.org/10.3390/ijgi8020099

[16] Environmental Protection Agency (2023) *How BenMAP-CE estimates health and economic effects of air pollution*. Available at: https://www.epa.gov/benmap/how-benmap-ce-estimates-health-and-economic-effects-air-pollution (Accessed: 10 February 2024).

[17] Eren, B., Aksangür, İ. and Erden, C. (2023) ‘Predicting next hour fine particulate matter (PM2.5) in the Istanbul Metropolitan City using deep learning algorithms with time windowing strategy’, *Urban Climate*, 48, p. 101418. Available at: https://doi.org/10.1016/j.uclim.2023.101418

[18] Faris, H., Alkasassbeh, M., Ghatasheh, N. and Harfoushi, O. (2014) 'PM_{10} prediction using genetic programming: A case study in Salt, Jordan', *Life Science Journal*, 11(2), pp. 86-92.

[19] FCCMG (2016) *4 ways poor air quality can negatively affect your health*. Available at: https://www.fccmg.com/blog/4-ways-poor-air-quality-can-negatively-affect-your-health (Accessed: 10 February 2024).

[20] Fenger, J. (2009) ‘Air pollution in the last 50 years - From local to global’, *Atmospheric Environment*, 43(1), pp. 13-22. Available at: https://doi.org/10.1016/j.atmosenv.2008.09.061

[21] Fiore, A. M., Naik, V. and Leibensperger, E. M. (2015) ‘Air quality and climate connections’, *Journal of the Air and Waste Management Association*, 65(6), pp. 645-685. Available at: https://doi.org/10.1080/10962247.2015.1040526

[22] Font, A., Guiseppin, L., Blangiardo, M., Ghersi, V. and Fuller, G. W. (2019) ‘A tale of two cities: Is air pollution improving in Paris and London?’, *Environmental Pollution*, 249, pp. 1-12. Available at: https://doi.org/10.1016/j.envpol.2019.01.040

[23] García Nieto, P. J., Combarro, E. F., del Coz Díaz, j. J. and Montañés, E. (2013) '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*, 219(17), pp. 8923-8937. Available at: https://doi.org/10.1016/j.amc.2013.03.018

[24] Graves, A. and Schmidhuber, J. (2005) ‘Framewise phoneme classification with bidirectional LSTM and other neural network architectures’, *Neural Networks*, 18(5-6), pp. 602-610. Available at: https://doi.org/10.1016/j.neunet.2005.06.042

[25] 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*, 40(7), pp. 1216-1229. Available at: https://doi.org/10.1016/j.atmosenv.2005.10.036

[26] Hall, J. V., Brajer, V. and Lurmann, F. W. (2010) ‘Air pollution, health and economic benefits: Lessons from 20 years of analysis’, *Ecological Economics*, 69(12), pp. 2590-2597. Available at: https://doi.org/10.1016/j.ecolecon.2010.08.003

[27] Hamilton, J. D. (1989) ‘A new approach to the economic analysis of nonstationary time series and the business cycle’, *Econometrica*, 57(2), pp. 357-384. Available at: https://doi.org/10.2307/1912559

[28] Hamilton, J. D. (1990) ‘Analysis of time series subject to changes in regime’, *Journal of Econometrics*, 45(1-2), pp. 39-70. Available at: https://doi.org/10.1016/0304-4076(90)90093-9

[29] Hamilton, J. D. (1994) *Time series analysis*. Princeton, NJ: Princeton University Press.

[30] Hu, J., Chen, Y., Wang, W., Zhang, S., Cui, C., Ding, W. and Fang, Y. (2023) ‘An optimized hybrid deep learning model for $PM_{2.5}$ and O_3 concentration prediction’, *Air Quality, Atmosphere and Health*, 16(4), pp. 857-871. Available at: https://doi.org/10.1007/s11869-023-01317-0

[31] Kang, G. K., Gao, J. Z., Chiao, S., Lu, S. and Xie, G. (2018) ‘Air quality prediction: Big Data and machine learning approaches’, *International Journal of Environmental*

Science and Development, 9(1), pp. 8-16. Available at: https://doi.org/10.18178/ijesd.2018.9.1.1066

[32] Kemeny, J. G. and Snell, J. L. (1983) *Finite Markov chains: With a new appendix*. New York: Springer.

