National College of Ireland

Air Pollutant Concentration Prediction Using Deep Learning Techniques for Smart City: Beijing

MSc Research Project
MSc in Data analytics

Sankalp Saoji
Student ID: X17154171

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sankalp Saoji

Student ID: X17154171

Programme: MSc. In Data analytics **Year:** 1st year

Module: MSc. Research project

Supervisor: Dr. Catherine Mulwa

Submission Due Date: 18th April 2019

Project Title: Air Pollutant Concentration Prediction Using Deep Learning Techniques for Smart City: Beijing

Word Count: 8548 **Page Count** 25

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: .....

Date: 18th April 2019

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):	

Air Pollutant Concentration Prediction Using Deep Learning Techniques for Smart City: Beijing

Sankalp Saoji
X17154171

Abstract

Air pollution has been identified as a global threat and leads to various respiratory diseases and effect the environment adversely, with the air quality degradation in China hitting fatal levels, the prediction of air quality index to take measures for reducing air pollution is a critical task. This research focuses on predicting the concentration of a primary pollutant, responsible for respiratory diseases and a leading cause of death among citizens; the PM2.5 particles which are fatal due to their small size. The research focuses on using neural networks using long short term memory (LSTM), gated recurrent units (GRU) and comparing their performance with the traditionally implemented machine learning Models. The research focuses on understanding state of the art models for prediction of PM2.5 concentration in upcoming time frames. The idea here is to be able to lower the concentration of emissions using machine learning models when meteorological data leads to conditions which could maximise the effects of air pollution, hence slowly eradicating the problem from its roots.

Keywords: Long short term memory (LSTM), PM2.5, pollutant, Beijing, gated recurrent units (GRU), neural networks

1 Introduction

Air pollution had an adverse effect on the health of the people who are exposed to high amounts of the same. This problem has now exceeded local health issues and taken up a more global impact. This is due to an alarming rise in global temperatures credited to the pollution and its constituents which affect different aspects of our lives in various ways. This research aims to tackle the problem of air pollution by implementing a forecasting model which is capable of suggesting a downtime to lower industrial emissions and avoiding further damage to the already suffocating smart city: Beijing (Yang, 2018).

Beijing has been targeted for this research considering the fact that it has been declared the highest emitter of air pollution in the last decade with fatal suspended particulate matter (PM 2.5) concentration. PM2.5 is responsible for causing many respiratory diseases and long-term effects being a leading factor causing lung cancer, this can be seen in the figure 1 (Berkeley earth), The figure 1 shows that Beijing is most one of the most polluted cities in China and has worst impacts on health quality compare to all other cities in China.

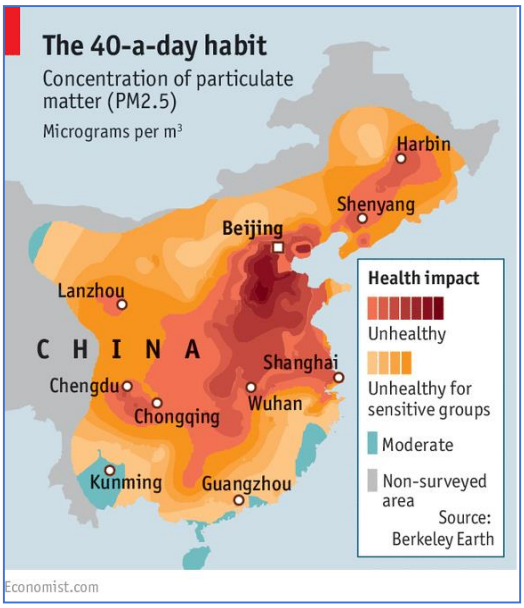

Figure 1: China air pollution

Air quality degradation has led to a major shift in the aura of industrialization with many industries going eco-friendly but there are still major players which cannot contribute due to massive losses any major transition could incur. These industries still cause pollution and this research can help them be a part of the change in a more ethical way without compromising on the money they make. The idea is to design a model that can predict when the air quality degradation could be accelerated and government bodies can make some rules to warns the industries to cut down emissions. This process would be done before hand to give the industries enough time to cope up with any deadlines or production updates.In a research conducted by (Lary, Lary, & Sattler, 2015), the use of machine learning to predict PM2.5 concentrations was focused on due to the high number of mortalities due to this air pollutant and hence, the era of more data oriented research in the field began to improve health and human life span.

The industries being targeted under this research are the highest producers of air pollution in Beijing (Hao, Wang, Shen, Li, & Hu, 2007), i.e. Thermal Power plants, Cement industries and Steel plants, as their requirements can be set before hand and targets can be achieved on a quarterly basis, making their constant contribution to air pollution a problem that can be resolved.

1.1 Research Question:

***RQ**: "How can we improve both health and air quality in Beijing city using multiple prediction model such as deep learning (convolutional neural networks, multi-layer perceptron, long short term memory, gated recurrent unit) with meteorological and air pollutants data to increase life expectancy?"*

Sub Research Question:

RQ: *"To what magnitude are the different meteorological features (temperature, pressure, humidity, wind speed) correlated with each other and with air pollutant concentration (PM2.5), measured using correlation matrix?"*

1.2 Research Objectives and Contributions

The main objectives surrounding this research are:

- Obj(1) - A critical review of literature on air pollution prediction (1999 - 2018).
- Obj(2) - To find the correlation between dependent variables in the dataset, especially meteorological factors causing rise in PM2.5 concentration levels.
- Obj(3) - Exploratory data analysis, feature selection and data preprocessing for implementing of deep learning models
- Obj(4a) - Implement, evaluate and results of MLP to determine PM2.5 concentrations with respect to patterns generated from data.
- Obj(4b) - Implement, evaluate and results of CNN to extract data and patterns which would determine the series of causation due to the factors in the data.
- Obj(4c) - Implement, evaluate and results of GRU to determine the effects of gated recurrent units on this research analysis.
- Obj(4d) - Implement, evaluate and results of LSTM to determine the effects of short memory analysis on the research.
- Obj(5) - Comparison of the all the above mentioned developed models.

There is a certain range of circumstances which lead to the maximization of the adverse effects of air pollution, at other times, the effects can be overseen. This leads the idea for this research as government bodies can advise industries to cut emissions during certain weather-based phenomena and be allowed to capitalize on better days, allowing a window for recovery to both, the industries and the atmosphere and help people to breath fresh air, lowering health problems and increasing life expectancy. This models if implemented accurately, could set the threshold for pollutant concentration and that can be lowered with improvement in air quality, hence allowing a timely change to the air quality of Beijing over a time period of a few years.

This research focuses on implementing the research question based on critically assessing the work done in the field, which has been reviewed in section 2. The next section is focused on showing the scientific methodology that has been followed for the research by showing the data mining methodology and the research flow from an architectural perspective. This is now succeeded by the implementation, evaluation and result section where each model has been run one by one and the results are displayed. To conclude the findings and understand the process outputs, a discussion section has been maintained which is followed by the future works.

2 Literature Review

2.1 Introduction to Literature

Air quality prediction has been a hot topic among data enthusiasts who want to bring a change to the global impact of air quality depreciation. This opens scope for various algorithms and scenario-based predictions to understand how air quality depreciates over time with minor and major changes in the climate and other factors which may or may not be natural. With a league of researchers focusing on time series-based prediction models to understand at what times the air quality may decline in near future, there is still a need of action. Not many industries could afford to shut down their production to save the environment as it would decimate their financial goals and hence, a more suitable approach that keeps sustainable development in mind is the need of the hour. Keeping this need ahead, recent developments have been made in the area of predicting time based attribute dependencies using Neural Networks (RNN, ANN and CNN).

This section focuses on the research done to improvise air quality prediction over time keeping meteorological aspects in mind. This starts with a brief explanation of how the Neural Network models work, along with a contrasting comparison between the models presented in a tabular manner.

2.2 Investigating the PM2.5

Air pollution is responsible for over 6.5 million premature deaths around the world annually (Priddle, 2016). This leads to a bunch of respiratory and cardiovascular diseases which lead to shortened human life expectancy numbers. Of all the major air pollutants, particulate matters (PM) with small diameters have claimed the highest number of victims.

