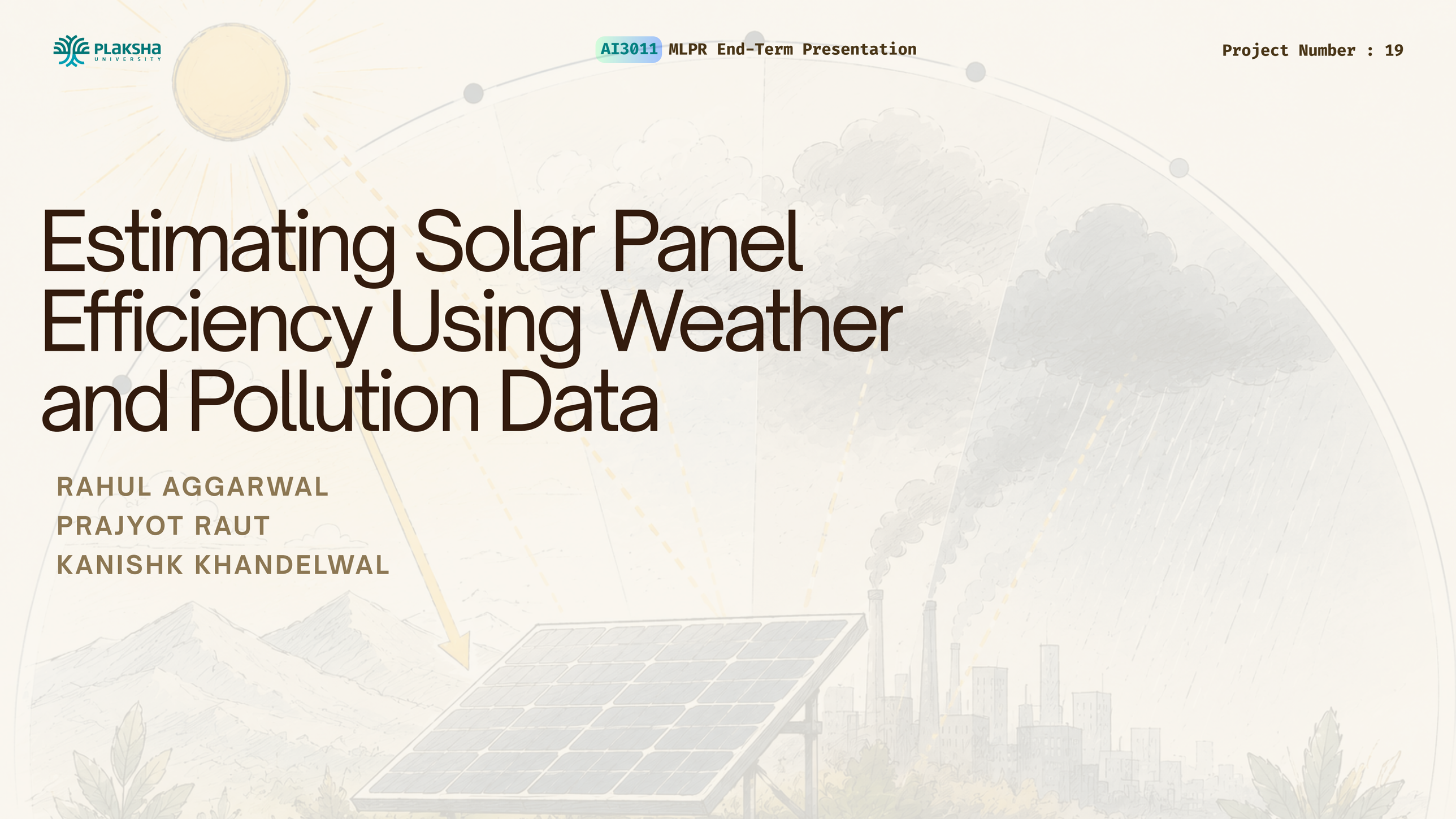
Estimating Solar Panel Efficiency Using Weather and Pollution Data

RAHUL AGGARWAL
PRAJYOT RAUT
KANISHK KHANDELWAL



Problem Statement

Solar panel efficiency depends on the amount of sunlight reaching the Earth's surface, which is strongly influenced by atmospheric conditions such as air pollution , sun position and weather. These environmental factors can reduce the solar radiation available for energy generation.

