ANALYSIS OF AMBIENT AIR QUALITY DATA USING MACHINE LEARNING ALGORITHMS AND LSTM ARCHITECTURE FOR TWO MAJOR TRAFFIC INTERSECTIONS OF KOLKATA CONSIDERING LICHEN AS A BIO-INDICATOR

SAPTAPARNI GHOSH MAJUMDAR

Registration no.: 160400

Exam Roll No.: M4EBT23002

MASTER OF TECHNOLOGY

IN

ENVIRONMENTAL BIOTECHNOLOGY

SCHOOL OF ENVIRONMENTAL STUDIES

FACULTY OF INTERDISCIPLINARY STUDIES, LAW AND MANAGEMENT (FISLM)

JADAVPUR UNIVERSITY

KOLKATA

JUNE 2023

Analysis Of Ambient Air Quality Data Using Machine Learning Algorithms And Lstm Architecture For Two Major Traffic Intersections Of Kolkata Considering Lichen As A Bio-Indicator

A thesis

Submitted in partial fulfilment for the award of the degree of

Master of Technology

In

Environmental Biotechnology

By

Saptaparni Ghosh Majumdar

Registration no.: 160400

Exam Roll No.: M4EBT23002

Under Supervision of

Dr. Anupam Debsarkar

Department of Civil Engineering

Jadavpur University, Kolkata

School of Environmental Studies

Faculty of Interdisciplinary Studies, Law And Management (FISLM)

Jadavpur University

Kolkata

Acknowledgement

I am deeply grateful to the individuals and institutions who have supported me throughout my thesis journey and have contributed to the successful completion of this work.

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr. Anupam Debsarkar and Dr. Rita Saha. Their expertise, guidance, and unwavering support have been invaluable. Their insightful feedback, patience, and encouragement have played a pivotal role in shaping the direction and quality of this thesis.

I am deeply indebted to my senior Ms. Bipasha Dinda, who graciously shared her time, knowledge, and experiences. Her contributions have added depth and credibility to my research, and I am sincerely grateful for her willingness to be a part of this project. I would also like to extend my sincere appreciation to the faculty member of the Department of Geography, Dr. Debjit Dutta for his valuable input and feedback during the various stages of my research.

I am profoundly thankful to my friends and family for their unconditional love, encouragement, and understanding. Their belief in my abilities, constant motivation, and moral support have been a source of strength throughout this journey.

Lastly, I would like to acknowledge my Institution Jadavpur University, Kolkata as it provided required support in laboratory facilities, accessing relevant resources, literature, and research materials that has been immensely helpful.

To all those who have contributed in any way, whether big or small, thank you for your valuable assistance and encouragement. Your support has been vital in the realisation of this thesis.

Declaration

I hereby declared that the work presented in this thesis report titled "*Analysis of ambient air quality data using Machine Learning algorithms and LSTM architecture for two major traffic intersections of Kolkata considering lichen as a Bio-indicator*" submitted to Jadavpur University, Kolkata in partial fulfilment of the requirements for the award of the degree of M.Tech is a bonafide record of the research work carried out under the supervision of Dr. Anupam Debsarkar and co-supervision of Dr. Rita Saha. The contents of Thesis report in parts, have not been submitted to and will not be submitted by me to any other Institute or University in India or abroad for the award of any degree or diploma.

Signature

Saptaparni Ghosh Majumdar

Saptaparni Ghosh Majumdar
(Roll no.: 002130904006)
Environmental Biotechnology
School of Environmental Studies
Jadavpur University, Kolkata

Certificate

This is to certify that the Thesis Report titled "*Analysis of ambient air quality data using Machine Learning algorithms and LSTM architecture for two major traffic intersections of Kolkata considering lichen as a Bio-indicator*", submitted by Saptaparni Ghosh Majumdar (002130904006) to Jadavpur University, Kolkata for the award of degree of M.Tech, is a bonafide record of the research work done by her under my supervision. To the best of my knowledge, the constituents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature

Anupam Debsarkar 19/06/2023

Thesis Supervisor
Dr. Anupam Debsarkar
Department of Civil Engineering
Jadavpur University, Kolkata

Rita Saha

Co-Supervisor
Dr. Rita Saha
Ex. Scientist-'E'
Central Pollution Control Board
Eastern Regional Directorate,Kolkata

TO WHOM IT MAY CONCERN

It is hereby notified that this thesis titled "*Analysis of ambient air quality data using Machine Learning algorithms and LSTM architecture for two major traffic intersections of Kolkata considering lichen as a Bio-indicator*", is prepared and submitted for the partial fulfilment of the continuous assessment of Master of Technology in Environmental Biotechnology course of Jadavpur University by **Saptaparni Ghosh Majumdar (002130904006)**, a student of the said course for session 2021-2023. It is also declared that no part of this thesis has been presented or published elsewhere.

Tushar Jash 14/6/202

(Signature with Seal)
Director
Tushar Jash
School of Environmental Studies
Jadavpur University, Kolkata

Anupam Debarkar 24/06/2023

(Signature with Seal)
Thesis Supervisor
Anupam Debarkar
Department of Civil Engineering
Jadavpur University, Kolkata

Tushar Jash 14/6/202

(Signature with Seal)
Dean
Tushar Jash
School of Environmental Studies
Jadavpur University, Kolkata

CERTIFICATE OF APPROVAL

This foregoing thesis is hereby approved as a credible study of an engineering subject carried out and presented in a manner satisfactorily to guarantee its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not endorse or approve any statement made or opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it has been submitted.

Final Examination for evaluation of Thesis

Signature of Examiners

Abstract

In this study, an attempt was made to consolidate outcomes acquired from monitoring and mapping of lichen with quantitative analysis of AQI data acquired from continuous monitoring stations. Moreover, ***Index of Atmospheric Purity*** was evaluated utilising grid mapping method, in terms of assessment of frequency and cover of foliose and crustose lichens observed on various tree specimens. The wind direction and speed were also assessed in the study area that possessed major traffic intersections nearby. Based on seasonal variation for study period of 2019-2023, windrose analysis was conducted to track correlation among prevalent wind directions and lichen growth. Furthermore, air quality data acquired from CPCB's continuous monitoring station at Jadavpur area were analysed using Machine Learning algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor and KNN Regressor. These models were compared on the basis of performance metrics R^2, RMSE, MSE and MAE in order to suggest an effective model for predicting AQI. It was observed that Random Forest Regressor outperformed other algorithms for predicting AQI values effectively. Moreover, for time series forecasting of predicted data RNN-LSTM model has been utilised as it enables the acquire dependencies on several pollutants. Multivariate LSTM analysis was conducted on 7 significant pollutant concentrations to acquire suitable AQI prediction. Moreover, daily data for a span of 2019 - 2023 was estimated to acquire future forecasting of available data. Although LSTM model depicted higher accuracy, as compared to other ML algorithms the performance metrics were less. Therefore, with higher volume of data the accuracy metrics of the model could be enhanced.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1. Introduction

Kolkata, a rapidly evolving metropolis in the Indian subcontinent has been exposed to significant air pollution levels over the decades. The overall population of Kolkata is over 15 million as of 2023 with a growth rate of 1.3% (Macrotrends.net, 2023). Moreover, with an increase in overall population, certain major contributors to air pollution in the city have constantly enhanced the overall pollution index (Sam, 2021). This involves vehicular pollution in major traffic intersections. The enhanced air pollution levels in Kolkata are primarily attributed to escalated vehicular population, resulting due to an average decadal growth rate for two-wheelers by 70%, four-wheelers by 20% specifically taxis and cars, nearly 14% for three-wheelers and about 10-13% for buses (Chowdhury, 2015). Based on WBPCB, the primary reason behind enhanced vehicular emissions is due to inflated average age of vehicles, surface condition of roads which involves high population density utilising similar road space, diversified vehicular mode including slow-moving vehicles that are accountable for congested traffic and lesser surface area of roads in the city.

According to Das et al., (2006), concerns regarding the rapidly deteriorating air quality in these Indian cities, notably due to air pollution from automobiles given the overall tonnage of contaminants in metropolitan areas, have grown in recent years. Road traffic-related air pollution poses a serious threat to both the environment and public health in today's society. On the other hand, Lelieveld et al., (2015) stated that Kolkata city suffered from significant air pollution. Between 1999 and 2011, the yearly average pollution level, particularly with reference to Respirable Suspended Particulate Matter (RSPM; (PM_{10})) and NOx, did not meet the national limit (CPCB, 2015). In recent years, yearly average NOx and PM_{10} levels had routinely exceeded the NAAQS, while wintertime $PM_{2.5}$ levels have been shown to be far higher than the requirement (Majumdar et al., 2020). Along with those from numerous small-scale enterprises, Kolkata is subjected to the daily emissions of 1.14 million vehicles (CPCB, 1999). The regular occurrence of temperature stratification and the relatively low wind speed during the winter season trap pollutants close to the ground, making the issue of air pollution more severe at this time of year (Majumdar et al., 2009). The national ambient air quality criteria are considerably exceeded by the average concentration of respiratory suspended particulate matter (RSPM) in this region.

1.2. Background of the Research

In this study, lichen has been utilised as a biomonitoring agent to detect overall air pollution persisting in urban environments specifically near traffic intersections. Monitoring biodiversity changes offers a practical way to assess the condition (quality) of an ecosystem. Loppi (2014) states that lichen species that are found in a given area exhibit measurable responses to environmental changes, and lichen biodiversity counts can be used as accurate indicators of the state of the environment, with high values signifying unpolluted or low-polluted conditions and low values signifying polluted ones. On the other hand, Bealey (2008) stated that lichens are bio-monitored in terms of detecting transformation in community composition, physiological status and constituents of trace elements, delivering effective evidence for temporal and spatial trends. In urban and industrialised settings, where the high density of various emitting sources makes monitoring air pollution with traditional physicochemical techniques very challenging owing to the diversity of contaminants, using lichens as biomonitors of air quality may prove to be highly effective (Loppi, 2019). Moreover, epiphytic lichens are appropriate organisms for monitoring overall air pollution in a specified area. Transformation in physiological parameters and diversity of lichens are utilised as indicators of environmental concerns in a significant region (Majumder et al., 2013). These organisms possess the ability to flourish in various forms of substrates and are outlined from different climatic conditions (Gupta et al., 2016).

Furthermore, forecasting air pollution is particularly helpful in letting policymakers know the amount of pollution that would enable them to take action to lessen its effects. To prevent high pollution incidents, for instance, traffic limits could be put in place. The Air Quality Index (AQI) is often used to represent the degree of air pollution. The following pollutant concentrations have a piece-wise linear relationship with AQI: ozone (O_3), particulate matter ($PM_{2.5}$, PM_{10}), sulphur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2) (Kumar & Pande, 2022). With an increase in industrial and motorised growth, monitoring and forecasting AQI, particularly in metropolitan areas, has become an essential and difficult undertaking. Although the concentration of the deadliest pollutant, $PM_{2.5}$, is found to be multiplied in developing countries, most studies and research focused on air quality targets in developed nations (Rybarczyk and Zalakeviciute 2021).

Through this study, a significant correlation has been drawn based on growth of lichen observed in study area with respect to prevalent air quality data acquired from continuous air quality monitoring by CPCB in major traffic intersections of Kolkata region.

1.3. Problem Statement

Degradation of ambient air quality in the Kolkata region has been a significant concern over the past years. Increasing population in urban centres and anthropogenic activities with transforming utilisation patterns of land have led to changes in local environmental aspects of this city. Significant pollutants that impact the overall air quality in this region are $PM_{2.5}$ (Particulate matter of size less than 2.5μ), PM_{10} (Particulate matter of size less than 10μ), NO (Nitrous Oxide), NO_2 (Nitrogen Dioxide), NOx, SO_2 (Sulphur Dioxide) and Ozone. A primary factor associated with an increase in air pollution levels in Kolkata is transportation. Enhanced utilisation of petrol fuel, ample amount of vehicles that are poorly maintained, and ineffective supervision is creating transportation as a significant air polluting segment. Moreover, due to the high amount of pollution caused by cars stopping and starting at traffic intersections, some parts of the road are seen as extremely important. As a result, intersections harm fuel combustion compared to other kinds of road components. The longer time spent in traffic, the more gasoline is used, which leads to higher fuel consumption and higher vehicle emissions of nitrogen oxides, carbon monoxide, carbon dioxide, and hydrocarbons (NOx).

Based on reports from the government of India, the problem has gotten worse due to the cities' dense car populations and high motor vehicle-to-population ratios (Loksabhadocs.nic.in, 2023). The increase in vehicle counts every year hinders appropriate implementation of measures for degrading vehicular emissions in traffic intersections. Furthermore, lack of proper land use planning during urban area growth results in increased car usage and strong population migration to urban areas because there are more tall buildings in cities, vehicle emissions get stuck at ground level and don't move around as well as they should. These aspects are addressed in order to decrease the overall pollution index in air and implement measures to improve the overall air quality due to distinct factors in study area. Moreover, lichen species are deployed as a bioindicator of air pollution in this segment as loss of biodiversity due to increasing levels of pollution is a significant aspect of climate change. Further, Carreras & Pignata (2007) illustrates that this symbiotic organism possesses the ability to assemble elements to an extent that exceeds their physiological requirements and has enabled them to serve as persistent pollutant accumulators and sensitive indicators. Thus, availability of lichen species in urban areas becomes challenging due to increased pollution levels, although in areas where modernised instruments could not be employed for tracking air quality these bioindicators provide a comprehensive idea of air quality based on their growth patterns and variety.

1.4. Significance of the study

The findings of this study would redound to the benefits of the society and further research in this segment. This research utilises machine learning algorithms and Long Short-Term Memory (LSTM) architecture to analyse and predict air quality data. This approach would provide a more accurate and reliable assessment of air quality compared to traditional statistical methods. Moreover, traffic intersections are known to be hotspots for air pollution due to vehicular emissions. This study focuses on Kolkata's traffic intersections and further provides insights into the air quality in these areas. It would further assist in deployment of effective mitigation strategies by analysing the issues with governing parameters. Utilising lichen as bioindicator in this study assists that these species are sensitive to air pollution and could be utilised as bioindicators to monitor air quality. Furthermore, it would also provide a cost-effective and efficient method for air quality monitoring. Therefore, this research area would provide significant policy implications for improving air quality in Kolkata's traffic intersections. The utilisation of ML algorithms and LSTM architecture would assist in fabrication of predictive models for air quality, which could provide policymakers with effective strategies. Overall, this research has significant implications for environmental monitoring, air quality management, and policymaking.

Chapter 2: Literature Review

2.1. Lichen as a biomonitor for air pollution

Lichen is a significant biological component in indicating and monitoring environmental quality, specifically air pollution levels. According to Abas (2021), certain lichen species are observed to possess a significant relationship with basidiomycete yeasts. Moreover, lichens do not possess defensive tissues and this enables them to absorb nutrients, gases and water directly from the environment. These species have been utilised as a bioindicator for detecting air quality since 1866, in which epiphytic lichens were utilised to analyse air pollution levels in local regions. On the other hand, Kuldeep and Prodyut (2015), state that transformation in lichen communities and loss of their diversity due to urbanisation, transformed climate and air pollution have been significantly witnessed in cities of Kolkata and Bangalore. Moreover, unregulated lichen harvesting has turned out to be a major hazard in Western Ghats and Himalayas. Knops *et al.*, (1991) state that lichens are significant organisms to capture nutrients from occult precipitation, wet deposition, impaction, gaseous uptake and sedimentation. Major nutrients captured through these procedures illustrate fresh nutrient inputs that would not be seized through the ecosystem. A segment of these nutrients is introduced in lichen biomass and is available upon decomposition and death, although a portion gets leached through soil surface deposition.

Based on research by Sujetovienė and Česynaitė, (2021), it was identified that transformation in assembling heavy metals and eco-physiological parameters in lichens transferred to a shooting range environment. It further states ***"Thalli"*** of epiphytic lichen were transported from a non-polluted area to a shooting range. ***"Thiobarbituric acid reactive substances"*** (TBARS), Chlorophyll fluorescence and destruction to cell membranes in lichen *Ramalina farinacea* and *Evernia prunastri* thalli were identified after 3 months of exposure indoors in a specified shooting range. The concentration of certain heavy metals including copper, cadmium, manganese, iron, lead, nickel, zinc and antimony evaluated in lichens depicted air pollution levels in a specified environment. It further portrayed stress symptoms as the existence of metal pollutants leading to integrity loss in cell membranes of lichens and bringing oxidative stress as an outcome of an enhanced level of TBARS. On the other hand, mosses and lichens are cryptogamic organisms that occur in all terrestrial ecosystems and possess the ability to tolerate prolonged drought which could further colonise areas with severe environmental conditions (Adamo et al., 2003). Due to its high surface: volume ratio, basic anatomy and deficiency of cuticles enable them to consolidate heavy metals, centralising them in tissues. Therefore, this ability enables them to depict an elemental composition that is demonstrated over dissolved gases, atmospheric metal ions, particulate matter and prolonged duration.

2.2. Diversity distribution of lichens in Kolkata region

Lichen diversity distribution is utilised as it is the cheapest form of air pollution evaluation mechanism. Surveying and mapping lichen diversity is inexpensive and does not require high-end costs like chemical analysis, permitting cost reduction and complicated sample gridding (Abas & Awang, 2017). Based on the studies of Mikhaylov (2020), lichen diversity distribution was primarily monitored in industrial and urban areas. Moreover, through mapping lichen diversity it is identified that greater diversity of lichen in a specified area depicts less pollution in the region. Several indices' values are utilised in studies to evaluate air pollution in terms of lichen diversity values for evaluation of air pollution in industrial areas. However, community diversity analysis is utilised in various cases for biomonitoring of lichen which involves segregation of lichen community into three segments that are moderate, tolerant and sensitive species (Sujetovien ̇e, 2017). Furthermore, surveys are administered to analyse growth forms in lichen that includes crustose, fruticose, and foliose that provide a value and higher value acquired from survey illustrates superior air quality (Dathong, 2016).

According to new records published in recent years, only 25 species of lichen were found in the Botanical Garden and surrounding areas of Kolkata (Upreti et al. 2005). In 1865, about 53 species of lichens were described as occurring in Kolkata (Nylander 1867). This demonstrates the severity of the lichen biota loss in Kolkata, which is presumably a side effect of the city's rising pollution. Since a long time ago, air pollution in urban environments has been frequently detected using cryptogams, particularly mosses and lichens (Conti & Cecchetti 2001). However, some lichen types that grow on tree bark are exposed directly to the atmosphere and receive nutrients only from air or precipitation, in contrast to lichens that grow on rocks or soil, which have the potential to interact with the substratum in addition to air. Thus, these epiphytic lichens are particularly well suited for air quality monitoring. Stress caused by the environment may be detected in a specific location by looking at the variety of epiphytic lichens and changes in physiological parameters (Paoli & Loppi 2008).

2.3. Classification of lichens in terms of air quality

Lichens are successful and unique alliances that constitute a significant biodiversity component and transpire on various sub-strata which involve twigs and barks of trees, leaves, soil and rocks in appropriate climatic conditions ranging from alpine to tropical regions and polar territories of Antarctica. Moreover, in India, these species are depicted in terms of 2532 species under 78 families and 324 genera involving 21.3% endemic species. Lichens depend on other forms of an organism for their survival (Bsienvis.nic.in, 2022). Organisms that assist lichens to survive involve photo bio-ant-like cyanobacteria or green algae and certain types of fungi. These organisms generate a synergetic association among themselves for their survival (Kuldeep and Prodyut, 2015). Certain forms of lichen possess extremely complicated structures as it comprises two organisms that are associated synergistically. Thus, various pollution stages are analysed by lichens as they possess capability of surviving for extremely long tenures. Moreover, lichens respond gradually to natural environmental transformations and are a significant indicator of sulphur oxide found in atmosphere. These species are able to detect overall pollution as they consume necessary nutrients and water from atmosphere and not soils (Preethaa., 2021). It also possesses ability to tether pollutants that are absorbed in fungal threads and possesses an enhanced ability to analyse radioactive metals and toxic elemental pollutants in surroundings.

According to Cobianchi (2003), lichens and mosses are predominantly scarce in urban areas and due to such reasons a "bag technique" has been fabricated for monitoring air pollution in city. These bags comprise a grid or mesh and contain water-washed lichens or mosses. It possesses significant advantages that involve uniformity of exposure period and entrapment surface, flexibility in several stations and site selection which illustrate original contaminant concentration in biomonitor and superior collection efficiency of most elements. Moreover, Tiwari (2008) states that lichens are extremely sensitive to a major variety of pollutants and significant factors according to their sensitivity. Gas and water are exchanged over complete lichen thallus. It further illustrates that during dehydration various contaminants and nutrients are majorly concentrated through conversion to slow-releasing forms which involve cloistered inside organelles, crystallised between cells, and being absorbed into cell walls. Moreover, during heavy rains pollutants and nutrients are leached gradually. Thus, a dynamic equilibrium persists among atmospheric pollutants and nutrient loss and accumulation, which evaluates lichen as a sensitive tool for a depiction of transformation in air quality.

