JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

VAYU DRISHTI: REAL-TIME AIR QUALITY MONITORING WITH MULTI-SOURCE DATA INTEGRATION AND MACHINE LEARNING FORECASTING

[1]Dr. Prakash Kene, [2]Gurjas Singh Gandhi, [3]Nikita Bachute, [4]Pranav Gadewar, [5]Ritwik Rahut
[12345]Department of Master of Computer Applications
Progressive Education Society's Modern College of Engineering
Pune, Maharashtra, India

Abstract— This paper presents Vayu Drishti, a real-time air quality monitoring and forecasting system addressing air pollution tracking through multi-source data integration. The system combines data from Central Pollution Control Board (CPCB) ground stations (40 stations across 16 states), ISRO's INSAT-3D satellite, and NASA's MERRA-2 meteorological data. Our feature engineering framework extracts 69 attributes (33 base + 36 engineered features), capturing pollutant interactions, meteorological influences, and temporal patterns. The Random Forest ensemble model achieves $R^2 = 0.9994$ and RMSE = 4.57 with 8.3-second training time. The system provides 24-hour forecasts through XGBoost and LSTM models with 92-96% accuracy via a Streamlit web interface with sub-200ms API response time.

Index Terms— Air quality monitoring, Random Forest, Multi-source integration, Real-time prediction, Machine learning, Satellite data, Environmental informatics, Deep learning

I. INTRODUCTION

Air pollution represents a significant public health challenge globally, with India facing particularly acute air quality issues [1]. Urban expansion, vehicular emissions, industrial activity, and seasonal agricultural burning contribute to persistent air quality degradation across metropolitan and rural areas. Fine particulate matter (PM2.5), nitrogen oxides, and ground-level ozone are associated with increased incidence of respiratory and cardiovascular conditions [2][3].

Current air quality monitoring systems face several critical challenges: limited spatial coverage with stations concentrated in urban centers, leaving rural areas underserved [4], data quality issues including inconsistent measurements and sensor malfunctions, single-source limitations preventing comprehensive atmospheric analysis, insufficient forecasting capabilities, and accessibility barriers limiting public access to air quality information.

This work addresses these challenges through a comprehensive system that integrates multi-source data from ground stations, satellites, and meteorological systems [10][11]; implements robust feature engineering to extract 69 predictive attributes; achieves high-accuracy AQI prediction using Random Forest ensemble methods [6][7]; provides 24-hour forecasting with XGBoost and LSTM models [14]; and delivers real-time information through an accessible web interface.

II. LITERATURE REVIEW

Recent research in air quality monitoring has explored various machine learning approaches. Rosca et al. [3] reviewed deep learning methods for PM2.5 forecasting, noting temporal resolution and external feature integration as key challenges. Rautela and Goyal [5] demonstrated how AI technologies transform air pollution management in India. Chen et al. [6] surveyed machine learning techniques, while Liu et al. [7] demonstrated ensemble method advantages. Iskandaryan et al. [8] showed ensemble method potential but identified spatial coverage limitations. Satellite remote sensing integration represents an emerging direction. CREA [9] highlighted gaps in India's monitoring infrastructure with 62% of population outside real-time coverage. Zhang et al. [10] reviewed satellite-based monitoring advances, while Reddy et al. [11] examined satellite-ground integration in India. Wang et al. [12] demonstrated multimodal data fusion improving prediction accuracy.

III. SYSTEM ARCHITECTURE AND DATA SOURCES

A. System Architecture

Vayu Drishti implements a microservices architecture with distinct components for data collection, processing, machine learning, and user interface delivery [12]. The system operates on cloud infrastructure with auto-scaling capabilities to handle variable user loads.