Pollution particles scatter and absorb sunlight

This makes solar energy harder to estimate

Clouds and humidity reduce surface radiation

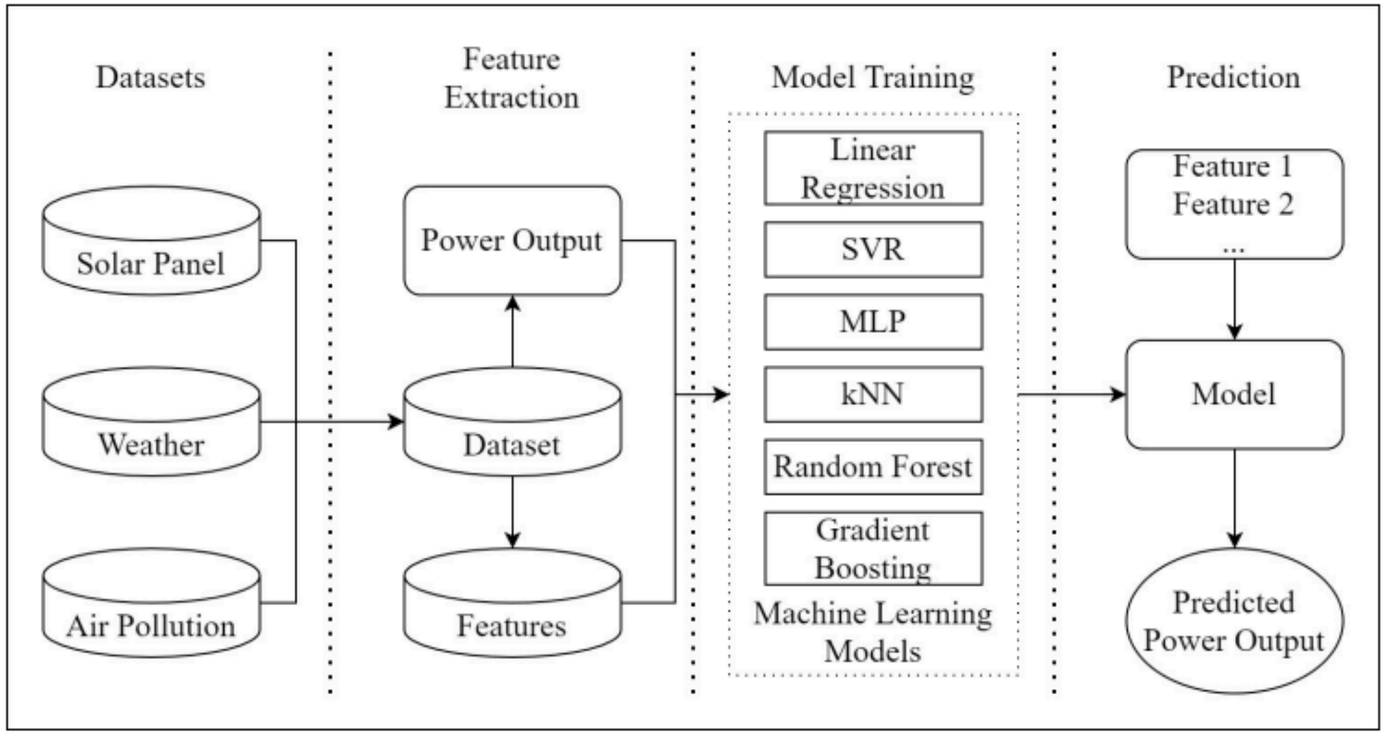
Literature Review

Solar Power Prediction Using Weather and Pollution Data (2021)

Author: Tserenpurev Chuluunsaikhan,Aziz Nasridinov, Woo Seok Choi, Da Bin Choi, Sang Hyun Choi, Young Myoung Kim

Dataset: Solar panel output data from Daeyeon C&I, Weather data from Korea Meteorological Administration, Air pollution data from Seoul Metropolitan Government

Results: Random Forest achieved the best performance with **≈98% R^2, RMSE ≈0.89.**



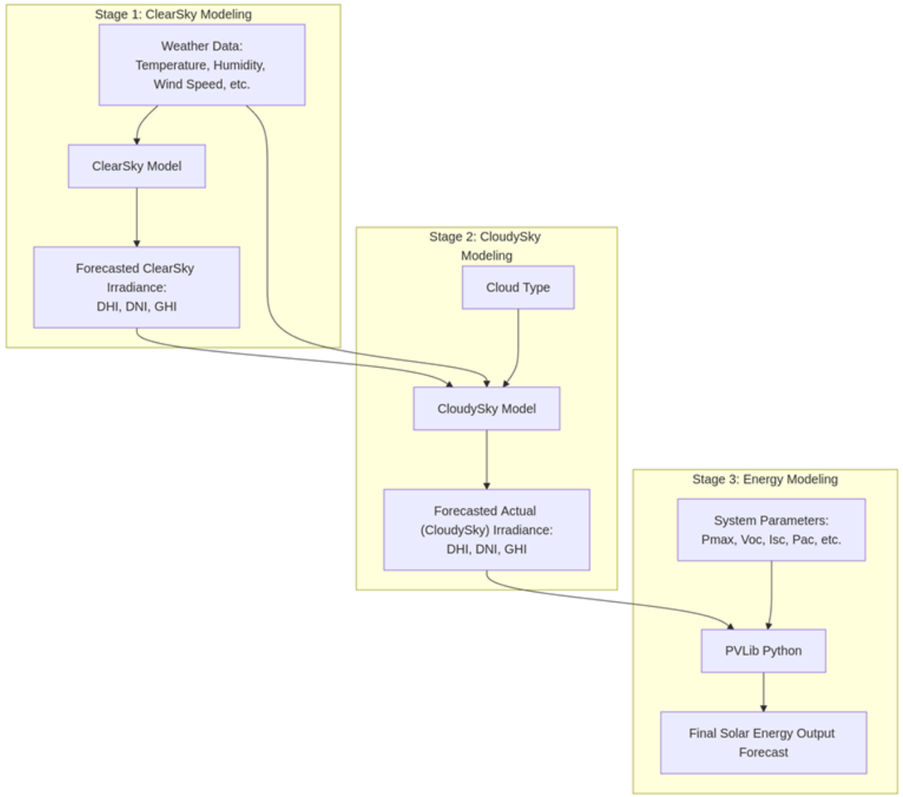
Forecasting solar power output in Ibadan: A machine learning approach leveraging weather data and system specifications

Author : Obarotu Peter Urhuerhi,Chrisopher Udomboso and Caston Sigauke.