PM2.5 has been declared as one of the most dangerous air pollutants (Castanas & Kampa, 2004) as it is not just a pollutant but due to its high internal impact range due to its small diameter of less than 2.5 micrometers, which means that this has the tendency to cause respiratory diseases by entering the lungs. Kampa also determined that particles sized above this are more likely filtered out due to our body's adaptability and evolution. This makes the research on PM2.5 concentration in the air a must.

2.3 Critiques of Techniques and Methods

2.3.1 Time Series Analysis

Prediction of air pollutant concentration, specially PM2.5 is a crucial step towards understanding how to eradicate the problem of air quality degradation in Beijing. In a research conducted by (Shen, 2012), it was concluded that in major cities such as Shanghai, Beijing and Chengdu (all in China), the amount of PM2.5 in the air goes beyond hazardous levels for over 50 days a year. This means that if these 50 days are predicted beforehand, precautionary measures can be taken to avoid any severe damage to the air quality. In this research, Jaiming has considered using a number of attributes and has taken the results in three different aspects, first being a trend prediction, where the forecasting is done for a certain time duration, value

prediction, where the amount of PM2.5 concentration is focused on and a hidden factor prediction that focuses on a prediction model for factors not directly derived by data.
Time series prediction has long been used in understanding the expectations using old statistics but with the increase in factors which have a high impact on the research, specially the hidden factors such as unexpected industrial rise, which is categorized as a foreseen aspect but the extent of which is unknown, makes the prediction vague.

2.3.2 Using Artificial Neural Networks and Fuzzy Logic

A unique approach was taken up by (Mishra, Goyal, & Upadhyay, 2015) where D.Mishra proposed the use of ANN and Fuzzy Logic to forecast PM2.5 concentrations in another Megacity: Delhi, India. The research focused on combining information purely from Industrial emission across the geographical state of Delhi, which could be considered as taking into account the wind speeds and atmospheric pressure indirectly. They have considered 2 highly measuring and operative areas, ITO (which is the hub of Industrial manufacturers in Delhi) and IGI (International Airport), where by the readings are taken for air traffic control. These are also geographically nearly at the 2 opposite poles of Delhi, making this research versatile enough to pass on as good.
Their approach proposes using ANN due to their ability to classify patterns (Gardner & Dorling, 1999) and using PM2.5 data for over 3 years, this becomes a self-training model which depended on the evaluation of the concentration on a daily basis. Another interesting approach adapted was the use of neuro-fuzzy logic (Mishra et al., 2015), where neural networks. Here, the approach focused on collectively combining the data and deriving a fuzzy relationship between all the aspects before analyzing them in neural networks. This means that suppose a certain time frame in the research did a pattern match among the leading factors, it wasn't necessary for the exact things to happen again for the phenomena of 'Haze' to re-occur, hence, a fuzzy algorithm was placed to help cover near matches and forecast for the next few days based on scenarios considering any approach to lower the concentrations was taken.

2.3.3 Using Multi-Source Data

A critical aspect to any good research lies in the number of ways one looks at it, with the best outcomes coming from researches implementing all possible methods to meet the desired outcome. One such approach was used by (Ni, Huang, & Du, 2017), where the focus relied on simple analysis for PM2.5 concentration based on data from various sources. The idea here was to cover maximum aspects and still be able to get an overall prediction on air pollution. The research was conducted simply by using data from air quality measure and internet blogs and feeds using simple web scrapping algorithms, stored into a data base where PM2.5 particles were evaluated separately and all other factors separately, the information was then run through multivariate statistical analysis and neural network approaches, the outcomes were based a follow up using timeseries, hence, there was established correlation and a forecast available for the PM2.5 concentration.

2.3.4 Using Recurrent Neural Networks

RNN is used to model time series in data, where there is a loop which works on a time based manner. This means that (Athira, Geetha, Vinayakumar, & Soman, 2018) once a pattern is recognized, temporal tasks are executed at the time and this information helps in identifying tasks that could come up in near future (LeCun, Bengio, & Hinton, 2015). RNN has been used to be able to identify any hidden sequences in the vector formation of the attributes. This means that if any patterns are identified, they will be repeatedly checked within the same dataset as well. Each node in the neural network at RNN is embedded with a characteristic time variant value which is activated in real time (Tsai, Zeng, & Chang, 2018) while performing the operation. RNN has the ability to handle long term dependencies, which makes it an ideal choice for this research, but it has been unreliable for the same if the time intervals for the data to be referenced are increased. This means that once the time intervals go up, the reliance on RNN goes down.

2.3.5 Using Convolutional Neural Networks

Another approach to predict the concentration of PM2.5 in near future was proposed by (Huang & Kuo, 2018), where the author applied the usage of CNN with LSTM to execute the research. This research is contributory to the current research in two ways:

- First, the research shows the impact of LSTM on an CNN in terms of improvement.
- Second, it has focused on city specific details and on the major sources of pollution in the cities.

As per research conducted by Chiou-Jye Huang, 22% of the pollution is caused by vehicles, 17% by coal combustion, 16% due to industrial discharge and over 25% due to factors like pollution being carried by other countries. This means that over 55% of the sources can be controlled and manipulated in a way to nullify the harmful effects of air pollution if precautionary measures are taken.

For this research, the algorithm generated a loss value every time a pattern was detected among the attributes. This would generate a new attribute index, which would be filled once either a new pattern is detected or a previous one is added. This would eventually lead to an overfit due to a high number of patterns. To fix this, the concept of regularization is introduced (Xie, Wang, Wei, Wang, & Tian, 2016).

This research concluded that CNN is the right choice for PM2.5 concentration prediction but it still lacks some dynamically adaptability issues which if can be overcome, would make Neural networks the best option to tackle the problem of air quality prediction.

2.3.6 Using Gated Recurrent Units

Gated recurrent unit (GRU) is a solution to solve the vanishing gradient problem in RNN (Dey & Salem, 2017). GRU is used as a theoretical mid-level fixture between RNN and LSTM. This implies that it has an update and a reset gate, hence, they can alter the information output form the RNN process, unlike in LSTM where the input can also be hampered using a forget gate. Due to this difference from LSTM, GRU has a final memory recording which is made at time steps relevant to the research as it does not have a forget gate, hence, stores the last calls made.

2.3.7 Using Long Short Term Memory

Long short-term memory (LSTM) networks have been designed for the purpose of classification, processing and predicting based on time series data. These serve as an improvisation over the RNN in terms of time-based dependencies (Li et al., 2017). This means that LSTM has the capabilities to handle the errors in RNN occurring due to long time intervals in data collection. LSTM has the ability to read, write and delete information, which makes it work somewhat as a processor which can cover up for missing inputs in RNN and being prepared for any drastic changes in the input. LSTM is easy to implement and can fix problems of dependencies, which in the context of this research is pivotal as the data has to be updated with the latest statistics and any comparison to an old input would be irrelevant as the sources of pollution and the severity of impact changes rapidly.

2.3.8 Using Support Vector Regression

Support vector machines are used in machine learning comprehensively for the purpose of analyzing data which would imply categorical values. In a subtle modification based on this research, since most of the variables being included are all numerical, we modify SVM to SVR (Support Vector Regression) (Liu, Binaykia, Chang, Tiwari, & Tsao, 2017).

The SVR model is implemented when the business requirements dictate the need of a specific function where only non-linear data needs to be mapped onto a high dimensional feature space. In forecasting, a hybridized portion of SVM (Fan & Chen, 2006) is maintained to be the best solution to contradict the problems faced which are seen in the form of SVM in its simplest form being difficult to evaluate, hence, somewhat making the research ambiguous.

2.4 Investigation of Existing Pitfalls, Issues and Challenges

On a general consideration, this research is effected by various factors which are nearly impossible to account for. For the scope of this research, we are considering Beijing, but many other cities could have different reasons for PM2.5 concentration. There is evidence to volcanic eruptions (Mass & Portman, 1989) leading to major climate change and air density changes. Such factors could lead to unexpected fault values in the data which need to be accounted for and either removed or placed as special circumstances. Some problems faced by research methods are shown below:

2.5 A Comparison of Reviewed Methods and Techniques

The table 1 below shows the comparison of algorithms and/or techniques used by different authors with their advantages and critics.

