Varun G | Kunal R | Advait S

Flood Prediction

Problem Statement

1. **What Problem Are We Solving?**
   - India faces devastating floods every year:
   - Average annual loss: ₹25,000 crore (US $3B+)
   - Lives lost: Over 1,600 people annually
   - Millions displaced and livelihoods destroyed

Current forecasting systems fall short:

- Rely on either satellite or ground data-not both
- Provide only “yes/no” alerts, not risk probabilities
- Short lead times, poor spatial detail, and limited adaptability

"BEHIND EVERY STATISTIC IS A SHATTERED HOME OR LOST FUTURE"

2. Why Does This Matter?

- Floods are increasing in frequency and intensity due to climate change and rapid urbanization
- Vulnerable communities suffer the most-trauma, lost education, and cycles of poverty

Tripura state declared natural calamity-affected area amid flood devastation

Tripura government has declared the entire state as a natural calamity-affected area following severe floods. The floods have resulted in significant loss of lives and extensive damage to properties.

10 Dead, No Connectivity, Flash Floods In Meghalaya's Garo Hills Cause Devastation

Home / India News / 19 dead, 17,000 evacuated as Andhra Pradesh, Telangana face flood fury

19 dead, 17,000 evacuated as Andhra Pradesh, Telangana face flood furv

Bihar under water: Why the state sees floods every year

Bihar is among India's most flood-prone states. The reasons for this are the state's unique geography, and a decades-old solution that has proved short-sighted. We explain.

3. What Will Our Solution Enable?

- Fuses 7-day Sentinel-2 satellite image sequences with ground data for richer, more reliable flood risk forecasts
- Predicts flood risk as probability intervals, not just binary alerts
- Scalable and adaptable across India's diverse flood types and regions

4. Potential Impact

- Saves lives through earlier, more precise warnings
- Reduces economic losses and protects critical infrastructure
- Empowers communities, planners, and agencies with actionable, uncertainty-aware insights

Literature Survey

S. NO	STUDY	METHOD / SOLUTION	DATA USED	KEY OUTCOME	GAP
1	Tiwari et al. (2020) – Kerala Floods	Sentinel-1 SAR + Otsu Thresholding	VV/VH time-series (GEE)	Acc > 94% for flood extent mapping	No use of optical/ground variables or learning model
2	Osman & Das (2024) – Damodar Basin	Naïve Bayes Tree + EBF/FR ensemble	15 static predictors + flood inventory	AUC = 0.85; RMSE < 0.4	Static risk maps only; no lead-time or rainfall input
3	Sushanth et al. (2023) – Jharkhand	LSTM + SHAP + conceptual reservoir block	Daily rainfall + inflow/outflow	NSE = 0.93; explainability via SHAP	No satellite input; doesn't generalize to ungauged basins
4	Shekar et al. (2024) – Madhya Pradesh	1D CNN + RNN sequence model	Monthly rainfall & streamflow	NSE = 0.97; accurate runoff learning	Coarse monthly data; lacks spatial flood extent
5	Barman et al. (2024) – Brahmaputra Basin	SWAT + NARX-ANN hybrid	GCMs + LULC CA-Markov + hydrology	R^2 = 0.74 vs 0.35 (SWAT baseline)	Complex setup; doesn't output flood probabilities

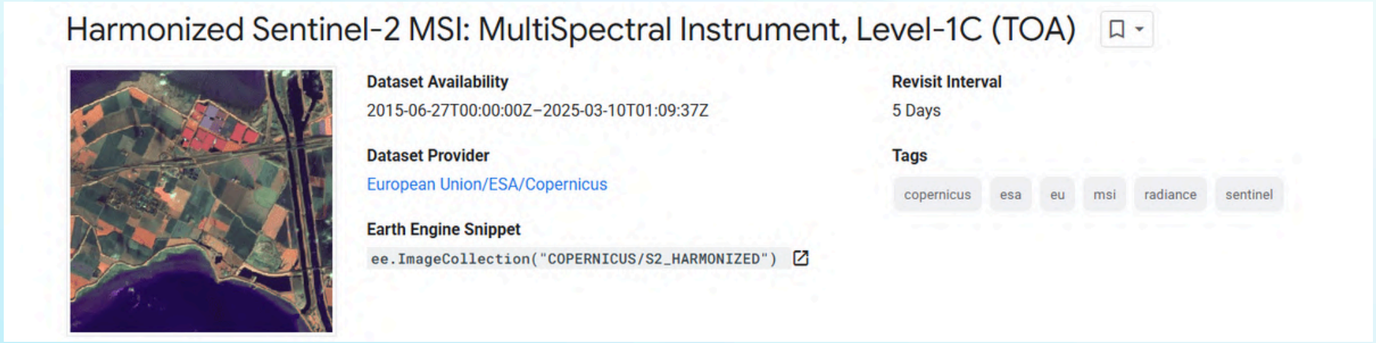
Literature Review (Summary)

Past research and industry efforts have largely focused on either satellite-based flood mapping or ground-data-driven forecasting using machine learning. While these approaches have shown promise in specific regions, they often rely heavily on dense sensor networks, offer deterministic outputs, and lack integration across data types.

Our Contribution

We extend this work by combining image features with tabular data using a class-wise sampling strategy, enabling richer learning from small datasets. Our use of quantile regression introduces uncertainty-aware forecasting, offering more reliable and actionable flood predictions.

Dataset 1 - Satellite Imagery



- Source: Sentinel-2 (COPERNICUS/S2_HARMONIZED) via Google Earth Engine
- Events Covered: 130 flood events across India from 2015 to present
- 7 day range leading up to events
- Image Type: Multispectral satellite images, including RGB, NIR, and SWIR bands which help differentiate better.
- No manual preprocessing required for these.

