NAME: ZAINAB FATIMA

INTERN: DATA SCIENCE

COHORT: 5

PROJECT: AQI-PREDICTOR FOR NEXT 3 DAYS

SUBMITTED ON: 18. AUG. 2025

Technical Report

Overview

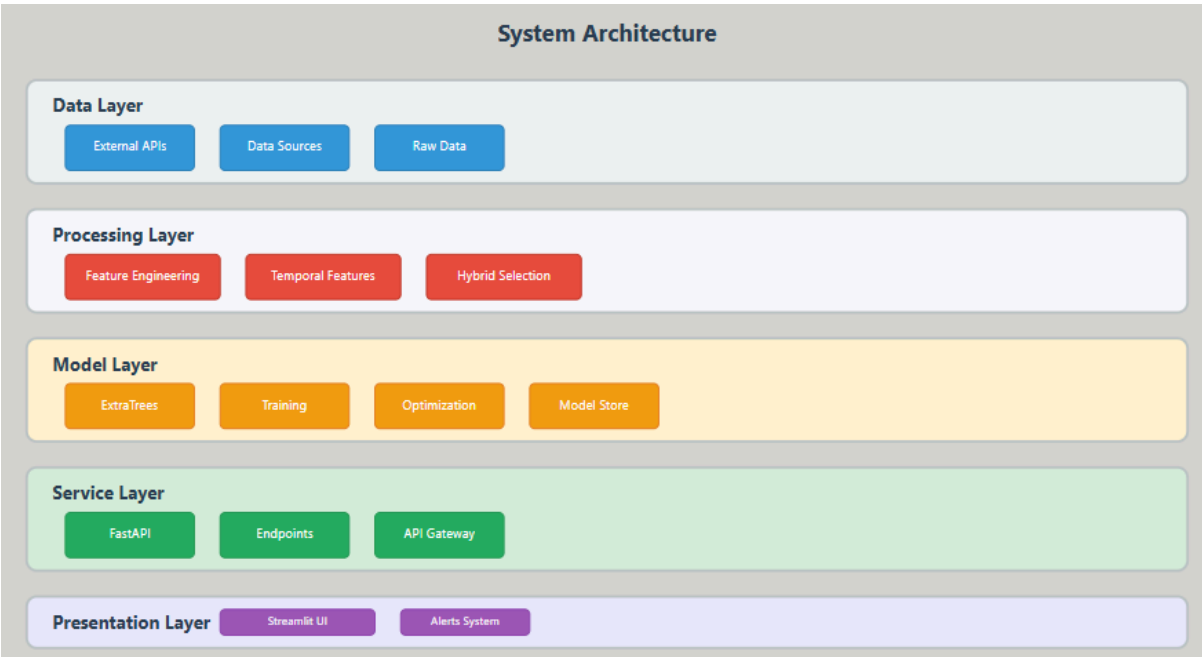
AirLens is a comprehensive air quality monitoring and prediction platform delivering 72-hour AQI forecasts for Karachi. The system integrates FastAPI backend, Streamlit frontend, and advanced ML models achieving R^2 scores of 0.5902 (24h), 0.616 (48h), and 0.217 (72h) using ExtraTrees algorithm.

System Architecture

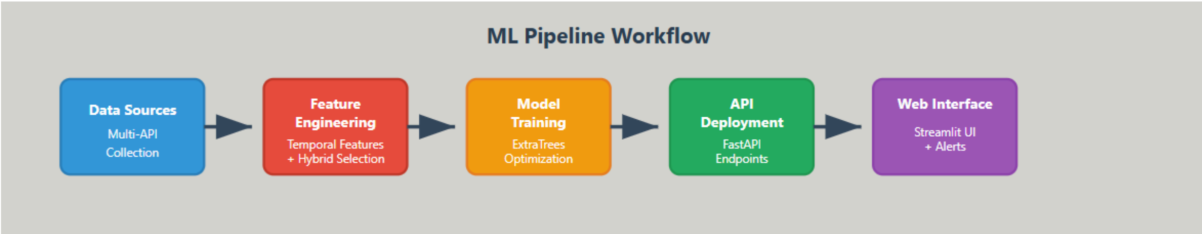
Core Components

- **Backend:** FastAPI with ML model integration and automated CI/CD
- **Frontend:** Streamlit with real-time alerts and responsive design
- **Data Pipeline:** Multi-source API integration with Hopsworks feature store
- **Deployment:** Docker containerization with GitHub Actions automation

SYSTEM ARCHITECTURE:



ML PIPELINE WORKFLOW:



