SMART INDIA HACKATHON 2025

- **Problem Statement ID -** 25178
- **Problem Statement Title -** Short term forecast of gaseous air pollutants (ground-level O3 and NO2) using satellite and reanalysis data
- **Theme -** Space Technology
- **PS Category-** Software
- **Team Name -** TORQUE23 / 108944

TORQUE23

IDEA TITLE

SMART INDIA HACKATHON 2025

Challenges

- Prevent leakage: Input CSVs were originally shuffled; all rows were organized by date-time and split chronologically to avoid future information bleeding into training.
- Handle sparsity/noise: Robust lags and rolling stats keep forecasts stable when inputs are missing or noisy.
- Nail peak hours: Strong temporal features are needed to capture short, sharp O3/NO2 spikes.

Proposed Solution

- Train one Optuna-tuned LightGBM on advanced temporal and error-aware signals to predict hourly NO2 24–48 h ahead.
- Use chronological train/test, persist scaler/model, and generate QA plots for deployment readiness.
- At inference, standardize inputs, predict, and export hourly NO2 with forecast-vs-predicted visuals.
- Run site-wise, then aggregate to station cards and a lightweight map for operations.

Uniqueness

- Optuna-tuned LightGBM: Systematic hyperparameter search for high accuracy with low compute.
- Error-aware features: Lagged forecast_error reduces recurring bias at critical hours.
- Temporal richness: Cyclic time encodings plus multi-scale lags/rollings model diurnal dynamics.

Value Proposition

- Robust and continuous: Works despite missing data using strong temporal encodings, rollings and lags.
- Efficient and scalable: High skill at low compute enables daily citywide operation.
- Easy adoption: Transparent metrics, artifacts, and simple CSV/PDF deliverables let agencies act immediately.

Key Features

- Single LightGBM forecaster: Early-stopped, RMSE-oriented training for hourly predictions.
- Automated feature builder: Datetime, sin/cos encodings, lags/rollings, wind speed synthesis.
- Reproducible outputs: Persisted scaler/model, aligned columns for unseen data, CSV and plots.

TORQUE23

TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

Programing Languages

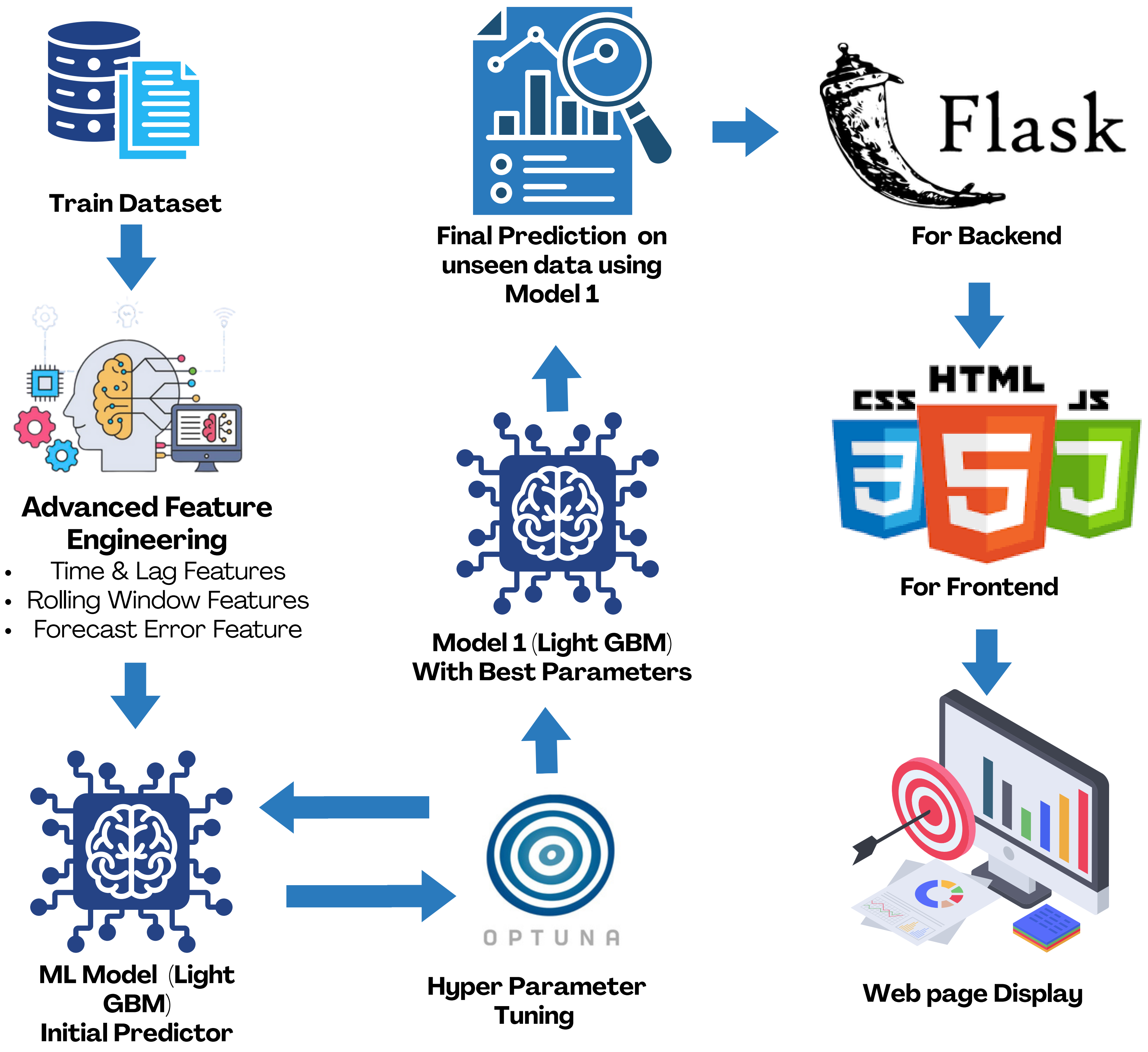
- **Python:** Used for all backend Machine Learning tasks like data preprocessing, feature engineering, model training and evaluation.
- **HTML, CSS and JavaScript:** For building the interactive frontend user interface.