Lichens exist primarily in four forms which include Fruticose lichens, Foliose lichens, Crustose lichens and Squamulose lichens. Based on air quality and sulphur dioxide SO2 concentration lichens are classified as, **Extremely clean air:** Sensitive species such as Usnea, Lobaria and Ramalina, **Clean air:** Fruticose and Foliose lichen, **Moderate air:** Potential and Squamulose lichen, **Polluted air:** Crustose lichen such as Lecanora, **Heavily polluted air:** No lichen only green algae. The mapping of lichen biodiversity is a common practice in many nations since it provides a gauge of the air quality based on the biological effects of air pollution. Also, there is evidence that lichens react swiftly to changing concentrations of air pollutants, even though this bioindicator needs a time lag for monitoring changes in air quality that must be long enough to account for community changes in species number and composition (Asta et al., 2002). Monitoring the variety of lichens is affordable and yields data that may be used to predict human health. Awasthi, (1988) states that lichens have developed effective systems for absorbing nutrients from their surroundings. Anions like nitrate and sulphate, have an active uptake mechanism in which the algal partner expends metabolic energy to take up these anions, which accumulate inside the cells. Lichens may trap minute pieces of rock, earth, or any other heavy metal contaminants inside their structure and absorb metal ions like Ca^{2+} through ion exchange (Negi, 2003). Some of the metabolites that lichen species break down these particles and release nutrients that may subsequently be absorbed by the lichen's cells.

2.4. Physiological and morphological transformations in lichen

Based on studies of Bokhorst (2015), lichen species play a significant factor in nutrient cycling and provide a basal trophic level for varied animals, including nematodes, gastropods, reindeer, mites and springtails. Lichen species indicate substantial variation in physiological and morphological traits that would possibly affect community composition. Moreover, lichens comprise cyanobacteria that enable fixation of atmospheric nitrogen (N) which is imparted to fungal tissue and thus, possess significant imputation for lichen thallus Nitrogen concentrations.

Lichens are extremely sensitive organisms and susceptible to environmental transformations that illustrate them as the most sensitive component in the ecosystem (Munzi et al., 2014). Thus, these organisms are treated as ecological indicators and their sensitivity has been utilised for mapping air pollution levels and also for estimating deposition and decay in SO_2. Green (2011) states that these species are poikilohydric and are unable to control their water content which varies to maintain equilibrium with adjacent environments. The water dependence externally controls physiological activity of lichens as damp environmental conditions activate and hydrate them in arid environmental conditions and it results in inactiveness (Green, Sancho & Pintado 2011). Aptroot & Van Herk (2007) states that lichens are considered to be slow-growing organisms and their dependence on atmosphere permits them to acknowledge climate change for 5 years in terms of sensitivity to climate change.

Morphological Performance and lichen morphology involve growth and this has been related to atmospheric sulphur dioxide and the prevalence of acidic precipitation (Seaward, 1976; Wolterbeek, 2003). Growth of lichen virtually ceases in winter season and there might be a substantial difference in growth rate of lichen species in winter and summer in temperate latitudes. According to Palomäki (1992), it has been observed that measurable growth impacts from ozone and fluoride have been evaluated to generate non-structural changes. On the other hand, Piervittori et al. (1997), stated that morphological transformations were identified due to acidity as lichen thalli get exposed to simulated acid rain generated due to progressive depletion of exterior amorphous layer. There is an indication of substantial specificity along with certain functional roles in the lichen symbiosis in the highly varied bacterial microbiome of lichens (Aschenbrenner et al., 2016). However, it has been proposed that the lichen symbiosis is distinct in that when grown in axenic culture, the mycobiont yields morphological characteristics that are neither known from other fungi nor from the mycobiont itself in order to form the "greenhouse" for the photobiont (Honegger, 2012). Moreover, Sanders (2001) states that due to this, mycobionts in axenic culture demonstrate completely distinct phenotypes from lichenized thalli, and the lichen architecture is only apparent when both bionts are present. Through this architecture, lichens behave in their surroundings as if they were a single creature, emulating plants in terms of both function and morphology. While considering cephalodia and so-called photosymbiodemes, in which a given fungus forms different structures or entirely different lichens when connected with either green algae or cyanobacteria, this morphological differentiation in the form of a photobiont is particularly striking.

Moreover, when lichens are exposed to pollutants it leads to stress production of ethylene, which generates electrolyte leakage as an outcome of cell membrane disruption, further increasing chlorophyll degradation and diminishing ATP content (Silberstein et al., 1996). Thus, it leads to degradation in photosynthesis, respiration and N2 fixation. Majumder (2013) states that these responses are extremely essential in terms of stress indication and further assist in spotting early indications of changing environmental circumstances.

2.5. Impact of vehicular emission on air quality

Air pollution involves a transformation of chemical, biological and physical properties in air. The transformation in air is majorly due to a variety of anthropogenic and natural activities. The emerging trends of metro cities pilot heavy manoeuvring of traffic, various constructional projects and aircraft. Thus, vehicular emissions with increased population leading to congested roads and major public transportation systems impact health standards of local people (Chakraborty, 2014). The air emitted from petrol as well as diesel run engine at the time of functioning of public transport plays a pivotal role in generation of traffic air pollution in cities. Traffic intersections in Kolkata city and their impact on ambient quality have been a significant area of concern as it affects various residential and commercial places with high-level traffic volume. Based on research by Chakraborty (2014), the primary factors associated with governing traffic intersections and vehicle emissions to evaluate overall air pollution in a specified area involve evaluation of particulate matter concentrations that adversely impact health of public. Ancient polluting vehicles have clogged major roads and traffic intersections of Kolkata city and enhanced levels of diesel vehicles operating in this city lead to increased levels of air pollution. Based on reports of WPCB and CPCB, a significant quantity of energy emissions is primarily due to CO_2 emission energy which has been observed in this city along with increased levels of particulate matter on a daily basis. Spiroska (2011) states that particulate is related to mortality and significance of sub-micron particles especially PM_{10} which are primarily contributed by motor vehicles. Certain industrial pollutants and automobile exhausts comprise NO_2 which through photochemical reaction generates O_3 and impacts allergic asthmatics through augmentation of allergic responses.

Based on the study of Kumar (2021), the transportation industry contributes significantly to air pollution; with the rise in the number of cars, the contaminants generated by vehicles also increase. One of the most important pollutants emitted by cars, which are generated by burning fossil fuels, has been identified as CO_2. Increased use of motor vehicles raises atmospheric CO_2 levels, which has the greenhouse effect and contributes to global warming. Recent studies on air pollution and vehicle emissions show that about 66% of air pollution comes from ground-based transportation, which is mostly made up of dangerous emissions from cars, trucks, and motorcycles (Siew et al., 2008). When the usual flow of traffic is disrupted and stopped by unforeseen delays, accidents, breakdowns, and unfavourable weather conditions, the air pollution problem becomes more acute. Such circumstances cause the normally smooth flow of vehicular traffic to be disturbed, particularly at crossroads, junctions, traffic lights, and accident locations (Anjum et al., 2019). The significant change in the air quality index is caused by this accumulation of automobiles as well as road conditions and traffic patterns (AQI).

2.6. Prediction of Air Quality Index Using Machine Learning Models

According to Mahalingam et al., (2019), Artificial Intelligence and Machine Learning play a significant role in a broad variety of critical implementations that involve image processing, expert systems and data mining. Moreover, in this case Support Vector Machine a supervised machine learning algorithm has been utilised for regression and classification analysis. It also utilises neural network algorithms such as ANN for predicting AQI values and eventually feeds the acquired data into SVM for assessing accuracy of predicted data. Various kernel functions were taken into consideration in SVM algorithm and maximum accuracy was obtained from ***'Medium Gaussian SVM'*** of 97.3%. On the other hand, Kalapanidas (2001) states that air pollution possesses significant impacts on meteorological features that involve wind, precipitation, temperature, humidity and solar radiation. It is further signified through utilising a case-based reasoning system with respect to different air pollution levels (alarm, high, med and low). Moreover, Khan et al., (2022) utilised index-based seasonal machine learning algorithms for predicting concentration of $PM_{2.5}$ in majorly polluted settings on hourly and daily scale.

Random forest and linear regression models have been deployed that illustrate satisfactory performance for daily and hourly $PM_{2.5}$ datasets possessing various land uses. Li et al., (2016) analysed spatial stability of STDL model and predictive performances of Guanyuan station and Zhiwuyuan station. It illustrated Guanyuan station R^2 value in testing phase depicted that 98.24% of explained variance has been analysed in this model. Furthermore, Zhiwuyuan station possessed a higher mean absolute percentage error value by 25% and highest relative error as this area was located in an urban area so only traffic pollutants were dominant.

In the studies of Nandini & Fathima (2019), K-mean algorithm, Decision Trees and Logistic Regression are analysed for prediction. Certain pollutants were considered while execution of model. Moreover, error and accuracy of the model are evaluated in terms of correct labelling of data with respect to AQI levels. It was observed that among implemented models' logistic regression performed efficiently as compared to decision tree models possessing an error rate of 0.44 and 0.66. On the other hand, Londhe (2021) analysed five distinct predictive models on 3 datasets. Additionally, for predicting AQI values, it has taken into consideration techniques of dimension reduction, among which the best suited model was PLS with ***"Leave One Out Cross Validation"*** and it considered fifth component from each model. Models were implemented on station-wise data of Indian cities for predicting and classifying AQI models, in which KNN model possessed repeated CV, possessing a tuning length 10 was considered for AUC and accuracy. According to proposed model assists in predicting time series analysis of AQI in sophisticated scenarios. Algorithms of dynamic decomposition in terms of CEEMDAN, EEMD and EMD are fabricated through utilisation of time window sliding. Sharma (2021) stated that association among AQI and corresponding factors impacting AQI were analysed utilising Artificial Neural Networks that illustrates non-linear association among AQI and contributing variables although prediction accuracy was less favourable. It depicted that Decision Tree Classifier provided better accuracy of 99.7% as compared to other models. Moreover, ARIMA models were utilised for time series forecasting, the overall computational time and prospective of implementing non-linear ML methods provided insights on accuracy of model.

2.7. Research Gap

An appropriate assessment based on the particulate levels found in major traffic intersections is essential to diminish unfavourable impacts on air quality aspects. Thus, effective real-time analysis of measured concentrations of $PM_{2.5}$ and PM_{10} and further predicting the projected levels would enable it to decrease its impact in future. Moreover, biological species such as lichen also assist in providing a significant idea regarding concentration of NO_2 and SO_2 in ambient air and these species are dependent on-air quality for their survival. Therefore, assessing the accumulation of heavy metal concentration in lichen would provide measures to diminish metal concentration levels in order to enhance the sustenance of lichens in atmosphere. Limited studies have been conducted in this area undertaking a comparative analysis between real-time monitoring of air quality data and further predicting the AQI levels through implementation of Machine learning algorithms.

Chapter 3: Objectives and Scopes of the study

3.1. Research Objectives

The research objective of this study is illustrated below:

- *To evaluate the impact of degrading ambient air quality in traffic intersections of Kolkata region on diversity rate of lichen species based on various lichen genera explored in study area.*
- *To analyse past 4 years CPCB data on air quality of the specified region with generic forecasting of air quality index using Machine Learning algorithms.*

3.2. Scope of the research

The scope of this research are discussed below:

- This study identifies the prevalent air quality scenario in major traffic intersections of Kolkata by utilising lichen as a bioindicator of environmental pollution.
- It takes into consideration prior research in terms of lichen adaptability in areas which are exposed to high pollution levels.
- The impacts on overall air quality of Kolkata region from rising pollution levels at major traffic intersections were examined from viewpoints of prior research and documentation.
- This study was limited to nearly 20 tree specimens from Jadavpur area of South 24 Parganas, in which presence of lichen was detected.
- The area was considered to capture the impacts of vehicular pollution on air quality, as two major traffic intersections of the city are located here.
- Overall growth of lichen in this region was thoroughly evaluated from 2022-2023, and segmented on the basis of growth achieved in various seasons due to differences in seasonal pollution levels.
- Secondary databases of air quality were utilised and obtained for this study from the ***"Central Pollution Control Board"*** website.
- Based on the air quality data prediction modelling has been conducted using machine learning algorithms to assess the future impacts of prevalent pollutant concentration.

- A significant comparison was drawn from pollutant concentration observed through continuous air quality monitoring and pollutant concentration in lichen thallus.

For this study, Jadavpur area traffic intersection has been taken into consideration as air quality in this region is significantly polluted and availability of lichen was observed in extremely marginal quantities.

Chapter 4: Methodology

4.1. Study Area

The investigated site considering major traffic intersections nearby to detect impact of vehicular pollution was chosen at Jadavpur (22.4955° N, 88.3709° E), Kolkata, India. The air quality data of Jadavpur area was acquired from CPCB's continuous air quality monitoring station positioned at CSIR (Indian Institute of Chemical Biology). While conducting this study, the ambient temperature of this region ranged between 30 to 35°C in summer; and between 21 to 27°C in winter. Average wind speed was observed to be 3.5–5.8 m/sec.

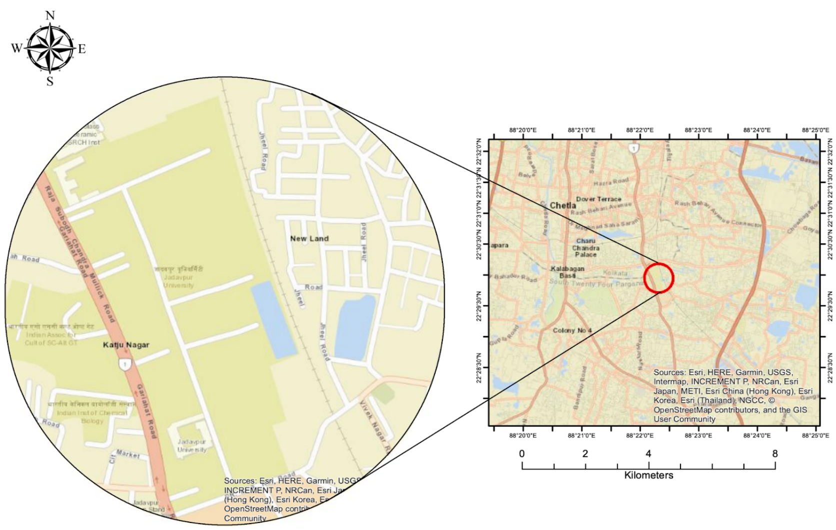


Figure 4.1: Study area of major traffic intersection near Jadavpur University

Table 4.1: Location, distance from road of area in which samples were found, and distance from mid-city of study area has been illustrated

Locations	Distance from road (Km.)	Distance from mid-city (Km.)	Remarks
Jadavpur area (Two significant traffic intersections of Jadavpur Thana and 8B Bus Stand)	0.04	9.8	Semi-exposed areas with major vehicular activity.

4.2. Sampling Technique

Lichen samples were gathered from Jadavpur area adjoining two major traffic intersections (22°20′N, 88°22′E) in Eastern segment of India from June 2022 to April 2023. Various tree specimens in which lichen growth was observed have been illustrated in the table below along with their geo-coordinates. Lichen species were predominantly collected from barks of tree species found in this region. Dominant tree specimens observed in this site were Royal Palm (*Roystonea regia*). The crustose and foliose variety of lichen were majorly found in this location which involved Caperat lichen (*Parmelia caperata*), Wreath lichen (*Cryptothecia subnidulans*) and *Diploicia canescens* on tree barks.



Figure 4.2: Images of Caperat lichen (Parmelia caperata), Wreath lichen (Cryptothecia subnidulans)

4.3. Biomonitoring of lichen species

Biomonitoring of lichen species were conducted as they are both sensitive and perennial. Transformation in environmental parameters was effectively analysed as lichens are subjected to elevated levels of pollution. For biomonitoring, individual lichens were tracked on a gap of 15 days for a year. Tracing of individual lichen species from various trees found in Jadavpur University area was done to analyse the overall growth of lichen in various seasons concerning prevalent air quality. The tracings were further placed in millimetre graphs to assess the surface area in terms of mm^2. Furthermore, each lichen tracing was scanned in 1:1 scale and digitised using QGIS software through georeferencing and converting raster data to vector data. The area of digitised lichen samples was calculated further using Digimizer 6 software. Moreover, by comparing the tracing data over the year the overall growth rate of lichens could be assessed and after scanning the tracings the number of pixels is counted that are interior to the tracings. This enables and acquires a fairly accurate rendition of the amount of area that a lichen occupies. Furthermore, correlation of size of the lichen with its age enables establishment factors that when a lichen is observed based on its size, its approximate age could be predicted. Lichens possess an extremely dynamic population and scenarios are also equivalently dynamic as a lot of lichen species are born every year and a lot of lichen species die every year. These are extremely slow-growing organisms as they work on a time scale of hundred years and they tend to swiftly adapt to change and take advantage of the change.

Field method considered for monitoring and recording data of epiphytic lichen species in terms of its stress factors for various trees found in Jadavpur area, predominantly Royal Palm species (*Roystonea regia*). Five substantial trees exposed to more or less comparable environmental circumstances were chosen for each location. These trees' trunks had lichens on them up to a height of 2 metres, which were gathered and recorded. The percentage of trees at a given site on which a species of lichen was physically present was used to compute frequency of occurrence (Fr), and the surface area occupied by the species at that location was used to calculate coverage (Co). A translucent sheet was used to trace the outline of each individual lichen patch, and graph paper was used to calculate its surface area. The total surface area of all the individual patches present at that site made up the coverage of any species (Das et al., 2013). By multiplying frequency (Fr) and coverage (Co), the frequency coverage (f) of each species was calculated and expressed as a number on a scale of 1 to 5. (LeBlanc et al. 1974).

4.4. Index of Atmospheric Purity

The sampling was done for an area of 2 kilometres mostly within Jadavpur University campus area and nearly 20 trees were approached from this site. This site was close to a continuous monitoring station by CPCB to record the air quality in terms of significant parameters. In various compass directions the abundances of lichen species were calculated and based on these results lichen indices were calculated in terms of lichen abundance and ecological sensitivity of individual species.

$$IAP = \sum_{1}^{n} \frac{(Q \times f)}{10}$$

Where, n = a number of species, Q = factor of accompanying species, f = cover and frequency of each species, IAP = Index of atmospheric purity.

Index of atmospheric purity enables monitoring of significant air pollutants like NOx and SOx, and concentration of heavy metals like Zinc, Ni, Cu, Mn, Cd, Pb, and Cr. It further assists in zoning and mapping regions adjacent to pollution genesis with a gradient. This method is primarily utilised for bryophytes and lichens that further enables an examination of pollution gradients across a line or belt transect. The abundance and frequency of species can be taken into consideration through the quantity of transects diffusing in various directions from the transect line and with enhanced distance from the pollution source (Das et al., 2021). Thus, it illustrates a significant picture of the long-term effect of pollution levels at a specified site.

4.5. Grid mapping technique for assessing lichen coverage

Sampling technique implemented in this segment was based on the methods illustrated in a study by Muhammad (2018), in which tree specimens were assessed through "***Index of Atmospheric Purity"*** (IAP) methodology. It is based on frequency, tolerance and number of lichen available in study area. In this segment, Royal Palm (*Roystonea regia*) species were selected as host trees that possess a circumference higher than 60 cm and are located at a distance of 5 metres from vehicular traffic movement. For our study, nearly 20 tree specimens were taken into consideration from Jadavpur University and a grid subdivided into 10 rectangles with size of 30 X 50 cm. Each rectangle comprised lichen species which were significantly identified and measured. Quantity of lichen available in each rectangular segment was analysed. This grid framework was placed over tree bark and was measured in terms of its height from the bottom of the tree. Amount of vehicular movement in adjacent areas were studied based on vehicles in number passing during peak hours. The primary pollutants in air were evaluated through data acquired from continuous monitoring stations and the overall exposure level was assessed. Moreover, depth of bark crevice (mm) and diameter of tree at breast height (cm) were measured. Based on study of Mežaka (2008), depth of bark crevice was evaluated at a height of 1.20m and significant tree bark samples were collected for pH measurements.

