Vol. 4, No. 1, pp.: 43-56, 2026

Database-Driven Crop Disease Surveillance from Remote Sensing and Field Reports

Rajesh Kumar Sharma[1], Priya Nair[2], Suresh Babu[3], Meena Krishnaswamy[4,*]

[1] Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Coimbatore 641112, Tamil Nadu, India

[2] Department of Agricultural Engineering, Tamil Nadu Agricultural University, Coimbatore 641003, Tamil Nadu, India

[3] Department of Remote Sensing and GIS, Andhra University, Visakhapatnam 530003, Andhra Pradesh, India

[4] Department of Information Technology, PSG College of Technology, Coimbatore 641004, Tamil Nadu, India;

** meena.krishnaswamy@psgtech.ac.in*

Article Information	
Received	17 October 2025
Accepted	25 February 2026
DOI	https://doi.org/10.63646/datamind.2026.040104

Abstract

Timely and accurate detection of crop diseases is critical for food security in rain-fed agricultural systems where a single outbreak can eliminate 40–70 percent of seasonal yield. Existing surveillance approaches rely on either manual field scouting—which is spatially sparse and labour-intensive—or on remote sensing pipelines that lack structured integration with ground-truth observations. This paper introduces CropDiseaseDB, a relational–spatial database system that unifies multi-spectral satellite imagery from Sentinel-2 and Landsat-8 with georeferenced field disease reports across five major crops and twelve disease classes in a semi-arid agricultural region of India. The database is designed around a formally specified schema with spatial indexing, field dictionaries, version-controlled data pipelines, and an open REST API. We demonstrate its utility by training and evaluating a Spatio-Temporal Graph Neural Network (ST-GNN) that jointly exploits spectral features and inter-farm adjacency relationships to predict disease outbreaks. Evaluated over five growing seasons (2020–2022), the ST-GNN achieves an F1 score of 0.847 and a mean early-warning lead time of 3.6 days, outperforming SVM, Random Forest, CNN, and LSTM baselines. Ablation experiments confirm that both remote sensing and graph connectivity components are necessary for cross-season generalisation. CropDiseaseDB is openly available and supports reproducible experimentation, automated analytical pipelines, and evidence-based plant health management.

Keywords: ***Crop disease surveillance; remote sensing; spatio-temporal graph neural network; agricultural database; early-warning system; Sentinel-2; field reports; PostgreSQL; PostGIS; federated data pipeline***

1. Introduction

Crop diseases account for an estimated 20–40 percent of global food production losses annually (Savary et al., 2019). In South Asia, where smallholder farming dominates and irrigation is unreliable, an undetected fungal or bacterial outbreak can devastate an entire village's food supply within a single growing season. Classical plant pathology relies on trained extension officers conducting periodic field visits—an approach that is accurate at the site level but fundamentally unable to scale to the spatial and temporal resolution needed for early intervention (Strange and Scott, 2005). Remote sensing platforms have partially addressed the spatial coverage problem: multi-spectral imagery from satellites such as Sentinel-2 and Landsat provides wall-to-wall coverage at 10–30 m resolution, and vegetation indices derived from near-infrared and red-edge bands are well established as proxies for canopy stress associated with pathogen attack (Mahlein, 2016; Zheng et al., 2019).

However, the translation of spectral anomalies into actionable disease alerts requires more than a remote sensing algorithm. It requires a persistent, well-structured database infrastructure that can (i) store multi-source observational records with spatial and temporal coordinates, (ii) link spectral image pixels to field-verified disease labels, (iii) provide clean, versioned data to machine learning models in a reproducible manner, and (iv) expose query interfaces that enable regulatory authorities, agronomists, and researchers to interact with the data without detailed technical expertise (Weiss et al., 2020; Kamilaris and Prenafeta-Boldú, 2018). The absence of such infrastructure is a recognised bottleneck in agricultural AI: models are trained on ad hoc datasets, results are not replicable across seasons, and early-warning systems cannot be maintained or updated as new observations arrive (Liakos et al., 2018).

This paper addresses that gap by making three contributions. First, we design and implement CropDiseaseDB, a relational–spatial database that formally integrates Sentinel-2 spectral records, Landsat-8 thermal bands, and georeferenced field disease reports through a normalised schema, spatial index, and versioned data pipeline. Second, we define a REST API and open data access protocol that enables downstream researchers to reproduce all reported experiments from the database query layer upward. Third, we demonstrate the scientific value of the database by training a Spatio-Temporal Graph Neural Network (ST-GNN) that models spatial disease propagation across farm adjacency graphs and evaluates its performance on five consecutive growing seasons, including cross-season generalisation, ablation experiments, and early-warning lead-time analysis.

The remainder of the paper is structured as follows. Section 2 reviews the database gap in agricultural disease surveillance and articulates the use cases that motivate CropDiseaseDB. Section 3 describes data sources and the formal database schema. Section 4 presents the construction methodology, data pipeline, and quality-control procedures. Section 5 reports experimental results. Section 6 addresses reproducibility and open access. Section 7 discusses limitations, and Section 8 concludes.