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

[34] Krakovna, V. and Doshi-Velez, F. (2016) 'Increasing the interpretability of recurrent neural networks using hidden Markov models', *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Available at: https://doi.org/10.48550/arXiv.1611.05934

[35] Krolzig, H.-M. (1997) 'The Markov-Switching Vector Autoregressive Model', in H.-M. Krolzig (ed) *Markov-Switching Vector Autoregressions: Modelling, statistical inference, and application to business cycle analysis*. Berlin, Heidelberg: Springer, pp. 6-28. Available at: https://doi.org/10.1007/978-3-642-51684-9_2

[36] Leal, P. H. and Marques, A. C. (2022) 'The evolution of the environmental Kuznets curve hypothesis assessment: A literature review under a critical analysis perspective', *Heliyon*, 8(11), p. e11521. Available at: https://doi.org/10.1016/j.heliyon.2022.e11521

[37] Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S. and Duan, C. (2021) 'Statistical approaches for forecasting primary air pollutants: A review', *Atmosphere*, 12(6), p. 686. Available at: https://doi.org/10.3390/atmos12060686

[38] Liu, B.-C., Binaykia, A., Chang, P.-C., Tiwari, M. K. and Tsao, C.-C. (2017) 'Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang', *PLoS ONE*, 12(7), p. e0179763. Available at: https://doi.org/10.1371/journal.pone.0179763

[39] Liu, X., Guo, C., Wu, Y., Huang, C., Lu, K., Zhang, Y., Duan, L., Cheng, M., Chai, F., Mei, F. and Dai, H. (2023) 'Evaluating cost and benefit of air pollution control policies in China: A systematic review', *Journal of Environmental Sciences*, 123, pp. 140-155. Available at: https://doi.org/10.1016/j.jes.2022.02.043

[40] Lokys, H. L., Junk, J. and Krein, A. (2015) 'Making air quality indices comparable - Assessment of 10 years of air pollutant levels in western Europe', *International Journal of Environmental Health Research*, 25(1), pp. 52-66. Available at: https://doi.org/10.1080/09603123.2014.893568

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

[42] MacDonald, I. L. and Zucchini, W. (2016) *Hidden Markov and other models for discrete- valued time series*. Boca Raton, FL: CRC Press.

[43] Mandal, T. K. and Gorai, A. K. (2014) 'Air quality indices: A literature review', *Journal of Environmental Science & Engineering*, 56(3), pp. 357-362.

[44] Martínez-España, R., Bueno-Crespo, A., Timon-Perez, I. M., Soto, J., Muñoz, A. and Cecilia, J. M. (2018) 'Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain', *Journal of Universal Computer Science*, 24, pp. 261-276.

[45] Masih, A. (2019) 'Machine learning algorithms in air quality modeling', *Global Journal of Environmental Science and Management*, 5(4), pp. 515-534. Available at: https://doi.org/10.22034/gjesm.2019.04.10

[46] Méndez, M., Merayo, M. G. and Núñez, M. (2023) 'Machine learning algorithms to forecast air quality: A survey', *Artificial Intelligence Review*, 56, pp. 10031-10066. Available at: https://doi.org/10.1007/s10462-023-10424-4

[47] Monteiro, A., Vieira, M., Gama, C. and Miranda, A. I. (2017) 'Towards an improved air quality index', *Air Quality, Atmosphere & Health*, 10(4), pp. 447-455. Available at: https://doi.org/10.1007/s11869-016-0435-y

[48] Moscoso-López, J. A., González-Enrique, J., Urda, D., Ruiz-Aguilar, J. J. and Turias, I. J. (2023) 'Hourly pollutants forecasting using a deep learning approach to obtain the AQI', *Logic Journal of the IGPL*, 31(4), pp. 722-738. Available at: https://doi.org/10.1093/jigpal/jzac035

[49] Navares, R. and Aznarte, J. L. (2020) 'Predicting air quality with deep learning LSTM: Towards comprehensive models', *Ecological Informatics*, 55, p. 101019. Available at: https://doi.org/10.1016/j.ecoinf.2019.101019

[50] Niharika and Rao, P. R. (2014) A survey on air quality forecasting techniques. *International Journal of Computer Science and Information Technologies*, 5(1), pp. 103-107.

