applied sciences

MDPI

Article

Breathable Cities: Dynamic Machine Learning Modelling Approaches for Advanced Air Pollution Control

Roba Zayed and Maysam Abbod *

Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, UK; roba.zayed@brunel.ac.uk
* Correspondence: maysam.abbod@brunel.ac.uk

Abstract: This paper discusses air quality index (AQI) representation using a fuzzy logic framework to cover the blurry areas of AQI where indices are in between ranges of values. After studying several standards for air quality prediction (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 represents a more dynamic way to support cross-country comparisons as well. This research expanded upon the United States Environmental Protection Agency (USEPA) standards to address their acknowledged limitations by constructing a fuzzy air quality levels prediction (FAQLP) model, which categorizes air quality into 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. The study aims to incorporate a comparative study of two urban settings providing dynamic machine-learning modeling approaches for advanced air pollution control. The DNN–Markov model is presented in this paper as the selected hybrid model for AQI prediction, and the adaptive neuro-fuzzy inference system (ANFIS) was used to represent AQI. This work presents a novel air quality index framework that consists of a DNN–Markov model for accurate hourly predictions and air quality level representations using ANFIS.

Keywords: air quality index; fuzzy logic; machine learning modeling; prediction; adaptive algorithms; air pollution forecasting; air quality monitoring; artificial intelligence; dynamic modeling; hybrid models

check for updates

Citation: Zayed, R.; Abbod, M. Breathable Cities: Dynamic Machine Learning Modelling Approaches for Advanced Air Pollution Control. *Appl. Sci.* **2024**, *14*, 5581. https://doi.org/10.3390/app14135581

Academic Editor: Yves Rybarczyk

Received: 20 May 2024
Revised: 19 June 2024
Accepted: 24 June 2024
Published: 27 June 2024

1. Introduction

Monitoring air quality is increasingly a necessity given continuously increasing pollution levels, with diverse and serious consequences on human health, social interactions, atmospheric impacts, and ultimately socio-economic development. Artificial intelligence (AI), machine learning (ML), and deep learning uses are advancing over time, which has been leveraged by this study aiming to offer enhanced predictive accuracy and a reliable tool for policymakers. This paper presents advanced ML approaches for air quality monitoring. This research is relevant from different perspectives, given the increasing impact of air quality on public health and policymaking. The methods used have demonstrated approaches for reliable accuracy and operational feasibility in urban settings, leveraging multivariate data to provide hourly forecasts for air quality in two diverse urban environments. 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.

The air quality index (AQI) has come to be used over the years to express the level of pollutant concentration over a period in a way that is understandable by the public and decision-makers [1]. Researchers have alluded to the lack of a systematic process for hybrid ML models and the possibility of generating an unlimited number of scenarios by combining different models to maximize potential; this could scale up AQP research while potentially precluding practical application [2]. Cities worldwide are affected by various consequences of air pollution, and in the worst cases, there may be scenarios where people are unable to undertake normal activities of daily living. 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 particulate matter (PM).

Observations arising from the analysis of the literature (as presented in the following section) reveal that several approaches can be followed for AQP modeling, and in most cases, there is no single method that fits all air quality domain prediction modeling 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 them. Different model topologies should be considered, and more specifically, integration methods (i.e., hybrid modeling) are of increasing research interest. Hybrid modeling can create extremely large numbers of possibilities between different model types, architectures, and topologies, as well as the number and ways in which different models can be integrated to offer more effective solutions.

Hence, the complexity of selecting one method over another entails numerous academic and practical trade-offs in managing general and particular difficulties. It is recommended to carefully study the selection of different variables of studies and evaluate the 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 [3]. Hybrid ML models have gained increasing attention in recent years due to superior performance and more accuracy. It was thought that there would generally be no intrinsic restrictions on how models could be combined; researchers have been innovative in the way they combine models, driven by the aims and objectives of the studies. Models, as such, generally depend on the architecture, parameters, or relevant variables, which could increase complexity; thus, specifying parameters is of extreme importance in building such hybrid models [4].

The study conducted several methods in an effort to contribute a new and efficient approach to predicting air quality for the next hour. Deep neural network (DNN) with Markov (i.e., 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 modeling 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 was generated using fuzzy logic 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 using fuzzy logic.

2. Related Work

2.1. AI and Neural Networks

The concept of AI was used historically before the term ‘Artificial Intelligence’ itself was coined. The essence of AI is the ability of a system, machine, or any computer-instructed program to undertake a certain functionality (or different functionalities) without human mediation. Despite the conceptualization of official AI capabilities during the 1950s, AI was not widely utilized until developments in ML during the 2000s, when algorithmic methods empowered computers to learn from data and provide different types of outputs based on applications. Over the last two decades, computers have been able to undertake more practical functionalities in doing tasks or building decision or prediction models without instructional human interference. The resultant outputs are a result of the computer (“machine”) learning from data. The discovery has evolved since then to take an in-depth dimension in the evolution of DNNs as an advanced solution to complex problems built up on Neural Networks theory, but with multi-layer architectures to support AI applications in classification and forecasting. To sum up, NN, DNN, and other algorithms can be identified as tools used for different levels of problems using mathematical models that vary based on the underlying complexity of the application. The main implications of notable recent works developing AQI systems and systematic reviews in this field are summarized in the following sections.

2.2. Reviewed AQI Systems

‘Comparison of the Revised Air Quality Index with the PSI and AQI indices’ [5] notes that the pollution standards index (PSI) was established due to 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. However, the cost of establishing a monitoring system covering PM2.5 is prohibitively expensive for many countries to implement, notwithstanding serious global O_3 problems.

‘Novel, fuzzy-based air quality index (FAQI) for air quality assessment’ [6] proposed FAQI, using fuzzy logic for different pollutants based on different weighting factors. The FAQI was suggested as a more sensitive tool to address limitations in legacy AQI. The results of FAQI were compared to those of AQI USEPA (of the ‘United States Environmental Protection Agency’), and the authors flagged FAQI as a comprehensive, reliable method for decision-makers.

‘Towards an improved air quality index’ [7] claimed that there are no universally significant methods that cover all specific situations of air quality and pointed out that methods of AQI could differentiate based on the number of pollutants, air quality levels categories, and boundaries points, and sampling period. The current AQI paradigm 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’), 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’ [8] suggested factor analysis of the national air quality index (NAQI) to be used to cover the gaps in the EPA’s system. The authors claimed that NAQI could be used to compare 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’ [9] 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 diverse parameters.

2.3. Systematic Reviews

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.

'Time series forecasting using artificial neural networks methodologies: A systematic review' [2] studied new ANN models (developed during the period 2006–2016) and presented evidence that the predictions of hybrid models were more accurate than those of traditional ANN models (such as back-propagation with single hidden layers), despite the lack of a systematic process for hybrid model development. The study explored some new models in terms of architecture, complexity, relevant variable selection, parameters estimation, and implementation and evaluation, and recommended more research to specify criteria for relevant variable selections (i.e., the basis for selection), methodological development for the selection of ANN architectures, the creation of evaluation models, and a methodology that tests the generalization of models.

'Machine learning approaches for outdoor air quality modeling: A systematic review' [10] analyzed 46 papers to determine why some algorithms are selected over others in prediction. It addressed the main need for ML-based statistical models to overcome the limitations of deterministic techniques, to model non-linear relationships between concentrations with required accuracy, and provided details about algorithms and how they are applied to enhance accuracy (principles of algorithms). Its main findings showed that estimation problems usually apply ensemble learning and regressions, forecasting problems mostly arise due to NNs and SVMs, and challenges exist for improving peak prediction and contaminants (such as nanoparticles). The authors claimed that ML research is mainly undertaken in Europe and North America and chiefly focuses on the estimation of pollutants' concentration (using ensemble learning and regression analysis) and forecasting problems (using NN and SVM), which 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 the drawbacks of being very computationally demanding. The study emphasized the suitability of ML to predict air quality, and traditional deterministic methods showed complexity in modeling fine PM, while ML approaches (estimation and forecasting) showed high accuracy relative to other emission gases; however, lower precision is noted for peak values. The study reported that accuracy is higher for medium and small peaks than high concentrations of pollutants (high peaks), while forecasting for some gases such as CO and NOx is limited in terms of performance. The assessed models showed better performance in peak weather conditions. The study suggested future directions, developing models that enhance pollution peak prediction and models that improve critical pollutants such as CO and NOx.