Dataset 2 - Ground Data

Dataset Overview

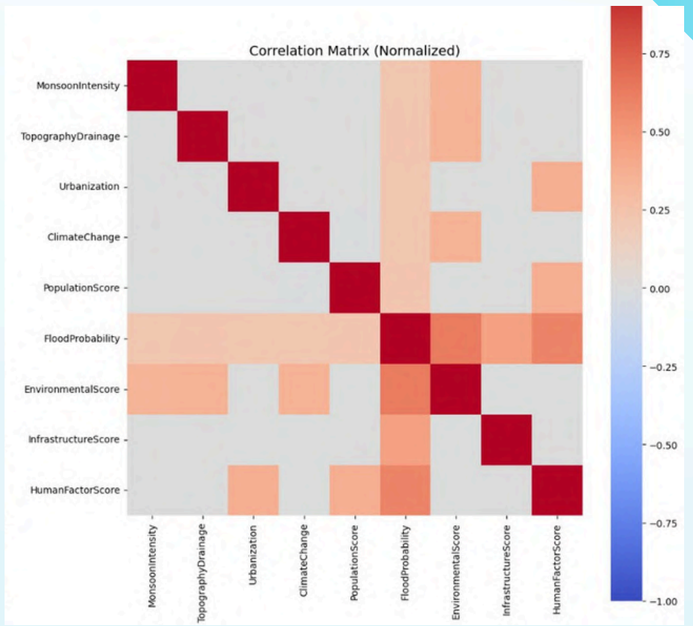
- 50,000 data points, 21 numerical features (no missing values)
- Usability: 9.41 on Kaggle
- Designed for flood probability prediction

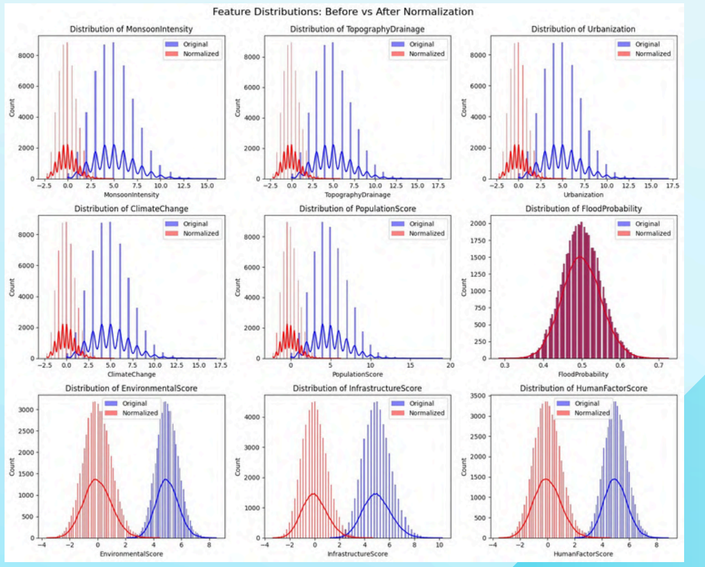
Key Features

- Environmental Factors – Monsoon Intensity, Climate Change, Deforestation, Coastal Vulnerability, Watersheds, Wetland Loss
- Human & Infrastructure Factors – Urbanization, River Management, Drainage Systems, Dams Quality, Population Density, Poor Planning
- Target Variable – Flood Probability (Likelihood of flooding)

Preprocessing

- No missing values in the dataset
- There are no categorical features, everything has been converted to numerical data
- Normalized the data
- All features were relevant so we did not drop any or reduce dimensions.

