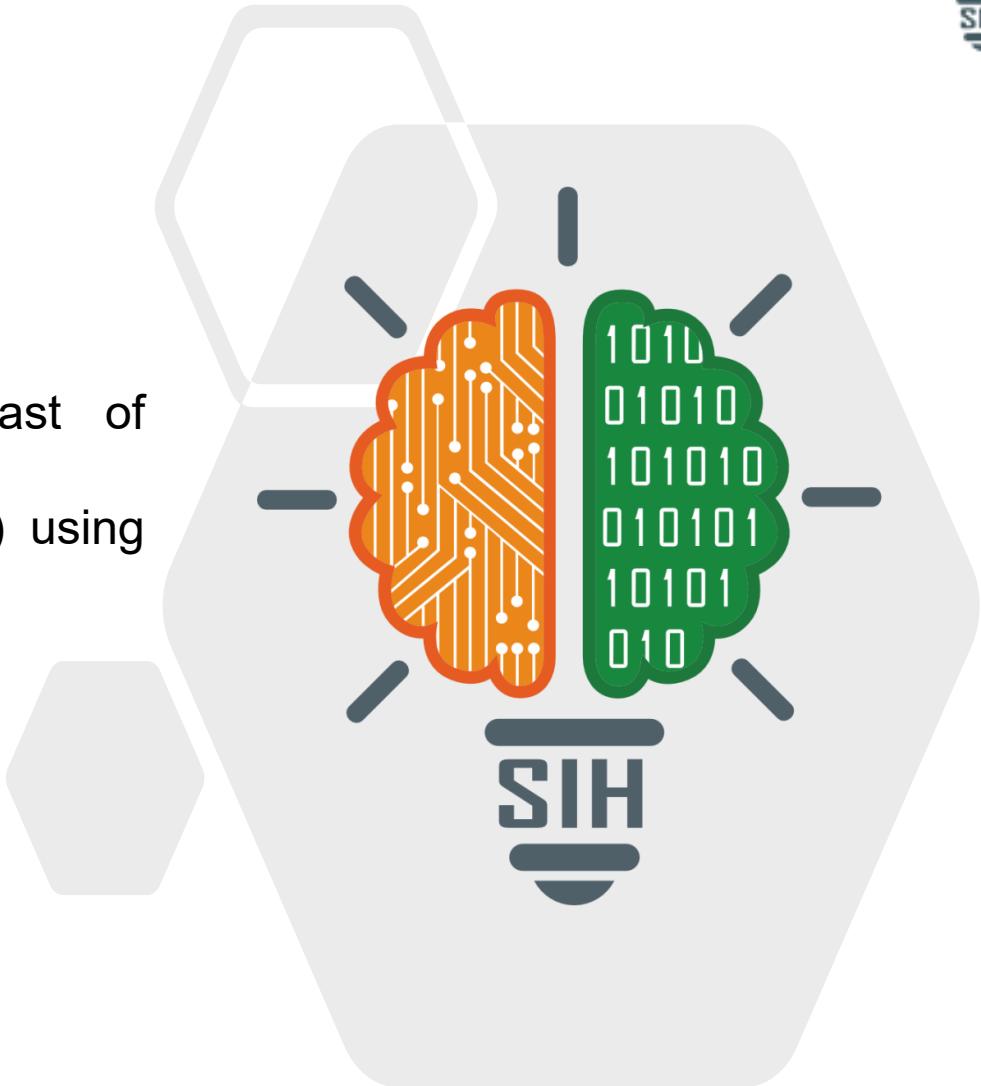


SMART INDIA HACKATHON 2025



- **Problem Statement ID – SIH25178**
- **Problem Statement Title-** Short term forecast of gaseous air pollutants (ground-level O₃ & NO₂) using satellite and reanalysis data
- **Theme-** Space Technology
- **PS Category-** Software
- **Team ID-** 114573
- **Team Name –** Skylytics1.0



SAFAL – System of Air Forecasting with Artificial intelligence and machine Learning the next generation SAFAR



Proposed Solution :

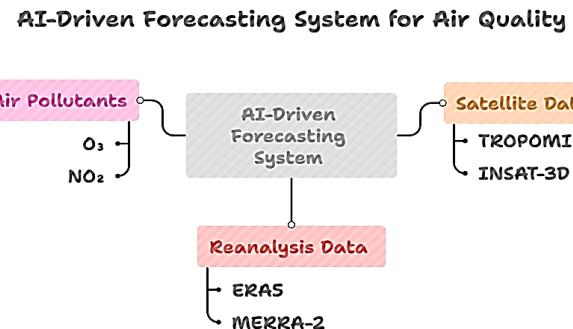
Idea & Solution:

Unified AI Pipeline

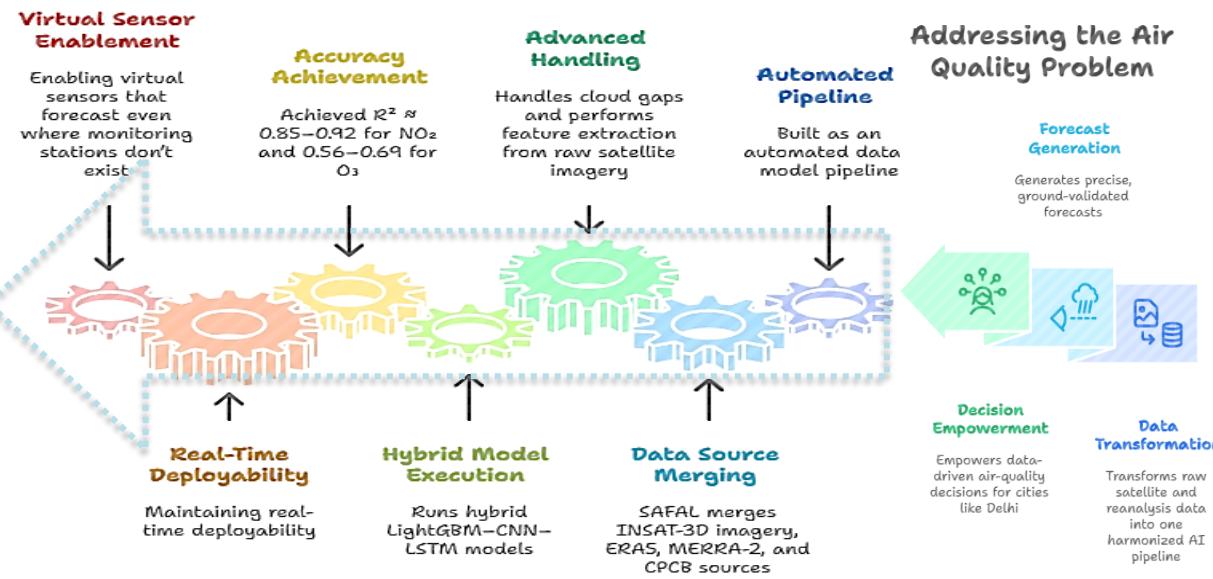
Multisource fusion for 24–48 h air-quality forecasts, upgrading the SAFAR system.



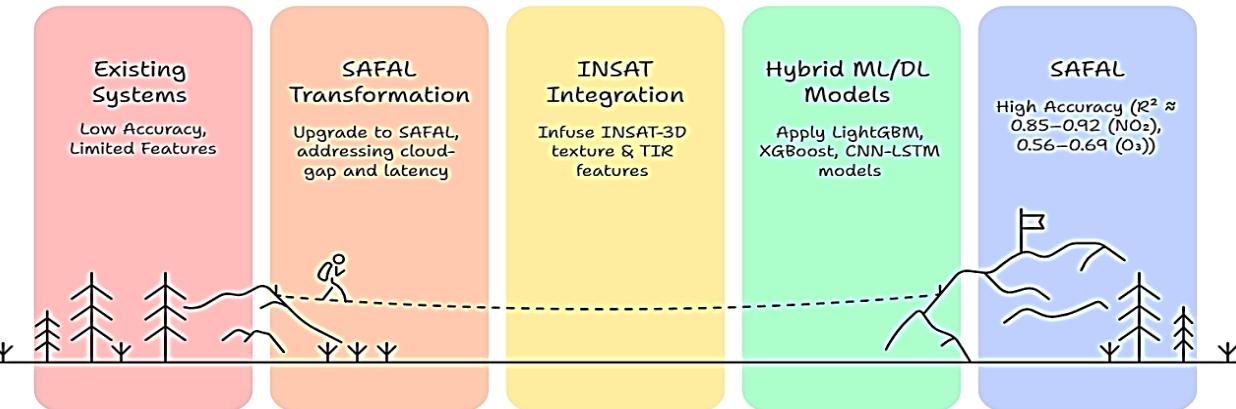
Fully automated — data ingestion → preprocessing → hybrid ML/DL (LightGBM, XGBoost, LSTM) → SHAP explainability.