Data Collection Pipeline

Source Integration Strategy

Primary Pipeline (Training): OpenMeteo API - 92 days historical data for ML training
Secondary Pipeline (Real-time): IQAir → AQICN → OpenWeather (priority fallback)

Target Cities & Coverage

- **Karachi:** 24.8607°N, 67.0011°E - Industrial/coastal pollution
- **Lahore:** 31.5204°N, 74.3587°E - Seasonal smog challenges
- **Islamabad:** 33.6844°N, 73.0479°E - Regional transport effects
- **Rate Limiting:** 1-second delays, exponential backoff, request tracking for API compliance

AQI Standardization

Implements EPA breakpoint formula to calculate AQI form collected pollutants:

```
AQI = ((IHi - ILo) / (BPHi - BPLo)) × (Cp - BPLo) + ILo
```

Where:

- **AQI** = Air Quality Index
- **Cp** = Pollutant concentration
- **BPHi** = Breakpoint concentration ≥ Cp
- **BPLo** = Breakpoint concentration ≤ Cp
- **IHi** = AQI value corresponding to BPHi
- **ILo** = AQI value corresponding to BPLo

Data Processing & Feature Engineering

Processing Pipeline

1. **Data Quality:** Duplicate removal, outlier capping (5th-95th percentiles), temporal validation
2. **Missing Values:** Forward/backward fill with 2-value limit, linear interpolation

3. **Temporal Features:** Hour/day/month cycles with sine/cosine transformations
4. **Lag Features:** Strategic intervals (72h, 84h, 96h, 120h, 144h, 168h) with horizon-based minimum lags
5. **Rolling Statistics:** 24/48/72-hour windows for means, std dev, maxima on lagged data
6. **Interaction Features:** Temperature-humidity interactions, wind decomposition, PM2.5/PM10 ratios
7. **Multicollinearity Reduction:** Correlation threshold 0.85 using triangle method

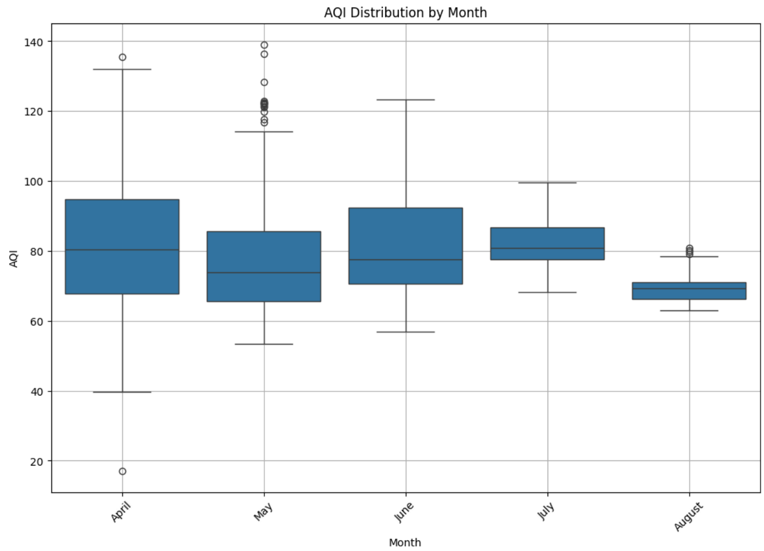
Advanced Features

- **Statistical Trends:** Rate-of-change features across multiple horizons
- **Pollutant-Specific Windows:** Adapted rolling periods for different atmospheric lifetimes

Exploratory Data Analysis:

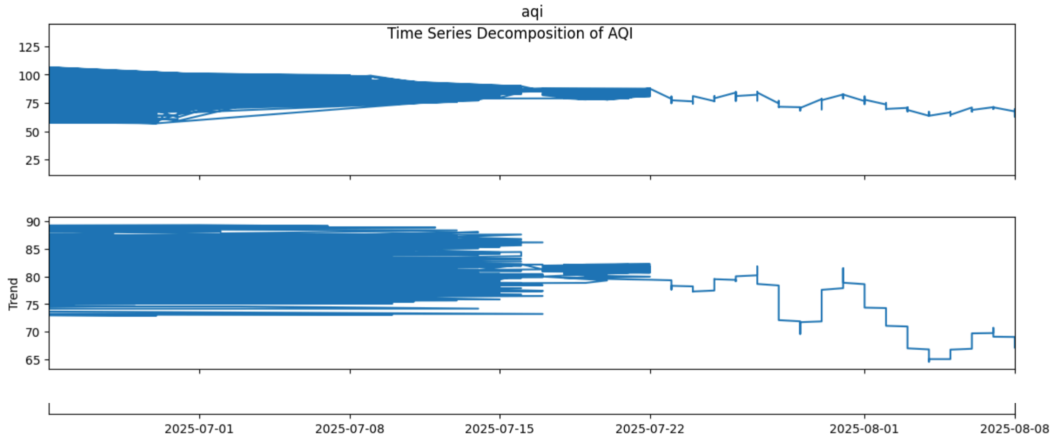
AQI Distribution by Month

This box plot visualizes the distribution of AQI across different months (April to August) in the original existing_df. It shows variations in the median AQI and the spread of values across these months, suggesting a seasonal pattern.



