Comparison Between Machine Learning Techniques for Early Air-Pollution Detection: A Case of Bojanala Platinum District Municipality (BPDM), South Africa

Koyana.Ntombikayise[1,*], Elisha D. Markus[1], Malusi Sibiya[2], and Adnan M. Abu-Mahfouz[3]

[1]Department of Electrical, Electronic and Computer Engineering, Central University of Technology Bloemfontein, South Africa
[2]Department of Computer Science, University of South Africa, Johannesburg, Florida campus
[3]Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
Email: ntombikayise.koyana0@gmail.com (K.N.); emarkus@ieee.org (E.D.M.); sibiym@unisa.ac.za (M.S.); a.abumahfouz@ieee.org (A.M.A-M.)
Manuscript received July 7, 2023; revised August 22, 2023; accepted November 9, 2023; published June 24, 2024

***Abstract*—Air pollution has been one of the major threats to modern livelihoods. It is affecting, health, economies, and social well-being and has even resulted in fatalities in certain instances. As part of the fourth industrial revolution in South Africa, there has been a recent focus on smart cities. Hence, exploring new ways of combating the menace of air pollution has become pertinent. Machine learning techniques have been applied to solving many modern problems. However, in South Africa, these technological solutions are still at their inception. This paper proposes an early air pollution detection technique for a city in South Africa. This city has experienced air pollution problems in the past owing to the presence of many mining and industrial activities. Past data collected in the city shows a pattern of air pollution threatening the city's fabric. If left unchecked, the result on health and livelihoods would be disastrous. Deep learning neural networks, multiple linear regression, and random forest trees for regression are used to model the pollution patterns, and a short-term prediction strategy was designed to warn residents ahead of impending catastrophes that could be damaging to their health. Based on the results, the random forest regressor model provided better predictions and is recommended for deployment. The results therefore show that the early air pollution detection strategy provides good outcomes and could effectively monitor and warn residents and authorities.**

***Keywords*—machine learning, air pollution, PM2.5, forecasting, regression model**

I. Introduction

In recent years, South Africa has experienced significant economic growth and development, characterized by rapid urbanization and industrialization. However, this progress has come at a cost, as the nation grapples with a pressing environmental challenge: air pollution. This issue has cast a shadow over the country's urban centres, posing a significant concern for both the government and its citizens. The burgeoning industrialization within cities has led to a parallel increase in population, vehicular traffic, and a heightened potential for deteriorating air quality. Moreover, the prevalent use of coal and wood as primary sources of heating and cooking in settlements and households has added another layer of complexity to the issue of air quality, warranting immediate attention [1].

Recognizing the gravity of this situation, the South African government has taken steps to address air pollution through the South African Air Quality Information System (SAAQIS). This system has been instrumental in monitoring and collecting air pollution data across the nation, underscoring the nation's commitment to mitigating the adverse effects of air pollution. As South Africa moves towards a future marked by smart cities and innovative technologies, there is a growing need to harness the power of machine learning to effectively detect and combat air pollution in real time [1, 2]

This paper delves into the critical issue of early air pollution detection in smart cities within the context of the Bojanala Platinum District Municipality in South Africa. By leveraging machine learning techniques, the study aims to shed light on innovative solutions that not only monitor air quality but also provide valuable insights for policy-making and urban planning.

II. Air Pollution in the Local Municipality

This section discusses the specific challenges of air pollution within the Bojanala Platinum District Municipality (BPDM) and what can be done to curb its effects. The BPDM serves as an illustrative case study due to its unique characteristics, including a history of extensive mining activities, which have contributed significantly to the air quality challenges within this geographical area. To enforce compliance with emission standards, monitoring of pollution is done with directives for atmospheric impact reports or pollution prevention plans conditions or requirements for Atmospheric Emission Licenses (AEL) and information provided to an Air Quality Office (AQO) at Bojanala Platinum District Municipality (BPDM). This is done in partnership with communities and stakeholders to ensure that healthy air quality is maintained. To ensure that the community breathes clean air that is not detrimental to the well-being of persons in the district. Additionally, measures have been taken to prevent the possibility of short- and long-term harm to delicate natural environmental systems from air pollution during future expansions in the areas of transportation, mining, housing, etc. [1]. However, the impact of air quality on health is covered by the National Ambient Air Quality Standards (NAAQS), but the impact of air quality on the global ecosystem is not known. The BPDM's most likely sources of air pollution, where they emit their emissions, and where they are most noticeable have all been identified in addition to reviewed in Table 1.

A. Major Impacts of Air Pollution in the Municipality

The Air Quality Management Plan (AQMP) was compiled to set out what should be done to achieve the prescribed air

quality standards. All municipalities were required to include an AQMP as part of their Integrated Development Plan. After the enactment of the Air Quality Act (AQA), the approach to monitoring air quality in South Africa has evolved from focusing solely on point sources to adopting a more comprehensive method that considers the effects on the receiving environment, including humans, plants, animals, and structures.

The foundation of this philosophy lies in proactive planning, which involves the development of air quality management plans for all municipal areas, the licensing of specific industrial processes, and the identification of priority areas where air quality falls short of the standards set for certain air pollutants [2].

Table 1. Synopsis of contamination sources, emission, and the affected area within the BPDM

Contamination source	Major pollutants produced	Crucial affected regions
Industrial activities encompass emissions from small-scale boiler sources, as well as those from more extensive sectors like steel processing and cement manufacturing.	PM10, PM2.5, CO, NO_x, SO_2, Pb	Most of the industries in the district are in Rustenburg and Madibeng. These two Municipalities account for over 90% of all reported industrial emissions in the BPDM
Agricultural activities:	PM10, PM2.5	Agriculture is a dominant land use within many areas of the BPDM, with subsistence farming occurring in the Moses Kotane and Moretele Local Municipalities. BPDM is home to the two largest platinum-producing mines in the world, with most of the mines in the district being in the Rustenburg and Madibeng local municipalities. Mines are also located in the Kgetlengrivier local municipality
The combustion of biomass	PM10, PM2.5, CO, NO_x, SO_2	The provincial emissions survey provides good estimates of veld fire emissions, and these indicate that the local municipality of Rustenburg is the main source of veld fire emissions.
Domestic fuel burning	PM10, PM2.5, CO, NO_x, SO_2	The use of fuels such as coal, wood, and paraffin for cooking and space-heating purposes occurs mainly within rural, informal, low-income, and densely populated settlements.
Vehicle tailpipe emissions	NO_x, CO, CO_2, SO_2, VOCs,	Most vehicle emissions within the BPDM happen in the N4 highway area compared to the other districts Rustenburg and Mjadibeng have a high impact on vehicle emissions because they are industrialized.
Transboundary Transport of Air Pollution	various	Since this air pollution originates outside of their borders of control, it presents a difficult problem for the authorities in charge of handling air quality issues in their respective regions.

The BPDM was identified as an area of poor air quality in the previous years. National Framework and the baseline assessment confirmed this statement, highlighting the industrial area which includes the Merensky Reef, which stretches from west of the Pilanesberg, southwards through the Bafokeng area and parallels the Magaliesberg towards Marikana and Brits in the east. There are several supplementary manufacturing industries co-located in the area. The two largest platinum mines in the world, Anglo Platinum, and Impala Platinum, are situated in the district BPDM. Chrome, lead, marble, granite, and slate are also produced in the area [3]. Air quality impacts were found to extend from Rustenburg towards Brits, up the eastern boundary. Most of the industrial and domestic fuel-burning sources are in this area. The other local municipalities were found to have low pollution loads. Emissions from industrial operations were attributed to small boiler sources and industries such as galvanizing works and autocatalytic manufacturing. The main pollutants generated from these processes are PM10, PM2.5, CO, NOx, SO_2, and Pb. Agriculture is the dominant land use in the region and these activities are associated with emissions of PM10 and PM2.5. Moses Kotane LM was found to be the largest contributor to biomass burning emissions, associated with emissions of PM10, PM2.5, CO, NO_x, and SO_2. The use of wood, coal, and paraffin was confined to low-income, densely populated rural areas and informal settlements. The main area of concern for motor vehicle emissions was noted to be along the N4 highway. High vehicle emissions were reported in Rustenburg and Madibeng LMs, as the most industrialized municipalities within the district. There are several landfill sites in operation in the district; these were identified as sources of heavy metals, dioxin, and furan emissions. Transboundary emissions were also identified as a source of pollutants in the district, but specific pollutants were not listed [3].