Table 1: Comparison of methods and techniques

Algorithms, methods and techniques used	Pros	Cons	Authors
Time series analysis	Takes into account a lot of factors, uses almost all the fields in the data set it implies.	Fails to elaborate on the hidden factors used by Jaiming, leaves a lot to the imagination of the reader, could fail unless very frequently modified.	(Shen, 2012)
Use Artificial Neural Networks and Fuzzy Logic	Neural networks are good at establishing patterns, fuzzy logic would help broaden the range of identifying many different possibilities.	Fuzzy logic increases the processing time and load, ANN has its independencies due to un-verified backend method, which would always have to be considered for something that needs daily algorithmic runs such as this research.	(Mishra et al., 2015)
Use Multi-source data	Multiple data sources meant a broader look and helped to train the model well as neural networks trains best with more data.	One source was 'blogs', where people do sometimes to garner viewers change statistics and fabricate some level of information, hence, not extremely reliable.	(Ni et al., 2017)
Use of Recurrent Neural Networks	RNN for the example of time series has the capability of detecting changes / patterns overtime.	It can handle long term dependencies, but if the time interval for data reference is increased it leads to exploding gradient problems.	(Athira et al., 2018)**,** (LeCun et al., 2015)**,** (Tsai et al., 2018)
Use of Convolutional Neural Networks	CNN can extract features automatically for any new data / task provided that all the local features are correlated with each other.	Every time a pattern is detected a loss value is generated which means there would be an over fit in future. CNN requires high computational cost and requires lot of training data to gain high accuracy.	(Huang & Kuo, 2018)**,** (Xie et al., 2016)
Use of Long Short Term Memory	Overcomes the long-time	Could change the results if historic data needs to be considered, hence, on application	(Li et al., 2017)

	dependencies in RNN	of LSTM, we cannot completely disregard the historic data till confirmed.	
Use of Support Vector Regression	Covers up for the lacking in SVM and allows numeric values.	Choosing the appropriate hyper parameters and measures is a hefty task. Choosing the right kernel for the job is also a dependent task.	(Liu et al., 2017), (Fan & Chen, 2006)

2.6 Conclusion

This section helped us narrow down the research methods implemented by other authors with respect to our research. To show how the methods of Support Vector Regression, Long Short Term Memory, Recurrent Neural Networks, Gated Recurrent Units and Time Series perform in overall scenarios, the assessment has been done which explains their strengths and weaknesses, which prepares us for the research implementation.

3 Scientific Methodology Approach and Project Design

3.1 Introduction

This research implies a fully functional approach to tackling the problem of air quality degradation in Beijing using meteorological data as the basis of the research. To explain how this is done, we need to break down the methodology into 2 major fields, the data mining field where we explain the textbook approach being implemented to execute this research, and the architecture kept in mind for acceptable execution of the research.

This section is dedicated towards understanding this process and paving the path for implementing the research in order to get the best outcomes from the data in terms of stake holder value as well as from a developer's perspective who would want to approach the research from scratch.

3.2 The Data Mining Model

The data mining model best suited for a research that relies on constantly changing business needs and stake holder requirements in terms of the pollution levels constantly changing due to ever increasing sources has been chosen to be CRISP DM (Wirth & Hipp, 2010). The diagram can be seen in figure 2.

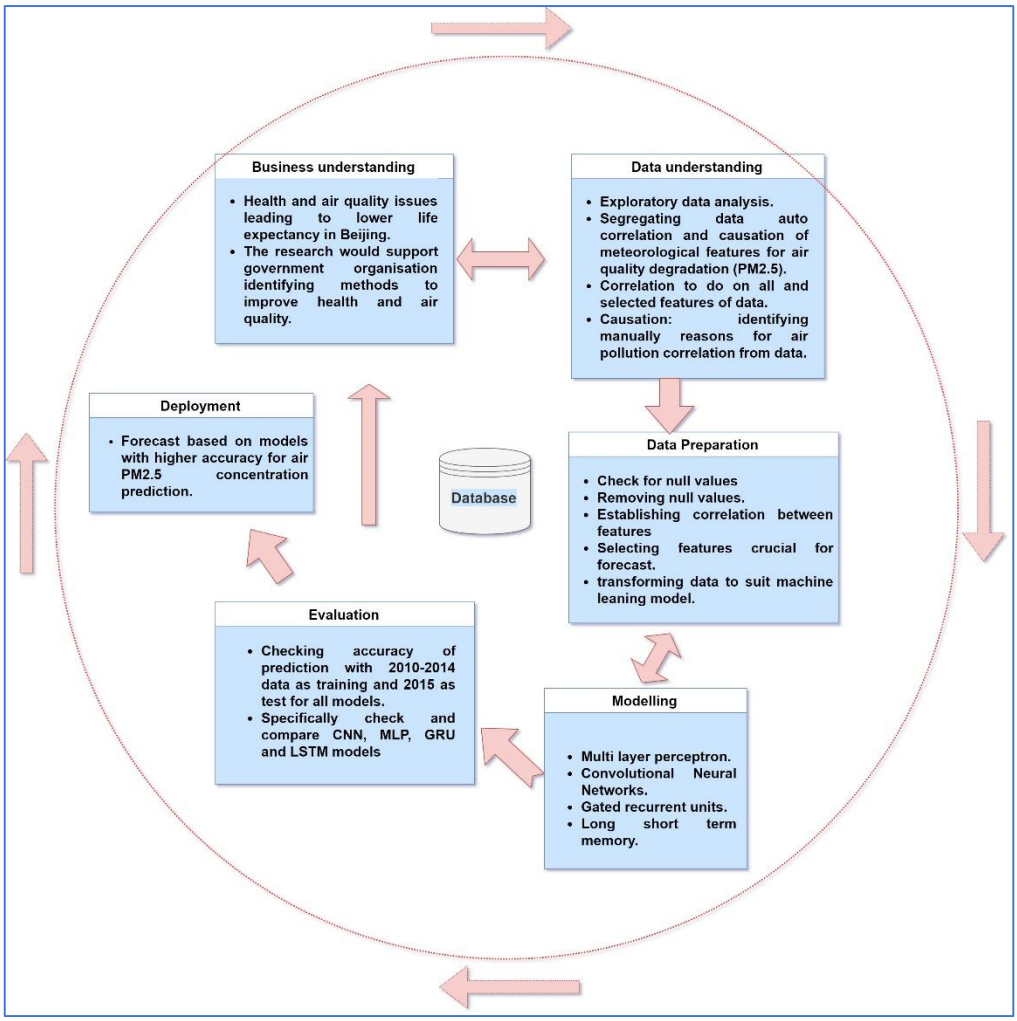

Figure 2: Scientific methodology approach

Once the research requirements are set on improving air quality degradation prediction in Beijing, we know that the stakes are high as this leads to improving life expectancy in the region of interest. This brings us to the second phase where we do a thorough research on the data to identify what variables and attributes are relevant and what are not. This phase involves doing an exploratory data analysis (EDA) and also understanding how many factors are just correlated to the high PM2.5 concentration and not a causation factor. This further involves narrowing down the data size from granularity perspective, where it initially covered hourly data but will now cover data on daily basis. This step doesn't only narrow down the complexity but also makes the data smaller, which reduced the processor workload overall. Once this process is completed, we come down to editing the data and making it more research friendly and interface prepared. Here, we remove values which are problematic such as nulls and replace any missing values with either means or eradicate them altogether. For this research, both these measures have been implemented in different parameters and the final features for forecast have been determined.

The models being implemented in this research are mentioned after this and can be seen clearly in figure 2. This is followed by an in detail evaluation that focuses on predicting the PM2.5 concentration for different durations based on the data collected form 2010 to 2014. This would allow the research to validate the accuracy of the forecast methods and see which models

perform best, especially evaluating the performance of LSTM with CNN, MLP, GRU as LSTM will be using a smaller chunk of data based on its forget function.

3.3 Project Design

One of the key principles of any research is its value and re-enact ability to the stake holder. To make this process easy, this research has followed a simple 2 tier architecture which can be seen in figure 3.

