AEROBRIDGE

Bridging Satellite Imagery & Pollution Measurements

*"Aerobridge" is part of a larger initiative — **Air Pollution Estimation based on Satellite Imagery (INSAT) & Air Quality Monitoring Station Data (CPCB)**.*
*This project is based on the **Bhartiya Antariksha Hackathon 2025** problem statement by ISRO:*

"Monitoring Air Pollution from Space using Satellite Observations, Ground-Based Measurements, Reanalysis Data, and AI/ML Techniques."

*When we started, we quickly realized the **true scale** of this challenge.*

- ***Pollution data (CPCB)** was messy, incomplete, and inconsistent.*
- ***Satellite imagery (INSAT)** was massive — over **100 TB** estimated for all-India coverage.*

*It became clear that a **full-scale, all-India implementation** would be a dream project — but unrealistic for a student team without a supercomputer.*
*So we **strategically scaled down** while keeping the **core technical challenges** intact: working with **large datasets** and **real-world messy data**.*

CPCB Data Collection

Our first hurdle: collecting the right CPCB datasets.

- *We eventually gathered data from **~450 stations across India**.*
- *Some stations had data from **2010–2023**, others only from **2018 onwards**.*
- *Many files were **missing values**, had **inconsistent formats**, and required **extensive cleaning**.*

For practical processing (and future scalability), we narrowed it down to:

- ***4 States**: Maharashtra, Delhi, Karnataka, Haryana.*

- ***5 Stations per state*** *(20 total).*

This allowed us to focus on ***quality over quantity*** *while still simulating a* ***multi-state integration challenge****.*

CPCB Final Dataset

We prepared the pollution dataset for ***integration with INSAT imagery*** *by:*

- *Cleaning and standardizing timestamps.*
- *Removing invalid/missing PM2.5 readings.*
- *Filtering to a consistent date/time window.*

Final CPCB dataset specs:

- *Period:* ***1 Sept – 30 Nov*** *(91 days) for both* ***2021*** *and* ***2022****.*
- *Time range:* ***08:00 AM – 04:00 PM IST****.*
- *Total:* ***728 hours per station × 20 stations = 14,560 hours*** *of pollution data.*

INSAT Imagery Collection

As mentioned earlier, our initial INSAT dataset estimate exceeded ***100 TB*** *for all-India, all six bands. This was* ***unmanageable*** *for our resources.*
We scaled down intelligently:

- *Selected only* ***2 bands****:* ***TIR1*** *(Thermal Infrared) and* ***WV*** *(Water Vapour) — most relevant for atmospheric analysis.*
- *Ordered* ***state-specific imagery*** *instead of national coverage.*
- *Targeted the* ***exact same period and time window*** *as CPCB: 1 Sept – 30 Nov, 08:00–16:00 IST.*

We sourced the imagery from ***ISRO's MOSDAC platform****, placing* ***hundreds of orders*** *(including test runs, failed requests, and successful downloads).*

INSAT data summary:

- ***2021:*** *3,580 images (expected ~2,912, but ISRO provided extras).*
- ***2022:*** *3,230 images (again, more than expected).*
- *Images were half-hourly, and each timestamp had* ***two bands****, doubling counts.*

Pipeline Development

With both CPCB and INSAT data ready, we built a ***two-stage pipeline****:*

1. ***Image Cropping & Merging***
    - *Takes raw INSAT .tif images.*
    - *Crops each image to a small bounding box around each station using lat/lon.*
    - *Matches cropped images to CPCB readings for the nearest timestamp.*
    - *Produces a* ***merged CPCB + INSAT dataset****.*
2. ***Feature Extraction & Final Dataset Creation***
    - *Calculates descriptive statistics from each cropped image (mean, std, min, max, median, percentiles, threshold counts, skewness, etc.).*
    - *Merges these features with CPCB pollution readings.*
    - *Produces the* ***final station-specific dataset*** *— location-aware, time-aligned, and ML-ready.*

The ***end result****: a high-quality dataset where* ***every row*** *represents a specific hour, with* ***ground-based PM2.5 data*** *from CPCB and* ***satellite-derived features*** *from INSAT imagery — the perfect foundation for machine learning models.*

After the first time both pipelines successfully produced our *final_dataset*, we decided to level up. Originally, we had processed data for only a 6-month window, but with the proof of concept working, I scaled the pipelines to handle **24 full months** — the entire years 2021 and 2022.