Libraries and Framework

- **LightGBM:** A high-performance gradient boosting framework used to train the core predictive models.
- **Optuna:** Employed for advanced hyperparameter optimisation (up to 1000 trials) to maximise the model performance.
- **Pandas, NumPy, and Matplotlib:** Fundamental libraries for data processing, numerical computation, and visualisation.
- **Joblib:** Used to efficiently save and load trained model and scalers for deployment.
- **Flask:** A lightweight backend framework for creating efficient APIs to serve model predictions.

Click Here- Github Repository/link
Youtube Video/link

Key Features:



TORQUE23

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2025

Analysis of the feasibility of the idea

- **Scalability**: Site-agnostic pipeline; add stations by dropping new CSVs and reusing the same inference flow.
- **Operational Readiness**: Daily PDFs, CSVs, enable immediate use without heavy IT.
- **Technical Feasibility:** Uses established ML (LightGBM), time-aware splitting, and automated feature engineering; runs reliably on standard CPUs/GPUs.
- **Market Viability**: Serves civic agencies needing hourly O3/NO2 guidance for health advisories, traffic, and construction windows.

Photochemistry dynamics

Complex, time-dependent chemical reactions of pollutants like NO_2 and O_3 create unpredictable behavior, hindering accurate forecasts of pollution peaks.

Solution:

Our model uses strong temporal features and photochemical principles to learn these nonlinear dynamics and precisely predict peak timing and magnitude.

Atmospheric Uncertainty

Rapid shifts in wind and boundary layer height can quickly dilute or trap pollutants, leading to highly volatile and uncertain forecast outcomes.

Solution:

We integrate real-time atmospheric data, like wind sectors and boundary proxies, to capture these sudden changes and produce more stable forecasts.

Data Inconsistency

Unreliable ground-truth data from monitoring stations due to outages, drift, and errors introduces significant label noise, degrading model performance.

Solution:

Our data pipeline uses quality control flags, smoothing, and robust loss functions to effectively filter this noise and train on clean, reliable data.

TORQUE23

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2025

1
Early Warning for Air Pollution Events
Public advisories can be released to limit outdoor exposure, particularly protecting vulnerable groups like children, elderly, and people

2
Compare Projections
Better preparedness in hospitals and public health systems; potential to save lives and reduce healthcare costs.

3
Analysis of Policy Outcomes
By comparing the predicted pollution (without intervention) to the actual observed pollution (with intervention), the model can be used to quantify the effectiveness of a policy change (e.g., a ban on biomass burning).

4
Data-Driven Policy
Allows the government to optimize environmental spending by showing which measures yield the greatest air quality improvement.

5
Always-On
When satellite swaths are missing or cloudy, the system falls back to meteorology and lagged signals, keeping forecasts uninterrupted for continuous service.

6
Ready to Pilot
From day one, agencies receive a one-page PDF brief, machine-readable CSVs, and a lightweight web map—outputs they can act on immediately without heavy IT integration

TORQUE23

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2025

Challenges Solved:

- **Temporal features for robustness:** Multi-scale lags and rolling stats stabilize learning under inconsistent/missing hourly data; prevents overreacting to short gaps.
- **Chronological splits (no leakage):** Train/valid/test are time-ordered with train-only scaling, so metrics reflect real deployment.
- **Model choice:** LightGBM for tabular forecasting balances accuracy, speed, and interpretability; handles mixed feature types well.

References:

- https://pmc.ncbi.nlm.nih.gov/articles/PMC11774898/
- https://www.researchgate.net/publication/336874395_Ground_Ozone_Level_Prediction_Using_Machine_Learning
- https://aaqr.org/articles/aaqr-20-07-oa-0471
- https://www.aqi.in/in/dashboard/india/delhi/new-delhi