IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AIR QUALITY FORECASTING AND EARLY WARNING SYSTEM

Ast. Prof. Dr. Santosh Singh[1], Sachin Manoj Chaurasiya[2], Sumit Dinesh Singh[3], 1

Assistant Professor, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India 2, 3 PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

Abstract

Air pollution poses a significant threat to public health and the environment, making accurate air quality forecasting essential. This study presents an Air Quality Forecasting and Early Warning System that leverages machine learning models and time series analysis to predict air pollution levels and issue timely alerts.

The system utilizes real-time and historical data from environmental sensors to forecast key pollutants such as PM2.5, PM10, CO, NO_2, and O_3. Advanced predictive models, including SARIMA, LSTM, and Random Forest, are employed to improve forecasting accuracy. A web-based interface provides visual analytics and automated alerts to authorities and the public, enabling proactive measures to reduce exposure and mitigate pollution effects.

Experimental results demonstrate the system's effectiveness in predicting pollution trends and enhancing early warning mechanisms, contributing to sustainable urban air quality management.

However, challenges persist in improving model accuracy and public compliance. Despite the availability of early warnings, public awareness and adherence to recommendations are often inadequate. Enhancing data integration from a wider range of sources and improving forecast resolution can increase the effectiveness of the system.

Keywords: LSTM (Long Short-Term Memory) Model, SARIMA (Seasonal Autoregressive Integrated Moving Average) Model, Air Quality Prediction, Air Quality Forecasting, Warning System, Air Quality Index(AQI), Mean Absolute Error (MAE), Mean Squared Error (MSE), R-Squared (R^2), Data Extraction, PM2.5: Particulate matter smaller than 2.5 micrometers (in µg/m^3), PM10: Particulate matter smaller than 10 micrometers (in µg/m^3).

Introduction:

Air pollution in urban areas has become a critical public health and environmental concern, particularly in metropolitan cities like Delhi, which consistently records hazardous Air Quality Index (AQI) levels.

Factors such as rapid industrialization, vehicular emissions, construction activities, and seasonal variations contribute to severe air pollution episodes, leading to respiratory diseases, cardiovascular issues, and reduced visibility. Timely and accurate air quality forecasting is essential to mitigate its adverse effects by enabling proactive measures.

This study aims to develop an Air Quality Forecasting and Early Warning System for Delhi using advanced time series forecasting models—Long Short-Term Memory (LSTM) and Seasonal Autoregressive Integrated Moving Average (SARIMA). These models leverage historical air pollution data and real-time sensor readings to predict the concentration of key pollutants, including PM2.5, PM10, NO_2, CO, and O_3.

Models that is used for "AIR QUALITY FORECASTING AND EARLY WARNING SYSTEM" are -

LSTM, a deep learning-based recurrent neural network, is well-suited for capturing complex, long-term dependencies in time-series data.
SARIMA, a statistical model, effectively captures seasonal variations and trends in air pollution levels.

By integrating these models, the proposed system aims to provide highly accurate short-term and long-term air quality predictions. Additionally, an early warning mechanism will generate real-time alerts, allowing government agencies and citizens to take preventive actions such as implementing traffic restrictions, increasing green cover, or using protective measures like masks and air purifiers.

To evaluate model performance, the study computes key error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics help assess prediction accuracy and model reliability in forecasting air quality trends. The system integrates real-time and historical air pollution data from official monitoring stations, ensuring robust predictive capability. Furthermore, an early warning mechanism is developed to issue alerts when pollution levels exceed safe thresholds, helping authorities and citizens take precautionary measures.

This research contributes to the growing need for data-driven air pollution management in Delhi by offering a scientifically robust forecasting framework that can help in policy-making, urban planning, and public awareness campaigns.

Literature Survey

Air quality forecasting and early warning systems (AQFEWS) are designed to predict pollution levels and issue alerts to mitigate the impact of poor air quality on public health and the environment. These systems integrate data from various sources, employ predictive models, and communicate information to stakeholders.

Early systems primarily relied on observational data and simple statistical methods. Over time, advancements in satellite technology, sensor networks, and computational models have significantly enhanced the capabilities of AQFEWS.