This meant ordering new INSAT imagery for the full period (32 large data orders in total) and adapting the pipelines to manage the increased volume. The full run wasn't trivial — the pipeline had to be executed three times due to errors before a clean output was achieved. The final successful run took **5 hours, 26 minutes, and 57 seconds** to complete, during which it:

- Processed **20 CPCB stations**
- Cropped **1,039,390 INSAT images**
- Merged **115,660 records**
- Cached **228,973 extracted features**

The result: a **rich 2-year combined dataset** ready for advanced modelling.

Alongside the pipelines, I also developed a lightweight internal application to monitor and manage the entire processing workflow. This app allowed us to:
• Track pipeline stages in real time
• Verify intermediate outputs before committing to long runs
• Quickly visualize sample integrations of CPCB readings with INSAT imagery

While not designed for public release, the app became an invaluable tool during development. It reduced the guesswork when handling multi-gigabyte datasets, cut down on repetitive manual checks, and made debugging far less painful — especially when processing hundreds of thousands of satellite images in a single run.

Future Work & Roadmap

While *Aerobridge* has achieved a significant milestone — integrating **CPCB ground data** with **INSAT satellite imagery** and building a fully automated pipeline — this is only the beginning.

In the coming phases, we plan to:

- **Integrate Atmospheric Reanalysis Data**
    - Incorporate meteorological variables from datasets such as **MERRA-2**, including wind speed, temperature, humidity, and pressure.
    - This will help improve the accuracy of pollution predictions by capturing more environmental context.
- **Advanced Machine Learning Models**
    - Begin experiments with **LightGBM** and **XGBoost** to predict **PM2.5 concentrations** using only satellite-derived features (e.g., Aerosol Optical Depth).
    - Explore deep learning architectures such as **Convolutional Neural Networks (CNNs)** for spatial feature extraction at scale.
- **End-to-End Prediction System**
    - Transition from dataset preparation to building a system that can ingest real-time satellite data and output PM2.5 predictions for specific locations without needing ground measurements.

We understand that this project is far from "finished." As **B.Sc. Data Science students and freshers**, our journey with *Aerobridge* has just begun — but even this first phase has been a remarkable achievement.

Reflections & Lessons Learned

This was one of the **largest and most complex projects** I have undertaken so far.
It took **over a month** to bring it to its current form, and every stage came with lessons that can't be learned in a classroom:

- **The Harsh Reality of Real-World Data**
    - Gigabytes of messy formats, inconsistent timestamps, and null values that didn't just "clean themselves" with one line of Pandas.

    - Learning how to make **strategic compromises** (e.g., scaling down from 100 TB) without losing the essence of the challenge.

- **Pipeline Engineering (…as a Data Science student!)**
    - Building a pipeline of **3,000+ lines of code** — despite not being a "software developer."
    - Testing and debugging repeatedly. Sometimes it failed in 90 minutes, meaning every error cost serious time.
    - Encountering errors from the **basic** (yes, even indentation errors) to the **truly obscure**.
- **Rediscovering the Joy of the Craft**
    - This project reminded me *why* I love data science — the problem-solving, the creativity, and the thrill of making data come alive.
    - I fell back in love with Python, and I'm still learning new tricks every day.

From the outside, *Aerobridge* might look like a "simple integration project" — but behind the scenes, it's a story of **persistence, trial-and-error, and genuine curiosity**. And the best part? We're not done yet.

