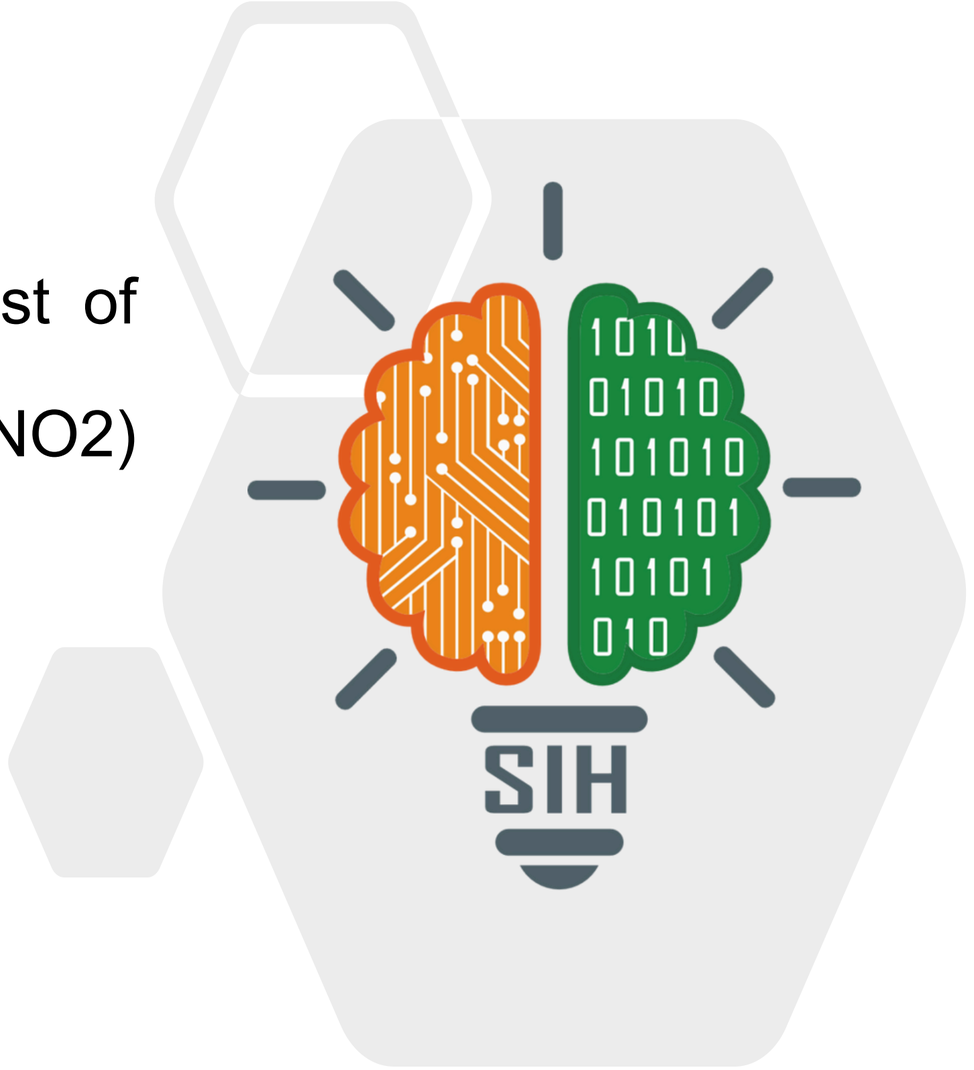
# SMART INDIA HACKATHON 2025

## TITLE PAGE

- **Problem Statement ID –** SIH25178
- **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 ID –** 65478
- **Team Name –** Dues-Ex-Machina