• Historical Developments: Initially, air quality monitoring focused on real-time measurements with limited forecasting. Early models used linear regression and basic statistical approaches (e.g., the work of Seinfeld and Pandis, 2006).

• Technological Evolution: The advent of remote sensing technologies and the expansion of ground-based sensor networks have greatly improved data collection and forecasting capabilities (e.g., Nowak et al., 2006).

Effective AQFEWS rely on diverse data sources, including:

• Ground-based Sensors: Provide high-resolution, localized data on various pollutants (e.g., PM2.5, NO2, O3).

• Satellite Observations: Offer broad spatial coverage and valuable data on aerosol optical depth and other atmospheric parameters (e.g., Martin et al., 2002).

• Meteorological Data: Essential for understanding the dispersion and transformation of pollutants (e.g., Baklanov et al., 2014).

Environmental pollution has been one of the most serious scourges because of the great damage it causes to economies and the lives of people. According to the World Health Organization (WHO) report from March 6, 2017, air pollution takes the lives of 1.7 million children under 5 years of age every year (http://www.who.int/mediacentre/news/releases/2017/pollution-child-death/en/).

With industrialization and urbanization, air pollution and hazy weather have increased at a great rate, especially in developing countries (Wu and Zhang, 2017). In recent years, China has experienced severe and persistent air pollution in the winter, which has attracted worldwide attention (Liu et al., 2017). Relevant research about this issue has also flourished (Ma et al., 2017; Zhuo et al., 2017).

The dominant pollution contaminants, such as particulate matter (PM), sulfur dioxide (SO2) and nitrogen oxides (NOx), pose significant risks to environments. These pollution contaminants are barometers of reaction to environmental problems (L. Zhang et al., 2017). In most air quality monitoring systems, carbon monoxide (CO) and ozone (O3) are the two main objective pollution contaminants (Maga et al., 2017). Additionally, other hazardous substances, such as radiation, soiling or specific toxic gases, cause great damage to air quality. In fact, finding the main factor that causes air pollution is the key to solving the problem. Moreover, precise prediction of pollution contaminants plays a vital role. Therefore, it is of great importance to propose an early-warning system and take effective corresponding protection measures.

There are a series of environmental factors that cause air pollution, and the flow and diffusion of pollution contaminants is a rather complex process (Vidale et al., 2017). In early stages, many scholars focused on the relationship between air pollution and its environmental influence factors (Pahlavani et al., 2017). Based on these pioneering works, air quality monitoring systems emerged for monitoring the main pollution contaminants. Recently, studies on the emission sources of pollution contaminants have become a hot topic, and early-warning systems have shown practical importance (Yin et al., 2015; Liora et al., 2016; L. Li et al., 2017). However,

little research has focused on the selection of pollutant indicators and the design of early-warning systems for particular cities.

On the other hand, a variety of models have been proposed to forecast pollution contaminants that can be roughly divided into two categories: statistical models and machine learning models. Traditional statistical approaches, including regression models (Lee et al., 2017), time series models (Nhung et al., 2017) and autoregressive moving average models (ARIMA) (Zafra et al., 2017), are applied in pollution contaminant prediction. Additionally, researchers have investigated the partial and multiple linear features of pollution contaminants and have shown different patterns of statistical models (Carlsen et al., 2018; Zhu et al., 2017). However, each of these approaches has difficulties in dealing with non-linear problems.

Methodology: LSTM (Long Short-Term Memory)

1. Data Preparation

Data Preprocessing: The time series data is cleaned, missing values are handled, and any outliers are treated.

Normalization/Scaling: Time series data is scaled to improve the model's convergence rate. Techniques like Min-Max Scaling or Standardization (Z-score normalization) are commonly used.

Sequence Generation: Since LSTMs require data to be fed as sequences (i.e., a window of previous time steps to predict the next), the data is split into input-output pairs. For example, if you use a window of size n, you would use the first n observations to predict the next value in the sequence.

2. Model Development

Model Architecture: An LSTM network is built using layers such as:

LSTM Layer(s): This is the core part of the network, which captures temporal dependencies.