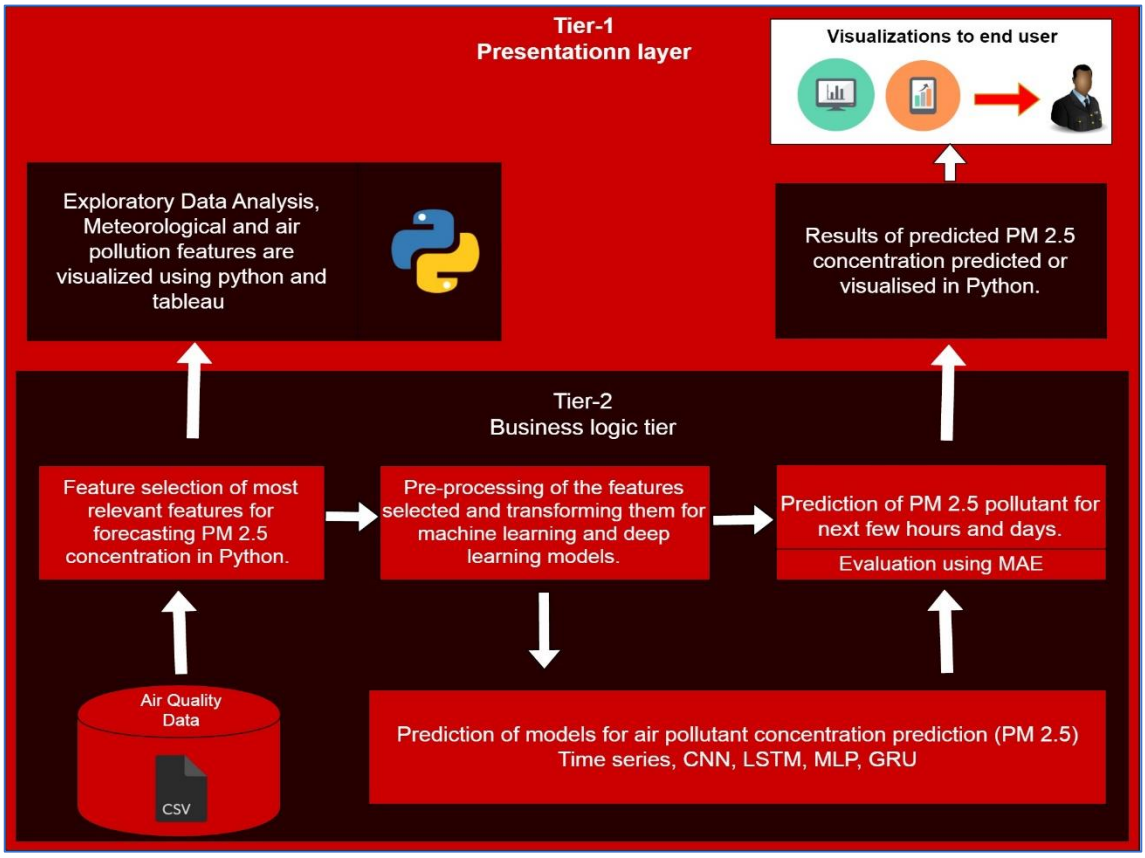

Figure 3: Design architecture

Tier1: Presentation Layer

This layer in the architecture represents the client site approach to this research using a presentation tier. Allowing the users to select what aspects of the data they want to visualize using Tableau and Python. This layer is connected to the business logic tier where all the logical permutations and modelling are implemented.

Tier 2: Business Logic Layer

This layer forms the backbone of this research. It carefully places all the implementation-based aspects of the research under one field. As seen in the figure, first the data is taken from a simple file and pre-processing is initiated, this leads us to the logic and forecasting bit where the prediction based models are run. This layer prepares the research for the output which can then be seen by the users in the presentation tier.

3.4 Conclusion

This section confirmed that a scientific methodology implemented would be CRISP - DM and there would be a two tier architecture implementation. This section helped the research by dividing the objectives defined in section 1 into presentation based and business based, which will now assist the research visually and using deep learning techniques implemented in the following section.

4 Implementation, Evaluation and Results of Air Pollutant Concentration Prediction Models

4.1 Introduction

This section focuses on the implementation aspect of the research by focusing on the innovation aspects followed by the dataset information. This is followed by a detailed exploratory data analysis process that helps us feed the data to the machine for implementing machine learning and deep learning models. The model execution has been implemented and represented in a simple format where all models have been classified well enough to obtain and visualize all results individually.

4.2 Innovation

In practicality, the project is focused on understanding that air pollution is a major problem in mega cities such as Beijing. The innovative angle constituted in this research is that we are focusing on the concentration of PM2.5 levels in the air pollution existing. This is being done to highlight the fact that life expectancy is shortened due to components of air pollution that are less than 2.5 micrometres in size, which settle within the respiratory system of humans leading to diseases. This project also elaborates on the issues faced by RNN as a machine learning model. One of the biggest problems with RNN modelling is that it is considered as a model which implies the use of hidden states generated by previous inputs and current input. Another issue in RNN is the vanishing and gradient exploding issue, which was detectable due to two symptoms; instability in the model causing large changes in output even after small updates and the model not being able to catch up to the training capabilities.

All this could lead to extremely mismatched forecast and hence cause disruption in the environment of the cities using the research. This undesirable scenario can be prevented using GRU and LSTM, which enables the RNN models to run over longer input ranges (in terms of time). Another reason for this is that LSTM has the ability to manipulate the previous inputs and outputs, hence, in any circumstances leading to changes in the data due to massive meteorological phenomena can be removed to make the model more stable.

4.3 Dataset Information

The data has been extracted from the UCI repository where the focus has been tilted towards using PM2.5 concentration levels. This contains information from 2010 to 2014 for air pollution concentration over Beijing. The information in the data is strictly meteorological and is updated every hour. The fields covered represent PM2.5 concentration by the hour and

focuses on understanding when the concentration is highest or lowest. The meteorological factors aforementioned are dew point (DEWP), temperature (TEMP), air pressure (PRES), combined wind direction (cbwd), cumulative wind speeds (lws), cumulative hours of snow (ls) and cumulative hours of rain (lr).
The data is presented in a tabular manner and even a visual glimpse of the table could help determine whether the factors collaboratively add on to the concentration of PM2.5 level.

4.4 Feature Selection Process

Feature selection has been implemented manually to a certain extent to focus on logicality of the research. This is driven by the factors such as air quality degradation due to PM2.5 concentration levels only. Hence, not including the chemical variants of air pollution which minutely contribute to the level of PM2.5 concentration.
Another feature selection method implemented here is to implement a simple correlation matrix using the relevant features left after the manual editing. The process of establishing a correlation matrix using heatmap imagery is shown in figure 4.

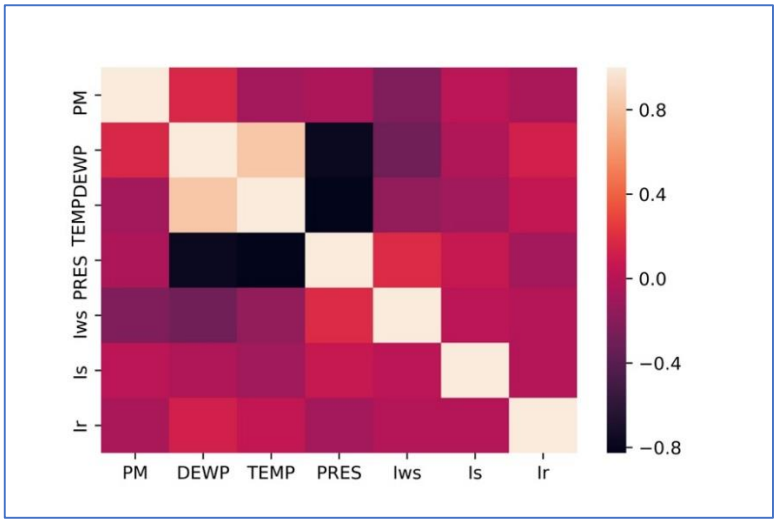

Figure 4: Heatmap depicting all features

As depicted, the first step was to establish a correlation matrix between all the variables in the data for feature understanding. This means that there is a low correlation among the variables is all the pollutants are considered. The fields shown in the above image are taken below the manual removal of less important features in done. Feature selection has been done and the variables constituting this research are now shown in figure 5. As seen, there is a slight correlation between the PM2.5 concentration and factors such as dew point, temperature and other features, therefore we can see that there is no causality of any factor on PM2.5 so will move further for the implementation of machine learning and deep learning models for air pollutants concentration prediction.