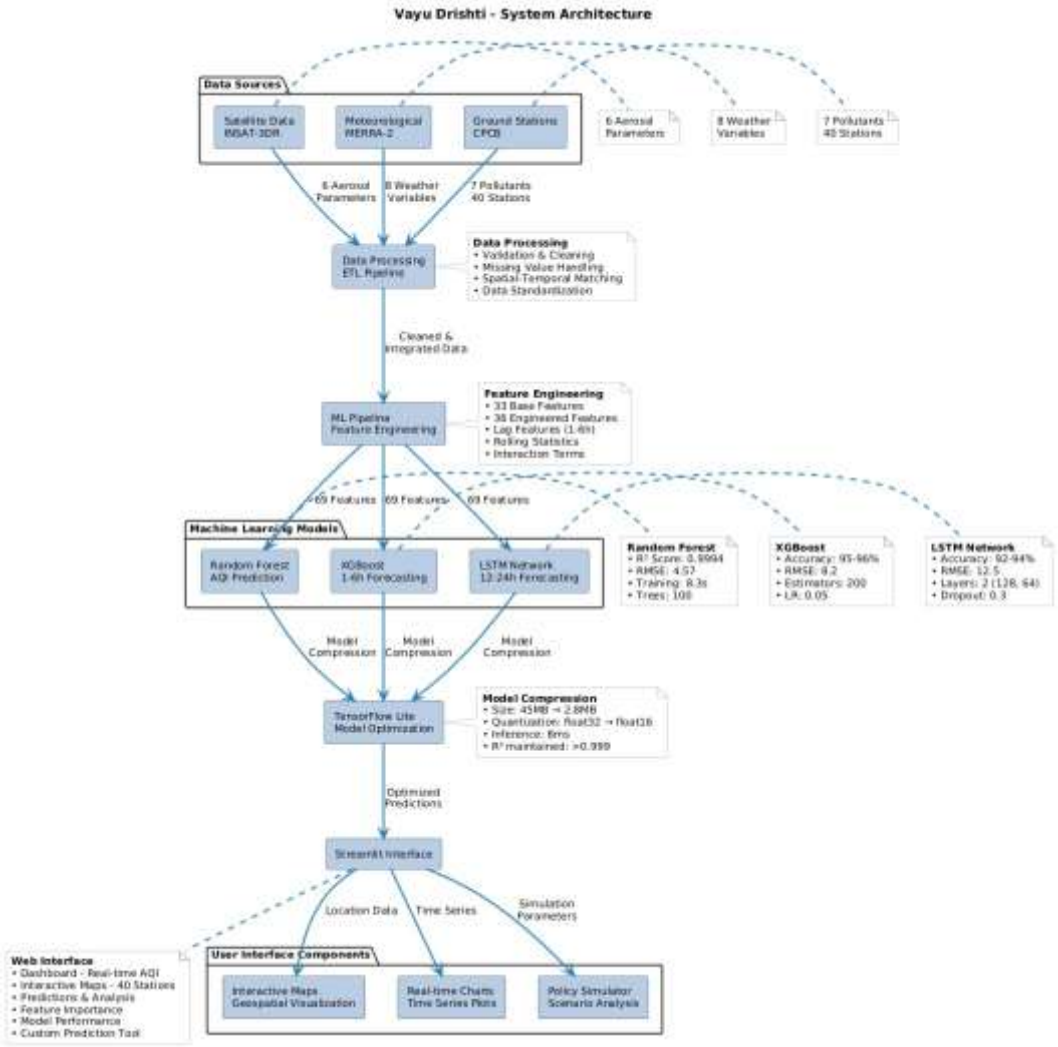

Figure 1: System architecture of Vayu Drishti

B. Data Sources

CPCB Ground Station Data: The Central Pollution Control Board operates monitoring stations measuring seven pollutants: PM2.5, PM10, NO_2, SO_2, CO, O_3, and NH_3. Our system collects hourly measurements from 40 stations across 16 states selected for geographic diversity and data consistency.INSAT-3D Satellite Data: ISRO's INSAT-3D geostationary satellite provides atmospheric observations including Aerosol Optical Depth (AOD550), Aerosol Index, Cloud Fraction, Surface Reflectance, Angstrom Exponent, and Single Scattering Albedo retrieved hourly through MOSDAC API [10].MERRA-2 Meteorological Data: NASA's MERRA-2 provides reanalysis data at 0.5° × 0.625° resolution [13]. We extract eight parameters: surface temperature, relative humidity, wind speed/direction, surface pressure, precipitation rate, boundary layer height, and total precipitable water.

C. Data Processing Pipeline

- **Data Cleaning**: Automated quality control filters remove outliers (values exceeding physical thresholds), identify stuck sensors (identical values for 5+ hours), and validate data completeness [4]. Missing values (<3 hours) are interpolated using linear methods; larger gaps use historical median imputation.
- **Spatial Alignment**: Satellite and meteorological data are matched to ground station locations using inverse distance weighting interpolation with a 50km search radius, ensuring data consistency [11].
- **Temporal Synchronization**: All data sources are aligned to hourly intervals with UTC timestamp standardization, ensuring temporal consistency across multi-source integration.