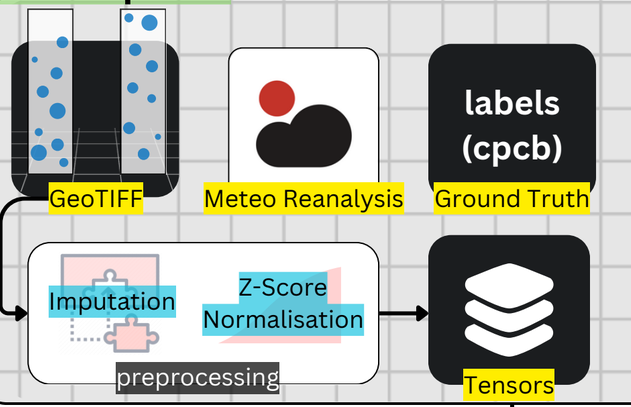


DUES-EX-MACHINA

# PROJECT CANARY

SMART INDIA HACKATHON 2025

## Data Pipeline

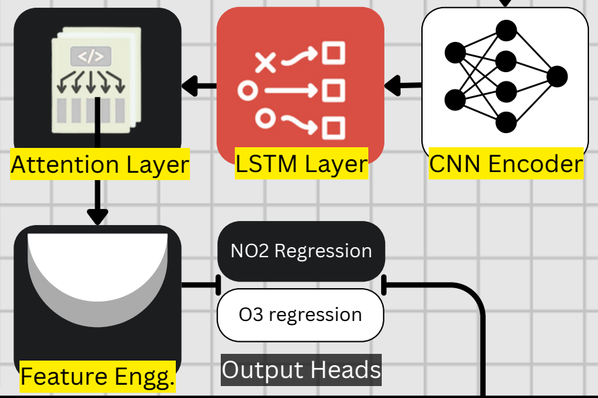



**Satellite Data (Sentinel-5P TROPOMI):** Tropospheric $NO_2$ and $O_3$ densities, processed from GeoTIFF to spatiotemporal tensors.

**Meteorological Reanalysis (ERA5 & CPCB):** Hourly variables interpolated to station coordinates.

**Ground Monitoring Data (CPCB):** Hourly station-level $NO_2$ and $O_3$, used as supervised labels.

## ML Architecture



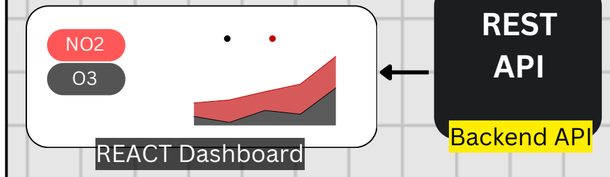

**CNN Encoder:** Extracts spatial features from gridded rasters.

**biLSTM Layer:** Captures temporal dependencies from the past 24 hours of pollutant + meteorology data.

**Attention Mechanism:** Dynamically weights recent timesteps to emphasize critical events (example, rush-hour emissions)

**Feature Engineering.** Normalized column densities by PBL height to approximate surface-level pollutant concentrations.

## Deployment




**Data & Model:** Processed datasets stored as .npy tensors → used to train PyTorch forecasting model.

**Serving & Visualization:** Trained model exposed via REST API backend → results displayed on lightweight dashboard.

## WHAT MAKES IT UNIQUE

### MULTI-SOURCE DATA FUSION

Conventional models: ground data only. We integrate: **satellite tropospheric columns + reanalysis meteo + ground truth labels**.

### DEPLOYMENT READY

Design allows integration with **real-time APIs** (OpenAQ, Copernicus ADS, GEE) for live deployment.

### SCALABILITY & GENERALIZATION

Architecture supports **transfer learning**, pretrained on Delhi, adaptable to other metros or Tier-2 cities with limited ground data.

### AUTOMATED RETRAINING

**Updates model weights** as new data arrives, ensuring continuous model **adaptability** without manual intervention.

## WHAT IT SOLVES

### SPARSE COVERAGE

Satellite products bridge gaps where **station density is insufficient**, while CPCB provides calibration.

### DECISION SUPPORT

Helps govt and orgs take **data-driven decisions** to help in policy making.

### MORE THAN JUST PREDICTION

The Hybrid model captures: pollutant spread across space (CNN), temporal evolution (LSTM), and key weather influences (Attention).