'Machine learning algorithms to forecast air quality: A survey' [11] reviewed 155 publications and demonstrated a direct correlation between the most polluted and the most studied countries, noting an increasing trend in the number of ML models used for pollution studies. For the studied pollutant measures, nearly half of the studied papers used AQI, and for air pollutant concentration, 54 papers showed that PM2.5 is the most predicted. The most used pollutant features were weather variables.

In terms of ML techniques, DL methods were more widely used than regression algorithms, and hybrid algorithms include both types. Specifically, the most used algorithms were LSTM and MLP, while CNN, RNN, GRU, and auto-encoders were used less commonly. The most used regression algorithms were SVR and RF, while the less frequently used ones comprised DT, ARIMA, KNN, and Boosting. The review noted the increasing trend of using deep transformer networks. It also observed that 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). Recently, graph NNs have become more popular for air quality forecasting, which could model dynamic interactions (e.g., different cities, neighborhoods, and streets) with distance-based weights. There are recent applications for using temporal convolutional networks (TCNs) specifically for PM2.5 and recent mention of the use of recent applications of complex event processing (CEP) for air quality forecasting.

'Statistical approaches for forecasting primary air pollutants: A review' [4] quantitatively analyzed research published between 1990 and 2018, identifying trends. It found that most papers mainly focused on air pollution and its relation to health diseases, urban pollution exposure models, and land use regression methods. PM, NOx, and O_3 were the most studied pollutants, and there was a marked preference for using ANN when studying PM and O_3, while LUR was mostly used in NOx studies. Hybrid methods (a combination of models) became the most used method between 2010 and 2018. The authors expected future 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, and there is an increasing trend for studying PM and the influence it has on air pollution. Reviewed research papers showed that PM is the most studied emission, followed by NOx and O_3, while 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, pointed out the increase in accuracy for AQP studies over years of efforts in the domain, and discussed that there are still gaps in the domains and work to be conducted in this regard. It highlighted the necessity to study the interaction or relation between air pollutants, human health, and the urban environment, including the interactions between pollutants, in particular PM-NOx and PM-O_3 (as the main combination of interest).

'A systematic literature review of deep learning neural network for time series air quality forecasting' [3] reviewed recent deep learning applications for time series air quality forecasting, and combinations of multiple components that produced hybrid forecasting models were suggested for potential superior performance and improved accuracy. Hybrid models may increase computational complexity and reduce the time efficiency of the models, which can be a downside of using hybrid models. The main components of deep learning studied were feature extraction, data decomposition, and spatiotemporal dependency. Various combinations of deep learning input parameters were presented for different problem requirements (different applications studied).

'Machine learning algorithms in air quality modeling' [12] analyzed 38 studies applying ML techniques and studied input predictors and the impact of inputs on prediction accuracy improvements, considering the geographical locations of studies. It explored techniques applied for pollutant concentration (forecasting/estimation) and linear regression, NN, SVM, and ensemble learning algorithms, etc. The study concluded that ML techniques are usually used and applied in North America and Europe, and multicomponent analysis (factorial analysis) showed that estimation for pollution was performed 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, noting estimation models' low variability and standard deviation. Forecasting was found to remain very limited with NN and SVM, and the study advised that other models and pollutants should be considered (specifically, NOx and SO_2; currently, there is more focus on PM10 and PM2.5). The authors also suggested considering other models (such as ensemble learning or others) to improve model accuracy.

The above analysis presents some highlights of the recent literature conducted to study the features used and models 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. The existing literature showed limitations in the presented comparative geographical contexts (e.g., comparing countries or cities) in the field of AQP analysis. Another major point to consider 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 an AQI framework. The current study seeks to contribute to emerging studies in this area.

3. Materials and Methods

3.1. Overview

First of all, the data part of the experiment consisted of data collection, data pre-processing and data preparation; Table 1 shows the main data sources.

Table 1. Main data sources.

Type	London	Jordan
Air quality station (data capturing)	Marylebone Road	Greater Amman Municipality
Collection years	2014–2018	2016–2018
Data points	43,824	26,268
Frequency	Hourly	Hourly

Hourly data in this research refers to the frequency at which air quality data are 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. It should be mentioned that the study for this work started in 2019, and historical data were requested from the Ministry of Environment. The data provided pertained to hourly data requests from 2016 to 2018. Data completeness was a very important factor in deciding on data selection. To conduct a comparable study with England's data and to observe the behavior of the modeling with larger datasets, data from 2014 to 2018 were selected. Additionally, data from Italy were used in this study as a new source of data (data not seen previously by the model) to ensure model generalization.

Data preparation was conducted using multiple methods, as described below.

Input Data: day, month, year, hour, humidity, temperature, wind speed, wind direction. Day, month, year, and hour columns were created by splitting the date time from the original raw data (from monitors) into separate columns. The year column was transformed to years 1, 2, 3, etc. (instead of 2014, 2015, 2016, etc.).

Output Data:

- England: CO (μg/m^3), NO (μg/m^3), NO$_2$ (μg/m^3), NOx (μg/m^3), O$_3$ (μg/m^3), PM10 (μg/m^3), SO$_2$ (μg/m^3).
- Jordan: PM10 (μg/m^3), NO$_2$ (ppb), CO (ppb), SO$_2$ (ppb).
- Figure 1 shows pollutants' concentration correlations.

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.

The flowchart of the experimental design stages shown in Figure 2 explains the major stages of the experimental work conducted in this paper. It can be seen that datasets were collected from three sources during the data collection process, and subsequent data preparation and pre-processing ensured data quality. Models were developed in phases based on the aim and objectives of this study. 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 2.

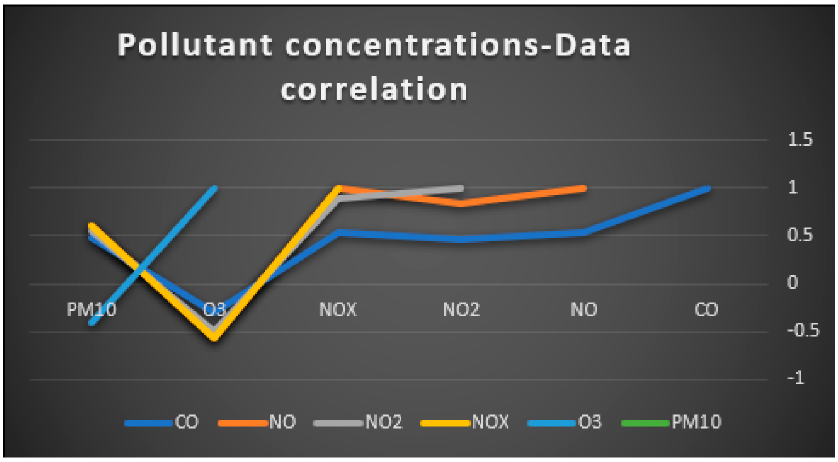


Figure 1. Data correlation.