Team AEROBRIDGE — *Lalit K Hire, Amit S Joshi, Nikita S Deshmukh, Siddhi S Gawade*

Contributions

- **Lalit K Hire** — Developer and data scientist; implemented the core pipeline, INSAT data processing, CPCB integration, feature extraction, and dataset creation. Working on Machine Learning, Deep Learning & CNN.
- **Amit S Joshi** — Developer and data analyst; Assisted with CPCB dataset processing and quality assurance. Develop automated system for auto processing & Cleaning of CPCB. Working on Reanalysis of Dataset, Working on Machine Learning, Deep Learning & CNN.
- **Nikita S Deshmukh -** contributed to planning discussions, Lead Initial Research, worked on INSAT & CPCB Data collection.
- **Siddhi S Gawade** — contributed to planning discussions, work on Researching other Satellite datasets, i.e. MODIS, S5P

Acknowledgements
This project was made possible through the combined efforts of our team. each member contributed in their own way — from processing initial datasets to participating in discussions that shaped the project's scope and direction. Every contribution, big or small, played a role in helping *Aerobridge* become what it is today.

1. Data Collection & Preparation

Raw CPCB Data:

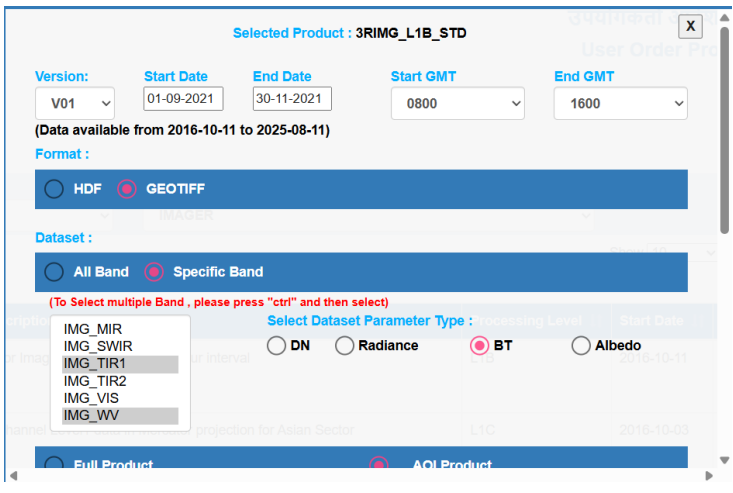
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	From Date	To Date	PM2.5 (ug/	PM10 (ug/	NO (ug/m3	NO2 (ug/m	NOx (ppb)	NH3 (ug/m	SO2 (ug/m	CO (mg/m	Ozone (ug/	Benzene (u	Toluene (u	Temp (deg	RH (%)	WS (m/s)	WD (deg)	SR (W/mt2	BP (mmHg	VWS (m/s)	Xylene (ug/
2	01-07-2016 10:00	01-07-2016 11:00	10.67	39	17.67	39.2	32.33	7.07	6.6	0.48	14.5	1	4.63	33.43	71.67	2.3	226.33	123.67		-0.1	0.1
3	01-07-2016 11:00	01-07-2016 12:00	2	39	20.5	41.9	35.8	7.4		0.49	15	0.7	4.5	33.7	70	2.5	223	186		-0.1	0.1
4	01-07-2016 12:00	01-07-2016 13:00																			
5	01-07-2016 13:00	01-07-2016 14:00																			
6	01-07-2016 14:00	01-07-2016 15:00	20.5	50	15.4	43.6	32.78	6.35	6.38	0.47	10.5	0.6	4.5	33.57	63.5	1.88	223	240.5		-0.1	0.1
7	01-07-2016 15:00	01-07-2016 16:00	15.25	59.5	24.3	45.12	40.12	6.65	6.53	0.51	6.6	0.77	5.33	33.55	63.75	1.4	213.25	152.25		-0.1	0.1
8	01-07-2016 16:00	01-07-2016 17:00	11.67	60	26.73	49.1	43.8	6.93	5.7	0.46	17.43	0.8	6.47	33.57	64	1.5	208.67	92.67		-0.1	0.23
9	01-07-2016 17:00	01-07-2016 18:00	11.75	57.5	19.1	46.33	36.82	8.12	6.23	0.44	19.98	0.7	6.4	33.48	64	1.95	222.75	35		-0.1	0.1
10	01-07-2016 18:00	01-07-2016 19:00	18	57.75	14.5	44.2	32.33	7.78	6.07	0.45	12.2	0.7	5.3	33.23	63.75	2.15	220.75	14		-0.1	0.1
11	01-07-2016 19:00	01-07-2016 20:00	12	63	10.43	41	27.83	8.13	6.2	0.41	17.5	0.7	5.03	33.27	64	2.4	220.67	7		-0.1	0.1
12	01-07-2016 20:00	01-07-2016 21:00	14.5	49.75	4.97	35.03	20.8	8.47	6.52	0.35	13.82	0.4	4.2	33.3	64.25	2.6	225.5	7		-0.1	0.1
13	01-07-2016 21:00	01-07-2016 22:00	12.5	45	5.45	20.23	13.93	9.1	5.65	0.36	12	0.25	2.88	34.77	64.75	2.68	224.5	7		-0.1	0.1
14	01-07-2016 22:00	01-07-2016 23:00	11	51	9.23	26.5	19.8	9.13	5.27	0.31	15.07	0.8	0.67	36.87	68	1.8	225.67	7		-0.1	0.1
15	01-07-2016 23:00	02-07-2016 00:00	7	38	5.05	29.9	18.35	7.8	8.15	0.26	13.4	0.7	2.5	33.55	70.5	2.05	237.5	7		-0.1	0.1
16	02-07-2016 00:00	02-07-2016 01:00																			
17	02-07-2016 01:00	02-07-2016 02:00																			
18	02-07-2016 02:00	02-07-2016 03:00																			
19	02-07-2016 03:00	02-07-2016 04:00																			