[51] Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J. and Kolehmainen, M. (2004) 'Evolving the neural network model for forecasting air pollution time series', *Engineering Applications of Artificial Intelligence*, 17(2), pp. 159-167. Available at: https://doi.org/10.1016/j.engappai.2004.02.002

[52] Patra, S. R. (2017) 'Time series forecasting of air pollutant concentration levels using machine learning', *Advances in Computer Science and Information Technology*, 4(5), pp. 280-284.

[53] Pinson, P. and Madsen, H. (2012) 'Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models', *Journal of Forecasting*, 31(4), pp. 281-313. Available at: https://doi.org/10.1002/for.1194

[54] Plaia, A. and Ruggieri, M. (2011) ‘Air quality indices: A review’, *Reviews in Environmental Science and Biotechnology*, 10(2), pp. 165-179. Available at: https://doi.org/10.1007/s11157-010-9227-2

[55] Polezer, G., Potgieter-Vermaak, S., Oliveira, A., Martins, L. D., Santos-Silva, J. C., Moreira, C. A. B., Pauliquevis, T., Godoi, A. F. L., Tadano, Y., Yamamoto, C. I. and Godoi, R. H. M. (2023) ‘The new WHO air quality guidelines for PM2.5: predicament for small/medium cities’, *Environmental Geochemistry and Health*, 45(5), pp. 1841-1860. Available at: https://doi.org/10.1007/s10653-022-01307-8.

[56] Rabiner, L. R. (1989) ‘A tutorial on hidden Markov models and selected applications in speech recognition’, *Proceedings of the IEEE*, 77(2), pp. 257-286. Available at: https://doi.org/10.1109/5.18626

[57] Rakholia, R., Le, Q., Ho, B. Q., Vu K. and Carbajo, R. S. (2023) ‘Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam’, *Environment International*, 173, p. 107848. Available at: https://doi.org/10.1016/j.envint.2023.107848

[58] Ripple, W. J., Wolf, C., Newsome, T. M., Barnard, P. and Moomaw, W. R. (2020) ‘World scientists’ warning of a climate emergency’, *BioScience*, 70(1), pp. 80-12. Available at: https://doi.org/10.1093/biosci/biz088

[59] Robinson, G. M. (2009) ‘Time series analysis’, in *International Encyclopedia of Human Geography*. Available at: https://doi.org/10.1016/B978-008044910-4.00546-0

[60] Rybarczyk, Y. and Zalakeviciute, R. (2018) ‘Machine learning approaches for outdoor air quality modelling: A systematic review’, *Applied Sciences*, 8(12), p. 2570. Available at: https://doi.org/10.3390/app8122570

[61] Scabini, L. F. S. and Bruno, O. M. (2023) ‘Structure and performance of fully connected neural networks: Emerging complex network properties’, *Physica A: Statistical Mechanics and its Applications*, 615, p. 128585. Available at: https://doi.org/10.1016/j.physa.2023.128585

[62] Shao, F. and Shen, Z. (2023) ‘How can artificial neural networks approximate the brain?’, *Frontiers in Psychology*, 13, p. 970214. Available at: https://doi.org/10.3389/fpsyg.2022.970214

[63] Shrestha, A. and Mahmood, A. (2019) ‘Review of deep learning algorithms and architectures’, *IEEE Access*, 7, pp. 53040-53065. Available at: https://doi.org/10.1109/ACCESS.2019.2912200

[64] Siami-Namini, S., Tavakoli, N. and Namin, A. S. (2019) ‘The performance of LSTM and BiLSTM in forecasting time series’, *Proceedings 2019 IEEE International Conference on Big Data*, pp. 3285-3292. Available at: https://doi.org/10.1109/BigData47090.2019.9005997

[65] Singpurwalla, N. D. and Booker, J. M. (2004) ‘Membership functions and probability measures of fuzzy sets’, *Journal of the American Statistical Association*, 99(467), pp. 867-877. Available at: https://doi.org/10.1198/016214504000001196