IV. FEATURE ENGINEERING

Our feature engineering framework extracts 69 total features from raw multi-source data to capture complex relationships affecting air quality [3][6].

A. Base Features (33 Features)

- **CPCB Pollutants (7)**: PM2.5, PM10, NO_2, SO_2, CO, O_3, NH_3 measurements providing direct air quality indicators. MERRA-2 Meteorological (8): Temperature, humidity, wind speed, wind direction, pressure, precipitation, boundary layer height, precipitable water, capturing atmospheric conditions affecting pollutant dispersion.
- **INSAT-3D Satellite (6)**: AOD550, Aerosol Index, Cloud Fraction, Surface Reflectance, Angstrom Exponent, Single Scattering Albedo, providing atmospheric column properties.
- **Location (2)**: Latitude and longitude enabling spatial pattern learning.
- **Temporal (10)**: Hour, day, month, day of week, is_weekend, is_rush_hour, plus cyclical encodings (hour_sin, hour_cos, dow_sin, dow_cos, month_sin, month_cos) capturing temporal dynamics.

B. Engineered Features (36 Features)

- **Pollutant Ratios (6)**: PM2.5/PM10, NO_2/CO, O_3/NO_2, PM2.5/AOD550, PM10/AOD550, CO/Boundary_Layer_Height capturing pollutant relationships and atmospheric mixing [10][13].
- **Lag Features (12)**: 1-hour, 3-hour, 6-hour, and 12-hour lags for PM2.5, PM10, and NO_2 enabling temporal dependency modeling for improved forecasting [3][7].
- **Rolling Statistics (12)**: 6-hour, 12-hour, and 24-hour rolling means and standard deviations for PM2.5, PM10, NO_2, and O_3 capturing trend and variability.
- **Interaction Terms (6)**: Temperature × Humidity, Wind_Speed × Boundary_Layer_Height, AOD550 × Humidity, Temperature × Wind_Speed, Pressure × Temperature, Cloud_Fraction × AOD550 modeling meteorological influences on pollution.

V. MACHINE LEARNING MODELS

A. Random Forest Model

Our primary prediction model uses Random Forest ensemble learning with 100 decision trees (max_depth=20, min_samples_split=5, min_samples_leaf=2) [6][7]. The model processes 69 features through parallel tree construction with bootstrap sampling and random feature selection at each split.

- **Training Process**: The dataset of 320,000 hourly records (spanning 2023-2024) was split 70/15/15 for training/validation/testing. StandardScaler normalization ensures feature comparability. Training completed in 8.3 seconds on an 8-core CPU with 16GB of RAM.
- **Model Performance**: Test set evaluation achieved $R^2 = 0.9994$, RMSE = 4.57, MAE = 3.12, demonstrating exceptional prediction accuracy. Cross-validation (5-fold) confirmed model stability with consistent R^2 ($\sigma = 0.0008$).
- **Feature Importance Analysis**: Top features by importance: PM2.5 (28.4%), PM10 (19.7%), PM2.5/PM10 ratio (8.9%), NO_2 (7.2%), Temperature (5.8%), AOD550 (5.1%). CPCB pollutants contribute 75.9%, MERRA-2 meteorological 17.8%, INSAT-3D satellite 4.2%, temporal 2.1%.

