National College of Ireland

Prediction an air quality index data using machine learning and deep learning

MSc Research Project
Data Analytics

Ruchita Patil
Student ID: 19197411

School of Computing
National College of Ireland

Supervisor: Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Ruchita Patil
Student ID:	19197411
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Bharathi Chakravarthi
Submission Due Date:	16/08/2021
Project Title:	Prediction an air quality index data using machine learning and deep learning
Word Count:	5994
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ruchita Patil
Date:	20th September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction an air quality index data using machine learning and deep learning

Ruchita Patil
19197411

Abstract

Environment sustainability has now become an important aspect of daily life. Air pollution is one of the most serious risks to the environment's long-term viability. Delhi, the national capital of India has been facing with the issue of poor air quality index for quite some years. The poor air quality has been negatively impacting the life of residents. As it is said, prevention is better than cure, it would be meaningful to predict the future scenarios beforehand to be better prepared to deal with it. This thesis uses various time series forecasting techniques to predict the Air Quality Index of Delhi for next few time periods. The pollutant levels for Particulate Matter (PM2.5, PM10), Sulphur Dioxide (SO2), Carbon Monoxide (CO), Nitrogen Dioxide (NO2) among other have been forecasted for a single chosen location in Delhi. The errors for various techniques have been reported. The findings of this research paper also include referring to other secondary sources that throws some light on the underlying issues of air pollution. This can thus be used for various future related studies that could come up in future and talks about Delhi's air pollution. For the suggested design, two models were integrated. The first layer used Gated Recurrent Unit, and all data was passed to the Long Short-Term Memory layer, which was followed by two dense layers. This proposed model is evaluated in comparison to the Long Short-Term Memory, Gated Recurrent Unit, Decision Tree, and Linear Regression models. Mean Square Error, Root Mean Square Error, and Mean Absolute Error are performance measures that are used to calculate error rates. When two models are combined, the overall performance of some factors improves.

1 Introduction

The Air Quality Index (AQI) is a statistic for assessing the quality of the air in our immediate surroundings. It measures how air pollution can affect an individual's health within a specified period of time. The common air pollutants are Particulate Matter (PM2.5, PM10), Nitrogen Dioxide (NO2) Carbon Monoxide (CO), Sulphur Dioxide (SO2) and Ozone (O3). There are 87 health risk factors calculated as the number of people who died in the year 2019, polluted air has High cholesterol, tobacco use, and dietary hazards are the top three global dangers. However, in India, air pollution has surpassed all other risk factors as the leading cause of premature mortality. More than 90 percent of people live in places where the World Health Organization's (WHO) guideline for healthy air is exceeded. More than half of the population lives in places that do not meet the WHO's toughest air quality standard. In 2019, India was again one of the top ten countries with

the most ozone (O3) exposure. In the last ten years, India has experienced the greatest increase (17 percent) in ozone (O3) concentrations. According to new WHO data, nine out of ten people breathe air that contains significant levels of contaminants. Each year, nearly 7 million people die as a result of breathing polluted air, according to the World Health Organization. As a result, monitoring the city's air quality and warning the public as soon as possible is vital.

In recent years, the air quality in most Indian cities has deteriorated dramatically. Aside from the standard pollutant like Carbon dioxide (CO2), many more recent pollutants like dioxide (NO2), Sulphur dioxide (SO2), Lead, Carbon monoxide (CO), Ozone (O3), Particulate Matter (PM10 and PM2.5) also has been added into the atmosphere. Most of the air pollutants are injurious to our health. But CO is that the most hazardous. it's also referred to as Because it takes life quietly and quickly, it is known as the Silent Killer. It enters blood cells directly and restores oxygen, depriving the brain and heart of the oxygen they require to function. If it's present within the air, it briskly enters the blood cells resulting in symptoms like headache, nausea, flu, confusion, dizziness etc. because the pollutant level increases, people get unconsciousness, vomiting and if exposure is simply too long it's going to result in damage of brain cells or perhaps causes death. The Environment is nothing but everything that surround us. The environment is getting polluted thanks to human activities and natural disaster, very dangerous among them is pollution. Meteorological elements such as atmospheric wind speed, wind direction, ratio, and temperature determine the concentration of air contaminants in the ambient air. If the humidity is increased, we feel much hotter because sweat won't evaporate into the atmosphere.

Raw data is difficult to grasp when it comes to assessing air pollution. It is for this reason that air quality indexes are established. For reporting daily air quality, the US Environmental Protection Agency (US EPA) devised an air quality index (AQI). The main focus of the AQI is on the health effect factor and the impact of polluted air. To prevent underestimating air pollution, government organizations should follow the EPA's instructions for AQI computations and communicate accurate and trustworthy AQIs to the public. AQI readings were calculated using the EPA's air quality index (AQI) method during the period studied. An air quality index is a grading system that illustrates how polluted the atmosphere is as well as the risks associated with each level. An air quality index (AQI) transforms numerical data into a qualitative grading scale that citizens of all ages may use to better understand the number of contaminants in the air they breathe.

To begin, average all contaminants using the EPA's rolling technique. The pollutant with the greatest average level was then discovered across all of the monitors. The observation concentration Cp is covered by two breakpoints. Where BP_Hi and BP_LO represent the higher and lower concentrations of the air pollutant, respectively, and I_Hi and I_Hi represent the AQI break point concentrations. The highest AQI rating from all air pollutants is then chosenYousefi and Hadei (2019).

$$I_P = \frac{I_Hi - I_Lo}{BP_Hi - BP_Lo}(C_p - BP_Lo) + I_Lo$$

1.1 Background and motivation

Environmental uncleanness, pollution, and contamination resulting from solid waste management are a universal issue that requires urgent investigation and solution. This urgent attention is necessary to address the current wave of pollution for economic sustainab-

ility. Much attention should be directed to emerging and developing nations on issues regarding waste management since there is uncoordinated day to day activities capable of prompting, encouraging and stimulating unsustainable management of waste. Developing facilities for managing waste is a function of the environment under consideration, and the nature of waste in such vicinity, be it in the rural settings or urban arrangement. Irrespective of the arrangement as mentioned earlier, negative operational precincts, political, technical and economic legislation's suffered in diverse operation systems. Developing and emerging countries suffer constant, continuous and uncontrolled waste disposal due to human activities that generate serious pollution, which affects the terrestrial and aquatic domain.

1.2 Research Question

This study is concerned with the research questions:

"How successfully can Deep Neural Networks and machine learning control air pollution and provide good health to living beings by forecasting air quality data?"

Table 1 shows the research questions objectives. Determine which model is best for data prediction based on data from India's air pollution. The second goal is to compare several models in order to determine the best forecasting model.

Table 1: Research Objective

Objective	**Implementation**
Objective1	Review of study related Air quality Index.
Objective 2	Selection and Pre-processing data related Air pollution.
Objective 3	Implement and evaluate forecasting model.
Objective 3.a)	IImplement and evaluate LSTM and GRU model.
Objective 3.b)	Implement and evaluate LR and DT model.
Objective 3.c)	Implement and evaluate proposed model.
Objective 4	Comparison of all model using various measures.

