COMP47250 Team Software Project
Project Plan - HAB Detection System
Team: Gradient Descent

Team Members:
Daniel Ilyin [18327256]
Sagar Satish Poojary [24205485]
Dharmik Arvind Vara [24215287]
Karthika Garikapati [24206355]
Kruthi Jangam [24206846]
Roshan Palem [24209558]

Contents

1 Project Objectives

This project aims to develop a web-based Harmful Algal Bloom (HAB) detection and prediction system based on the HABNet architecture proposed by Hill et al. (2020). HABs pose significant risks to marine ecosystems, aquaculture operations, and public health, with traditional detection methods relying on periodic manual sampling that leads to delayed response times.

Our system will implement a spatiotemporal "datacube" approach combined with deep neural networks to achieve high-accuracy HAB detection and prediction capabilities. The primary objectives include:

Core Technical Objectives:

- Develop an automated datacube generator for processing remote sensing data from MODIS-Aqua/Terra satellites and Sentinel-3 sensors
- Implement and train deep learning models (NASNet-Mobile backbone with LSTM) for HAB event classification
- Create a minimum viable reproduction achieving ¿90% detection accuracy and 80% prediction accuracy
- Build a real-time web-based dashboard for HAB monitoring and prediction

User-Centered Objectives:

- Provide marine biologists and environmental agencies with early warning capabilities (up to 8 days ahead)

- Enable aquaculture operators to implement timely mitigation measures
- Support public health officials in issuing timely advisories for recreational water use

Success will be measured through technical performance metrics (detection/prediction accuracy), user evaluation feedback, and system scalability demonstrations using cloud infrastructure.

2 Project Plan

2.1 Sprint Overview

Our development follows an agile methodology with 2-week sprints aligned with project milestones:

Sprint	**Dates**	**Deliverables & Goals**
Sprint 1	20/5 - 2/6	Team formation, environment setup, initial data source exploration
Sprint 2	3/6 - 16/6	MVP development, basic datacube pipeline, simple classifier
Sprint 3	17/6 - 30/6	Model training, interim presentation preparation, web interface prototype
Sprint 4	1/7 - 14/7	Advanced model implementation, cloud deployment, user testing framework
Sprint 5	15/7 - 28/7	System integration, performance optimization, comprehensive evaluation
Sprint 6	29/7 - 11/8	Final testing, documentation, presentation preparation
Sprint 7	12/8 - 19/8	Final report completion, system refinements

2.2 Key Dates

- **9/6/2025:** Project Plan submission
- **23/6/2025:** Interim Presentation & MVP Demo
- **4/8/2025:** Final Presentation & Complete System Demo
- **19/8/2025:** Final Report submission

3 Roles

- **Project Manager (Kruthi):** Tracks milestones, maintains sprint board, leads team syncs.
- **ML Lead (Roshan):** Leads model selection, training, and tuning.

- **Data Engineer (Karthika):** Manages dataset acquisition, cleaning, and ETL.
- **Frontend Developer (Dharmik):** Builds the Streamlit dashboard and integrates backend.
- **DevOps & Cloud Lead (Daniel):** Sets up and maintains cloud infrastructure and deployment.
- **UX & Testing Lead (Sagar):** Handles UI testing, SHAP analysis, and final user evaluations.