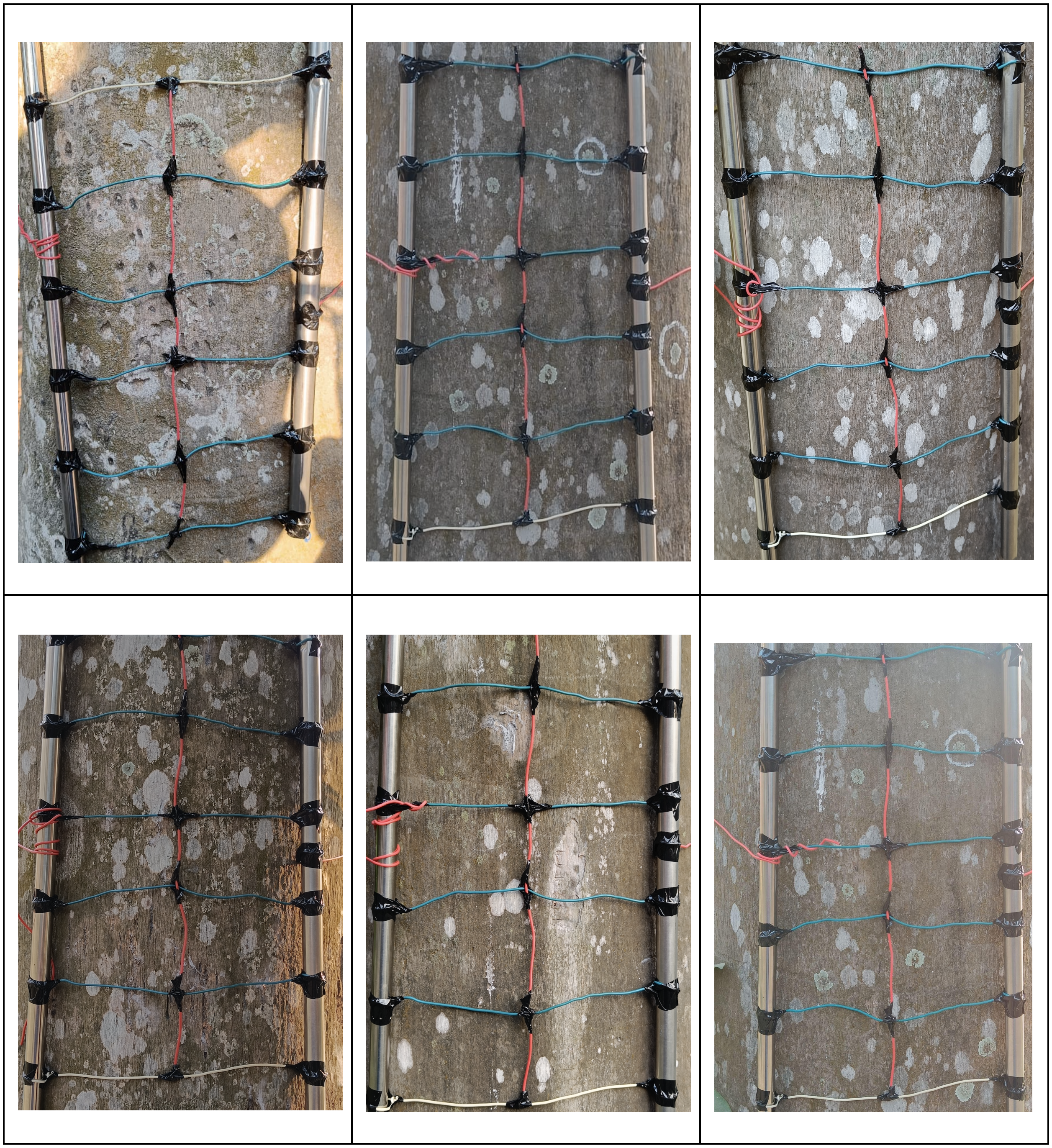


Figure 4.3: Grid mapping of tree specimens to identify the concentration of lichen growth in tree bark

Furthermore, *Roystonea regia* and *Wodyetia bifurcata* palm varieties of tree specimens were taken into consideration for grid mapping of lichen species. In the studies of Gottardini (1999), it was observed that pollution level could be effectively predicted utilising frequency method as it possesses a certainty of 97%. The frequency of species present in palm trees is analysed through the area covered by lichen species in each grid. Thus, similar tree species in study area were identified for ensuring homogenous observations. For selecting trees circumference of bark (min. 70 cm), inclination of trunk (< 10%) and damage in bark of trees are considered (Conti & Cecchetti, 2001). It is observed that major correlation of pollution data is considered with frequency of lichen species that can be illustrated through formula:

$$IAP = \sum F_i$$

Moreover, frequency value of each species is documented on a monthly and weekly basis. However, frequency of each species corresponds with amount of sub-units present in the network which varies at a range (minimum=1 and maximum=10). This assists in computing IAP of the specified tree.

Table 4.2: Diameter of breast height and height from bottom of tree to grid placement of certain Roystonea regia and Wodyetia bifurcata trees found in this region.

Tree No.	Tree Name	Diameter of Breast Height (cm)	Height from bottom of tree to grid placement (cm)
J-1037	*Roystonea regia*	173.5	75
J-1036	*Roystonea regia*	132	116.5
J-1035	*Roystonea regia*	131.5	65
J-1029	*Wodyetia bifurcata*	157	86
J-1016	*Wodyetia bifurcata*	148	72

For sampling, every lichen canopy cover was identified and every tree selected was evaluated utilising grids of canopy cover possessing 100 blank squares. The grid sheet for assessing canopy cover was placed on group below canopy of sampling tree. Thus, number of shaded squares provided percentage canopy cover (Yatawara & Dayananda, 2019).

4.6. Identification of trees possessing lichen species

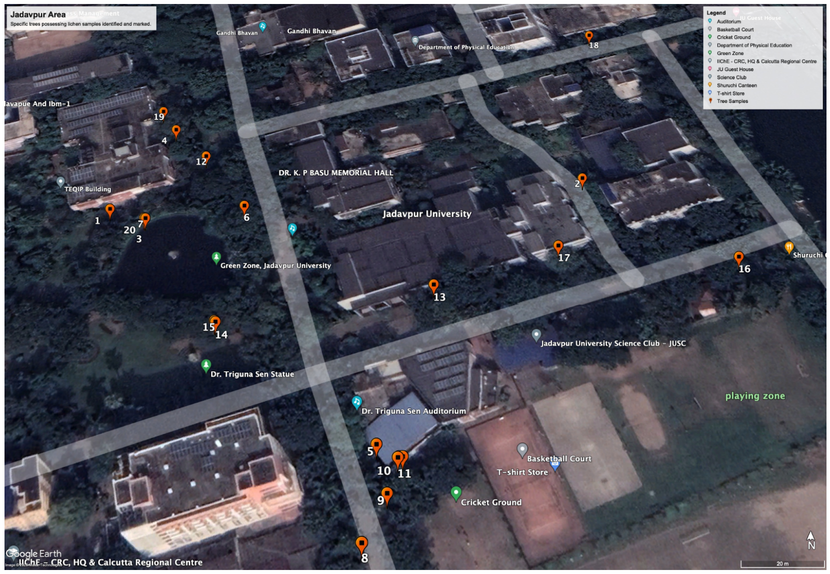


Figure 4.4: Trees in which lichen samples were found in Jadavpur, Kolkata region.

Fig. 4.4 illustrates sampling locations of lichen specimens and precisely the significant trees in which constant growth of lichen species were observed at a span of 9 months of monitoring. Additionally, geographic coordinates of the trees along with their scientific name has been illustrated in Table 4.3. It is observed that *Roystonea regia*, *Dypsis lutescens* and *Wodyetia bifurcata* are dominant varieties of palm trees in which lichen growth has been assessed.

Table 4.3: Geographic coordinates of trees with lichen growth on their bark

Tree No.	Tree Name	Latitude	Longitude
1	*Mimusops elengi*	22.49888	88.370454
2	*Mimusops elengi*	22.498988	88.37189
3	*Roystonea regia*	22.498847	88.370565
4	*Roystonea regia*	22.499168	88.370602
5	*Roystonea regia*	22.498211	88.371276
6	*Roystonea regia*	22.49888	88.370862
7	*Dypsis lutescens*	22.498846	88.370569
8	*Dypsis lutescens*	22.497993	88.371248
9	*Dypsis lutescens*	22.4981	88.371305
10	*Dypsis lutescens*	22.498183	88.371342
11	*Dypsis lutescens*	22.498181	88.37133
12	*Dypsis lutescens*	22.499062	88.370717
13	*Wodyetia bifurcata*	22.498627	88.371425
14	*Wodyetia bifurcata*	22.498519	88.370831
15	*Wodyetia bifurcata*	22.498522	88.370828
16	*Wodyetia bifurcata*	22.498723	88.372309
17	*Couroupita guianensis*	22.498754	88.371789
18	*Mesua ferrea L.*	22.499567	88.371985
19	*Polyalthia longifolia Sonn.*	22.499241	88.370547
20	*Phoenix dactylifera*	22.498846	88.370569

4.7. Air Quality Data collection

The data on concentration of various air pollutants was collected from CPCB's continuous air quality monitoring station in Jadavpur area. Daily data was acquired separated through parameters or pollutants being measured which involves $PM_{2.5}$, PM_{10}, NO, NO2, NOx, SO2, Ozone, relative humidity, Benzoyl pyrene, wind speed and wind direction. The daily data were acquired between June 2019 to April 2023. A total of nearly 16800 data samples were recorded and utilised for assessment. The initial 70% of a dataset has been utilised as training data and the remaining 30% as a testing dataset. The overall dataset utilised for air quality analysis prediction and further descriptive statistics of the data frame are illustrated in Fig 4.5 and Fig 4.6.

	Date	PM2.5	PM10	NO	NO2	NOx	SO2	Ozone	RH	BP	WD	AQI	AQI_bucket
0	06-01-2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	06-02-2019	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	06-03-2019	28.63	39.56	1.92	14.14	16.28	0.65	110.32	62.56	1000.58	137.13	NaN	NaN
3	06-04-2019	24.09	43.41	3.00	10.53	13.54	0.85	53.37	73.63	1000.45	123.46	NaN	NaN
4	06-05-2019	26.82	33.82	3.89	15.21	19.08	1.01	36.84	89.06	999.52	120.78	41.0	Good

Figure 4.5: Raw dataset with $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, SO_2, Ozone, RH, BP, WD, AQI, AQI bucket features

	PM2.5	PM10	NO	NO2	NOx	SO2	Ozone	RH	BP	WD	AQI
count	1301.000000	1300.000000	1291.000000	1292.000000	1292.000000	1301.000000	1285.000000	1301.000000	638.000000	1301.000000	1229.000000
mean	49.303643	97.149485	25.720209	25.891680	50.333019	6.457756	22.294630	80.249700	989.738401	170.177879	106.274207
std	40.633445	73.898872	34.810827	19.231718	48.067605	2.740452	14.039847	13.941522	45.770995	61.211276	85.658669
min	4.340000	11.420000	1.070000	3.850000	6.420000	0.650000	2.810000	34.980000	725.000000	73.400000	17.000000
25%	17.720000	36.447500	5.970000	12.707500	19.787500	4.590000	12.960000	70.510000	996.002500	113.060000	40.000000
50%	32.450000	64.470000	10.450000	19.590000	31.295000	5.970000	18.610000	81.700000	998.550000	154.090000	68.000000
75%	74.630000	154.282500	32.630000	33.265000	62.922500	7.800000	27.400000	92.380000	999.665000	231.320000	158.000000
max	218.910000	374.770000	343.800000	148.850000	386.500000	23.340000	110.320000	99.910000	1013.590000	330.930000	369.000000

Figure 4.6: Description of data frame

In this study, python programming language has been utilised to acquire descriptive statistics. The library functions in Python utilised for data manipulation and data statistics are NumPy and Pandas. Further, based on data pre-processing appropriate visualisation and correlation analysis of dataframe is acquired.

4.8 Data Pre-processing

Data quality plays an effective role in significantly generating and visualising structured machine learning models. The pre-processing process enables in diminishing noise persistent in data that further escalates generalisation capability and processing speed of machine learning algorithms. While reading the dataset, Unicode escape encoding has been used to acquire a uniform dataset. Moreover, missing values or null data are the most common errors observed in monitoring applications and data extraction. In order to obtain appropriate prediction in machine learning models, data is required with suitable value for all the features in every observation. Thus, missing data or values can lead to bias in estimating parameters and further negotiate accuracy of ML models. The missing value percentage of various features in the dataset has been illustrated in the diagram below. It has been observed that among various features Benzoyl Pyrene has maximum missing values and wind direction has least possible missing values. A significant number of missing data is persistent due to various factors, which involves if a station is able to sense data although it does not have a significant device to record the same for a specific interval.

	Total	Percent
BP	669	51.185922
AQI	78	5.967865
AQI_bucket	78	5.967865
Ozone	22	1.683244
NO	16	1.224178
NO2	15	1.147666
NOx	15	1.147666
PM10	7	0.535578
PM2.5	6	0.459067
SO2	6	0.459067
RH	6	0.459067
WD	6	0.459067
Date	0	0.000000
year	0	0.000000

Figure 4.7: Percentage of missing values in various features

Furthermore, to acquire appropriate interpretation from the overall dataset and utilise effective techniques, imputation of missing values is a suitable measure to decrease biases. The percentage of missing values is calculated for deciding if a specific column could be kept, dropped or imputed. Moreover, when more than 55% to 60% of the data is missing in a specific case then the entire column is chosen to be dropped, although, for our use case, the maximum amount of missing values is observed in Benzoyl Pyrene with nearly 52% of data.

4.9. Imputation of Missing value

In this segment, the missing values are imputed through an ***iterative imputer*** numerical data as the dataset possesses multivariate data. This enables the use of available data in different features for estimating missing value imputation as it estimates various features of several independent variables. It is based on the strategy of imputing missing data through modelling various features with missing data as an outcome of other features. Moreover, categorical data has been imputed through a mode that involves replacing through ***"most frequent category value"***. Thus, after missing data is appropriately imputed and a revised data frame is prepared, a normalisation process is implemented to standardise data. This ensures that variable significance is unchanged by their units or ranges (Kumar & Pande, 2022). Normalisation procedure enables reorganising databases to a common measurement scale. It is highly significant for training machine learning models and boosting performance. Therefore, this enables normalising data types of various features present in the data frame such as for data time data *dtype* is normalised using *datetime* python library.

4.10. Selection of features

In this study CPCB dataset includes significant parameters that include $PM_{2.5}$, PM_{10}, NO, NO2, NOx, SO2, Ozone, relative humidity, wind direction, and benzoyl pyrene. Central and state agencies utilise this parameter to make individuals aware of quality of air. Selection of features is a significant pre-processing tread and it assists in zoning relevant variables in a dataset, further eliminating collinear variables (Github.io, 2023). Furthermore, for a significant understanding of features in the dataset feature extraction and feature engineering of both numerical and categorical data present are conducted. Thus, effectively understanding correlation among different features ***Pearson's Correlation Coefficient*** is utilised to evaluate relationship between each feature in terms of target variable (Air Quality Index). This further illustrates correlation based on collinearity among target value and other features. The range of ***Pearson's Correlation Coefficient*** varies between -1 to +1. However, it is observed that if two variables don't share any correlation then this coefficient is zero. On the other hand, if ***Pearson's Correlation Coefficient*** range is -1 or +1, it illustrates that association among two variables is stronger (Cohen et al., 2009). Thus, the formula for calculating ***Pearson's Correlation Coefficient*** has been illustrated below:

$$r = \frac{n(\Xi xy) - (\Xi x)(\Xi y)}{\sqrt{[n\Xi x^2 - (\Xi x)^2][n\Xi y^2 - (\Xi y)^2]}}$$

Where,

r represents correlation coefficient, n represents sample size, x represents a value of first variable, y represents a value of second variable

Table 4.4: Correlation analysis of various pollutants concerning AQI

Sl No.	Features	Correlation Coefficient (r)
1.	$PM_{2.5}$	0.979954
2.	PM_{10}	0.971962
3.	NO	0.631457
4.	NO_2	0.655242
5.	NOx	0.680807
6.	SO_2	0.278604
7.	Ozone	-0.034560
8.	RH	-0.558522
9.	BP	-0.014323
10	WD	0.691648
11.	AQI	1.000000

It is observed in Table 4.4, that correlation coefficients of $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, and WD are highest (higher than 0.55). However, other features (SO2, Ozone, Relative humidity, and Benzoyl pyrene) are comparatively less correlated with the value of AQI as its Pearson's coefficient value ranges between -0.01 to -0.56. Moreover, the value of month and year attributes in this segment is observed to be 0.035 and 0.112. Therefore, it is observed that in the dataset several features barely contribute to AQI prediction and for such reasons are discarded and important features are taken into consideration (Van et al., 2023). Fig. 4.8 illustrates correlation among various features in the form of a heatmap.

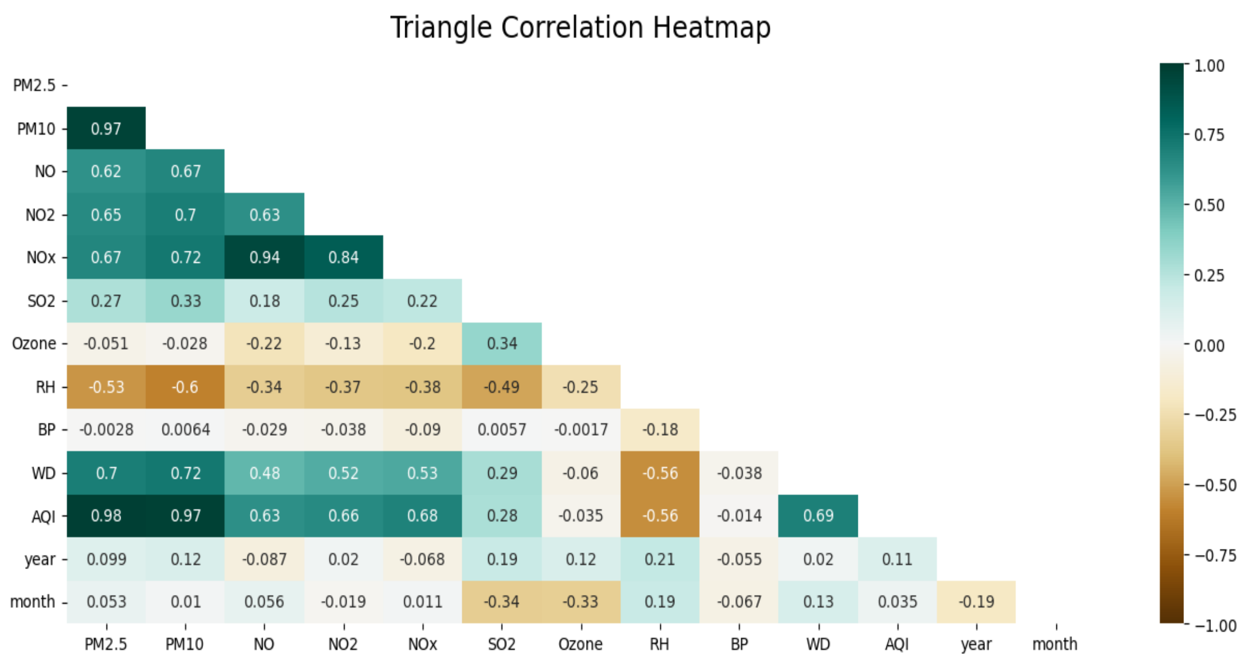


Figure 4.8: Correlation analysis of various features through Heatmap

The heatmap above represents correlation among various features, illustrated through specific colour patterns. The flamboyant colours represent direct association among every feature. It is observed that $PM_{2.5}$ and PM_{10} have the highest correlation with AQI which involves a value of 0.98 and 0.97 although it depicts a minute correlation with relative humidity. Further, relative humidity also possesses an extremely minute correlation with wind direction and AQI. On the other hand, other features depict moderate association with each other as portrayed in the heatmap above.

4.11. Meteorological features

Weather parameters such as wind direction, wind speed and relative humidity directly impact air quality. Moreover, higher wind speed tends to reduce PM2.5 concentration and higher levels of humidity degrades the air pollution levels. Thus, meteorological parameters are extremely significant for air quality index forecasting.

4.12. Prediction Methodologies

The primary pollutant emission in Indian sub-continent is generated due to road traffic, road dust, energy-producing industry, soil, power plants, waste incineration, and open burning of waste. This study possesses significant data from a span of 4 years (June '2019 - May '2023) acquired from ***Central Pollution Control Board***, India. Air quality analysis of substantial air pollutants present in Jadavpur area, such as $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, SO_2, Ozone, etc. have been considered for formulation of ***"Air Quality Index"*** and further prediction concerning aspects. For implementation of machine learning model, the dataset is split into independent and dependent columns, where $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, SO_2, and Ozone are independent columns and AQI is the dependent column as the value of AQI is being predicted through implementation of suitable algorithms. Furthermore, the data is separated into training and testing, the train test split function has been utilised in this segment. Thus, an independent and target column has been considered and for our model, 80% of the data has been used for ***training purposes*** and the remaining 20% is used for ***testing purposes***. For our study, the various regression and classification models have been utilised for prediction and their accuracies have been compared.