1) Measuring air pollution in the local municipality

The purpose of the Air Quality Act was to implement Section 24 of the Constitution by offering reasonable safeguards against air pollution. According to the Air Quality Act, AQMPs must be created by local governments and included in their integrated development plans. Several municipalities in South Africa, including Rustenburg Local Municipality, have addressed their obligations and created AQMPs. Information from the Northwest Provincial AQMP was used in this article. The designation of priority areas where the air quality is thought to be poor and harmful to both human health and the environment is another feature of the Air Quality Act. The Minister of Environmental Affairs and Tourism in South Africa declared the Vaal Triangle to be the priority area. However, the initiatives are aimed at managing poor air quality in air pollution hotspot areas that cross the Gauteng and Free State provincial boundaries in the case of the Vaal Triangle Airshed Priority Area (VTAPA). The second priority area to be declared in South Africa was the Highveld Priority Area. Parts of the provinces of Gauteng and Mpumalanga are included in the Highveld Priority Area (HPA), which also comprises three district municipalities (Sedibeng, Gert Sibande, and Nkangala) and one metropolitan municipality (Ekurhuleni). Currently, work is being done on the Highveld Priority Area Air Quality

Management Plan [4].

As a result, the Minister designated the Waterberg–Bojanala Priority Area (WBPA), which spans the Northwest and Limpopo Provinces, as the third National Priority Area. BPDM has been designated as an area with poor air quality by the Republic of South Africa's National Framework for Air Quality Management. Therefore, this study focuses on the deposition of particulate matter (PM10, PM2.5) and sulphur dioxide (SO_2) within Rustenburg in the Waterberg-Bojanala Priority Area [5].

2) Inadequacies in the legacy systems

Lack of awareness of air quality is another issue that has been identified, but funding for air quality projects is also limited because air quality monitoring stations require a lot of expensive equipment. Some of these must be imported into South Africa, because emissions are still a problem even when data is not loaded into the database for early warning. Due to serious health apprehensions, atmospheric pollution has developed as a major basis of premature mortality among the public by causing millions of deaths each year World Health Organization (WHO). Nevertheless, Limb stated no city area completely follows air quality guidelines set by WHO [6]. Separately from those who suffer from asthma, cardiovascular problems and respiratory issues, children and ageing people are at elevated risk of being disposed to the negative effects of atmospheric pollution [7]. South African Air Quality Information System (SAAQIS), a centralized database will be developed at the South African Weather Services (SAWS) to which all verified ambient monitoring data will be transferred to a database. Some source and emissions data recorded within each Municipality and Province will be incorporated into a national electronic database, allowing easy access and manipulation of data from any sphere of government. The BPDM will need to ensure that their current emissions database is regularly updated to maintain accurate data for the stations at SAAQIS [1]. Fig. 1 illustrates a typical air quality monitoring station, indicating analysers for different pollutants.

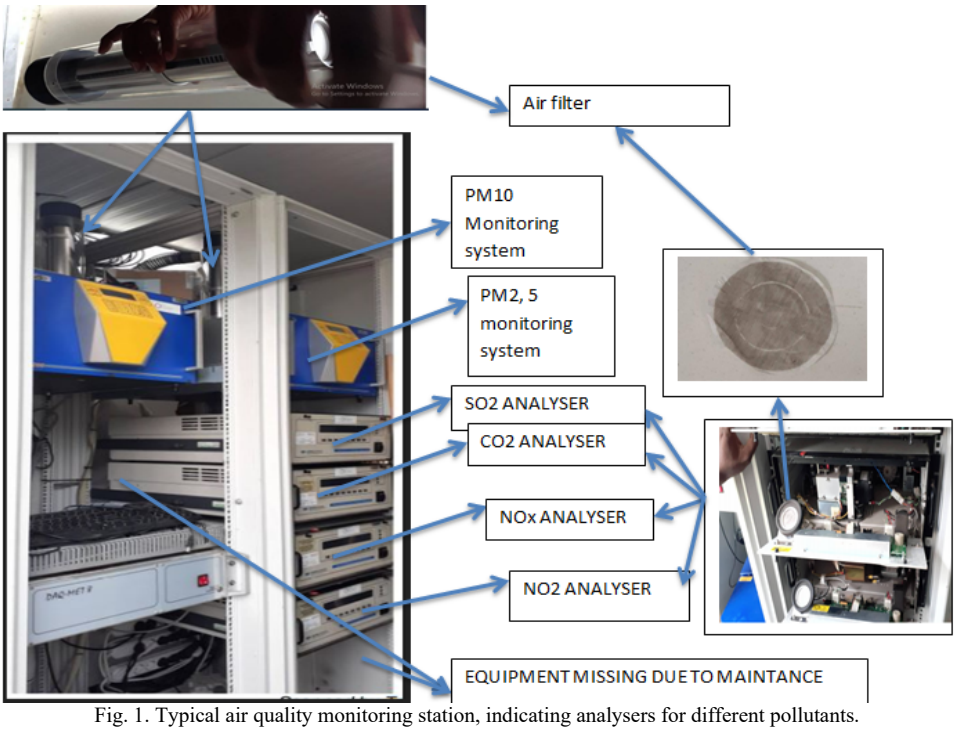


Fig. 1. Typical air quality monitoring station, indicating analysers for different pollutants.

III. Related Work

A. Early Air Pollution Detection

Several authors have proposed a review of the performance of Multiple Linear Regression (MLR) and Multi-layer Perceptron (MLP) intended for forecasting SO_2 awareness in the air [8]. Different strictures to be precise climatological strictures, city traffic data, urban green space information, and time strictures were elected for the prediction of SO_2 daily awareness to analyse and improve predicting air pollution and air quality management. Bellinger *et al*. [9] conducted a systematic literature review on the application of data mining and machine learning methods in air pollution epidemiology. The survey showed that the majority of the studies have been conducted in Europe, China, and the USA and that data mining is becoming an increasingly common tool in environmental health. The authors recognized that deep learning can be used for forecasting air pollution and generating hypotheses. Artificial neural networks are preferred over decision trees. Machine learning algorithms, both classifiers and regressors, have been applied to forecasting and prediction problems. Clustering algorithms, such as K-Means and hierarchical clustering, have been applied for knowledge discovery and source apportionment. The results show that data mining is increasingly useful in air pollution epidemiology.

Kang *et al*. [10] studied various big-data and machine learning-based techniques for air quality forecasting and assessment using artificial intelligence methods, such as decision trees and deep learning. With the development of IoT infrastructures, big data technologies, and machine learning techniques, real-time air quality monitoring and assessment are essential for future smart cities. In another study, Alyousifi *et al.* [11] proposed a novel Markov-

weighted fuzzy time series model and an innovative cross-forecasting model for predicting daily Air Pollution index (API) data or any type of time series, based on selecting the optimal divider method. The partition method consisted of two stages: in the first stage, the grid partition technique was used to determine the appropriate number of partitions of the universe of discourse. In the second stage, different partition methods were used, including the grid technique with the optimal partition number from the first stage. The best partition method among those approaches was chosen. Thus, both stage partition approaches avoided random selection of intervals, which significantly improved the model's accuracy.

The use of machine learning and deep learning approaches has been proposed by many studies for the prediction of air pollution and the improvement of environmental health. Time-series reading estimation is one such approach that has been utilized. For instance, Bell *et al.* [12], conducted a review of methods, and the findings of the time-series readings estimate health risks associated with interim exposure to particulate matter (PM). The findings provided strong evidence supporting an association between PM levels and adverse public health impacts. Mahalingam *et al.* [13] developed a model to predict the AQI of smart cities and tested it in Delhi, India. The authors reported that the medium Gaussian Support Vector Machine (SVM) exhibited maximum accuracy. The authors claim that their model can be used in other smart cities too.

Similarly, Hu *et al.* [14] presented a machine-learning model that combines sparse fixed-station data with dense mobile sensor data to estimate the air pollution surface for any given hour on any given day in Sydney. The system was evaluated using seven regression models and tenfold cross-validation. The results showed that the estimation accuracy of Support Vector Regression (SVR) is comparable to decision tree regression and random forest regression, and higher than extreme gradient boosting, multi-layer perceptrons, linear regression, and adaptive boosting regression. Validation of air pollution estimates was conducted through field trials, and the result showed that SVR not only produces high spatial resolution estimates that match well with the pollution surface obtained from fixed and mobile sensor monitoring systems but also identifies the boundaries of polluted areas better than other regression models. Results were demonstrated using a Web-based application customized for metropolitan Sydney. The authors claimed that the continuous estimates provided by the proposed system could better inform air pollution exposure and its impact on human health.

Martínez-España *et al.* [15] proposed air pollution prediction in smart cities using machine learning methods. The authors analysed different machine-learning techniques for predicting ozone levels and determined the best model for ozone prediction. A study by Cecilia *et al.* [16] proposed a high-throughput solution for real-time air pollution monitoring and assessment in the field of V2I communication. The paper introduced a hardware-software infrastructure for providing novel cooperative Intelligent Transportation Systems (ITS) services based on big data processing from vehicles. The solution is a geo-located air quality service that uses a fuzzy clustering technique on heterogeneous servers with CPU and multiple GPUs to predict ozone levels and identify polluting traffic areas and drivers. The study evaluated the performance and scalability of the proposed infrastructure under high loads and showed that it can handle realistic and challenging deployments. Dominici *et al.* [17] proposed semi-parametric regression directly relevant to risk estimation in time series studies of air pollution and health. Time-series data on air pollution and mortality were analysed by using semi-parametric regression models; where the linear component measures the air pollution effects and the smooth component adjusts for confounding by time-varying factors such as season and weather. Although the reviewed studies proposed various techniques for air pollution monitoring and assessment, the use of deep learning and time series machine learning algorithms was not considered.