Dense Layer(s): After the LSTM layer(s), a fully connected (dense) layer may be used to output the predicted value.

Activation Function: Typically, the ReLU or tanh activation function is used in the LSTM layer.

Dropout Layer: Dropout may be applied to avoid overfitting.

Compilation and Optimization: The model is compiled using an optimizer (e.g., Adam, RMSProp) and loss function (e.g., Mean Squared Error).

Training: The model is trained on the training dataset with a suitable batch size and number of epochs. The training process involves adjusting weights to minimize the error between predicted and actual values.

3. Model Evaluation

The model's performance is evaluated using metrics like RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), or MAPE (Mean Absolute Percentage Error).

Cross-validation techniques like Time Series Cross-Validation can be used to avoid overfitting and ensure the model generalizes well.

4. Forecasting

Once trained, the model can be used to predict future values by providing the most recent data point(s) as input and using the model to forecast the next time step(s).

Methodology: SARIMA (Seasonal AutoRegressive Integrated Moving Average)

1. Data Preparation

- Data Preprocessing: Handle missing values, outliers, and ensure that the time series is stationary. If the data isn't stationary, transformations like differencing may be applied.
- Seasonality Identification: SARIMA models are particularly useful for time series that exhibit clear seasonal patterns. Identifying the seasonality period (e.g., daily, weekly, monthly, yearly) is important for building the model.

2. Model Identification

- Model Parameters: The SARIMA model has the following general form: SARIMA(p,d,q)(P,D,Q)s\text{SARIMA}(p,d,q)(P,D,Q)_sSARIMA(p,d,q)(P,D,Q)s

where:

p, d, q are the parameters for the non-seasonal part (AutoRegressive, Integrated, Moving Average).

P, D, Q are the seasonal components of the model (AutoRegressive, Integrated, Moving Average for seasonality).

s is the number of periods in each season (e.g., 12 for monthly data with yearly seasonality).

- Autocorrelation and Partial Autocorrelation: Use ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots to identify possible values for p, d, q (for non-seasonal part), and seasonal parameters.
- Stationarity: If the series is not stationary, apply differencing (both seasonal and non-seasonal) until the series becomes stationary.

3. Model Fitting

- Model Estimation: Fit the SARIMA model to the time series data using the identified parameters. This can be done using statistical software like Statsmodels in Python.
- Optimization: The model's parameters are optimized using methods like Maximum Likelihood Estimation (MLE).

4. Model Evaluation

- Residual Diagnostics: Evaluate the residuals (the difference between actual and predicted values) to ensure that they resemble white noise. This can be done by examining the autocorrelation of the residuals and performing a Ljung-Box test.

- Performance Metrics: Use metrics like AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and RMSE to assess model fit.

5. Forecasting

- The SARIMA model is then used to make forecasts by extrapolating the past data, taking into account both non-seasonal and seasonal components.

3. *Early Warning and Public Dissemination*

• User-friendly Platforms: Create a centralized digital platform (web/mobile app) that disseminates real-time air quality data and forecasts. The platform should be accessible and easily understandable by the general public.

• Air Quality Index (AQI): Use a clear AQI system with color codes and health advisories based on forecasted air quality levels. Provide guidance on protective actions (e.g., avoiding outdoor activities, wearing masks).