4 Architecture

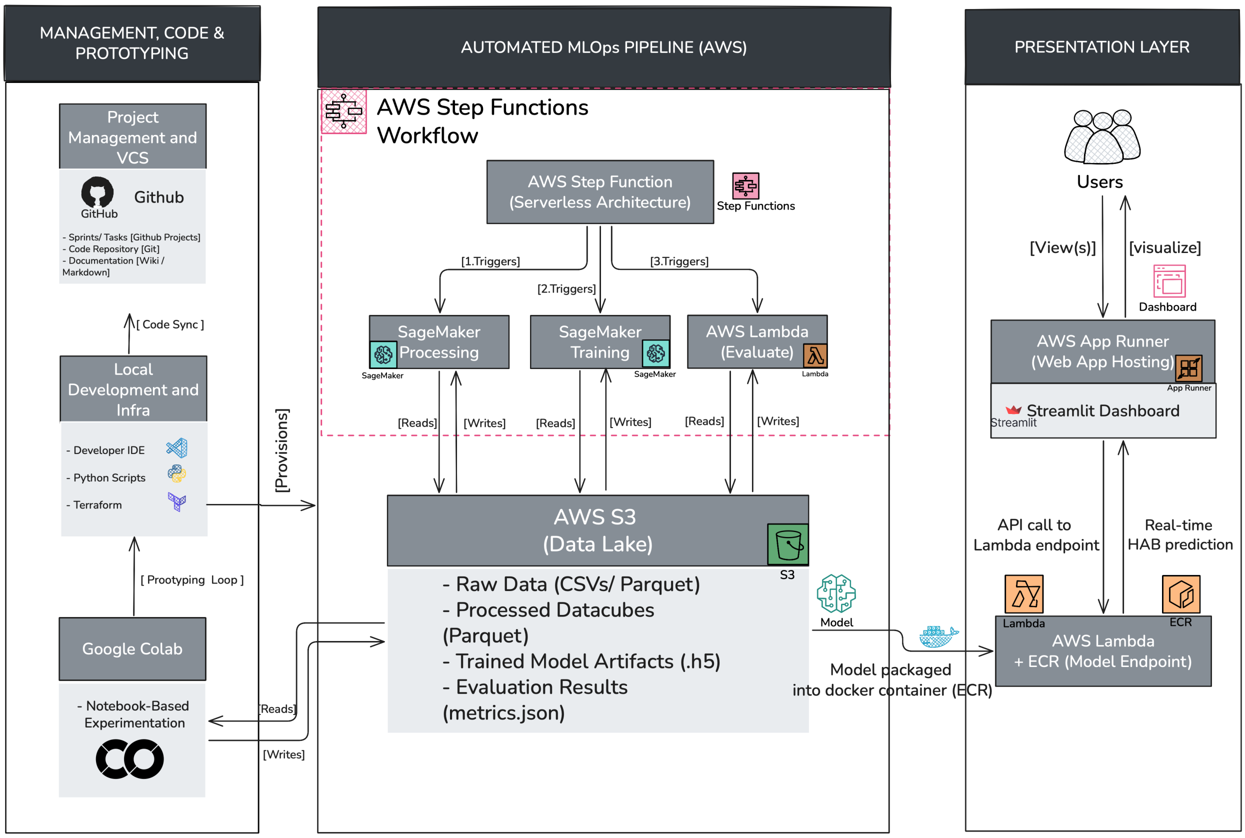

Figure 1: System architecture for HAB detection and prediction. The pipeline integrates local development, version control, and notebook experimentation with an automated AWS-based MLOps workflow using Step Functions, SageMaker, and Lambda. Outputs are stored in S3 and served through a Streamlit dashboard hosted on AWS App Runner for real-time HAB prediction.

5 Data Plan

5.1 Data Collection Strategy

5.1.1 Ground Truth Data Sources

The primary source for labeled HAB events will be obtained from HAEDAT (Harmful Algal Event Database), which provides comprehensive data including:

- Geographic coordinates of HAB events
- Temporal information (timestamps)
- Validated HAB event classifications

5.1.2 Remote Sensing Image Data

NASA MODIS satellite constellation serves as the primary source for remote sensing imagery:

- **MODIS-Aqua**: Ocean color and temperature data
- **MODIS-Terra**: Land and atmospheric observations

5.2 Data Preprocessing Pipeline

The preprocessing workflow follows a systematic approach:

Raw Data → Resampling → Feature Extraction → Datacube Construction

5.2.1 Image Preprocessing Steps

- **Raw Data Acquisition**: Collection of satellite imagery from MODIS sensors
- **Resampling**: Standardization of spatial and temporal resolution
- **Feature Extraction**: Identification of relevant spectral and spatial features
- **Datacube Construction**: Organization of multi-dimensional data arrays