This research attempts to develop air pollution monitoring architectures using deep learning and compare them with other existing approaches to highlight the importance of deep learning in this field. Deep learning has recently shown a significant impact on air pollution forecasting, and its use can reduce the lack of significance in this area and lead to more projects. When deep learning models for air pollution estimation are compared with other methods such as artificial neural networks and fuzzy logic, they often produce more accurate results. Several studies have proposed deep learning architectures for air pollution prediction, such as Long Short-Term Memory (LSTM) and Denoising Autoencoders (DAE) [18], Other models explored in these studies encompass LSTM, Space-Time Deep Learning (STDL), Dual Attention Learning (DAL), and Convolutional Neural Network (CNN). Hybrid deep learning frameworks [19, 20], and deep learning models that learn spatial-temporal correlation features [21]. These models have shown promising results in predicting various pollutants such as PM2.5, NO_2, SO_2, and O_3 [18–21]. Kok *et al*. [22] proposed a deep learning model for analyzing IoT smart city data. This model uses Long Short-Term Memory networks to predict future values of air quality in a smart city. The prediction results of the proposed model proved to be favourable and showed that the model can be applied to other smart city scenarios such as air pollution. In Ref. [23], the authors proposed the development and evaluation of forecasting models for NO_2 pollutants in the air using deep learning and time series methods. The development involves the implementation of combined AutoRegressive Integrated Moving Average (ARIMA), Seasonal AutoRegressive Integrated Moving Average (SARIMA), Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend (TBATS) and neural networks. The results showed that the neural network with stacked LSTM outperformed all other models.

In a study by Shams *et al*. [8] Multiple Linear Regression (MLR) and Multi-layer Perceptron (MLP) were compared for predicting SO_2 daily concentration in the air using atmospheric parameters, urban traffic data, green space information, and time parameters. The evaluation of the accuracy, effectiveness, and simplicity of the MLR and MLP models for air pollution forecasting showed that the MLP model performed much better than the MLR model. The authors also acknowledged that the Artificial Neural Network (ANN) model can be widely used in environmental research, including air pollution prediction, due to its impact on human

health and the environment. Iskandaryan *et al.* [24] conducted a review on air pollution prediction using machine learning algorithms based on sensor data in smart cities. The study highlighted some important discoveries: Firstly, 38% of the studies employed more advanced and sophisticated techniques like 1D Convnets and Bidirectional Gated Recurrent Units (GRU), Deep CNN-LSTM and other neural network models. Secondly, about 63% of the literature reviewed was focused on China. This highlights the need for more research focused on applying advanced techniques to improve air pollution prediction and environmental health in South Africa. Thirdly, since most studies used existing datasets, external factors were not considered. The authors recommended the inclusion of external factors to reflect more practical scenarios.

Shahid *et al.* [25] introduced a framework for traffic forecasting using an existing dataset on air pollution. The reason for the choice of data was that it captures air pollution associated with traffic congestion. Initially, a virtual analysis of seven regression models was performed to find out which model gives better accuracy. The result showed that Multi-Layer Perceptron gave the best metrics score. The authors then suggested a framework using regression models in which the initial regression model result is increased using the boosting collaborative technique and showed that the proposed framework gave more substantial outcomes than the above 7 regression models.

A literature review conducted by Koyana *et al*. [26] discussed the current state of air pollution monitoring in South Africa and the potential of intelligent techniques to address the challenges. The study showed that there is a need to reassess the existing monitoring stations and to measure a complete set of pollutants, rather than focusing on specific ones. The authors suggested that using machine learning models to monitor air pollutants can reduce the cost and provide long-term insights into the environmental and health impacts of air pollution in the province. Kow *et al*. [27] proposed a hybrid model that combines a Convolutional Neural Network and a Back Propagation Neural Network to forecast PM2.5 concentrations for multiple locations simultaneously. The hybrid model was evaluated against three types of machine learning models (Backpropagation Neural Network (BPNN), Random Forest (RF), and dynamic LSTM) and the result showed that it achieves the best results in terms of the RMSE metrics. The study also showed that the hybrid model can significantly improve the accuracy and reliability of long-term PM2.5 forecasting.

From the foregoing, previous studies have indicated significant use of machine learning techniques for air pollution monitoring globally. However, the use of machine learning and deep learning to improve air pollution monitoring and environmental health in South Africa is still in its nascent stage. Just a few studies like that of Chiwewe *et al.* [28] and Shikwambana *et al*. [29] have been conducted in South Africa. Although, Chiwewe *et al.*[28] proposed a big data modeling approach to predict the ground-level Ozone levels based on the cross-correlation bandwidth readings for multiple stations in Gauteng province in South Africa, Shikwambana *et al*. [29] introduced a similar approach to monitor the temporal patterns of PM10 concentration using an observation camera and a regression algorithm calibrated by the atmospheric reflectivity and the measured air quality data. These studies only provided a partial view of the viability of these techniques for air pollution monitoring. This study has provided a comprehensive comparison of various machine-learning techniques for air pollution monitoring.

IV. Prediction Models and Data Analysis

A. Data Pre-processing

In this study, the Boitekong station and Marikana CC-NAQI station were selected for the Rustenburg province. Fig. 1 presents a typical air quality monitoring station, indicating analyzers for different pollutants.

Understanding the structure of data is particularly important in machine learning projects. If the data are understood, machine learning algorithms with optimized hyperparameters may be achieved. For the dataset obtained from the selected stations, to be usable for prediction, data pre-processing has to be carried out to eliminate outliers and ensure all observations are recorded with the correct datatype. Data pre-processing techniques like data wrangling, were employed as an important step in the cleaning of the data. Using different pre-processing methods in the Python Integrated Development Environment (IDE), the data were pre-processed with the help of Python libraries that were imported to the Python IDE. Fig. 2 illustrates how the dataset was pre-processed before it was deemed ready for use by the three machine-learning models.

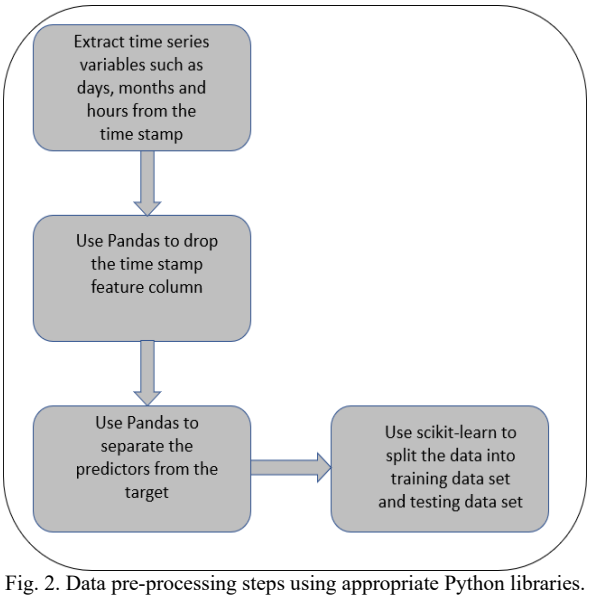


Fig. 2. Data pre-processing steps using appropriate Python libraries.

Table 2. A select few Python libraries that were used in the development of machine learning algorithms for the prediction of PM2.5 gas

Number	Name of the library	Purpose of the library
1	NumPy	To perform array operations
2	Pandas	To provide data framework and manipulation
3	Scikit-learn	To provide machine learning algorithms
4	Matplotlib and Seaborn	To perform data plots
5	Keras-Tensorflow	To provide a neural network platform

The selected dataset was for a total of about four years of observation by the selected station and consisted of thirteen gas input features and a total of 12,807 samples after the

dataset was cleaned. The dataset was used to develop models based on three machine learning algorithms to predict the concentration of hazardous PM2.5 gas in the air. After the split, the training data set comprised 10,246 observations while 2,562 observations were reserved for the testing data test. Table 2 lists a few important Python libraries that were used for various purposes in this study.

Table 3 shows the first eleven observations of the pre-processed dataset with 13 predictors and one target feature which is the PM2.5 column.