Prototype & How it addresses the problem:



Innovation & Uniqueness:



Detailed Explanation of proposed solution:

What it Does

24–48 h NO_2 & O_3 forecasts via multi-source fusion — INSAT-3D, ERA5, MERRA-2, CPCB, Given data.

How it Works

Auto-ingests by coordinates, syncs time, crops INSAT tiles, extracts TIR features, merges hourly.

Feature & Model Engine

used approx. 160 features; LightGBM/XGBoost + LSTM/CNN ensemble for spatiotemporal learning.

Outputs & Explainability

Point forecasts + uncertainty bounds + SHAP-based interpretability.

Innovation & Impact

AI-driven cloud-gap infill, virtual sensors, real-time reliable coverage, updating existing system.

Validation & Readiness

trained & tested on 87600 hours of data. Fine tuned for 5 cities of Delhi.

TECHNICAL APPROACH



Prototype Summary — Delhi Focus

Data: 2 yrs hourly Delhi data

Features: 160+ total, top 20 used

Technical Approach

ERA5/MERRA-2: Auto-fetched via API + NetCDF4

INSAT-3D: Cropped for Delhi (rasterio, geopandas)

Sync: INSAT (½ hr) | ERA/MERRA (UTC) | CPCB (IST)

Gaps: 36K hrs recovered (5 sites, 2021–22)

Geo-Aware: Station-specific pipelines

Model Performance

Multi-model trials + feature tuning

NO₂: 85–91% | **O₃:** 59–69%

Validated on unseen data (no leakage)

Next Steps

Cloud-gapping: INSAT WV/TIR + CNN/GNN

Virtual Sensors: Predict non-station zones

All requirements met

Tech stack:



Python 3.11



Pandas 2.1.1



NumPy 1.26.4



Scikit-learn 1.4.2



Matplotlib 3.8.2



Plotly 5.22.0



Seaborn 0.13.2



Tensorflow 2.16.1



XGBoost 2.0.3



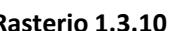
LightGBM 4.3.0



PyTorch 2.3.1



GeoPandas 0.14.3



Rasterio 1.3.10



Power BI



NetCDF4 1.6.3



cdsapi 0.7.6



xarray 2025.10



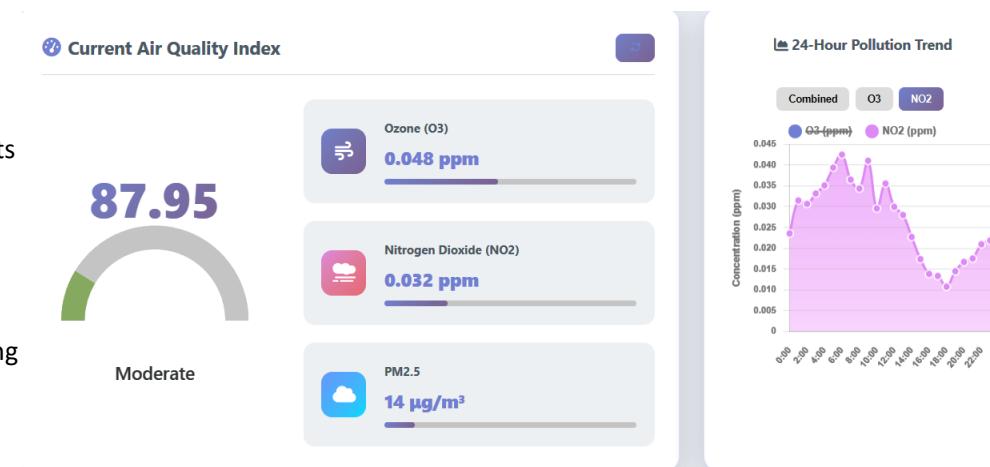
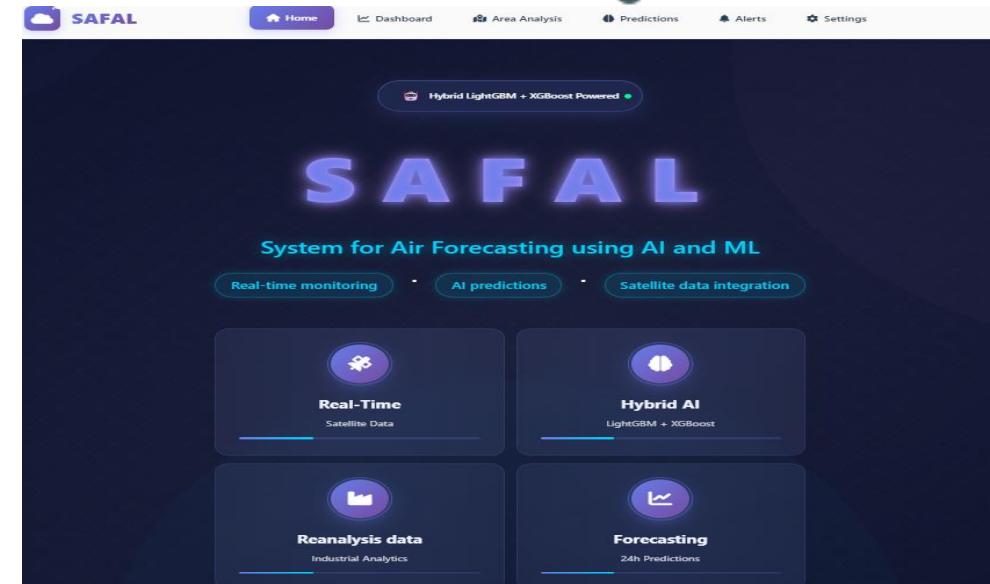
PyYAML 6.0