5.3 Data Structure and Labeling

The dataset will be structured into two primary categories:

- **Positive Samples**: Confirmed HAB events with corresponding satellite imagery
- **Negative Samples**: Non-HAB oceanic conditions for comparison

5.3.1 Data Augmentation Strategy

To enhance model robustness and increase training data diversity:

- **Rotation**: Apply various rotational transformations
- **Flipping**: Horizontal and vertical image flipping

5.4 Pretrained Model Implementation Plan

5.4.1 Primary Architecture

The implementation will utilize a transfer learning approach:

Pretrained CNN Backbone → Feature Extraction → Classification

5.4.2 Pretrained Models

Three candidate architectures will be evaluated:

- **NASNet Mobile** (Recommended by original research)
- **EfficientNet-B0/B1** (Alternative approach)
- **ResNet-50** (Baseline comparison)

5.4.3 Model Adaptation Strategies

The adaptation process will follow a progressive approach:

- **Feature Extraction**: Initial freezing of pretrained weights
- **Fine-tuning**: Gradual unfreezing of top layers
- **Layer Replacement**: Modification of final classification layers

5.5 Training Strategy

5.5.1 Data Partitioning

The dataset will be divided according to standard machine learning practices:

Dataset Partition	Percentage
Training Set	70%
Validation Set	10%
Testing Set	20%

5.5.2 Performance Evaluation

Model performance will be monitored using established classification metrics:

- **Precision**: $P = \frac{TP}{TP+FP}$
- **Recall**: $R = \frac{TP}{TP+FN}$
- **F1-Score**: $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

6 GitHub Repository

Our GitHub repository can be found at https://github.com/danieli1245/Harmful-Algal-Bloom-Detection-System. Our repository demonstrates consistent team engagement with members contributing to project documentation, planning and initial development setup.

6.1 Contribution Evidence

Commit Activity: Figure 2 demonstrates the contributions from team members during the initial project setup phase. The repository shows meaningful commits including project structure setup, project plan creation, and demo material.