Table 3. First eleven observations of the air pollution data set (Time series) with predictors and PM2.5 target feature

No.	SO_2	NO_2	NO	NO_x	PM2.5	PM10	Amb_Wspeed	Amb_WDirection	Temperature	Amb_ReHum	Amb_Pressure	Days	Hours
0	4	2	0	2	22	36	1.6	76	16	55	882	1	1
1	5	2	0	2	19	32	1.9	68	16	56	883	1	2
2	3	1	0	2	13	23	2	57	16	57	883	1	3
3	2	2	1	2	12	21	2.2	57	15	61	883	1	4
4	3	2	0	2	11	20	2.4	66	14	66	883	1	5
5	3	2	0	2	10	18	2.3	59	14	68	883	1	6
6	2	2	1	2	9	18	2.5	51	14	67	883	1	7
7	1	2	1	2	9	20	3	43	17	59	884	1	8
8	1	2	1	3	9	24	3.3	43	19	51	884	1	9
9	1	1	1	2	10	23	3.5	39	21	42	883	1	10
10	1	1	1	2	11	21	3	42	24	33	883	1	11

B. Data Analysis

Using Seaborn and Matplotlib libraries, the histograms and scatter plots were generated to understand the feature distribution in the data set. These plots are shown in Figs. 3 and 4. For instance, if information about the SO_2 feature were to be understood, we would look at the SO_2 histogram in Fig. 3. A closer look at this histogram shows that SO_2 has more than 10,000 negative values in the range of zero and −20. Fig. 4 shows a scatter matrix plot of the data features. The purpose of this plot is to understand the relationship of data distribution among the data features. For instance, some features are linearly correlated with each other. These linearly correlated features as shown in Fig. 4, are NO_X vs NO_2, NO vs NO_X, and finally PM10 vs PM2.5.

As shown in Fig. 4 there are linear correlated features in the data set, it is therefore important to understand the degree to which these features are correlated. Data correlation is a way to understand the relationship between multiple variables and features in the dataset. Positive Correlation means that if feature A increases, then feature B also increases or if feature A decreases then feature B also decreases. So, both features move in tandem, and they have a linear relationship. The latter- mentioned features which are linearly correlated, all have a positive correlation. Negative Correlation means that if feature A increases, then feature B decreases and vice versa. In the data set used in this study, the features that tend to have a negative correlation as shown in Fig. 4 are temperature (Temperature) and relative humidity (Amb ReHum).

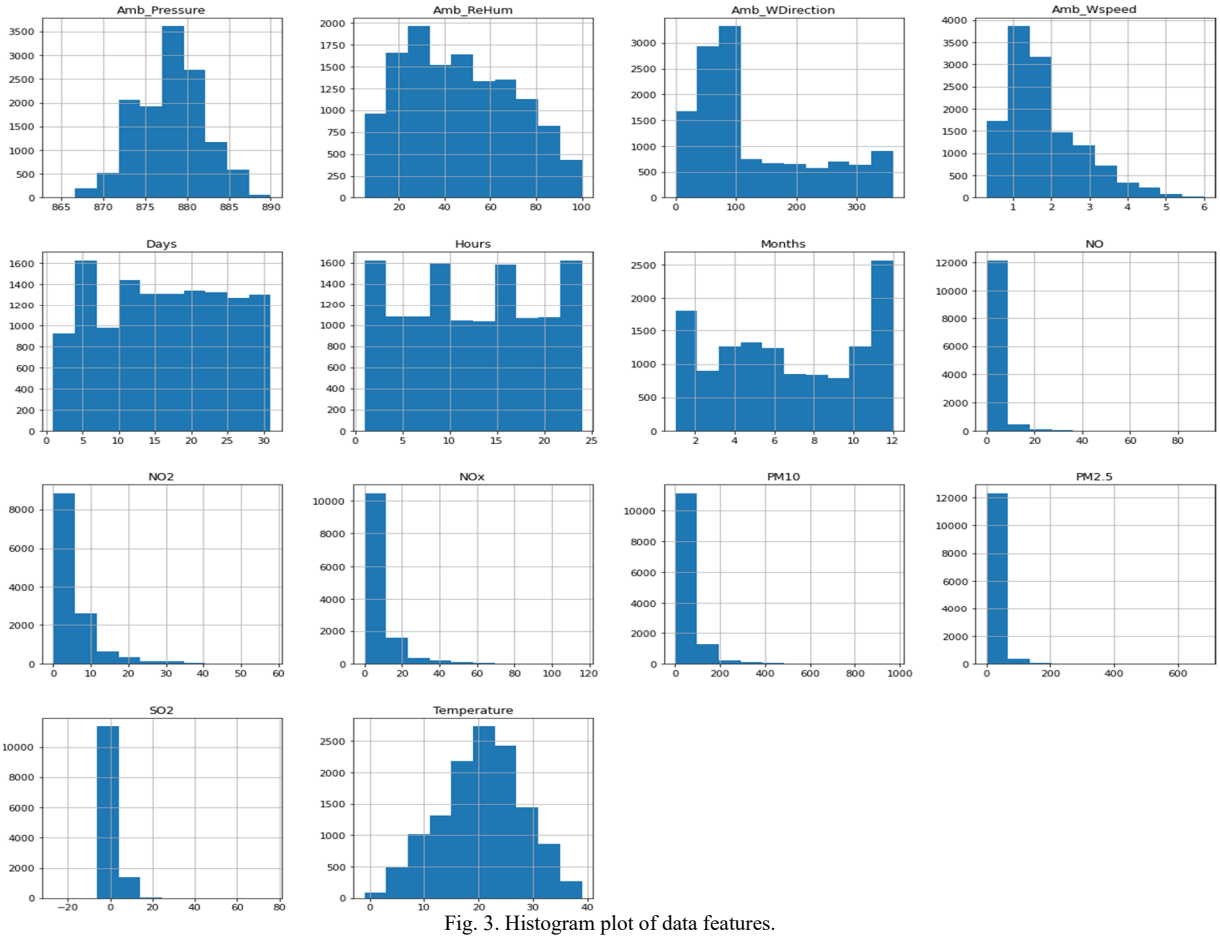


Fig. 3. Histogram plot of data features.

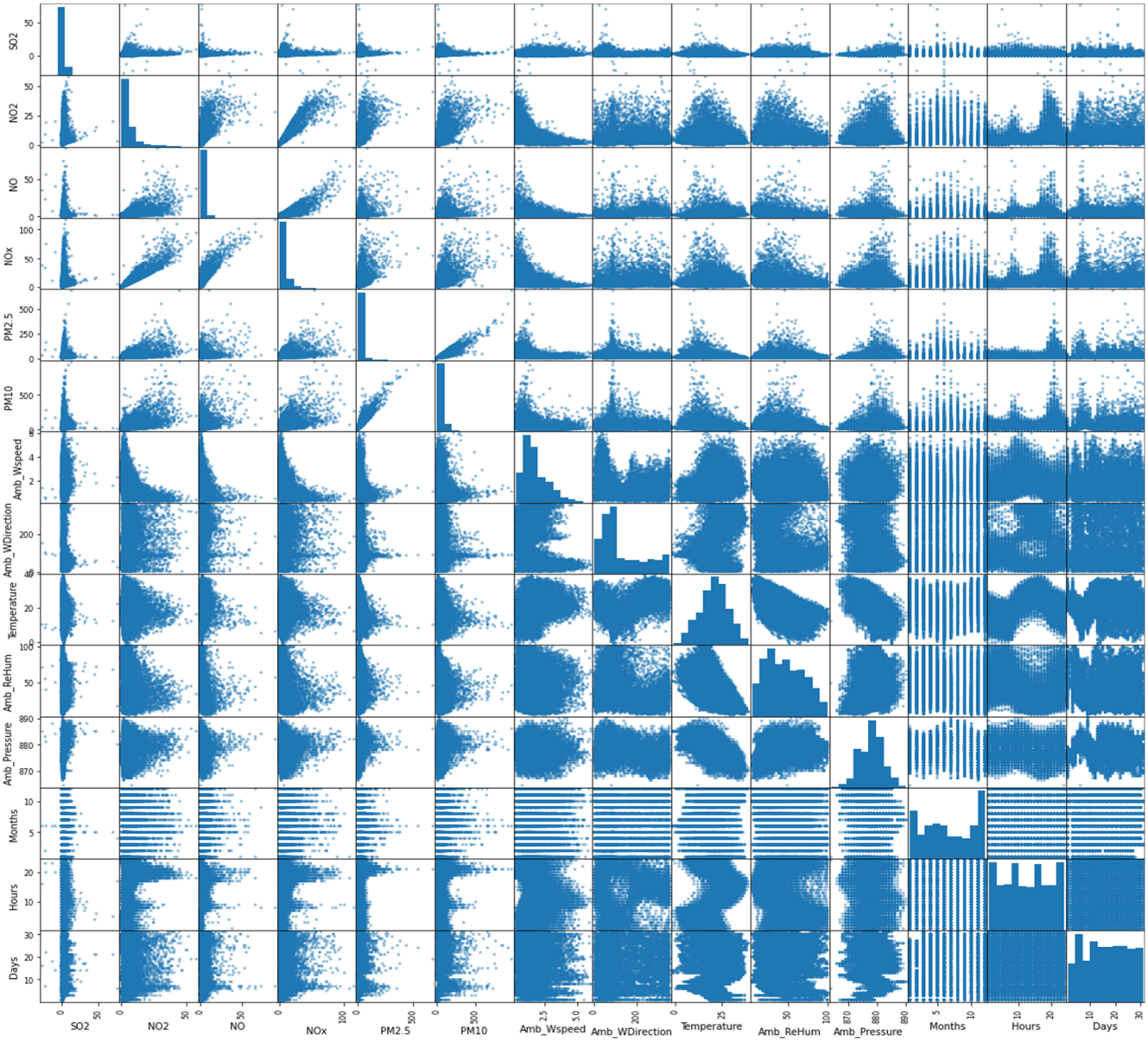


Fig. 4. Scatter matrix plot of the data features.