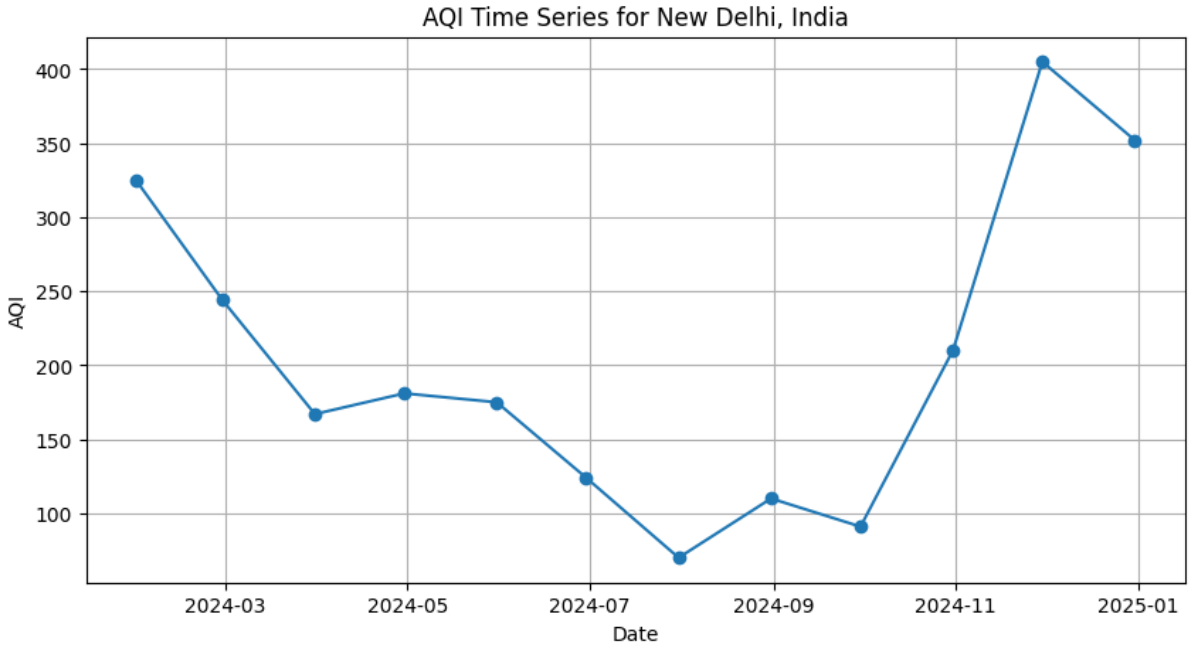
• Push Notifications: Enable early warning systems to send notifications or alerts to the public based on forecasted spikes in pollution. This can be done via SMS, app notifications, or other popular communication channels.

• Integration with Smart Cities: Collaborate with existing smart city initiatives to integrate air quality data with urban planning, traffic control systems, and public health responses.

• Social Media Integration: Regular updates on air quality forecasts and warnings should be shared via social media channels, targeting various demographics.

Result- SARIMA:

1.