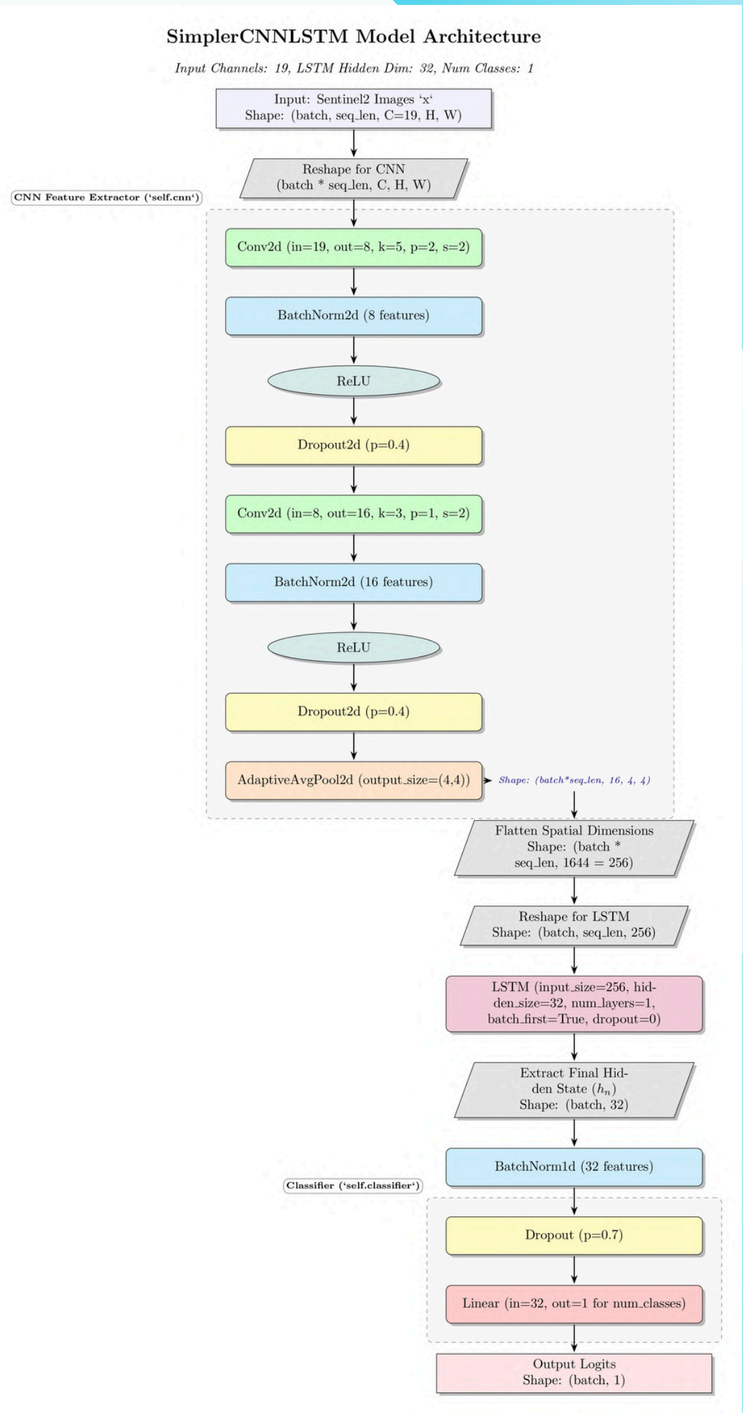




Methodology

SimplerCNN-LSTM Architecture from scratch

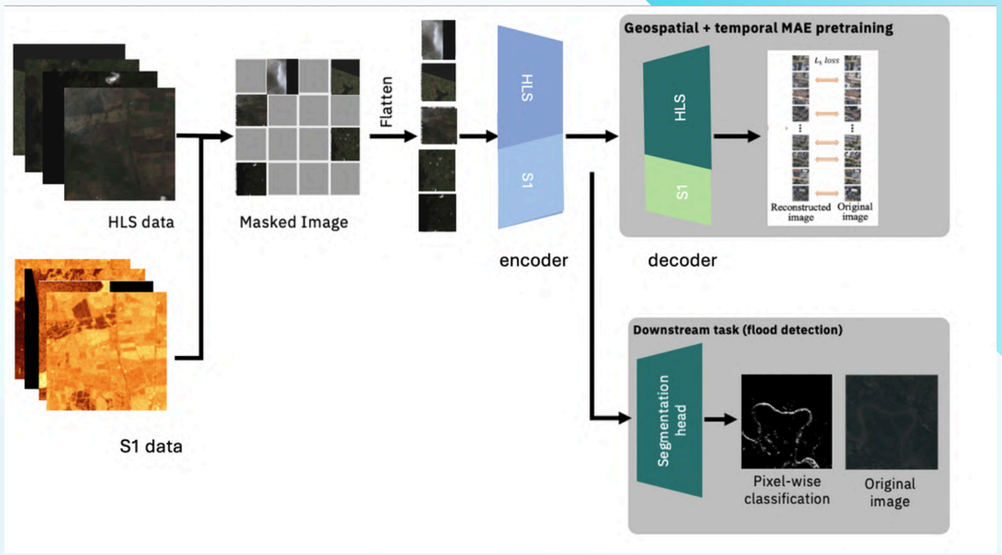
- Goal: Image sequence classification.
- Method: CNN for spatial features, LSTM for temporal sequence.



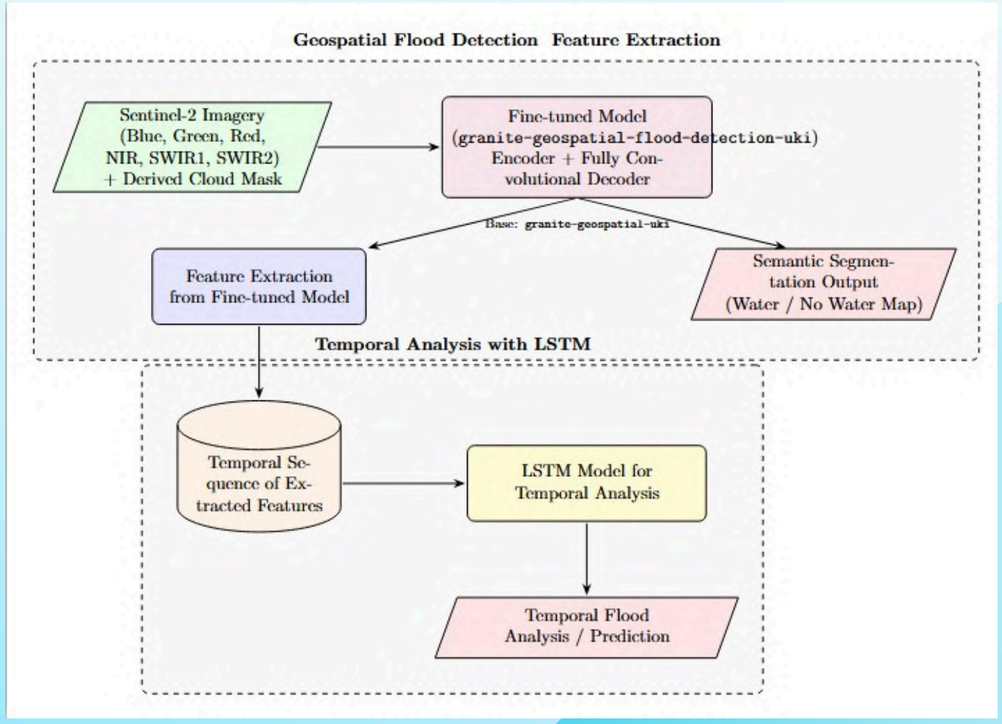
Methodology

Geospatial Flood Detection Architecture

- Overall Goal: Temporal flood analysis and prediction using satellite imagery.
- Two Main Stages:
  - Feature Extraction: Using a fine-tuned geospatial model.
  - Temporal Analysis: Using an LSTM on features extracted over time.



https://huggingface.co/ibm-granite/granite-geospatial-uki-flooddetection



Methodology

MLP Architecture for Flood Prediction

- Goal: Predict flood probability using input features.
- Type: Multi-Layer Perceptron (Feedforward Neural Network).