Git 2.40+



VS code | Jupyter



Layers of Prototype:

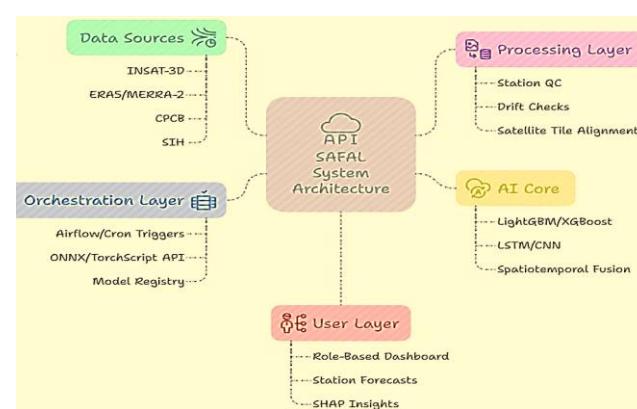
Data Sources: INSAT-3D + ERA5/MERRA-2 + CPCB → links sky measurements with ground NO₂/O₃.

Processing: Auto Quality Control, anomaly/drift checks, satellite-station alignment; hourly model-ready streams.

AI Core: LightGBM/XGBoost + LSTM-CNN hybrid; captures spatial-temporal patterns.

Automation: Airflow pipelines for ETL → train → deploy, with version tracking & alerts.

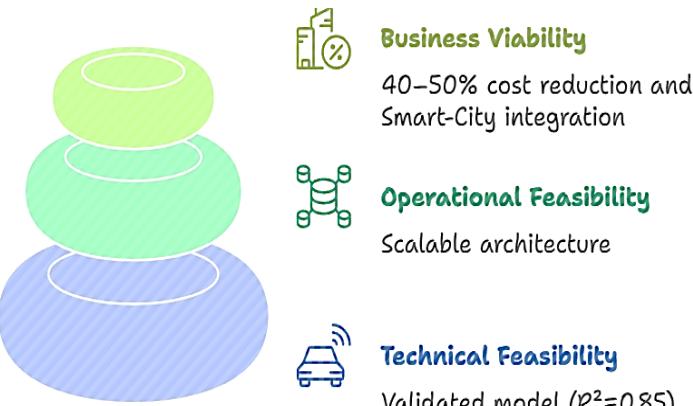
User Layer: Role-based dashboard + APIs → 24-48 hr forecasts with SHAP explanations for regulators.



FEASIBILITY AND VIABILITY



Analysis of the feasibility of the idea:



Technical Feasibility:

85%+ Prediction Accuracy: Hybrid LSTM-CNN achieving superior performance vs SAFAR's 70% with real-time multisource fusion (satellite + ground + reanalysis data)

Operational Feasibility & Scalability

10 × faster rollout — 2–4 weeks / city vs SAFAR's 6–12 months, 50% lower infra costs, and scalable to 100+ cities via cloud-native, modular pipelines. Potential extend further for 10-15 years.

GNN | Multimodels | fine-tunable models — With data for over decade, we can create a dataset using which we can use for many purposes.

