**Literature Analysis**

**1) A Review on Prediction of Air Quality Index and Forecasting using Machine Learning Algorithms”** by Vivek Kumar Yadav and Prof. Vishwa Gupta, published in International Journal of Research, Vol. 9 Issue 10, October 2022:

**1. Dataset(s) Used**

This paper is a literature review and compiles findings from various studies. Key datasets referenced include:

| **Study** | **Dataset Details** |
| --- | --- |
| **Liu et al. (2019)** | - Beijing: 1738 instances (Dec 2013–Aug 2018), hourly AQI + PM2.5, O3, SO2, PM10, NO2 - Italian city: 9358 instances (Mar 2004–Feb 2005), hourly CO, NOx, NO2 |
| **Ziyue Guan et al. (2018)** | - PM2.5 data from EPA (Melbourne) + Airbeam mobile sensor |
| **Heidar Maleki et al. (2019)** | - Hourly data from 4 stations in Ahvaz, Iran (Aug 2009–Aug 2010): NO2, SO2, PM10, PM2.5, CO, O3 |
| **Nidhi Sharma et al. (2018)** | - Delhi air quality (2009–2017): SO2, NO2, PM, O3, CO, Benzene |
| **Mohamed Shakir et al. (2018)** | - Karnataka Pollution Control Board data (NO, NO2, CO, PM10, SO2, temp, humidity, wind) |

**2. Features (Input Variables) Used**

Across studies, the following features were used:

* **Air Pollutants:** PM2.5, PM10, NO2, NOx, CO, SO2, O3, Benzene, NH3
* **Meteorological Data:** Temperature, humidity, wind speed
* **Temporal Features:** Date, hour, day of week
* **Health Indicators (in some studies):** Mortality data, health burden due to pollutants

**3. Models Used**

| **Study** | **Model(s) Used** | **Purpose** |
| --- | --- | --- |
| **Anikender Kumar et al. (2011)** | Principal Component Regression (PCR), Multiple Linear Regression (MLR) | Forecast AQI |
| **Huixiang Liu et al. (2019)** | Support Vector Regression (SVR), Random Forest Regression (RFR) | SVR for AQI, RFR for NOx |
| **Ziyue Guan et al. (2018)** | ANN, Linear Regression, LSTM | Predict PM2.5 (LSTM best) |
| **Heidar Maleki et al. (2019)** | Artificial Neural Network (ANN) | Predict pollutant concentrations + AQI |
| **Aditya C R et al. (2018)** | Logistic Regression, Auto Regression | Detect/predict PM2.5 |
| **Mohamed Shakir et al. (2018)** | ZeroR, K-means Clustering (WEKA tool) | Pollution pattern discovery |
| **S. Tikhe Shruti et al. (2013)** | ANN, Genetic Programming (GP) | GP outperformed ANN |
| **Archontoula Chaloulakou et al. (2003)** | ANN, MLR | ANN outperformed MLR for PM10 prediction |

**4. Best Performing Model**

| **Best Model** | **Context** | **Why Best** |
| --- | --- | --- |
| **LSTM** (Ziyue Guan et al.) | PM2.5 Prediction | Handled sequential dependencies and gave accurate results |
| **SVR** (Liu et al.) | AQI Prediction | Outperformed Random Forest in AQI context |
| **GP (Genetic Programming)** | Future pollutant concentration in Pune | Better accuracy over ANN |
| **ANN** (multiple studies) | AQI & pollutant prediction | Strong pattern learning capabilities, especially with rich datasets |

**5. Key Takeaways / Findings**

* **ANN, SVR, LSTM**, and **GP** consistently deliver high accuracy in AQI or pollutant forecasting.
* **Meteorological data** greatly enhances prediction accuracy.
* **LSTM** is especially strong for **time-series** pollutant data.
* Feature selection and input variable engineering (e.g., PCA in PCR) are critical.
* Predictive models can inform **public health** decisions and **policy** actions.

**6. Strengths of the Paper**

* Consolidates findings from **diverse geographies** (India, China, Iran, Italy, Australia).
* Compares performance of models across **multiple pollutants and timescales**.
* Highlights the **importance of combining pollutant and weather data**.

**7. Limitations & Suggestions for Future Work**

* Lacks original model experimentation—review-based only.
* Suggests integrating **all pollutants and meteorological factors** in future hybrid models.
* Advocates for the use of **real-time sensors and mobile data collection**.
* Emphasizes the need for **multi-city, multi-season** models for generalization.