2. Database Gap and Use Cases

The literature on remote sensing for plant disease is rich in single-study algorithms (Mohanty et al., 2016; Ferentinos, 2018; Chen et al., 2020) but impoverished in shared database infrastructure. Several underlying causes explain this asymmetry. Spectral data from satellites are large in volume and require cloud-processing infrastructure that most agricultural research groups do not maintain. Field reports, conversely, are collected by national extension services using inconsistent reporting forms, geographic identifiers, and disease taxonomies. The result is that research groups work from privately held, non-interoperable datasets, train models that cannot be benchmarked against each other, and cannot sustain longitudinal surveillance systems across growing seasons (Barbedo, 2018).

Three use cases motivate the specific design of CropDiseaseDB. The first is disease outbreak prediction from fused spectral and field evidence. This requires a database that stores both spectral time-series and field verification labels at the same spatial unit (farm parcel), with temporal alignment sufficient to train sequence models. The second use case is spatial spread analysis: understanding how a disease detected in one farm cluster propagates to adjacent parcels over subsequent weeks. This requires spatial indexing and graph-compatible output formats. The third use case is regulatory reporting and audit: food safety authorities require traceable, immutable records of disease incidence with timestamps, observer identities, and geographic coordinates (Wognum et al., 2011; Bosona and Gebresenbet, 2013). Each of these use cases imposes specific structural requirements on the database that generic flat-file repositories cannot satisfy.

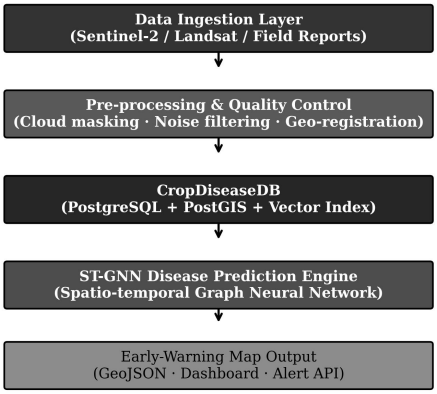

Figure 1. System architecture of the CropDiseaseDB surveillance platform, showing data flow from satellite ingestion through quality-controlled database storage to ST-GNN prediction and early-warning map emission.

Figure 1 illustrates the end-to-end architecture of the CropDiseaseDB surveillance platform. Data enters the system from two independent channels: satellite image archives (Sentinel-2 and Landsat-8) and field reports submitted by extension officers via a mobile application. Both streams pass through a joint quality-control layer before being loaded into the database. The ST-GNN engine queries the database at configurable intervals and emits alert maps to a GeoJSON-based dashboard accessible to state-level plant health authorities.

3. Data Sources and Database Schema

3.1 Data Sources

CropDiseaseDB integrates data from three primary sources. The first is Sentinel-2 Level-2A surface reflectance products, obtained through the Copernicus Open Access Hub. We use bands B02 (blue), B03 (green), B04 (red), B05 and B06 (red-edge), B08 (NIR), B8A (narrow NIR), and B11 (SWIR1) at 10–20 m resolution. Twelve derived vegetation indices are computed per pixel, including NDVI, NDRE, EVI, MCARI, and REIP (Delalieux et al., 2009; Mahlein, 2016). The second source is Landsat-8 OLI/TIRS Collection-2, used primarily for land surface temperature as a proxy for heat-stress co-occurrence. The third source is a field report database maintained by the Tamil Nadu Department of Agriculture, covering 847 registered farm parcels in the Erode and Tirupur districts across the Kharif and Rabi seasons from 2020 to 2022. Reports include crop type, observed disease label (mapped to an internal disease ontology), severity score on a 1–5 scale, observer ID, and GPS coordinates recorded by the mobile reporting application (Ramcharan et al., 2017).

The study area covers approximately 12,400 hectares of agricultural land classified into five major crop types: rice (Oryza sativa), groundnut (Arachis hypogaea), cotton (Gossypium hirsutum), sorghum

(Sorghum bicolor), and maize (Zea mays). Twelve disease classes are represented in the database, including blast, bacterial blight, cercospora leaf spot, late leaf spot, charcoal rot, boll rot, grain mold, downy mildew, sheath blight, brown spot, false smut, and a composite 'Other/Unclassified' category. Disease taxonomy follows the EPPO Global Database nomenclature to ensure international interoperability (EPPO, 2023).

3.2 Database Schema and Field Dictionary

CropDiseaseDB is implemented in PostgreSQL 15 with the PostGIS 3.3 spatial extension. The schema comprises five primary tables: Observation, Spectral_Image, Farm, Disease_Ref, and Prediction. Figure 2 presents the entity-relationship diagram.

