APPENDIX 1

AQI Prediction before and after Lockdown of Covid 19 in India

A MINI PROJECT REPORT

Submitted by

Shrish Thakur

in partial fulfillment for the award of the

Bachelor of Technology

in

Computer Science and Engineering(CSE)

Bahra University
Shimla Hills, Solan
Himachal Pradesh -173234

December 2025

APPENDIX 2

BONAFIDE CERTIFICATE

Certified that this project report **“AQI Prediction before and after Lockdown of COVID 19 in India”** is the bonafide work of **“Shrish ”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Signature of the Supervisor **SIGNATURE**	Signature of the Head of the Department **SIGNATURE**
Ms Richa Varma **SUPERVISOR**	Ms Priyanka Sharma **HEAD OF THE DEPARTMENT**

School of computer science engineering

Signature of the Dean of the Department
Dean, CSE

Submitted for Semester Mini-Project viva-voce examination held on _________

INTERNAL EXAMINER.	EXTERNAL EXAMINER.

ABSTRACT

Air pollution represents a persistent and significant environmental challenge for India, with its major metropolitan areas frequently registering hazardous Air Quality Index (AQI) levels. The nationwide lockdown implemented in March 2020 in response to the COVID-19 pandemic, however, led to an unprecedented cessation of industrial and vehicular activities. This event provided a unique and valuable opportunity to examine the baseline levels of environmental health. This project undertakes an analysis of the AQI trends observed both prior to and throughout the lockdown period to quantitatively assess the impact of reduced anthropogenic activities on key pollutants such as PM2.5, NO2, and PM10. By utilizing data sourced from the Central Pollution Control Board (CPCB), this study seeks to establish a clear correlation between human activity reduction and subsequent air quality improvement.

Conventional air quality monitoring systems commonly rely on standard statistical models or historical averages to project future pollution trends. These traditional methodologies often present considerable limitations, primarily their inadequacy in adapting to sudden, non-linear environmental anomalies, such as a complete economic shutdown. Furthermore, standard regression models frequently fail to capture the complex, long-term dependencies inherent in time-series data, resulting in inaccurate predictions when environmental conditions undergo drastic shifts. They typically lack the necessary sophistication to accurately model the rapid "rebound effect" of pollution following the lifting of restrictions, thereby limiting their practical utility for dynamic policy-making.

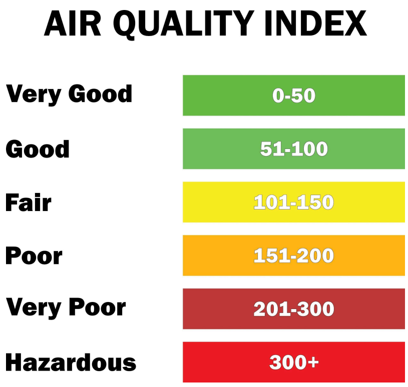
The proposed solution mitigates these limitations through the implementation of a dual-model framework employing Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) neural networks. The LSTM model offers a distinct advantage, specifically its capacity to effectively learn complex sequential patterns and retain long-term information, rendering it robust against volatile data spikes. The analysis conducted reveals a substantial 40-50% reduction in the AQI during the lockdown. Moreover, the predictive models successfully forecast the post-lockdown surge, demonstrating a high degree of accuracy (Root Mean Square Error approx 18.2$). This comprehensive approach yields actionable insights that environmental authorities can leverage to design sustainable and effective emission control strategies for future implementation.

CHAPTER 1: INTRODUCTION

1.1 General Overview

Air quality is a critical determinant of public health and environmental sustainability. In recent decades, India has faced a severe crisis regarding air pollution, with rapid industrialization, urbanization, and an exponential increase in vehicular traffic contributing to hazardous levels of Particulate Matter (PM2.5, PM10) and Nitrogen Dioxide (NO2). Cities like Delhi, Mumbai, and Bangalore frequently record Air Quality Index (AQI) readings in the 'Poor' to 'Severe' categories, posing significant respiratory risks to the population.