DUES-EX-MACHINA

# TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

## DATA PREPROCESSING

Daily $NO_2$/$O_3$ 10x11 Raster Grids

**Hourly Variables**: AT, RH, Precipitation, PBL Height, Wind, Solar Radiation

**Derived Variables**: Temporal Indices, u/v Wind, Dewpoint, Time Cycle Tags

Hourly NO2/O3 **Labels** — CPCB Stations

Temporal Interpolation → KNN → Imputation

**70** Train / **30** Test — Split

**Standard Scaler**: Fit On Train Set Only — Mean = 0, STD = 1

Final Inputs:
- **Satellite** (T,10,11,2)
- **Station** (T,11)
- **Labels** (T,2) for NO2 & O3

## MODEL ARCHITECTURE

**Input 1** — **9-Channel Grids**: $NO_2$ $SO_2$ CO HCHO $O_3$ $H_2O$ Cloud $CH_4$ Aerosol

CNN: Convolution → B. Normalize → ReLU → Max Pooling (Stack x3) → **64**-Dim Satellite Embedding

**Input 2** — **11 Tabular Features**

MLP: FC: 11 → 128 → ReLU → FC: 128 → 64 → **64**-Dim Station Embedding

Fusion (Concat): Satellite 64 + Station 64 → **128** Dim

biLSTM: 128 → 128

**256**-Dim Temporal Embedding

Attention Layer: Bahdanau

**256**-Dim Context Vector

**Output Heads**: NO2, O3 — 24h Forecast

## IMPLEMENTATION PROCESS

### Data Preprocessing

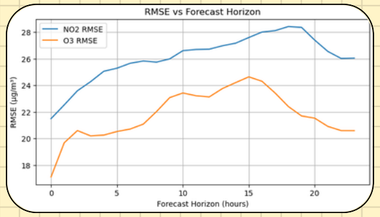
- **Data Integration:** Combined Sentinel-5P satellite grids ($NO_2$, $O_3$) with ERA5 weather data and CPCB ground-station readings.
- **Imputation:** Handled gaps using imputation (prev-day fill, KNN), normalized with StandardScaler, and reduced using PCA.
- **Final inputs:** Satellite tensors (T,9,10,2), station features (T,18), and hourly labels (T,2) for $NO_2$ & $O_3$.

### model architecture

- **CNN Encoder:** Two convolutional layers with pooling extract spatial patterns in $NO_2$ and $O_3$ grids, forming 32-dim feature maps.
- **Temporal Modeling (BiLSTM):** Captures both short-term (hourly) and long-term (seasonal) dependencies w/ biLSTM (hidden = 256).
- **Feature Fusion:** Combines CNN and LSTM outputs with ERA5-derived station features for richer contextual learning.
- **Attention Mechanism:** Dynamically weights meteorological drivers such as wind and boundary-layer height for explainable forecasts.
- **Dual Regression Heads:** Predict hourly $NO_2$ and $O_3$ concentrations simultaneously across the next forecast window.

## THE PROTOTYPE

### Prediction Results




Cick To View

| | MAE | RMSE | $R^2$ |
|---|---|---|---|
| NO2 | 12.35 | 17.89 | 0.755 |
| O3 | 14.65 | 20.91 | 0.666 |

### Web Dashboard


Click To Visit

Repository Link

## AUTO RETRAINING

Auto Fetch Live Data → Preprocess Pipeline → MLflow Tracking

MLflow Tracking: Trigger Retraining, Log Parameters, Version Models

Model Evaluation: New vs Baseline (R², RMSE)

Promote best model to Production

## TECH STACK

| | |
|---|---|
| MLflow | Automate retraining for continuous model improvement. |
| Scikit-learn | To preprocess data (imputation, scaling, PCA) |
| PyTorch | To implement model, training & inference. |
| Pandas | Handling tabular data and time-series alignment. |
| NumPy | Scientific calculations & processing data. |
| Rasterio | Reading and processing GeoTIFF satellite rasters. |
| Django | REST API backend for serving forecasts. |
| React.js | Interactive dashboard for visualization. |

DUES-EX-MACHINA

# FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2025

## FEASIBILITY

### DATA AVAILABILITY

Uses **open-source**, publicly available datasets (Sentinel-5P, ERA5, CPCB).

### DIVERSE COVERAGE

Covers **diverse variables** across multi-year timeframes for robust training.

### EASY INTEGRATION

**Easily integrates** with dashboards, APIs, and existing AQI systems.

## VIABILITY

### AUTO RETRAINING

Supports **automated retraining** with incoming data streams (MLflow).

### ADAPTABILITY

**Extendable** to additional pollutants ($PM_{2.5}$, $SO_2$, CO) without redesign.

### COST EFFICIENT

Ensures **low operational costs** (open data + light deployment).

