***A Project Report On***

**COMPARATIVE ANALYSIS OF DEEP LEARNING AND STATISTICAL MODELS FOR AIR POLLUTANTS PREDICTION IN URBAN AREAS**

*Submitted in partial fulfillment of the requirement for the award*

The degree of

# MASTER OF TECHNOLOGY

**In**

**DATA SCIENCE**

Submitted By

# CH.SAI GOPALA KRISHNA DEEKSHIT

# Regd.No. VP22CSEN0200011

### Under the guidance

*Of*

# Dr. Pragnyaban Mishra

### Professor



DEPARTMENTOFCOMPUTER SCIENCE &ENGINEERING,

GITAM SCHOOL OF TECHNOLOGY, GITAM (DEEMED TO Be UNIVERSITY), RUSHIKONDA, VISAKHAPATNAM 530045 ANDHRAPRADESH, INDIA.

**2024**

# DECLARATION

This is to certify that the project titled **“**Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas**”** submitted by me to the Department of Computer Science &Engineering, GITAM School of Technology, GITAM Deemed to be University, Visakhapatnam, India, for the award of the degree of Master of Technology in DATA SCIENCE is a record of project work carried out by me. This work has not been submitted to any other university or institution for the award of any degree.

**CH.SAI GOPALA KRISHNA DEEKSHIT**

Station: Visakhapatnam

DEPARTMENTOFCOMPUTER SCIENCE&ENGINEERING

GITAM SCHOOL OF TECHNOLOGY GITAM(DEEMEDTOBEUNIVERSITY)

RUSHIKONDA, VISAKHAPATNAM-530045.

**CERTIFICATE**



This is to certify that the Project report entitled “**Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas”** submitted by **CH.SAI GOPALA KRISHNA DEEKSHIT Regd.No. VP22CSEN0200011** to the Department of Computer Science &Engineering, GITAM School of Technology, GITAM Deemed to be University, Visakhapatnam, INDIA, for the award of the degree of **MASTER OF TECHNOLOGY in DATA SCIENCE** is a bonafide record of project carried out by him under my supervision. The contents of the project report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

# Dr. Pragnyaban Mishra Dr. R. SIREESHA

# Professor Professor

PROJECT GUIDE HEAD OF THE DEPARTMENT

**EVALUATION REPORT**

Name of the candidate : CH. Sai Gopala Krishna Deekshit

Name of the guide : Dr. Pragnyaban Mishra

Title of the project : Comparative Analysis of Deep Learning and Statistical Models for Air Pollutants Prediction in Urban Areas

Specialization in MTech :DATA SCIENCE

Date of examination/Viva-voce:

This dissertation is approved by the Board of Examiners. Chairman, Board of examiners :

Chairman,Board of studies :

Internal examiner :

# ACKNOWLEDGEMENTS

I am grateful to GITAM (DEEMED TO BEUNIVERSITY) for providing mewith the resources and facilities necessary to conduct my research. The DEPARTMENT OFCOMPUTER SCIENCE & ENGINEERING staff and faculty have been extremely helpful in providing mewith guidance and support, and I appreciate their assistance throughout my research journey. I would like to express my deepest gratitude to my project guide for their guidance, support, and encouragement throughout the entire Project process. Their in valuable insights, constructive feedback, and unwavering dedication have been instrumental in shaping this Project .I would also like to thank the members of all my faculty members from the DEPARTMENT OFCOMPUTER SCIENCE & ENGINEERING, for their insightful comments and suggestions, which havehelpedmerefinemy research and improvethe qualityofthis Project.