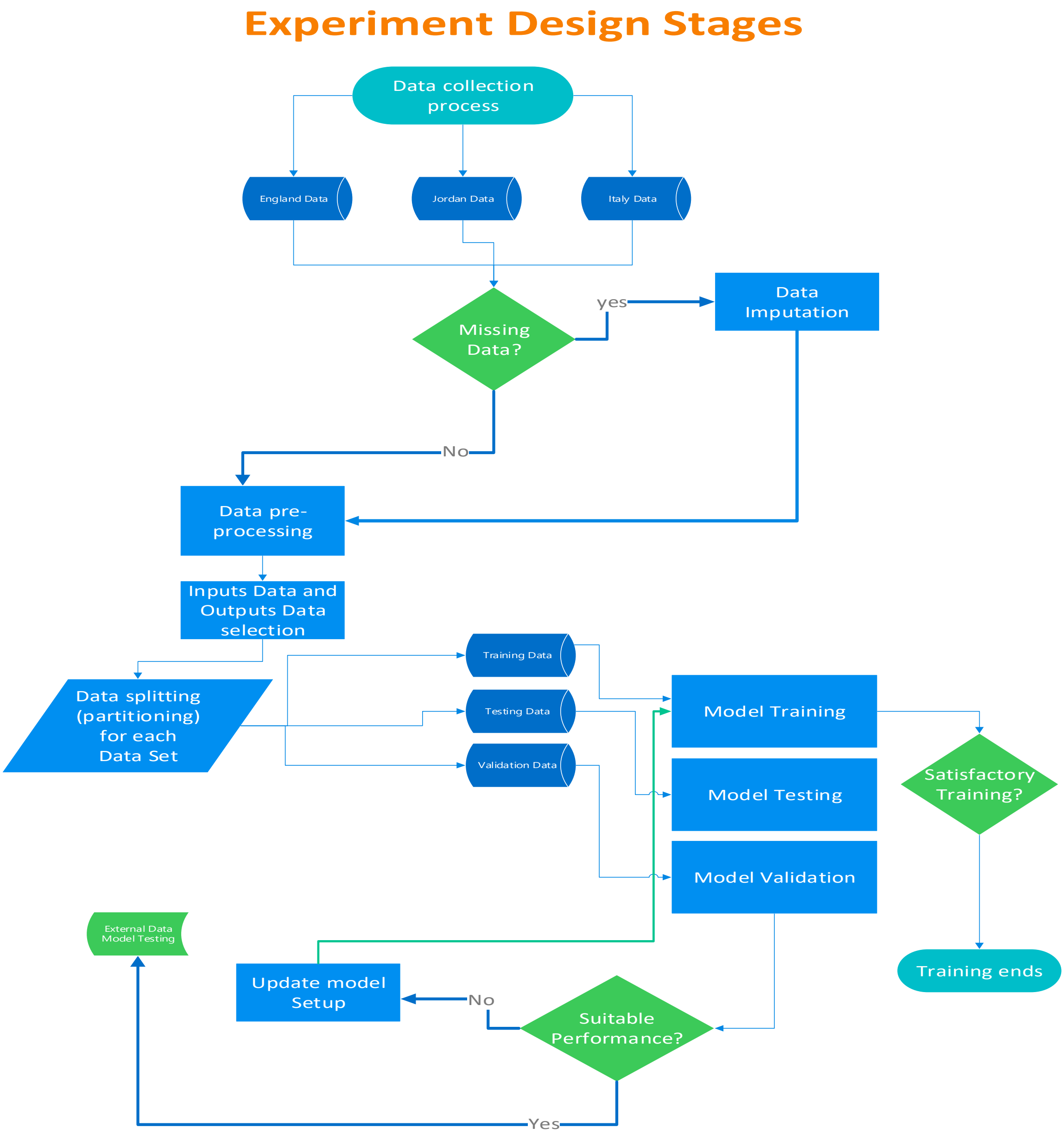


Figure 2. Experimental design stages.

The setup of model parameters is ultimately one of the most important parts of this research experiment, as attaining the desired reliability and accuracy of the model depends on fine-tuning parameters to fulfill the objectives of the study. DNNs and Markov-switching dynamic regression models have different sets of parameters, and the experiments were run for each standalone model until satisfactory results were obtained, after which they were combined to reach even better results than each standalone model. The following subsections present more about the parameters setup for each standalone model.

3.2. DNN

DNN was used with different setups (see Tables 2–4) to test multiple possible scenarios and select the topology that provides suitable results for the study in terms of accuracy and reliability. The DNN model was initially built using two LSTM layers architecture and was then optimized using the following techniques:

Table 2. DNN training options—England data.

England Data	
numHiddenUnits	800
maxepochs	1000
miniBatchSize	900

Table 3. DNN training options—Jordan data.

Jordan Data	
numHiddenUnits	300
maxepochs	500
miniBatchSize	256

Table 4. DNN layers architecture—England data.

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

Adaptive moment estimation (ADAM) is a common optimization algorithm specifically for DNN and, in particular, for neural networks. It provides an adaptive technique to adjust the learning rate for each parameter. It has been used in this experiment and proved enhancement in DNN model results for both England and Jordan data.

Initial learning rate is a hyperparameter that defines the size of the step taken during the optimization process to minimize the loss function. In other words, it controls how much weights are adjusted during training. Suitable initial rate values help balance the stability of the training to avoid getting stuck in the local minima. In this experiment, the initial learning rate parameter was set to 1×10^{-2}.

Learning rate schedule is useful for training efficiency, convergence, and model generalization. It is a pre-defined schedule for adjusting the learning rate during the training. In this experiment, the learning rate schedule was set to piecewise.

Learning rate drop factor helps generalization and convergence; it is a factor where the learning rate is reduced. It is used when techniques such as learning schedule is used. In this experiment, the learning rate drop factor was set to 1.

Learning rate drop period is the frequency at which the learning rate is reduced during training. It is used when techniques such as learning schedules are used. In this experiment, the learning rate drop period was set to 125.

3.3. Markov-Switching Dynamic Regression Model

The Markov-switching dynamic regression model built in this experiment has different components: autoregressive integrated moving average (ARIMA), random walks, probability, discrete-time Markov chain (DTMC), state transition, and also the simulation of inputs and outputs.

3.3.1. Autoregressive Integrated Moving Average (ARIMA)

ARIMA model uses the auto-regressive component (number of delays in the model), the integrated component (the differentiation degree for converting time series into stationary series), and the moving average component (past errors required to explain current error value). ARIMA performs prediction based on past and present data. The states of the Markov-switching regression model in the experiment were identified based on the number of inputs, each of which was converted to the ARIMA model with the experimental number of parameters. The following inputs were used as parameters for the ARIMA model:

- Hour: the hour of the day (for 24 h)
- Day: the day of the month
- Month: the month of the year
- Year: the year
- Wind speed
- Wind direction
- Temperature
- Humidity

3.3.2. 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 modeling, random walk evolves according to the parameters associated with each regime.

3.3.3. Probability Matrix

To create a probability matrix, a state vector is initiated with zeros for eight 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 is used to create a Markov probability matrix using the 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, as shown below:

$$p = \begin{bmatrix} p_{11} & \cdots & p_{12} \\ \vdots & \ddots & \vdots \\ p_{21} & \cdots & p_{22} \end{bmatrix} \tag{1}$$

3.3.4. Discrete-Time Markov Chain (DTMC)

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 a Markov chain, while the sequence of transitions from one state to another can be described as a stochastic matrix (i.e., indicating the probability of states transitioning).

3.3.5. State Transitioning

State transitioning or regime-dependent covariance was based on states being defined using eight variables (MdlX1, MdlX2... MdlX8), each of which represents inputs as states for the Markov model. Each variable is assigned to the ARIMA model, whose parameters are 'AR', 'beta', 'constant', and '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))

3.3.6. Parameters

The ARIMA model consists of the AR parameter, beta parameter, constant parameter, and variance.

AR equals 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 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 beta parameter of the ARIMA model 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 stored in one matrix variable, MdlX.

3.4. Input Simulation

Inputs (input1, input2... input 8) were simulated using the simulation function based on equal probability for each of the four states (transition probability).

3.5. 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 a simulate function with Mdl parameter, with the number of observations represented by a number of rows of outputs and the observed output data. Each output is represented by a simulated 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).

3.6. Multi-Input Multi-Output Hybrid Deep Neural Network Markov (DNNM) Model

3.6.1. Overview

The proposed multi-input, multi-output hybrid DNNM model achieves reliable accuracy of hourly time-series data and provides a large dataset for 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 by selecting the best ML methods to support air quality analysis.
- Increased reliability and accurate modeling to predict air quality.
- An effective AQI model for policy and regulation, supporting health and climate change issues.
- Considering transportation/traffic factors.