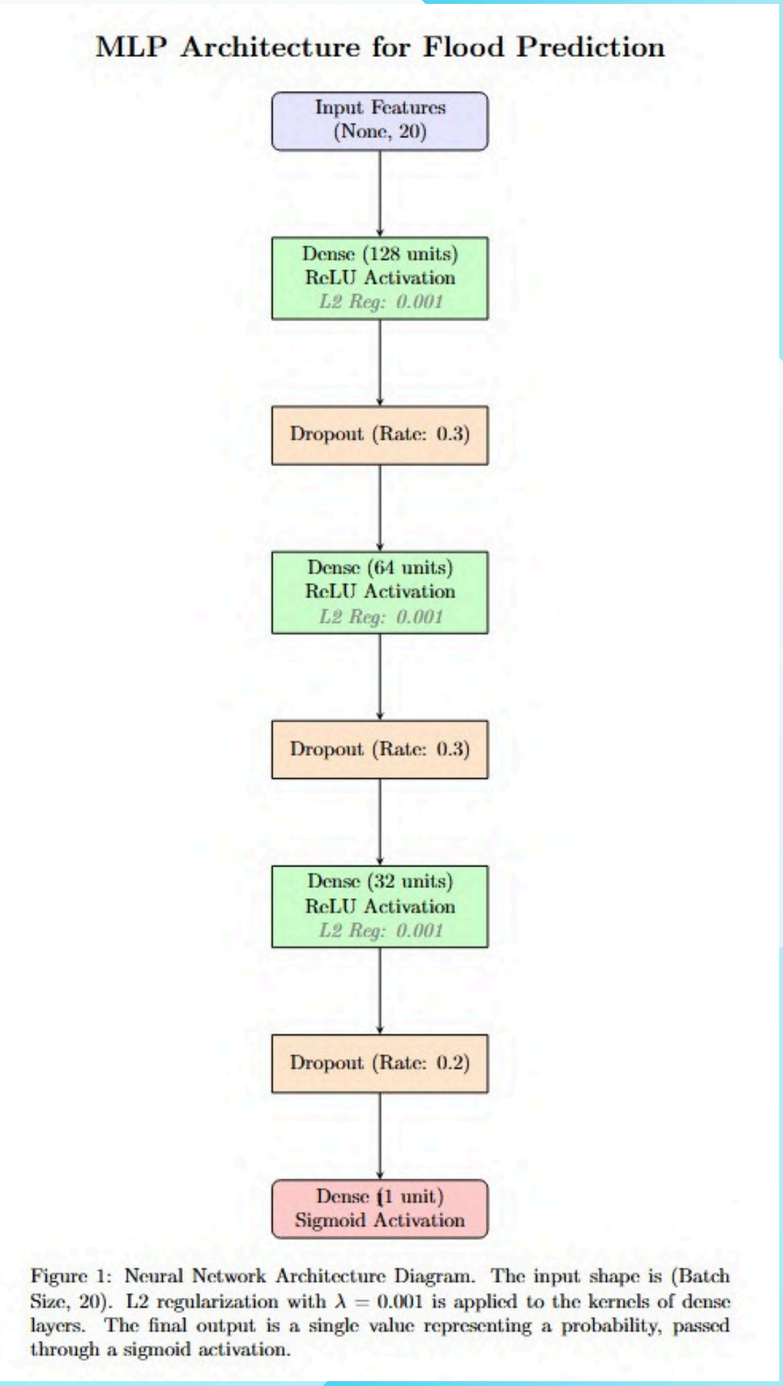


Figure 1: Neural Network Architecture Diagram. The input shape is (Batch Size, 20). L2 regularization with $\lambda = 0.001$ is applied to the kernels of dense layers. The final output is a single value representing a probability, passed through a sigmoid activation.

Methodology

Pipeline: Combined Feature Flood Prediction

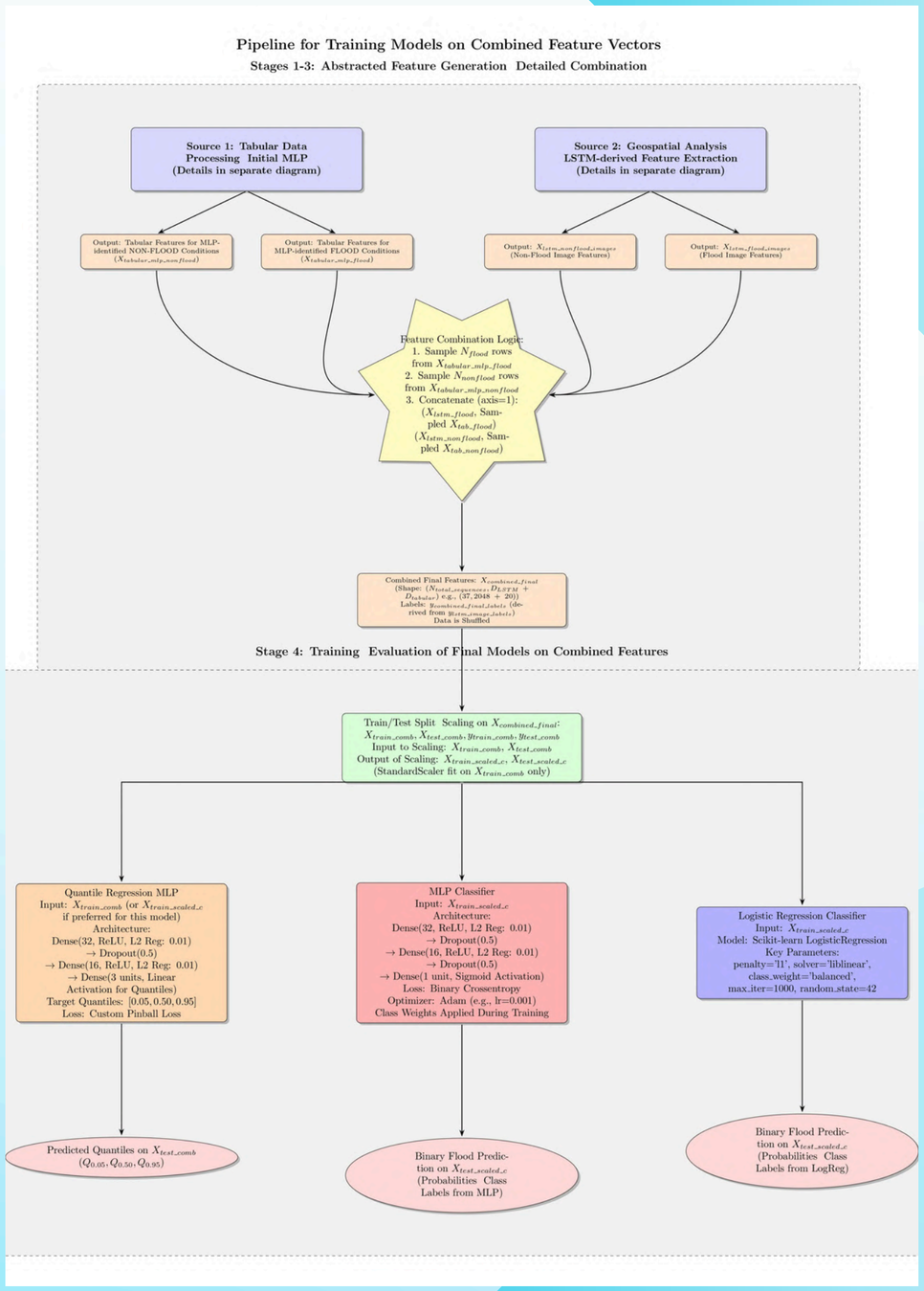
- Goal: Train models on fused Tabular & Geospatial (Image) data.