The results will be analysed and underlying causes for this state can be studied in detail. Following this, a roadmap can be created for dealing with this situation in a pragmatic manner. For finding out the underlying causes, data and matter will be collected and studied from reliable sources for meaningful deductions.

This research thesis has been divided in segments as per the following sequence- Part 2: "Literature Review" "Research Gap" which talks about the similar work earlier done and what new aspect this thesis is trying to bring, Part 3: "Research Methodology" which explains about the data collection and respective analysis that this thesis attempts to bring, Part 4: "Evaluation" that deep dives into the time series forecasting of Air Quality Indexes, Part 5: "Results Discussion".

2 Literature Review

2.1 Introduction

This section contains information on a quick explanation of how air pollution data is predicted using various models and what work has been done with these data. This

section separated into three parts; the first part discusses the how many types of deep learning approaches that are exposed to forecast the air quality index, in second part research about machine learning techniques. The final section contains related work completed in India.

2.2 Deep Learning techniques

Forecasting air quality is useful for preventing and reducing pollution. This research focuses on a multi-time forecasting model that was applied to data from Beijing's air quality. Beijing's data consider various pollutant, meteorological data and spatiotemporal data. The researchers compared the results models for deep learning such as LSTM, CNN, and BPNN. As a result, LSTM performs better, allowing enabling data forecasting using LSTM. Overall, either spatial or seasonal clustering-based forecasting is better appropriate for improving forecasting in a particular cluster or season Yan and Li (2021).

India consist Central and state control board using monitoring program they cover 240 cities air quality data. 342 monitoring stations were used by the board, which loaded data hourly and daily. We can analyze those data. The central and state control boards gave statistics for the entire country, but the researcher focused solely on the city of Chennai. After gathering data from Chennai, use preprocessing techniques to eliminate missing variables. The SVR and LSTM models were applied for AQI classify the data. The deep learning process reliably forecasts AQI values and aids in the planning of a metropolitan city for long-term sustainability. By adding traffic signal synchronization, promoting the use of public transportation, and Increasing the number of trees in specific places, the projected AQI value can limit pollution levels Janarthanan R (2021).

2.3 Machine Learning techniques

In this study, machine learning techniques such as SVM, BTR, and ANN are used to reduce PM10 and PM2.5 pollution levels in a reduction scenario. Data was collected from the AQM site, which comprised polluting gases and temperature. They employed the principal component analysis method to pre-process data for feature selection. They employed a variety of metrics to assess the project, including RMSE, Coefficient of Efficiency, and Coefficient of Correlation. Examine both the testing and training datasets and compare the results for the best results. The traffic composition of each vehicle, such as an electric car, a diesel car, a bus, or a coach, is studied by researchers Suleiman A (2019).

Humans and organizations are increasingly being warned about the dangers of public health pollution, which is contributing to worldwide air pollution issues. Air and air pollution, as well as meteorological molecules, are discharged into the air by enterprises such as gaseous fuel. Air pollution, specifically PM10 and PM2.5, is a primary pollutant that has a negative impact on human health. According to other publications, the Taiwan Air Quality Collection focuses at air molecules. This study used two methods for missing value pre-processing: spline multinomial and Fourier arrangement. TAQMN data was used, which was for 6 years of information. For forecasting data DT, RF, Gradient Boosting Regressor and LR methods was used. The performance of the modules Gradient Boosting Regressor was evaluated using RMSE, MAE, MSE, and Coefficient of Determination, and all of them outperformed the others Harishkumar and Gad (2020).

2.4 A comparative review of air pollution in India

Summary of here, the researchers have used two methodologies for forecasting the AQI (Air Quality Index). Firstly, they have used Single Exponential Smoothing (SES) technique and secondly, the researchers have used Double Exponential Smoothing technique that allowed the existence of trends in data. The dataset used for forecasting was information gathered from the official website of CPCB (Central Pollution Control Board) for the city of Kolkata (India). The data used was a half-hourly data obtained over a period of 24 hours from 17th October 2017 to 18th October 2017. The researchers have considered forecasting for significant contaminants in the air, namely, PM 10 (Particulate Matter 10), Carbon Monoxide (CO), Ozone (O3), Nitrogen Dioxide and Sulphur Dioxide. For the purpose of forecasting using exponential smoothing, the researchers have used three values of smoothing constant (), =0.1, =0.5 and =0.9. The above-mentioned paper didn't discuss about the major air polluting substance: PM 2.5 Bose and Sarddar (2020).

The researchers considered the Air Quality of Hong Kong for air pollutants PM 2.5, NO2 and O3. Their study considered 'Daily average', 'Hourly average' and 'Daily 1-hr Maximum average O3' and 'Daily 8-hr Maximum average for O3'. The paper used ARIMAX (Where, X $\rightarrow$ Output from Numerical Models) for forecasting. The analysis was done on R platform using a suitable library. There were three sites of Hong Kong that were used in the analysis- roadside, urban and rural sites namely. Ultimately the paper provides 1-day forecast, 3-day forecast and some other forecast. Liu and Fung (2018).

This study was carried out in Bulgaria and the data-set for 1-year period was considered starting from 1st September 2011 to 31st August 2012. The paper used ''Factor Analysis' and 'Principal Component Analysis' to find col-linearity between the pollutants and found high correlation in six of them. The researchers then grouped together the pollutants in three categories and interpreted them as three sources of pollution. Finally, the researchers carried out time series analysis for hourly data using SARIMA Models Gocheva-Ilieva and Boyadzhiev (2014).

Here, the researchers carried out time series analysis using ARIMA model to forecast daily mean of air pollutants at ITO, Delhi for the pollutants O3, CO, NO and NO2. The data were collected by the researchers for a duration of almost 350 days starting from 3rd Aug 2006 to 17th July 2007. The detailed analysis were computed on MATLAB Kumar and Jain (2010).

In this evaluation we got to know that this document aims to provide information on the current state of air quality in India's cities, as well as the consequences of air pollution. The NAQI allows for comparisons between cities, allowing for the development of new regulations to Particulate matter in the air should be reduced. The frequency of urban air pollution in India has been increasing in recent years, resulting in a perilous web of particulate matter and poisonous compounds in the air. He came to the conclusion that the level of RSPM AND SPM is exceedingly high based on his studies. Air pollution is developing as a result of vehicles and automobiles. To reduce air pollution, more efforts should be made. Instead of giving the possibility to create pollution, it is preferable to use measure the safety step and also take a strong step to prevent it. On a regular basis, an awareness program on the effects of pollution should be presented Nigam and Mhaisalkar (2015).

After evaluating we got to know that motor parks area in Ilorin Metroolis, Kwara state, Nigeria isalso being polluted due to particulate pollution. The motor park area

air quality should also need to be regularly survey to establish conformity to approve adjustment. He concluded from his research that main cause to decrease the air quality is due to PM2.5 and PM10. The main reason of increasing particulate matter is due to burning of firewood, wind blow dust, use of generators and using and mixing different chemical in atmosphere. High particulate matter can leads to respiratory and cardiovascular diseases. This study examines how air pollution affects the health. There is a robust connection between the health and air pollution; hence this research is based on alternative hypothesis because the researcher expects to find a link between air pollution and its effect on human health. It is hypothesized that the effects of air contamination on human health, warrants affirmative actions to protect human life and the environment **?**.