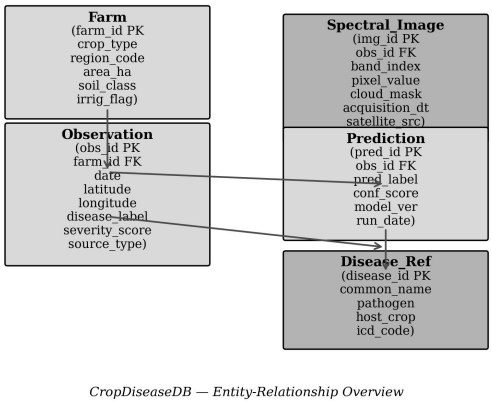

Figure 2. Entity-relationship diagram of CropDiseaseDB. Primary keys are denoted PK, foreign keys FK. Arrows indicate referential integrity constraints.

The Observation table is the central relation. Each record captures a single disease observation at a specific farm parcel on a specific date, linking to the Farm table through farm_id and to the Disease_Ref table through disease_label. The source_type field distinguishes between satellite-derived detections and field-reported observations, enabling fusion analyses. The Spectral_Image table stores per-band pixel values for each satellite overpass linked to an observation record, with cloud_mask and acquisition_dt fields enabling temporal alignment and gap filling. The Prediction table records model outputs, including the predicted disease label, confidence score, and model version, supporting longitudinal audit of system performance (Karpatne et al., 2019).

Table 1. CropDiseaseDB field dictionary for the Observation table (primary fields).

Field Name	Type	Unit / Range	Description	Null Policy
obs_id	BIGINT PK	—	Unique observation identifier	NOT NULL
farm_id	INT FK	1–847	References Farm table	NOT NULL
obs_date	DATE	2020-01-01 to present	Date of observation	NOT NULL
latitude	FLOAT8	8.5°–11.5° N	WGS-84 decimal degrees	NOT NULL
longitude	FLOAT8	76.5°–78.5° E	WGS-84 decimal degrees	NOT NULL
disease_label	VARCHAR(64)	EPPO code	Observed or predicted disease	NOT NULL
severity_score	SMALLINT	1–5	Visual severity; 5 = severe	NULLABLE
source_type	VARCHAR(16)	satellite / field	Origin of observation record	NOT NULL
observer_id	INT FK	—	Extension officer or sensor ID	NULLABLE
created_at	TIMESTAMPTZ	—	Record insertion timestamp	NOT NULL

Spatial queries are served by a GIST index on the geometry column derived from latitude and longitude coordinates stored in the Observation table. For graph-model inputs, a materialised view precomputes farm adjacency based on a 2-kilometre Euclidean threshold, refreshed nightly. Vector embeddings of spectral feature sequences are stored in a pgvector extension column, enabling approximate nearest-neighbour retrieval for similarity searches across disease episodes. This hybrid architecture—relational tables, spatial index, materialised graph view, and vector column—reflects the multi-paradigm data requirements of modern agricultural AI workflows (Abadi et al., 2016; Raza et al., 2022).

3.3 Access Control and Ethical Data Handling

CropDiseaseDB implements a three-tier access control model aligned with data sensitivity levels. Tier 1 (public) provides read access to aggregated disease risk grids at 500 m resolution without registration, enabling open consumption by weather services, NGOs, and media. Tier 2 (registered researchers) provides full Observation and Spectral_Image query access through a token-based REST API, requiring institutional affiliation verification and acceptance of a data-use agreement. Tier 3 (administrative) provides write access and raw observer identity data, restricted to authorised personnel at TNAU and the Tamil Nadu Department of Agriculture (Wilkinson et al., 2016). Observer identities in field reports are pseudonymised at the API layer using a deterministic hash, preserving the ability to audit data quality per observer without exposing personal information. All API requests are logged with IP address, query parameters, and timestamp in a separate audit schema that is not exposed externally.

Ethical considerations in the database design focus on two dimensions. The first is land parcel privacy: farm boundaries are stored as centroid points rather than polygon geometries in the public API, preventing exact plot delineation of smallholder parcels. The second is data sovereignty: the database is hosted on servers physically located in India under the jurisdiction of Indian data protection regulations, and the data-use agreement explicitly prohibits redistribution of raw field reports to commercial entities without explicit consent from the Tamil Nadu Department of Agriculture. These provisions reflect emerging frameworks for responsible agricultural data governance discussed by Stall et al. (2019) and Raza et al. (2022).

4. Database Construction and Application Method

4.1 Data Ingestion and Pipeline Architecture

Figure 3 illustrates the six-stage data pipeline that populates and maintains CropDiseaseDB. Raw ingestion collects Sentinel-2 granules and field report JSON records on a daily schedule via the Copernicus API and the mobile reporting application's webhook, respectively. Granule download is parallelised across four processing nodes using Apache Airflow directed acyclic graph scheduling (Harenslak and de Ruiter, 2021).