Stages 1-3: Feature Generation & Fusion

- Source 1 (Tabular): Processed tabular data -> Condition-based features (Flood/Non-Flood).
- Source 2 (Geospatial/Image): Image analysis (LSTM/CNN) -> Image features & True flood labels.
    - Fusion Logic:Input: Features & labels from Source 1 & 2.
    - Method: Pair image features with sampled condition-matched tabular features.
    - Output: X_combined_final (fused), y_combined_final_labels (true image labels).

Stage 4: Model Training & Evaluation (on X_combined_final)

- Preprocessing: Train/Test Split, Scaling.
    a. Models Trained: Quantile Regression MLP: Predicts flood probability quantiles (0.05, 0.50, 0.95).
    b. MLP Classifier: Binary flood prediction (Sigmoid output).
    c. Logistic Regression: Binary flood prediction (L1 penalty, balanced weights).
- Outputs: Model-specific flood predictions/probabilities.



Performance Metrics

Initial CNN+LSTM

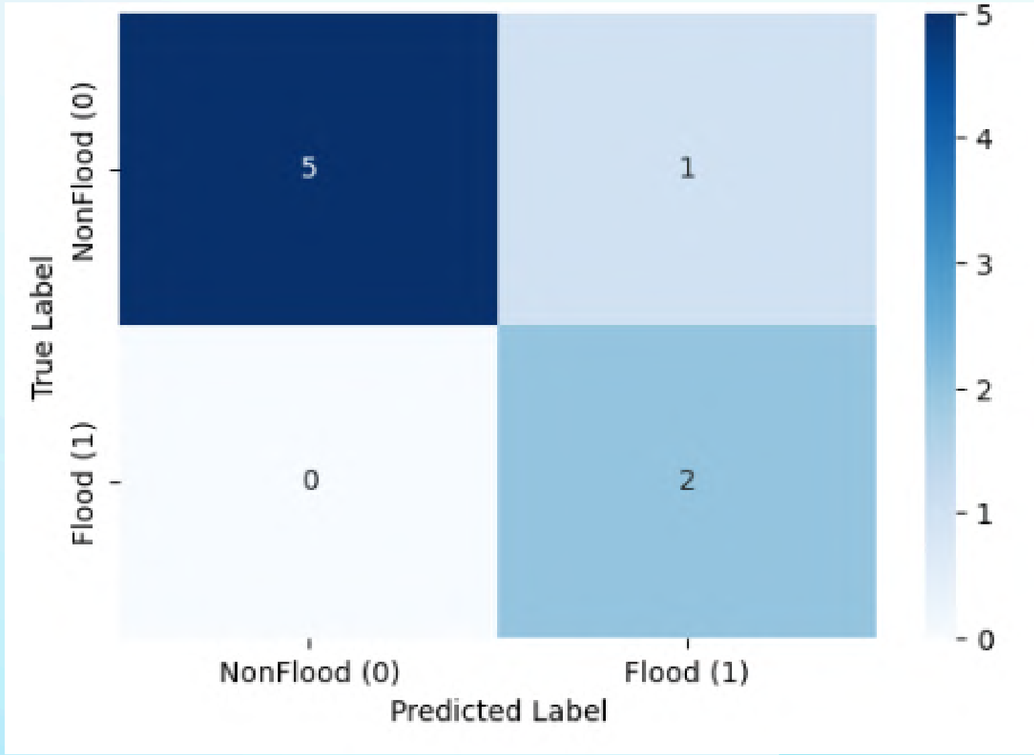
83.33% Accuracy

```
Final Test Accuracy with threshold 0.30: 83.33%
Test Precision: 0.6667, Recall: 1.0000, F1: 0.8000
Test Confusion Matrix:
[[3 1]
 [0 2]]
```

Performance Metrics

Pretrained Model*+LSTM

```
--- Validation Set Metrics ---
Accuracy:  0.8750
Precision: 0.6667
Recall:    1.0000
F1-score:  0.8000

Classification Report:
              precision    recall  f1-score   support

NonFlood (0)       1.00      0.83      0.91         6
   Flood (1)       0.67      1.00      0.80         2

    accuracy                           0.88         8
   macro avg       0.83      0.92      0.85         8
weighted avg       0.92      0.88      0.88         8
```