2.5 Identified Gaps and conclusion

Many academics use deep learning and machine learning approaches to forecast data, according to a literature study. Some researchers used all pollutants to predict values, whereas others focused solely on PM2.5 and PM10, which have an impact on human health. Human health is crucial. People can take precautions and survive if we forecast the values. Forecasting of both outdoor and indoor air pollution is mentioned in few papers. It is beneficial to govern the environment as well as preserve the lives of living beings if we can predict air quality.

Table 2: Objectives of review papers

Models	**Evaluation Parameters**	**Problem Addressed**	**Factors applied**	**Author Name**
LSTM, CNN	RMSE and IA	Compare seasonal and spatial cluster forecasting.	Air quality data, Meteorological data	Rui Yan a,b , Jiaqiang Liao a,c
SVR and LSTM	RMSE and R-Square	Compare existing techniques with proposed model.	PM2.5, NO2, SO2, CO	R. Janarthanan a, P. Partheeban
BRT, SVM	FAC2, R, COE, RMSE	Measure the effectiveness of particle reduction.	PM10 and PM2.5	A. Suleiman
LR, RF, DT	RMSE, MAE, MSE	Compare result with traditional model.	PM10 and PM2.5	Doreswamy

3 Methodology

Data mining is very crucial in determining the patterns especially from huge amounts of data. Data mining usually give insights as well as a correlation for data that was ignored or went unrecognized. CRISP-DM has been constantly in use especially in the development and optimization of materials. CRISP-DM impacts a lot in various fields where it is used. It has six phases namely business understanding, data or info understanding, data preparation, modelling, data evaluation as well as data deployment. At every phase, there are distinct processes and algorithms that are normally incorporated into the system. CRISP-DM technique has been employed in different fields such as financial, engineering as well as the medical world. Huge data can be managed using this system bearing in mind the complexity of the systems involved. CRISP-DM has enabled organizations to make vital decisions regarding their operations and this normally yields profits for the firm.

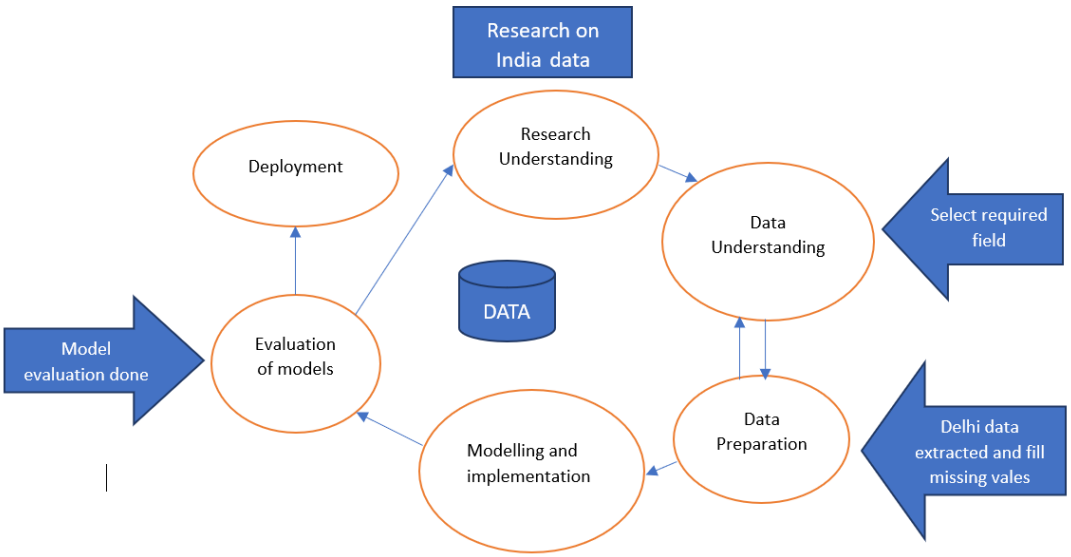

Figure 1: CRISP-DM Model

3.1 Business Understanding

Before any analysis process can begin, the data scientist must first gain a thorough grasp of the business. Air pollution is the most severe worry in today's world, owing to industrial pollutants that pollute the atmosphere. Environmental difficulties arise as a result of pollution-related health problems. Examining the significance of air pollution predicting air quality data is beneficial for future use, such as limiting air pollution and preventing health problems.

3.2 Data Understanding

The first step in understanding data is to figure out what's there in the data that the client already achieved. The main aim in this stage is to understand the customers' information

comprising of his/her geographical location, past purchases helping also to determine the customer's interest that may be used even later on as the business progresses.

This stage is normally carried out successfully beginning with collecting the initial data, analysing and giving the report of that data, description of the data collected and giving its report, exploration of the discussed data and giving the report on the same and finally conducting out verification of data quality as well as giving result on the same marking the end of the data understanding phaseda Rocha and de Sousa Junior (2010).

3.3 Data Preparation

The data preparation process entails cleaning as well as preparation of raw data from the business understanding. The data sampled and prepared are transmitted and subjected to the data mining algorithms where the outcome of the developed and the executed solutions depends on the quality of the data. The phase works out by changing the format of the data obtained either in a structured or unstructured manner as well as using the imputation method to handle the omitted dataDåderman and Rosander (2018).

3.4 Modelling

Fourthly, is the data modelling; in this stage various modelling ideas are chosen and used for the data feature execution, model data, turn on the model, and to calibrate its parameters to medium figures. Thus, the stage majorly involves application of a desirable data mining or the machine training algorithms to the given dataset and gives solution to most business problems by either use of a single approach or a composite of several techniques to evaluate the model and cub the issue at hand.

It's a statistical technique for analysing the pattern of data points in order to predict the future, data is gathered across time. Time series forecasting is a technique that uses time series data to estimate future values. Data obtained on a certain feature over a length of time at a fixed/regular interval is referred to as time series data. Error measures: (1) mean absolute error (MAE), (2) Root mean square error (RMSE) and (3) mean square error (MSE). Recurrent Neural Networks (RNN) and their special types, Gated Recurrent Units and Long-short Term Memory (LSTM), we implement hybrid model for better resultErskine and Grimaila (2010).

3.4.1 Long short-term memory

Input Gate, Forget Gate and Output Gate are the three types of gates available in LSTM model. The output of the input modulation gate is transferred to the memory cell, it is in charge of receiving all additional info from the outside world. The forget gate determines which data to keep and which to discard in the next generation. It selects the best delays for the supplied data series this way. The calculated results are fed into the output gate as input. The Long Short-Term Memory cell's output is generated by the output gate. A softmax layer is stacked on top of the LSTM's output layer in most language models. Our method, on the other hand, piles a dense layer on top of the LSTM cell's output layer.

The most significant things in LSTM are the cell states, which stands by the horizontal line at the top part of Figure. The cell state behaves like a conveyor belt so it keeps running in the chain of LSTM with some linear functions' transformation, which retains

the previous information. An activation function layer (typically sigmoid) and a point-wise multiplication operation are commonly used to produce the three gates that allow the LSTM is used to add or remove data from a cell's state.