2) **Air Quality Index Prediction in Six Major Chinese Urban Agglomerations: A Comparative Study of Single Machine Learning Model, Ensemble Model, and Hybrid Model”**,

**1. Dataset Used**

* **Sources:**
  + **Air Quality Data:**  
    China National Environmental Monitoring Center (CNEMC)  
    ➤ Hourly data for AQI and pollutant concentrations (PM2.5, PM10, CO, SO₂, NO₂, O₃)
  + **Meteorological Data:**  
    National Climatic Data Center (NCDC), USA  
    ➤ Every 3-hour data (Temperature, Pressure, Dew Point, Wind Speed, Precipitation)
* **Geographical Scope:**  
  Six major Chinese urban agglomerations:
  + BTH-UA (Beijing–Tianjin–Hebei)
  + CP-UA (Central Plains)
  + YRD-UA (Yangtze River Delta)
  + YRMR-UA (Middle Reaches of the Yangtze)
  + CY-UA (Chengdu–Chongqing)
  + PRD-UA (Pearl River Delta)
* **Time Frame:**  
  January 2017 – December 2020
* **Data Size:**
  + ~134,658 data points across 95 cities
  + AQI + 6 pollutants + 5 meteorological features
  + Final format: **daily averages**

**2. Features Used**

* **Pollutants:** PM2.5, PM10, CO, SO₂, NO₂, O₃
* **Meteorological Factors:** Temperature (T), Pressure (P), Dew Point (TD), Wind Speed (WS), Precipitation (Pre)
* **Target Variable:** Air Quality Index (AQI)

**3. Models Used**

| **Type** | **Model** | **Description** |
| --- | --- | --- |
| **Single** | Linear Regression (LR), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Long Short-Term Memory (LSTM) | Baseline and deep models for time-series regression |
| **Ensemble** | Random Forest (RF), XGBoost (XGBT), LightGBM (LGBM) | Ensemble decision tree-based models |
| **Hybrid** | LSTM-SVR | LSTM for initial prediction + SVR for residual correction |

**4. Best Performing Model**

| **Model** | **Why It’s Best** |
| --- | --- |
| **LSTM-SVR (Hybrid)** | Achieved the **highest R² (0.993)** and **lowest RMSE (2.67)** on average, particularly better in **high pollution zones** like BTH-UA and CP-UA |
| **RF (Ensemble)** | Performed nearly as well (R² = 0.989) and slightly better than LSTM-SVR on **MAE (0.64)**, especially in low-pollution areas like PRD-UA |

**Evaluation Metrics Used:**

* **R² (Coefficient of Determination)**
* **RMSE (Root Mean Square Error)**
* **MAE (Mean Absolute Error)**

**5. Key Results (Mean Performance Across All Cities)**

| **Model** | **R²** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| **LSTM-SVR** | **0.993** | **2.67** | 1.29 |
| RF | 0.989 | 3.22 | **0.64** |
| XGBT | 0.988 | 3.56 | 1.42 |
| LGBM | 0.985 | 3.88 | 1.64 |
| LSTM | 0.977 | 4.99 | 3.14 |
| SVR | 0.930 | 8.81 | 4.32 |
| KNN | 0.946 | 7.84 | 5.13 |
| LR | 0.860 | 12.35 | 9.25 |

**6. Important Observations**

* **Higher AQI = Harder to Predict:** CP-UA had the highest AQI and lowest model accuracy.
* **LSTM-SVR excelled in high-pollution cities**, compensating LSTM’s bias using SVR.
* **RF excelled in low-pollution areas**, being faster than hybrid models.
* **Data split method matters:** **Time-series split (36:12)** gave better results than random splits.
* **Standardization (Z-score)** was applied to all inputs due to variance in feature scales.

**7. Strengths of the Study**

* Thorough **comparison of 8 models** across 6 regions using 3 robust metrics.
* Introduced a **hybrid model** that significantly improved predictions in polluted cities.
* Used **real multi-year, multi-source data** with environmental and weather factors.
* Performed **cross-regional analysis**, revealing model stability and adaptability.