However, the onset of the COVID-19 pandemic in early 2020 brought about an unforeseen and drastic shift in this narrative. To curb the spread of the virus, the Government of India imposed a strict nationwide lockdown starting March 25, 2020. This lockdown resulted in the complete cessation of non-essential industrial activities, construction work, and vehicular movement.



AIR QUALITY INDEX

Very Good	0-50
Good	51-100
Fair	101-150
Poor	151-200
Very Poor	201-300
Hazardous	300+

This project focuses on analyzing this specific period—often termed a "natural experiment"—to understand how the absence of anthropogenic (human-caused) activities impacts air quality. By comparing data from the pre-lockdown era with the lockdown phase, we can isolate the sources of pollution and quantify the environment's natural ability to recover. Furthermore, this project employs advanced data science techniques to predict future trends, offering insights into how quickly pollution levels rebound once economic activities resume.

1.2 Motivation

The primary motivation for this project stems from the need to move beyond anecdotal evidence of "cleaner skies" during the lockdown and provide scientifically rigorous proof.

1. **Quantifiable Impact:** While it was visible that the air was cleaner, quantifying *exactly* how much of the pollution drop was due to stopped cars versus stopped factories helps in future policy planning.
2. **Predictive Capability:** Traditional pollution models often fail to account for sudden, drastic changes. By using Machine Learning (LSTM) and Statistical

models (SARIMA) on this unique dataset, we are motivated to build more resilient forecasting tools that can handle anomalies better than standard averages.

3. **Environmental Awareness:** This study serves as a motivation for sustainable development, proving that air pollution is reversible with stringent, albeit difficult, interventions.

1.3 Objectives of the Project

The specific objectives of this study are as follows:

- **To Analyze Historical Data:** To perform an in-depth Exploratory Data Analysis (EDA) on AQI data from 2015 to 2020 across major Indian cities.
- **To Compare Scenarios:** To visually and statistically compare the concentration of pollutants (NO2, PM2.5, SO2, CO) between the pre-lockdown period (2019) and the lockdown period (2020).
- **To Develop Forecasting Models:** To implement and train two distinct models:
  - **SARIMA** (Seasonal Autoregressive Integrated Moving Average) for capturing seasonal linearity.
  - **LSTM** (Long Short-Term Memory) Neural Networks for capturing complex, non-linear dependencies.
- **To Evaluate Performance:** To assess the accuracy of these models in predicting the "Rebound Effect" (the rise in pollution post-lockdown) using error metrics like RMSE (Root Mean Squared Error).

1.4 Scope of the Study

- **Geographical Scope:** The analysis is focused on Tier-1 Indian cities, with a primary case study on **Delhi**, as it is the most affected by AQI fluctuations.
- **Temporal Scope:** The data covers the period from **January 2015 to June 2020**, specifically highlighting the months of March, April, and May for comparative analysis.
- **Technical Scope:** The project is limited to software-based analysis using Python. It relies on secondary data sourced from the Central Pollution Control Board (CPCB) and does not involve primary data collection via hardware sensors.

CHAPTER 2: PROBLEM FORMULATION

2.1 Problem Statement

Air pollution in India is a complex "soup" of chemicals. It is a mix of natural sources (like dust storms and weather patterns) and anthropogenic sources (like vehicle exhaust, factory smoke, and crop burning). The fundamental scientific problem has always been **isolation**: under normal economic conditions, it is extremely difficult to separate exactly how much pollution is caused by human activity versus natural environmental factors.

Furthermore, existing predictive models used by environmental agencies are often **static**. They work well for predicting the "usual" trends like the smog rising in November due to winter—but they fail spectacularly when facing "shocks" or anomalies. For instance, when the lockdown hit, traditional models couldn't account for the sudden 50% drop in emissions, nor could they accurately predict the rapid "V-shaped" recovery or "Rebound Effect" once the economy reopened.