4.12.1 Linear Regression

Linear regression is the simplest algorithm in machine learning. It is a statistical model that illustrates a relationship among two variables in terms of a linear equation. It is a supervised machine learning model that observes best-fit linear lines among dependent and independent variables. This algorithm assists in performing a task for predicting value of a dependent variable (y) in terms of an independent variable (x). Thus, linear regression could be illustrated through equation below.

$$y = a_0 + a_1 x + e$$

Where,

$$a_0 = intercept\ of\ regression\ line\ along\ vertical\ axis$$

$$a_1 = regression\ coefficient$$

$$e = random\ error$$

Therefore, linear regression is utilised to acquire the best-fit line, that signifies error among actual and predicted values are required to be minimised. A best-fit line possesses the least error, and different values for coefficients or weights of a line (a_0, a_1) provides various lines of regression.

4.12.2 Decision Tree Algorithm

Decision tree is a prominent and primary algorithm of machine learning. This enables representing decision logic which involves tests and correlates significant outcomes for data classification in a tree-like framework. Furthermore, nodes of decision trees possess multiple levels in which the first node is depicted as a root node. However, various internal nodes depict tests on input attributes or variables. Based on test results, classification algorithm offshoots towards respective internal nodes where branching and test procedures are repeated till it reaches leaf node. The terminal or leaf node correlates with decision outcomes. Moreover, its regressor model has been used for prediction in the x-train and y-train. Further, it has been predicted on the x-test and y-test, which have evaluated the overall performance and accuracy of the model by using evaluation metrics (Root mean squared error, Mean squared error and R-square value).

4.12.3 Random Forest Algorithm

Random forest is signified as an ensemble classifier and comprises numerous decision trees analogous to the technique of forests which is an assemblage of numerous trees. When decision trees are developed extremely deeply, the training data are often overfitting, leading to a significant fluctuation in classification resulting in little change in the input data. They are prone to errors on the test dataset because they are particularly sensitive to their training data. The various decision trees of a random forest are trained using the various training dataset components. The input vector of a fresh sample must pass down with each decision tree of the forest in order to be classified. Each decision tree then takes into account a distinct aspect of the input vector and provides a classification result. The classification that receives the most "votes" (for a discrete classification result) or the average of all the trees in the forest is then chosen by the forest (for numeric classification outcome). The random forest method may decrease the variance caused by the consideration of a single decision tree for the same dataset since it takes results from several distinct decision trees into account.

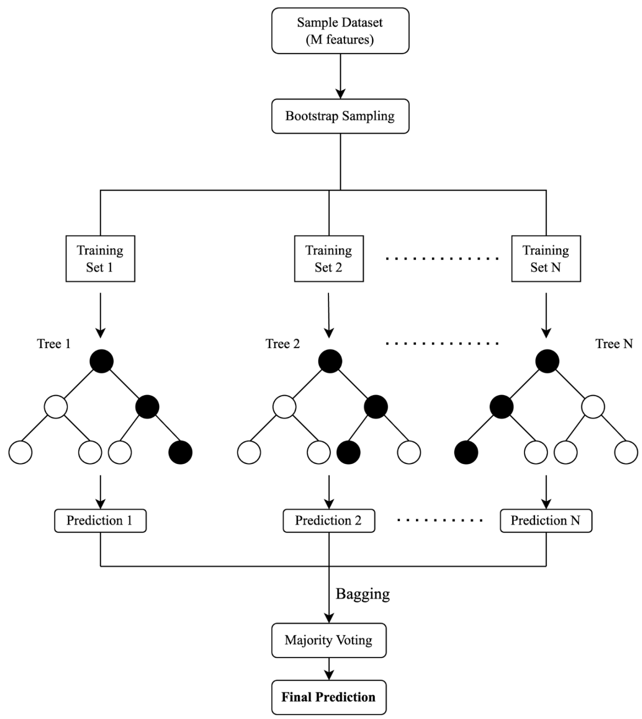


Figure 4.9: Random Forest Regressor Mechanism

4.12.4 XGBoost Algorithm

XGBoost or Gradient Boosted is an ensemble learning algorithm fabricated using several weak learners (simple models) or shallow trees. It is a gradient-boosting library optimally distributed, and fabricated for scalable and efficient training of machine learning models. Moreover, every tree in this segment possesses an in-depth between one to five and effectively performs in predicting dataset segments and further resolves inaccuracies present in previous trees. Therefore, enumeration of weak learners enables enhancing overall performance of model. Gradient boosting technique is effective for enormous datasets that possess high accuracy. This algorithm avoids overfitting through features regularisation, tree pruning, effective handling of missing data, and parallelization. In terms of speed and effectiveness when compared to other models, XGBoost is among the finest algorithms. Moreover, the sequential trees are processed by XGBoost using a parallelized implementation technique. For boosting techniques, the decision tree gets trained based on residual values.

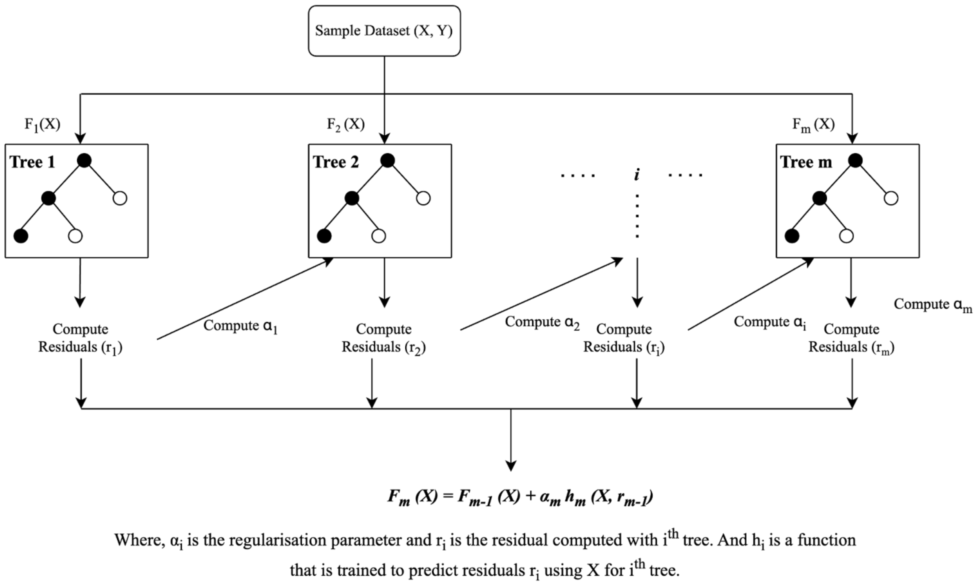


Figure 4.10: XGBoost Regressor Mechanism

4.12.5 KNearest Neighbours Algorithm

The KNN algorithm is a supervised machine learning technique that makes the assumption that a point's value is comparable to its neighbours' values. Moreover, it is built on the idea of calculating value of a specific point using its nearest neighbours in the dataset, ordered by distance. Accordingly, it operates on the tenet of selecting the value of K (neighbours) close to the area of interest and casting a vote for the most prevalent class. Thus, local anomalies increase inaccuracy in the direction of the decision boundaries while the large number of K lowers noise. The main problem with KNN is the way it becomes slower as the amount of data increases.

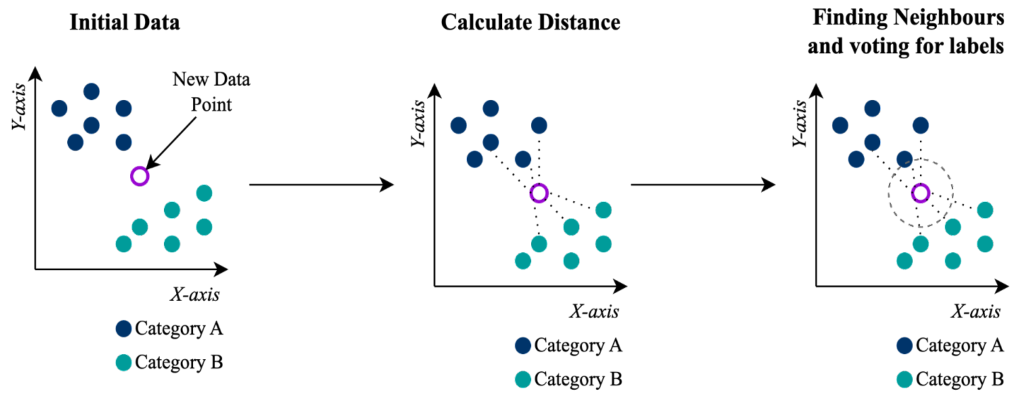


Figure 4.11: KNearest Neighbour Regression Mechanism

The KNN algorithm undertakes K nearest neighbour training data points and further detects the closest instances of k. It is a non-parametric and more flexible approach for utilising the average of neighbourhood X in terms of estimate $\hat{Y}$.

$$Y = f(x) + \xi$$

$$\hat{f}(x) = \frac{1}{K} \sum_{x \, in \, N} y_j$$

This algorithm is implemented when the distance (Mahalanobis distance, Euclidean distance, etc.) is provided along with a k-value, this algorithm evaluates distance among points in the training dataset and data point for adopting k nearest values and demonstrates their average as a prediction (Méndez et al., 2023).

4.13. Performance Index

The performance of the above model has been evaluated by adopting RMSE (Root-mean squared error), R^2 value (coefficient of determination), Accuracy value and Kappa score, which are the regression and classification performance metrics analysed in this model. However, for depicting performance for prediction of air quality index, a significant function is required. Thus, coefficient of determination or R^2 value for testing set in regression models has been considered as a fitness function. On the other hand, classification algorithm accuracy for testing data has been considered a fitness function. The model that possesses a higher R^2 value or accuracy is likely to have better performance. Various metrics utilised for modelling have been calculated through formula provided below:

$$\text{RMSE} = \sqrt{\frac{\sum_{1}^{n}(y_i - y_{\hat{i}})^2}{n}} \qquad R^2 = 1 - \frac{\sum |y_i - y_{\hat{i}}|}{\sum |y_i - \underline{y}|}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad Kappa\ Coefficient\ (K) = \frac{P_0 - P_e}{1 - P_e}$$

Where y_i depicts actual air quality, $y_{\hat{i}}$ depicts predicted air quality and n is the amount of samples evaluated. The RMSE value was utilised to analyse absolute error and it assists in providing true measurements at every predicted data point. In addition, R^2 value enables us to evaluate the best-suited regression line for observed data. Moreover, the Kappa score enables reliability measures among two independent parameters of the target variable. The accuracy illustrates the amount of correct predictions prevalent concerning total predictions. Therefore, a best-suited model for prediction of air quality index was analysed through higher R^2 value and accuracy.

4.14. Deep Learning Algorithm

Deep learning algorithms are primarily a subset of machine learning algorithms and focus on function and structure of brain known as artificial neural networks. Deep neural networks comprise numerous layers of interrelated nodes, where every node is fabricated upon previous layer in order to optimise or refine categorisation or prediction. Thus, progress of computations in this network is stated as forward propagation. Moreover, another process in which this mechanism works is known as back propagation and utilises algorithms such as gradient descent for evaluating errors found in prediction and further adjusting weights and biases of function by operating backwards through layers, for training the model (Ibm.com, 2023). Therefore, deep learning algorithms are subdivided into various forms of neural networks such as *Convolution Neural Network (CNNs)* and *Recurrent Neural Networks (RNNs)* for addressing significant issues or datasets.

4.14.1 LSTMs or Long short-term memory neural network

Long short-term memory is a variety of cyclic neural network that assists evaluating input information in terms of time series (Meng et al., 2020). It further examines correlation among the present input information and the information in the next moment. According to Zhang et al., (2020), LSTM is an artificial recurrent neural network (ARNN) employed in the deep learning industry. The RNN considers an ANN type in which linkages between hidden nodes produce a recurrent structure. LSTMs are capable of memorising data for any period of time. The cell state, which carries information throughout the data processing, is the essential component. Moreover, LSTM network comprises a forgetting gate, an input gate and an output gate (Yuan et al., 2021). Each of them uses a sigmoid activation mechanism to regulate the information that ought to be present in the cell state. The forget gate chooses which information from the previous state should be forgotten. The fresh information that will be utilised to update the memory is decided by the input gate. It produces a vector candidate to be added to the cell state using a hyperbolic-tangent function (Xu et al., 2021). The output gate is the last gate, and it employs a *tanh* activation function to choose which component of the modified cell state will be the output.

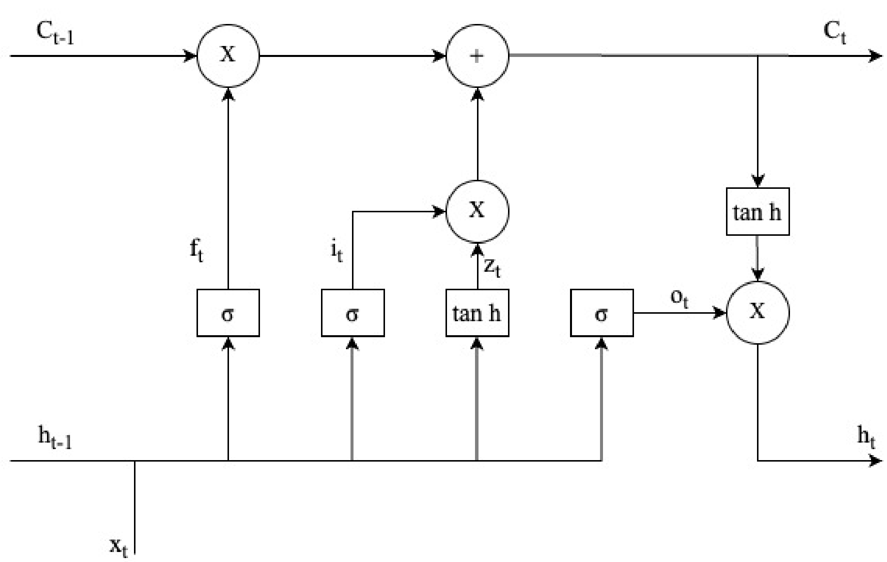


Figure 4.12: Structure of Long Short-Term Memory

Where, C_t = Current unit state, C_{t-1} = status of previous unit, h_{t-1} = hidden state of previous unit, h_t = hidden state of current unit, i_t = input gate, o_t = output gate, f_t = forget gate, x_t = unit input, z_t = status of temporary unit, W_i = weight of input gate, b_i = offset of input gate, W_f = weight of forget gate, b_f = offset of forget gate, W_o = weight of output gate, b_o = offset of output gate, W_z = weight of temporary unit, b_z = offset of temporary unit.

The various parameters could be expressed as (Zhang et al., 2019),

$$C_t = \left(C_{t-1} \times f_t\right) + \left(z_t \times i_t\right), \qquad h_t = tanh\left(C_t\right) \times \left(o_t\right),$$

$$z_t = tanh\left(W_z \times \left[h_{t-1}, x_t\right] + b_z\right), \qquad o_t = \sigma\left(W_o \times \left[h_{t-1}, x_t\right] + b_o\right),$$

$$i_t = \sigma\left(W_i \times \left[h_{t-1}, x_t\right] + b_i\right), \qquad f_t = \sigma\left(W_f \times \left[h_{t-1}, x_t\right] + b_f\right),$$

4.15. Traffic Data

The data was collected by installing cameras at two major traffic intersections of Jadavpur area of Kolkata city. Duration of 1 hour was taken into consideration from 11 a.m. to 12 p.m., which was observed to have a higher flow of traffic. This data was acquired for a span of one week and the traffic volume was observed to be similar with a difference of 50-60 vehicles depending on the day of week.

Table 4.5: Traffic volume analysed at study sites of Kolkata (The traffic data provided below is subject to approximations*)**

Sl. No.	Study Site	Traffic volume (vehicles/h) *	Number of heavy traffic (bus or trucks/h) *
1.	8B Bus stand traffic intersection	1,850	235
2.	Jadavpur Police Station traffic intersection	2165	320

4.16 Workflow Diagram

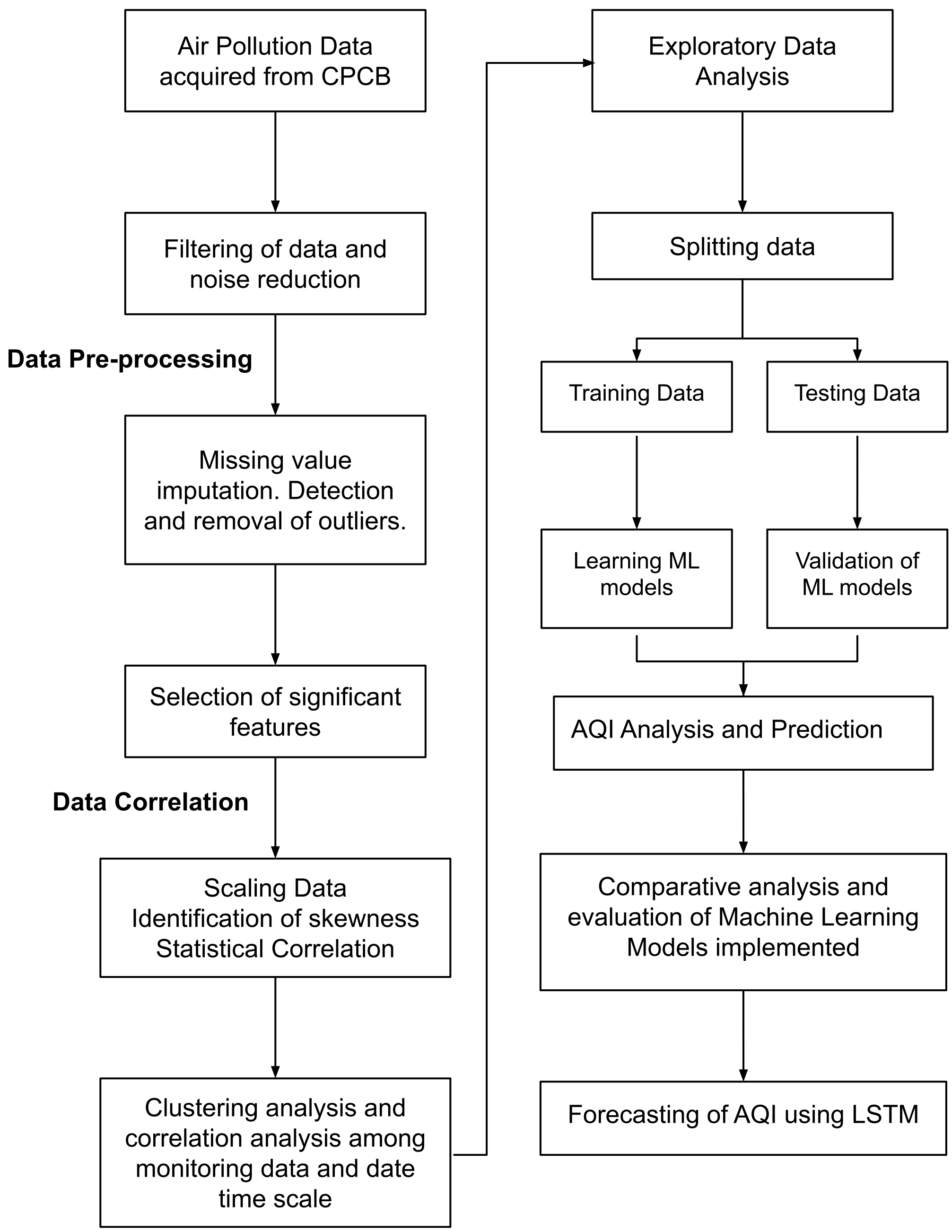


Figure 4.13: Flow-diagram of proposed model for air quality analysis

Chapter 5: Results and Discussion

5.1. Lichen Data Evaluation

Based on the growth aspects of lichen depicted in Table 5.1, it is observed that with respect to time several varieties of lichen species found in Jadavpur area show a constant increase in their overall surface area. Growth rates in 20 lichen species were summarised for a span of 9 months. Direct lichenometry method has been considered as direct measurements of lichen growth were taken (Armstrong & Bradwell, 2010).