Business Viability

₹387M market — targeting ₹50–75 Cr from 90+ cities via scalable, low-cost cloud pipelines, free satellite/APIs, and reusable ML models with 70–80% margins.

Challenges & Solutions:

Cloud Gaps → INSAT WV/TIR Interpolation

Dense clouds create missing satellite pixels → gap-fill using **INSAT-3D WV/TIR + ERA5/WRF surrogates**.

Sparse Ground Data → Transfer Learning

Few/no CPCB stations limit local calibration → transfer learning from data-rich cities + virtual sensors for stabilization.

API Delays → Caching & Auto-Retries

API latency disrupts ingestion → layered caching, TTL, exponential backoff, and fallback to recent forecasts.

Seasonal Drift → Monthly Retraining

Seasonal shifts lower accuracy → monthly retraining with seasonal indices, regional calibration, and drift-triggered updates.

Quality Control → Station Drift Checks

Sensor drift introduces bias → automated Quality Analysis & Control, anomaly screening, dual raw/validated streams per CPCB protocols.

Latency & Scale → Real-time Processing

Handling <5-min updates for multiple cities increases compute load → solved using stream-first ETL, micro-batching, Airflow/Kubernetes pipelines, and ONNX-accelerated model inference.

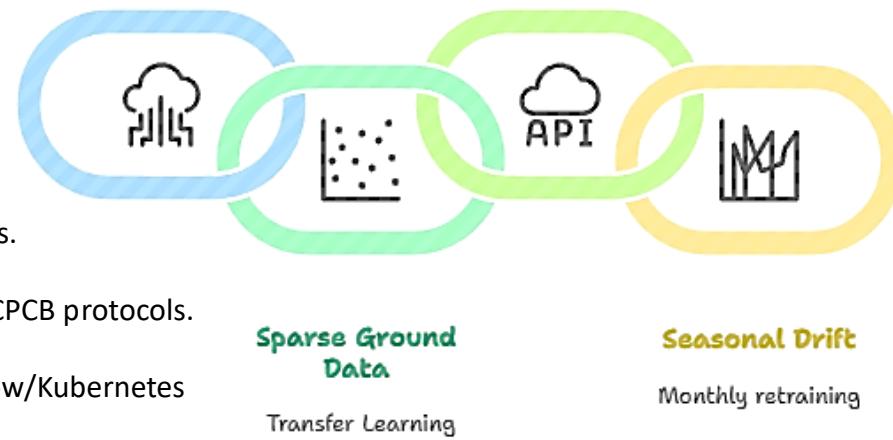
Characteristic	Data Source	Scalability	Infrastructure	Integration
Technical	Public Datasets	Modular Architecture	Prototype Validation	Compatible APIs
Operational & Implementation	Public Datasets	Automated Data Syncs	Mid-tier Cloud Instances	End-to-End Pipeline

Cloud Gaps

INSAT WV/TIR interpolation

API Delays

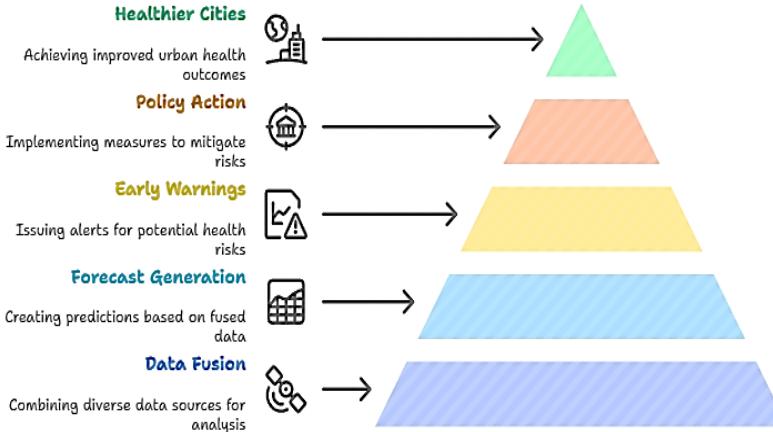
Caching & Auto-retries



IMPACT AND BENEFITS



SAFAL Impact Chain



Citizens & Communities

- **Health:** 24-hr alerts → ~30% risk reduction
- **Empowerment:** 1 km forecasts guide daily choices
- **Participation:** Community feedback & data contributions improve forecasts
- **Safety:** Alerts for vulnerable groups