-CH SAI GOPALA KRISHNA DEEKSHIT

# ABSTRACT

Air pollution and poor air quality are caused by the rapid rise of urbanization and industrialization. It has been deemed a "silent public health emergency" because of its detrimental impact on both the environment and human health. Accurate air pollution prediction is crucial for stakeholders to take the necessary steps to address this worldwide issue. Deep learning-based forecasting algorithms have recently shown promise for more accurate and efficient air quality forecasting than conventional methods. In this work, we made a comparative analysis of various deep learning-based single-step forecasting models such as long short-term memory (LSTM), gated recurrent unit (GRU), and a statistical model to predict five air pollutants namely Nitrogen Dioxide (NO2), Ozone (O3), Sulphur Dioxide (SO2), and Particulate Matter (PM2.5, and PM10). For empirical evaluation, we used a publicly available dataset collected in Northern Ireland, using an air quality monitoring station situated in Belfast city centre. It measures the concentration of air pollutants. The performance of forecasting models is evaluated based on three performance metrics such as root mean square error (RMSE), mean absolute error (MAE) and R-squared (R2 ). The result shows that deep learning models consistently achieved the least RMSE compared to the statistical models. In addition, the deep learning model is also found to have the highest R2 score.

**Keywords: Air quality, machine learning, deep learning, predictive models, statistical methods.**

# LISTOFCONTENTS

|  |  |  |  |
| --- | --- | --- | --- |
| **TopicNo.** | | **Topic** | **Page No.** |
| 1. | | Introduction | 8-9 |
|  | 1.1 | Motivation | 9 |
|  | 1.2 | Scope | 9 |
|  | 1.3 | Area Description | 9 |
| 2 | | Problem Statement & Objective | 10 |
|  | 2.1 | Problem Statement | 10 |
|  | 2.2 | Objective | 10 |
| 3 | | Project Description& Project Analysis | 11-14 |
|  | 3.1 | Project Description | 11 |
|  | 3.2 | Project Analysis | 11-14 |
| 4 | | Literature Survey | 15-18 |
| 5 | | References | 19 |

**1. INTRODUCTION**

Over the past few years, air pollution has become a major global challenge. Air pollution has a direct impact not only on the environment but also on human health and well-being. It has been observed that air pollution leads to increase mortality and morbidity such as respiratory diseases, impaired cognitive function, cardiovascular diseases, and cancer. There are multiple factors involved in deteriorating air quality such as manufacturing, industrial emissions, transportation (in land, air and sea) emissions, dust, and coal consumption. Education and raising public awareness relating to this issue requires interdisciplinary approaches with professionals and other stakeholders. Local councils are playing their part and have set up many air quality monitoring stations throughout the country to monitor the concentration of air pollutants. Data collected from such monitoring stations can be used in the prediction of pollutants. Prediction of air quality is important to control air pollution and to identify areas which require solutions to overcome air pollution and related impacts. However, how to model air quality accurately is a challenge on its own and depends on available data and modelling approaches. In recent years, deep learning-based forecasting models show promise for more effective and efficient forecasting of air quality than other approaches.

Education and raising public awareness relating to this issue requires interdisciplinary approaches with professionals and other stakeholders. Local councils are playing their part and have set up many air quality monitoring stations throughout the country to monitor the concentration of air pollutants. Data collected from such monitoring stations can be used in the prediction of pollutants. Prediction of air quality is important to control air pollution and to identify areas which require solutions to overcome air pollution and related impacts. However, how to model air quality accurately is a challenge on its own and depends on available data and modelling approaches. The main contributions of this study include: (1) We propose a combination of meteorological parameters such as temperature, wind speed and wind direction with lagged air quality feature which is based on the concentration value of the previous hour for the pollutant being predicted for the next hour. We also make use of the datetime index by splitting it into hour, day, and month for additional features to improve prediction accuracy and reduction of error. (2) We propose a comprehensive study to predict the five major air pollutants and a comparison of deep learning base models with statistical model. (3) We provide detail architectural and parameters information for both deep learning and statistical models, to predict each pollutant, which can be useful in various applications in the scope of smart cities or can provide benchmarking for more better and accurate models

**1.1 Motivation:**

Air pollution and poor air quality are caused by the rapid growth of urbanization and industrialization. It has been deemed a "silent public health emergency" due to its detrimental effects on both human health and the environment. Accurate air pollution prediction is essential for stakeholders to take the necessary action in order to address this global challenge. Compared to other methods, deep learning-based forecasting models have shown promise in recent years for more accurate and efficient air quality forecasting.