Fig. 5 illustrates the difference between negative and positive correlation. Each of these correlation types can exist in a range represented by values from 0 to 1 where slightly or highly positive correlation features can be of values in the range of 0.5 or 0.7. If there is a strong and perfect positive correlation, then the result is represented by a correlation score value of 0.9 or 1. If there is a strong negative correlation, then it will be represented by a value of −1. Highly correlated features may lead to skewed or misleading results due to a problem called multicollinearity. To overcome this, among a few available methods, ensemble tree models such as random forests may be used as they are immune to this problem (This was applied in this study).

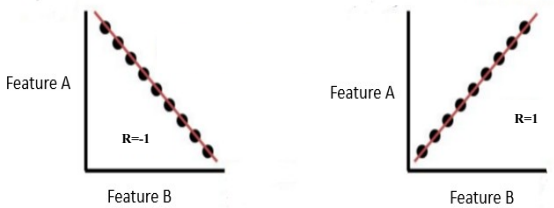


Fig. 5. Negative correlation (Left) and Positive correlation (Right).

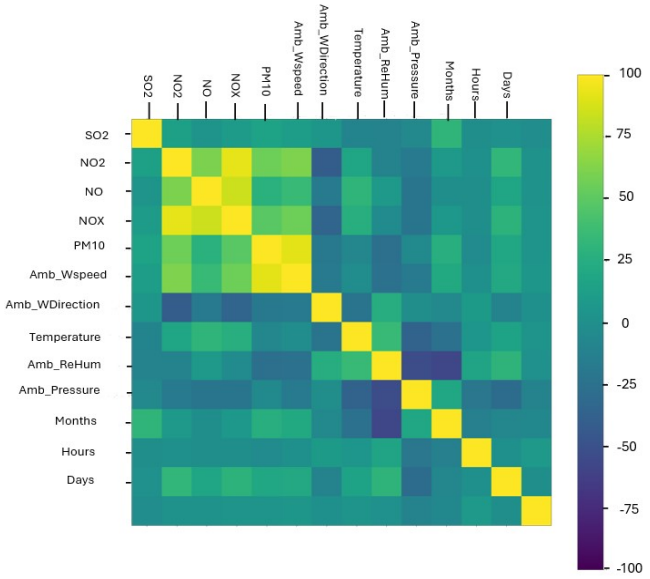


Fig. 6. Correlation plot of the data features.

To understand the degree of correlation between features, a correlation plot was used as illustrated in Fig. 6. The previously mentioned features that were said to be having

positive linear correlation relationship, have yellowish regions in the correlation plot. For instance, using the Pearson Correlation Coefficient in Python IDE, the calculated correlation between NO_2 and NO_X was 0.92. Among the three machine learning models used in this study, the multiple linear regression model ("Linear Regression") which is not immune to multicollinearity (caused by highly correlated features), scored a variance score of 0.86. This variance score was less than that of the "Random Forrest Regressor" model which scored a variance of 0.93. The variance of the "Random Forrest Regressor" was higher because it is immune to multicollinearity (caused by highly correlated features).

C. *Model Selection Procedure*

The neural network was initially proposed for this study due to its ability to oversee any data with non-linearities. Since a "No free lunch theorem" is highly considered in machine learning engineering, the two other machine learning algorithms were proposed after they passed a machine learning model selection test that was proposed by [30]. These two additional models that are available in the sci-kit-learn library are "Linear Regression" and "Random Forest Regressor." Fig. 7 shows the model selection procedure that was used for the selection of the two latter mentioned models. The two models were used with their default parameters. The parameter settings of the neural network which formed the third model are summarised below.

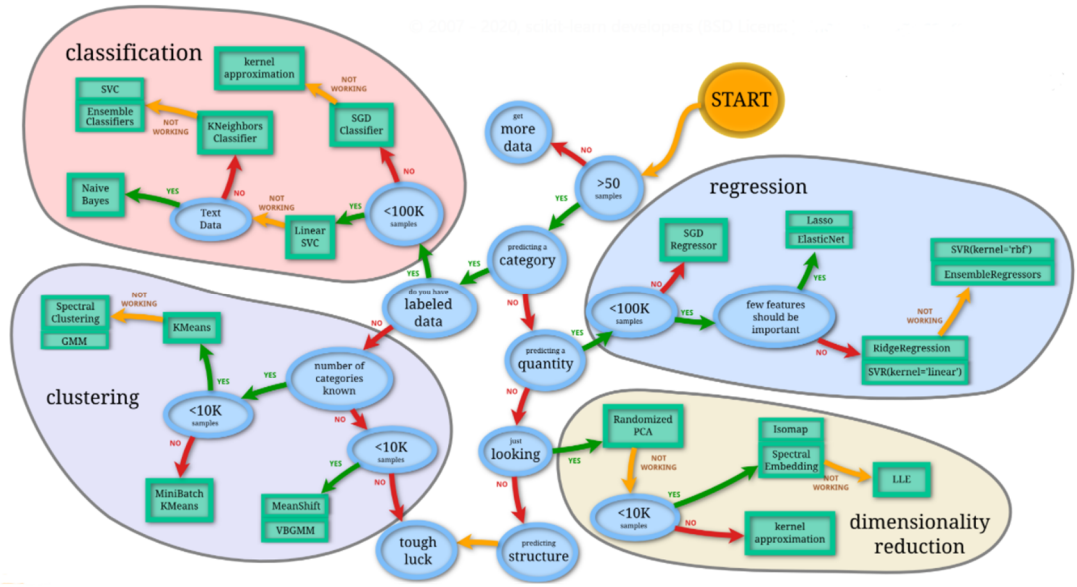


Fig. 7. Model selection procedure [30].

1) *Neural network model parameter settings*

The neural network model was built on the Tensorflow-Keras library with the model architecture.

The input layer consisted of thirteen input dimensions since there were 13 predictors in the dataset. One output node was used because there was only one output target feature. The input layer and two hidden nodes used rectified linear unit (relu) activations. Each hidden layer L consisted of seven neurons N that were obtained by using the general rule of thumb presented by Eq. (1).

$$N^l = \frac{x_i + y_i}{\sum_{i=0,\dots,I}^{I} L_i} \quad (1)$$

where:

N^l: represents the neurons N in each hidden layer l. Where l is the number of the current hidden layer.

x_i: is the input node for node number i

y_i: is the output node for node number i

L_i: is the hidden layer L for layer i and

$\sum_{i=0,\dots,I}^{I} L_i$: is the total number of hidden layers.