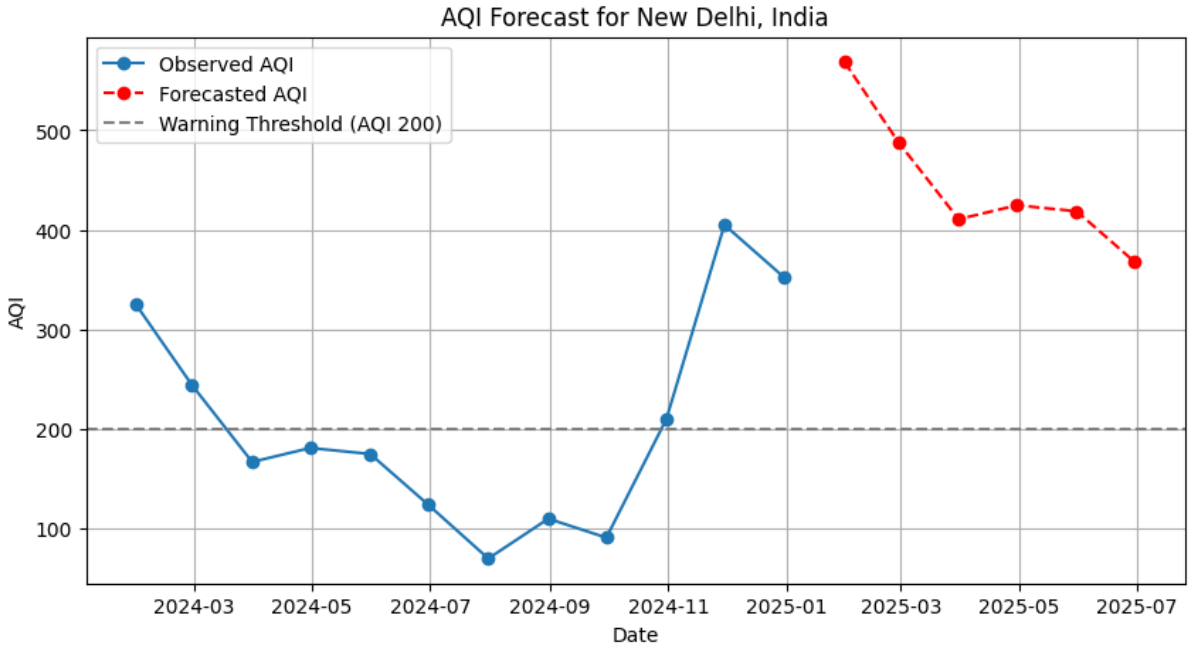


FutureWarning: 'M' is deprecated and will be removed in a future version, please use 'ME' instead.

df_city.index = pd.date_range(start="2024-01", periods=12, freq="M")

Assuming latest year

2.



UserWarning: Too few observations to estimate starting parameters for ARMA and trend. All parameters except for variances will be set to zeros.

warn('Too few observations to estimate starting parameters%s.'

UserWarning: Too few observations to estimate starting parameters for seasonal ARMA. All parameters except for variances will be set to zeros.

warn('Too few observations to estimate starting parameters%s.'

FutureWarning: 'M' is deprecated and will be removed in a future version, please use 'ME' instead.

forecast_index = pd.date_range(start=df_city.index[-1] + pd.DateOffset(months=1),

periods=forecast_steps, freq="M")

3. MAE, MSE, RMSE, R^2

MAE: 297.56673160636103

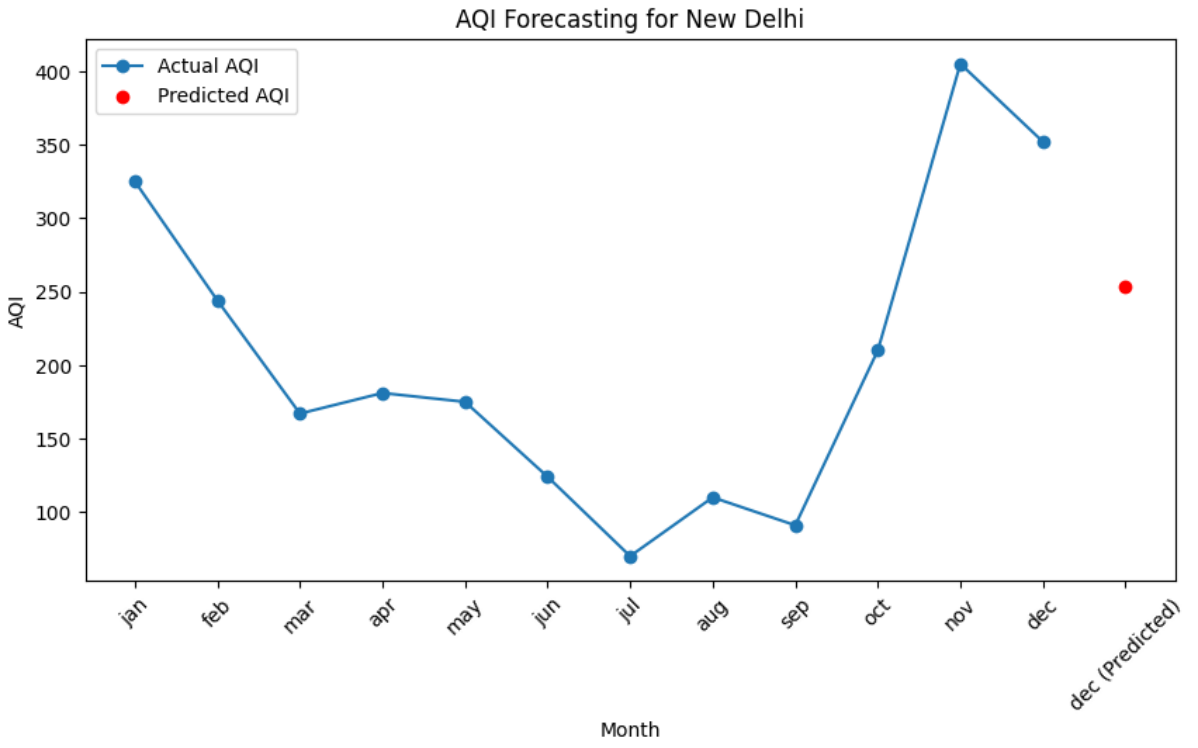
MSE: 95027.91772150889

RMSE: 308.26598534627345

R-squared: -13.514731984970384

LSTM:

1.