Out[95]:

	PM	DEWP	TEMP	PRES	Iws	Is	Ir
PM	1	0.2	-0.09	-0.05	-0.2	0.02	-0.05
DEWP	0.2	1	0.8	-0.8	-0.3	-0.03	0.1
TEMP	-0.09	0.8	1	-0.8	-0.2	-0.09	0.05
PRES	-0.05	-0.8	-0.8	1	0.2	0.07	-0.08
Iws	-0.2	-0.3	-0.2	0.2	1	0.02	-0.01
Is	0.02	-0.03	-0.09	0.07	0.02	1	-0.01
Ir	-0.05	0.1	0.05	-0.08	-0.01	-0.01	1

Figure 5: Correlation with selected features

The features such as Dew point and temperature have a strong correlation with each other, while pressure has strong inverse correlation with temperature and dew point. This is evident of a research scope that was identified and highlighted during this process.

4.5 Data Pre-Processing

Data pre-processing is an important aspect of any data analytic research. In this research, the focus has been to use the most relevant aspect of the data and implement the machine learning techniques. Data pre-processing is critical to this research as we are using Neural Networks for training the data. The step one to make the data more friendly to machine learning models was to eradicate columns such as years, month, day and hours. This has been done to reduce the computation time during modelling and also reduce the number of variables to be assessed and ease down the process of correlation matrix. This segregation of the three fields into one done using the datatime function has been shown in figure 6.

Out[36]:

sr.number	PM	DEWP	TEMP	PRES	Iws	Is	Ir	Date
0	98.613215	-21	-11.0	1021.0	1.79	0	0	2010-01-01 00:00:00
1	98.613215	-21	-12.0	1020.0	4.92	0	0	2010-01-01 01:00:00
2	98.613215	-21	-11.0	1019.0	6.71	0	0	2010-01-01 02:00:00
3	98.613215	-21	-14.0	1019.0	9.84	0	0	2010-01-01 03:00:00
4	98.613215	-20	-12.0	1018.0	12.97	0	0	2010-01-01 04:00:00

Figure 6: Dataset after pre processing

The second step towards pre-processing the data was to perform an Exploratory Data analysis (EDA) on the data. This step covers the critical part of checking the data description using the data.describe function in python. Once executed, we obtain count, mean, standard deviations and minimum of the values in the object. The output file for this is seen in figure 7.

Out[24]:

	pm2.5	DEWP	TEMP	PRES	Iws	Is	Ir
count	41757.000000	43824.000000	43824.000000	43824.000000	43824.000000	43824.000000	43824.000000
mean	98.613215	1.817246	12.448521	1016.447654	23.889140	0.052734	0.194916
std	92.050387	14.433440	12.198613	10.268698	50.010635	0.760375	1.415867
min	0.000000	-40.000000	-19.000000	991.000000	0.450000	0.000000	0.000000
25%	29.000000	-10.000000	2.000000	1008.000000	1.790000	0.000000	0.000000
50%	72.000000	2.000000	14.000000	1016.000000	5.370000	0.000000	0.000000
75%	137.000000	15.000000	23.000000	1025.000000	21.910000	0.000000	0.000000
max	994.000000	28.000000	42.000000	1046.000000	585.600000	27.000000	36.000000

Figure 7: Dataset description

The standard values have been depicted in the figure and help us understand the impact of null values on the data. On the first look, we can clearly see that the count values of all fields are varying, which means that there are many null values which need to be handled.

Next step in the process is to check all the null values in the data. These values are seen as nan, as per python syntax and since there are too many of these. To check for all the null values together, the is.null function in Python has been used. This shows all the rows with any null values. The next options here were to either add average values of the columns to the nulls or remove them altogether. The reason for this is that on a regular skim through the data, some null values were coming in due to zero as a value. This means that a better option is to delete all null values. This has been done using the drop.na function, which removes any rows with nulls.

The next process undertaken was to check for outliers in the data, which would help us determine whether there are too many disrupted readings in the data. The boxplot for the data has been shown in figure 8. As we can see, there are some outliers in the data which could be considered during training the models.

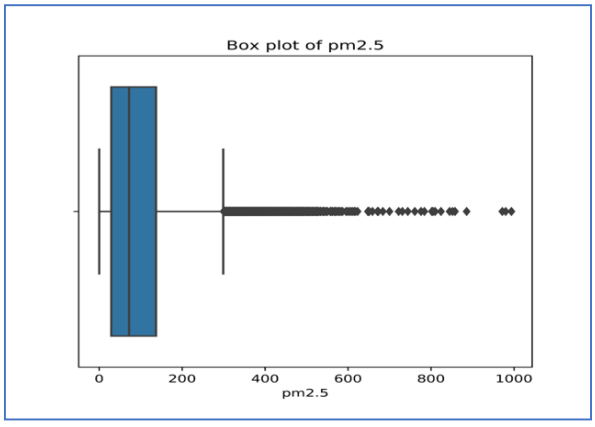

Figure 8: Box plot for PM2.5 concentration

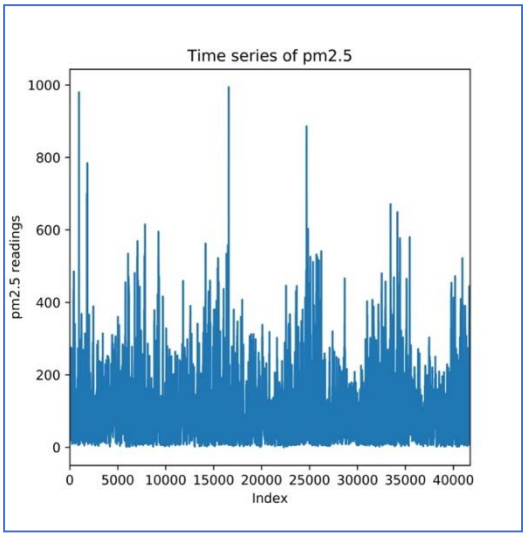

Figure 9: Overall distribution of PM2.5

Next up is a quick analysis of the data in terms of linearity. This is done to ensure that the data is not overly consistent, which may lead to an overfit once the machine learning models are run. The overall linearity of this has been shown in figure 9. This depicts that when all the data is considered there is no linearity.
Figure 10 shows the linearity check for yearly data. The image shows data distribution throughout the year for 2010. It can be seen that there is no linear data distribution. The plot in figure 11 shows a similar analysis done for the month of January 2010.