Since the nature of the problem dealt with in this study is a regression problem, the 'linear' activation function was used in the output node of the neural network. To train the network, one hundred training epochs were used and a total of 344 parameters were trainable as shown in Fig. 8.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 13)                182
_________________________________________________________________
dense_1 (Dense)              (None, 7)                 98
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 56
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 8
=================================================================
Total params: 344
Trainable params: 344
Non-trainable params: 0
_________________________________________________________________
```

Fig. 8. Neural network model architecture.

V. Performance Evaluation

Given that pollutant parameters were specifically monitored over a four-year period, there was a need to review and evaluate the effectiveness of the suggested models. Here, the MLR is used in a combination of several independently trained two-layer MLR networks. The learning algorithm of the scaled conjugated gradient and transfer functions of the sigmoid (hidden layer) and linear (output layer) is used. A network of its own is created and trained for each missing data pattern using an early stopping approach (to avoid over-fitting) with a test set of training in a proportion of 1/5. The number of inputs matches the number of values that are currently accessible, and the number of outputs matches the

number of values that are missing from the data rows with the missing pattern. N_{mlp}, the number of hidden neurons is determined through an experimental process.

$$N_o = round(2xI_o) + 1 \quad (2)$$

$$N_i = round(2xI_i) + 1 \quad (3)$$

$$(If N_o < N_i \; then \; N_{mlp} = N_i, then \; N_{mlp} = N_o) \quad (4)$$

where N_i and N_o are the number of hidden neurons defined from the inputs and outputs (o), I_o and I_i are the number of neurons in the input and output layers, and round is the rounding towards to nearest integer.

The metric R^2 and the root mean square error RMSE has been calculated using Eq. (5). The most common indicators of imputation ability are the correlation coefficient R and its square: coefficient of determination R^2n i.e., the variance explained, which is limited to a range between 0 and 1

$$R^2 = \left[\frac{1}{n}\frac{\sum_{i=1}^{n}[(P_i-\bar{P})-(O_i-\bar{O})]}{\sigma_P\,\sigma_o}\right]^2 \quad (5)$$

where N is the number of imputations, O_i the observed data point, P_i the imputed data point, $\bar{O}$ is the average of observed data, $\bar{P}$ as an average of imputed data σ_P the standard deviation of the imputed data and σ_o the standard deviation of the observed data.

The root mean squared error RMSE which summarises the difference between the observed and imputed concentrations was used to provide the average error of the model.

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{n}[[P_i - O_i]^2]\right)^{\frac{1}{2}} \quad (6)$$

where n is the number of observations, O_i is the observed parameter, P_i is the calculated parameter, O is the mean of the observed parameter, P is the average of the calculated parameter, is the standard deviation of the observations, and is the standard deviation of the calculations. A separate prediction is conducted for each pollutant (PM10 and PM2.5), using Multiple Linear Regression (MLR), Support Vector Regression (SVR) and Random Forest

$$MAE = \frac{1}{N}\sum_{i=1}^{n}[P_i - O_i] \quad (7)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left[\frac{P_i-O_i}{P_i}\right] \quad (8)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}[[P_i - O_i]^2] \quad (9)$$

where P_i represent the real values while, O_i represents the predicted values of ambient pollutants, which to make model results more intuitive and reliable, we performed a comparative analysis by employing Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Squared Error MSE with state-of-the-art frameworks such as:

Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR), Multilayer Perception (MLP), Long short-term Memory (LSTM) using Multiple Linear Regression (MLR) and Stacked Long short-term Memory (SLSTM). The MAE, MAPE, and MSE are used to calculate the prediction error, and the lower the value means, the higher the prediction accuracy. The performance metrics are calculated using Equations.

VI. Results And Discussion

This section discusses the results obtained from various machine learning simulations performed to determine the effectiveness of the models developed for the prediction process.

A. Neural Networks

The validation loss of 23×10^{-5} was achieved and this was in line with the training loss as shown in Fig. 9. To check the model performance, the results of the predicted output were plotted against the test data of the target, see Fig. 10. The linear relationship between the predicted output and the test target data demonstrated a model that was predicting well because the predicted output results were achieved by using the test data of the input features that were not used during training.

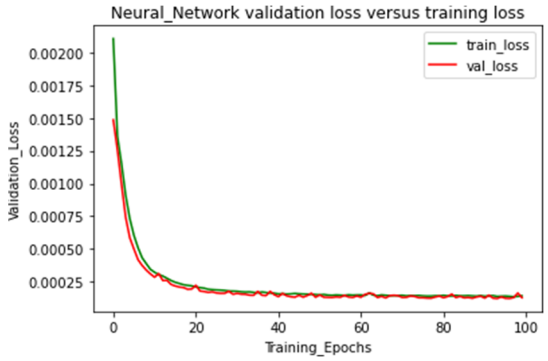


Fig. 9. Training loss compared to validation loss during training.

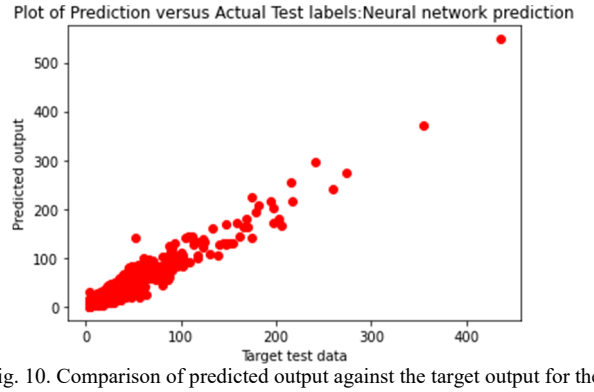


Fig. 10. Comparison of predicted output against the target output for the neural network model.

B. Multiple Linear Regression

Using the scatter matrix plot of Fig. 4, the dataset consisted of highly correlated features. The linear regression model simply suggested that we could not take it as the final model whereas according to the literature, there are models that can oversee the multicollinearity (caused by high correlated features) such as the random forest trees.

However, after experimenting with the Linear regression model on the same data set, a variance score of 0.86 resulted. For a model that was not hyperparameter-tuned, this was not a bad score at all. Next, to check the model performance, the results of the predicted output were plotted against the test data of the target, see Fig. 11. The linear relationship between the predicted output and the test target data demonstrated a model that was predicting well because the predicted output results were achieved by using the test data of the input features that were not used during training.

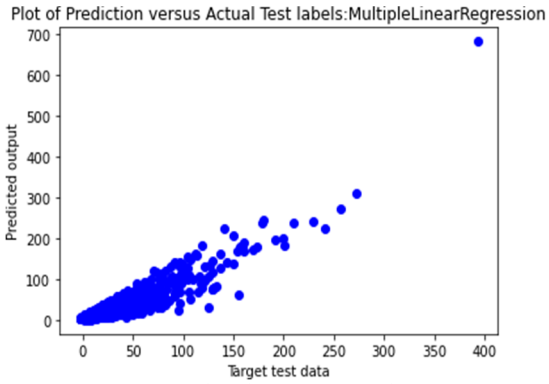


Fig. 11. Comparison of predicted output against the target output for the multiple linear regression model.

C. Random Forest Regression

In comparison with the variance score that was achieved by the linear regression model ("Linear Regression"), the "Random Forest Regressor" model achieved a variance score of 0.93. This proved the fact that the ensembled tree models can oversee the multicollinearity (caused by highly correlated features) problem. Next, to check the model performance, the results of the predicted output were plotted against the test data of the target (see Fig. 12). The linear relationship between the predicted output and the test target data demonstrated a model that was predicting well because the predicted output results were achieved by using the test data of the input features that were not used during training.

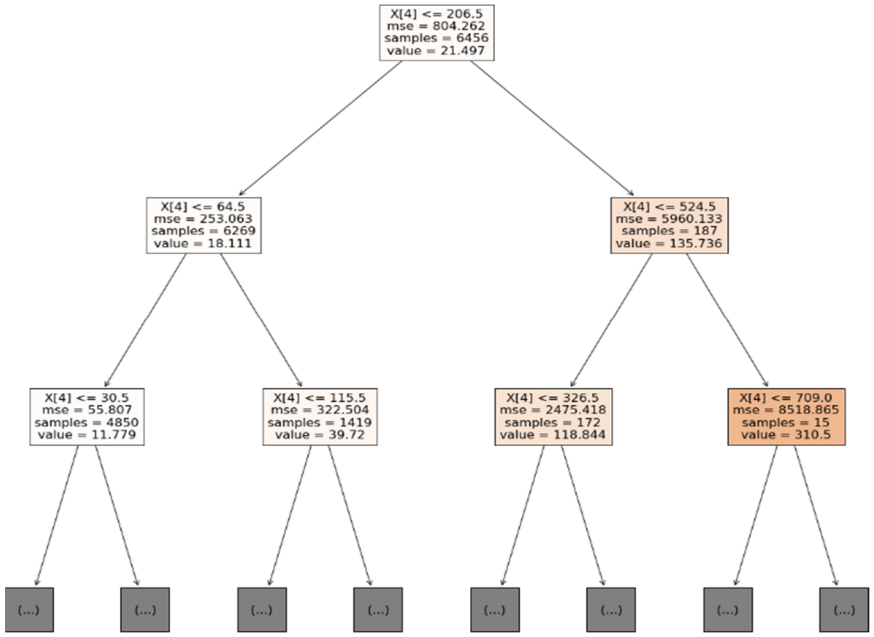
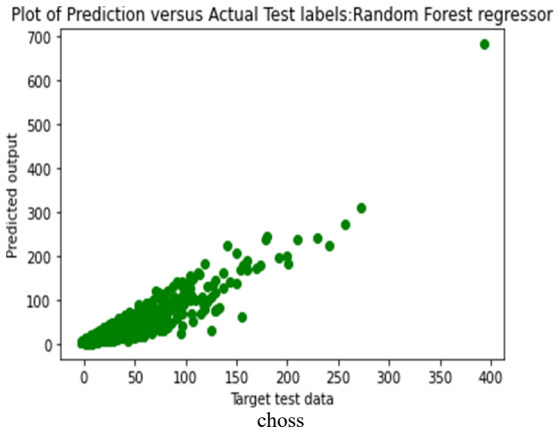


Fig. 12. A first tree estimator extracted from one hundred tree estimators used by the random forest regressor model.



choss

Fig. 13. Comparison of predicted output against the target output for the random forest regressor model.

The "Random Forest Regressor" model used one hundred tree estimators from the default settings of the model. Fig. 13 shows the first tree and the predictions made from the root node to the leaf nodes.

D. Time Series Predictions

Based on the time series data set, Table 4 compares the performance of the three models by comparing the predicted outputs against the actual test target outputs that were reserved during the split of the data for this purpose.

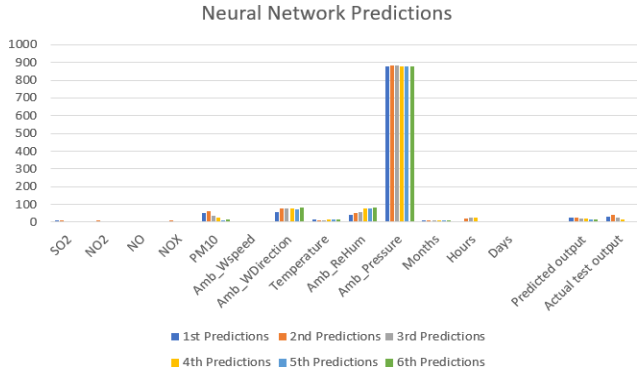


Fig. 14. Neural network predictions based on test data sets of Table 4.

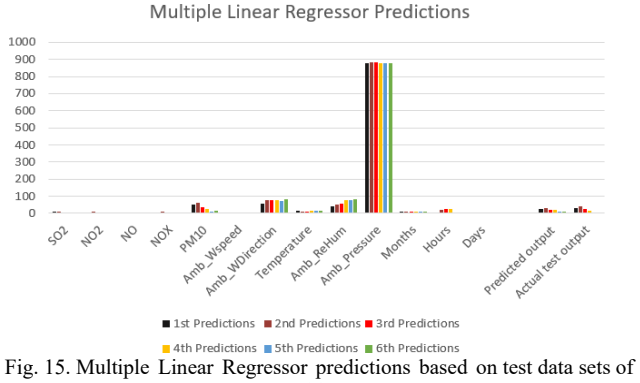


Fig. 15. Multiple Linear Regressor predictions based on test data sets of Table 4.

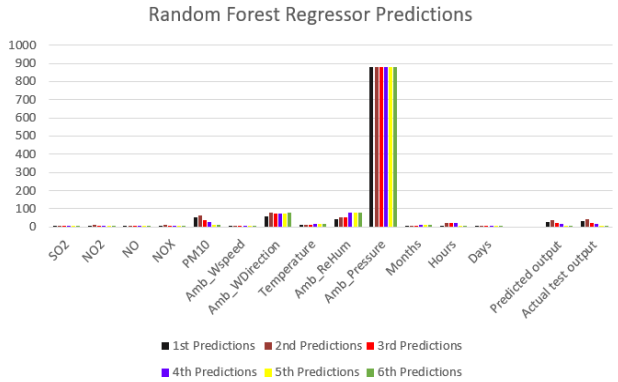


Fig. 16. Random Forest Regressor predictions based on test data sets of Table 4.

The performance results tabulated in Table 4 are visually illustrated using histograms as shown in Figs. 14–16. The "Random Forest Regressor" model outperformed the other two models when the same input features of the test data set were used to make predictions. Since the models were operated at their default parameters and no hyperparameter tuning was made, it is assumed that the models would improve in their performance. For instance, in the case of neural networks, the improvement would include hyperparameter tuning such as considering the network with additional hidden layers, applying the dropout in the hidden layers, and reducing the learning rate. Figs. 17–19 show the code snippets of the results tabulated in Table 4.

Table 4. Time series model predictions using a select few of the testing data sets

SO_2	NO_2	NO	NO_X	PM10	Amb_Wspeed	Amb_WDirection	Temperature	Amb_ReHum	Amb_Pressure	Months	Hours	Days	Predicted Output	Actual test Output
Neural network Performance														
8	6	1	7	53	1.3	56	13	43	880	8	21	6	25.64	31
8	11	1	12	63	1.1	78	11	52	881	8	22	6	27.53	41
7	5	2	7	38	1.2	76	10	54	881	8	23	6	22.39	24
1	2	1	2	27	1	75	17	79	879	12	24	4	20.66	16
1	1	1	1	12	1.2	73	17	78	878	12	1	4	16.56	7
1	1	1	1	13	1.1	81	16	81	878	12	2	4	16.83	7
Multiple Linear Regression Performance														
8	6	1	7	53	1.3	56	13	43	880	8	21	6	24.44	31
8	11	1	12	63	1.1	78	11	52	881	8	22	6	29.78	41
7	5	2	7	38	1.2	76	10	54	881	8	23	6	19.09	24
1	2	1	2	27	1	75	17	79	879	12	24	4	18.29	16
1	1	1	1	12	1.2	73	17	78	878	12	1	4	9.50	7
1	1	1	1	13	1.1	81	16	81	878	12	2	4	10.22	7
Random Forest Regressor Performance														
8	6	1	7	53	1.3	56	13	43	880	8	21	6	28.15	31
8	11	1	12	63	1.1	78	11	52	881	8	22	6	34.94	41
7	5	2	7	38	1.2	76	10	54	881	8	23	6	21.5	24
1	2	1	2	27	1	75	17	79	879	12	24	4	17.07	16
1	1	1	1	12	1.2	73	17	78	878	12	1	4	6.9	7
1	1	1	1	13	1.1	81	16	81	878	12	2	4	7.69	7