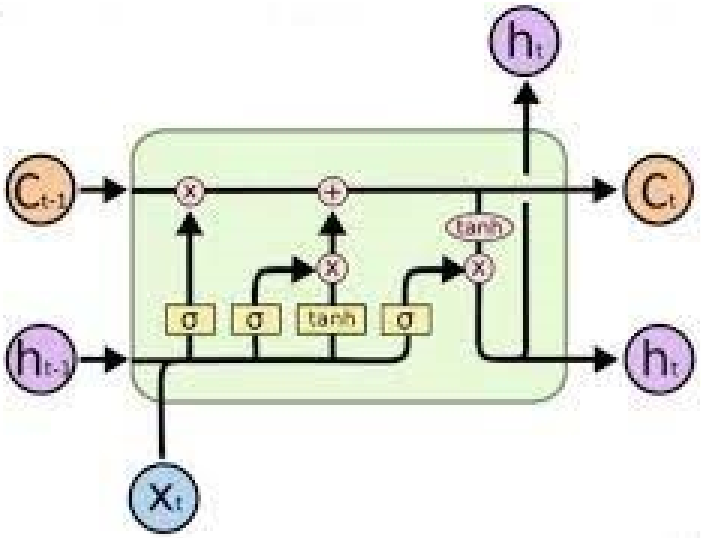

Figure 2: LSTM architecture Olah (2015)

3.4.2 Gated recurrent units

Due to the fact that the term LSTM Auto-encoder is commonly used throughout the machine-learning community, Seq2Seq models are not limited to LSTM units. LSTM units come in a variety of shapes and sizes. Now-a-days Gated Recurrent Unit is most commonly used. The unit is based on the LSTM unit, and it combines the input and forget gates into a

$$Z_t$$

update gate.

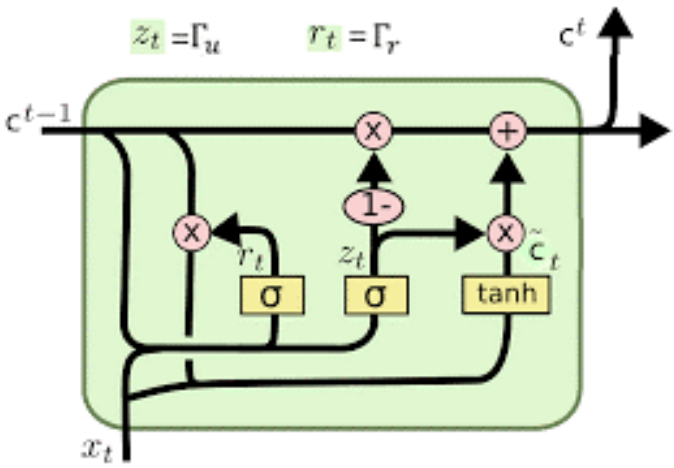

Figure 3: GRU architecture Olah (2015)

This phase determines how much data passes from previous hidden levels to subsequent hidden layers. Because there are only a few parameters, this unit is easy to calculate. There was also a redundant data reset gate that determined how much of the prior hidden state should be forgottenOlah (2015).

3.4.3 Linear Regression

The most fundamental and widely used type of data analytics is linear regression Regression's purpose is to look at two things: (1) Is it possible to determine an outcome (dependent) variable making use of a group of predictors variables? (2) Which characteristics in specific are significantly predictive of the dependent variables, and What effect do they have on the outcome variable–as indicated by the beta approximation's size and sign? Determining the intensity of predictors, anticipating an impact, and trend forecasting are three primary applications of regression analysis.

$$Y = B_o + B_1 X + E$$

3.4.4 Decision Tree

It's a classification algorithm of some sort. It can handle categorical as well as continuous inputs and outputs. We divide the data set into two or more homogeneous sets using a Decision Tree based on the best difference throughout the response variable. It contains a number of intriguing characteristics that address issues such as missing values, outliers, and determining the most significant dimensions; nevertheless, it does not function as well with continuous target variables as it does with categorical ones. It is beneficial for data exploration since it requires less data cleansing, the data type does not have to be constant, and it is a non-parametric method.

A decision tree is a tool for making decisions that must be projected in the form of a tree. The tool is extremely useful for debugging algorithms with conditional control statements. ID3 is the main algorithm for creating decision trees. To build a decision tree, ID3 uses Entropy and Information Gain. ID3 is the most often used algorithm for creating decision trees. To build a decision tree, ID3 uses Entropy and Information Gain.

$$E(S) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

3.5 Evaluation

Data evaluation is the second last stage which only acts as a monitoring stage to the modelling stage as the model building requires to be carefully and thoroughly looked upon before Data evaluation is the second last stage which only acts as a monitoring stage to the modelling stage as the model building requires to be carefully and thoroughly looked upon before. Finally, is the deployment stage which is the last stage that works out once the model building and data evaluation are completed and everyone is convinced and satisfied by the whole process.

R-squared is the percentage of variance in the outcome that can be described by the dependent variables (R2).

The Root Mean Squared Error is a measure of a model's average error in making predictions given observation (RMSE). In mathematics, the RMSE is the square root of the mean squared error (MSE), which is the minimum absolute discrepancy between observed actual output values and predicted values by the model.

The Mean Absolute Error (MAE), like the RMSE, quantifies the MAE prediction error. MAE is less vulnerable to outliers than RMSE. Kalgotra and Sharda (2016).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_{ACT(i)} - x_{PRED(i)})^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{ACT(i)} - x_{PRED(i)})^2}$$

4 Design Specification

The entire project is divided into several parts. First, we choose data to understand business and environmental needs. Next, we select air quality data, which is important for both the environment and human life. Predicting air quality is important to industry and pollution management. The second stage of data pre-processing Whatever data we have is raw, and if we apply it to a model straight, the output could be incorrect. Pre-processing is an important stage in data forecasting. When the values are too large for the model to understand, a transformation step is required. Following data preparation, we must use a variety of models. The next stage is evaluation, which is critical since whatever model we use, we must compare the results on an error basis in order to justify the better model.

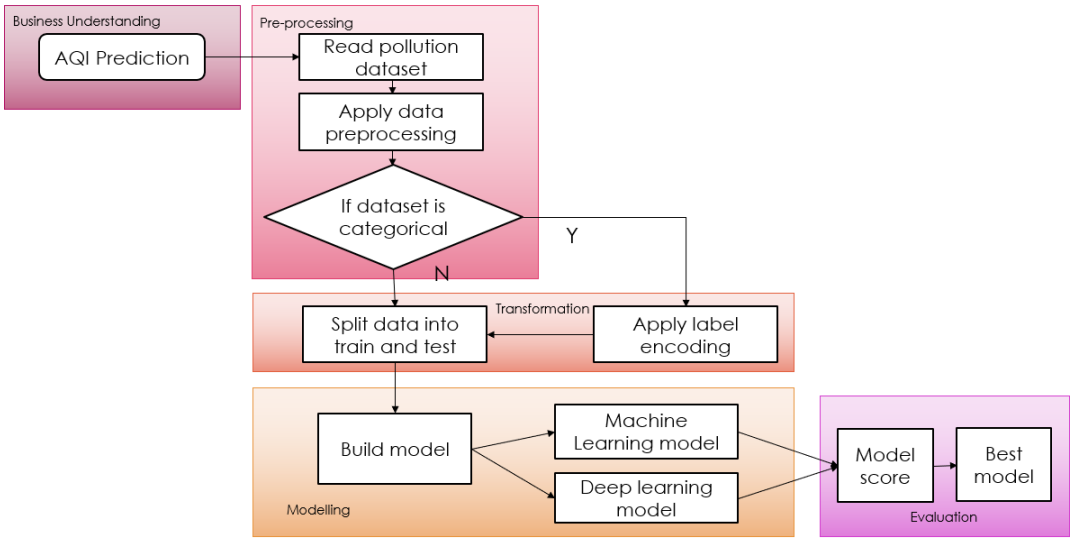

Figure 4: Flowchart of Proposed model

5 Implementation

5.1 Environmental setup

The project was created in the Python programming language. I utilized Google Colab to implement LSTM and GRU, and the main library included tersorflow, keras, dense, and sequencial, which are useful for neural network algorithms.