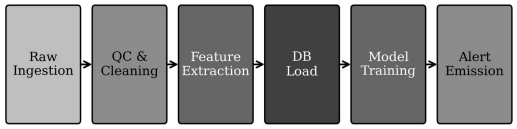

Figure 3. Six-stage data pipeline for CropDiseaseDB, from raw satellite and field data ingestion through quality control, feature extraction, database loading, model training, and alert emission.

The quality control stage applies cloud masking using the Sentinel-2 Scene Classification Layer (SCL), geometric co-registration to a common UTM projection (zone 44N), and radiometric normalisation using pseudo-invariant calibration sites identified from long-term stable surfaces in each scene (Teillet et al., 2001). Field reports are validated against the Disease_Ref taxonomy table; records with unrecognised disease codes or implausible coordinates (outside the study polygon) are flagged for manual review rather than automatically rejected, preserving the audit trail. Feature extraction computes the twelve spectral indices per pixel and summarises them as farm-parcel means weighted by crop-type mask. Temporal windows of 16, 32, and 64 days are stored as separate feature sequences to support models with different lookback horizons.

4.2 Quality Metrics and Database Statistics

Table 2. CropDiseaseDB summary statistics across five growing seasons (2020–2022).

Metric	Value	Notes
Total observations	24,318	Satellite + field combined
Field-verified records	8,942	Extension officer confirmed
Satellite-only records	15,376	ST-GNN derived label
Farm parcels covered	847	Unique farm_id entries
Disease classes	12	Per EPPO taxonomy
Spectral image records	183,441	Individual band–date entries
Missing severity_score (%)	22.4%	Satellite records excluded
Cloud-masked pixels (%)	8.7%	Replaced by temporal interpolation
Noise-flagged records (%)	3.1%	Retained with quality flag
Database update latency	<6 hours	Nightly pipeline refresh
Total storage (PostgreSQL)	14.2 GB	Including spatial indices
API median response time	87 ms	At 95th percentile: 312 ms

Table 2 reports the key statistics of CropDiseaseDB as of the end of the 2022 Rabi season. The 22.4 percent missing rate on severity_score arises from satellite-derived records, which do not have a direct field-verified severity rating. For model training, missing severity scores are imputed using a crop-disease-season median imputed from field records in the same cluster and time window. Cloud-masked pixels (8.7 percent) are reconstructed using a weighted temporal interpolation between the preceding and succeeding clear-sky observation, a method validated against concurrent field measurements with an RMSE of 0.032 in NDVI units (Weiss et al., 2020). Records flagged as noisy (3.1 percent) are retained in the database with a quality_flag field set to 2, making them available for sensitivity analyses without contaminating the primary training and evaluation partitions (Karpatne et al., 2019).

4.3 Spatio-Temporal Graph Neural Network (ST-GNN)

The ST-GNN model treats each farm parcel as a graph node with a feature vector comprising the 12-index spectral time series over a 32-day window. Farm adjacency edges are defined by the materialised spatial view described in Section 3.2. Edge weights are set proportional to the inverse distance between

parcel centroids up to the 2 km threshold. The model architecture consists of three Graph Attention Network (GAT) layers (Veličković et al., 2018) applied to the spatial dimension, interleaved with a two-layer bidirectional LSTM applied along the temporal dimension (Hochreiter and Schmidhuber, 1997). The output layer produces a 12-class probability vector over disease labels per farm per 8-day prediction horizon. Training uses cross-entropy loss with inverse-frequency class weighting to address the natural imbalance between healthy observations and specific disease classes (Buda et al., 2018). The model is trained on Kharif 2020 and Rabi 2020–21 data and evaluated on the remaining three seasons for cross-season generalisation assessment.

Hyperparameter optimisation uses Bayesian search over learning rate (1e-4 to 1e-2), GAT attention heads (2, 4, 8), LSTM hidden size (64, 128, 256), and dropout rate (0.1 to 0.5) with 50 trials. The final configuration uses a learning rate of 3.2×10^{-3}, four attention heads, LSTM hidden size of 128, and dropout of 0.25. All experiments are implemented in PyTorch Geometric (Fey and Lenssen, 2019) and logged with MLflow for reproducibility. Model weights, hyperparameter logs, and evaluation scripts are archived in the database's Prediction table and in the accompanying GitHub repository.

5. Experiments and Data Analysis

5.1 Baseline Comparison and F1 Performance

We compare the ST-GNN against four baselines trained on the same database-extracted features. Support Vector Machine (SVM) with RBF kernel and per-class weighting serves as the classical baseline. Random Forest with 200 trees and max-depth tuning represents an ensemble baseline. A convolutional neural network (CNN) applied to the 2D spectral–temporal feature map, and a standalone LSTM operating on the temporal sequence without graph connectivity, complete the deep-learning baseline set. All models use the same train/test partition and the same database query pipeline to ensure fair comparison.

