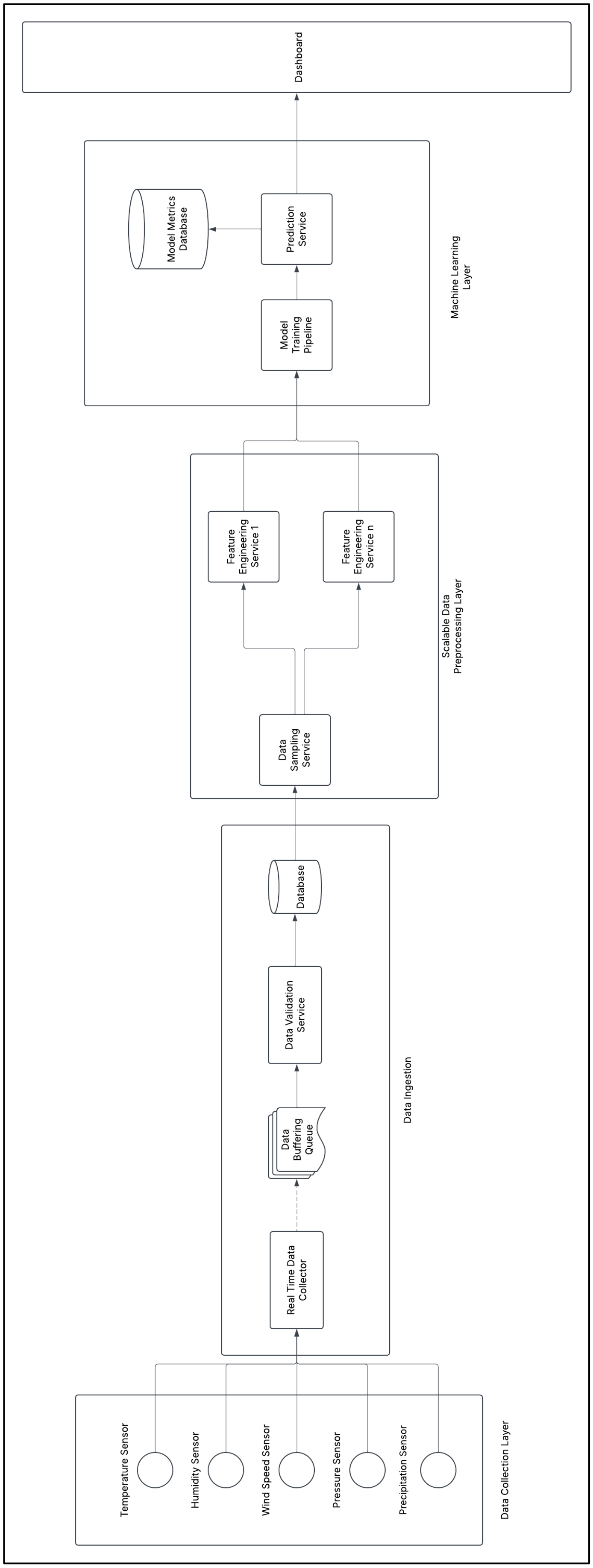
Report 2

Task 1 - Weather Forecasting

Team Semi Colon

1. **System Design Diagram** [Link for the diagram]



2. System Components Explanation

2.1 Data Collection Layer

IoT Sensors: Multiple sensors (temperature, humidity, wind speed, precipitation, pressure) collecting data at 1-minute intervals
External Weather APIs: Secondary data sources for validation and to fill gaps when sensors malfunction
Redundancy: Multiple sensors of the same type in different locations to ensure data reliability

2.2 Data Ingestion Layer

Real-time Data Collector: Service that collects data from all sources
Data Validation Service: Performs initial checks for anomalies, outlier detection, and format validation
Data Buffering Queue: Message queue (like Kafka or RabbitMQ) to handle sensor outages and prevent data loss
Data Storage Database:

- Validated data is persistently stored in a time-series database (like InfluxDB or TimescaleDB)
- Handles high-frequency writes efficiently with appropriate indexing
- Stores both raw and validated data for auditability and backup purposes
- Implements partitioning strategies for faster query performance on historical data

2.3 Data Processing Layer

Missing Value Imputation: Handles missing data using appropriate statistical methods or interpolation
Feature Engineering: Creates additional features that may improve prediction accuracy
Data Normalization: Standardizes data for model consumption

2.4 Machine Learning Layer

Model Training Pipeline: Automatically retrains models with new data on a scheduled basis

- Implemented using Dagster for workflow orchestration

- Dagster pipelines handle the end-to-end ML lifecycle including data extraction, preprocessing, training, and evaluation
- Scheduled automatic retraining with new data on daily/weekly basis
- Supports dependency management between different pipeline steps
- Provides built-in monitoring, logging, and error handling
- Enables conditional execution paths for different model types or data conditions

Model Registry: Stores model versions with performance metrics
Prediction Service: Generates 21-day forecasts using the best performing model
Model Metrics: Monitors model performance with real-time feedback loop

2.5 API/Dashboard Layer

Presents forecast information to farmers through APIs and a user-friendly dashboard

Error Handling & Sensor Malfunction Mitigation

Data Validation Checks: Automatically flags suspicious readings

Fallback Mechanisms:

- When sensors malfunction, system uses historical patterns
- External weather API data supplements missing sensor data

Anomaly Detection: Machine learning identifies sensor malfunctions in real-time

Confidence Scores: Each prediction includes reliability rating based on data quality

Alert System: Notifies maintenance team of sensor issues for timely repairs