[66] Sowlat, M. H., Gharibi, H., Yunesian, M., Mahmoudi, M. T. and Lotfi, S. (2011) ‘A novel, fuzzy-based air quality index (FAQI) for air quality assessment’, *Atmospheric Environment*, 45(12), pp. 2050-2059. Available at: https://doi.org/10.1016/j.atmosenv.2011.01.060

[67] Staudemeyer, R. C. and Morris, E. R. (2019) ‘Understanding LSTM: A tutorial into Long Short-Term Memory Recurrent Neural Networks’, *arXiv*, p. 1909.09586. Available at: https://doi.org/10.48550/arXiv.1909.09586

[68] Su, F., Li, M., Shanthini, A. and Parthasarathy, R. (2021) ‘Systematic effective modeling and geospatial information framework for environmental issues in urban areas’, *Environmental Impact Assessment Review*, 86, p. 106507. Available at: https://doi.org/10.1016/j.eiar.2020.106507

[69] Tdunning (2012) *Hidden Markov model*. Available at: https://en.wikipedia.org/wiki/Hidden_Markov_model#/media/File:HiddenMarkovModel.svg (Accessed: 10 February 2024).

[70] Tealab, A. (2018) ‘Time series forecasting using artificial neural networks methodologies: A systematic review’, *Future Computing and Informatics Journal*, 3(2), pp. 334-340. Available at: https://doi.org/10.1016/j.fcij.2018.10.003

[71] Tripathi, K. and Pathak, P. (2021) ‘Deep learning techniques for air pollution’, *Proceedings IEEE 2021 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2021*, II, pp. 1013-1020. Available at: https://doi.org/10.1109/ICCCIS51004.2021.9397130

[72] van den Elshout, S., Léger, K. and Nussio, F. (2008) ‘Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative’, *Environment International*, 34(5), pp. 720-726. Available at: https://doi.org/10.1016/j.envint.2007.12.011

[73] Vito, S. (2016) *Air quality*. UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C59K5F

[74] Wang, P., Wang, H., Zhang, H., Lu F. and Wu, S. (2019) ‘A hybrid markov and LSTM model for indoor location prediction’, *IEEE Access*, 7, pp. 185928-185940. Available at: https://doi.org/10.1109/ACCESS.2019.2961559

[75] Wang, Z., Tan, Y., Guo, M., Cheng, M., Gu, Y., Chen, S., Wu, X. and Chai, F. (2023) ‘Prospect of China’s ambient air quality standards’, *Journal of Environmental Sciences*. 123, pp. 255-269. Available at: https://doi.org/10.1016/j.jes.2022.03.036.

[76] Weston, P. (2019) ‘Global climate emergency: 11,000 scientists from across the world unite to issue unprecedented declaration’, *The Independent*, 5 November. Available at: https://www.independent.co.uk/environment/climate-emergency-scientists-emissions-letter-climate-change-a9185786.html?fbclid=IwAR1WNs5HLQaGxIac50scgXXayAhJTE42nofsKwrTIyT0IIyJZulCokXlzF0 (Accessed: 10 February 2024).

[77] Zaini, N., Ean, L. W., Ahmed, A. N. and Malek, M. A. (2022) ‘A systematic literature review of deep learning neural network for time series air quality forecasting’, *Environmental Science and Pollution Research*, 29(4), pp. 4958-4990. Available at: https://doi.org/10.1007/s11356-021-17442-1

[78] Zakaria, N. N., Othman, M., Sokkalingam, R., Daud, H., Abdullah, L. and Abdul Kadir, E. (2019) ‘Markov chain model development for forecasting air pollution index of Miri, Sarawak’, *Sustainability*, 11(19), p. 5190. Available at: https://doi.org/10.3390/su11195190

[79] Zhang, J. and 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), p. 114. Available at: https://doi.org/10.3390/ijerph14020114

[80] Zhao, X., Zhang, R., Wu, J.-L. and Chang, P.-C. (2018) ‘A deep recurrent neural network for air quality classification’, *Journal of Information Hiding and Multimedia Signal Processing,* 9(2), pp. 346-354

[81] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E. and Li, T. (2015) ‘Forecasting fine-grained air quality based on Big Data’, *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2267-2276. Available at: https://doi.org/10.1145/2783258.2788573

[82] Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I.-F. and 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, pp. 134-145. Available at: https://doi.org/10.1016/j.jclepro.2018.10.243