Dataset: Solar radiation and weather data from the NSRDB database(2005–2022) for Ibadan, Nigeria

Results: Random Forest performed best, achieving **nRMSE ≈ 0.19**

Key Insight: Combining machine learning with physical PV models improves prediction reliability.



Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features

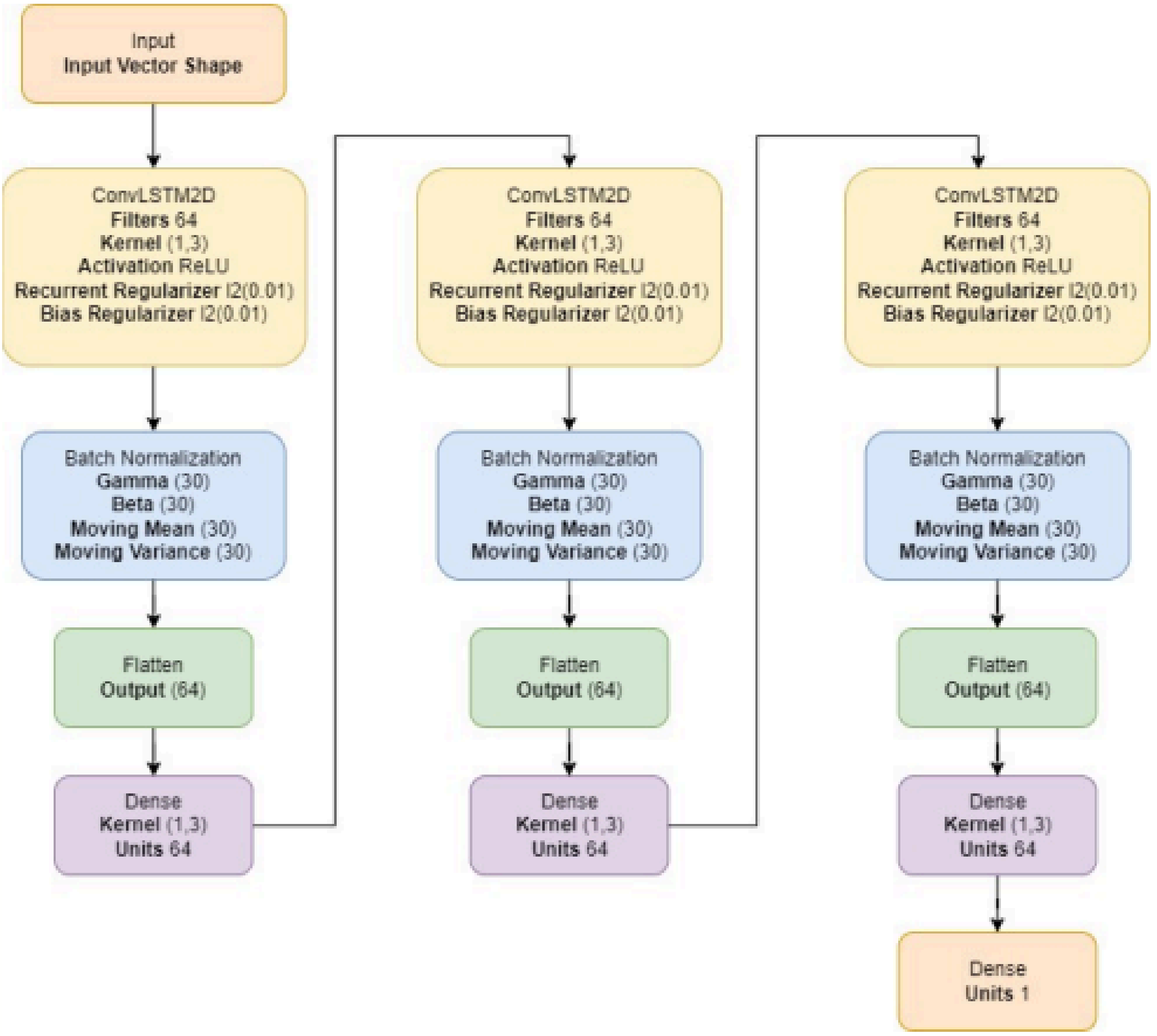
Authors: Arjun Shah, Varun Viswanath, Kashish Gandhi, Dr. Nilesh Madhukar Patil

Dataset was sourced from (La Trobe University) along with nearest AQI station (Macleod, Victoria)

R2 Score: 0.9691

MAE (Mean Absolute Error): 0.18.