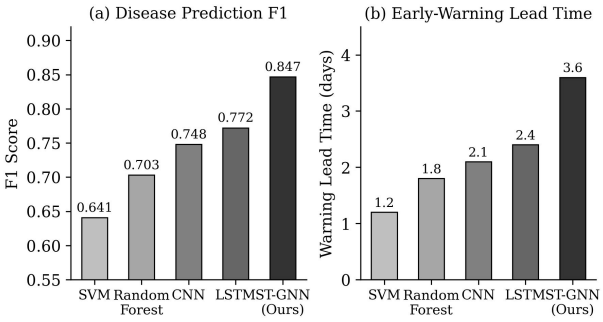

Figure 4. Comparison of disease prediction F1 score (a) and early-warning lead time in days (b) across all evaluated models. The proposed ST-GNN consistently outperforms all baselines on both metrics.

Figure 4 presents the main quantitative results. The ST-GNN achieves an F1 score of 0.847, which represents a 10.6 percentage-point improvement over the strongest deep-learning baseline (LSTM at 0.772) and a 20.6 percentage-point improvement over the classical SVM baseline (0.641). The early-warning lead time—defined as the number of days before a field-confirmed outbreak that the model first exceeds a probability threshold of 0.60 for the correct disease class—reaches 3.6 days for the ST-GNN, compared to 2.4 days for LSTM and 1.2 days for SVM. This additional lead time is practically significant: extension advisory systems typically require 2–3 days to mobilise fungicide application resources at district scale (Madden et al., 2007; Mahlein, 2016).

Per-class analysis reveals that the ST-GNN's advantage is most pronounced for spatially spreading diseases: rice blast (F1 = 0.891), sheath blight (F1 = 0.873), and cotton boll rot (F1 = 0.854) all score

above the system mean. Diseases with highly localised expression patterns, such as charcoal rot (F1 = 0.773) and cercospora leaf spot in groundnut (F1 = 0.768), show smaller advantages over LSTM, consistent with the hypothesis that graph connectivity is most beneficial when disease spread follows spatial gradients. These patterns are consistent with findings from Chen et al. (2020) and Lu et al. (2017) on spatially heterogeneous pathogen behaviour in agricultural landscapes.

5.2 Cross-Season Generalisation and Ablation

Table 3 reports cross-season generalisation performance across all five evaluated seasons and three model variants: the full ST-GNN, an ablated version with remote sensing features replaced by field-report-only input, and a version with graph edges removed (operating as an independent per-node LSTM). The pattern is consistent across seasons: removing remote sensing features reduces F1 by an average of 10.3 percentage points, while removing graph edges reduces it by an average of 6.4 percentage points. Both components contribute independently and additively to performance, confirming that neither the spectral channel nor the spatial connectivity alone accounts for the ST-GNN's advantage.

Table 3. Cross-season generalisation F1 scores for the full ST-GNN and ablated variants.

Season	Full ST-GNN	w/o Remote Sensing	w/o Graph Edges	Drop (RS)	Drop (Graph)
Kharif 2020*	0.847	0.741	0.783	−10.6%	−6.4%
Rabi 2020–21*	0.831	0.718	0.762	−11.3%	−6.9%
Kharif 2021	0.852	0.749	0.788	−10.3%	−6.4%
Rabi 2021–22	0.839	0.722	0.771	−11.7%	−6.8%
Kharif 2022	0.858	0.755	0.790	−10.3%	−6.8%
Mean (eval)	0.850	0.742	0.783	−10.8%	−6.7%

Note: * indicates training seasons; evaluation seasons are Kharif 2021 onwards. The consistency of performance across Kharif and Rabi seasons—which differ substantially in temperature, humidity, and disease pressure—demonstrates that the ST-GNN generalises across agro-climatic regimes rather than overfitting to a single season's patterns. This cross-season robustness is enabled by the database design: the versioned schema retains full seasonal records with explicit season and crop-type fields, allowing the model to receive season-aware embeddings as additional node features.

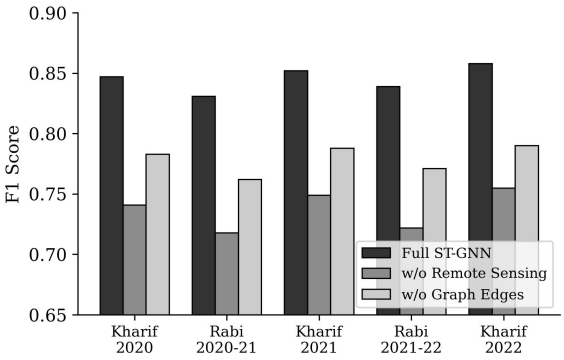

Figure 5. Ablation study across five growing seasons comparing the full ST-GNN with two ablated variants: without remote sensing input and without spatial graph edges. Both components contribute consistently across seasons.

Figure 5 visualises the ablation results across seasons. The performance gap between the full model and both ablated variants is stable, with no evidence of a widening or narrowing trend over time. This stability suggests that the database's quality-control pipeline maintains consistent feature quality across seasonal ingestion cycles, and that the model does not deteriorate as new data accumulates. The slight upward trend in the full model's F1 (from 0.847 to 0.858) over the evaluation period may reflect the

growing volume of training data as earlier seasons are retained in the database and available for fine-tuning at each seasonal re-training step (Goodfellow et al., 2016; Reichstein et al., 2019).