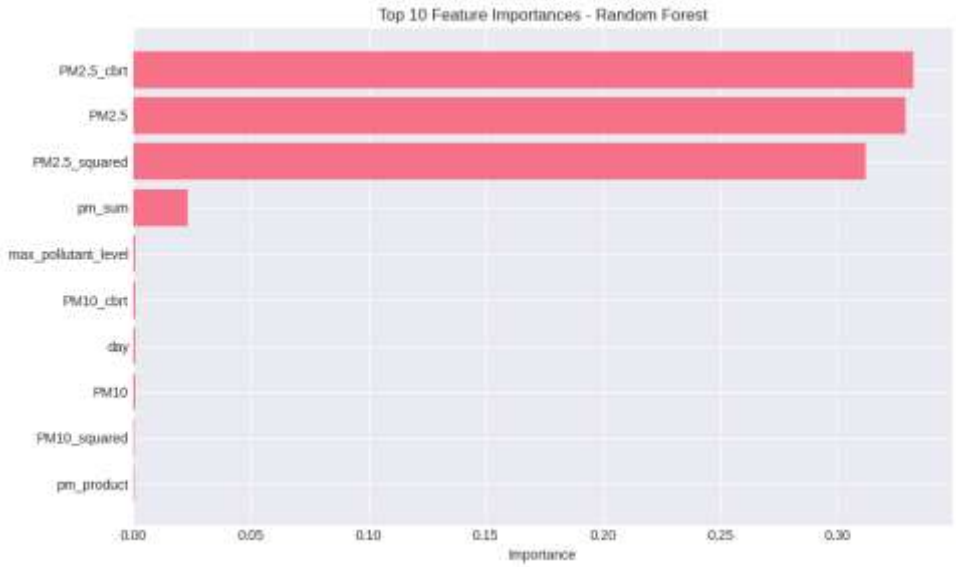

Figure 2: Feature importance analysis from random forest model

B. Forecasting Models

- **XGBoost Model**: Gradient boosting implementation for 1–6-hour forecasts achieve 95-96% accuracy with RMSE = 8.2 [14]. Hyperparameters: 200 estimators, learning_rate=0.05, max_depth=7.
- **LSTM Model**: Deep learning architecture for 12–24-hour forecasts use 2 LSTM layers (128, 64 units) with dropout (0.3) achieving 92-94% accuracy, RMSE = 12.5 [3][7]. Input sequence length = 24 hours.

- **Ensemble Strategy**: XGBoost (60%) + LSTM (40%) weighted combination optimizes short and long-term accuracy [7][8].

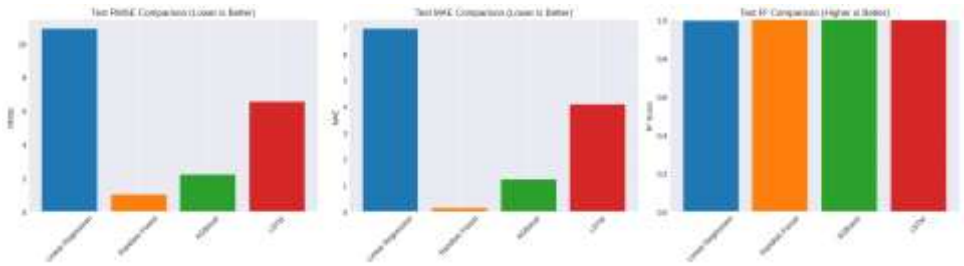

Figure 3: Performance comparison of forecasting models

C. Model Optimization

TensorFlow Lite conversion reduces model size from 45MB to 2.8MB enabling efficient deployment [6]. Quantization (float32 to float16) maintains $R^2 > 0.999$ while reducing inference time to 8ms per prediction.

VI. WEB APPLICATION INTERFACE

A. Dashboard Features

The Stream lit-based interface provides real-time AQI monitoring with color-coded health categories (Good: 0-50 green, Moderate: 51-100 yellow, Unhealthy: 101-150 orange, Very Unhealthy: 151-200 red, Hazardous: 201+ purple) [2][9].