Time Series Decomposition of AQI

This plot decomposes the AQI time series into its trend, seasonal, and residual components. The trend component shows the overall long-term movement of AQI, while the seasonal component highlights repeating patterns within a fixed period (24 hours in this case).



Model Development Approaches

Approach 1: Direct Forecasting with All Features

Method: Separate models per horizon using hundreds of variables

Result: Good linear regression baseline but overfitting with tree-based models

Issue: High-dimensional feature space caused instability and poor generalization

Approach 2: Single-Method Feature Selection

Method: Mutual Information only, then Random Forest importance only

Result: Improved but suboptimal performance, nearly identical results between methods

Issue: MI selected redundant features; RF missed subtle non-linear relationships

Approach 3: Manual Feature Selection

Method: Domain knowledge-based selection with volatility features

Result: Stable R^2 0.39-0.45, RMSE ~14, good interpretability

Issue: Not improved generalisation and poor RMSE

Approach 4: Delta Prediction with Cascading

Method: Predict AQI changes rather than absolute values

Result: Error propagation issues when converting back to absolute AQI

Issue: Increased dimensionality without performance improvements

Final Approach: Enhanced Direct Forecasting with Sliding Window

Method: Direct AQI prediction with forecasted AQI values as features + hybrid MI-RF selection

Breakthrough: Incorporating T, T+48, T+72 AQI values as features transformed performance

Result: Cross-validated test R^2 improved from 0.40 to 0.6, training R^2 0.80-0.85

Validation: Sliding window approach with 8-week training, 4-week test windows

Feature Selection: Hybrid MI-RF Methodology

Innovation

Combines Mutual Information (captures non-linear relationships) with Random Forest importance (weighted impurity reduction) through cross-validated aggregation across TimeSeriesSplit folds.

Process:

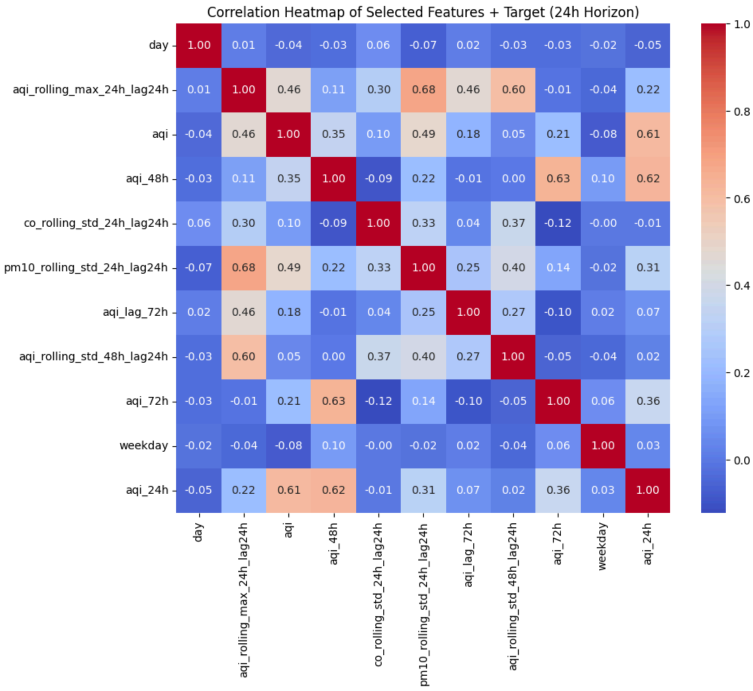
1. 3-fold TimeSeriesSplit validation
2. Calculate MI and RF scores per fold
3. Normalize and aggregate scores
4. Select top features based on combined ranking

Advantage: MI alone selected redundant features; RF alone missed subtle relationships; hybrid approach leveraged complementary strengths.