Cleaned CPCB Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	From Date	PM2.5 (ug/	PM10 (ug/	NO (ug/m3	NO2 (ug/m	NOx (ppb)	NH3 (ug/m	SO2 (ug/m	CO (mg/m	Ozone (ug/	Temp (deg	RH (%)	WS (m/s)	WD (deg)
2	01-01-2021 08:00	311.95	439.18	113.22	83.4	196.56	27.77504	12.56297	2.59	16.27	27.45444	65.51356	1.54102	180.1127
3	01-01-2021 09:00	441.33	577.21	149.8	164.26	314.03	27.77504	12.56297	2.85	22.07	27.45444	65.51356	1.54102	180.1127
4	01-01-2021 10:00	628.92	938.73	97.33	281.43	378.63	27.77504	12.56297	2.82	31.85	27.45444	65.51356	1.54102	180.1127
5	01-01-2021 11:00	505.02	719.91	24.47	170.46	194.9	27.77504	12.56297	3.26	43.1	27.45444	65.51356	1.54102	180.1127
6	01-01-2021 12:00	468.68	582.5	14.03	124.11	138.04	27.77504	12.56297	3.05	61.48	27.45444	65.51356	1.54102	180.1127
7	01-01-2021 13:00	478.13	601.08	14.34	140.25	154.64	27.77504	12.56297	2.9	58.35	27.45444	65.51356	1.54102	180.1127
8	01-01-2021 14:00	441.76	519.24	10.74	91.34	102.08	27.77504	12.56297	2.76	57.85	27.45444	65.51356	1.54102	180.1127
9	01-01-2021 15:00	261.44	329.08	8.82	79.61	88.43	27.77504	12.56297	2.5	62.27	27.45444	65.51356	1.54102	180.1127
10	01-01-2021 16:00	243.64	331.45	8.44	73.69	82.13	27.77504	12.56297	2.28	52.4	27.45444	65.51356	1.54102	180.1127
11	02-01-2021 08:00	268.29	420.25	22.87	80.97	103.86	27.77504	12.56297	2.02	16.96	27.45444	65.51356	1.54102	180.1127
12	02-01-2021 09:00	258.82	346.28	24.69	77.37	102.04	27.77504	12.56297	1.92	12.75	27.45444	65.51356	1.54102	180.1127
13	02-01-2021 10:00	220.44	323.44	28.81	91.96	120.8	27.77504	12.56297	1.87	13.31	27.45444	65.51356	1.54102	180.1127