See Figures 3 and 4 for the deep learning model flow chart and Markov chain model flow chart.

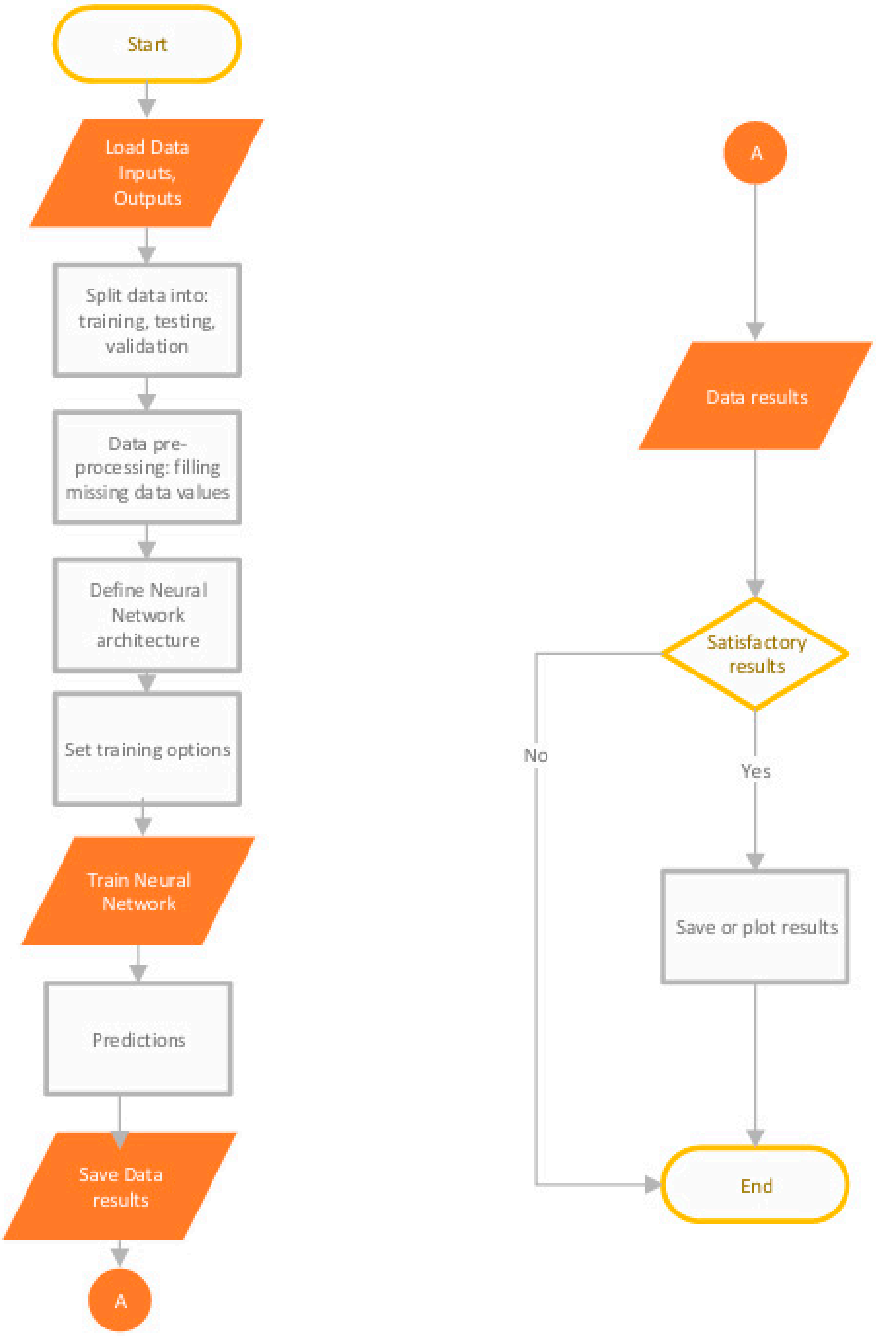

Start
Load Data Inputs, Outputs
Split data into: training, testing, validation
Data pre-processing: filling missing data values
Define Neural Network architecture
Set training options
Train Neural Network
Predictions
Save Data results
A
A
Data results
Satisfactory results
No
Yes
Save or plot results
End

Figure 3. Deep learning model flowchart.

This paper presents fuzzy logic as a classifier method for air quality levels; the method builds on findings of continued research in the area of air quality predictions. Previous work [13,14] contributed a hybrid DNN–Markov model. After numerous trials, the authors concluded their output of the study by combining the DNN model as a dynamic method to cover the data complexity and the Markov simple model to assist in more interpretable data and to back for errors that occur from the DNN model by simulating data through Markov switching dynamic regression type of models using ARIMA models to build the states (i.e., wind speed, wind direction, temperature, and humidity) and then outputs (which represent gases) are simulated for the hybrid model results. The final results of the hybrid model were then implemented using the fuzzy logic model (as post-prediction categorization).

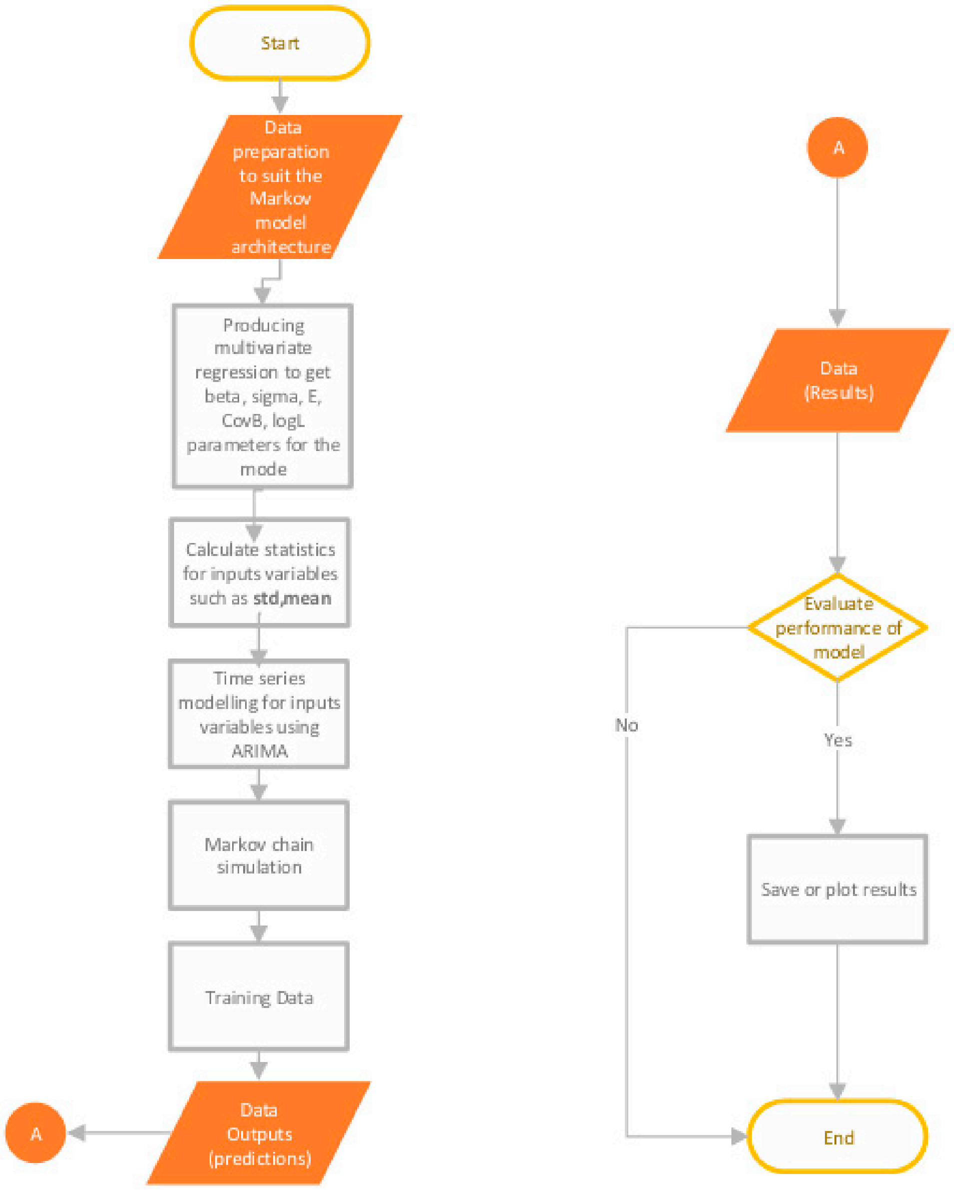

Start
Data preparation to suit the Markov model architecture
Producing multivariate regression to get beta, sigma, E, CovB, logL parameters for the mode
Calculate statistics for inputs variables such as **std,mean**
Time series modelling for inputs variables using ARIMA
Markov chain simulation
Training Data
Data Outputs (predictions)
A
A
Data (Results)
Evaluate performance of model
No
Yes
Save or plot results
End

Figure 4. Markov chain model flowchart.