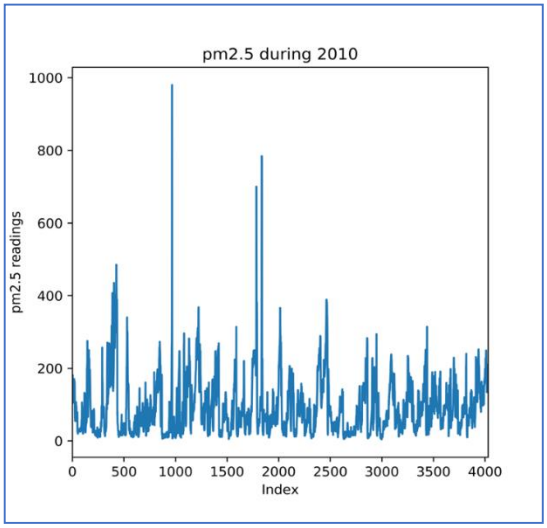

Figure 10: Distribution of PM2.5 for one year

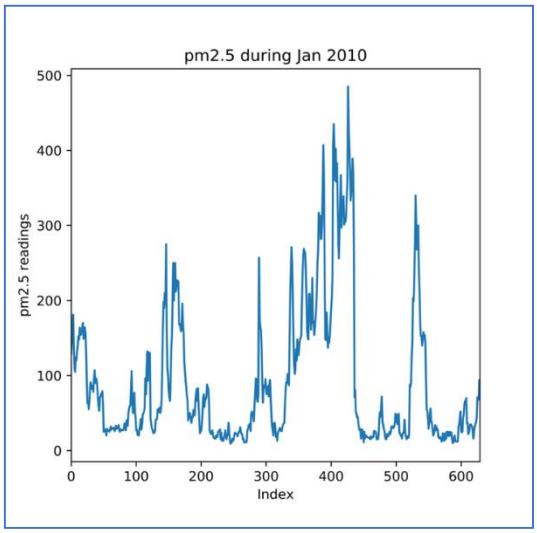

Figure 11: Distribution of PM2.5 for one month

This also depicts no linearity in the data. This analysis based on the linearity check of the data assures that a granular level training of the models is possible, which would allow for higher accuracy achievement once the models are trained.
The data used for this research project is a multivariate data and for time series modelling first step is to check for Granger-causal relationship (Farhani & Ozturk, 2015), which is a no in this case, then the second step is to check if there is any linear relationship assumed so from Figure 9; we can see that there is no linear relationship in our data so, this research further moves to implement neural networks.
Scaling of data is done for implementation of neural network models using the min-max scalar function (sk-learn python) to remove ambiguities in data and provide a normalized form of the data. Once the process is done, we split the data into test and training sets. The training data is 80% data that is from January 2010 to December 2013 for training the model and implement that to check the accuracy on the 20% testing dataset that is the data for year 2014. We generated regressors for training and testing dataset, we created the 2-D array of regressors and 1-D array of target variable from the original 1-D array of our target variable which is PM2.5. The pattern of research that has been followed here is to use the data for 7 days and implement that to check the accuracy on the 8th day. This research is done with respect to time series model. This gives us a window of operation which is not too narrow as to not let the model learn but just get an overfit, nor too wide which would now allow us to understand the patterns in meteorological events. Figure 12 shows how the data has been scaled for test (red) and train (Blue) sets.

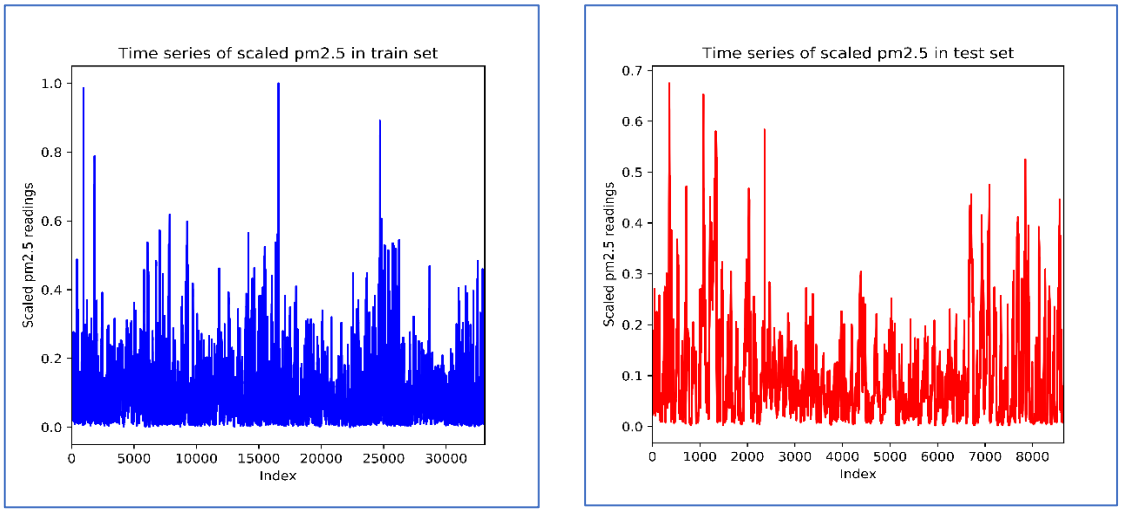

Figure 12: Scaled train and test data

4.6 Understanding Hidden Patterns

While performing the initial data analysis, we came across a shift in the PM2.5 concentration based on the quarter of the years. This has been shown in figure 13.

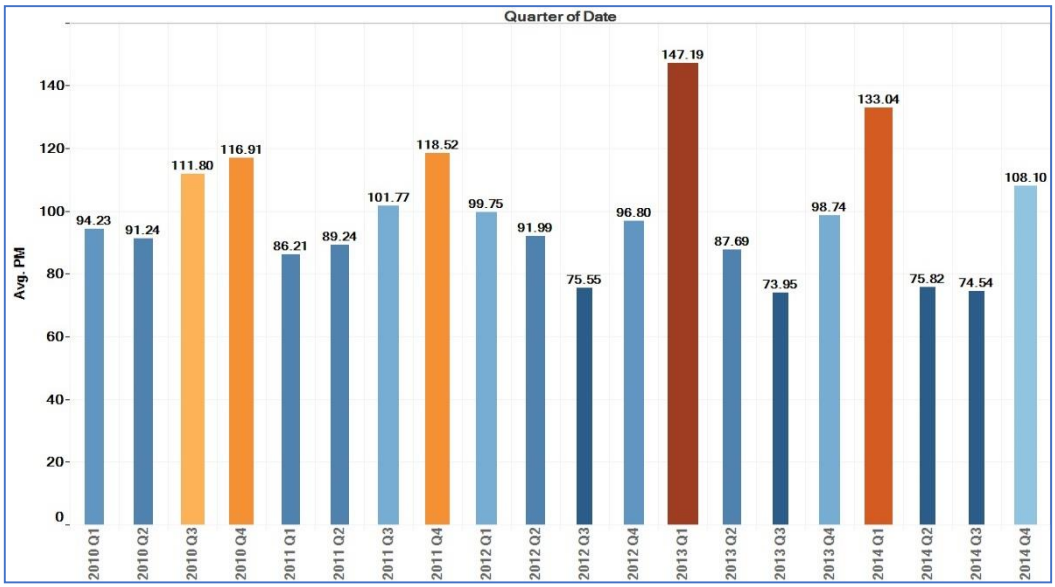

Figure 13: Quarterly distribution of PM2.5

Here, data on highest PM2.5 concentration has been taken and during the years 2010 and 2011, the highest concentration has been observed during the 4th quarter of the year, which is the time when winter is at its peak and factors such as wind speed and low temperatures are present.
During 2013, the peak was observed in quarter 1, completely skipping 2012. This trend doesn't show a lack of pollution during the time, but shows more of the pollution depending on temperature as a variant as in 2013, the first major shift in temperature cycle change around the globe was observed as winter times somewhat shifted form the peak being in December to January and February. To confirm this hypothesis, we looked into 2014, and the results were astonishingly similar with the first quarter peaking in PM2.5 concentration.

4.7 Software, Libraries and Formulae Used

The implementation of this research has been done using Python on Jupyter Notebook. Tableau is used for data visualization due to its adaptability towards estimating correlations in the data. The libraries used in Python 3.7 (64-bit) are Pandas, NumPy, Matplot Lib, Seaborn, Sklearn and Keras.

To evaluate the results, we used mean absolute error (MAE) calculations, the reason for doing so is that due to outliers in the data being used for the research, there is a high probability of deviations due to fluctuating results which is undesirable (Cort & Kenji, 2005). This error calculation gives precise results with outliers as compared to the squared deviations.

$$\boldsymbol{Mean\ absolute\ error} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} |\boldsymbol{y_j} - \boldsymbol{y}|$$

Where: **n** = the number of errors,

$|\mathbf{y_j} - \mathbf{y}|$ = the absolute errors

4.8 Implementation, Evaluation and Result of Deep Learning Models for PM2.5 Concentration Prediction

4.8.1 Implementation, Evaluation and Results of Multi-Layer Perceptron

Implementation: MLP comprises of a minimum of three layers of nodes; input, hidden and the output layer. It is a type of deep ANN. The input layer has multiple input layers, and this continues with the hidden layers as well. As seen in figure 14(a), the initial forward feed is set to take the input of 7 time steps. Now, the first layer of inputs taken has 32 feed forward elements, which generates 256 parameters. The next layer then implies another 16 elements which cumulates the previous time steps as well and leads to 528 parameters. This leads to the third time step utilizing less time steps as previous ones have been occupied and the final parameters are calculated to 272.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 7)	0
dense_1 (Dense)	(None, 32)	256
dense_2 (Dense)	(None, 16)	528
dense_3 (Dense)	(None, 16)	272
dropout_1 (Dropout)	(None, 16)	0
dense_4 (Dense)	(None, 1)	17

Total params: 1,073
Trainable params: 1,073
Non-trainable params: 0

Figure 14(a): Hyper parameters for MLP

Evaluation: The regularization method implemented is dropout, which implies the reduction of overfit due to insignificant hidden layers being dropped out. The network weights are

optimized using “adam” optimizer due to its capability to use variating learning weights for each weight individually and using that to segregate updates in the training process.
As seen in figure 14(b), the evaluation of predicted behaviour using MLP is not very different from the actual reading.