**1.2 Scope:**

We suggest combining lagged air quality feature, which is based on the concentration value of the previous hour for the pollutant being predicted for the next hour with meteorological parameters like temperature, wind speed, and wind direction. Additionally, we utilize the date time index by dividing it into hour, day, and month segments for extra features that enhance prediction precision and minimize error.

We suggest a thorough investigation to forecast the five main air pollutants as well as a comparison between statistical and deep learning foundation models.

In order to predict each pollutant, we offer detailed architectural and parameter information for both statistical and deep learning models. This information can be used for benchmarking purposes or in a variety of smart city applications.

**1.3 Area Description:**

A number of factors, including manufacturing, industrial emissions, dust, coal consumption, and transportation (by land, air, and sea), are contributing to the declining quality of the air. The release of toxic materials and gases into the atmosphere is known as air pollution, and it is a major health risk to people and other living things. Pollutants are these dangerous substances (solids, liquids, or gases). The quality of our environment is diminished and there are major negative health effects when certain pollutants, such as particulate matter, are produced in higher concentrations than usual.

**2. Problem Statement & Objective**

**2.1 Problem Statement:**

To analyze the influential factors of air quality on a national scale, a non-linear framework based on Geographic Information System (GIS) and Extreme Gradient Boosting (XGBoost) was introduced in the literature. They looked into the most crucial elements affecting air quality by gathering and analyzing 171 features, including environmental, demographic, economic, meteorological, and energy-related ones. To model the relationship and determine the variable importance, XGBoost, a non-linear machine learning algorithm, was applied. To pre-process the diverse variables and visualize the outcomes, GIS was used.

**2.2 Objective of the Project:**

For empirical assessment, we employed a publicly available dataset collected in Northern Ireland using an air quality monitoring station situated in Belfast city center. It determines the amount of air pollutants that are present. Forecasting models are assessed using three performance metrics, including R-squared (R2), mean absolute error (MAE), and root mean square error (RMSE).

**3. Project Description& Project Analysis**

**3.1 Project Description**

This project aims to utilize deep learning techniques to forecast five air pollutants: nitrogen dioxide (NO2), ozone (O3), sulphur dioxide (SO2), and particulate matter (PM2.5 and PM10). The deep learning models contains recurrent neural network (RNN), which is based on sequence or time series data and found to provide improved performance in applications like natural language processing and speech recognition. The RNN model has memory units to capture and learn dependencies between input and output over short or long term. However, with the increase in the network layers and iterations, RNN tends to forget the dependencies and suffer with vanishing gradient problem.

* Data Collection
* Data Pre-processing
* Deep Learning Model
* Feature Extraction
* Model Training and Evaluation
* Implementation
* Ethical Considerations
* Conclusion

**3.2 Project Analysis**

**3.2.1Existing System:**

In literature they introduced a non-linear framework based on Extreme Gradient Boosting (XGBoost) and Geographic Information System (GIS) to analyze the influential factors of air quality on a national scale. They collected and analyzed 171 features ranging from environmental, demographical, economic, meteorological, and energy to investigate the most important factors of air quality. XGBoost, is a non-linear machine learning algorithm, was utilized to model the relationship and measure the variable importance. GIS was employed to preprocess the diversified variables and visualize the results. Their methodology framework also compares the classification performance of XGBoost with other machine learning models to show the reasonability of choice. A case study was conducted in America to validate the effectiveness of their framework. Experimental results show that six kinds of factors have the largest impact on air quality, and practical suggestions are proposed from the six aspects to control and prevent air pollution.

Disadvantages:

* The existing work employs the XGBoost algorithm for analyzing influential factors, which may not be suitable for time series data.
* The existing work doesn't conduct a comparative analysis of different approaches. This omission could lead to a lack of understanding about the relative strengths and weaknesses of various techniques.
* The existing work evaluates its results based on classification performance, which might not directly align with forecasting accuracy.