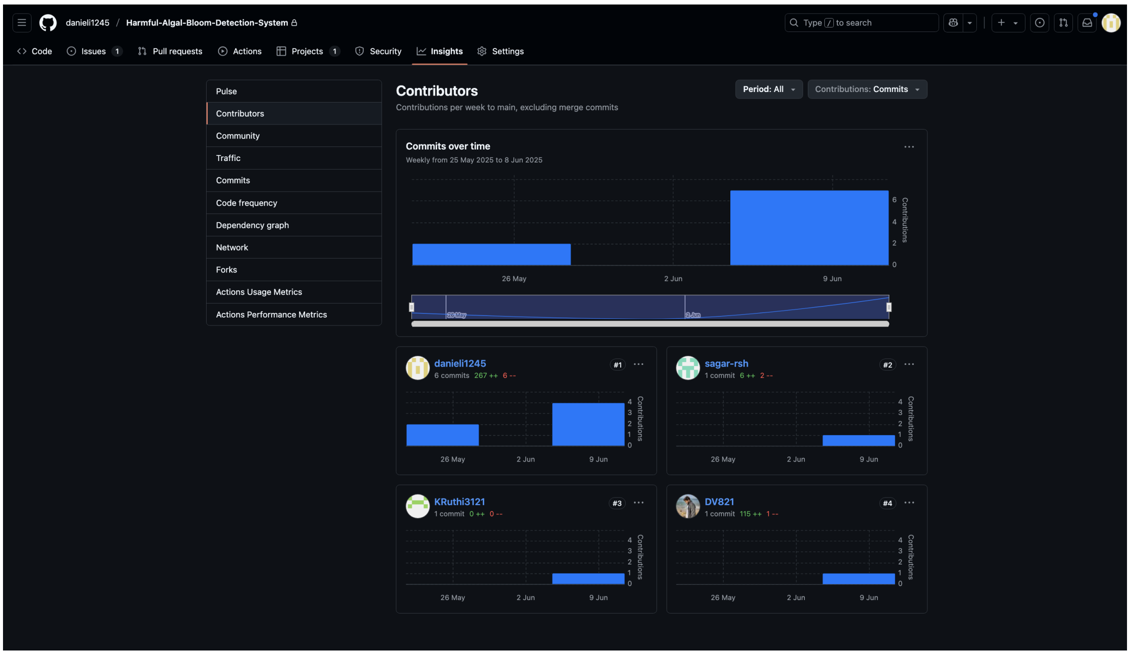

Figure 2: Evidence of regular commits from team members

6.2 Team Meeting Documentation

Regular team meetings are documented and minutes are created through zoom

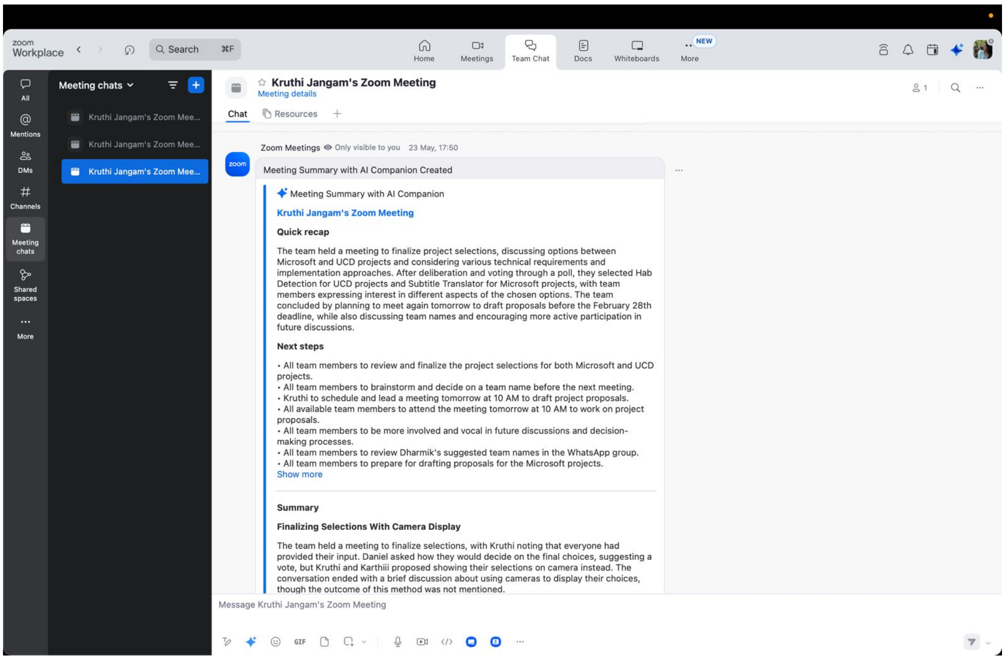

Figure 3: Meeting minutes on zoom

Then we upload the minutes to to our Github Projects dashboard

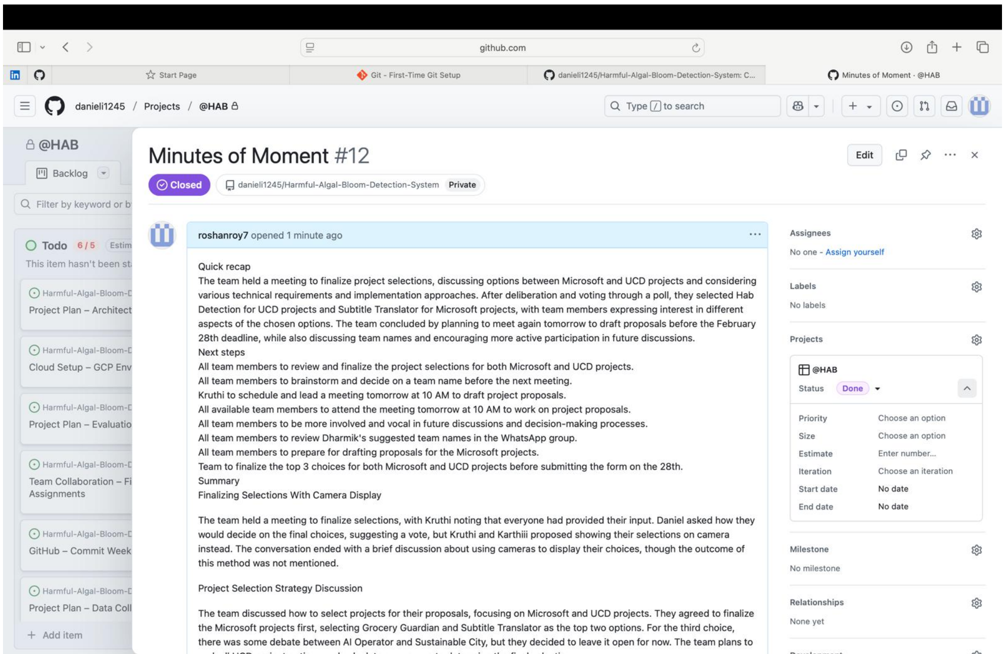

Figure 4: Meeting minutes on Github Project

7 Team Management

7.1 Meeting Format

Meeting Format: Our meetings are conducted via Zoom with structured agendas using Github Projects covering progress updates, blocker identification, and task assignments. All meetings are documented for future reference. We hold a main meeting every Monday, with other follow up meetings throughout the week, as needed.