```
y1_predict = model.predict(np.array([[8,6,1,7,53,1.3,56,13,43,880,8,21,6]])) # Actual is 31
print("The predicted output is {} and the actual output is 31 ".format (y1_predict))
y2_predict = model.predict(np.array([[8,11,1,12,63,1.1,78,11,52,881,8,22,6]])) # Actual is 41
print("The predicted output is {} and the actual output is 41 ".format (y2_predict))
y3_predict= model.predict(np.array([[7,5,2,7,38,1.2,76,10,54,881,8,23,6]])) # Actual is 24
print("The predicted output is {} and the actual output is 24 ".format (y3_predict))
y4_predict = model.predict(np.array([[1,2,1,2,27,1,75,17,79,879,12,24,4]])) # Actual is 16
print("The predicted output is {} and the actual output is 16 ".format (y4_predict))
y5_predict = model.predict(np.array([[1,1,1,1,12,1.2,73,17,78,878,12,1,4]])) # Actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y5_predict))
y6_predict = model.predict(np.array([[1,1,1,1,13,1.1,81,16,81,878,12,2,4]])) # actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y6_predict))
```

```
The predicted output is [[26.643219]] and the actual output is 31
The predicted output is [[27.536377]] and the actual output is 41
The predicted output is [[22.397385]] and the actual output is 24
The predicted output is [[20.669685]] and the actual output is 16
The predicted output is [[16.566565]] and the actual output is 7
The predicted output is [[16.83355]] and the actual output is 7
```

Fig. 17. Neural network time-series predictions.

```
y1_predict = regressor.predict(np.array([[8,6,1,7,53,1.3,56,13,43,880,8,21,6]])) # Actual is 31
print("The predicted output is {} and the actual output is 31 ".format (y1_predict))
y2_predict = regressor.predict(np.array([[8,11,1,12,63,1.1,78,11,52,881,8,22,6]])) # Actual is 41
print("The predicted output is {} and the actual output is 41 ".format (y2_predict))
y3_predict= regressor.predict(np.array([[7,5,2,7,38,1.2,76,10,54,881,8,23,6]])) # Actual is 24
print("The predicted output is {} and the actual output is 24 ".format (y3_predict))
y4_predict = regressor.predict(np.array([[1,2,1,2,27,1,75,17,79,879,12,24,4]])) # Actual is 16
print("The predicted output is {} and the actual output is 16 ".format (y4_predict))
y5_predict = regressor.predict(np.array([[1,1,1,1,12,1.2,73,17,78,878,12,1,4]])) # Actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y5_predict))
y6_predict = regressor.predict(np.array([[1,1,1,1,13,1.1,81,16,81,878,12,2,4]])) # actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y6_predict))
```

```
The predicted output is [[24.44692444]] and the actual output is 31
The predicted output is [[29.78587615]] and the actual output is 41
The predicted output is [[19.09284263]] and the actual output is 24
The predicted output is [[18.294313]] and the actual output is 16
The predicted output is [[9.50120114]] and the actual output is 7
The predicted output is [[10.22394335]] and the actual output is 7
```

Fig. 18. Multiple linear regression time-series predictions.