Table 5.1: Growth rate of lichen samples found in Jadavpur area

Monitoring Dates	Growth in surface area of Lichen samples (mm^2)									
	1	2	3	4	5	6	7	8	9	10
03/07/22	3.476	0.895	2.492	0.809	7.493	5.248	3.395	2.291	1.511	0.975
18/07/22	3.498	0.914	2.510	0.821	7.524	5.356	3.418	2.310	1.532	1.004
04/08/22	3.510	0.936	2.539	0.836	7.573	5.416	3.462	2.332	1.546	1.013
20/08/22	3.524	0.943	2.575	0.855	7.615	5.427	3.503	2.356	1.590	1.028
05/09/22	3.539	0.957	2.598	0.863	7.669	5.431	3.521	2.378	1.605	1.041
15/09/22	3.556	1.094	2.624	0.878	7.687	5.456	3.537	2.403	1.616	1.054
25/09/22	3.649	1.262	2.662	0.898	7.724	5.487	3.549	2.428	1.625	1.073
05/10/22	3.672	1.292	2.697	0.911	7.754	5.508	3.562	2.458	1.634	1.092
15/10/22	3.681	1.301	2.721	0.923	7.797	5.519	3.586	2.517	1.655	1.105
22/10/22	3.730	1.308	2.795	0.967	7.891	5.538	3.603	2.525	1.677	1.114
04/11/22	3.730	1.317	2.883	1.027	8.040	5.555	3.620	2.541	1.693	1.227
22/11/22	3.963	1.330	2.910	1.039	8.174	5.617	3.641	2.560	1.714	1.258
03/12/22	3.970	1.358	2.936	1.067	8.289	5.740	3.705	2.580	1.722	1.322
20/12/22	3.989	1.385	2.964	1.098	8.293	5.834	3.734	2.608	1.750	1.371
11/01/23	4.015	1.391	2.971	1.112	8.383	5.844	3.783	2.658	1.826	1.412
21/01/23	4.023	1.398	2.979	1.129	8.426	5.929	3.905	2.667	1.830	1.427
01/02/23	4.030	1.399	2.987	1.167	8.443	6.042	3.951	2.677	1.835	1.441
15/02/22	4.112	1.412	2.991	1.178	8.465	6.104	3.963	2.764	1.863	1.452
03/03/23	4.138	1.419	2.998	1.187	8.479	6.165	3.986	2.858	1.886	1.476
19/03/23	4.156	1.427	3.011	1.195	8.499	6.257	4.008	2.879	2.024	1.487
27/03/23	4.189	1.433	3.090	1.203	8.509	6.304	4.214	2.916	2.057	1.498

(Contd.) *[Refer to Appendix 1]*

The *Parmelia caperata* variety of foliose lichen species develop consistently in this region at a growth rate of 0.30 mm^2 every 15 days. This crustose variety of lichen is observed to have a significant increase in growth rate during the months of September, October and November. In certain lichen species such as Lichen 2, Lichen 5, Lichen 7, Lichen 9, Lichen 12, Lichen 13, Lichen 14, Lichen 16, Lichen 17 and Lichen 19 has portrayed a faster growth rate as compared to other lichen species found in the region.

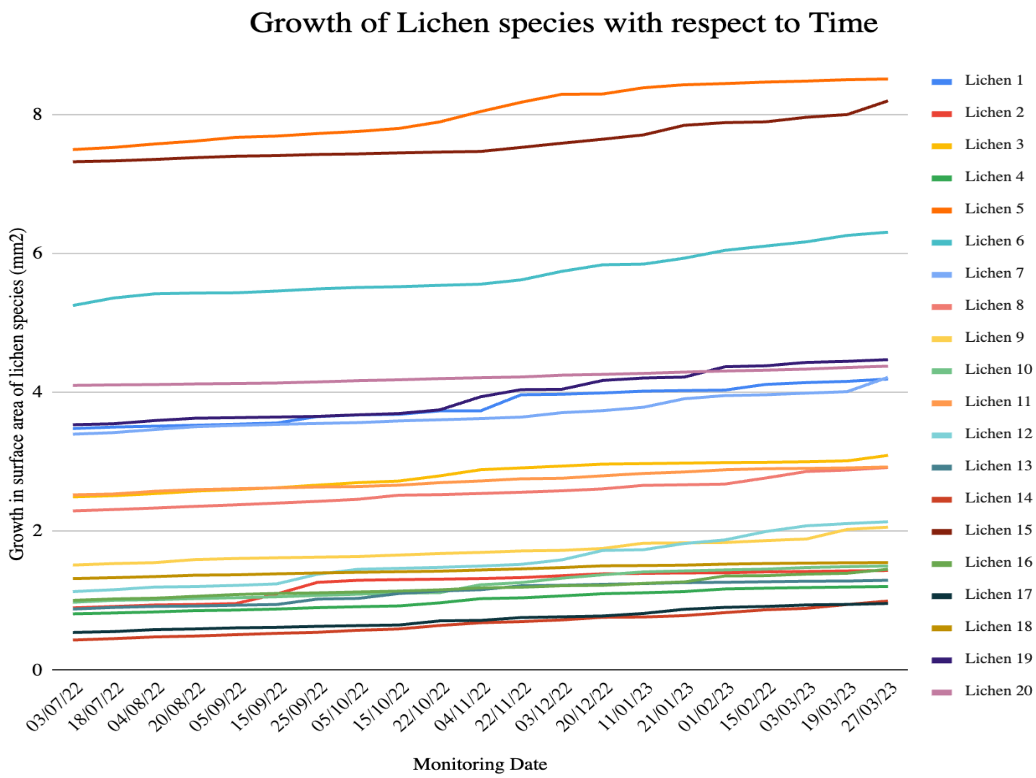


Figure 5.1: Growth of lichen species surface area for monitoring time (July '22 to March '23)

From monitoring lichen species found in Jadavpur area, near major traffic intersections it was observed that initial growth was steady although it was accompanied by a period in which a linear growth of thallus area was observed for certain lichen species (Armstrong & Bradwell, 2011). Thus, a direct association among size and age of lichen species has been depicted above through graphical representation. This enabled reporting major change in growth pattern of certain lichen species based on seasonal aspects during monitoring tenure. Therefore, due to major change in surface area of specific lichen species during post-monsoon season, it could be stated that overall pollution level prevalent in this area significantly decreased during these months enabling sustenance of lichen species.

5.2 Evaluation of Influence of Atmospheric purity

The compositional transformation in lichen communities were coordinated with transformation in atmospheric pollution levels. Moreover, the ecological index (Q) for specific species was evaluated through number of prevalent species (growing) at study area and further dividing the sum through number of sampling sites in which the species have occurred. The frequency or cover of every species have been evaluated utilising a transparent quadrant which was also required for lichen sampling (LeBlanc et al., 1974). Therefore, for Jadavpur traffic intersection area the value of number of species (n), factor of accompanying species (Q), frequency of every species (f) was utilised for calculating IAP value and it is represented in the Table 5.2.

Table 5.2: Calculation for Index of Atmospheric Purity lichen species in Roystonea regia and Wodyetia bifurcata trees.

Tree No.	Tree Name	Number of species (n)	Factor of accompanying species (Q)	Frequency of each species (f)	Influence of atmospheric purity
J-1037	*Roystonea regia*	2	12	61	73.2
J-1036	*Roystonea regia*	2	11	32	35.2
J-1035	*Roystonea regia*	3	9	52	46.8
J-1029	*Wodyetia bifurcata*	1	13	29	37.7
J-1016	*Wodyetia bifurcata*	2	17	15	25.5

Table 5.3: IAP value classified into five categories in terms of air quality

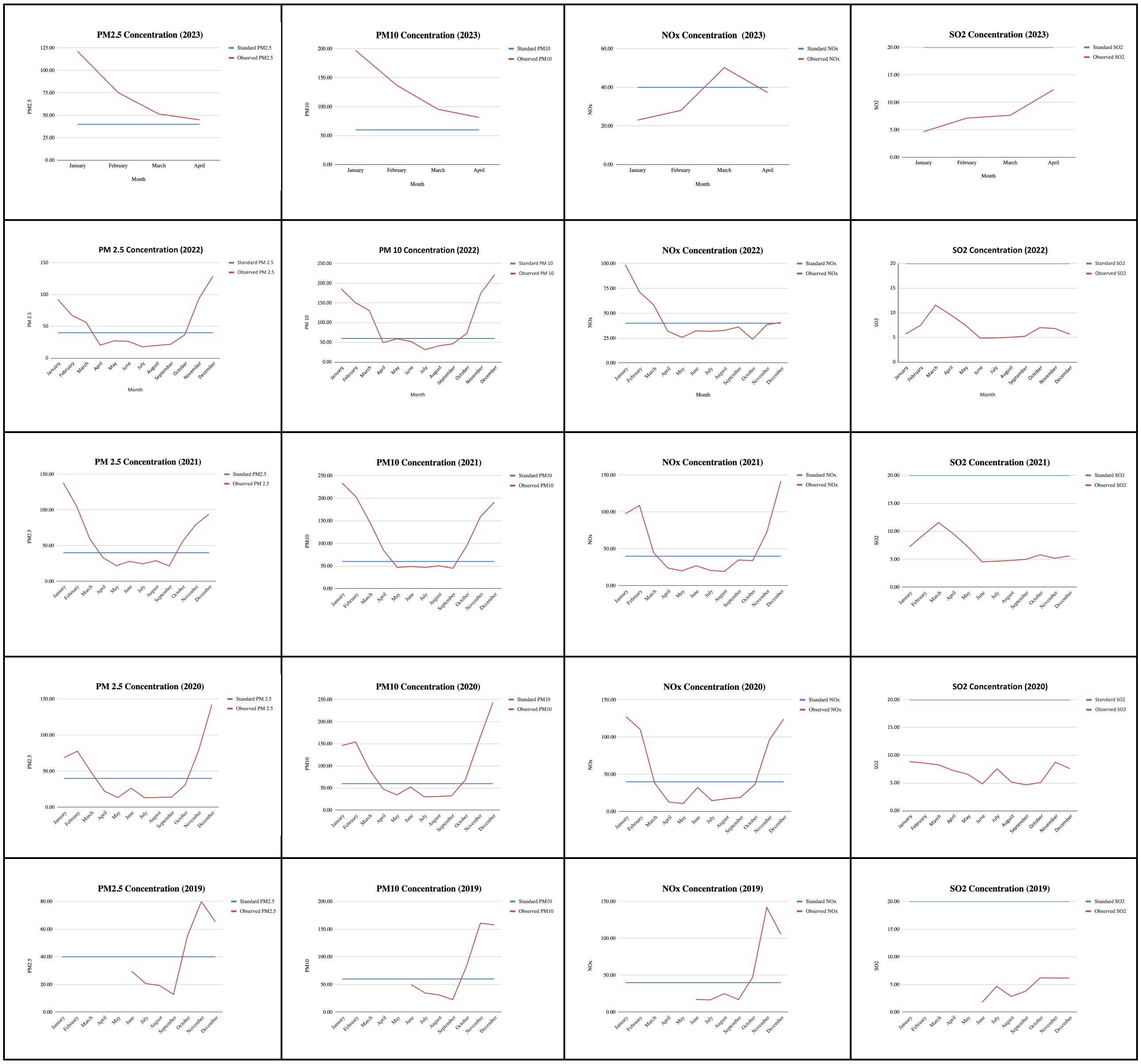
Quality Levels	IAP Range	Pollution Level
Level A	0 - 12.5	Extremely high pollution level
Level B	12.5 - 25	High pollution level
Level C	25 - 37.5	Moderate pollution level
Level D	37.5 - 50	Low pollution level
Level E	>50	Extremely low pollution level

From Table 5.2, the Influence of atmospheric purity is calculated and the permissible range of IAP in terms of pollution levels has been analysed from Table 5.3. Thus, based on the results it is observed that in our study area the overall pollution level in terms of tree specimens studied for IAP calculation, is moderate. Mostly, the values obtained from *Roystonea regia* and *Wodyetia bifurcata* tree specimen belongs to quality level C and predominant lichen species in these trees were observed to be Caperat lichen (*Parmelia caperata*), Wreath lichen (*Cryptothecia subnidulans*). However, for major lichen growth has been observed in tree J-1037 as its IAP range exceeded Level E indicating extremely low pollution level. From our study, lichen desert has not been observed in this area, which signifies there has been presence of lichen species in the area and the IAP value is not 0.

5.3. Variation in pollutant Concentration

For illustrating scenario of data utilised, the four principal air pollutant concentrations in Jadavpur area were plotted in the form of a line graph, and acquired results are illustrated in the table above. Moreover, these pollutants depicted apparent periodicity i.e., $PM_{2.5}$, PM_{10}, NOx and SO_2 portrayed a seasonal trend of decrease in pollutant concentration levels during summer and monsoon months although a significant increase in pollutant concentration is observed during winters. $PM_{2.5}$, PM_{10}, NO_x and SO_2 concentration attained their maximum values during the month of December and minimum values were observed during September every year, illustrating a ‘U’ structure. Thus, during summer due to powerful solar radiation, surface temperature increases distinctly and heats the surface air. This generates enhanced precipitation and transmission that hasten deposition and diffusion of atmospheric pollutants (Zhang et al., 2019). On the other hand, during winters surface temperature is low and generates surface inversion, also meteorological conditions are unfavourable for vertical transmission. Thus, air pollution is significantly greater near surface, as all four pollutants possess similar trends and influencing factors, although remaining pollutants such as Ozone and Benzoyl Pyrene are required to be consolidated for predicting overall Air Quality Index.

Table 5.4: Yearly (June' 2019 - April' 2023) variation of various pollutant ($PM_{2.5}$, PM_{10}, NO_x, SO_2) concentration with respect to standard values prescribed by CPCB.



The above graphical representation illustrates that overall $PM_{2.5}$, PM_{10} and NO_x levels tend to decrease during month of April and further remain within standard limit till September. However, a significant increase in levels of $PM_{2.5}$, PM_{10} and NOx is observed during September and it exceeds the standard concentration level during months October, November, December, January, February and March for all years taken into consideration. Moreover, SO_2 concentration in this region remains within limits of standard concentration value prescribed by CPCB.

Although a significant increase in SO_2 concentration was observed during month of March in the year 2021, 2022 and 2023.

5.4. Meteorological parameters

Fig. 5.2 to Fig. 5.8 depicts a windrose diagram for June 2019 to March 2023 for Jadavpur area in Kolkata, India. The wind patterns were analysed in terms of seasonal variation with March, April, May, June considered as summer months, July, August, September, October are considered as Monsoon months and November, December, January, February as winter months. The wind pattern analysis was comparatively similar every year based on overall wind speed with few variations in wind direction patterns. The predominant wind direction patterns observed during summer months are South, South-West and South-East with 0.7-0.8 m/s average wind speed. During monsoons predominant wind direction is observed to be South, South-East with 0.6-0.7 m/s average wind speed. In winters predominant wind direction West, South-West, South-East, East with 0.35-0.5 m/s average wind speed. Moreover, temperature variations in various seasons also significantly affect wind patterns (Sharma et al., 2020). Thus, it can be deduced that mostly metrology of each season during analysis period from 2019-2023 was observed to be similar on a yearly basis. The seasonal dataset utilised for formulation of windrose model from year 2019 - 2023 has been provided in Appendix. *[Refer to Appendix 2]*