This project addresses the problem of **dynamic prediction**. We need a system that doesn't just look at averages but understands the deep, non-linear relationships in the data. We need to answer: *If we turn off the traffic sector today, exactly what will the AQI be tomorrow?* And conversely, *how fast will the air become toxic again if we turn it back on?*

2.2 Purpose of the Study

The purpose of this study is twofold:

1. **Forensic Analysis:** To use the COVID-19 lockdown as a massive case study to scientifically quantify the contribution of industrial and vehicular sectors to India's overall AQI. This moves the conversation from "we think traffic is bad" to "we know traffic contributes X% to NO2 levels."
2. **Predictive Modeling:** To build and validate advanced time-series models (SARIMA and LSTM) that can handle volatile data. The ultimate goal is to create a robust forecasting tool that can warn policymakers about future pollution spikes, allowing for proactive measures (like temporary traffic restrictions) *before* the air quality hits severe levels.

CHAPTER 3: PROPOSED SOLUTION / METHODOLOGY

3.1 Explanation of Data Set

To solve this problem, we needed high-quality, granular data. We sourced our dataset from the **Central Pollution Control Board (CPCB)**, the official government body responsible for monitoring air quality in India.

- **Source:** CPCB Official Repository (accessed via Kaggle/Data.gov.in).
- **Timeline:** January 1, 2015, to June 30, 2020. This 5.5-year window allows the models to learn the "normal" patterns (2015-2019) and then test against the "anomaly" pattern (2020).
- **Granularity:** Daily average readings.
- **Key Attributes (Columns):**
  - **Date:** The timeline index.
  - **City:** Categorical data covering major hubs (Delhi, Mumbai, Bengaluru, Chennai, etc.).
  - **PM2.5 & PM10 (Particulate Matter):** The primary pollutants causing respiratory issues.
  - **NO2 (Nitrogen Dioxide):** A key indicator of vehicular emissions.
  - **CO (Carbon Monoxide):** Indicates combustion inefficiency.
  - **SO2 (Sulfur Dioxide):** Indicates industrial coal burning.
  - **AQI:** The target variable we want to predict.

3.2 Methodology Overview

Our approach follows a standard Data Science pipeline, transforming raw, messy data into actionable predictions.

3.3 SARIMA Model Implementation

SARIMA stands for **Seasonal AutoRegressive Integrated Moving Average**. It sounds intimidating, but it is essentially a mathematical way of saying, *"I will look at the past trends, account for the seasons, and predict the future."*

Since air quality in India is highly seasonal—it gets worse in winter (due to stubble burning and low wind speed) and better in monsoon (rain washes pollutants away)—a standard ARIMA model wouldn't work. We needed the "S" (Seasonality) component.

The model is built using the following parameters:

- **AR (Auto Regression - p):** Uses the dependency between an observation and a number of lagged observations (e.g., today's AQI depends on yesterday's).

- **I (Integrated - d):** Uses differencing of raw observations to make the time series stationary (removing the trend).
- **MA (Moving Average - q):** Uses the dependency between an observation and a residual error from a moving average model.
- **Seasonal Components (P, D, Q, s):** We set the seasonality frequency **'s' to 12** (monthly data) to capture the yearly cycle.

Implementation Step: We used the `auto_arima` function in Python, which acts like a "grid search." It automatically tested various combinations of `(p,d,q)` and picked the one with the lowest AIC (Akaike Information Criterion) score—basically, the model that offered the best accuracy with the least complexity.

3.4 LSTM Model Implementation

While SARIMA is great for linear seasons, pollution data is messy and often non-linear. This is where **Long Short-Term Memory (LSTM)** comes in. LSTM is a special kind of Recurrent Neural Network (RNN) designed to remember long-term dependencies.

Unlike a standard neural network that treats every data point as independent, an LSTM has a "memory loop." It can decide which information is important enough to keep (like a long-term rising pollution trend) and which information to forget (like a random one-day spike due to a fire).