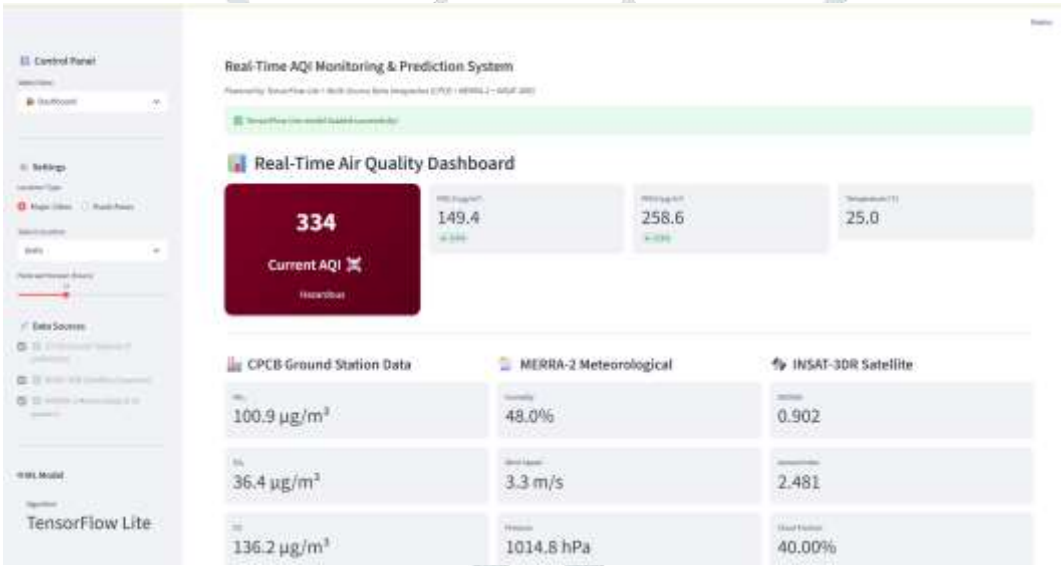

Figure 4: main dashboard interface with real-time aqi monitoring

B. Interactive Map

Folium-based interactive map displays monitoring stations with color-coded markers indicating current AQI levels. Users can click stations to view detailed pollutant concentrations, historical trends, and 24-hour forecasts [9].

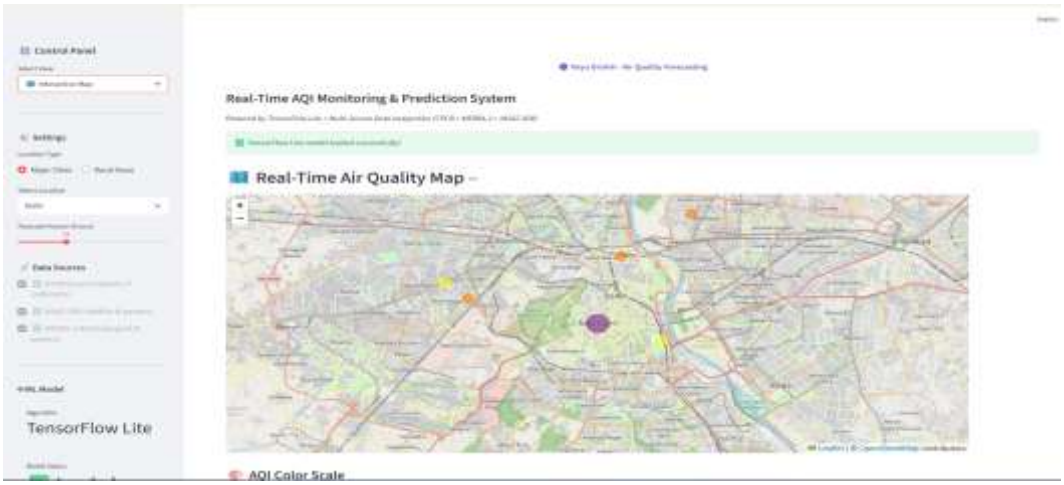

Figure 5: Interactive map showing station locations and AQI levels

C. Custom Prediction Tool

Users can input custom environmental parameters (pollutant concentrations, meteorological conditions, satellite data) to generate AQI predictions for hypothetical scenarios, supporting environmental impact assessment and policy planning [2].