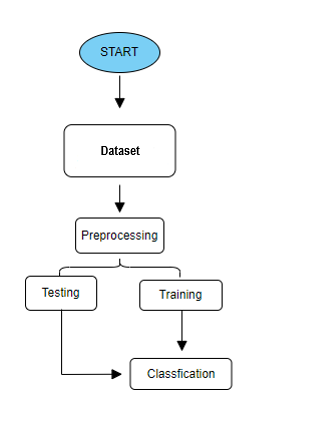
**3.2.2 Proposed method**

We conducted a comparative analysis of several single step forecasting models based on deep learning, including long short term memory (LSTM), gated recurrent unit (GRU), and An auto regressive integrated moving average (ARIMA) used as a statistical model to forecast five air pollutants: nitrogen dioxide (NO2), ozone (O3), sulphur dioxide (SO2), and particulate matter (PM2.5 and PM10). We used a publicly accessible dataset gathered in Northern Ireland utilizing an air quality monitoring station located in Belfast city center for empirical evaluation. It calculates the quantity of air contaminants present. The performance of forecasting models is evaluated based on three performance metrics such as root mean square error (RMSE), mean absolute error (MAE) and R-squared (R2 ).

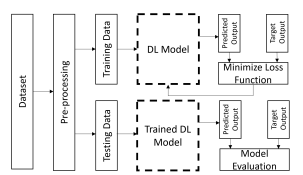
Advantages:

1. Our work employs a range of advanced forecasting models including LSTM, GRU, and ARIMA, which are specifically designed for time series data. This diversity allows for a more comprehensive exploration of forecasting techniques, potentially capturing complex temporal patterns in air pollutant levels.
2. Our work, on the other hand, utilizes RMSE, MAE, and R-squared as robust metrics specifically tailored for forecasting evaluations.
3. We directly address the challenge of forecasting air pollutant levels. This provides actionable insights for future pollution levels, enabling proactive interventions and planning for pollution control measures.
4. We explicitly accounts for temporal variations and trends in air pollutant levels. This ensures a more accurate understanding of short-term fluctuations and long-term patterns, which can greatly enhance the ability to predict pollution levels.
5. The inclusion of deep learning techniques like LSTM and GRU allows the present work to capture intricate temporal relationships in the data. These techniques can better handle complex patterns, leading to potentially more accurate forecasting results.

**Work Flow of Proposed system**

****

**Project Architecture**

****

**4. LITERATURE SURVEY**

1. M. Stafoggia and T. Bellander, ‘‘Short-term effects of air pollutants on daily mortality in the Stockholm county—A spatiotemporal analysis,’’ Environ. Res., vol. 188, Sep. 2020, Art. no. 109854.

Short-term exposure to air pollutants has been extensively related to daily mortality, however most of the evidence comes from studies conducted in major cities, and little is known on the extent of the spatial heterogeneity in the effects within areas including both urban and non-urban settings. We aimed to investigate the short-term association of air pollutants with daily cause-specific mortality in the Stockholm county, and to test whether an association exists also outside the metropolitan area. We used a spatiotemporal random forest model to predict daily concentrations of fine and inhalable particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2) and ozone (O3) at 1-km spatial resolution over Sweden for 2005–2016. We collected data on daily mortality for each small area for market statistics (SAMS) of the Stockholm county, to which we matched daily exposures to air pollutants and air temperature. We applied a case-crossover design to investigate the short-term association between the four pollutants and mortality from non-accidental, cardiovascular and respiratory causes. We compared the associations in and out the Stockholm urban area, by SAMS population density and across the 26 municipalities of the county. We found weak effects of most air pollutants on cause-specific mortality in the full year analysis, with estimates much larger and significant only during the warmer months (April to September): non-accidental mortality increased by 4.58% (95% confidence interval – 95% CI: 0.89%, 8.41%) and by 2.21% (95% CI: 0.71%, 3.73%) per 10 μg/m3 increase in lag 0–1 PM2.5 and O3, respectively. Associations were in general higher in the Stockholm city and in SAMS with high population density. When comparing the 26 municipalities, we didn't detect a significant heterogeneity in the short-term associations with air pollutants. In conclusion, we found a suggestion of a harmful role of air pollution also in non-urban areas, but the study was underpowered to draw firm conclusions. We consider this study as a pilot to investigate the spatial heterogeneity of the association between daily air pollution and mortality at the national level in Sweden.