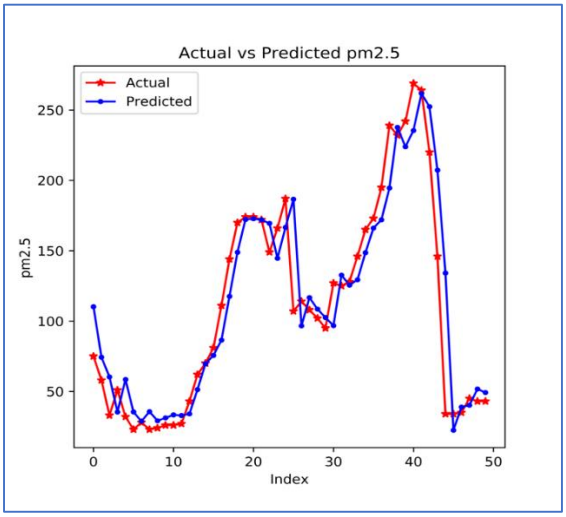

Figure 14(b): Actual vs. predicted graph for MLP

Result: The mean absolute error for MLP model is 13.272 for the test data.

4.8.2 Implementation, Evaluation and Results of Convolutional Neural Networks

Implementation: CNN is a deep learning technique commonly applied to predict the occurrence of events using multiple variables. It is used to extract features from data automatically. Figure 15(a) shows the layers involved in CNN implementation. The input shape represents the number of samples, number of time steps and the number of features per time step considered.
The zero padding layer is applied to ensure that any downsizing that may be applied to the convolution layer is not effective on the output sequences. Conv_1d represents the number of features in the output and the length of one dimensional window. The average pooling layer in this case takes moving averages on rolling window of three time units. The flatten layer reshaped the provided input from samples and timesteps to number of features in timesteps, hence reducing the execution complexity of the layer.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 7, 1)	0
zero_padding1d_1 (ZeroPaddin	(None, 9, 1)	0
conv1d_1 (Conv1D)	(None, 7, 64)	256
conv1d_2 (Conv1D)	(None, 5, 32)	6176
average_pooling1d_1 (Average	(None, 3, 32)	0
flatten_1 (Flatten)	(None, 96)	0
dropout_1 (Dropout)	(None, 96)	0
dense_3 (Dense)	(None, 1)	97

Total params: 6,529
Trainable params: 6,529
Non-trainable params: 0

Figure 15(a): Hyper parameters for CNN

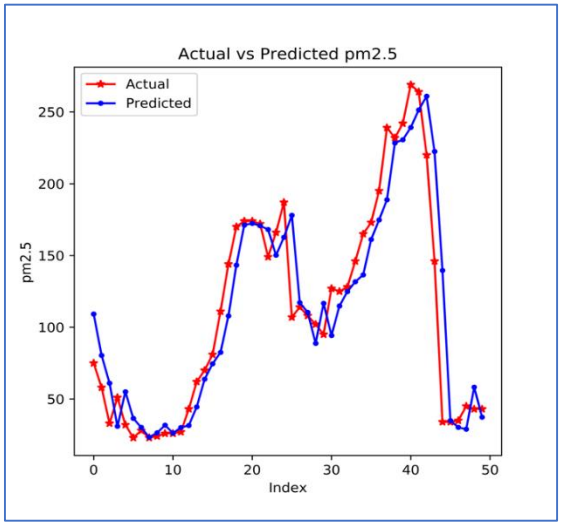

Figure 15(b): Actual vs. predicted graph for CNN

Evaluation: The red line in the graph depicts actual behaviour of the model using pre-processed data. It can be determined that it is not significantly different from the predicted one. This indicated no change going up the pm concentration scale on an average but slight changes going along the number of data considered.

Result: The mean absolute error observed for CNN is 13.1868 for the test data.

4.8.3 Implementation, Evaluation and Results of Gated Recurrent Units

Implementation: GRU is an improved version of RNN with their update and reset date. They can be called on as a middle step between the RNN and LSTM to evaluate whether the process is sufficient as per the data science problem at hand.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 7, 1)	0
gru_1 (GRU)	(None, 7, 64)	12672
gru_2 (GRU)	(None, 32)	9312
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 1)	33

Total params: 22,017
Trainable params: 22,017
Non-trainable params: 0

Figure 16(a): Hyper parameters for GRU

The input layer involves the number of samples, timesteps and number of features considered per time step. GRU timesteps considered are 7. The layers of the GRU model are stacked sequentially. The first layer is seen returning the outputs form each of the timesteps. This output is now used as the input to the next GRU. There are 64 neurons in the timestep. The regularization process here is same as in section 4.7.1, the dropout regularization.

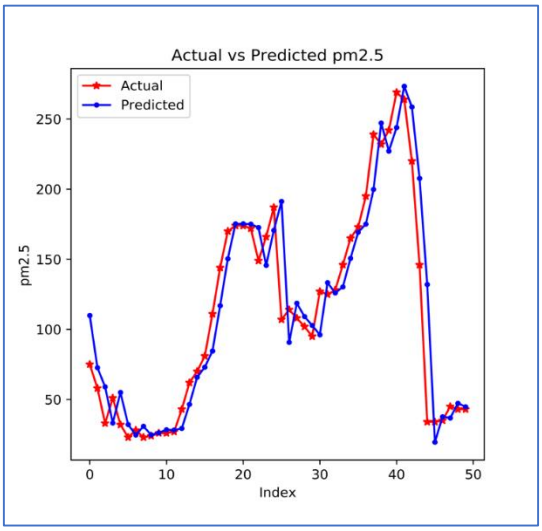

Figure 16(b): Actual vs. predicted graph for GRU

Evaluation: This model initially predicted the pm2.5 concentration high but the model slowly caught up and there aren't any significant differences from the actual readings.
Result: The mean absolute error (MAE) observed is 12.3795, which is again lesser than the previous 2. Indicating improvement over the previous implemented models.