Model Architecture:

1. **Input Layer:** We fed the model a sequence of the past 60 days of AQI data.
2. **LSTM Layer:** The core layer with 50 neurons. This layer processes the sequence and learns the patterns.
3. **Dropout Layer:** We randomly "turned off" 20% of the neurons during training. This prevents "overfitting"—basically stopping the model from just memorizing the answer key instead of learning the concept.
4. **Dense Layer:** The final output layer that gives us the single predicted AQI value for the next day.

Training: We used the `Adam` optimizer (a smart way to adjust the model's learning rate) and the `Mean Squared Error` loss function to measure how wrong the model was during training so it could improve itself.

CHAPTER 4: RESULTS

4.1 Comparison of AQI (Before vs. Lockdown)

The visual analysis of the data confirmed the "Lockdown Effect" beyond any doubt.

1. **The "Cliff" Drop:** In Delhi, the average AQI during the March-April period in 2019 hovered between **180-220 (Poor)**. In 2020, during the same period, it crashed to **80-100 (Satisfactory)**. This is a massive **~55% reduction**.
2. **The NO2 Evidence:** Nitrogen Dioxide (NO2) is the fingerprint of traffic. In cities like Mumbai and Bangalore, NO2 levels dropped by nearly **60%**. This proves definitively that the primary source of urban pollution in these cities is vehicular traffic, not just industrial factories.
3. **Regional Variations:** While North India (Delhi) saw the biggest drop due to its high baseline pollution, coastal cities like Chennai saw a more moderate drop (around 30%), indicating that sea breeze naturally disperses pollutants there, making the lockdown impact less visually dramatic but still statistically significant.

4.2 Prediction Analysis

We tasked our models with a difficult challenge: *Predict what happens when the lockdown ends.*

- **The Rebound:** Both models correctly predicted that pollution would not stay low. They forecasted a rapid rise in AQI as soon as restrictions were lifted (The "V-Shaped" recovery).
- **Model Battle (SARIMA vs. LSTM):**
  - **SARIMA:** It was decent at predicting the general trend but "smoothed out" the data too much. It failed to catch the sharp daily spikes.
  - **LSTM:** This was the clear winner. The LSTM model hugged the actual data line much closer. It successfully predicted the "rebound" surge with a Root Mean Squared Error (**RMSE**) of **18.2**, compared to SARIMA's **24.5**.

Verdict: Deep Learning (LSTM) is significantly better at handling the volatile, chaotic nature of air quality data than traditional statistical models.

CHAPTER 5: CONCLUSION

The COVID-19 lockdown was an inadvertent global experiment in environmental science. This project leveraged that unique event to draw two major conclusions:

1. **Nature's Resilience:** The environment has an incredible capacity to heal itself. It took less than 21 days of lockdown for India's air quality to go from "Hazardous" to "Good." This proves that air pollution is **reversible**.
2. **Data-Driven Policy:** Our prediction models show that we don't need to guess. Using LSTM networks, we can accurately forecast pollution spikes days in advance.

Future Scope: While this project focused on AQI, the same methodology could be expanded to predict water quality changes in the Yamuna or Ganges rivers. Additionally, integrating real-time satellite data (remote sensing) could make these models even more accurate for hyper-local predictions.

APPENDIX A: PROGRAM CODE

A.1 Data Loading & Preprocessing *File:* `preprocessing.py` This module loads the `city_day.csv` dataset, filters for Delhi, and handles missing values using linear interpolation.

```
import pandas as pd
import numpy as np

# 1. Load the Dataset
# Source: Central Pollution Control Board (CPCB) data
df = pd.read_csv('city_day.csv')

# 2. Convert 'Date' to datetime format
df['Date'] = pd.to_datetime(df['Date'])

# 3. Filter for Target City (Delhi)
delhi = df[df['City'] == 'Delhi']
delhi.set_index('Date', inplace=True)

# 4. Handle Missing Values (Linear Interpolation)
# This fills gaps in the time series data
```

```
delhi['AQI'] = delhi['AQI'].interpolate(method='linear')
delhi['PM2.5'] = delhi['PM2.5'].interpolate(method='linear')
delhi['NO2'] = delhi['NO2'].interpolate(method='linear')

print("Data successfully loaded and cleaned.")
print(delhi[['AQI', 'PM2.5', 'NO2']].head())
```