2. K. Abutalip, A. Al-Lahham, and A. El Saddik, ‘‘Digital twin of atmospheric environment: Sensory data fusion for high-resolution PM2.5 estimation and action policies recommendation,’’ IEEE Access, vol. 11, pp. 14448–14457, 2023.

Particulate matter smaller than 2.5 microns (PM2.5) is one of the main pollutants that has considerable detrimental effects on human health. Estimating its concentration levels with ground monitors is inefficient for several reasons. In this study, we build a digital twin (DT) of an atmospheric environment by fusing remote sensing and observational data. An integral part of the DT pipeline is the presence of feedback that can influence future input data. Estimated values of PM2.5 obtained from an ensemble of Random Forest and Gradient Boosting are used to provide recommendations for decreasing the agglomeration levels. We formulate a simple optimization problem for suggesting the recommendations and identify possible action policies, such as cloud seeding, scheduling of air filtering, and SMS notifications. The PM2.5 estimation part of the proposed DT pipeline has achieved RMSE and R2 of 38.94 and 0.728 (95%, CI 0.717-0.740). In addition, we investigate different approaches for quantitatively examining the contribution of each independent variable.

3. J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, Y. Tan, V. J. L. Gan, and Z. Wan, ‘‘Identification of high impact factors of air quality on a national scale using big data and machine learning techniques,’’ J. Cleaner Prod., vol. 244, Jan. 2020, Art. no. 118955.

To effectively control and prevent air pollution, it is necessary to study the influential factors of air quality. A number of previous studies have explored the relationships between air pollution and related factors. However, the methods currently used either cannot well address the multicollinearity problem or fail to explain the importance of the influential factors. Moreover, most of the existing literature limited their studied area in a city or a small region and studied factors in one aspect. There is a lack of studies that analyze the influential factors from the perspective of a country or take into consideration multiple variables. To fill the research gap, this paper proposes a multivariate analysis in the national scale to investigate the most important factors of air quality. In order to study as much influential factors as possible, 171 features ranging from environmental, demographical, economic, meteorological, and energy, were collected and analyzed. To tackle such a “big data” problem, a non-linear machine learning algorithm namely Extreme Gradient Boosting (XGBoost) is utilized to model the relationship and measure the variable importance. Geographical Information System (GIS) is employed to pre-process the diversified variables and visualize the results. Performance of XGBoost is compared with other models and its parameters are tuned using Bayesian Optimization. Experimental results of a case study in the U.S. show that our methodology framework can effectively uncover the important factors of air quality. Six kinds of factors are found to have the largest impact on air quality. Practical suggestions are also proposed from the six aspects to control and prevent air pollution.

4. I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, ‘‘Environmental and health impacts of air pollution: A review,’’ Frontiers Public Health, vol. 8, p. 14, 2020.