RMSE (Root Mean Squared Error): 0.10.



The Gap

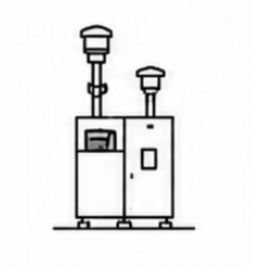
No study combines pollution, weather, and solar geometry together. Existing work uses only pollution + weather or weather + geometry, causing systematic prediction errors.

No India-wide GHI forecasting pipeline exists. Existing studies cover only single sites or small city sets, leaving India’s pollution-driven solar variability largely unaddressed.

Existing models use generic AQI data, not ground-sensor resolution. None use station-level CPCB hourly PM2.5/PM10 telemetry for localized aerosol attenuation.

Dataset

Ground Data



340+ CPCB Stations
PM2.5, PM10 (hourly)
GHI (ground measurements)

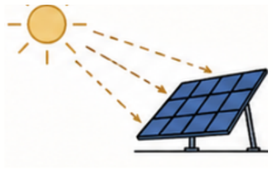
Satellite Datta (ERA-5)

Temperature, Humidity,Wind, Pressure,Cloud Cover, Precipitatin

Spartial Allignment

Solar Zenith Angle , Solar Azimuth Angle , Earth–Sun Distance

Solar Geometry



340+ CPCB Stations
PM2.5, PM10 (hourly)
GHI (ground measurements)

Final Dataset

Hourly Data
2023|2024|2025

Ethical concerns: All three (Pollution - CPCB , Weather - ERA5) are open government or institutional datasets with no personally identifiable information. Therefore , There were no significant ethical concerns in this study.

Features Preprocessing

Data Processing Pipeline

Data Collected

Collected nationwide CPCB station data from 2020–2025.

Station Filtering

Selected top 3 stations with minimum missing values.

Quality vs Scale

Reduced dataset size while preserving nationwide data diversity.

25 Lakh Observations

Features Preprocessing

Cyclical Time Encoding

Hour and month features were cyclically encoded to preserve solar and seasonal patterns.

Temporal Lag Feature Engineering

Created 1-hour and 24-hour lag features to capture temporal dependencies and recurring daily patterns.

Feature Distribution Normalization

Applied Yeo-Johnson Power Transformation for feature scaling to reduce skewness, stabilize variance, and improve overall model training performance.

Features Preprocessing

Challanges Faced

Data Integration Challenges

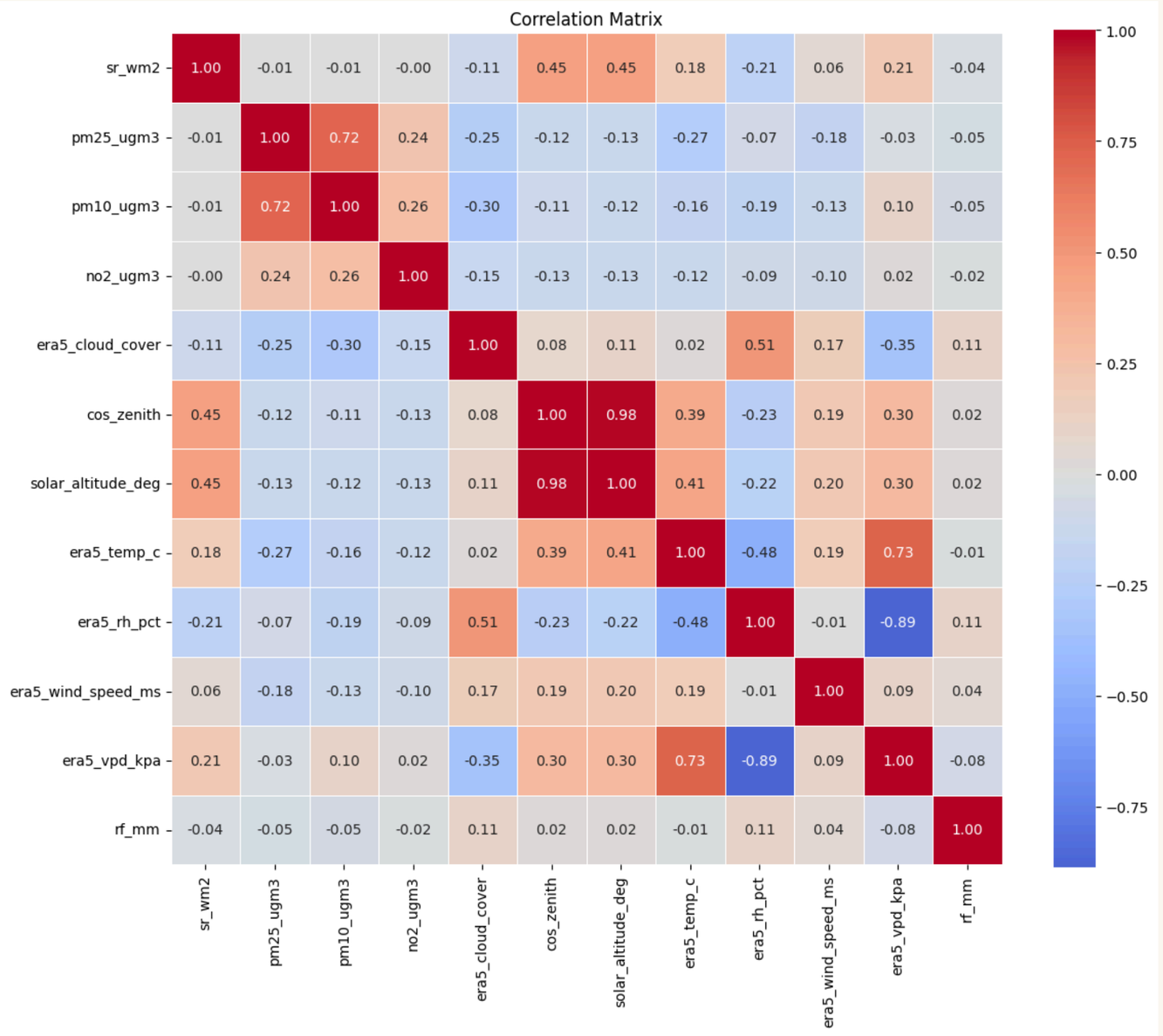
Combining pollution, solar, and weather datasets required aligning timestamps, locations.