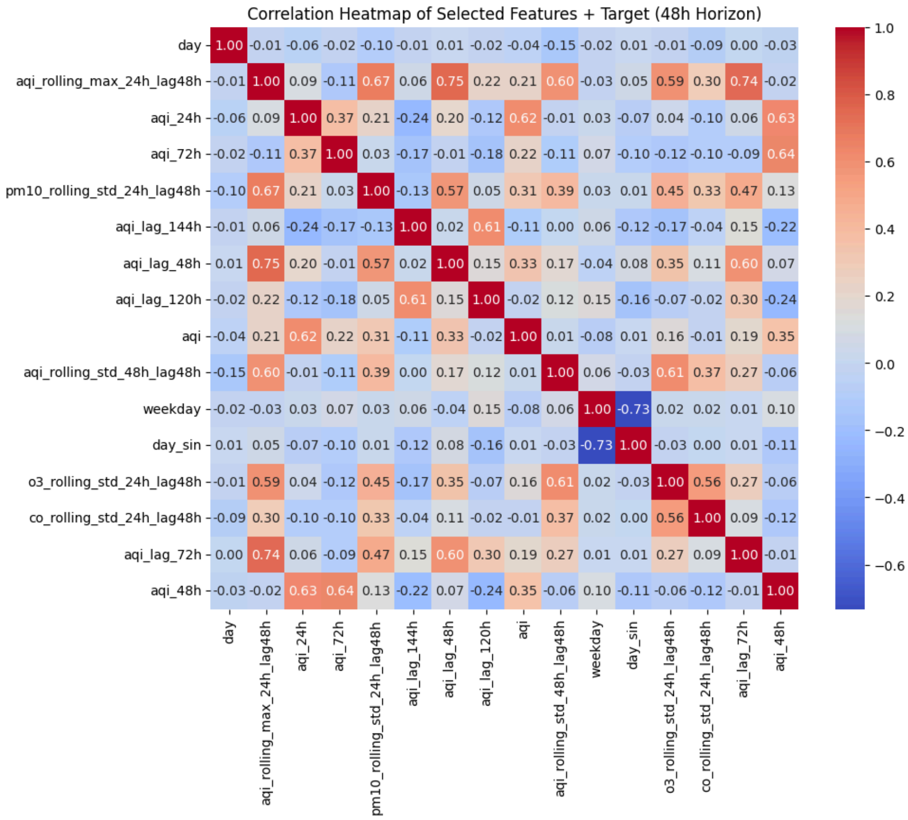
Correlation Heatmap of Numerical Columns

This heatmap shows the pairwise correlations between numerical features with their respective target columns:

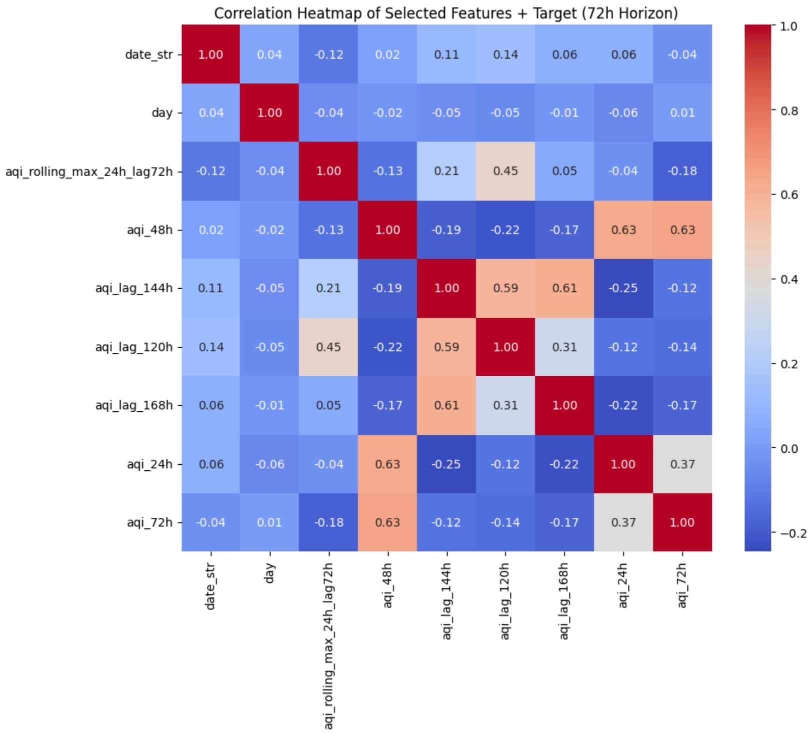
Target: aqi_24h



Target: aqi_48h



Target: aqi_72h



Model Architecture & Results

Algorithm Selection

Tested 7 tree-based algorithms: CatBoost, XGBoost, LightGBM, RandomForest, ExtraTrees, GradientBoosting, DecisionTree

Hyperparameter Optimization: Optuna TPE with R^2 objective, early stopping, cross-validated functions

Performance Results

Horizon	Best Model	Test R^2	Test RMSE	Test MAE	Test MAPE
24h	ExtraTrees	0.59	4.88	3.85	4.71%
48h	ExtraTrees	0.61	4.55	3.67	4.23%
72h	ExtraTrees	0.21	6.52	5.39	6.75%

Key Finding: ExtraTrees consistently outperformed due to additional randomization providing better generalization for Karachi's variable air quality patterns.