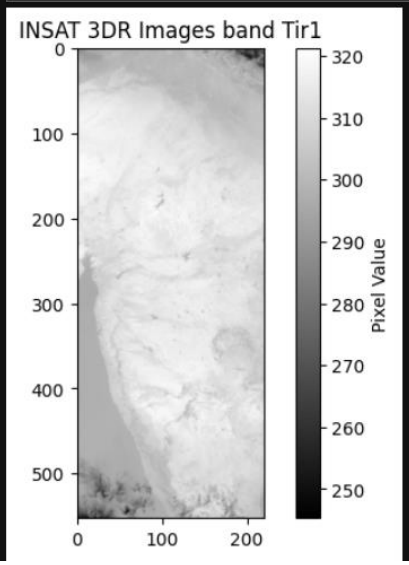
2. Satellite Imagery Acquisition

MOSDAC data ordering:

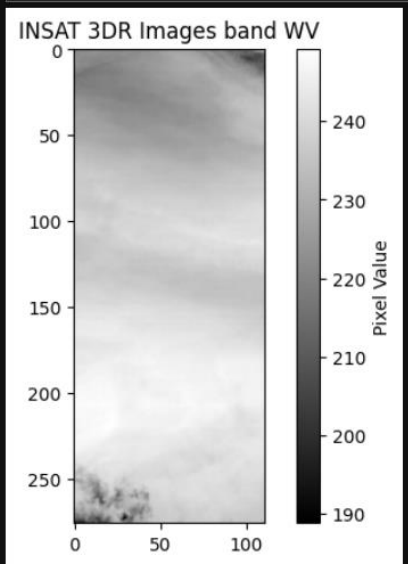


Sample of Satellite Images (Raw)

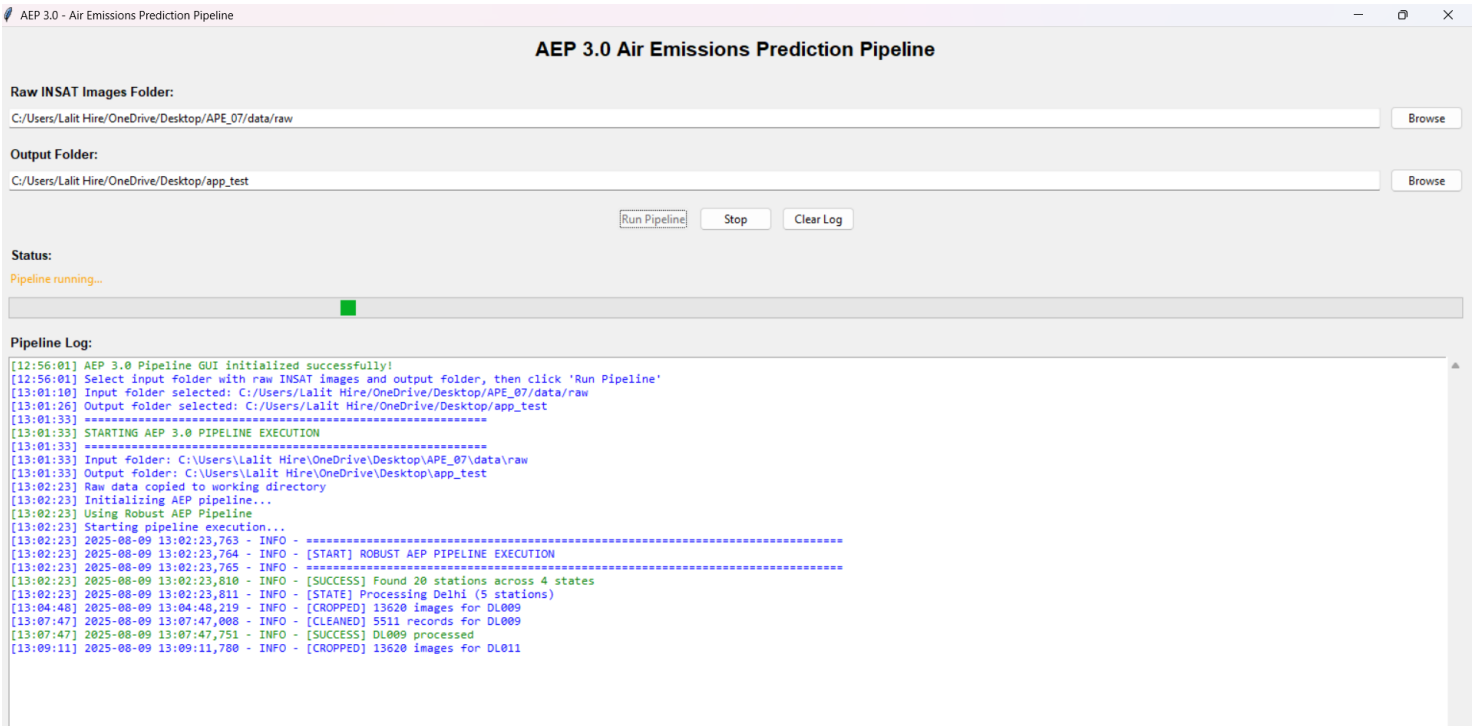
[6]:
```
import rasterio
import matplotlib.pyplot as plt

# === Load the GeoTIFF ===
file_path = '3RIMG_01APR2021_0815_L1B_STD_V01R00_IMG_TIR1.tif'

with rasterio.open(file_path) as src:
    image = src.read(1)
    plt.imshow(image, cmap='gray')
    plt.title("INSAT 3DR Images band Tir1")
    plt.colorbar(label="Pixel Value")
    plt.show()
```