Fuzzy logic was inspired by the pioneering work of Zadeh [15], who came up with the concept of extending the Boolean (zero or one) logic by allowing a degree of truth and values to be between zero and one. Fuzzy logic is based on the idea that in real-world applications, possibilities are not true or false, and fuzzy logic membership functions define the degree of membership or, in other words, how much an element belongs to a fuzzy set [16]. The reviewed literature explained how fuzzy logic was used by researchers as a comprehensive and reliable tool to further enhance the AQI predictions. It should be mentioned that the ANFIS (adaptive neuro-fuzzy inference system) MATLAB toolbox was used for the experiments reported in this work.

3.6.2. Process

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) [17] standard, using the flow logic shown in Figure 3. 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 5.

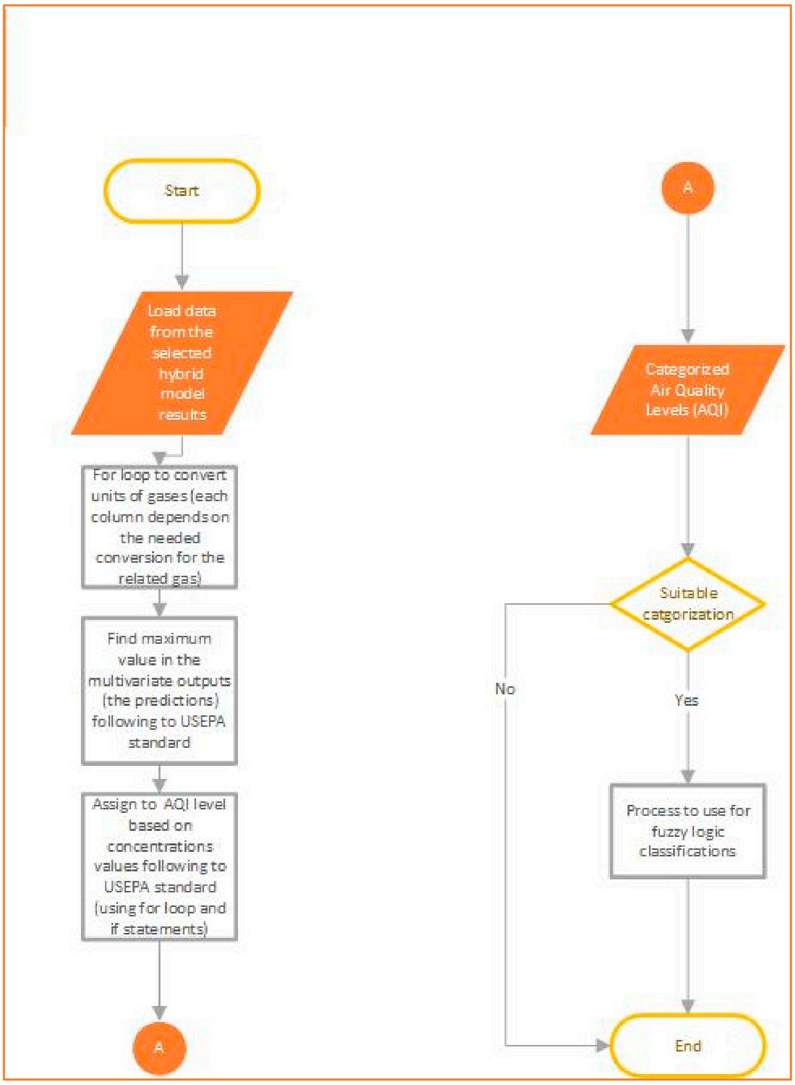

Start
Load data from the selected hybrid model results
For loop to convert units of gases (each column depends on the needed conversion for the related gas)
Find maximum value in the multivariate outputs (the predictions) following to USEPA standard
Assign to AQI level based on concentrations values following to USEPA standard (using for loop and if statements)
A
A
Categorized Air Quality Levels (AQI)
Suitable catgorization
No
Yes
Process to use for fuzzy logic classifications
End

Figure 5. AQI level calculations flowchart.

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 shows the whole AQI prediction process presented in this paper as part of continuous research performed in the air quality prediction field by the authors.

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 0 to 7 in the MATLAB code (e.g., 0 when it is less than zero and 7 when it is more than 500).

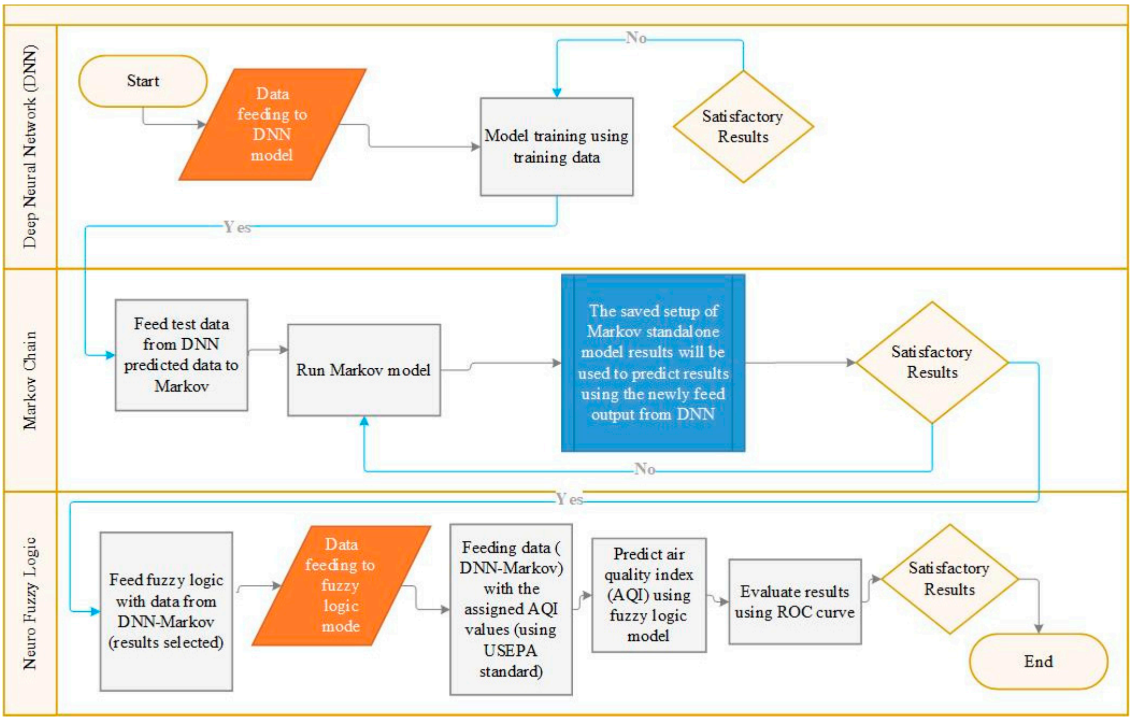

Deep Neural Network (DNN)
Start
Data feeding to DNN model
Model training using training data
Satisfactory Results
No
Yes
Markov Chain
Feed test data from DNN predicted data to Markov
Run Markov model
The saved setup of Markov standalone model results will be used to predict results using the newly feed output from DNN
Satisfactory Results
No
Yes
Neuro Fuzzy Logic
Feed fuzzy logic with data from DNN-Markov (results selected)
Data feeding to fuzzy logic mode
Feeding data (DNN-Markov) with the assigned AQI values (using USEPA standard)
Predict air quality index (AQI) using fuzzy logic model
Evaluate results using ROC curve
Satisfactory Results
End

Figure 6. AQI prediction process.