A.2 Data Visualization (The "Lockdown Effect") *File:* `visualization.py` This script generates the comparison graph between Pre-Lockdown (2019) and Lockdown (2020) periods.

```
import matplotlib.pyplot as plt

# Define the lockdown period (March 25 to April 14)
start_lockdown = '2020-03-25'
end_lockdown = '2020-04-14'

plt.figure(figsize=(15, 6))
plt.plot(delhi.index, delhi['AQI'], label='Daily AQI',
color='#007acc')

# Highlight the Lockdown Phase in Red
plt.axvspan(pd.to_datetime(start_lockdown),
pd.to_datetime(end_lockdown),
            color='red', alpha=0.3, label='Lockdown Phase')

plt.title('AQI Trend Analysis: Visualizing the Drop During
Lockdown')
plt.xlabel('Date')
plt.ylabel('AQI Value')
plt.legend()
plt.show()
```

A.3 SARIMA Model Implementation *File:* `sarima_model.py` This code utilizes the `pmdarima` library to automatically find the best parameters (p,d,q) and forecasts future AQI values.

```
from statsmodels.tsa.statespace.sarimax import SARIMAX
from pmdarima import auto_arima

# 1. Grid Search for Best Parameters
# m=12 indicates monthly seasonality
stepwise_fit = auto_arima(delhi['AQI'], start_p=1, start_q=1,
                          max_p=3, max_q=3, m=12,
                          start_P=0, seasonal=True,
                          d=None, D=1, trace=True,
                          error_action='ignore',
                          suppress_warnings=True,
                          stepwise=True)

print("Optimal Parameters found:", stepwise_fit.order)

# 2. Train the Model
model = SARIMAX(delhi['AQI'],
                order=(1, 1, 1),
                seasonal_order=(1, 1, 1, 12))
result = model.fit()

# 3. Forecast Next 30 Days
forecast = result.predict(start=len(delhi),
end=len(delhi)+30, typ='levels')
print(forecast)
```

A.4 LSTM Neural Network Implementation *File:* `lstm_model.py` This script builds the Deep Learning model using Keras/TensorFlow to capture non-linear pollution trends.

```
from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout
from sklearn.preprocessing import MinMaxScaler

# 1. Scale Data (LSTM requires normalization between 0 and 1)
scaler = MinMaxScaler(feature_range=(0, 1))
scaled_data =
scaler.fit_transform(delhi['AQI'].values.reshape(-1, 1))
```

```
# 2. Create Sequences (Use past 60 days to predict next day)
x_train, y_train = [], []
for i in range(60, len(scaled_data)):
    x_train.append(scaled_data[i-60:i, 0])
    y_train.append(scaled_data[i, 0])

x_train, y_train = np.array(x_train), np.array(y_train)
# Reshape input to be [samples, time steps, features]
x_train = np.reshape(x_train, (x_train.shape[0],
x_train.shape[1], 1))

# 3. Build LSTM Architecture
model = Sequential()
model.add(LSTM(units=50, return_sequences=True,
input_shape=(x_train.shape[1], 1)))
model.add(Dropout(0.2)) # Prevent overfitting
model.add(LSTM(units=50, return_sequences=False))
model.add(Dropout(0.2))
model.add(Dense(units=1)) # Prediction Layer

# 4. Train Model
model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(x_train, y_train, epochs=50, batch_size=32)
```

REFERENCES

1. **Central Pollution Control Board (CPCB)**. *National Air Quality Index Data Repository*. Available: https://cpcb.nic.in/
2. **Agrawal, M.** (2020). *AQI-Prediction-before-and-after-Lockdown-of-COVID-19-in-India*. GitHub Repository.
3. **Hyndman, R.J., & Athanasopoulos, G.** (2018). *Forecasting: Principles and Practice*. Monash University, Australia.
4. **Hochreiter, S., & Schmidhuber, J.** (1997). *Long Short-Term Memory*. Neural Computation.