Predicted AQI for next month: 253.27

2. MAE, MSE, RMSE, R^2

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3s 3ms/step - loss: 0.1065

Epoch 2/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0862

Epoch 3/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0547

Epoch 4/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0502

Epoch 5/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0367

Epoch 6/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0396

Epoch 7/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0345

Epoch 8/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0300

Epoch 9/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0293

Epoch 10/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0230

Epoch 11/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0210

Epoch 12/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0197

Epoch 13/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0170

Epoch 14/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0159

Epoch 15/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0160

Epoch 16/50

18/18 ——————————————————— 0s 4ms/step - loss: 0.0183

Epoch 17/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0152

Epoch 18/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0128

Epoch 19/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0149

Epoch 20/50

18/18 ——————————————————— 0s 3ms/step - loss: 0.0152

Epoch 21/50

18/18 ——————————————————— 0s 15ms/step - loss: 0.0174

```
Epoch 22/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0135
Epoch 23/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0163
Epoch 24/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0187
Epoch 25/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0153
Epoch 26/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 0.0136
Epoch 27/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0129
Epoch 28/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0133
Epoch 29/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0128
Epoch 30/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0147
Epoch 31/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0162
Epoch 32/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0135
Epoch 33/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0149
Epoch 34/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0145
Epoch 35/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0155
Epoch 36/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0136
Epoch 37/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 0.0137
Epoch 38/50
18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - loss: 0.0172
Epoch 39/50
```