**8. Limitations & Future Work**

* Hybrid models like LSTM-SVR are **computationally expensive** (longest training time).
* Performance may vary with **other cities or unseen pollution patterns**.
* Potential future enhancements include:
  + **Incorporating satellite data**
  + **Real-time prediction systems**
  + Using **transfer learning** to adapt models across regions

3) **“A Literature Review on Prediction of Air Quality Index and Forecasting Ambient Air Pollutants using Machine Learning Algorithms”**  
Published in International Journal of Innovative Science and Research Technology (IJISRT), Volume 5, Issue 8, August 2020 by **Radhika M. Patil, Dr. H.T. Dinde, and Sonali K. Powar**.

### ****Dataset(s) Used****

As this is a **literature review**, the paper discusses multiple datasets from different research studies:

| **Study** | **Dataset Summary** |
| --- | --- |
| **Liu et al. (2019)** | - Beijing dataset (Dec 2013–Aug 2018): 1738 records, hourly AQI and PM2.5, PM10, O₃, SO₂, NO₂  - Italian city (Mar 2004–Feb 2005): 9358 records with CO, Benzene, NOx, NO₂ |
| **Ziyue Guan et al. (2018)** | PM2.5 data from EPA (Melbourne) and Airbeam (mobile device) |
| **Maleki et al. (2019)** | 2009–2010 hourly data from 4 Iranian stations (NO₂, SO₂, PM10, PM2.5, CO, O₃) |
| **Nidhi Sharma et al. (2018)** | Delhi data (2009–2017) on SO₂, NO₂, PM, O₃, CO, Benzene |
| **Shakir & Rakesh (2018)** | Karnataka Pollution Control Board (NO, NO₂, CO, PM10, SO₂, Temperature, Humidity, Wind Speed) |
| **Others** | Studies using AirQ software (WHO) for Tehran, Kermanshah, and Salem cities for health impact due to pollutants |

### 2. ****Features Used****

* **Pollutants:** PM2.5, PM10, NOx, NO₂, SO₂, CO, O₃, Benzene, NH₃, Pb, Ni, As, Benzo(a)pyrene
* **Meteorological Variables:** Temperature, Wind Speed, Humidity, Time, Date
* **Target Variable:** AQI (Air Quality Index), pollutant concentration, AQHI (Health Index)

### 3. ****Models Discussed****

| **Model** | **Used By** | **Purpose** |
| --- | --- | --- |
| **Principal Component Regression (PCR)** | Anikender Kumar et al. | Forecast AQI (Delhi, seasonal) |
| **Multiple Linear Regression (MLR)** | Used with PCR | Compared with PCR for AQI |
| **Support Vector Regression (SVR)** | Liu et al. | Predict AQI |
| **Random Forest Regression (RFR)** | Liu et al. | Predict NOx |
| **ANN (Artificial Neural Network)** | Multiple studies | Predict pollutant concentrations and AQI |
| **LSTM (RNN)** | Guan & Sinnot | Time-series PM2.5 prediction |
| **Auto Regression** | Aditya C R et al. | PM2.5 forecast |
| **Logistic Regression** | Aditya C R et al. | Classify polluted vs. non-polluted |
| **Genetic Programming (GP)** | Tikhe Shruti et al. | Compare with ANN for SOx, NOx, RSPM |
| **ZeroR & K-means (WEKA)** | Shakir & Rakesh | Pattern analysis and clustering pollutants with environment |

### 4. ****Best Performing Models****

| **Model** | **Context** | **Performance Summary** |
| --- | --- | --- |
| **LSTM** | PM2.5 prediction (Melbourne) | Outperformed LR and ANN due to time-series ability |
| **SVR** | AQI forecast (Beijing) | More accurate for AQI than RFR |
| **RFR** | NOx prediction (Italian city) | Outperformed SVR for NOx |
| **GP** | Pollution prediction (Pune) | Better than ANN on accuracy |
| **ANN** | PM10 (Athens), AQI (Ahvaz) | Better than MLR, accurate with proper training |

### 5. ****Key Findings****

* **Time-series models like LSTM** are ideal for sequential pollutant data.
* **Hybrid modeling or combining ANN with feature reduction (like PCA)** improves results.
* Meteorological features **significantly impact** prediction accuracy.
* **ANN and regression models** are widely used and perform well with sufficient training.
* GP and other soft computing techniques can **outperform traditional ANN** in some contexts.

### 6. ****Strengths of the Paper****

* Wide range of studies covered (India, Iran, China, Italy, Greece, Australia).
* Detailed breakdown of both AQI calculation and prediction strategies.
* Emphasis on both **pollutant forecasting** and **health impact assessments** (AQHI).