5.2 Selection of Data

In our research, the data is available on the Central Pollution Control Board's website, https://cpcb.nic.in/, which is the government of India's official portal. Dataset included different parameters such as particulate matter (PM2.5 and PM10), Nitrogen monoxide (NO), nitrogen dioxide (NO2), Nitrogen oxides (NOx), Ammonia (NH3), carbon monoxide (CO) sulphur dioxide (SO2), Ozone (O3).

5.3 Data Pre-processing

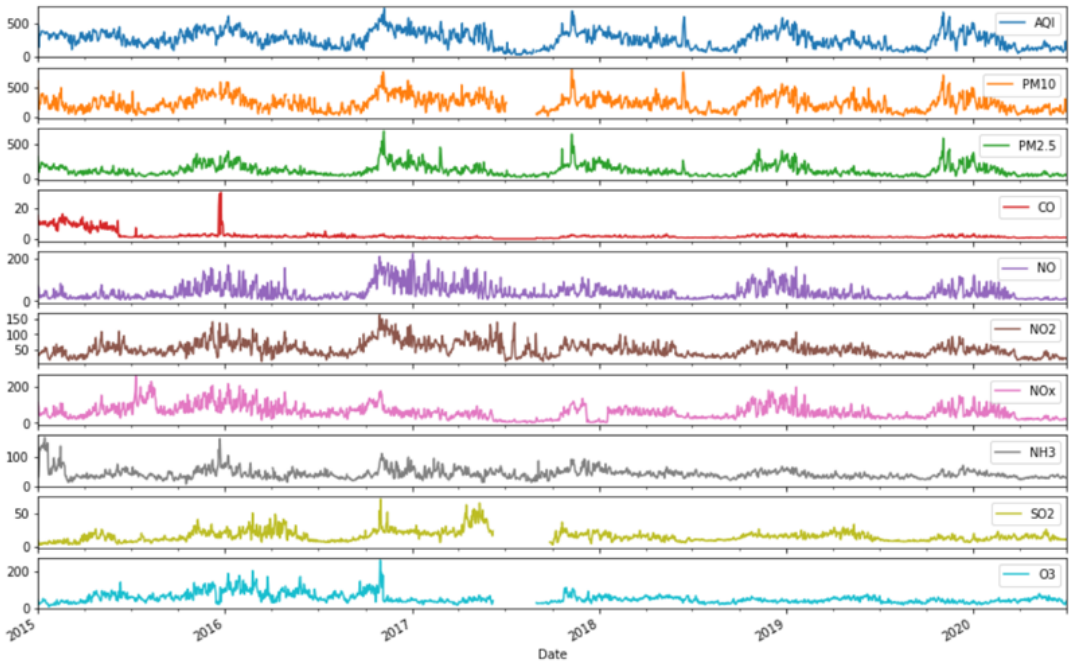

Figure 5: Each pollutant graph representation

On Google Colab, I installed all essential libraries and packages to handle various types of functions. All data was extracted from the website of the Central Pollution Control Board, which contains 342 monitoring stations. Because the data was so large, I constrained to data from the Delhi area.

For data analysis, plot each pollutant year by year so that we can see how each pollutant increases and decreases. Except for CO, almost every pollutant in the figure has a seasonal graph. The major variable here is the air quality index (AQI), which changes seasonally. We also noticed missing data in the graphs for PM10, SO2, and O3

pollution between 2017 and 2018. We can fill in missing data for good results using a variety of methods.

5.3.1 Data Cleaning

Data cleaning is an important part of the data preparation process. If we don't clean the raw data, the model's results may suffer, and the model may not suit the dataset. Cleaning data manually is insufficient; thus, I use the procedures below to clean data:

For improved results, check for missing numbers and fill all null values with the median function.

Check for duplicate values and eliminate them.

5.3.2 Data Transformation

We must check the datatype of all variables since some datatypes are incompatible with the operation. Date is already in object format in our scenario, however when we plot graph, we want date to be in datetime format. As a result, we change the date variable's format to Datetime. When applying a module to data, we must ensure that the array

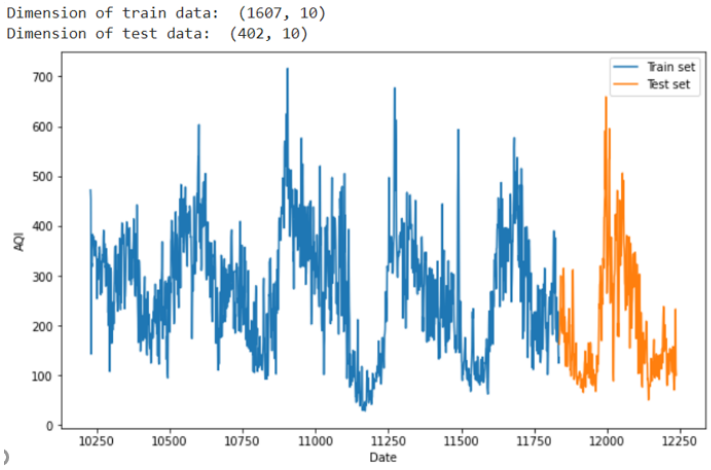

Figure 6: Split data into train and test dataset

dimension is compatible with our model implementation. We use a Min-Max scaler to alter data for a nice result. We used the drop function to separate data into X and Y variables. The input variable is X, and the output variable is Y. We have ten variables, but nine are dependents and one is independent, so we use the Drop function to insert all nine parameters into the x variable and only the AQI into the y variable.

For applying various types of models, we must divide data into train and test datasets. In this case, I chose 80 percent data for the train set and 20percent data for the test set. Prediction techniques operate in this manner: they first train the module with datasets, and then they need datasets for testing. After testing the model, we obtain the accuracy or error rate, which allows us to determine whether the model worked well or not.

5.4 Model Implementation and Evaluation

5.4.1 Implementation using LSTM

The first portion determines whether the prior timestamp's content should be retained or is unnecessary and can be ignored. The cell attempts to learn new knowledge from the input in the second section. Finally, the cell transfers modified information regarding the current timestamp towards the next timestamp in the third component. The gates are the three components of an LSTM cell. The Forget gate is the first section, the Input gate is the second, and the Output gate is the third.