7.2 Meeting Documentation

Regular team meetings are documented with full minutes capturing the key decisions and progress updates. Figure 4 shows our Zoom-based meeting format.
As referenced in Section 6, we maintain detailed meeting records that are then moved to our GitHub Projects dashboard for task tracking and sprint management (Figure ??).

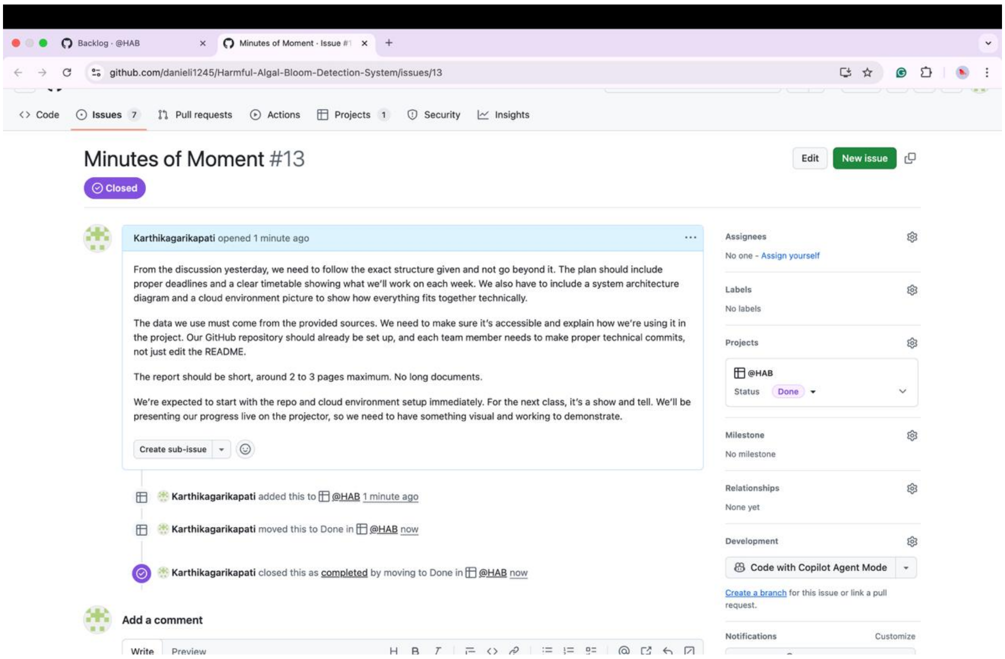

Figure 5: Additional team meeting documentation

7.3 Project Management Tools

GitHub Projects Dashboard: We use GitHub Projects as our primary project management tool for sprint planning and progress tracking. Figure 6 demonstrates our current board organisation with task categorisation and progress visibility.