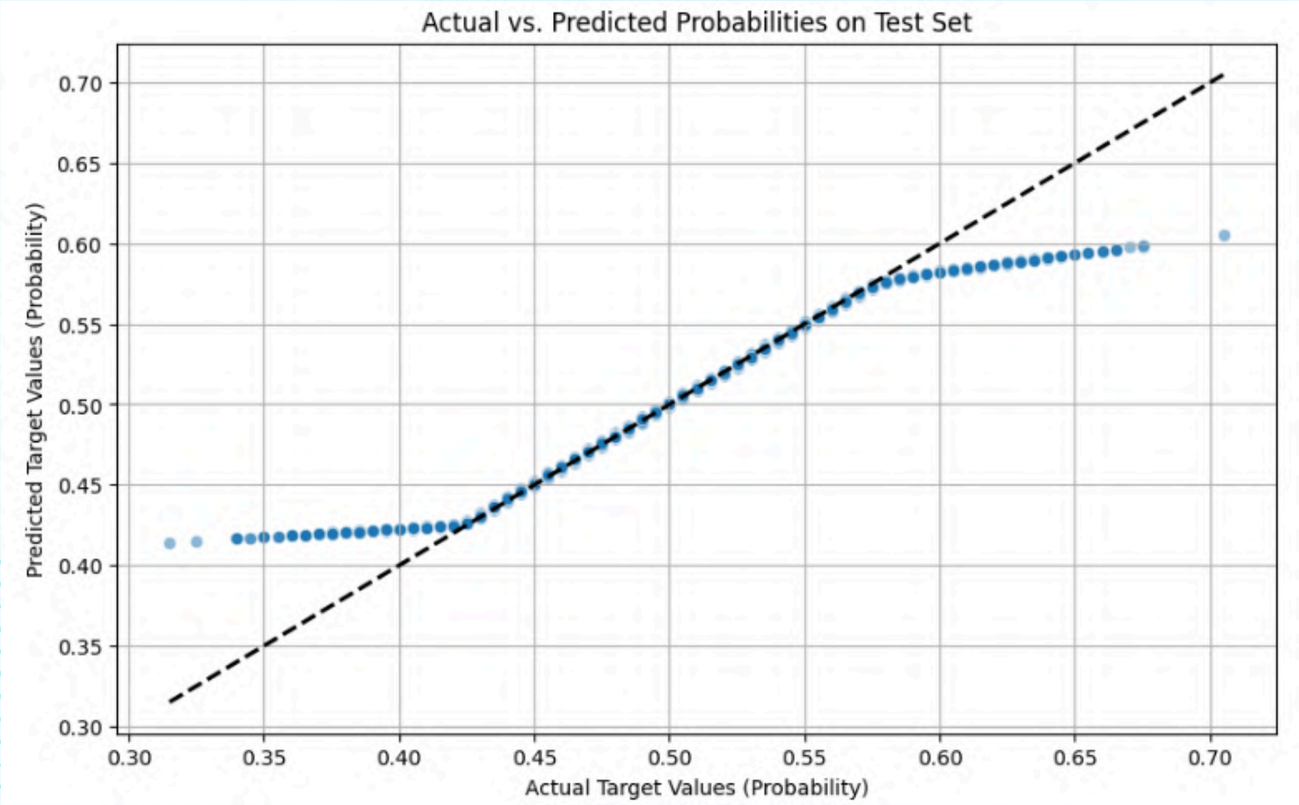


*https://huggingface.co/ibm-granite/granite-geospatial-uki-flooddetection

Performance Metrics

MLP on tabular data

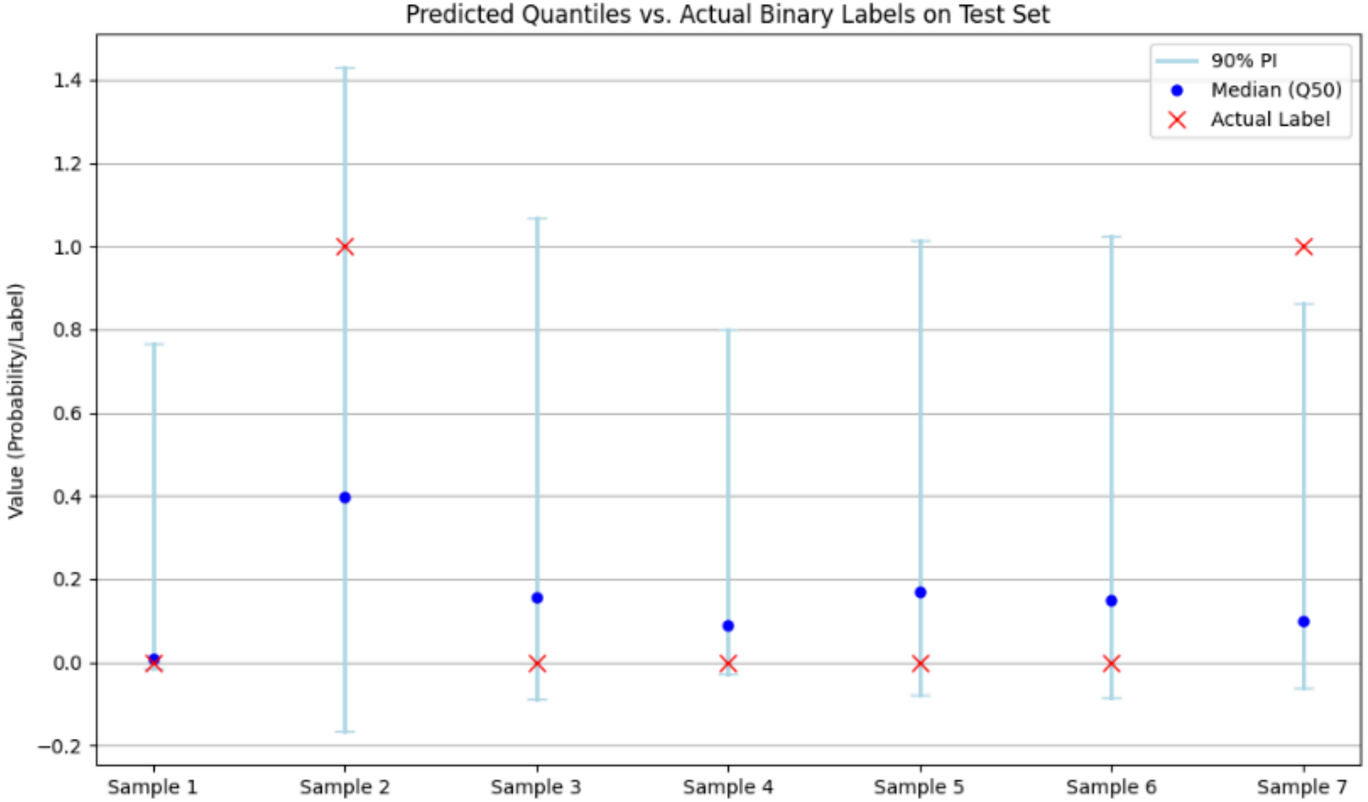
```
Scikit-learn Test MSE: 0.0001
Scikit-learn Test MAE: 0.0029
Scikit-learn Test R-squared: 0.9697
```