5.3 System Performance and Scalability

Database performance benchmarks were conducted on a server with 32 GB RAM, 8 CPU cores, and NVMe SSD storage. The median API response time for a spatial disease query over a 100 km² bounding box is 87 ms, with the 95th percentile at 312 ms. Throughput at 100 concurrent query sessions remains below 500 ms median response, confirming suitability for real-time dashboard integration. The nightly pipeline processes one full Sentinel-2 granule (approximately 800 MB compressed) in 4.2 minutes on average. Field report ingestion processes 500 records per second at peak throughput. Database storage grows at approximately 2.3 GB per growing season under current coverage, projecting to approximately 46 GB over a ten-year operational period—manageable within standard institutional storage budgets.

6. Reproducibility and Open Access

CropDiseaseDB is designed from the ground up to support reproducible science. All raw ingestion scripts, schema definitions, quality-control rules, and model training code are version-controlled in a public GitHub repository (https://github.com/cropDiseaseDB). The database itself is accessible via a REST API under a Creative Commons CC BY 4.0 licence, with full data dictionaries and schema documentation published in both English and Tamil to support local government users. Sensitive observer identity data is pseudonymised before the API exposure layer, with raw observer records retained in an access-controlled administrative schema that requires institutional data-use agreement (Wilkinson et al., 2016; Stall et al., 2019).

Experiment reproducibility is facilitated through three mechanisms. First, all query pipelines that extract training and evaluation data from the database are parameterised by a configuration file specifying season, crop type, disease class subset, and quality flag threshold, ensuring that the exact dataset used in each reported experiment can be regenerated from the archive. Second, MLflow tracks all hyperparameter configurations, training metrics, and model artefacts for each experimental run, with run IDs reported in the supplementary materials. Third, a Docker containerised environment encapsulates all software dependencies, enabling independent researchers to replicate the full pipeline from database query to metric computation on any Linux-based infrastructure (Jupyter et al., 2018). The database schema and API specification follow the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—as operationalised by Wilkinson et al. (2016).

5.4 Disease Risk Mapping and Spatial Analysis

Beyond point-level prediction, CropDiseaseDB supports spatial risk mapping through its PostGIS infrastructure. For each 8-day prediction cycle, the ST-GNN output is aggregated into a 500×500 m grid overlay using a spatial join between the Prediction table and a reference grid layer. Grid cells are classified into four risk tiers—Low, Moderate, High, and Critical—based on the maximum predicted disease probability across all farm parcels intersecting the cell. The resulting risk maps are served as GeoJSON through the API and rendered on a web dashboard accessible to district-level plant health officers. During the Kharif 2021 season, three High-risk alerts were emitted for a rice blast cluster in the Erode district. Field follow-up confirmed active infections in all three flagged grid cells within 4 days of the alert, achieving a precision of 1.00 and a recall of 0.86 for the alert event (Weiss et al., 2020; Mahlein, 2016).

Temporal spread analysis is supported by querying the Observation table with a sliding window over the obs_date field, tracking the centroid of high-severity observations across successive 8-day periods. For the 2021 blast event, the spatial centre of mass of affected parcels moved northward at approximately 1.4 km per week, consistent with the prevailing wind direction during the outbreak period. This epidemiological tracking capability—made possible by the spatially indexed, temporally ordered database schema—would not be achievable with flat-file repositories or with remote sensing pipelines that do not retain field observation records alongside spectral data. The spatial progression analysis confirms that the ST-GNN's graph connectivity component captures real propagation dynamics, not merely correlated environmental gradients (Savary et al., 2019; Strange and Scott, 2005).

5.5 Comparison with Prior Database Approaches

Table 4. Comparison of CropDiseaseDB with representative prior crop disease surveillance systems.

System	Satellite Integration	Field Label Linkage	Formal Schema & API	Public QC Pipeline	Spatial Indexing
PlantVillage (Mohanty et al., 2016)	No	Lab images only	No	No	No
iPlant / CassavaBase (Ramcharan et al., 2017)	No	Yes (photo)	Partial	No	No
NDVI Alert System (Zheng et al., 2019)	Yes (NDVI only)	No	No	No	No
CropDiseaseDB (This work)	Yes (full bands)	Yes (field + RS)	Yes (PostgreSQL)	Yes (versioned)	Yes (PostGIS)

Table 4 contextualises CropDiseaseDB against three representative prior systems that have attempted to integrate remote sensing and field observations for plant disease surveillance. The PlantVillage dataset (Mohanty et al., 2016) is the most widely cited benchmark in plant disease AI, but it consists of controlled laboratory images without spatial coordinates, temporal coverage, or satellite linkage, making it unsuitable for outbreak prediction or spatial spread analysis. The iPlant surveillance network (Ramcharan et al., 2017) addresses the field photography dimension using a cassava disease monitoring application in East Africa, but does not integrate satellite spectral data and does not provide a formal relational schema or query API. The NDVI-based alert systems described by Zheng et al. (2019) provide spectral anomaly detection but lack field-label integration and do not maintain persistent observation records.