One of our era's greatest scourges is air pollution, on account not only of its impact on climate change but also its impact on public and individual health due to increasing morbidity and mortality. There are many pollutants that are major factors in disease in humans. Among them, Particulate Matter (PM), particles of variable but very small diameter, penetrate the respiratory system via inhalation, causing respiratory and cardiovascular diseases, reproductive and central nervous system dysfunctions, and cancer. Despite the fact that ozone in the stratosphere plays a protective role against ultraviolet irradiation, it is harmful when in high concentration at ground level, also affecting the respiratory and cardiovascular system. Furthermore, nitrogen oxide, sulfur dioxide, Volatile Organic Compounds (VOCs), dioxins, and polycyclic aromatic hydrocarbons (PAHs) are all considered air pollutants that are harmful to humans. Carbon monoxide can even provoke direct poisoning when breathed in at high levels. Heavy metals such as lead, when absorbed into the human body, can lead to direct poisoning or chronic intoxication, depending on exposure. Diseases occurring from the aforementioned substances include principally respiratory problems such as Chronic Obstructive Pulmonary Disease (COPD), asthma, bronchiolitis, and also lung cancer, cardiovascular events, central nervous system dysfunctions, and cutaneous diseases. Last but not least, climate change resulting from environmental pollution affects the geographical distribution of many infectious diseases, as do natural disasters. The only way to tackle this problem is through public awareness coupled with a multidisciplinary approach by scientific experts; national and international organizations must address the emergence of this threat and propose sustainable solutions.

5. S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, ‘‘Comparative analysis of machine learning techniques for predicting air quality in smart cities,’’ IEEE Access, vol. 7, pp. 128325–128338, 2019.

Dealing with air pollution presents a major environmental challenge in smart city environments. Real-time monitoring of pollution data enables local authorities to analyze the current traffic situation of the city and make decisions accordingly. Deployment of the Internet of Things-based sensors has considerably changed the dynamics of predicting air quality. Existing research has used different machine learning tools for pollution prediction; however, comparative analysis of these techniques is required to have a better understanding of their processing time for multiple datasets. In this paper, we have performed pollution prediction using four advanced regression techniques and present a comparative study to determine the best model for accurately predicting air quality with reference to data size and processing time. We have conducted experiments using Apache Spark and performed pollution estimation using multiple datasets. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have been used as evaluation criteria for the comparison of these regression models. Furthermore, the processing time of each technique through standalone learning and through fitting the hyperparameter tuning on Apache Spark has also been calculated to find the best-fit model in terms of processing time and lowest error rate.

# 5.REFERENCES

# [1] R. D. Brook, ‘‘Cardiovascular effects of air pollution,’’ Clin. Sci., vol. 115, no. 6, pp. 175–187, Sep. 2008.

# [2] M. Stafoggia and T. Bellander, ‘‘Short-term effects of air pollutants on daily mortality in the Stockholm county—A spatiotemporal analysis,’’ Environ. Res., vol. 188, Sep. 2020, Art. no. 109854.

# [3] WHO. Household Air Pollution and Health. Accessed: Dec. 29, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ household-air-pollution-and-health

# [4] WHO. Air Quality and Health. Accessed: Jan. 10, 2023. [Online]. Available: <https://www.who.int/teams/environment-climate-change-andhealth/air-quality-and-health/policy-progress/sustainable-developmentgoals-air-pollution>

# [5] B. Paul and S. Louise. (2022). Air Quality: Policies, Proposals and Concerns—House of Commons Library. Accessed: Jan. 10, 2023. [Online]. Available: <https://commonslibrary.parliament.uk/researchbriefings/cbp-9600/>

# [6] P. J. Landrigan, ‘‘Air pollution and health,’’ Lancet Public Health, vol. 2, pp. e4–e5, Jan. 2017.

# [7] K. Abutalip, A. Al-Lahham, and A. El Saddik, ‘‘Digital twin of atmospheric environment: Sensory data fusion for high-resolution PM2.5 estimation and action policies recommendation,’’ IEEE Access, vol. 11, pp. 14448–14457, 2023.

# [8] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, Y. Tan, V. J. L. Gan, and Z. Wan, ‘‘Identification of high impact factors of air quality on a national scale using big data and machine learning techniques,’’ J. Cleaner Prod., vol. 244, Jan. 2020, Art. no. 118955.

# [9] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, ‘‘Environmental and health impacts of air pollution: A review,’’ Frontiers Public Health, vol. 8, p. 14, 2020.

# [10] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, ‘‘Comparative analysis of machine learning techniques for predicting air quality in smart cities,’’ IEEE Access, vol. 7, pp. 128325–128338, 2019.