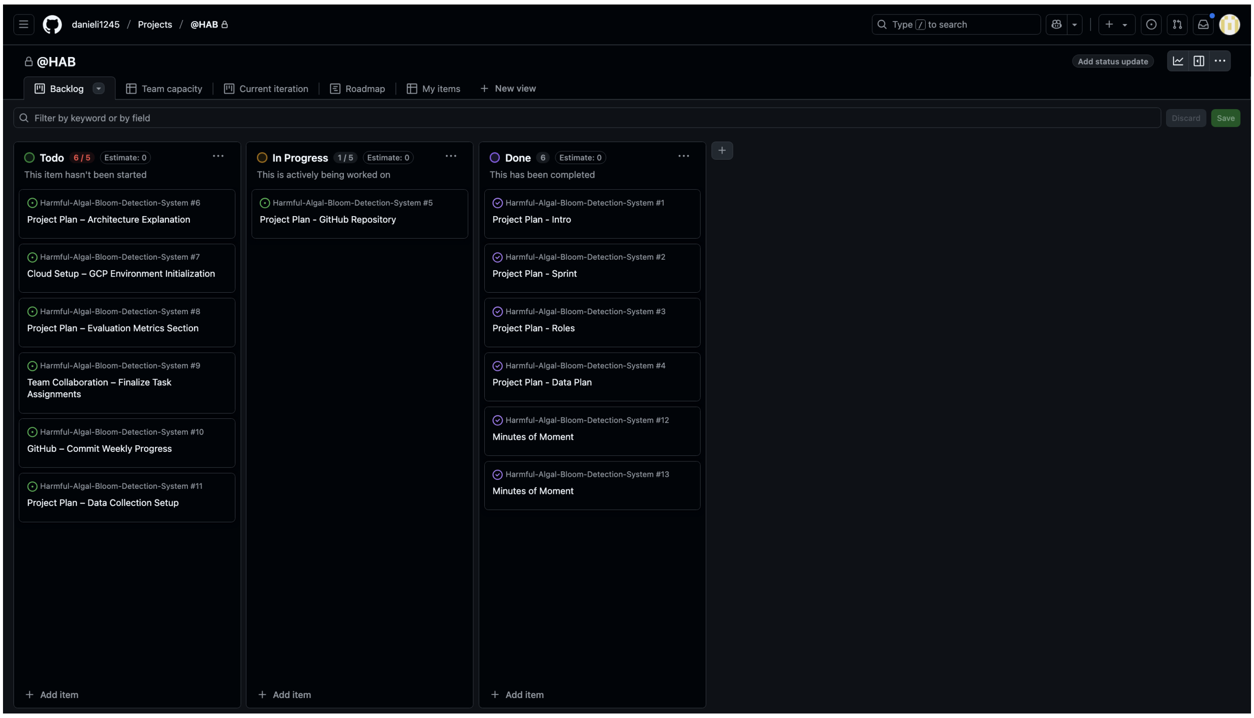

Figure 6: GitHub Projects dashboard showing task sprint planning and progress tracking with To Do, In Progress, and Completed items

Accountability:

- **Sprint Planning:** Tasks are defined and assigned during our meeting planning sessions
- **Progress Tracking:** Daily updates via Whatsapp and status changes on GitHub Projects
- **Blockers:** Issues are brought forward and addressed during the weekly team meetings

8 Current Status and Future Directions

The project is currently in the data acquisition phase. This document presents the foundational strategy for HAB detection implementation. The approach will be refined and optimized based on discoveries made during the data collection and initial experimentation phases.

The methodology will be continuously evaluated and improved as new insights emerge from the data analysis process.