### 7. ****Limitations & Future Work****

* As a literature review, no **original experimentation** or **model benchmarking**.
* Emphasizes future models should include:
  + All pollutant types (not just PM10/PM2.5)
  + Meteorological parameters
  + More robust training-validation cycles
  + Predictive models focused on health outcomes (AQHI)

4) **Air Quality Index Prediction Using Machine Learning”**  
Published in the International Journal of Research Publication and Reviews, Vol. 5, No. 1, January 2024.

### 1. ****Dataset Used****

* **Source:** Kaggle (specific dataset not named)
* **Type:** Tabular CSV format
* **Size:** Not specified
* **Attributes Included:**

| **Attribute** | **Type** | **Description** |
| --- | --- | --- |
| City | Text | Location name |
| Date | Date | Observation date |
| PM2.5, PM10 | Numeric | Particulate matter concentrations |
| NO, NO₂, NOx | Numeric | Nitrogen oxides |
| NH₃, CO, SO₂ | Numeric | Common gaseous pollutants |
| O₃ | Numeric | Ozone |
| Benzene, Toluene, Xylene | Numeric | Industrial volatile pollutants |
| AQI | Numeric | Target variable (continuous) |
| AQI\_Bucket | Text | AQI category label (e.g., Good, Poor, etc.) |

### 2. ****Features Used****

* **Pollutant Features:** PM2.5, PM10, NO, NO₂, NOx, NH₃, CO, SO₂, O₃, Benzene, Toluene, Xylene
* **Categorical Info:** City, Date (not used directly for modeling)
* **Target Variable:** AQI (Numeric), AQI\_Bucket (Categorical)

### 3. ****Models Used****

| **Model** | **Description** |
| --- | --- |
| **Linear Regression** | Predicts AQI using a linear relationship with features |
| **Lasso Regression** | Linear model with L1 regularization for feature selection and reducing overfitting |

### 4. ****Best Performing Model****

| **Metric** | **Linear Regression** | **Lasso Regression** |
| --- | --- | --- |
| **MAE** | 0.42 | 0.42 |
| **RMSE** | 0.57 | 0.61 |
| **R²** | 0.71 | **0.77** |

* **Conclusion:** Lasso Regression was **preferred** due to higher R², indicating better explanation of variance.

### 5. ****Key Findings****

* **Lasso Regression outperforms** linear regression in terms of generalization.
* Effective for predicting **both numeric AQI values** and **AQI category labels**.
* **Early-warning** systems and **public health awareness** can benefit from accurate AQI predictions.
* The **R² value of 0.77** for Lasso suggests a reasonably strong model fit, though there's room for improvement with advanced models.

### 6. ****Strengths****

* Clear **comparative evaluation** of two regression techniques.
* Uses **real-world data** from Kaggle with broad pollutant coverage.
* Includes **AQI categories**, making it suitable for classification-based expansion.
* Clean **methodology workflow:** Data preprocessing → Train/test split → Model training → Evaluation.

### 7. ****Limitations & Future Work****

* **Only basic regression models** used; advanced models (Random Forest, XGBoost, LSTM) not explored.
* **No mention of feature scaling**, outlier handling, or cross-validation.
* No temporal (seasonal/time-series) modeling despite date being available.
* **Future scope:**
  + Integrate **real-time cloud-based updates**
  + Use **more complex ML and deep learning models**
  + Include **health risk indicators or AQHI** (Air Quality Health Index)

5) **“Predicting Air Quality Index using Machine Learning”**  
Published in *International Journal of Research Publication and Reviews*, Vol. 5, No. 11, November 2024 by **Suru Sirisha**.

### 1. ****Dataset Used****

* **Source:** Kaggle
* **Features Included:**
  + Pollutants: CO, NO, NO₂, O₃, SO₂, PM2.5, PM10, NH₃
  + Date (converted into year, month, day)
  + City (one-hot encoded)
  + Target: AQI categories (Good, Moderate, Unhealthy, etc.)