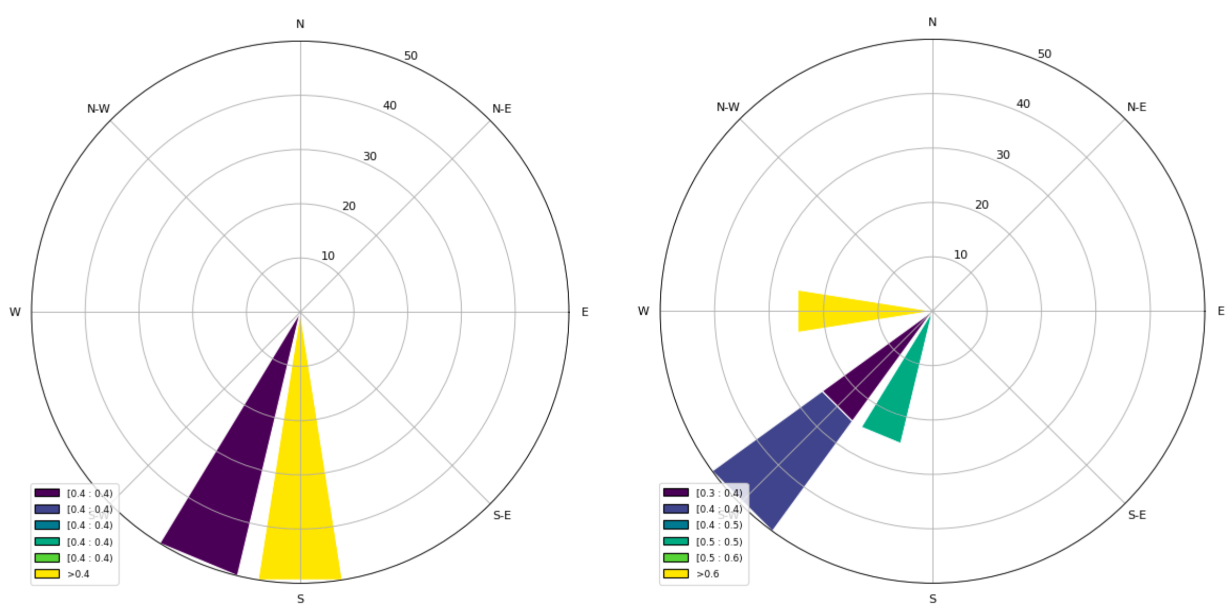


Figure 5.2: Windrose Diagram for Summer ’23 and Winter ’22

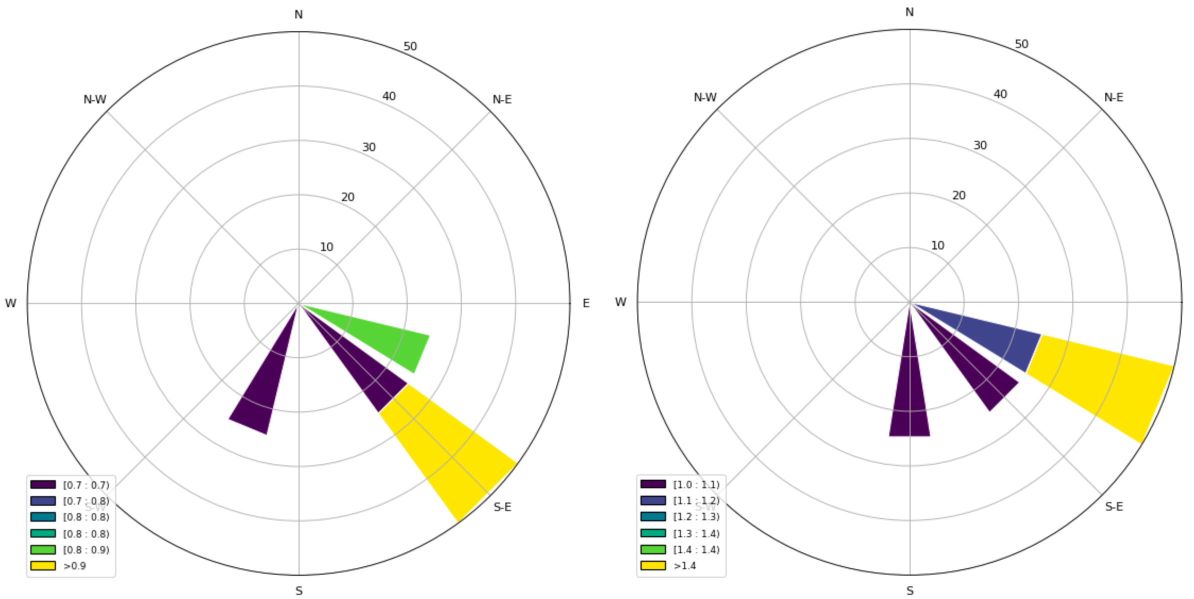


Figure 5.3: Windrose Diagram for Monsoon '22 and Summer '22

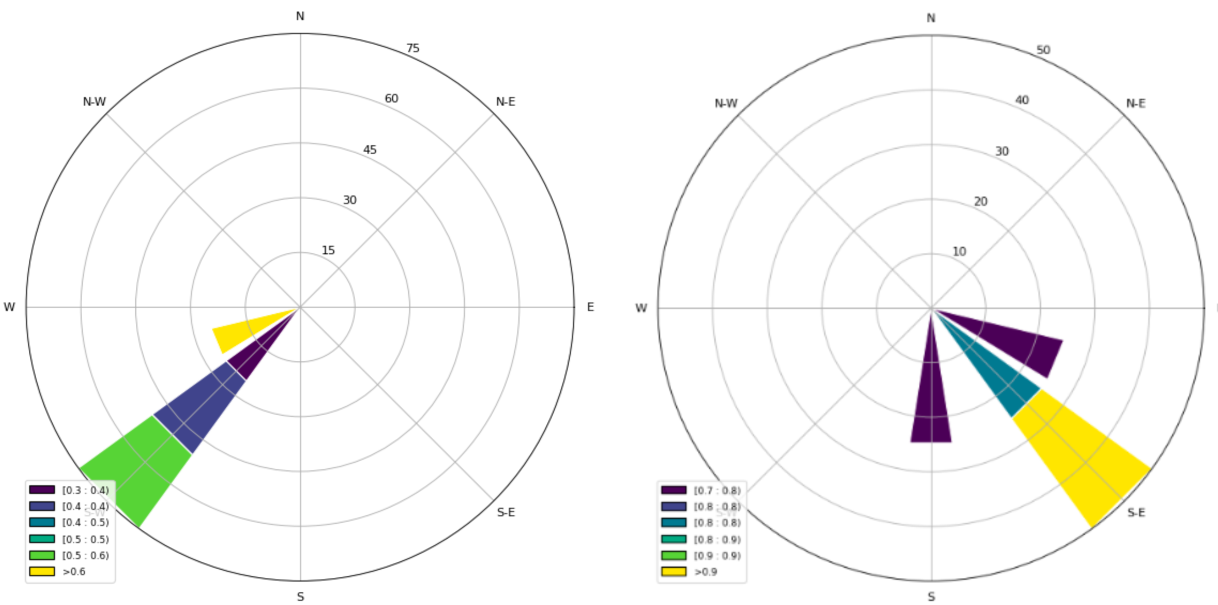


Figure 5.4: Windrose Diagram for Winter '21 and Monsoon '21

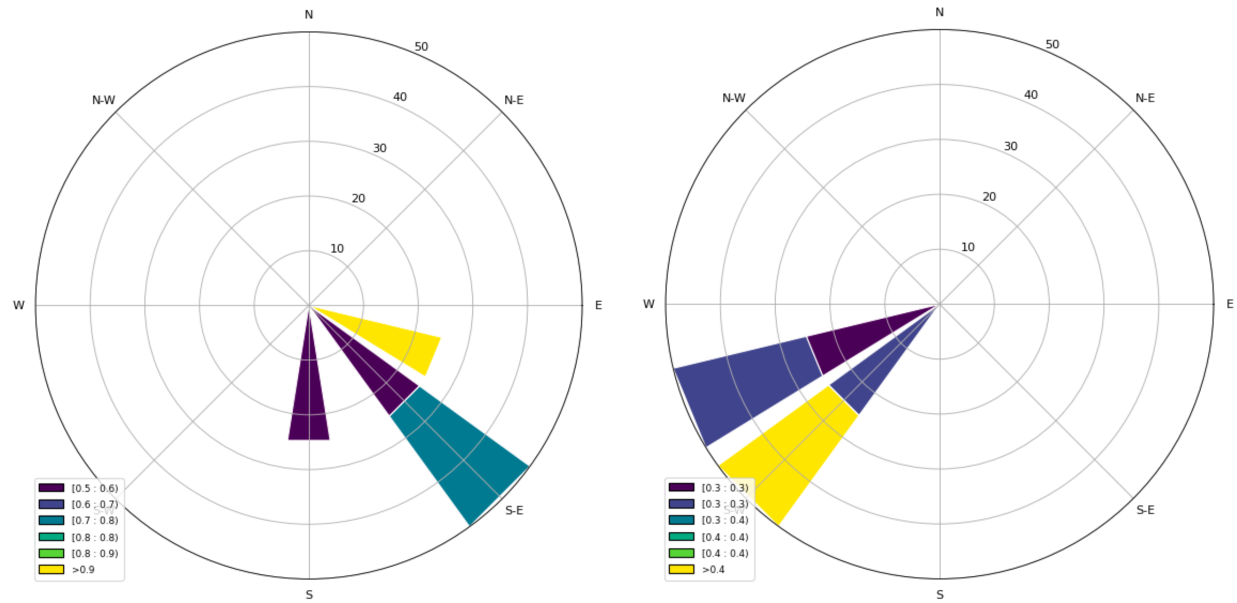


Figure 5.5: Windrose Diagram for Summer '21 and Winter '20

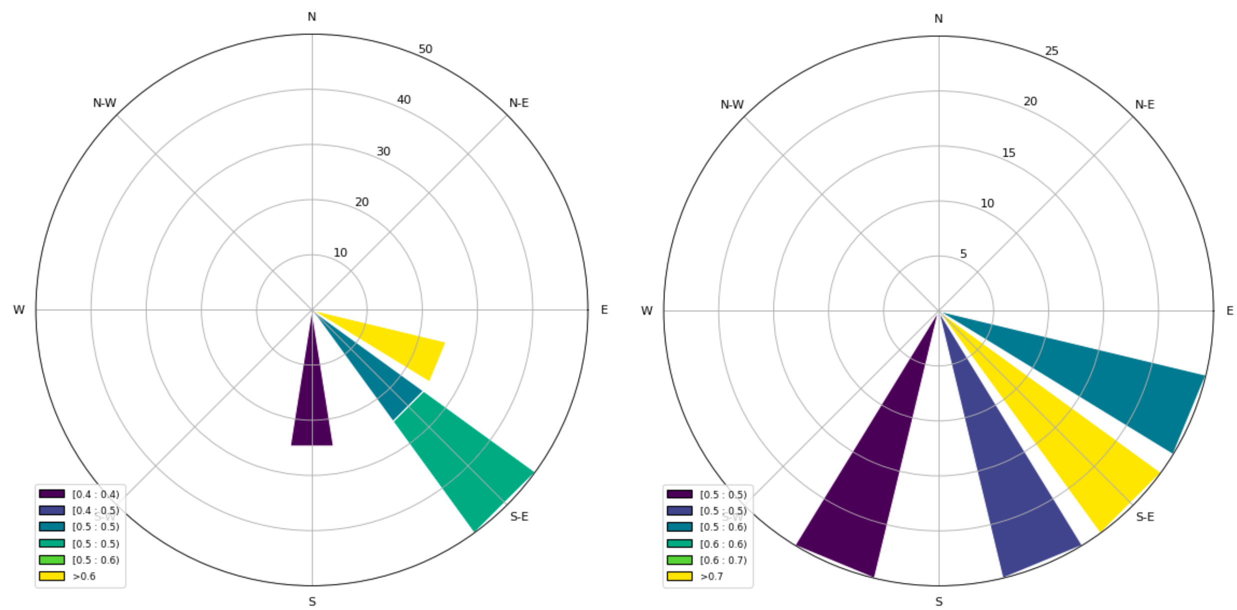


Figure 5.6: Windrose Diagram for Monsoon '20 and Summer '20

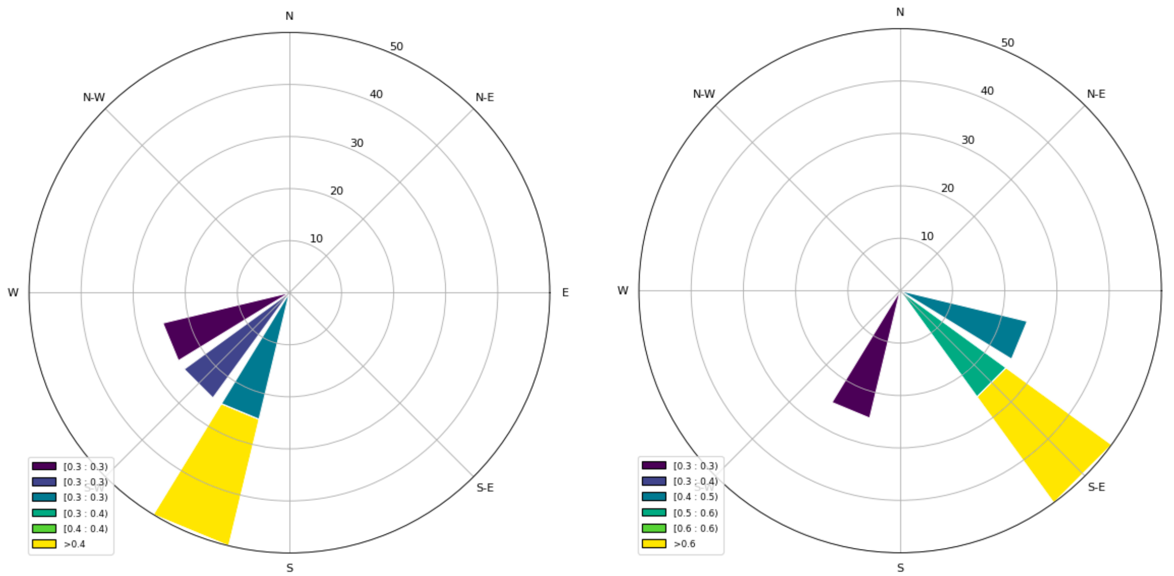


Figure 5.7: Windrose Diagram for Winter '19 and Monsoon '19

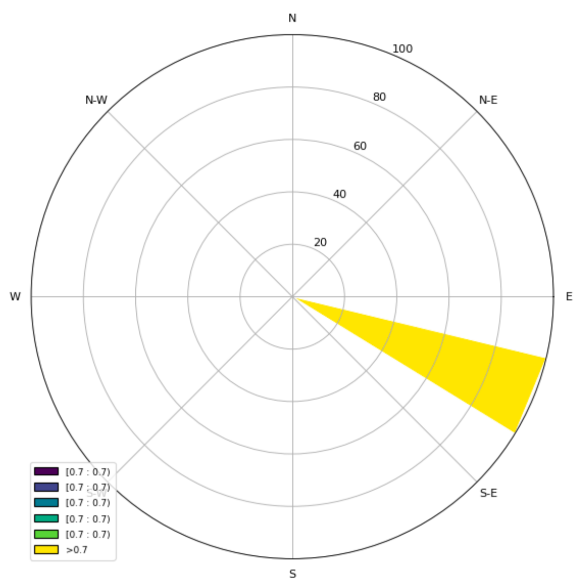


Figure 5.8: Windrose Diagram for Summer '19

5.5. Air Quality Index analysis using Machine Learning Models

Our air quality model has been fabricated utilising various packages of python that involve NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras and TensorFlow. Furthermore, in this study Air quality index prediction was simulated utilising various machine learning algorithms and deep learning architecture such as LSTM (Google Collab and Jupyter Notebook platforms has been utilised for model fabrication).

5.5.1 Performance Metrics of dataset

In this study, the model is employed to predict Air Quality Index of Jadavpur area. Moreover, cross validation has been utilised to evaluate accuracy of the model and this process randomly segregates dataset into various subsets and further runs numerous intervals in order to elude overfitting issues. Although higher computation costs are generated as the higher running time is required. This work utilised five prediction models comprising Linear Regression, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbours and time series analysis was conducted using LSTM architecture. Here, seven features ($PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, SO_2, Ozone) in the dataset were utilised as training features and AQI were utilised as a target feature.

5.5.2 Using Linear Regression for AQI Prediction

This supervised machine learning algorithm, provides a continuous predicted output and it possesses a constant slope. Moreover, the linear regression model was predicted in X_train and X_test. The evaluation of this model was done utilising evaluation metrics such as root mean square error, mean squared error, R^2 and training score. Thus, lower value of root mean square error and higher R^2 value closer to 100 signifies better model performance. In this case for assessing linear regression models the RMSE value is observed to be 14.58 at training and 13.1586 at testing. The R^2 value is observed to be 97% in training and 97.5% in testing for linear regression.

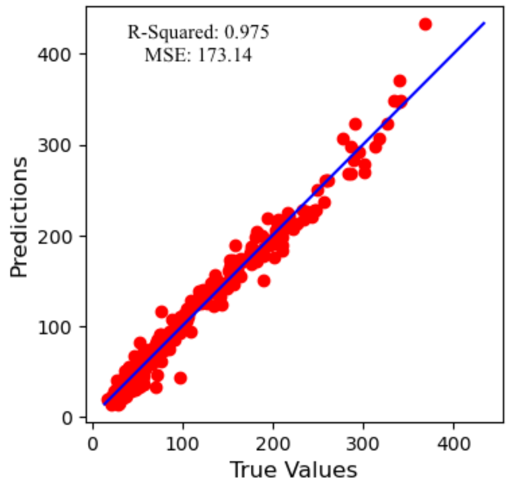


Figure 5.9: Scatter plot for True versus Predicted values variation for Linear Regression

5.5.3 Using Decision Tree Regression for AQI Prediction

The Decision tree regression model was fitted in X_train and Y_train. This regressor generates a model in the form of tree structure and breaks down the dataset into smaller and smaller subsets. It is further predicted on X_train and X_test. The evaluation of this model was done utilising evaluation metrics such as root mean square error, mean squared error, R^2 and training score. In this case for assessing decision tree regression models the RMSE value is observed to be 0 at training and 12 at testing. The R^2 value is observed to be 100% in training and 97.9% in testing for decision tree regression. Thus, the decision tree model is overfitting and it's not an ideal model for prediction.

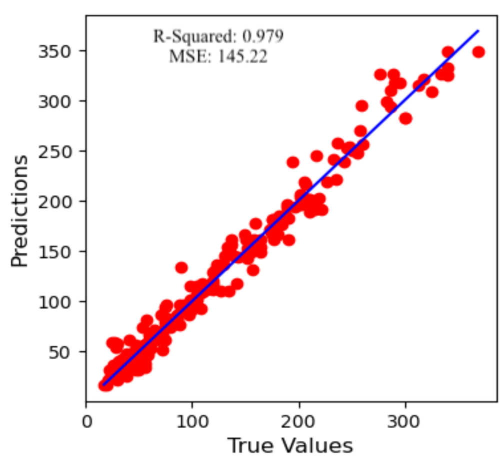


Figure 5.10: Scatter plot for True versus Predicted values variation for Decision Tree Regressor

5.5.4 Using Random Forest Regression for AQI Prediction

The Random Forest regressor model was fitted in X_train and Y_train. This is a meta-estimator that fits classifying decision trees on various sub-samples of the dataset and it utilises averaging that would enable enhancing predictive accuracy. It further follows a bagging mechanism that helps in reducing overfitting and bias. Moreover, this regression model would not overfit like decision trees. It is further predicted on X_train and X_test. The evaluation of this model was done utilising evaluation metrics such as root mean square error, mean squared error, R^2 and training score. In this case for assessing Random Forest regressor model the RMSE value is observed to be 3.94 at training and 9.35 at testing. The R^2 value is observed to be 99.7% in training and 98.7% in testing for random forest regression. Thus, this model is observed to provide better performance and accuracy as compared to other models.

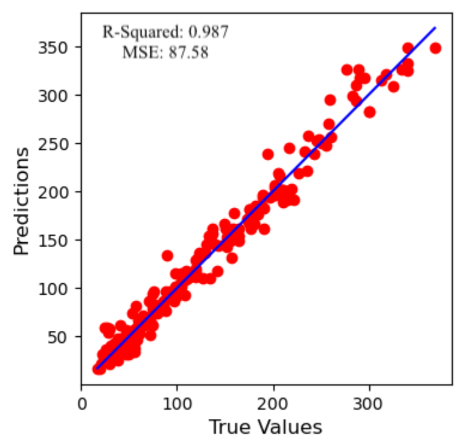


Figure 5.11: Scatter plot for True versus Predicted values variation for Random Forest Regressor

5.5.5 Using K-Nearest Neighbours Regression for AQI Prediction

The K-Nearest Neighbours regressor model nearest neighbours of the model is defined as 2 and it was further fitted in X_train and Y_train. This algorithm enables approximating relation among independent variables and continuous outcome through averaging observations of similar neighbourhoods. It is further predicted on X_train and X_test. The evaluation of this model was done utilising evaluation metrics such as root mean square error, mean squared error, R^2 and training score. In this case for assessing K-Nearest Neighbours regression model the RMSE value is observed to be 7.31 at training and 11.85 at testing. The R^2 value is observed to be 99% in training and 97.9% in testing for K-Nearest Neighbours regression.

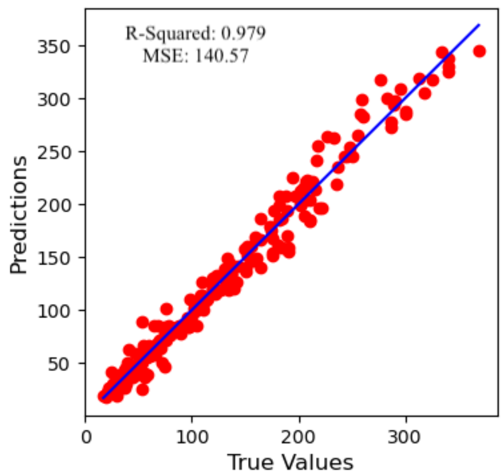


Figure 5.12: Scatter plot for True versus Predicted values variation for KNN Regressor

5.5.6 Using XGBoost Regression for AQI Prediction

The XGBoost regressor model was fitted in X_train and Y_train. This regression algorithm is a distributed and scalable gradient boosted decision tree and produces parallel tree boosting. Moreover, the boosting mechanism assists in reducing underfitting and bias. The evaluation of this model was done utilising evaluation metrics such as root mean square error, mean squared error, R^2 and training score. In this case for assessing XGBoost regression model the RMSE value is observed to be 0.7 at training and 9.53 at testing. The R^2 value is observed to be 99.9% in training and 98.6% in testing for XGBoost regression. Therefore, this model provides highest performance and accuracy after Random Forest regressor and can be taken into consideration for predictive analysis.

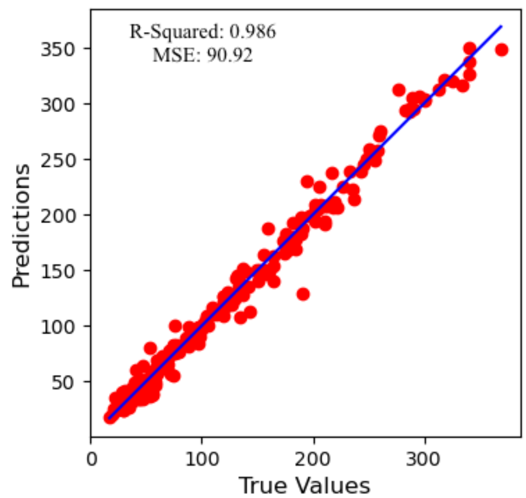


Figure 5.13: Scatter plot for True versus Predicted values variation for XGBoost Regressor

Table 5.5: Performance metrics of Machine Learning Regression models

Regression Algorithm	**Mean Absolute Error**	R^2 **Score**	**Mean Squared Error**	**Root Mean Squared Error**	**Training Score**
Linear Regression	9.7682	0.975	173.149	13.1586	0.9704
Decision Tree	7.9873	0.9792	145.2211	12.0507	1.0
Random Forest	5.9259	0.9874	87.5852	9.3586	0.9978
KNN	8.0088	0.9798	140.5702	11.8562	0.9925
XGBoost	6.2253	0.9869	90.9203	9.5352	0.9999

Based on the performance metrics discussed above, the Random Forest algorithm acquired slightly better results as compared to XGBoost. Linear Regression, Decision Tree and KNN algorithm models lacked efficiency, with high RMSE, MAE and lower R^2 value. Therefore, Random Forest algorithm performed efficiently, with RMSE = 9.3586, MAE = 5.9259 and R^2 = 0.9874. However, it was observed that the validation scores among testing set and training set were marginally different.

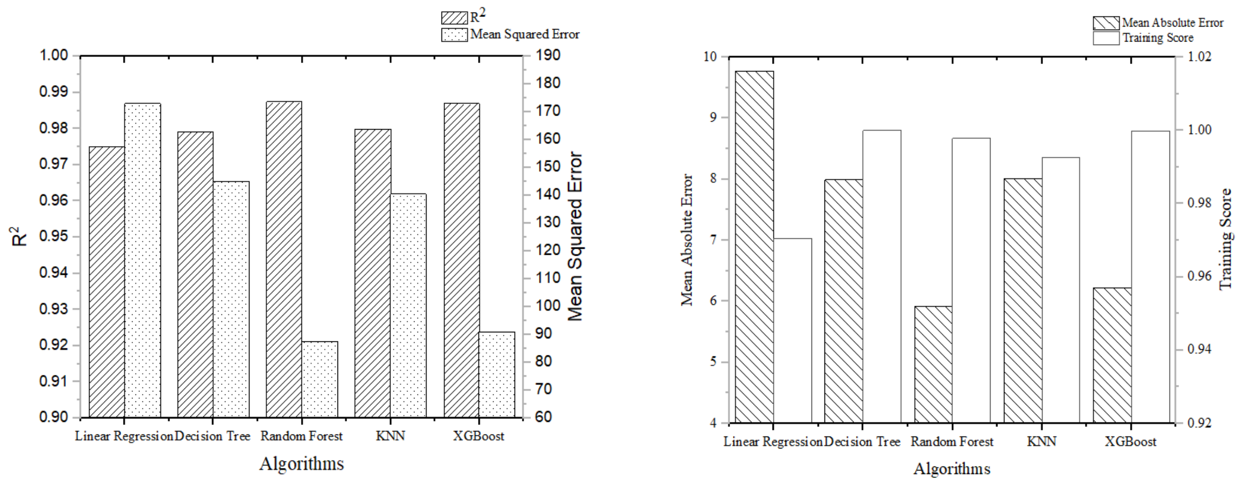


Figure 5.14: Comparison of performance metrics of various Machine Learning Algorithms

5.6 Time Series Analysis Using LSTM

The dataset possesses 7 features and the indexing of various features is done based on their data type, such as index of Date and Time is fixed through datetime datatype format. This enabled preparing time series dataset for implementation of LSTM modelling. The dataset is further segmented into two parts 80% for training and 20% for testing. After dividing training data and testing data, the data is scaled by normalisation method using MinMax Scaler. Thus, the value here is scaled between 0 and 1, it assists in understanding the patterns of the data effectively and further avoids the outliers. From here, the total day for train is 41.375 and total day for test is 13.083. Moreover, the number of time steps in input data has been considered to be 6 while number of time steps in output data is taken as 3 for this model. Following which the train and test data was converted to input and output data utilising split_sequence function.