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0141

Epoch 40/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0151

Epoch 41/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0153

Epoch 42/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0167

Epoch 43/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0181

Epoch 44/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0164

Epoch 45/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0147

Epoch 46/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0170

Epoch 47/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0137

Epoch 48/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0129

Epoch 49/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0148

Epoch 50/50

18/18 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0158

3/3 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0s 104ms/step

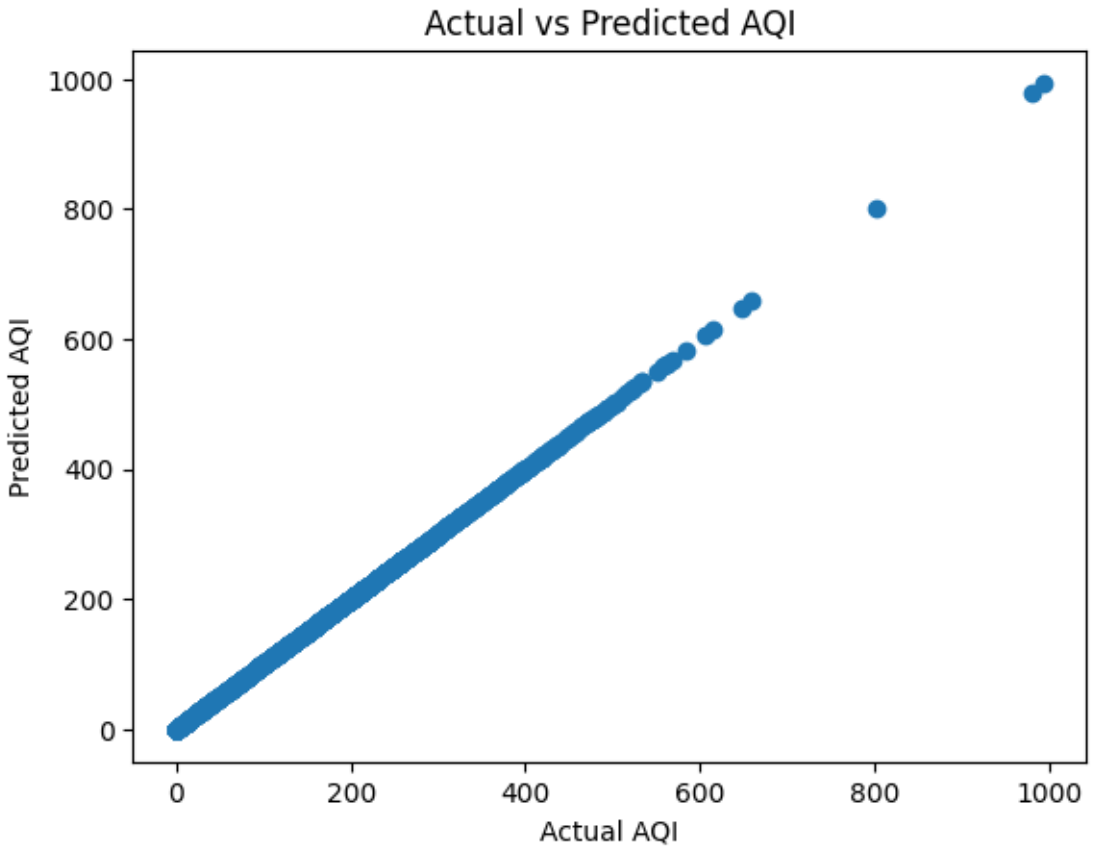
MAE: 0.11427239217966835

MSE: 0.02550878917596472

RMSE: 0.15971471183320815

R^2: 0.2319073793672617

3. *Actual vs Predicted AQI(Air Quality Index)*



Expected Outcomes:

1. When implementing LSTM (Long Short-Term Memory) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models for time series forecasting, each model provides distinct outcomes based on its strengths and limitations.
2. Accurate forecasting of time series data that exhibits complex, non-linear relationships. This could be particularly useful for applications like stock prices, weather forecasting, or any domain where there are intricate relationships between data points.
3. The LSTM model will learn from the long-term dependencies in the data, providing a better forecast over long horizons (e.g., months or years) compared to simpler models.
4. The SARIMA model will produce accurate forecasts for time series that exhibit clear seasonality and trends. For example, retail sales data with monthly seasonality, electricity demand with yearly seasonality, etc.
5. When properly tuned, the SARIMA model should provide reliable forecasts with clear seasonal patterns. However, if the data exhibits non-linear patterns or irregularities, SARIMA may struggle to capture these accurately.
6. The SARIMA model will work well when the time series exhibits relatively simple linear trends, and when the seasonality is strong and consistent over time.

Conclusion:

Therefore, this paper presented a highly efficient machine learning model combining Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machine (SVM) to differentiate between healthy and unhealthy nails. Such an accuracy in the detection of nail diseases is achieved by this model because feature extraction gets done by CNN; and the initial classification gets done by RF, while the results get refined by SVM. This solution shall help scale early disease detection, especially for countries like India that have less access to health care, thus enabling the proper identification and treatment of nail conditions at the proper time.

7. Although this model seems to be promising for encouraging outcomes, further refinements can be made by extending the size of the dataset and adding some sophisticated techniques. Having it implemented into real-life scenarios and integrated with applications on mobile devices will make it more accessible-especially in areas of remote places. This work leads to setting a premise for health solutions reliant on AI technologies that enhance early detection of diseases in resource-constrained settings.

REFERENCES

Baklanov, A., et al. (2014). "Modeling of atmospheric composition: A review of current practices and research directions." Atmospheric Chemistry and Physics.

DonnellyA. et al. "Real time air quality forecasting using integrated parametric and non-parametric regression techniques"

CobournW.G. "An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations"

BaiY. et al. "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions"

Cohan, D. S., et al. (2006). "Improved forecast of PM2.5 using the Community Multiscale Air Quality (CMAQ) model." Journal of Air & Waste Management Association.

Dowell, M. D., et al. (2011). "Real-time air quality forecasting: An overview of current methods and emerging technologies." Environmental Monitoring and Assessment.