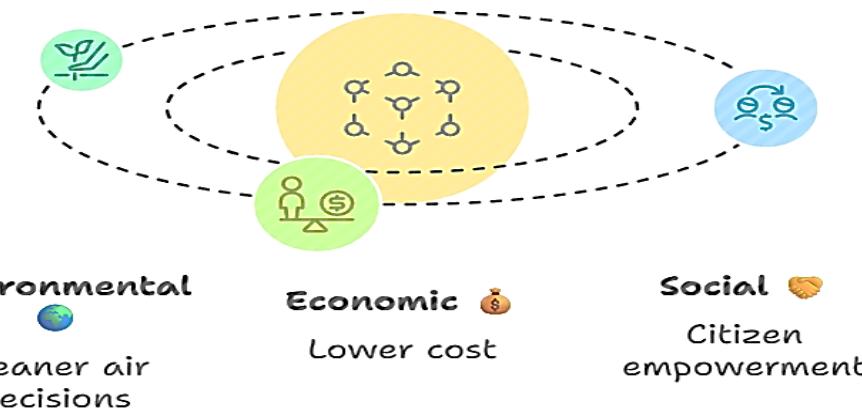
Researchers & Academia

- **Research:** Multisource datasets for studies
- **Publications:** 85%+ R², robust results
- **Collaboration:** APIs for cross-city & international research
- **Innovation:** Supports new AI/ML models

Government & Policymakers

- **Decisions:** Real-time hotspot mapping
- **Compliance:** Automated alerts → ₹10-50 L savings
- **Policy:** Quantifies Clean Air Program impact
- **Planning:** Enables evidence-based urban/environment planning

Benefits of SAFAL



Multi-Dimensional Benefits & Impact

Social – 1,200–1,400 lives saved/city; ₹2–5 cr annual health savings; alerts for vulnerable groups

Economic – Enables ₹10–50 L annual savings per industry via proactive alerts; reduces pollution-linked healthcare & crop losses.

Environmental – NO₂/O₃ ↓10–15%; targeted pollution interventions; supports net-zero goals

Quantified Impact

• Year 1: 400M+ citizens, 23,000+ lives saved, ₹3–16 B prevented losses

• 5 Years: 100+ publications, 20–30% pollution reduction, export to 20+ countries

Strategic Alignment – Sustainable Development Goals, 3/11/13; ₹4,400 Cr Clean Air Programme & 100 Smart Cities; India as AI-driven environmental leader

RESEARCH AND REFERENCES



Core datasets & Protocols:

- CPCB CAAQMS real-time data portal. [CPCB Data link](#)
- Copernicus ERA5 hourly reanalysis. [ERA5 Data link](#)
- Dataset given by Department of Space for SIH . [SIH](#)
- Data Access, GES DISC Merra 2 dataset. [Merra2 data access](#)
- INSAT-3D data [MOSDAC](#)

Landmark Publications:

- Singh et al., "Transforming Air Pollution Management in India with AI and Multisource Data," Nature Sci Rep 2024. [direct link to Article](#)
- Kumar et al., "Advanced Air Quality Prediction Using Multimodal Data and Deep Learning," Nature Sci Rep 2025. [direct link to article](#)

Algorithms & Tools:

- Apache Airflow Mops orchestration. [direct link](#)
- ONNX Runtime high-performance inference. [direct link](#)

Standards & Guidelines:

- CPCB Quality analysis/Quality Control protocols [direct link](#)
- WHO Air Quality Guidelines [direct link](#)

Prototype:

- Deployed prototype (Model Results only) [direct link](#)
- YouTube Video [direct link](#)

SAFAR vs SAFAL

Characteristic	Old SAFAR ✗	Next-Gen SAFAL ✓
Forecast Scope	Daily, coarse resolution	24/48-hour hourly, station level
Data Inputs	Limited ground + model data	Multi-source fusion: INSAT-3D + ERAS/MERRA-2 + CPCB
Cloud Gap Handling	Missing satellite data on cloudy days	Cloud-gap infilling using INSAT WV/TIR signals
Forecast Quality	Static model, no uncertainty checks	ML-based ensemble with drift checks + retraining
Deployment	Manual updates, no version tracking	Automated versioned feature store + retrain pipeline
Output Transparency	No interpretability or error reporting	SHAP-based interpretability + confidence metrics
Scalability	Hard-coded for few cities	Modular, station-scoped design for pan-India rollout
User Impact	Coarse-grained public info only	Actionable, station-level insights for planners + citizens