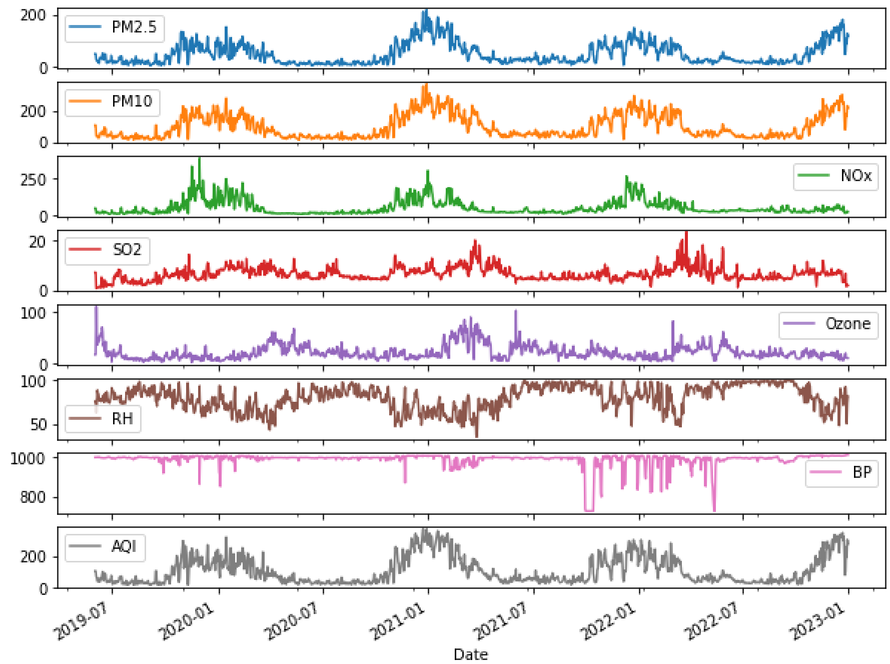


Figure 5.15: Variation of various pollutant with respect to time after predictive analysis

Furthermore, for model training LSTM layer is added with 200 units, including activation relu with input features. The repeat vector is also added along with output steps and the LSTM function is further defined for 200 units in order to acquire the last predicted value in the sequence. Moreover, for model fitting epoch is set as 50 and with Adam Optimizer loss, Mean Square Error, Mean Absolute Error is evaluated. The history of the model is acquired through validation of x_test and y_test. Further resizing and reshaping of predicted data is done to acquire suitable predicted value based on real data. The forecasting of various pollutants and AQI is acquired through this model. Although the accuracy of this model could be improved through hyperparameter tuning, epoch structure, layer, and model structure. The table below provides performance metrics of various pollutants using LSTM architecture.

Table 5.6: Performance metrics of various pollutant features using RNN-LSTM model

Sl. No.	Pollutant	RMSE	MSE	MAE	MAPE
1.	PM 2.5	14.2442	202.8980	10.1579	0.7716
2.	PM 10	46.4687	2159.3398	35.8662	0.6836
3.	Ozone	7.4429	55.3963	5.6175	0.2754
4.	SO2	14.9201	222.6106	10.5815	0.3179
5.	NO2	2.8121	7.9081	2.0128	0.3635
6.	NOx	8.9163	79.5021	6.4597	0.3435
7.	AQI	22.7757	518.7345	17.3519	0.6655

For analysis of overall air quality utilising RNN-LSTM, index for error evaluation error evaluation of predicted model is *Root Mean Squared Error (RMSE)*. The range of RMSE varies from zero to positive infinity, and smaller value indicates higher accuracy in prediction results. Moreover, after performing multiple iterations, repeated experiments and updated parameters, the value of error evaluation index for prediction of RNN-LSTM model for 0-24 hrs are illustrated below.

Table 5.7: Performance metrics of RNN-LSTM model

Evaluation Parameter	Value for LSTM Model
MAE	15.78
MSE	669.63
RMSE	25.87

The overall outcome of RNN-LSTM model is observed to attain higher accuracy for prediction. Moreover, significant disadvantages associated with vanishing gradients are resolved by using LSTM as it retains enough steep gradients and illustrates high accuracy and shorter duration for training. The performance metrics of RNN-LSTM model has been observed to be significantly efficient as the RMSE value was low with 25.87 although there has been scope of improvement by working with more data points and increasing batch size and number of epochs. It also depicted a significant mean absolute error value illustrating greater accuracy.

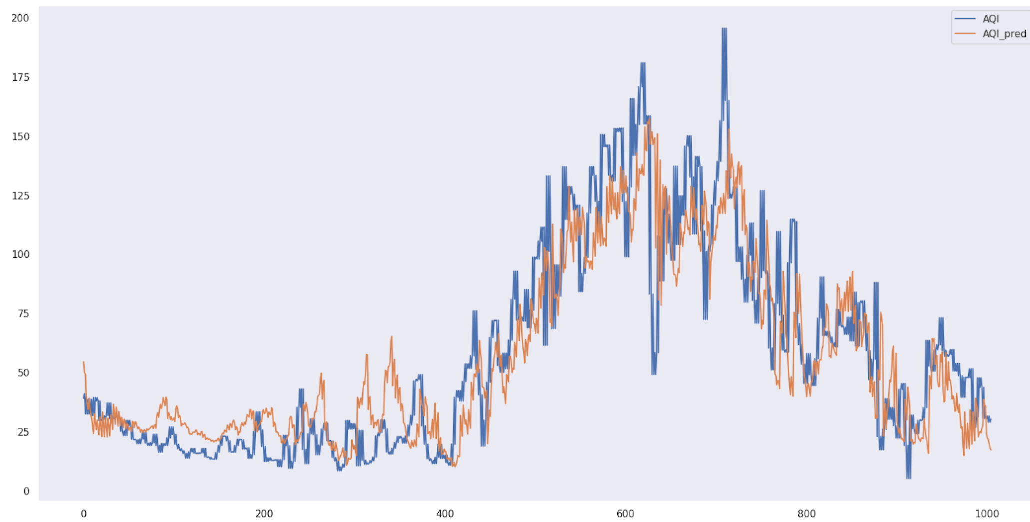


Figure 5.16: AQI Forecasting of predicted data after LSTM analysis

As observed in Fig. 5.16, RNN-LSTM curve fluctuation is stable and small, although it is hard to predict the transformation trend of concentration of AQI. Moreover, prediction accuracy of LSTM model reduced gradually with respect to time for next 48 hrs. However, it was able to accurately predict pollutant concentration trends for the future. From the above AQI forecasting curve, it is observed that trend of blue observation curve and yellow prediction curve are compatible. Furthermore, the formulated model demonstrated that for long-term prediction of pollutant concentration, concentration of various pollutants depict a strong correlation trend as compared to actual trend.

Chapter 6: Conclusion

This study takes into consideration lichen as a biomonitor of air pollution and implements a systematic approach for monitoring lichen biomass and cover in Jadavpur area of Kolkata which possess two significant traffic intersections nearby. It further depicts that there has been a significant increase in growth of lichen species in terms of surface area, present in roadside trees during monitoring duration of nine months. However, variety and quantity of lichen species observed in this region was less, indicating a moderate air quality in this region. The lichen species observed in this region were Caperat lichen (*Parmelia caperata*), Wreath lichen (*Cryptothecia subnidulans*) and *Diploicia canescens,* these species were predominantly observed more in barks of palm trees. Moreover, coverage of lichen species was identified employing a grid mapping method and further based on coverage and frequency of lichen species assisted in calculating ***index of atmospheric purity***. Therefore, IAP range calculated in significant palm trees has depicted a moderate pollution level for our study at the range 25 - 37.5. Furthermore, prevalent air quality in this region was also assessed through acquiring data from continuous monitoring stations of ***Central Pollution Control Board***, to determine concentration of particulates, SO_x, NO_x concentration and its impact on overall air quality.

Based on this study lichen diversity and growth in this region is hindered due increasing levels of pollutant concentration in this region as observed through CPCB data. The roadside plants are constantly exposed to enhanced levels of vehicular traffic. The seasonal variation in pollutant concentration and lichen growth depicts that during winter months the growth observed in lichen is negligible due to degrading air quality while lichen growth was observed to be at a faster pace during pre-monsoon and post-monsoon months. Prevalent impacts on lichen growth have also been closely observed due to changes in wind direction during various seasons. Moreover, this study formulated an effective model for predicting and forecasting air quality index. It is illustrated utilising dataset of various pollutants and air quality index for Jadavpur, Kolkata acquired through CPCB website. Moreover, various methods of pre-processing have been utilised for enhancing data representation that includes data normalisation, outlier removal, missing value imputation and feature selection. Machine learning and deep learning models employed for predicting are linear regression, decision tree regressor, random forest regressor, xgboost regressor, KNN regressor and RNN-LSTM. Among these models for predicting AQI Random Forest and XGBoost regressor illustrated better performance metrics as compared to other machine learning models. However, RNN-LSTM model was utilised for time series forecasting of predicted data and depicted an R^2 score of 76.44% and RMSE of 21. Thus, LSTM model provides output of pollutant concentration by providing input of temporal sequential data. In this segment a multivariate modelling of 7 pollutants taken into consideration has been done for time series forecasting. Moreover, performance metrics and accuracy obtained from this model could be stated as effective for forecasting of AQI. Although certain lag has been observed while prediction, there is scope of refinement of data through hyperparameter tuning and reduction in data imbalance. The proposed time series forecasting model could also be extended through bi-directional mechanism.

The present study illustrates the extent of pollutant concentration and further assesses the overall Air Quality Index through prediction modelling and time series forecasting to generate deeper understanding of the prevalent air quality scenario of Jadavpur area. This is further correlated with observed growth and diversity of lichen species found in this area, and whose sustenance is dependent on-air quality. Therefore, this study illustrates the scope of AQI and its possible menaces on abundance and diversity of lichen species found near significant traffic intersections of Kolkata region.

References

Abas, A. (2021) ‘A systematic review on biomonitoring using lichen as the biological indicator: A decade of practices, progress and challenges’, *Ecological Indicators*, 121, p. 107197. Available at: https://doi.org/10.1016/j.ecolind.2020.107197.

Abas, A. Z. L. A. N., & Awang, A. Z. A. H. A. N. (2017). Air pollution assessment using lichen biodiversity index (LBI) in Kuala Lumpur, Malaysia. *Pollution research*, *36*(2), 242-249.

Adamo, P., Giordano, S., Vingiani, S., Cobianchi, R. C., & Violante, P. (2003). Trace element accumulation by moss and lichen exposed in bags in the city of Naples (Italy). *Environmental pollution*, *122*(1), 91-103.

Anjum, S. S., Noor, R. M., Aghamohammadi, N., Ahmedy, I., Kiah, L. M., Hussin, N., ... & Qureshi, M. A. (2019). Modeling traffic congestion based on air quality for greener environment: an empirical study. *IEEE Access*, *7*, 57100-57119.

Aptroot, A., & Van Herk, C. M. (2007). Further evidence of the effects of global warming on lichens, particularly those with Trentepohlia phycobionts. *Environmental pollution*, *146*(2), 293-298.

Aptroot, A., & Van Herk, C. M. (2007). Further evidence of the effects of global warming on lichens, particularly those with Trentepohlia phycobionts. *Environmental pollution*, *146*(2), 293-298.

Armstrong, R. A., & Bradwell, T. (2011). Growth of foliose lichens: a review. *Symbiosis*, *53*, 1-16.

Armstrong, R., & Bradwell, T. (2010). Growth of crustose lichens: a review. *Geografiska Annaler: Series A, Physical Geography*, *92*(1), 3-17.

Aschenbrenner, I. A., Cernava, T., Berg, G., & Grube, M. (2016). Understanding microbial multi-species symbioses. *Frontiers in Microbiology*, *7*, 180.

Asta, J., Erhardt, W., Ferretti, M., Fornasier, F., Kirschbaum, U., Nimis, P. L., ... & Wirth, V. (2002). Mapping lichen diversity as an indicator of environmental quality. *Monitoring with lichens—monitoring lichens*, 273-279.

Augusto, S., Máguas, C., Matos, J., Pereira, M. J., Soares, A., & Branquinho, C. (2009). Spatial modeling of PAHs in lichens for fingerprinting of multisource atmospheric pollution. *Environmental science & technology*, *43*(20), 7762-7769.

Awasthi, D. D. (1988). A key to the macrolichens of India and Nepal. *The Journal of the Hattori Botanical Laboratory*, *65*, 207-302.

Bajpai, R., Upreti, D. K., & Dwivedi, S. K. (2010). Passive monitoring of atmospheric heavy metals in a historical city of central India by Lepraria lobificans Nyl. *Environmental monitoring and assessment*, *166*, 477-484.

BALASUBRAMANIAM, R. (2005). Do lichens still grow in Kolkata City?. *Current Science*, *88*(3), 338.

Bealey, W.J.; Long, S.; Spurgeon, D.J.; Leith, I.; Cape, J.N. Review and Implementation Study of Biomonitoring for Assessment of Air Quality Outcomes; Science Report SC030175/SR2; Environment Agency: Bristol, UK, 2008; pp. 1–170.

Begum, A., Ramaiah, M., Khan, I., & Veena, K. (2009). Analysis of heavy metals concentration in soil and litchens from various localities of Hosur road, Bangalore, India. E-journal of chemistry, 6(1), 13-22.

Bokhorst, S., Asplund, J., Kardol, P., & Wardle, D. A. (2015). Lichen physiological traits and growth forms affect communities of associated invertebrates. *Ecology*, *96*(9), 2394-2407.

Bsienvis.nic.in, (2022). "*LIST OF FAMILIES AND GENERA OF LICHEN IN INDIA"*. http://bsienvis.nic.in/Database/Lichen_22590.aspx

Budzenski, H., Jones, I., Bellocq, J., Pierad, C., & Garrigues, P. (1997). Evaluation of sediment contamination by polycyclic aromatic hydrocarbons in the Gironde estuary. Marine Chemistry, 58, 85–97.

Carreras, H. A., & Pignata, M. L. (2007). Effects of the heavy metals Cu2+, Ni2+, Pb2+, and Zn2+ on some physiological parameters of the lichen Usnea amblyoclada. *Ecotoxicology and Environmental Safety*, *67*(1), 59-66.

Chakraborty, A. (2014). Effects of air pollution on public health: the case of vital traffic junctions under Kolkata Municipal Corporation. *J Stud Dynam Change*, *1*(3), 125-133.

Chowdhury, I. R. (2015). Scenario of vehicular emissions and its effect on human health in Kolkata city. *International Journal of Humanities and Social Science Invention*, *4*(5), 1-9.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., ... & Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, 1-4.

Conti, M. E., & Cecchetti, G. (2001). Biological monitoring: lichens as bioindicators of air pollution assessment—a review. *Environmental pollution*, *114*(3), 471-492.

Conti, M. E., & Cecchetti, G. (2001). Biological monitoring: lichens as bioindicators of air pollution assessment—a review. *Environmental pollution*, *114*(3), 471-492.

CPCB. Parivesh Newsletter, Auto Emissions. 1999, 6(1).

Das, K., Nikita, B. P., Rani, A., & Uniyal, P. L. (2021). Lichens as bioindicators and biomonitoring agents. *Environ Intl J Sci Technology*, *15*, 18-25.

Das, M., Maiti, S. K., & Mukhopadhyay, U. (2006). Distribution of PM 2.5 and PM 10-2.5 in PM 10 fraction in ambient air due to vehicular pollution in Kolkata megacity. *Environmental Monitoring and Assessment*, *122*, 111-123.

Das, P., Joshi, S., Rout, J., & Upreti, D. K. (2013). Lichen diversity for environmental stress study: Application of index of atmospheric purity (IAP) and mapping around a paper mill in Barak Valley, Assam, northeast India. *Tropical ecology*, *54*(3).

Dathong, W. (2016). Epiphytic lichen diversity in different areas of Nakhon Ratchasima, Thailand.

Flesher, J. W., Horn, J., & Lehner, A. F. (2002). Role of the Bay-and L-regions in the metabolic activation and carcinogenicity of Picene and Dibenz [a, h] anthracene. *Polycyclic Aromatic Compounds*, *22*(3-4), 737-745.

Github.io, 2023, Mutual information-based feature selection, https://thuijskens.github.io/2017/10/07/feature-selection/

Glick, B. R. (2003). Phytoremediation: Synergistic use of plants and bacteria to clean up the environment. Biotechnology Advance, 21, 383–393.

Gottardini, E. L. E. N. A., Cristofolini, F. A. B. I. A. N. A., & Marchetti, F. L. A. V. I. O. (1999). Biomonitoraggio della qualità dell'aria della città di Trento tramite licheni epifiti. *Acqua & aria*, *30*(4), 67-71.

Green, T. A., Sancho, L. G., & Pintado, A. (2011). Ecophysiology of desiccation/rehydration cycles in mosses and lichens. *Plant desiccation tolerance*, 89-120.

Green, T. G., Sancho, L. G., & Pintado, A. (2011). Ecophysiology of desiccation/rehydration cycles in mosses and lichens. In *Plant desiccation tolerance* (pp. 89-120). Springer, Berlin, Heidelberg.

Gupta, S., Khare, R., Bajpai, O., Rai, H., Upreti, D. K., Gupta, R. K., & Sharma, P. K., 2016. Lichen As Bioindicator For Monitoring Environmental Status In Western Himalaya, India. *INTERNATIONAL JOURNAL OF ENVIRONMENT*.

Honegger, R. (2012). 15 The symbiotic phenotype of lichen-forming Ascomycetes and their endo-and epibionts. *Fungal associations*, 287-339.

Ibm.com, 2023, What Is Deep Learning, https://www.ibm.com/topics/deep-learning#:~:text=the%20next%20step-,What%20is%20deep%20learning%3F,from%20large%20amounts%20of%20data.

Kalapanidas, E., & Avouris, N. (2001). Short-term air quality prediction using a case-based classifier. Environmental Modelling & Software, 16(3), 263-272.

Khalili NR, Scheff PA, Holsen TM (1995) PAH source fingerprints for coke oven, diesel, and gasoline engines highway tunnels and wood combustion emissions. Atmos Environ 29(4):533–542

Khan, A., Sharma, S., Chowdhury, K. R., & Sharma, P. (2022). A novel seasonal index–based machine learning approach for air pollution forecasting. *Environmental Monitoring and Assessment*, *194*(6), 429.

Knops, J.M.H. *et al.* (1991) 'Mineral Cycling and Epiphytic Lichens: Implications at the Ecosystem Level', *The Lichenologist*, 23(3), pp. 309–321. Available at: https://doi.org/10.1017/S0024282991000452.

Kuldeep, S. and Prodyut, B. (2015) 'Lichen as a Bio-Indicator Tool for Assessment of Climate and Air Pollution Vulnerability: Review', 4.

Kumar, K., & Pande, B. P. (2022). Air pollution prediction with machine learning: A case study of Indian cities. *International Journal of Environmental Science and Technology*, 1-16.

Kumar, K., & Pande, B. P. (2022). Air pollution prediction with machine learning: A case study of Indian cities. *International Journal of Environmental Science and Technology*, 1-16.

Kumar, P. G., Lekhana, P., Tejaswi, M., & Chandrakala, S. (2021). Effects of vehicular emissions on the urban environment-a state of the art. *Materials Today: Proceedings*, *45*, 6314-6320.

LeBlanc, F., G. Robitaille & D. N. Rao. 1974. Biological response of lichens and bryophytes to environmental pollution in the Murdochville copper mine area, Quebec. Hattori Botanical Laboratory 38: 405-433.

LeBlanc, F., Robitaille, G., & Rao, D. N. (1974). Biological response of lichens and bryophytes to environmental pollution in the Murdochville copper mine area, Quebec. *J. Hattori Bot. Lab.;(Japan)*, *38*.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, *525*(7569), 367-371.

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, *23*, 22408-22417.

Loksabhadocs.nic.in, 2023, Vehicular Pollution In India, https://loksabhadocs.nic.in/Refinput/New_Reference_Notes/English/Vehicular%20pollution%20in%20India.pdf

Londhe, M. (2021). Data mining and machine learning approach for air quality index prediction. *International Journal of Engineering and Applied Physics*, *1*(2), 136-153.

Loppi, S. (2014). Lichens as sentinels for air pollution at remote alpine areas (Italy). *Environmental Science and Pollution Research*, *21*, 2563-2571.