The air quality index prediction process (Figure 6) 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.

4. Results

4.1. Overview

This section presents description of the experimental results, as well as the experimental conclusions drawn from this research. The fuzzy logic classifier was built on the data models from the first part of the experiment of this work. The hybrid model (DNN–Markov) results are shown in Tables 5 and 6, and Figures 7 and 8. More details about the fuzzy logic classifier are explained in the following subsections.

Table 5. Modeling results: England.

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

Table 6. Modeling results: Jordan.

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

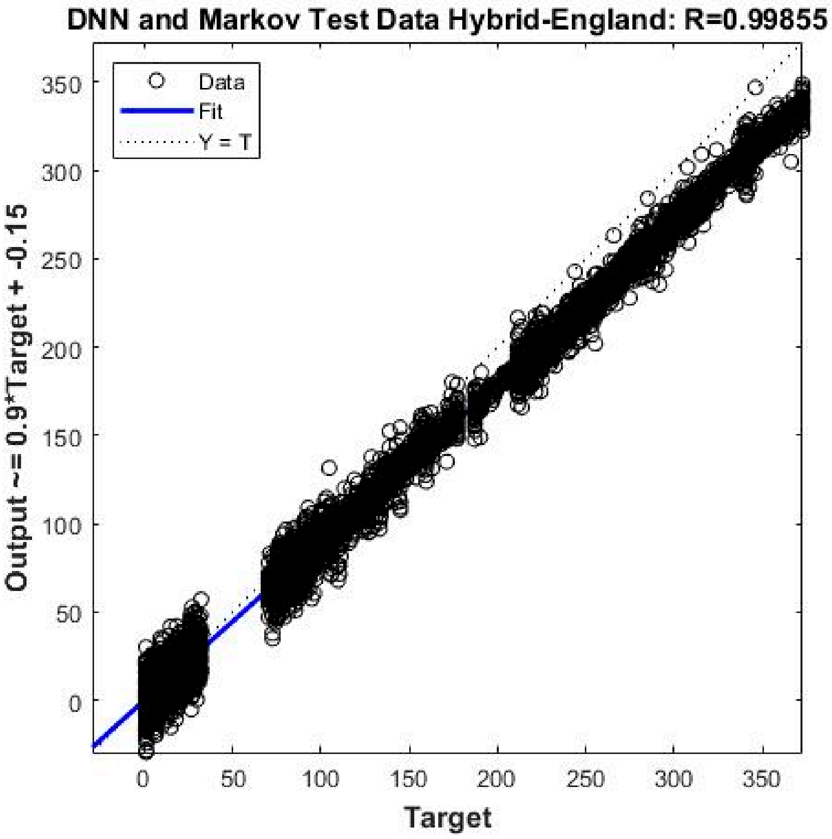


Figure 7. England air quality index prediction process.

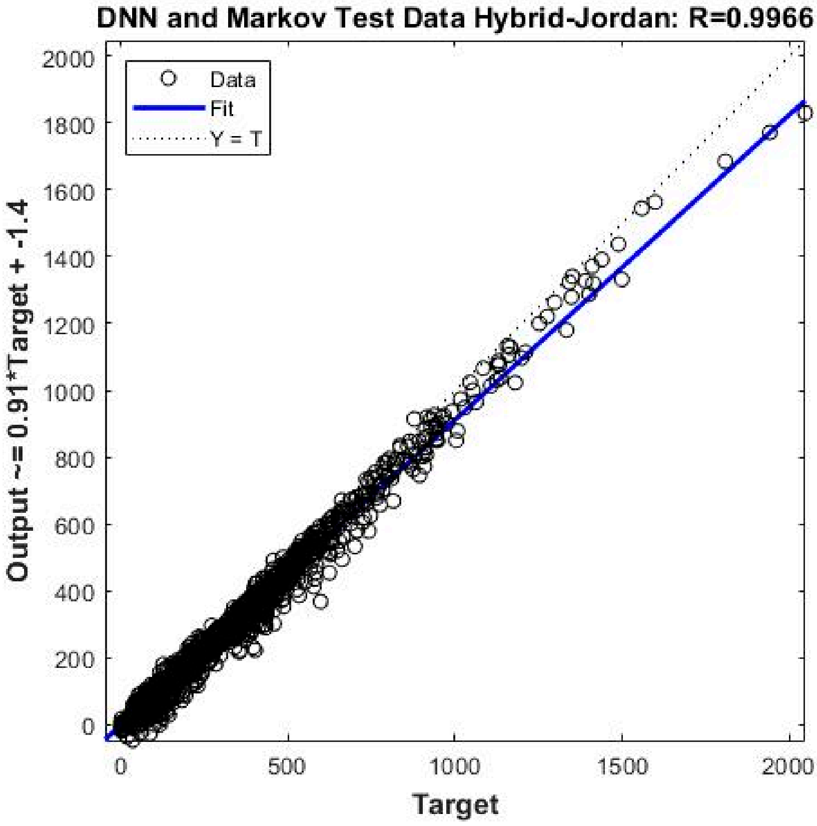


Figure 8. Jordan air quality index prediction process.

The results achieved from DNN–Markov (see Tables 5 and 6) indicated the best performance in this work experiment, and they were used to feed the fuzzy logic classifier for AQI representation.

Table 7 shows the model's generalization ability by achieving good accuracy using a completely new source of data (Italy data). The results displayed in Tables 5 and 6 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 7. Modeling validation (new data source from Italy).

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

There is obvious evidence from previous studies indicating the need for models that provide reliable, high-accuracy predictions while dealing with large datasets [2,11–13]. Additionally, there is a lack of ML models offering reliably accurate forecasting when predicting using Big Data. Put simply, there is an absence of reliable methods. The aim of this work is to present reliable and accurate machine learning methods for air quality predictions in light of the Big Data used in this experiment to test and verify the proposed methods. Hence, the methods provided in this paper demonstrate suitable accuracy and are recommended for use in the air quality prediction domain.

4.2. Fuzzy Logic for Air Quality Prediction

Steps to produce fuzzy logic classifier results for AQP:

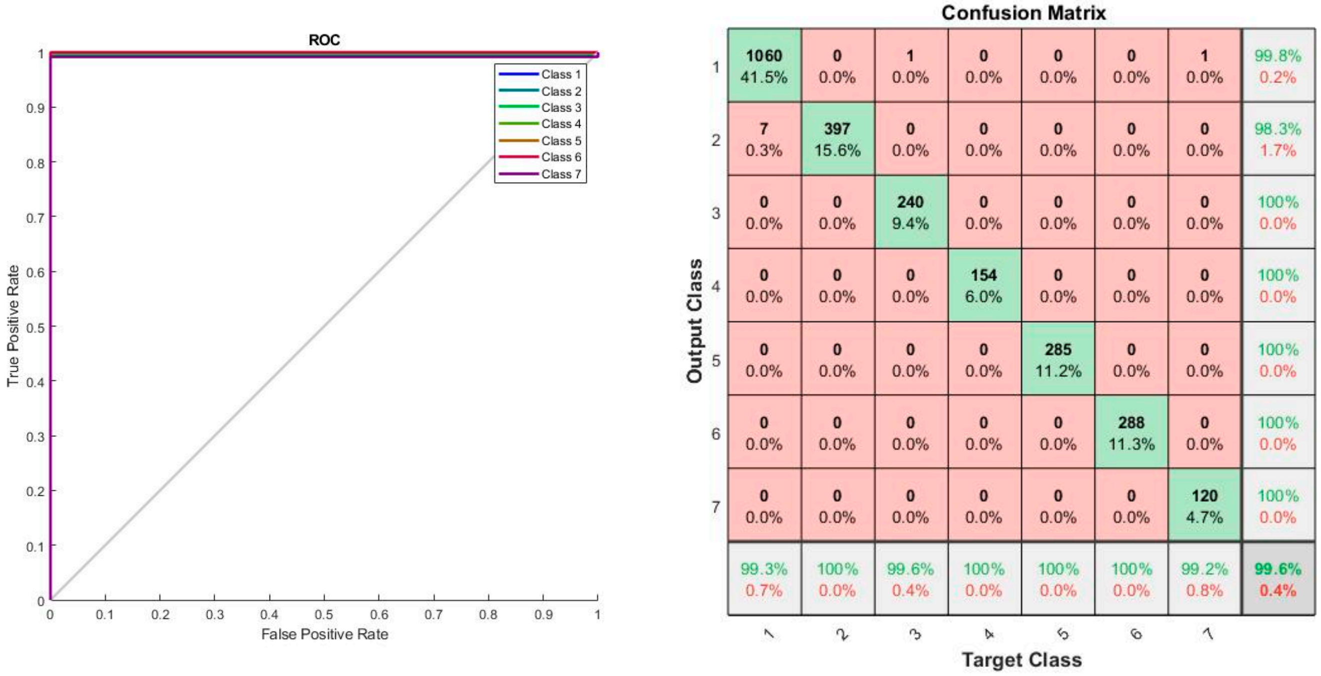
- Data trained using fuzzy logic with data (raw data outputs with corresponding AQI).
- Test data loaded in the fuzzy logic interface using (predicted data outputs with AQI value assigned).
- Rules exported and then executed the model exported using output predicted from DNN–Markov (the same data used in the previous step but without the assigned AQI).

Figure 9 shows 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 10 shows the fuzzy logic results (ROC and confusion matrices) for the England model, subject to the following data:

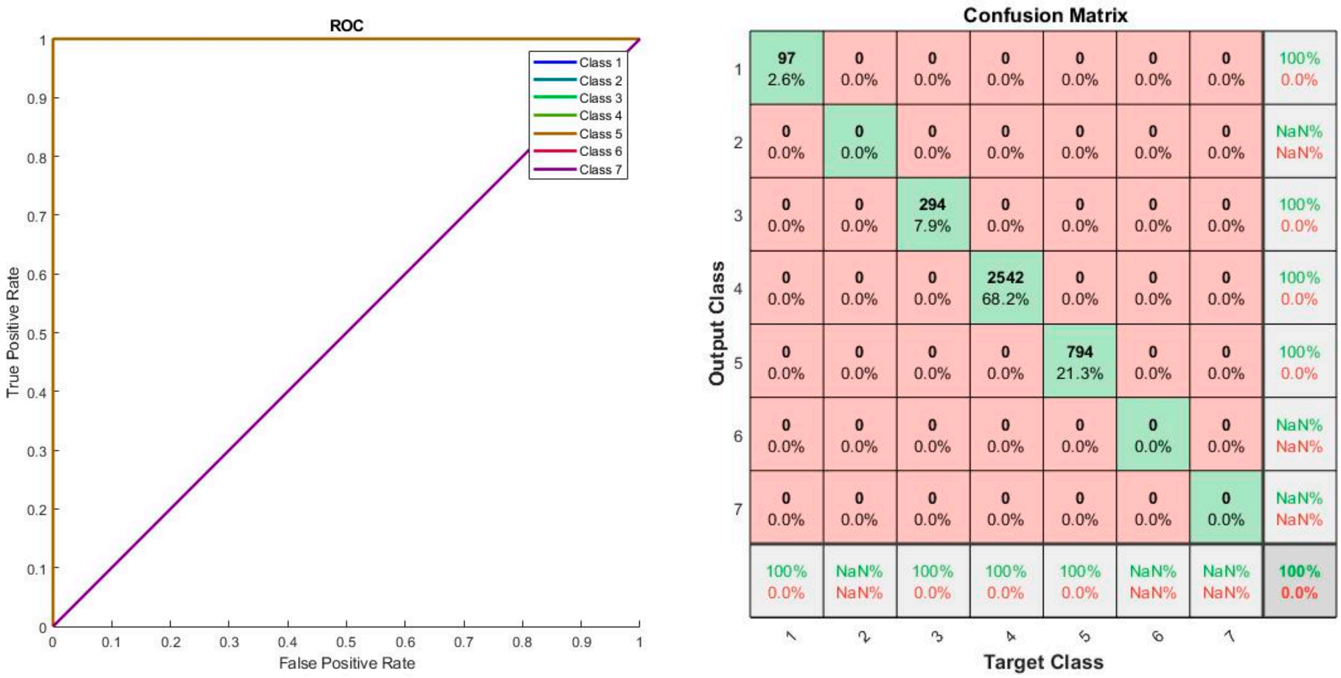
- Number of nodes: 294
- Number of linear parameters: 1024
- Number of nonlinear parameters: 28
- Total number of parameters: 1052
- Number of training data pairs: 4382
- Number of checking data pairs: 0
- Number of fuzzy rules: 128



	1	2	3	4	5	6	7	
1	1060 41.5%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	99.8% 0.2%
2	7 0.3%	397 15.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.3% 1.7%
3	0 0.0%	0 0.0%	240 9.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	154 6.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	285 11.2%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	288 11.3%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	120 4.7%	100% 0.0%
	99.3% 0.7%	100% 0.0%	99.6% 0.4%	100% 0.0%	100% 0.0%	100% 0.0%	99.2% 0.8%	99.6% 0.4%

(**a**) (**b**)

Figure 9. Fuzzy logic results for Jordan data: (**a**) ROC for Jordan test data and (**b**) confusion matrix for Jordan test data.

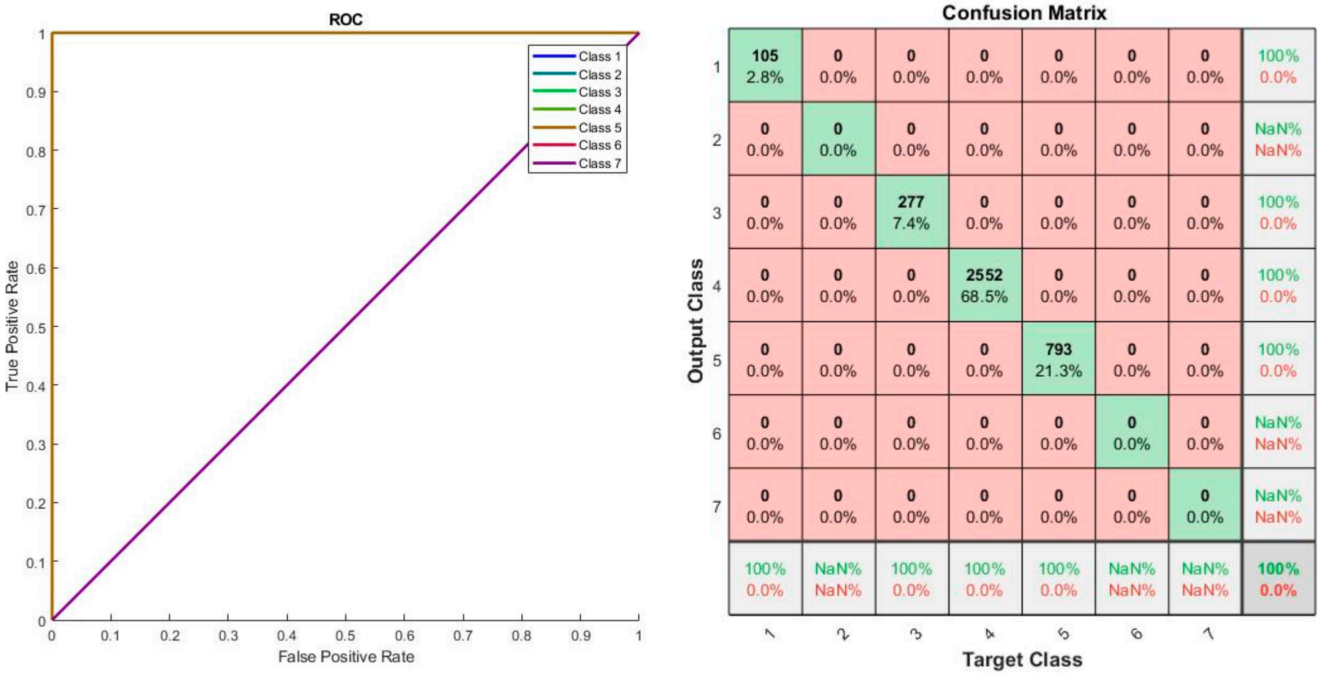


	1	2	3	4	5	6	7	
1	97 2.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
3	0 0.0%	0 0.0%	294 7.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	2542 68.2%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	794 21.3%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
	100% 0.0%	NaN% NaN%	100% 0.0%	100% 0.0%	100% 0.0%	NaN% NaN%	NaN% NaN%	100% 0.0%