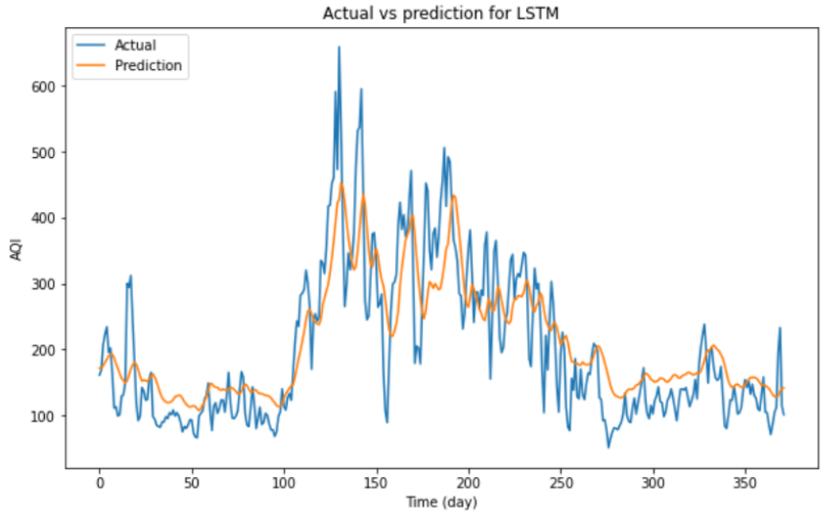

Figure 7: Actual vs Prediction LSTM

Prior to LSTM modelling, data scaling is a key consideration. When the neural network is exposed to the same scaled features, it converges faster and provides better accuracy. As a result, we split the data into a train and test dataset and used the Min-Max scaler approach to keep attribute variance within the same range. In addition, if one feature has more variance than others, it can affect the forecasting model's performance. We forecasted the data using AQI because the dataset is multivariate and the date is an index variable. As a result, the air quality attributes were scaled using Python's MinMaxScaler() method for maintaining records within a relevant interval and evaluate the model's correctness.

Table 3: Performance Comparison

Model	**R- square**	**RMSE**	**MAE**
LSTM	0.68	66.23	52.78

The model was built with an input layer, and then the next LSTM layer received sequence data rather than random data. To avoid overfitting the model, a dropout layer is used. Finally, we output one hot encoded result using a dense layer. Set up the model and begin training with a checkpoint and early stopping. When the monitored loss exceeds the tolerance, early stopping ends training, and checkpoint stores the model weight when it reaches the minimal loss.

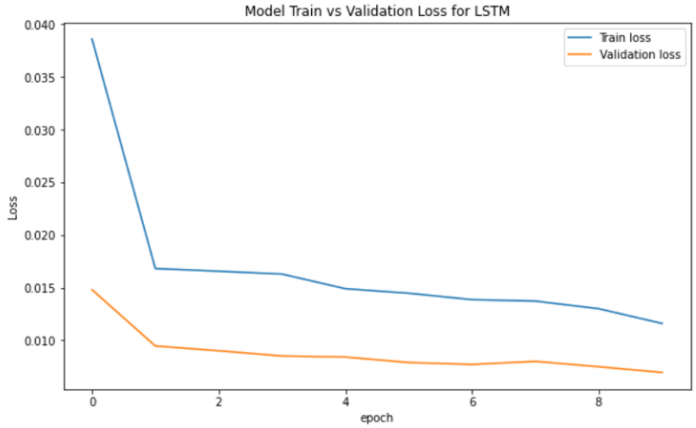

Figure 8: Train vs Validation Loss LSTM

5.4.2 Implementation using GRU

The GRU model uses the same dataset as the LSTM model. The MinMax function is used to scale data in LSTM. We start by passing data in a sequential order. In the GRU model, we pass 64 units.

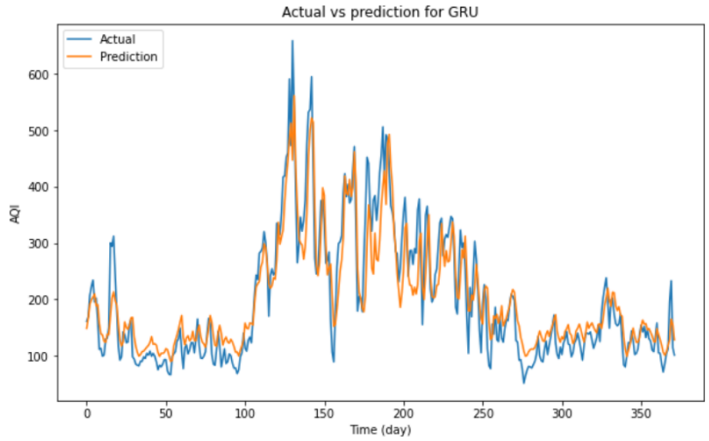

Figure 9: Actual vs Prediction GRU

We utilized the dropout function to avoid overfitting the data. We employed a dense layer with one unit to get a single variable output. Finally, we compile the module with the optimizer "adam" and use Mean Square Error to verify for loss.

Table 4: Performance Comparison

Model	**R- square**	**RMSE**	**MAE**
GRU	0.85	45.49	37.55

In one epoch, the full dataset is fed forward and backpropagated through the model. It's tricky to figure out how many epochs to use so that the model fits the data well without becoming overfitted. A model is considered to be underfitted if it fails to capture

the correlation between the independent (x) and dependent (y) variables. When a model is overfitted on the training data, it fails to generalize to the test data, which is known as overfitting.

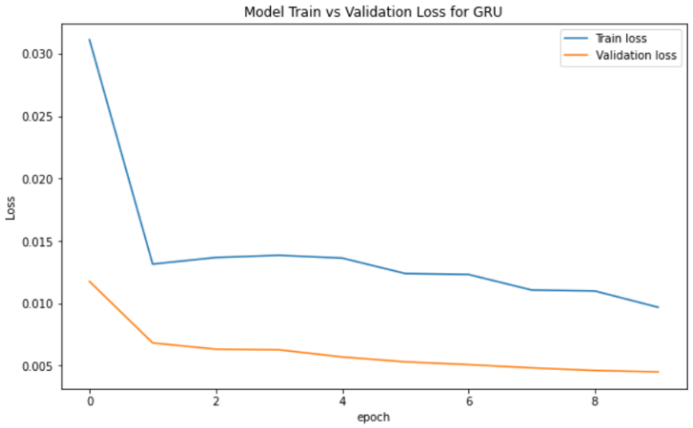

Figure 10: Train vs Validation Loss GRU

5.4.3 Implementation using LR

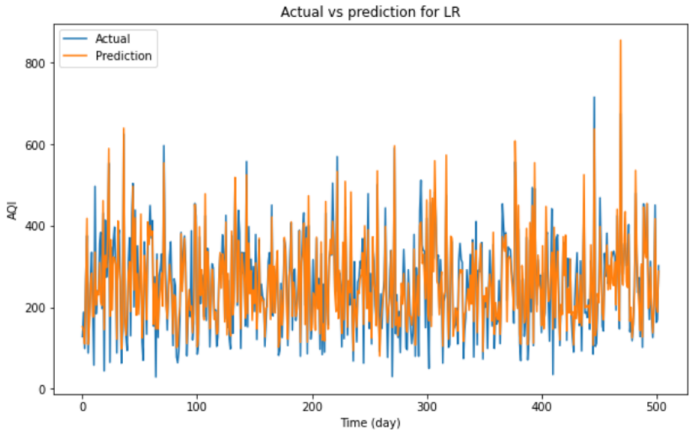

Figure 11: Actual vs Prediction LR

For visualization plot actual verses predicted graph for better understanding.