4.8.4 Implementation, Evaluation and Results of Long Short Term Memory

Implementation: LSTM is an upgraded version of RNN which implies the use of three gates with the ability to read, write and update any data depending on its requirements by the research using a memory based functionality. This makes LSTM a good option for research involving memory based operations.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 7, 1)              0
_________________________________________________________________
lstm_3 (LSTM)                (None, 7, 64)             16896
_________________________________________________________________
lstm_4 (LSTM)                (None, 32)                12416
_________________________________________________________________
dropout_2 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 33
=================================================================
Total params: 29,345
Trainable params: 29,345
Non-trainable params: 0
_________________________________________________________________
```

Figure 17(a): Hyper parameters for LSTM

The LSTM layers have been stacked as shown in figure 17(a). The first layer shows the output form all the layers. The layer description for this is pretty much same as in GRU except that there is an output gate along with the input and update ones.

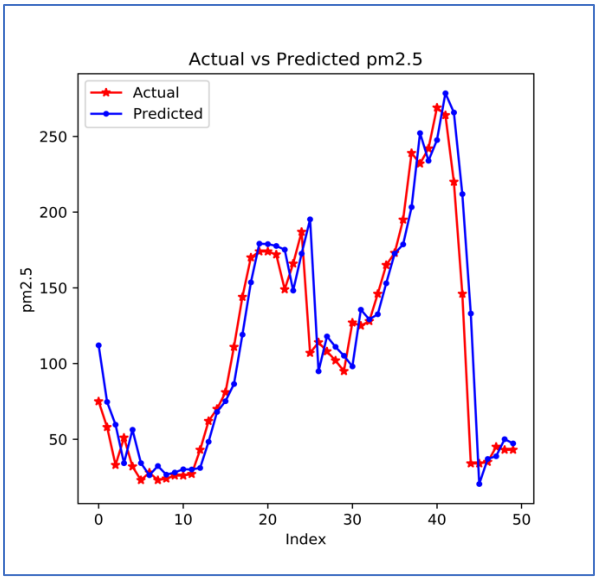

Figure 17(b): Actual vs. predicted graph for LSTM

Evaluation: There are slight discrepancies in the predicted but the overall aggregate remains somewhat unchanged as compared to the actual concentration. The slight discrepancies could be due to outliers in the data which were not replaced by the LSTM update.

Result: The MAE error for LSTM is 11.9904. This shows that the scope of error when using LSTM for this approach is minimum as compared to the other models.

4.9 Conclusion

This section shows that all the models implemented and the scientific methodologies used. In entirety, the models have performed exceptionally well considering a low error rate. Yet, CRISP DM methodology has the scope for improvement considering that it has a reverse back to business understanding where any error causes can be identified and removed. Thus, this implementation chapter have solved the research question and the sub research question framed

in chapter 1, all the research objectives have been implemented successfully which will impact in increasing the life expectancy in Beijing.

5 Discussion

On executing the models, if a sequential analysis is made on the basis of graph shown in figure 18, MLP failed in comparison to all the models which have been implemented in terms of error. The reason for failure of MLP to deliver despite being able to perform better with hidden layers is that it is not memory based. CNN margin of error is slightly less as compared to MLP, which has performed worse. CNN model might have not worked well due to the fact that the data being used is not sequential and also, not a specific type of pattern detection is required in the data, despite these non-recommended parameters. The other models implemented have the option to update data based on the requirements of the research and even remove any outliers from the data if needed. GRU and LSTM, both performing on neural networks have the options to update the data incoming based on the requirements of the project. LSTM has outperformed all the other models in the research due to its high adaptability and ability to change input values and also logically defining the forget gate. The aforementioned work concludes that objective 5 has been successfully solved by comparing all the implemented models.

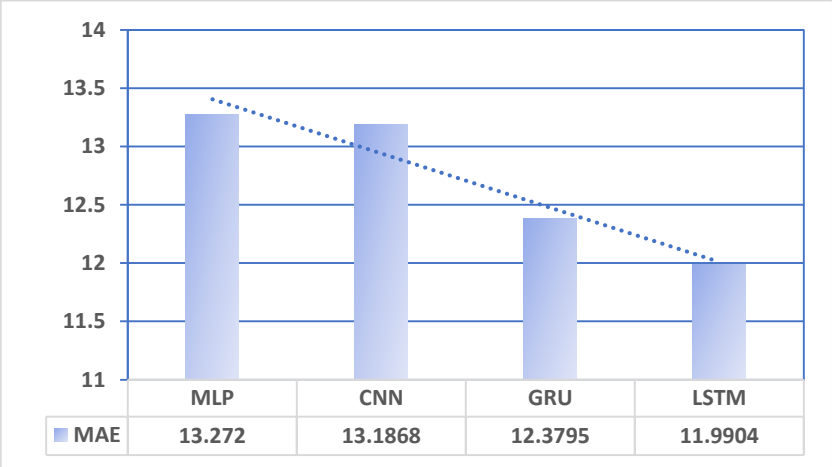

Figure 18: Comparison of MAE for all implemented models

6 Conclusion and Future Work

The research in conclusion was a success as the research question is answered as we can help to improve the air quality using LSTM model outputs. This can be done by reducing produce at times when meteorological aspects are at their peak and effecting the PM2.5 concentration in the environment. This leads us to believe there are measures of containing the concentration levels of PM2.5 as we know the times they can increase. The research concluded that for Neural Networks, the performance of all the models is seen as expected theoretically that Recurrent Neural Network was outperformed by GRU and LSTM. However, a key aspect to this is that the data sample taken for training is small, hence, the results are good for a small duration, but not a long period of time.

If extended, the future works of this research could implement up-sampling to train the data better. Another extension could be to change the hyper parameters of the research using LSTM and GRU to see how the models respond to that change. To make the research scope suit the time allocated to this project, small amount of data was used. On extending the research, the model can now be implemented with more data on a higher processing graphical processing unit (GPU), hence, there could be better processing abilities which could lead to larger training of data in less time and furthermore, better accuracies in result. Hence further research can be done to help learn to effectively detect and control the concentration of air pollutants with more meteorological features.

7 Acknowledgment

I would like to specially thanks my supervisor Dr. Catherine Mulwa for her guidance, supervision, help and support throughout the whole research project. I would also like to acknowledge my parents for always supporting me and having their trust in me.

References

Athira, V., Geetha, P., Vinayakumar, R., & Soman, K. P. (2018). DeepAirNet: Applying Recurrent Networks for Air Quality Prediction. *Procedia Computer Science*, *132*, 1394–1403.

Castanas, E., & Kampa, M. (2004). Human Health effects of air pollution. *Environmental Pollution*, *2*(151), 362–367.

Cort, J. W., & Kenji, M. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.

Dey, R., & Salem, F. M. (2017). Gate-variants of Gated Recurrent Unit (GRU) neural networks. *Midwest Symposium on Circuits and Systems*, 1597–1600.

Fan, S., & Chen, L. (2006). Short-term load forecasting based on an adaptive hybrid method. *IEEE Transactions on Power Systems*, *21*(1), 392–401.

Farhani, S., & Ozturk, I. (2015). Causal relationship between CO2emissions, real GDP, energy consumption, financial development, trade openness, and urbanization in Tunisia. *Environmental Science and Pollution Research*, *22*(20), 15663–15676.

Gardner, M. W., & Dorling, S. R. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment*, *33*(5), 709–719.

Hao, J., Wang, L., Shen, M., Li, L., & Hu, J. (2007). Air quality impacts of power plant emissions in Beijing. *Environmental Pollution*, *147*(2), 401–408.

Huang, C. J., & Kuo, P. H. (2018). A deep CNN-LSTM model for particulate matter (Pm2.5) forecasting in smart cities. *Sensors*, *18*(7).

Lary, D. J., Lary, T., & Sattler, B. (2015). Using Machine Learning to Estimate Global PM 2.5 for Environmental Health Studies. *Environmental Health Insights*, *9*(S1), 41–52.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*(September), 997–1004.

Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K., & Tsao, C. C. (2017). Urban air quality forecasting based on multidimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PLoS ONE*, *12*(7), 1–17.

Mass, C., & Portman, D. (1989). Major Volcanic Eruption and Climate: A Critical Evaluation. *Journal of Climate*, *2*, 566–593.

Mishra, D., Goyal, P., & Upadhyay, A. (2015). Artificial intelligence based approach to forecast PM2.5 during haze episodes: A case study of Delhi, India. *Atmospheric Environment*, *102*, 239–248.

Ni, X. Y., Huang, H., & Du, W. P. (2017). Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. *Atmospheric Environment*, *150*, 146–161.

Priddle, R. (2016). World Energy Outlook - Special Report Energy and Air Pollution. In *World Energy Outlook - Special Report*.

Shen, J. (2012). *PM 2.5 concentration prediction using times series based data mining*.

Tsai, Y. T., Zeng, Y. R., & Chang, Y. S. (2018). Air pollution forecasting using RNN with LSTM. *Proceedings - IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, IEEE 16th International Conference on Pervasive Intelligence and Computing, IEEE 4th International Conference on Big Data Intelligence and Computing and IEEE 3*, 1074–1079.

Wirth, R., & Hipp, J. (2010). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, (24959).

Xie, L., Wang, J., Wei, Z., Wang, M., & Tian, Q. (2016). DisturbLabel: Regularizing CNN on the loss layer. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4753–4762.

Yang, M. (2018). *A Machine Learning Approach to Evaluate Beijing Air Quality*.