(**a**) (**b**)

Figure 10. Fuzzy logic results for England data: (**a**) ROC for England test data and (**b**) confusion matrix for England test data.

Figure 11 shows the fuzzy logic results (ROC and confusion matrices) for the Italy model. It should be mentioned that the Italy data were a new source of ancillary data used for the purpose of validating the developed models.



(**a**) (**b**)

Figure 11. Fuzzy logic results for Italy data: (**a**) ROC for Italy test data and (**b**) confusion matrix for Italy test data.

4.3. Summary

As demonstrated in this section, the results were obtained through multiple steps and using several methods. Initially, a DNN was employed until satisfactory results were achieved, then a Markov model was utilized. Finally, both models were combined to enhance accuracy. The results from the combined model were then input into an ANFIS model to represent the AQI. 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 are 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 an approximation for the values within the specified rules, which makes the whole system more reliable for use in AQI. This paper 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. The experimental investigations presented in this work used statistical methods for prediction, aiming to select the best statistical model and refinement method for air pollution and meteorological data modeling in order to predict and categorize 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. The generated models are able to address missing data problems, complex gas predictions, and accuracy issues. The research experiment design and results were based on the data studied in this work, and the results reflect the parameters discussed in this paper.

5. Conclusions

This paper argues that the developed predictive framework for air quality indices offers an effective method of measuring healthy levels of the air we breathe every day. This research introduces viable and effective methods to predict hourly emissions concentrations and produce related AQI levels as a control system for vulnerable areas facing high toxic pollutant exposure. The paper's experimental scenarios affirm the efficacy of the studied methodologies, architectures, and optimization of standalone models, and the performed hybridization presents the best-reach scenario of the hybrid model and the selection along with results and performance of models.

It is 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. This paper presents an AQI framework using neuro-fuzzy logic, moving research forward concerning the conversions needed for the results and connecting the dots in proposing neuro-fuzzy logic to illustrate the refined outputs. The main limitations of this study are noted below.

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. The limited availability of air quality data, particularly in developed countries, poses a challenge.

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.

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.

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 to Jordan and the Middle East) could impact the precision of AQPs.

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 [18].

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.

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.

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 analyzing trends from 1990 to 2000 [19].

This work discussed several methods to address the gaps in accuracy and reliability when predicting with Big Data and identified a method for AQI interpretation from the predicted data. It examined the representation of pollution levels, highlighting the need for a global framework to unify the measurement of air quality indicators, given the varied standards in different countries and regions and the lack of supportive global standards for pollution level comparisons. Future research is needed to progress towards a more global unified framework for AQI prediction, starting with selected cities due to the challenging nature and scale of such research.

Author Contributions: Conceptualization, R.Z. and M.A.; methodology, R.Z. and M.A.; software, R.Z. and M.A.; writing—original draft preparation, R.Z.; writing—review and editing, R.Z. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Brunel University London.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, M.A., upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Plaia, A.; Ruggieri, M. Air Quality Indices: A Review. *Rev. Environ. Sci. Biotechnol.* **2011**, *10*, 165–179. [CrossRef]
2. Tealab, A. Time Series Forecasting Using Artificial Neural Networks Methodologies: A Systematic Review. *Future Comput. Inf. J.* **2018**, *3*, 334–340. [CrossRef]
3. Zaini, N.; Ean, L.W.; Ahmed, A.N.; Malek, M.A. A Systematic Literature Review of Deep Learning Neural Network for Time Series Air Quality Forecasting. *Environ. Sci. Pollut. Res.* **2022**, *29*, 4958–4990. [CrossRef] [PubMed]
4. Liao, K.; Huang, X.; Dang, H.; Ren, Y.; Zuo, S.; Duan, C. Statistical Approaches for Forecasting Primary Air Pollutants: A Review. *Atmosphere* **2021**, *12*, 686. [CrossRef]
5. Cheng, W.-L.; Chen, Y.-S.; Zhang, J.; Lyons, T.J.; Pai, J.L.; Chang, S.-H. Comparison of the Revised Air Quality Index with The PSI and AQI indices. *Sci. Total Environ.* **2007**, *382*, 191–198. [CrossRef] [PubMed]
6. Sowlat, M.H.; Gharibi, H.; Yunesian, M.; Mahmoudi, M.T.; Lotfi, S. A Novel, Fuzzy-Based Air Quality Index (FAQI) for Air Quality Assessment. *Atmos. Environ.* **2011**, *45*, 2050–2059. [CrossRef]
7. Monteiro, A.; Vieira, M.; Gama, C.; Miranda, A.I. Towards an Improved Air Quality Index. *Air Qual. Atmos. Health* **2017**, *10*, 447–455. [CrossRef]
8. Bishoi, B.; Prakash, A.; Jain, V.K. A Comparative Study of Air Quality Index Based on Factor Analysis and EPA Methods for an Urban Environment. *Aerosol Air Qual. Res.* **2009**, *9*, 1–17. [CrossRef]
9. Van Den Elshout, S.; Léger, K.; Nussio, F. Comparing Urban Air Quality in Europe in Real Time: A Review of Existing Air Quality Indices and the Proposal of a Common Alternative. *Environ. Int.* **2008**, *34*, 720–726. [CrossRef] [PubMed]
10. Rybarczyk, Y.; Zalakeviciute, R. Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Appl. Sci.* **2018**, *8*, 2570. [CrossRef]
11. Méndez, M.; Merayo, M.G.; Núñez, M. Machine Learning Algorithms to Forecast Air Quality: A Survey. *Artif. Intell. Rev.* **2023**, *56*, 10031–10066. [CrossRef] [PubMed]
12. Masih, A. Machine Learning Algorithms in Air Quality Modeling. *Glob. J. Environ. Sci. Manag.* **2019**, *5*, 515–534. [CrossRef]
13. Zayed, R.; Abbod, M. Big Data AI System for Air Quality Prediction. *Data Sci. Appl.* **2022**, *4*, 5–10.
14. Zayed, R.; Abbod, M. Hybrid Intelligent Modelling for Air Quality Prediction Deep Learning and Markov Chain Unconventional Framework. *Int. J. Simul. Syst. Sci. Technol.* **2022**, *23*, 3.1–3.6. [CrossRef]
15. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
16. Zadeh, L.A. Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets Sys.* **1999**, *100*, 9–34. [CrossRef]
17. Environmental Protection Agency. How Benmap-CE Estimates Health and Economic Effects of Air Pollution. Available online: https://www.epa.gov/benmap/how-benmap-ce-estimates-health-and-economic-effects-air-pollution (accessed on 10 February 2024).
18. Chapman, L. Transport and Climate Change: A Review. *J. Trans. Geogr.* **2007**, *15*, 354–367. [CrossRef]
19. Fenger, J. Air Pollution in the Last 50 Years—From Local to Global. *Atmos. Environ.* **2009**, *43*, 13–22. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.