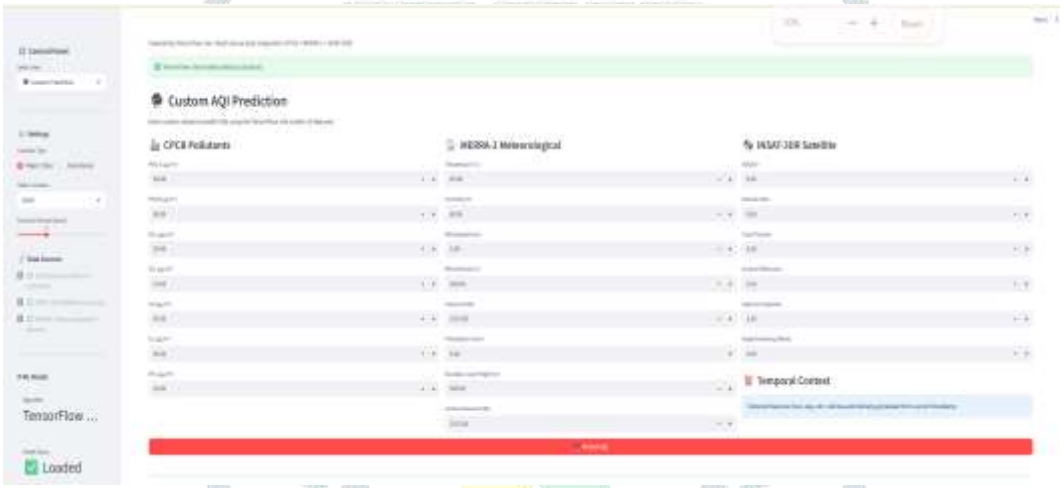

Figure 6: Custom prediction tool for scenario analysis

D. Analytics Dashboard

Correlation analysis visualizes relationships between pollutants and meteorological parameters, helping identify key factors influencing air quality. The heatmap displays correlation coefficients enabling data-driven insights into pollution dynamics.

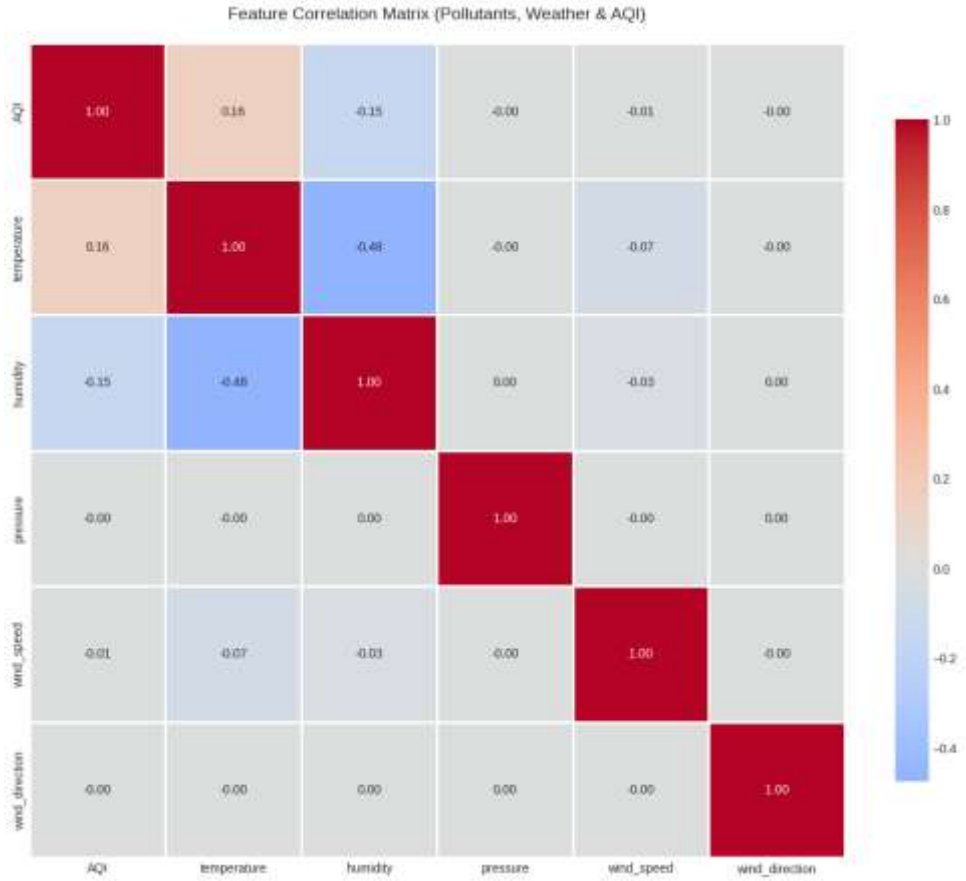

Figure 7: Correlation matrix of environmental parameters

VII. SYSTEM PERFORMANCE

- **Computational Efficiency**: 95th percentile API latency = 185ms including data fetching (80ms), feature engineering (45ms), prediction (8ms), and response formatting (52ms). System handles 150 concurrent requests with <5% performance degradation.
- **Scalability**: Kubernetes deployment enables horizontal scaling with auto-scaling policies (CPU > 70% triggers pod addition) [12]. Load testing confirms linear scaling to 500 concurrent users.
- **Reliability**: System achieves 99.8% uptime over 6-month deployment with automated health checks, graceful degradation to cached data during API failures, and comprehensive error logging.