### 2. ****Features Used****

* **Air Pollutants:** CO, NO, NO₂, O₃, SO₂, PM2.5, PM10, NH₃
* **Temporal Features:** Year, month, and day extracted from the date
* **Categorical Feature:** City (one-hot encoded)
* **Target Variable:** AQI Category (classification labels)

### 3. ****Models Used****

| **Model** | **Description** |
| --- | --- |
| **XGBoost Classifier** | Gradient boosting model used as **base learner** |
| **Support Vector Classifier (SVC)** | **Meta-learner** in the stacked model setup |
| **Stacked Model** | Combines predictions of XGBoost and SVC for final classification |

### 4. ****Best Performing Model****

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| **Stacked Model** (XGBoost + SVC) | **99.55%** | 99.34% | 99.19% | **99.27%** |

* Evaluation Metrics: Accuracy, Precision, Recall, F1-Score
* **Confusion matrix** and **category distribution** used for further analysis

### 5. ****Key Findings****

* **Stacked classifier significantly outperformed individual models** with over **99% accuracy**.
* Most misclassifications occurred in mid-range AQI categories.
* **Hazardous** category had the highest count in predictions, indicating severe urban pollution cases.
* Time-series plot showed **seasonal AQI variation**, highlighting importance of temporal features.

### 6. ****Strengths****

* Use of **stacking ensemble model** improves generalization and accuracy.
* **Hyperparameter tuning** with GridSearchCV enhanced XGBoost base performance.
* Addresses **imbalanced classification problem** by combining strong learners.
* Includes **explanation of evaluation metrics** and visualizations (confusion matrix, bar plot).

### 7. ****Limitations & Future Work****

* Only **classification model** used; suggests adding **regression models** to predict exact AQI values.
* **Real-time streaming or temporal modeling** (e.g., LSTM) not explored.
* Future scope:
  + Implement **XGBoostRegressor**, **SVR**, or **deep learning models**
  + Use **real-time cloud-based AQI updates**
  + Consider long-term trends via **time-series forecasting**

6) Predicting AQI using Attention Hybrid Deep Learning & QPSO

**Paper Title:** Predicting Air Quality Index using Attention Hybrid Deep Learning and Quantum-Inspired Particle Swarm Optimization  
**Authors:** Anh Tuan Nguyen et al.  
**Published in:** Journal of Big Data (2024)

### 1. ****Dataset Used****

* **Source:** Seoul Air Data Registry
* **Time Period:** January 1, 2021 – December 31, 2022
* **Location:** Seoul, South Korea (25 monitoring stations)
* **Frequency:** Hourly measurements
* **Pollutants Measured:** PM2.5, PM10, NO₂, SO₂, CO, O₃

### 2. ****Features Used****

* **Primary Pollutants:** PM2.5, PM10, NO₂, SO₂, CO, O₃
* **Temporal Features:** Hour, Date, Station location
* **Target Variable:** AQI (calculated per pollutant and aggregated using standard formula)

### 3. ****Model Used****

**Proposed Hybrid Model:**

**ARIMA + ACNN + QPSO-LSTM + XGBoost**

| **Component** | **Role** |
| --- | --- |
| **ARIMA** | Captures linear trends and seasonality |
| **ACNN (Attention CNN)** | Captures spatial and local temporal features |
| **LSTM (optimized by QPSO)** | Models long-term temporal dependencies |
| **QPSO** | Fine-tunes LSTM hyperparameters (e.g., neurons, batch size) |
| **XGBoost** | Fine-tunes and synthesizes final predictions |

Architecture: Encoder-Decoder with ACNN as encoder and Bi-LSTM as decoder, followed by XGBoost.

### 4. ****Best Model Performance****

| **Model** | **MSE (PM2.5)** | **MAE (PM2.5)** | **R² (PM2.5)** |
| --- | --- | --- | --- |
| **Proposed Hybrid** | **9.02** | **2.17** | **0.9579** |
| Best DL Baseline (Bi-LSTM) | 11.05 | 2.33 | 0.9445 |

✅ **Improvements over Bi-LSTM:**

* 31.13% reduction in **MSE**
* 10% reduction in **MAE**
* 1.64% increase in **R²**

**Performance for Station-Specific Forecasting (Nowon-gu):**

* **MSE:** 6.78
* **MAE:** 1.24
* **R²:** **0.99385** (near-perfect fit)

### 5. ****Key Contributions****

* Integrates statistical and deep learning approaches.
* Optimizes LSTM using **Quantum Particle Swarm Optimization (QPSO)**.
* Uses **Attention-based CNN** to extract local/global dependencies.
* Demonstrates robust forecasting across pollutants and locations.

### 6. ****Strengths****

* Handles **nonlinearity**, **noise**, and **short sequences** effectively.
* Performs better than individual models (ARIMA, LSTM, CNN-LSTM, XGBoost).
* Adaptable and scalable for **real-time deployment** in urban air monitoring systems.
* Considers **spatial station distribution** by selecting max AQI across districts.