Table 5: Performance Comparison

Model	**R- square**	**RMSE**	**MAE**
LR	0.85	45.23	34.82

For evaluation data, I utilized the LinearRegression function in Python to fit the model with the train dataset and forecast the model with the test dataset, as well as the sklearn metrics library. I used the mean absolute error, the R-square function, and the passed test and projected values as metrics.

5.4.4 Implementation using DT

I used the DecisionTreeRegressor function to implement the decision tree, and I set the random state parameter to zero. The train data was fitted using the Fit algorithm. Additionally, test data was used to forecast the data. Linear regression is used in the evaluation process. The root mean square error was calculated using the sqrt library from math.

Table 6: Performance Comparison

Model	R- square	RMSE	MAE
DT	0.84	47.26	33.22

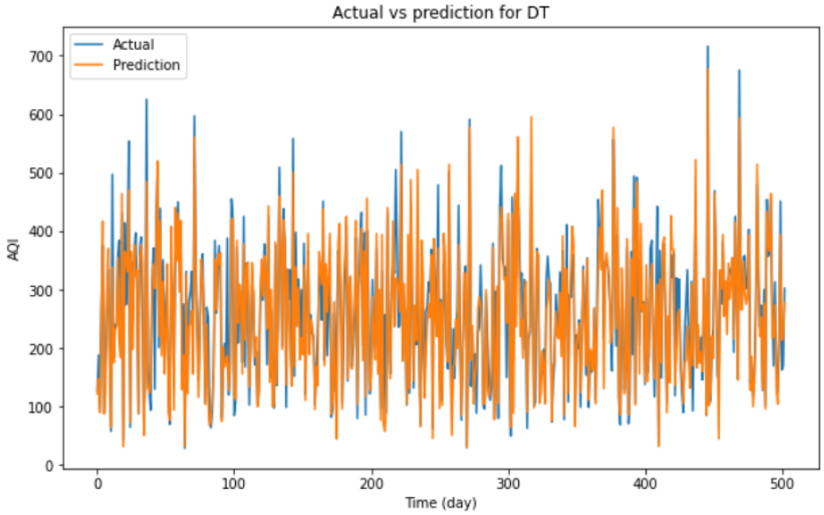

Figure 12: Actual vs Prediction DT

5.4.5 Implementation using GRU-LSTM

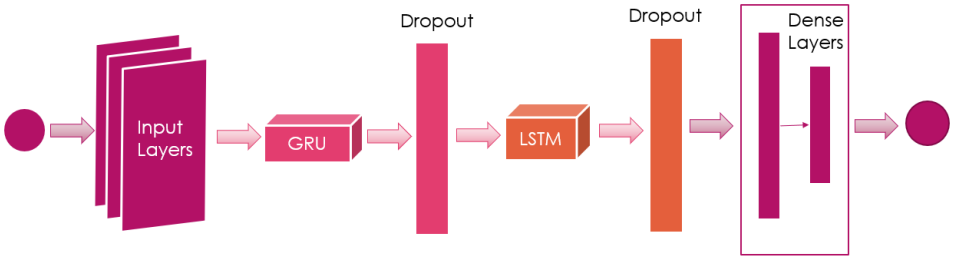

Figure 13: Proposed Model

We created two stage frameworks for forecasting data on air quality in this study. In the first phase, pre-processing input data, check for missing values and fill them in with the median function. For normalized input datasets, we developed minmax scaler

algorithms. Then we created a hybrid model that combined the GRU and LSTM deep learning models.

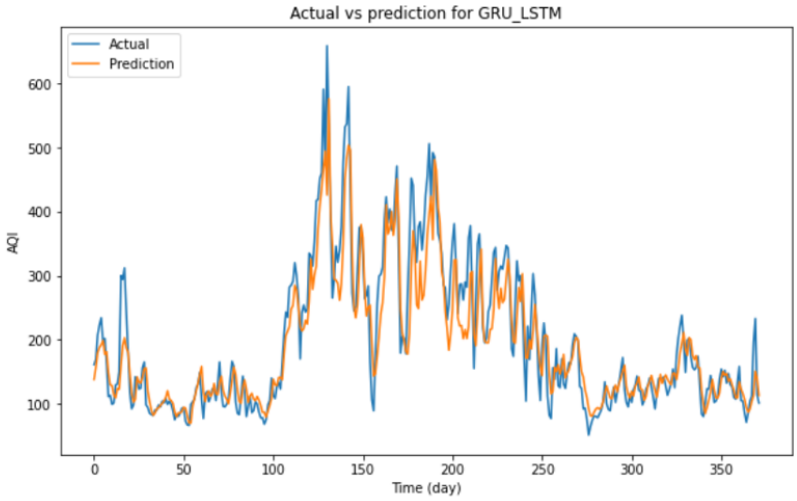

Figure 14: Actual vs Prediction GRU-LSTM

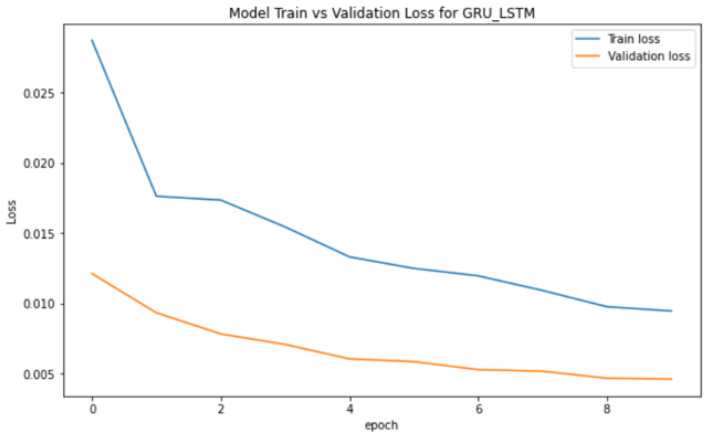

Figure 15: Train vs Validation Loss GRU-LSTM

For model preparation, we have data from the previous five years. If there is more data, the model can be improved. In our hybrid model, the GRU layer accepts input data and passes it to the LSTM layer, which uses a dropout function to avoid data overfitting. For optimal performance, two dense layers were applied, with the last dense layer providing one input value. The train dataset was then fitted into the model using the fit function. Run 50 epochs for assessment, with earlystopping used to avoid additional epochs.

We plot the train and validation loss graphs; we employed dropout in our model since train loss is bigger than validation loss. Although a portion of the features are set to zero, the validation process is strong. In the test set, all neurons are employed. This entire scenario is suitable for use in a prediction model.