Performance Metrics

Quantile Regression on Combined Features

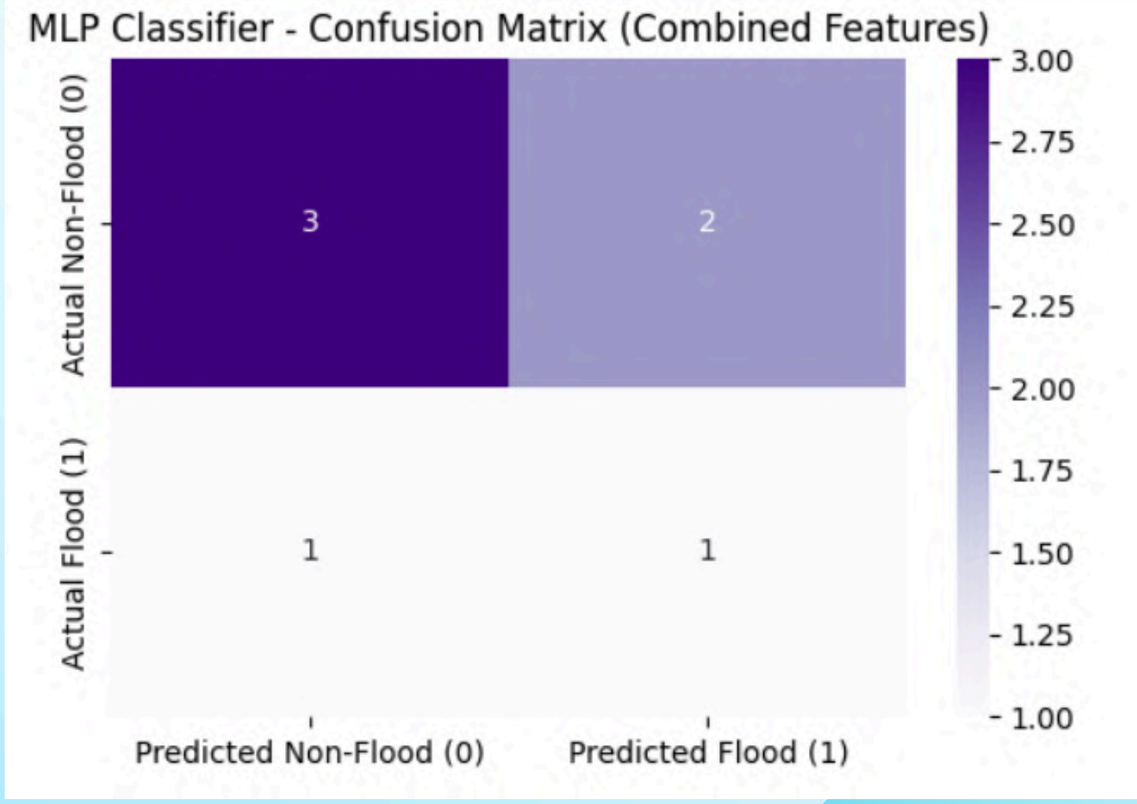
Coverage of 90% Prediction Interval: 85.71%
(6 out of 7 actual values fell within their predicted 90% interval)



Performance Metrics

MLP on the Combined Feature Vector

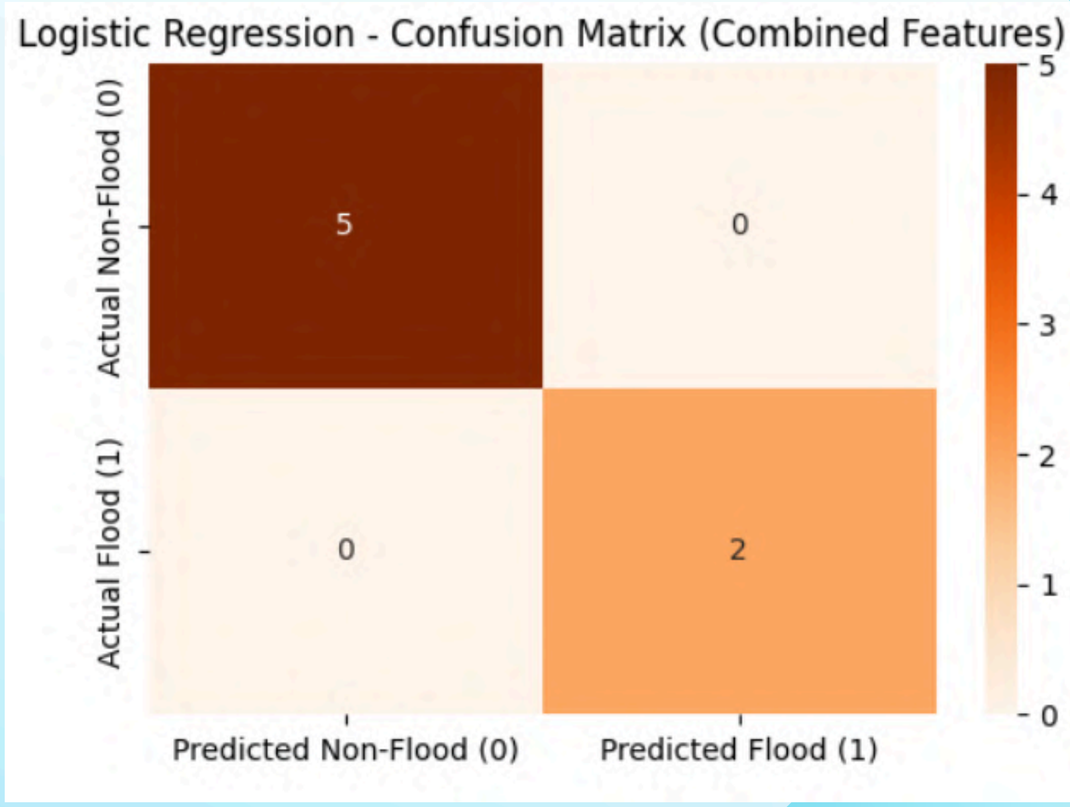
```
--- MLP Classifier Evaluation on Test Set ---
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 80ms/step
              precision    recall  f1-score   support

           0       0.75      0.60      0.67         5
           1       0.33      0.50      0.40         2

    accuracy                           0.57         7
   macro avg       0.54      0.55      0.53         7
weighted avg       0.63      0.57      0.59         7

AUC Score: 0.7000
```