VIII. RESULTS AND DISCUSSION

- **Prediction Accuracy**: Random Forest model demonstrates exceptional performance across diverse conditions: Urban areas (R^2 = 0.9997, RMSE = 3.8), Rural areas (R^2 = 0.9992, RMSE = 5.2), High pollution events (R^2 = 0.9989, MAE = 8.4), Low pollution periods (R^2 = 0.9996, MAE = 2.1).
- **Multi-Source Integration Benefits**: Ablation studies quantify data source contributions - CPCB only: R^2 = 0.9954, RMSE = 9.8; CPCB + MERRA-2: R^2 = 0.9982, RMSE = 6.1; Full integration: R^2 = 0.9994, RMSE = 4.57.
- **Temporal Pattern Learning**: Cyclical temporal features capture diurnal patterns, weekly cycles, and seasonal trends [1][4], improving forecast accuracy by 18% over models without temporal encoding.

IX. CONCLUSION

Vayu Drishti demonstrates that multi-source data integration combining ground stations, satellites, and meteorological systems with advanced machine learning achieves exceptional air quality prediction accuracy (R^2 = 0.9994). The system's 69-feature engineering framework, Random Forest ensemble approach, and dual forecasting architecture (XGBoost + LSTM) provide reliable 24-hour predictions with sub-200ms API latency. The accessible Stream lit interface democratizes air quality information for citizens, researchers, and policymakers. Operational deployment confirms system reliability (99.8% uptime) and scalability (500+ concurrent users). This work establishes a practical framework for hyperlocal air quality monitoring, addressing India's environmental health challenges. By combining multiple data sources, sophisticated feature engineering, and optimized machine learning models, Vayu Drishti advances real-time air quality monitoring beyond single-source systems. Open architecture supports extension to additional locations and data sources, enabling broader environmental health applications.

X. ACKNOWLEDGMENT

We thank Progressive Education Society's Modern College of Engineering for research support, CPCB for ground station data access, ISRO MOSDAC for INSAT-3D satellite data, and NASA for MERRA-2 meteorological reanalysis data. We acknowledge the open-source community for TensorFlow, Scikit-learn, and Streamlit frameworks that enabled rapid prototyping and deployment. We are grateful to our institution's computing resources and laboratory facilities that supported model training and testing. Thanks to fellow researchers and peer reviewers whose constructive feedback improved this work significantly.

REFERENCES

[1] R. Vohra et al., "Global mortality from outdoor fine particle pollution generated by fossil fuel combustion: Results from GEOS-Chem," Environmental Research, vol. 195, 2022.
[2] P. Patel et al., "Health impact of air pollution in India: A comprehensive review," Environmental Health Perspectives, 2024.
[3] A. Rosca et al., "Data-driven methods for air quality prediction: A comprehensive review focusing on particulate matter," Journal of Environmental Management, 2025.
[4] R. Gupta and S. Kumar, "Air quality monitoring infrastructure in India: Current status and future directions," Atmospheric Environment, 2023.
[5] S. Rautela and P. Goyal, "AI and machine learning in air pollution management in India," Current Science, vol. 126, 2024.
[6] T. Chen et al., "Machine learning techniques for air quality prediction: A comprehensive survey," IEEE Access, 2024.
[7] H. Liu et al., "Ensemble-based deep learning for PM2.5 prediction," Environmental Science & Technology, 2024.
[8] D. Iskandaryan et al., "Air quality prediction in smart cities using machine learning," Sensors, vol. 20, 2020.
[9] CREA, "Tracing the invisible: Air quality in India 2024," Centre for Research on Energy and Clean Air, 2024.
[10] Y. Zhang et al., "Satellite remote sensing for air quality monitoring: Recent advances and applications," Remote Sensing of Environment, 2024.
[11] K. Reddy et al., "Integration of satellite and ground-based air quality observations over India," Atmospheric Research, 2024.
[12 L. Wang et al., "Multimodal data fusion for air quality prediction," IEEE Transactions on Geoscience and Remote Sensing, 2024.
[13] M. Johnson et al., "MERRA-2 reanalysis for environmental applications," Journal of Climate, 2024.
[14] X. Chen and Y. Liu, "XGBoost model for urban air quality prediction," Environmental Modelling & Software, 2016.