### IT IS SCALABLE

Scales from city-level to **pan-India** forecasts with minimal effort.

## CURRENT CHALLENGES

## PLANNED SOLUTIONS

| | |
|---|---|
| **Taking live data as inputs** | Transitioning to live API libraries of Sentinel and ERA5 for **real-time** ingestion. |
| **Low res of ERA5 data** | Aggregate **multiple CPCB stations** within each cell by averaging to align with ERA5 resolution. |
| **High computational load** | Optimize w/ **model pruning**, distributed training & GPU accn. to cut train time and inference costs. |

## PAST CHALLENGES

## HOW WE SOLVED THEM

| | |
|---|---|
| **Spatial alignment in sat. data** | Rasters cropped to Delhi AOI & resampled to fixed 10×11 grid for consistent pixel tracking. |
| **Temporal mismatch bw datasets** | Built a robust data pipeline for timestamp alignment and resampling across sources. |
| **Joint encoding of $NO_2$ & $O_3$ is hard due to differing chemistry** | Optimize w/ **model pruning**, distributed training & GPU accn. to cut train time and inference costs. |

DUES-EX-MACHINA

# IMPACTS AND BENEFITS

SMART INDIA HACKATHON 2025

## BENEFITS

### REDUCE STATION DEPENDECY

Satellite–ERA5 fusion enables forecasts even in areas with **sparse monitoring** coverage.

### SCALABILITY

Same pipeline can **extend** from Delhi to **pan-India** with minimal retraining.

### COST-EFFECTIVENESS

**Open-access** datasets + API deployment means **lower cost** than expanding ground station infrastructure.

### DECISION SUPPORT

Helps orgs & govt bodies make **data-driven decisions**, to guide policy, and response actions.

### DATA DEMOCRATIZATION

Converts complex data into **accessible**, **interpretable** insights for researchers.

## IMPACTS

### PUBLIC AWARENESS

Forecast-driven **AQI dashboards** provide citizens with actionable early warnings.

### MODEL EXPLAINABILITY

Attention-based feature weighting highlights critical meteo drivers , improving **interpretability** for experts.

### RESEARCH ACCELERATION

Creates high-res, multi-source **spatiotemporal datasets** for academic and policy studies.

## USE CASES

**Researchers** use the high-res, explainable dataset to analyze pollutant dynamics.

**Citizens & policymakers** get to monitor pollution trends effectively.

**Health departments** get predictive support, alert system and IOT integration.

DUES-EX-MACHINA

# RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2025

## RESEARCH

**Research paper**

Attention mechanism based CNN-LSTM hybrid deep learning model for atmospheric ozone concentration prediction.
by: Jiang Yuan, Hua Dengxin, Wang Yufeng, Yang Xueting, Di Huige & Yan Qing

**Research paper**

Urban Air Pollution Forecasting: a Machine Learning Approach leveraging Satellite Observations and Meteorological Forecasts.
by: Giacomo Blanco, Luca Barco, Lorenzo Innocenti & Claudio Rossi

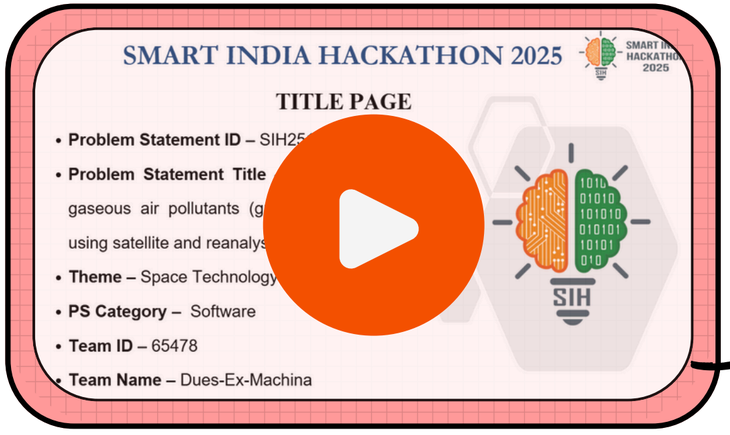
## REFERENCES

[1] Detailed idea approach

[2] Detailed technical approach

[3] GitHub repository of the project

Datasets: [4] raw data [5] processed data




## EXPLANATORY VIDEO