Performance Metrics

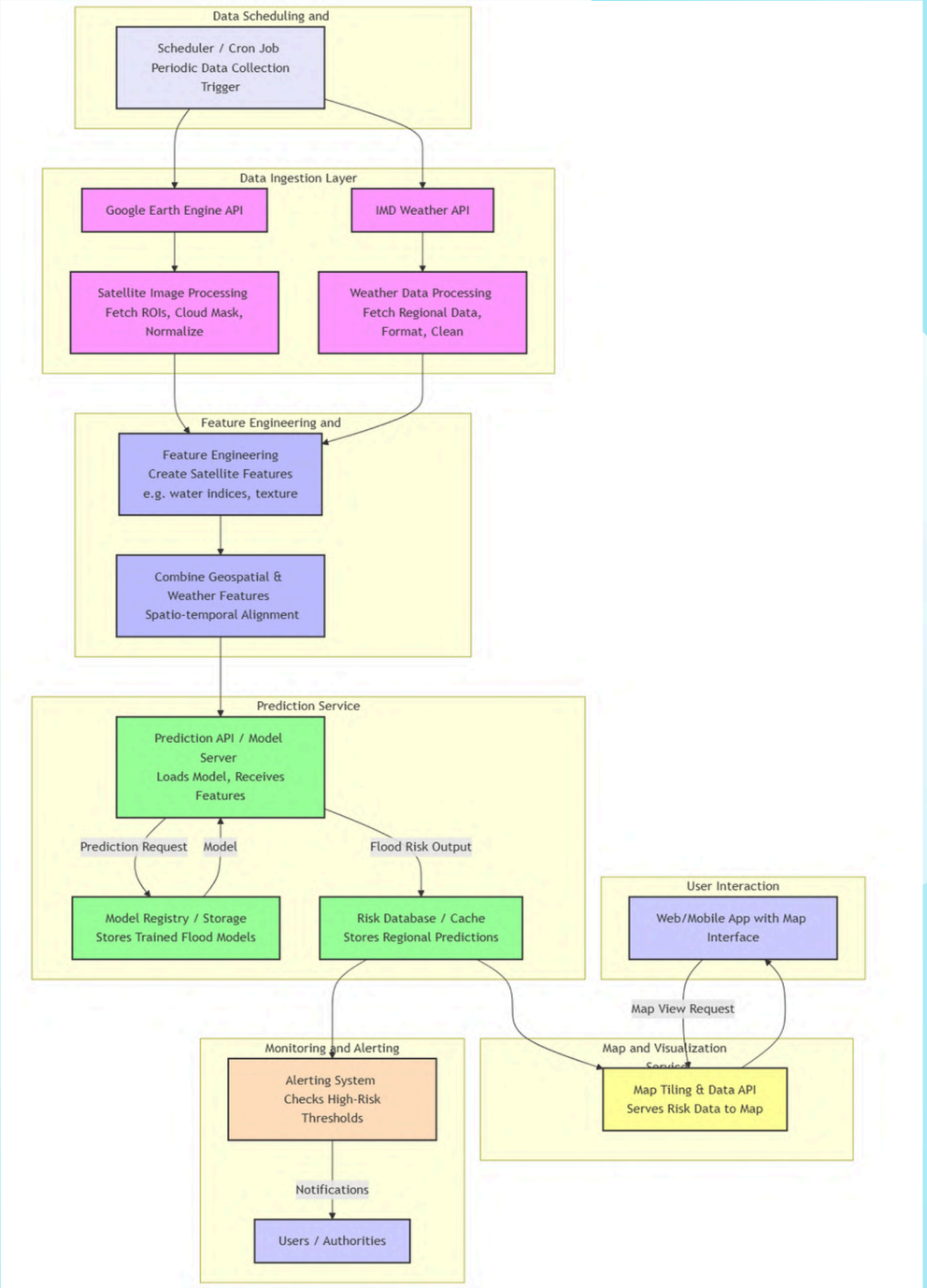
Logistic Regression on the Combined Feature Vector

```
--- Logistic Regression Evaluation on Test Set ---
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00         2

    accuracy                           1.00         7
   macro avg       1.00      1.00      1.00         7
weighted avg       1.00      1.00      1.00         7

AUC Score: 1.0000
```



Deployability

Ideally if our model generalised well, we would've hosted it and displayed a map that would display flood risks for regions in India. This would be possible by collecting satellite images using the earth engine api and weather data using the IMD api.



Thank You