Table 7: Performance Comparison

Model	R- square	RMSE	MAE
Proposed model	0.84	44.49	34.33

6 Model performance Comparison

The actual value has an approximate error as calculated in the above sheet. Thus, the forecast value can be taken into consideration for future forecasting of Air Quality Index (AQI). Furthermore, double and triple exponential smoothing methods can be used on the data to minimize errors and improve accuracy of the forecasts. Evaluation is important part after implementation model, we can determine best model. The analysis has been carried out using three measures. Finally, the computations were completed by comparing the actual value to the projected outcomes. The Root Mean Square Error (RSME), R-Square, and Mean Absolute Error (MAE) were used to evaluate the model performance in this investigation. While there have been many debates over whether to use a technique or a certain strategy to access model performance, there have also been many disagreements.

Table 8: Performance Comparison

Model	R- square	RMSE	MAE
Proposed model	0.84	44.49	34.33
LSTM	0.68	66.23	52.78
GRU	0.85	45.49	37.55
LR	0.85	45.23	34.82
DT	0.84	47.26	33.22

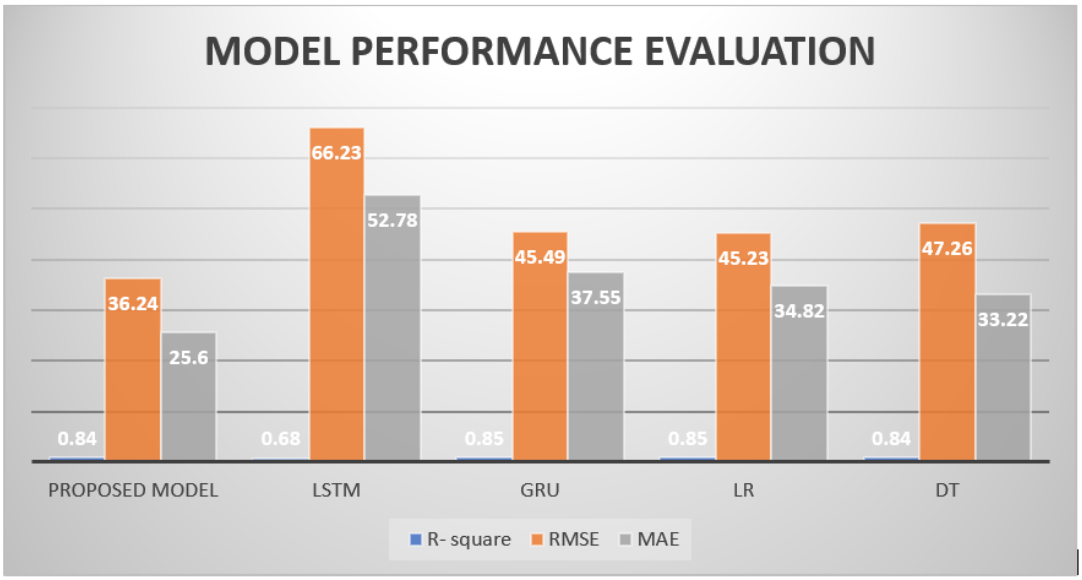

Figure 16: Performance Graph

By contrasting the accuracy measure in the table. It is proven that the proposed

model MAE 25.60 is less than other models, implying that the proposed model is good for projecting air quality data. However, the proposed R-square model has an R-square of 0.84 and LSTM has an R-square of 0.68, indicating that LSTM performs better in this case. In comparison to other models, the outcomes of some measures are lower.

7 Conclusion and Future Work

Combining deep learning with machine learning or by combining machine learning models the performance of the model increases. In this model the data is static. However, the government updates the data hourly. For better performance, real time data analysis using the cloud can produce better results. For further process, the AQI values predicted can be classified as per AQI and health standards and this can briefly talk about whether the air quality is hazardous or not. Based on this prediction it is useful for future researches as well as the measures to be taken.

The resultant pollution of ambient air, including highly harmful toxic substances, had a long-term effect on people living around these dumpsites. The suspended particulate matter (SPM), Hydrogen sulphide (H2S) and Oxides of Nitrogen (NOx) need urgent attention for remediation of the contaminated sites and new technology for the management of waste should be adapted.

In this study, a hybrid model combining GRU and LSTM was suggested to predict AQI in severely polluted cities. We can control pollution if we predict the AQI correctly. The MAE and RMSE parameters were used to compare several models such as DT, LR, LSTM, and GRU. The findings reveal that the suggested hybrid model generates fewer errors than the standalone models, demonstrating its superiority.

We can utilize the proposed approach to forecast data from other cities in the future. We can also determine the contaminated area and the cause of the pollution using prediction. Some pollutants are hazardous to human health, posing a major threat in the future.

References

Bose, R., D. R. R. S. and Sarddar (2020). Time series forecasting using double exponential smoothing for predicting the major ambient air pollutants, *In Information and communication technology for sustainable development,Springer,Singapore* pp. 603–613.

da Rocha, B. and de Sousa Junior, R. (2010). Identifying bank frauds using crisp-dm and decision trees., *International Journal of Computer Science and Information Technology* pp. 162–169.

Dåderman, A. and Rosander, S. (2018). Evaluating frameworks for implementing machine learning in signal processing: A comparative study of crisp-dm, semma and kdd.

Erskine, J.R., P. G. M. B. and Grimaila, M. (2010). Developing cyberspace data understanding: Using crisp-dm for host-based ids feature mining., *In Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research* pp. 1–4.

Gocheva-Ilieva, S.G., I. A. V. D. and Boyadzhiev, D. (2014). Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach., *Stochastic environmental research and risk assessment* **28(4)**: 1045–1060.

Harishkumar, K.S., Y. K. and Gad, I. (2020). Forecasting air pollution particulate matter (pm2. 5) using machine learning regression models., *Procedia Computer Science* pp. 2057–2066.

Janarthanan R, Partheeban P, S. K. E. P. (2021). A deep learning approach for prediction of air quality index in a metropolitan city., *Sustainable Cities and Society* pp. 1–67.

Kalgotra, P. and Sharda, R. (2016). Progression analysis of signals: Extending crisp-dm to stream analytics., *IEEE International Conference on Big Data (Big Data)* pp. 2880–2885.

Kumar, U. and Jain, V. (2010). Arima forecasting of ambient air pollutants (o 3, no, no 2 and co)., *Stochastic Environmental Research and Risk Assessment* **24(5)**: 751–760.

Liu, T., L. A. S. K. and Fung, J. (2018). Time series forecasting of air quality based on regional numerical modeling in hong kong., *Journal of Geophysical Research: Atmospheres* **123(8)**: 4175–4196.

Nigam, S., R. B. K. N. and Mhaisalkar, V. (2015). Air quality index-a comparative study for assessing the status of air quality., *Research Journal of Engineering and Technology* pp. 267–274.

Olah, C. (2015). Understanding lstm networks, 2015., *URL http://colah. github. io/posts/2015-08-Understanding-LSTMs* pp. 1–19.

Suleiman A, Tight MR, Q. A. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (pm10 and pm2. 5)., *Atmospheric Pollution Research.* pp. 134–44.

Yan, R., L. J. Y. J. S. W. N. M. and Li, F. (2021). Multi-hour and multi-site air quality index forecasting in beijing using cnn, lstm, cnn-lstm, and spatiotemporal clustering., *Expert Systems with Applications* pp. 114–513.

Yousefi, S., S. A. and Hadei, M. (2019). Applying epa's instruction to calculate air quality index (aqi) in tehran, *Journal of Air Pollution and Health* pp. 81–6.