CI/CD Pipeline

Automated Workflows (GitHub Actions)

1. **Data Collection** (5 AM UTC) - Multi-source API data gathering
2. **Feature Preprocessing** (6 AM UTC) - Feature engineering pipeline
3. **Model Training** (7 AM UTC) - ML training with Hopsworks integration
4. **Model Updates** (Hourly/Daily) - Download models, update local data
5. **Feature Extraction** (8 AM UTC) - Export horizon-specific datasets

Integration: Hopsworks serves as a central feature store and model registry with automated versioning.

FastAPI Backend

Architecture

- **Models:** Multi-format support (CatBoost .cbm, XGBoost .json, LightGBM .txt, Scikit-learn .pkl)
- **Performance Optimization:** Reduced memory from >1GB to <200MB through local model storage
- **Security:** Rate limiting (10-20 req/min), CORS configuration, Pydantic validation

Key Endpoints

Endpoint	Purpose	Rate Limit
`/forecast`	AQI predictions	10/min
`/forecast/hourly`	Hour-by-hour interpolation	10/min
`/historical/{location}`	Historical data (1-21 days)	10/min
`/dashboard/overview`	Current status + trends	10/min

Streamlit Frontend

Features

- **Real-time Dashboard:** Live AQI monitoring with interactive charts
- **AI Predictions:** 72-hour forecasts with hourly resolution
- **Advanced Alerts:** Automatic hazardous condition monitoring with pulsing animations
- **Multi-city Comparison:** Ranking and comparative analysis
- **Responsive Design:** Mobile-first CSS with Google Fonts integration

Alert System

Monitors AQI thresholds (301+ Hazardous, 201+ Very Unhealthy) with contextual health recommendations and persistent session state management.

Deep Learning Experiments

Preliminary LSTM Testing

- **Architecture:** Basic LSTM models for 24h/48h/72h horizons
- **Loss Function:** MSE with time-series split validation
- **Status:** Exploratory phase, not yet optimized or deployed
- **Purpose:** Feasibility validation for future single-model architecture

Deployment & Performance

Current Architecture

- **Docker:** Multi-service containerization (frontend:8501, backend:8000)
- **Startup:** 1-2 minutes (80% improvement from original)

Challenges Overcome

- **Vercel Limitations:** 512MB memory limit exceeded, moved to Docker
- **Cold Start Issues:** Solved through local model storage strategy
- **Model Loading:** Multiple format support with automatic detection

Future Enhancements

Planned Deep Learning Transition

Current Challenge: 21 separate models (7 algorithms × 3 horizons) = several GB storage

Solution:

- Single neural network for all horizons
- Incremental training on existing weights
- Model size reduction: Several GB → 10-100 MB
- Transfer learning for faster convergence

Implementation Roadmap

1. Multi-horizon neural architecture design
2. Weight migration from existing models
3. Incremental training pipeline
4. Automated container deployment
5. Performance monitoring post-transition

Key Insights & Lessons Learned

1. **Forecasted AQI as Features:** Essential breakthrough - without T+48, T+72 AQI values, models showed negative R^2
2. **Hybrid Feature Selection:** Outperformed single-method approaches consistently
3. **Sliding Window Validation:** Critical for realistic performance estimation in atmospheric forecasting
4. **ExtraTrees Superiority:** Additional randomization provided best generalization for variable air quality patterns
5. **Architecture Optimization:** Local storage reduced memory 80% and startup time 80%
6. **Time Series Methodology:** Proper temporal validation essential - standard CV methods failed

Conclusion

AirLens successfully demonstrates production-ready air quality forecasting with significant technical innovations in feature selection methodology and system optimization. The hybrid MI-RF approach and architectural optimizations provide a scalable foundation for environmental monitoring applications, while the planned deep learning transition addresses current storage and deployment challenges.

The system delivers immediate public health value through accurate predictions and real-time alerts, establishing a comprehensive framework for urban air quality management in Pakistan's major cities.