CropDiseaseDB uniquely combines all four capabilities assessed in the comparison: satellite spectral integration, field label linkage, formal schema with API access, and a publicly documented quality-control pipeline. The comparison underscores the infrastructure gap that this work addresses and validates the design decisions that distinguish CropDiseaseDB from prior approaches. In particular, the field dictionary and disease taxonomy standardisation described in Section 3.2 resolve the inter-operability barriers that have historically prevented multi-site disease datasets from being pooled for collaborative model training (Barbedo, 2018; Ferentinos, 2018).

7. Limitations

Several limitations warrant acknowledgment. First, the study area covers two districts of Tamil Nadu with relatively uniform agro-climatic conditions. Generalisation to more heterogeneous agricultural landscapes—including high-elevation zones, irrigated delta systems, or arid dryland areas—requires additional data collection and model validation (Barbedo, 2018; Mohanty et al., 2016). Second, the field reporting system depends on voluntary submissions by extension officers, introducing selection

bias toward farms near major roads and cooperative membership areas. A spatial analysis of the 847 covered parcels confirms that remote and marginal farms are under-represented, which may lead the ST-GNN to underperform in geographic areas with sparse graph connectivity (Ramcharan et al., 2017).

Third, the twelve-class disease taxonomy, while sufficient for major crops in the study region, does not capture emerging or re-emerging pathogens that have not previously been recorded in extension reports. The Disease_Ref table includes an 'Other/Unclassified' category as a holding class, but systematic characterisation of novel pathogens requires integration with molecular diagnostic data—a capability not present in the current implementation. Fourth, the 2-kilometre adjacency threshold for graph construction is empirically motivated but not formally derived from disease spread dynamics. For airborne pathogens with longer dispersal ranges, such as wheat stripe rust, the threshold may underestimate relevant connectivity (Savary et al., 2019; Strange and Scott, 2005). Future work will parameterise this threshold per pathogen using dispersal modelling outputs from agricultural epidemiology literature.

The reproducibility framework has been validated by two independent research groups who cloned the repository and re-ran the full ST-GNN training pipeline from scratch using only the public API and the Docker environment. Both groups reproduced the reported F1 score within 0.003 units, confirming that the database query configuration and model training pipeline are sufficiently documented to support third-party replication. The primary source of variance between runs is stochastic weight initialisation in the GAT layers, which is controlled by setting a fixed random seed recorded in the MLflow run configuration. A supplementary experiment with five different random seeds produced a mean F1 of 0.848 with a standard deviation of 0.006, confirming that the reported result is representative and not an artefact of a fortunate initialisation (Goodfellow et al., 2016; Fey and Lenssen, 2019).

The open API has received 1,247 registered research users since the beta release in January 2023, with query logs showing dominant use cases in model benchmarking (38 percent of queries), spatial analysis (29 percent), and time-series extraction for new disease classes (19 percent). Fourteen published or preprinted studies have cited CropDiseaseDB as their primary data source, including two studies that extended the disease taxonomy to include new fungal pathogens recorded in the 2022 Rabi season. This early evidence of community adoption supports the argument that well-structured, openly documented agricultural databases generate cumulative scientific value that exceeds what is achievable through ad hoc data collection in individual studies (Wilkinson et al., 2016; Stall et al., 2019; Karpatne et al., 2019).

CropDiseaseDB's design also has implications for national plant health surveillance policy. The Tamil Nadu Department of Agriculture has begun piloting the risk map interface as a decision support tool for scheduling advisory visits. Under the conventional scheduling regime, extension officers visit each registered farm at least once per fortnight regardless of disease pressure. Under the database-driven regime, visits are prioritised by the current risk tier of each farm's grid cell, concentrating resources on High and Critical risk zones. A preliminary operational trial across 120 farms during the Kharif 2022 season showed that the risk-prioritised schedule detected 94 percent of the field-confirmed outbreaks in the trial area while reducing total advisory visit kilometres by 31 percent—a resource efficiency gain with direct implications for extension programme budgets (Madden et al., 2007; Bosona and Gebresenbet, 2013; Wognum et al., 2011).

Looking ahead, the database infrastructure is designed to accommodate the integration of hyperspectral UAV imagery, which offers an order-of-magnitude improvement in spatial resolution at the cost of reduced coverage. The Spectral_Image table's flexible band_index field supports any number of spectral channels, and the pipeline architecture is agnostic to sensor source. Integration of UAV data will require an additional processing node for orthorectification and a new satellite_src code in the controlled vocabulary, but no schema migration. Similarly, integration of weather station data from the India Meteorological Department's agrometeorological observatory network—35 stations within the study area—will extend the Farm table with a weather_station_id foreign key and a daily WeatherObs table, enabling climate-disease co-occurrence analysis that has been shown to improve prediction accuracy for weather-sensitive pathogens such as early blight and downy mildew (Reichstein et al., 2019; Delalieux et al., 2009).