### 7. ****Limitations & Future Work****

* Currently limited to Seoul — generalization to other cities is required.
* External factors like **meteorological conditions** not yet integrated.
* Long-sequence forecasting (e.g., monthly or seasonal AQI trends) to be explored.
* Further comparisons with **transformer-based models** recommended.

7) *Air Quality Prediction Using Deep Learning – A Review*  
**Authors:** Ayush Kumar Gupta et al.

### 1. ****Dataset(s) Mentioned****

Various datasets were discussed from:

* **Beijing UCI dataset**
* **Kaggle datasets**
* **Microsoft Research Urban Air**
* **Remote sensing & meteorological datasets**
* **Official CPCB datasets**

### 2. ****Features Used in Literature****

* **Air pollutants:** PM2.5, PM10, NO₂, SO₂, CO, O₃, Benzene, Xylene
* **Meteorological variables:** Temperature, humidity, wind speed, pressure
* **Temporal features:** Date, time, station ID

### 3. ****Models Reviewed****

| **Model Type** | **Techniques** |
| --- | --- |
| **Machine Learning** | Random Forest, XGBoost, AdaBoost, SVR, KNN, LASSO, Ridge, Decision Trees |
| **Deep Learning** | CNN, LSTM, Bi-LSTM, GRU, ConvLSTM, Autoencoders |
| **Hybrid/Optimized** | CNN-LSTM, PSO-BP, GA-LSTM, DAQFF, IMDA-VAE |

### 4. ****Best Performing Techniques (Across Reviewed Studies)****

| **Model** | **Notable Findings** |
| --- | --- |
| **ConvLSTM** | Strong spatiotemporal feature learning |
| **CNN-LSTM** | Outperformed standalone LSTM in PM2.5 prediction |
| **DAQFF** | Better than traditional shallow & deep models |
| **IMDA-VAE** | Most accurate for multiple pollutant forecasts |
| **GRNN** | Better than SVR in sensor-based environments |
| **PSO-BP** | Enhanced learning efficiency and accuracy of BP NN |

### 5. ****Key Observations****

* DL models handle **temporal and spatial complexity** better than traditional ML models.
* **Hybrid and ensemble** methods yield superior performance.
* Optimization techniques (GA, PSO, etc.) improve DL training outcomes.
* Challenges include **computational cost**, **data imbalance**, and **missing values**.

**8) Optimized Machine Learning Model for Air Quality Index Prediction in Major Cities in India**

### 1. ****Dataset Used****

* **Source:** Kaggle AQI dataset (2015–2020)
* **Cities:** Delhi, Bangalore, Chennai, Hyderabad, Visakhapatnam, Kolkata
* **Features:** PM2.5, PM10, NO, NO₂, NOx, NH₃, CO, SO₂, O₃, Benzene, Toluene, AQI, AQI\_Bucket

### 2. ****Features Used****

* **Pollutants:** PM2.5, PM10, NOx, SO₂, CO, NH₃, Benzene, etc.
* **AQI Buckets:** Good, Satisfactory, Moderate, Poor, Very Poor, Severe
* **SMOTE** used to balance categorical AQI classes.

### 3. ****Model Proposed****

**Grey Wolf Optimization + Decision Tree Regression**

| **Component** | **Role** |
| --- | --- |
| **GWO** | Feature selection & dimensionality reduction |
| **Decision Tree Regression** | Final AQI prediction |

### 4. ****Performance Results (Balanced Dataset)****

| **City** | **R²** | **MAE** | **MSE** | **Accuracy** |
| --- | --- | --- | --- | --- |
| New Delhi | 0.9485 | 0.1102 | 0.0722 | 88.98% |
| Bangalore | 0.7466 | 0.0851 | 0.2865 | 91.49% |
| Chennai | 0.886 | 0.0478 | 0.1224 | 95.22% |
| Hyderabad | 0.898 | 0.0234 | 0.1322 | 97.66% |
| Visakhapatnam | 0.9024 | 0.0232 | 0.1421 | 97.68% |
| Kolkata | **0.9874** | 0.0552 | **0.0178** | 94.48% |

### 5. ****Key Observations****

* **Kolkata** had the best overall performance across all metrics.
* **SMOTE** was crucial in improving accuracy by balancing underrepresented AQI classes.
* The hybrid **GWO + DT** model outperformed Random Forest, SVR, and KNN baselines.