```
y1_predict = rf.predict(np.array([[8,6,1,7,53,1.3,56,13,43,880,8,21,6]])) # Actual is 31
print("The predicted output is {} and the actual output is 31 ".format (y1_predict))
y2_predict = rf.predict(np.array([[8,11,1,12,63,1.1,78,11,52,881,8,22,6]])) # Actual is 41
print("The predicted output is {} and the actual output is 41 ".format (y2_predict))
y3_predict= rf.predict(np.array([[7,5,2,7,38,1.2,76,10,54,881,8,23,6]])) # Actual is 24
print("The predicted output is {} and the actual output is 24 ".format (y3_predict))
y4_predict = rf.predict(np.array([[1,2,1,2,27,1,75,17,79,879,12,24,4]])) # Actual is 16
print("The predicted output is {} and the actual output is 16 ".format (y4_predict))
y5_predict = rf.predict(np.array([[1,1,1,1,12,1.2,73,17,78,878,12,1,4]])) # Actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y5_predict))
y6_predict = rf.predict(np.array([[1,1,1,1,13,1.1,81,16,81,878,12,2,4]])) # actual is 7
print("The predicted output is {} and the actual output is 7 ".format (y6_predict))
```

```
The predicted output is [28.15] and the actual output is 31
The predicted output is [34.94] and the actual output is 41
The predicted output is [21.5] and the actual output is 24
The predicted output is [17.07] and the actual output is 16
The predicted output is [6.9] and the actual output is 7
The predicted output is [7.69] and the actual output is 7
```

Fig. 19. Random Forest regressor time-series predictions.

VII. Conclusion and Future Work

This paper has presented a machine learning technique for early air-Pollution detection in a local municipality in South Africa. Real time data were collected from SAAQIS, the local service used for air monitoring within the municipality. The data collected had several discrepancies and the state of air pollution was sparsely represented. Using various machine learning techniques, a section of the data was used to train the models and another set of data was used to predict the occurrence of air pollution in the short term. Furthermore,

various validation methods were used to evaluate the efficacy of the predictions. Based on the results, the random forest regressor model provided better predictions and is recommended for deployment. Future work would involve comparing various models with tuned hyperparameters.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Prof. ED Markus: Supervison. Prof. ED Markus and Ms N Koyana: Conceptulization, software, Writing the first draft. Ms N Koyana: Methodology, Visualization, collecting data, writing review. Dr M Sibiya: Software, analyzed the data, validating. Prof. AM. Abu_Mahfouz: conducted the research, manage. All authors had approved the final version.

References

[1] B. Naude, N. Mandela, and M. L. Sefike, "The rustenburg local municipality air quality monitoring summary report," Compiled for: Rustenburg Local Municipality, August 2019.

[2] G. Gazette and G. Notice, "National environment management: Air quality act, 2004," *Government Gazette Republic of South Africa*, vol. 476, no. 39, February 2005.

[3] T. Meyer and D. Cilliers, "Bojanala platinum district municipality environmental management framework final EMF report," *Centre for Environmental Management*, vol. 11, July 2018.

[4] G. Gazette, "Manual for air quality management planning," *Environ Aff*, no. 39, April 2012.

[5] M. Limb, "Half of wealthy and 98% of poorer cities breach air quality guidelines," *BMJ*, vol. 353, i2730, May 2016. doi: 10.1136/BMJ.I2730

[6] R. N. Phala, "Using an inferential model to estimate dry deposition of SO_2 and NO_X (as NO_2) in lephalale in the waterberg-bojanala priority area," *Academia*, January 2016.

[7] A. Masih, "Modelling the atmospheric concentration of carbon monoxide by using ensemble learning algorithms," *CEUR Workshop Proceedings*, vol. 2298, CEUR-WS, February 2018.

[8] S. R. Shams, A. Jahani, S. Kalantary, M. Moeinaddini, N. Khorasani, "The evaluation on Artificial Neural Networks (ANN) and Multiple Linear Regressions (MLR) models for predicting SO_2 concentration," *Urban Clim.*, vol. 37, 100837, May 2021. doi: 10.1016/J.UCLIM.2021.100837

[9] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 1, pp. 1–19, November 2017. doi: 10.1186/S12889-017-4914-3/TABLES/7

[10] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018. doi: 10.18178/IJESD.2018.9.1.1066

[11] Y. Alyousifi, M. Othman, I. Faye, R. Sokkalingam, and P. C. L. Silva, "Markov weighted fuzzy time-series model based on an optimum partition method for forecasting air pollution," *International Journal of Fuzzy Systems*, vol. 22, no. 5, pp. 1468–1486, July 2020. doi: 10.1007/S40815-020-00841-W

[12] M. L. Bell, J. M. Samet, and F. Dominici, "Time-series studies of particulate matter," *Annu Rev Public Health*, vol. 25, pp. 247–280, 2004. doi: 10.1146/ANNUREV.PUBLHEALTH.25.102802.124329

[13] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A machine learning model for air quality prediction for smart cities," in *Proc. 2019 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2019,* March 2019, pp. 452–457. doi: 10.1109/WISPNET45539.2019.9032734

[14] K. Hu, A. Rahman, H. Bhrugubanda, and V. Sivaraman, "HazeEst: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors," *IEEE Sens J.*, vol. 17, no. 11, pp. 3517–3525, June 2017. doi: 10.1109/JSEN.2017.2690975

[15] R. Martínez-España, A. Bueno-Crespo, I. Timón, J. Soto, A. Muñoz, and J. M. Cecilia, "Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain," *Journal of Universal Computer Science*, vol. 24, no. 3, pp. 261–276, 2018.

[16] J. M. Cecilia, I. Timon, J. Soto, J. Santa, F. Pereniguez, and A. Munoz, "High-throughput infrastructure for advanced ITS services: A case study on air pollution monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 7, pp. 2246–2257, July 2018. doi: 10.1109/TITS.2018.2816741

[17] F. Dominici, A. Mcdermott, and T. J. Hastie, "Improved SEMI-parametric time series models of air pollution and mortality," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 938–948, 2004. doi: 10.1198/016214504000000656

[18] V. Gugnani and R. K. Singh, "Analysis of deep learning approaches for air pollution prediction," *Multimed Tools Appl.*, vol. 81, no. 4, pp. 6031–6049, February 2022. doi: 10.1007/S11042-021-11734-X/METRICS

[19] Y. A. Ayturan, Z. C. Ayturan, and H. O. Altun, "Air pollution modelling with deep learning: A review," *Int. J. of Environmental Pollution & Environmental Modelling*, vol. 1, no. 3, pp. 58–62, 2018.

[20] S. Du, T. Li, Y. Yang, and S. J. Horng, "Deep Air quality forecasting using hybrid deep learning framework," *IEEE Trans Knowl Data Eng.*, vol. 33, no. 6, pp. 2412–2424, June 2021. doi: 10.1109/TKDE.2019.2954510

[21] A. Heydari, M. M. Nezhad, D. A. Garcia, and F. Keynia, and L. De Santoli, "Air pollution forecasting application based on deep learning model and optimization algorithm," *Clean Technol Environ Policy*, vol. 24, no. 2, pp. 607–621, March 2022. doi: 10.1007/S10098-021-02080-5/FIGURES/7

[22] I. Kök, M. U. Şimşek, and S. Özdemir, "A deep learning model for air quality prediction in smart cities," in *Proc. 2017 IEEE International Conference on Big Data, Big Data 2017*, July 2017, pp. 1983–1990. doi: 10.1109/BIGDATA.2017.8258144

[23] Y. Vijaywargiya, "Forecasting of air pollution in united kingdom using deep learning and time series methods," MSc Research Project, School of Computing, National College of Ireland.

[24] D. Iskandaryan, F. Ramos, and S. Trilles, "Air Quality prediction in smart cities using machine learning technologies based on sensor data: A review," *Applied Sciences*, vol. 10, no. 7, 2401, April 2020, doi: 10.3390/APP10072401

[25] N. Shahid, M. A. Shah, A. Khan, C. Maple, G. Jeon, "Towards greener smart cities and road traffic forecasting using air pollution data," *Sustain Cities Soc.*, vol. 72, 103062, September 2021. doi: 10.1016/J.SCS.2021.103062

[26] N. Koyana, E. D. Markus, and A. M. Abu-Mahfouz, "A survey of network and intelligent air pollution monitoring in South Africa," in *Proc. 2019 International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2019*, November 2019, doi: 10.1109/IMITEC45504.2019.9015916

[27] P. Y. Kow, Y. S. Wang, Y. Zhou, I. F. Kao, M. Issermann, L.C. Chang, and F. J. Chang, "Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2.5 forecasting," *J Clean Prod.*, vol. 261, 121285, July 2020. doi: 10.1016/J.JCLEPRO.2020.121285

[28] T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations," in *Proc. IEEE International Conference on Industrial Informatics (INDIN)*, July 2016, pp. 58–63. doi: 10.1109/INDIN.2016.7819134

[29] L. Shikwambana, P. Mhangara, and N. Mbatha, "Trend analysis and first time observations of sulphur dioxide and nitrogen dioxide in South Africa using TROPOMI/Sentinel-5 P data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 91, 102130, September 2020. doi: 10.1016/J.JAG.2020.102130

[30] F. P. Fabianpedregosa, V. Michel, O. G. Oliviergrisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, *et al.*, "Scikit-learn: Machine learning in python gaël Varoquaux bertrand thirion vincent dubourg alexandre passos pedregosa, Varoquaux, Gramfort *et al.* matthieu perrot," *Journal of Machine Learning Research,* vol. 12, pp. 2825–2830, 2011.