[8]:
```
import rasterio
import matplotlib.pyplot as plt

# === Load the GeoTIFF ===
file_path = '3RIMG_01APR2021_0815_L1B_STD_V01R00_IMG_WV.tif'

with rasterio.open(file_path) as src:
    image = src.read(1)
    plt.imshow(image, cmap='gray')
    plt.title("INSAT 3DR Images band WV")
    plt.colorbar(label="Pixel Value")
    plt.show()
```



3. Pipeline in Action



4. Feature Extraction & Final Dataset

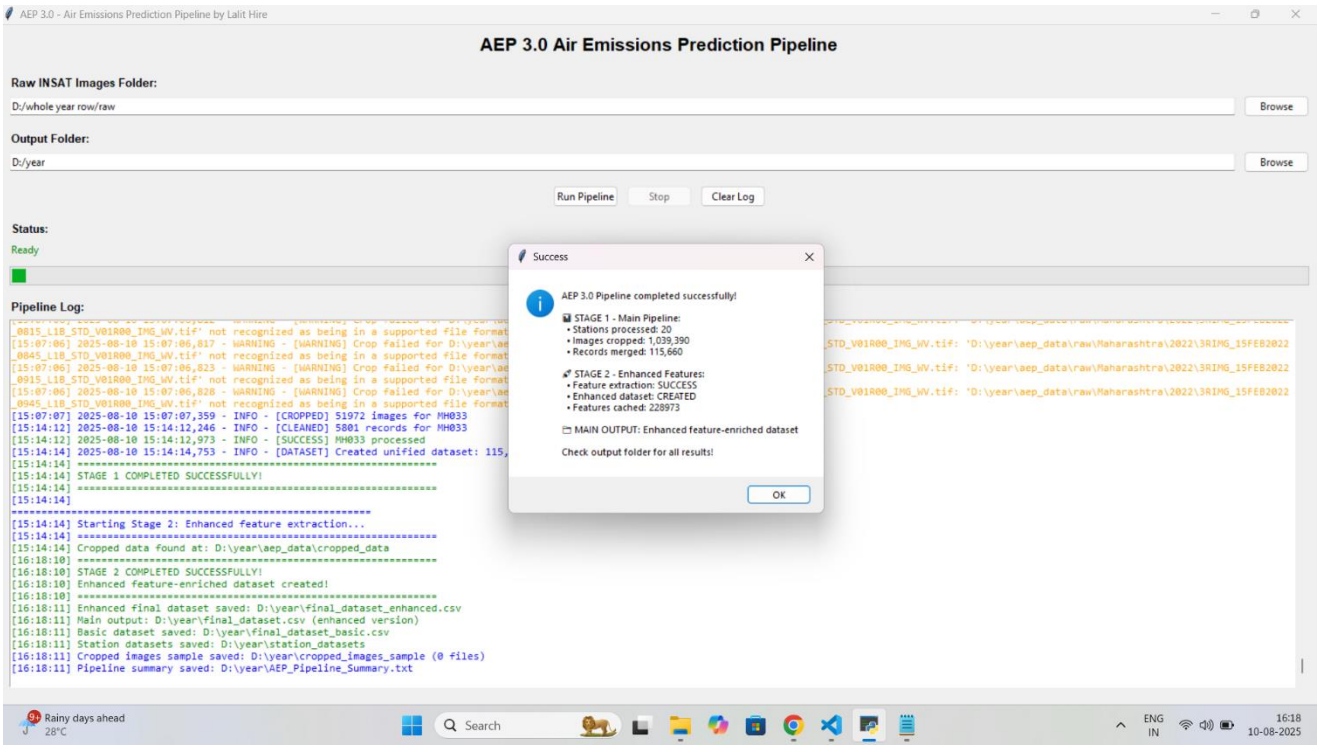
First merged dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	state	station_id	station_lo	latitude	longitude	timestamp	PM2.5	img_mean	img_std	img_min	img_max	img_median	
2	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	311.95	303.5073	0.805942	300.5585	304.4886	303.6605	
3	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	441.33	303.5073	0.805942	300.5585	304.4886	303.6605	
4	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	628.92	303.5073	0.805942	300.5585	304.4886	303.6605	
5	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	505.02	303.5073	0.805942	300.5585	304.4886	303.6605	
6	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	468.68	303.5073	0.805942	300.5585	304.4886	303.6605	
7	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	478.13	303.5073	0.805942	300.5585	304.4886	303.6605	