Loppi, S. (2019). May the diversity of epiphytic lichens be used in environmental forensics?. *Diversity*, *11*(3), 36.

Loppi, S., Frati, L., Paoli, L., Bigagli, V., Rossetti, C., Bruscoli, C., & Corsini, A. (2004). Biodiversity of epiphytic lichens and heavy metal contents of Flavoparmelia caperata thalli as indicators of temporal variations of air pollution in the town of Montecatini Terme (central Italy). *Science of the total environment*, *326*(1-3), 113-122.

Ma, L. L., Chu, S. G., Wang, X. T., Cheng, H. X., & Lin, X. Xu. (2005). Polycyclic aromatic hydrocarbons in the surface soils from outskirts of Beijing, China. Chemosphere, 58, 1355–1363.

Macrotrends.net, 2023, *Calcutta, India Metro Area Population 1950-2023,* https://www.macrotrends.net/cities/21211/calcutta/population#:~:text=The%20metro%20area%20population%20of,a%200.64%25%20increase%20from%202019.

Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019, March). A machine learning model for air quality prediction for smart cities. In

2019 International conference on wireless communications signal processing and networking (WiSPNET) (pp. 452-457). IEEE.

Majumdar, D., Purohit, P., Bhanarkar, A. D., Rao, P. S., Rafaj, P., Amann, M., ... & Srivastava, A. (2020). Managing future air quality in megacities: Emission inventory and scenario analysis for the Kolkata Metropolitan City, India. *Atmospheric Environment*, *222*, 117135.

Majumdar, D., Ray, S., Chakraborty, S., Rao, P. S., Akolkar, A. B., Chowdhury, M., & Srivastava, A. (2014). Emission, speciation, and evaluation of impacts of non-methane volatile organic compounds from open dump site. *Journal of the Air & Waste Management Association*, *64*(7), 834-845.

Majumdar, S., Ram, S. S., Jana, N. K., Santra, S., Chakraborty, A., & Sudarshan, M. (2009). Accumulation of minor and trace elements in lichens in and around Kolkata, India: an application of X-ray fluorescence technique to air pollution monitoring. *X-Ray Spectrometry: An International Journal*, *38*(6), 469-473.

Majumder, S., Mishra, D., Ram, S. S., Jana, N. K., Santra, S., Sudarshan, M., & Chakraborty, A. (2013). Physiological and chemical response of the lichen, Flavoparmelia caperata (L.) Hale, to the urban environment of Kolkata, India. *Environmental Science and Pollution Research*, *20*, 3077-3085.

Majumder, S., Mishra, D., Ram, S. S., Jana, N. K., Santra, S., Sudarshan, M., & Chakraborty, A. (2013). Physiological and chemical response of the lichen, Flavoparmelia caperata (L.) Hale, to the urban environment of Kolkata, India. *Environmental Science and Pollution Research*, *20*, 3077-3085.

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 1-36.

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 1-36.

Meng, X., Fu, H., Peng, L., Liu, G., Yu, Y., Wang, Z., & Chen, E. (2020). D-LSTM: short-term road traffic speed prediction model based on GPS positioning data. *IEEE Transactions on Intelligent Transportation Systems*, *23*(3), 2021-2030.

Mežaka, A., Brūmelis, G., & Piterāns, A. (2008). The distribution of epiphytic bryophyte and lichen species in relation to phorophyte characters in Latvian natural old-growth broad leaved forests. *Folia Cryptogamica Estonica*, *44*, 89-99.

Mikhaylov, A. (2020). Lichens as indicators of atmospheric pollution in urban ecosystems. *Israel Journal of Ecology and Evolution*, *67*(1-2), 60-68.

Muhammad, N., Hashim, N. H., Khairuddin, N. A., Yusof, H., Jusoh, S., Abas, A., ... & Din, L. (2018). Identification of most tolerant lichen species to vehicular traffic's pollutants: a case study at Batu Pahat. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences*, *42*(1), 57-64.

Munzi, S., Correia, O., Silva, P., Lopes, N., Freitas, C., Branquinho, C., & Pinho, P. (2014). Lichens as ecological indicators in urban areas: beyond the effects of pollutants. *Journal of Applied Ecology*, *51*(6), 1750-1757.

Nandini, K., & Fathima, G. (2019, March). Urban Air Quality Analysis and Prediction Using Machine Learning. In *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)* (pp. 98-102). IEEE.

Negi, H. R. (2003). Lichens: a valuable bioresource for environmental monitoring and sustainable development. *Resonance*, *8*, 51-58.

Nylander, W. (1867). *Lichenes Kurziani e Calcutta…*

Palomäki, V., Tynnyrinen, S., Holopainen, T., 1992. Lichen transplantation in monitoring fluoride and sulfur deposition in the surroundings of a fertilizer plant and a strip mine at Siilinjarvi. Annales Botanici Fennici 29, 25–34.

Paoli, L., & Loppi, S. (2008). A biological method to monitor early effects of the air pollution caused by the industrial exploitation of geothermal energy. *Environmental Pollution*, *155*(2), 383-388.

Pasupuleti, V. R., Kalyan, P., & Reddy, H. K. (2020, March). Air quality prediction of data log by machine learning. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1395-1399). IEEE.

Phruksahiran, N. (2021). Improvement of air quality index prediction using geographically weighted predictor methodology. *Urban Climate*, *38*, 100890.

Piervittori, R., Usai, L., Alessio, F., Maffei, M., 1997. The effect of simulated acid rain on surface morphology and n-alkane composition of Pseudevernia furfuracea. Lichenologist 29, 191–198.

Preethaa, K.S. (2021). Lichen Element Based Autonomous Air Pollution Monitoring Around Smart Cities–A Deep Learning Approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 151-161.

Rybarczyk, Y., & Zalakeviciute, R. (2021). Assessing the COVID-19 impact on air quality: A machine learning approach. *Geophysical Research Letters*, *48*(4), e2020GL091202.

Sanders, W. B. (2001). Lichens: The Interface between Mycology and Plant Morphology: Whereas most other fungi live as an absorptive mycelium inside their food substrate, the lichen fungi construct a plant-like body within which photosynthetic algal symbionts are cultivated. *Bioscience*, *51*(12), 1025-1035

Seaward, M.R.D., 1976. Performance of Lecanora muralis in an urban environment. In: Brown, D.H., Hawksworth, D.L., Bailey, R.H. (Eds), Systematics Association, Special Volume 8, Lichenology: Progress and Problems. Academic Press, London, pp. 323–357.

Sharma, M., Jain, S., Mittal, S., & Sheikh, T. H. (2021). Forecasting And Prediction Of Air Pollutants Concentrates Using Machine Learning Techniques: The Case Of India. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012123). IOP Publishing.

Sharma, S., Zhang, M., Gao, J., Zhang, H., & Kota, S. H. (2020). Effect of restricted emissions during COVID-19 on air quality in India. *Science of the total environment*, *728*, 138878.

Shukla V, Upreti DK (2009) Polycyclic Aromatic Hydrocarbon (PAH) accumulation in lichen, Phaeophyscia hispidula of DehraDun city, Garhwal Himalayas. Environ Monit Assess 149(1–4):1–7

Shukla, V., & Upreti, D. K. (2007). Physiological response of the lichen Phaeophyscia hispidula (Ach.) Essl., to the urban environment of Pauri and Srinagar (Garhwal), Himalayas, India. Environmental Pollution, 150, 295–299. doi:10.1016/j.envpol.2007. 02.010.

Shukla, V., & Upreti, D. K. (2011). Changing lichen diversity in and around urban settlements of Garhwal Himalayas due to increasing anthropogenic activities. *Environmental monitoring and assessment*, *174*(1), 439-444.

Shukla, V., Patel, D. K., Upreti, D. K., & Yunus, M. (2012). Lichens to distinguish urban from industrial PAHs. *Environmental chemistry letters*, *10*(2), 159-164.

Siew, L. Y., Chin, L. Y., & Wee, P. M. J. (2008). ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. *Malaysian Journal of Analytical Sciences*, *12*(1), 257-263.

Silberstein, L., Siegel, B. Z., Siegel, S. M., Mukhtar, A., & Galun, M. (1996). Comparative Studies Onxanthoria Parietina, A Pollution-Resistant Lichen, Andramalina Duriaei, A Sensitive Species. Ii. Evaluation of Possible Air Pollution-Protection Mechanisms. *The Lichenologist*, *28*(4), 367-383.

Spiroska, J., Rahman, A., & Pal, S. (2011). Air pollution in Kolkata: An analysis of current status and interrelation between different factors. *SEEU Review*, *8*(1), 182-214.

Sujetovienė, G., & Česynaitė, J. (2021). Assessment of air pollution at the indoor environment of a shooting range using lichens as biomonitors. *Journal of Toxicology and Environmental Health, Part A*, *84*(7), 273-278.

Tiwari, P. K. (2008). Lichens as indicator for Air Pollution: A Review. *Indian Journal of J.*

Pollut. Control, *1*, 8-17.

Upreti, D. K., & Patel, D. K. (2012). Rinodina sophodes (Ach.) Massal.: a bioaccumulator of polycyclic aromatic hydrocarbons (PAHs) in Kanpur City, India. *Environmental monitoring and assessment*, *184*, 229-238.

Van, N. H., Van Thanh, P., Tran, D. N., & Tran, D. T. (2023). A new model of air quality prediction using lightweight machine learning. *International Journal of Environmental Science and Technology*, *20*(3), 2983-2994.

Williams, L., Colesie, C., Ullmann, A., Westberg, M., Wedin, M., & Büdel, B. (2017). Lichen acclimation to changing environments: Photobiont switching vs. climate-specific uniqueness in Psora decipiens. *Ecology and Evolution*, *7*(8), 2560-2574.

Wolterbeek, H. T., Garty, J., Reis, M. A., & Freitas, M. C. (2003). Biomonitors in use: lichens and metal air pollution. In *Trace metals and other contaminants in the environment* (Vol. 6, pp. 377-419). Elsevier.

Xu, X., Xu, Y. L., Ren, Y., & Huang, Q. (2021). Site-specific extreme load estimation of a long-span cable-stayed bridge. *Journal of Bridge Engineering*, *26*(4), 05021001.

Yatawara, M., & Dayananda, N. (2019). Use of corticolous lichens for the assessment of ambient air quality along rural–urban ecosystems of tropics: a study in Sri Lanka. *Environmental monitoring and assessment*, *191*, 1-14.

Yuan, Z., Liu, J., Zhang, Q., Liu, Y., Yuan, Y., & Li, Z. (2021). Prediction and optimisation of fuel consumption for inland ships considering real-time status and environmental factors. *Ocean Engineering*, *221*, 108530.

Zhang, Q., Gao, T., Liu, X., & Zheng, Y. (2020). Public environment emotion prediction model using LSTM network. *Sustainability*, *12*(4), 1665.

Zhang, Z., Ye, L., Qin, H., Liu, Y., Wang, C., Yu, X., ... & Li, J. (2019). Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression. *Applied energy*, *247*, 270-284.

Appendix

Table 1: Lichen Monitoring data

Monitoring Dates	Growth in surface area of Lichen samples (mm^2)									
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
03/07/22	2.522	1.129	0.878	0.431	7.317	1.003	0.541	1.317	3.532	4.097
18/07/22	2.534	1.156	0.901	0.452	7.330	1.023	0.553	1.329	3.545	4.105
04/08/22	2.573	1.194	0.915	0.476	7.351	1.034	0.582	1.345	3.591	4.110
20/08/22	2.597	1.202	0.921	0.489	7.377	1.061	0.591	1.365	3.625	4.119
05/09/22	2.609	1.218	0.933	0.509	7.397	1.087	0.607	1.370	3.633	4.125
15/09/22	2.622	1.240	0.943	0.528	7.406	1.102	0.615	1.385	3.642	4.132
25/09/22	2.635	1.378	1.021	0.543	7.423	1.111	0.629	1.397	3.652	4.148
05/10/22	2.642	1.450	1.031	0.572	7.431	1.124	0.639	1.407	3.674	4.166
15/10/22	2.661	1.464	1.101	0.591	7.445	1.136	0.648	1.415	3.693	4.178
22/10/22	2.697	1.478	1.137	0.641	7.457	1.154	0.707	1.423	3.745	4.196
04/11/22	2.722	1.497	1.156	0.680	7.466	1.182	0.714	1.440	3.933	4.208
22/11/22	2.753	1.521	1.212	0.697	7.524	1.192	0.755	1.456	4.037	4.219
03/12/22	2.761	1.586	1.220	0.721	7.585	1.213	0.766	1.475	4.042	4.245
20/12/22	2.796	1.721	1.234	0.759	7.642	1.219	0.778	1.500	4.169	4.256
11/01/23	2.830	1.732	1.244	0.763	7.704	1.247	0.814	1.504	4.204	4.271
21/01/23	2.851	1.821	1.258	0.783	7.842	1.269	0.874	1.510	4.217	4.289
01/02/23	2.883	1.873	1.263	0.826	7.881	1.354	0.902	1.524	4.366	4.305
15/02/22	2.898	1.993	1.271	0.867	7.892	1.359	0.915	1.533	4.380	4.317
03/03/23	2.903	2.076	1.279	0.888	7.958	1.380	0.937	1.539	4.429	4.332
19/03/23	2.910	2.108	1.281	0.943	7.997	1.393	0.944	1.544	4.444	4.356
27/03/23	2.923	2.135	1.292	0.993	8.194	1.450	0.958	1.547	4.469	4.374

Table 2: Seasonal Dataset of Air Quality parameters from 2019-2023

Summer'19

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
June	29.55	49.74	4.04	13.23	17.25	1.79	6.15	37.75	79.10	998.66	120.13	0.69	56.32	Satisfactory

Monsoon'19

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
July	20.67	34.52	5.07	11.50	16.57	4.66	21.42	16.83	80.61	997.91	126.73	0.64	37.57	Good
August	19.24	31.22	10.42	14.63	25.07	2.84	14.94	8.58	88.10	998.48	128.56	0.51	36.11	Good
September	12.82	22.48	7.39	9.62	17.00	3.79	17.02	10.10	87.72	998.36	120.30	0.42	26.08	Good
October	55.01	83.39	26.18	21.03	47.21	6.20	13.89	15.15	82.57	NaN	192.24	0.27	97.86	Satisfactory

Winter'19

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
November	79.98	160.77	102.23	41.73	141.95	6.16	24.25	13.92	74.83	999.26	238.00	0.27	180.27	Moderate
December	65.40	157.79	62.17	47.51	105.57	6.16	25.92	11.61	69.46	NaN	215.16	0.30	158.64	Moderate
January	68.75	145.89	82.40	47.91	127.43	8.86	32.11	9.99	73.89	NaN	209.41	0.33	151.03	Moderate
February	77.74	154.30	61.48	50.95	110.23	8.61	38.16	15.00	62.33	NaN	210.66	0.38	164.07	Moderate

Figure 1: Seasonal dataset of year 2019

Summer'20

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
March	49.28	92.13	14.16	24.24	38.39	8.27	18.68	27.86	62.06	NaN	202.63	0.46	98.29	Satisfactory
April	22.56	48.07	3.35	9.46	12.81	7.26	8.16	39.25	70.43	NaN	146.62	0.53	62.72	Satisfactory
May	13.36	34.58	3.31	7.61	10.93	6.57	5.90	35.02	77.32	997.73	140.96	0.68	46.90	Good
June	12.28	31.37	3.80	10.31	14.11	5.28	5.50	25.43	84.63	997.95	116.34	0.58	34.77	Good

Monsoon'20

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
July	13.17	30.22	4.65	10.06	14.71	7.56	5.32	22.20	85.06	998.04	114.15	0.58	30.71	Good
August	13.71	31.02	5.96	11.37	17.34	5.17	6.62	20.15	89.83	996.16	130.81	0.51	33.87	Good
September	14.03	32.69	6.87	12.28	19.14	4.68	7.38	16.03	85.85	998.71	130.24	0.47	31.96	Good
October	30.73	68.75	11.86	24.62	36.47	5.10	11.12	15.51	79.87	999.42	175.12	0.39	67.38	Satisfactory

Winter'20

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
November	77.81	158.33	46.09	51.69	96.04	8.76	51.05	17.07	61.42	NaN	238.02	0.32	171.07	Moderate
December	142.10	243.54	74.03	52.00	124.01	7.60	35.73	14.22	67.46	NaN	238.51	0.29	293.87	Poor
January	138.09	233.53	49.45	49.82	97.62	7.22	30.98	22.30	64.19	NaN	221.69	0.33	292.03	Poor
February	105.50	203.41	55.68	53.12	108.85	9.44	29.27	49.31	58.11	NaN	221.05	0.41	220.33	Poor

Figure 2: Seasonal dataset of year 2020

SUMMER'21

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
March	59.81	147.29	15.07	29.73	44.80	11.57	16.27	55.53	61.71	996.52	174.92	0.45	130.13	Moderate
April	33.20	84.58	7.09	16.96	24.05	9.59	8.75	33.51	69.41	999.02	144.06	0.54	84.54	Satisfactory
May	21.99	46.82	6.44	13.73	20.10	7.27	8.04	14.18	79.69	997.80	143.62	0.69	47.00	Good
June	28.04	49.03	6.53	20.60	27.12	4.53	12.10	37.83	90.96	997.02	123.50	0.95	54.20	Satisfactory

MONSOON'21

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
July	24.60	46.98	8.43	12.43	20.85	4.66	10.63	27.74	93.48	996.04	135.20	0.93	48.97	Good
August	29.02	50.50	9.34	11.34	19.36	4.78	7.82	23.84	94.89	998.13	136.07	0.81	52.35	Satisfactory
September	21.50	45.13	15.18	19.93	34.91	4.97	12.43	17.42	95.08	985.45	122.81	0.73	45.47	Good
October	55.77	93.88	19.51	15.29	33.97	5.78	57.85	25.17	89.99	777.67	190.50	0.71	119.86	Moderate

WINTER'21

Month	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	AQI_Bucket
November	79.31	158.97	48.68	26.19	72.46	5.17	88.06	18.99	75.43	NaN	231.60	0.56	166.43	Moderate
December	94.10	191.08	63.20	80.25	142.03	5.58	18.66	12.33	82.71	NaN	236.86	0.59	204.00	Poor
January	138.09	233.53	49.45	49.82	97.62	7.22	30.98	22.30	64.19	NaN	221.69	0.33	292.03	Poor
February	105.50	203.41	55.68	53.12	108.85	9.44	29.27	49.31	58.11	NaN	221.05	0.41	220.33	Poor

Figure 3: Seasonal dataset of year 2021

Summer'22

MONTH	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	
March	56.38	131.05	34.68	25.46	58.45	11.55	13.61	32.45	69.23	1000.00	185.39	1.02	122.90	Moderate
April	20.44	49.51	9.64	22.73	31.99	9.61	5.48	30.81	89.92	999.31	113.44	1.45	49.53	Good
May	27.26	59.02	6.54	19.20	25.73	7.52	6.64	29.72	90.98	951.86	130.27	1.02	57.50	Satisfactory
June	26.47	52.34	11.41	20.92	32.32	4.85	5.97	25.19	96.38	996.84	122.66	1.19	53.79	Satisfactory

Monsoon'22

MONTH	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	
July	17.76	31.45	7.36	24.46	31.81	4.90	3.26	17.81	98.27	996.70	104.82	0.87	33.90	Good
August	19.93	41.10	14.83	18.72	32.76	5.03	9.74	19.52	97.81	997.32	133.77	0.88	42.36	Good
September	21.78	46.37	15.46	20.76	36.16	5.24	63.92	19.14	98.23	980.50	131.66	0.73	46.81	Good
October	36.79	72.78	9.85	13.96	23.81	7.00	57.85	17.31	86.57	1004.44	193.46	0.70	75.67	Satisfactory

Winter'22

MONTH	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	
November	93.32	174.38	25.38	22.22	38.55	6.81	88.06	17.91	68.28	1008.23	260.04	0.56	193.81	Moderate
December	128.82	222.16	37.27	22.30	40.72	5.68	81.33	15.09	72.93	1009.02	235.89	0.32	272.68	Poor
January	121.18	197.07	15.25	21.41	23.05	4.67	16.60	11.94	79.78	1,010.99	214.87	0.37	264.15	Poor
February	75.21	137.07	16.52	14.69	28.06	7.15	14.15	20.75	72.33	1,008.54	198.63	0.47	154.57	Moderate

Figure 3: Seasonal dataset of year 2022

Summer'23

MONTH	PM2.5	PM10	NO	NO2	NOx	SO2	NH3	Ozone	RH	BP	WD	WS	AQI	
March	51.44	95.51	22.38	29.53	50.22	7.65	74.33	31.48	71.39	1,007.13	181.86	0.41	103.70	Moderate
April	45.01	81.27	14.00	23.99	37.31	12.32	99.61	53.84	71.82	1,003.27	197.56	0.38	88.93	Satisfactory

Figure 4: Seasonal dataset of year 2023