Data Reduction vs Robustness

Balancing dataset reduction with nationwide model robustness became a major computational challenge.

Correlation Matrix

The correlation matrix validates our hypothesis : solar geometry drives GHI directly (~0.45), while pollution and weather variables introduce attenuation effects, confirming that all three data streams are necessary for accurate prediction.



Machine Learning Methodology

Model 1 : XGBOOST

XGBoost Hyperparameters

- Objective Function: reg:squarederror
- Evaluation Metric: RMSE
- Learning Rate: Optimized using Optuna
- **Maximum Tree Depth: 8**
- **Subsample Ratio: 0.9**
- **Boosting Rounds: 1000**
- **Early Stopping Rounds: 50**

Results

R2 : 0.8135

RMSE : 92.0942 W/m²

MAE : 44.0910 W/m²

Machine Learning Methodology

Model 2 : LightGBM

Light GBM Parameters

- objective: regression
- metric used : rmse
- boosting type: gbdt
- **learning rate: 0.05**
- **number of leaves: 63**
- **max depth: 8**
- **feature_fraction: 0.8**

Results

R^2: 0.8194

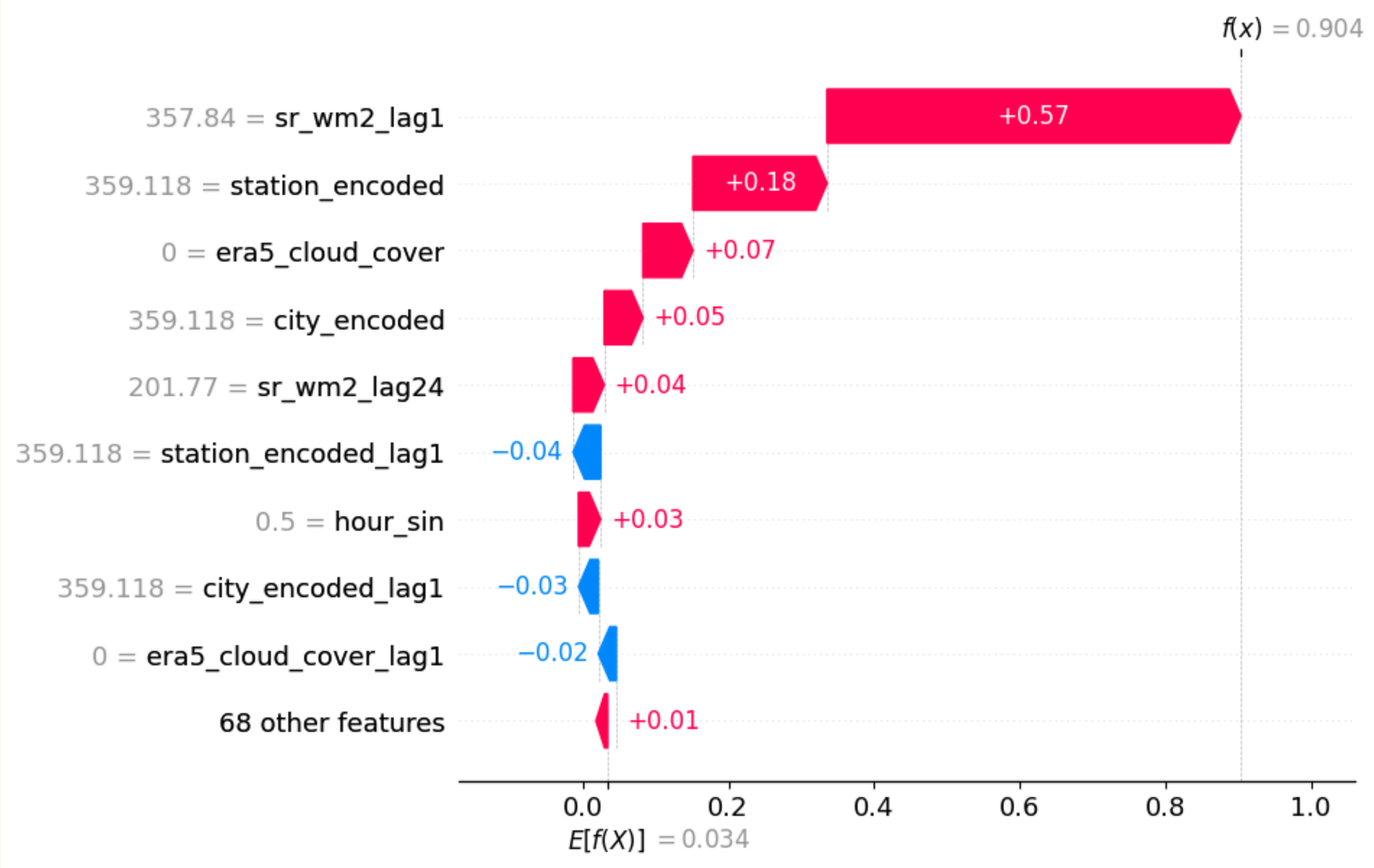
MAE: 44.35 W/m²

RMSE: 90.62 W/m²

Machine Learning Methodology

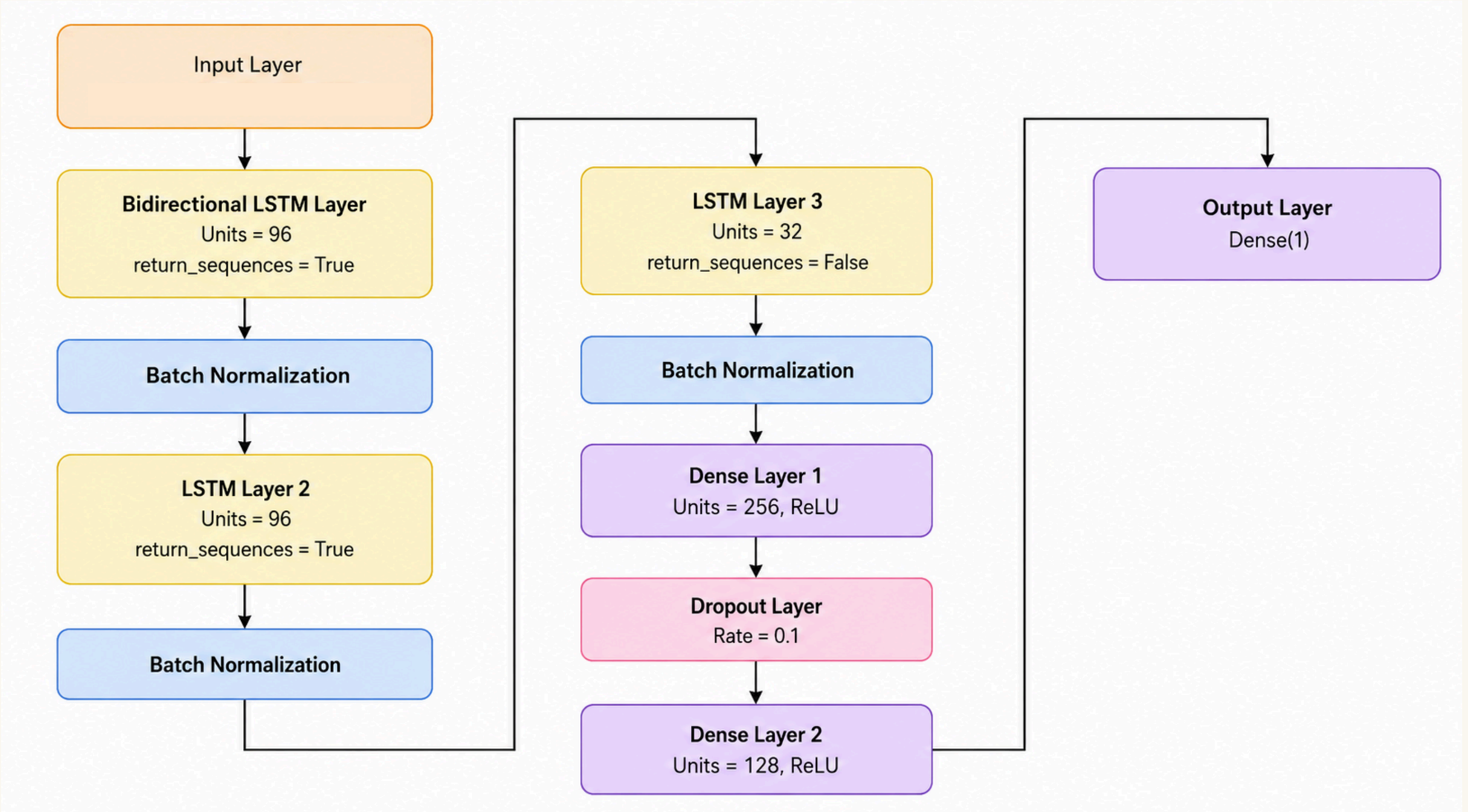
Model 2 : LightGBM

SHAP Analysis:

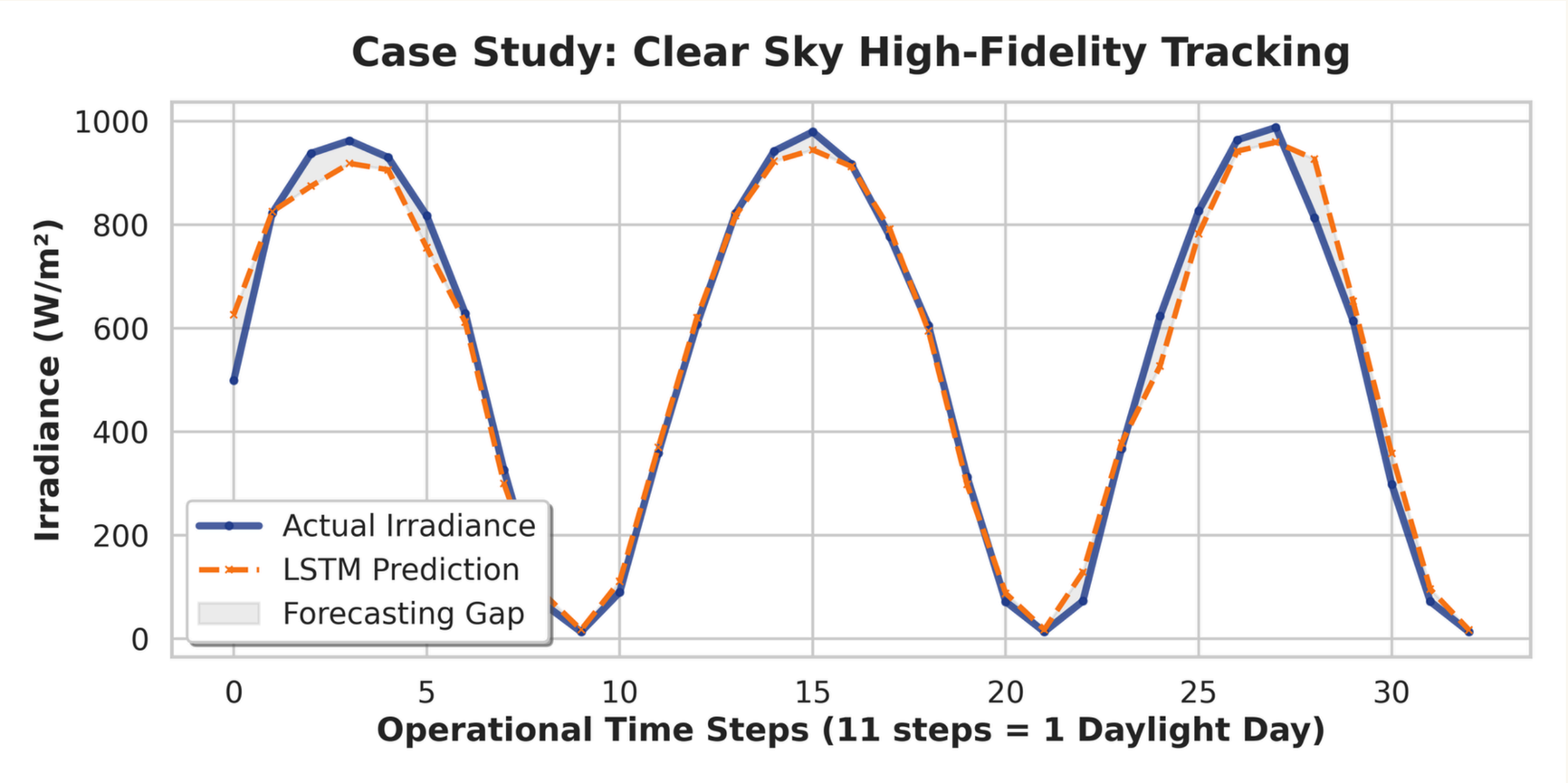


Machine Learning Methodology

Model 3 : LSTM



Machine Learning Methodology



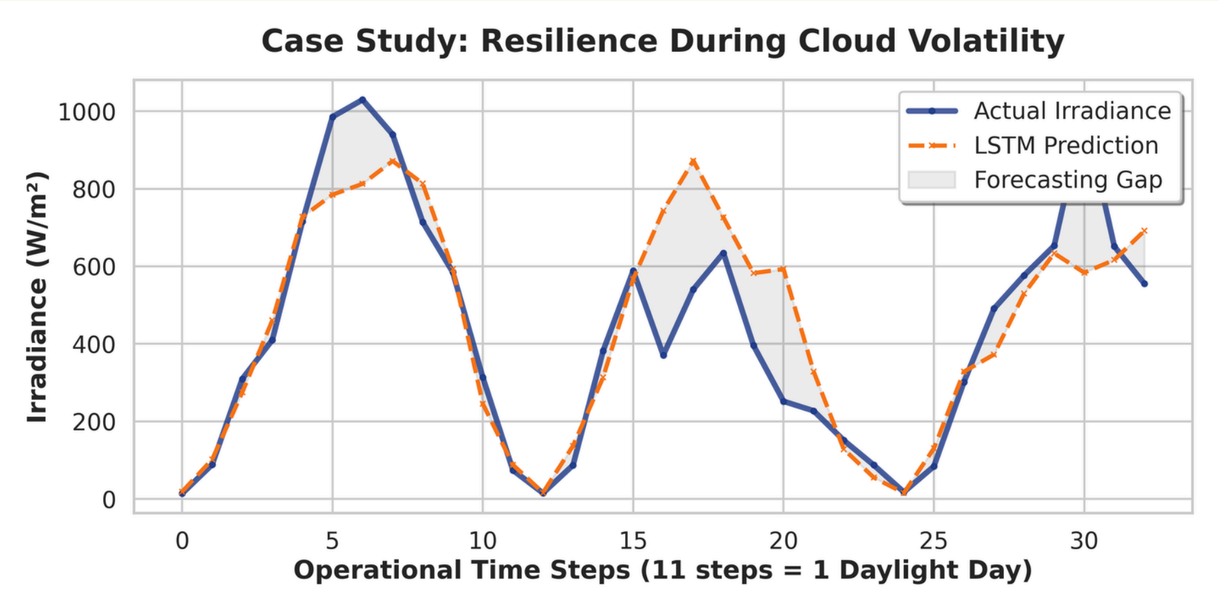
Model 3 : LSTM

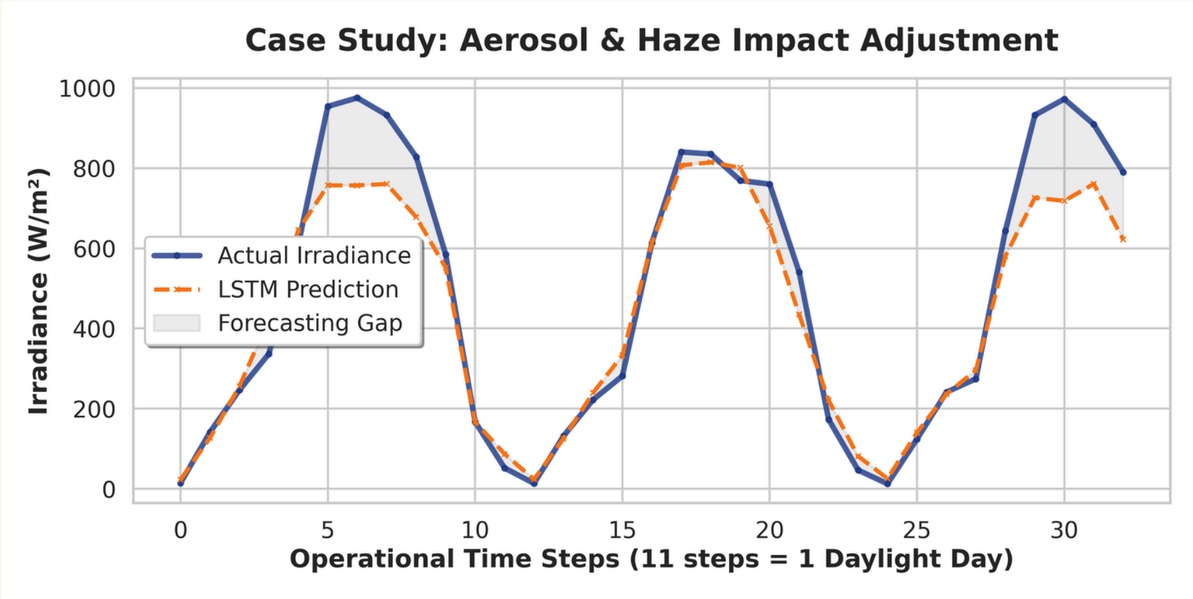
Results

R2 Score: 0.8669
MAE: 39.1769 W/m²

Machine Learning Methodology

Model 3 : LSTM





Metrics and Benchmarking

Paper	Benchmark
Chuluunsaikhan et al., Solar Power Prediction Using Weather and Pollution Data, 2021.	Random Forest ≈98% R^2, RMSE ≈0.89.
Urhuerhi et al., Forecasting Solar Power Output in Ibadan: A Machine Learning Approach Leveraging Weather Data and System Specifications.	Random Forest nRMSE ≈ 0.19
Shah et al., Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features.	R2 Score: 0.9691 MAE: 0.18. RMSE: 0.10.

OUR LSTM

R2 Score:0.8669
nMAE: 0.19

Future Scope & next steps

Scaling Challenges

- Distribution shift: A model trained on 2023–2025 data may drift as climate patterns shift
- Compute at inference: Running hourly predictions across hundreds of CPCB stations simultaneously requires Large Compute.

Future Scope

- Integrate satellite aerosol data for improved atmospheric and pollution-aware solar forecasting.
- Develop advanced hybrid deep learning models for more accurate large-scale predictions.

Can It Be Deployed at Plaksha?

Yes, It can be Deployed to a cloud infrastructure for inference.
Every hour CPCB pollution readings from nearest station in Mohali and ERA-5 weather forecasts are fed into the trained LSTM, which outputs a GHI forecast for the next 1–24 hours

Thank You