Second dataset with Extracted Features:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	state	station_id	station_lo	latitude	longitude	timestamp	PM2.5	img_mean	img_std	img_min	img_max	img_media	mean_TIR	std_TIR1	min_TIR1	max_TIR1	median_TI	p25_TIR1	p75_TIR1	p90_TIR1	p95_TIR1	pct_above	pct_below	pc
2	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	311.95	303.5073	0.805942	300.5585	304.4886	303.6605	290.3413	0.687775	288.9057	291.6345	290.4578	289.9089	290.8988	291.0214	291.1357	100	0	
3	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	441.33	303.5073	0.805942	300.5585	304.4886	303.6605	290.0905	0.52684	288.9982	290.8399	290.0976	289.8001	290.5652	290.6891	290.7129	100	0	
4	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	628.92	303.5073	0.805942	300.5585	304.4886	303.6605	288.587	0.607873	287.6714	289.6136	288.704	288.0165	288.9811	289.4156	289.5634	100	0	
5	Delhi	DL009	Pusa, Delh	28.6374	77.1577	2021-01-0	505.02	303.5073	0.805942	300.5585	304.4886	303.6605	286.8068	0.399559	286.03	287.6473	286.7565	286.5202	287.0516	287.3809	287.4858	100	0	

5. Scale-Up & Results



Closing Note

These screenshots show just a glimpse of Aerobridge in action—from messy CPCB files to INSAT imagery and the pipelines that connect them.

What began as a hackathon idea became a deep dive into large-scale data engineering and satellite analytics. Aerobridge is far from finished, but the foundation we've built proves that even freshers can tackle problems at scale with enough curiosity and persistence. Thank You!