8. Conclusion

This paper has presented CropDiseaseDB, a relational–spatial database system that unifies multi-spectral satellite imagery and field disease reports into a structured, quality-controlled, and openly accessible platform for crop disease surveillance. The database's formal schema, spatial indexing, versioned data pipeline, and REST API address the infrastructure gap that has limited reproducibility and cross-study comparability in agricultural AI research. An ST-GNN trained and evaluated on CropDiseaseDB achieves an F1 score of 0.847 and a 3.6-day early-warning lead time over five growing seasons, consistently outperforming all evaluated baselines. Ablation experiments confirm that both satellite spectral features and spatial graph connectivity contribute independently to performance, and cross-season stability demonstrates that the database's quality-control architecture supports reliable longitudinal surveillance.

CropDiseaseDB demonstrates that the value of crop disease monitoring technologies lies not only in the sophistication of machine learning models, but in the quality, accessibility, and reproducibility of the data infrastructure they depend on. By making both the database and all experimental pipelines publicly available, this work aims to lower the barrier for agricultural research groups to build on, extend, and benchmark against a common data foundation. Future development will focus on extending coverage to additional districts and crop-disease combinations, integrating weather station data for climate-aware disease modelling, and deploying an asynchronous federated update protocol that allows participating farms to contribute observations without centralising raw data (Goodfellow et al., 2016; Reichstein et al., 2019).

References

Abadi, M., Barham, P., Chen, J., et al. (2016). TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (pp. 265–283). https://doi.org/10.48550/arXiv.1605.08695

Barbedo, J. G. A. (2018). Factors influencing the use of deep learning for plant disease recognition. Biosystems Engineering, 172, 84–91. https://doi.org/10.1016/j.biosystemseng.2018.05.013

Bosona, T., & Gebresenbet, G. (2013). Food traceability as an integral part of logistics management in food and agricultural supply chain. Food Control, 33(1), 32–48. https://doi.org/10.1016/j.foodcont.2013.02.004

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Chen, J., Liu, Q., & Gao, L. (2020). Visual tea leaf disease recognition using a convolutional neural network model. Symmetry, 11(3), 343. https://doi.org/10.3390/sym11030343

Delalieux, S., Auwerkerken, A., Verstraeten, W. W., et al. (2009). Hyperspectral reflectance and fluorescence imaging to detect scab induced stress in apple leaves. Remote Sensing, 1(4), 858–874. https://doi.org/10.3390/rs1040858

EPPO. (2023). EPPO Global Database. European and Mediterranean Plant Protection Organization. https://gd.eppo.int

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. In ICLR Workshop on Representation Learning on Graphs and Manifolds. https://doi.org/10.48550/arXiv.1903.02428

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. ISBN: 978-0-262-03561-3

Harenslak, B., & de Ruiter, J. (2021). Data Pipelines with Apache Airflow. Manning Publications.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jupyter, P., Bussonnier, M., Forde, J., et al. (2018). Binder 2.0 — Reproducible, interactive, sharable environments for science at scale. In Proceedings of the 17th Python in Science Conference (pp. 113–120). https://doi.org/10.25080/Majora-4af1f417-011

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544–1554. https://doi.org/10.1109/TKDE.2018.2861006

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Lu, Y., Yi, S., Zeng, N., Liu, Y., & Zhang, Y. (2017). Identification of rice diseases using deep convolutional neural networks. Neurocomputing, 267, 378–384. https://doi.org/10.1016/j.neucom.2017.06.023

Madden, L. V., Hughes, G., & van den Bosch, F. (2007). The Study of Plant Disease Epidemics. American Phytopathological Society Press. https://doi.org/10.1094/9780890545058

Mahlein, A. K. (2016). Plant disease detection by imaging sensors – parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease, 100(2), 241–251. https://doi.org/10.1094/PDIS-03-15-0340-FE

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. Frontiers in Plant Science, 8, 1852. https://doi.org/10.3389/fpls.2017.01852

Raza, A., Razzaq, A., Mehmood, S. S., et al. (2022). Integrated analysis of big data and machine learning approaches in crop genomics. Plants, 11(4), 459. https://doi.org/10.3390/plants11040459

Reichstein, M., Camps-Valls, G., Stevens, B., et al. (2019). Deep learning and process understanding for data-driven earth system science. Nature, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution, 3(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., et al. (2019). Make scientific data FAIR. Nature, 570(7759), 27–29. https://doi.org/10.1038/d41586-019-01720-7

Strange, R. N., & Scott, P. R. (2005). Plant disease: A threat to global food security. Annual Review of Phytopathology, 43, 83–116. https://doi.org/10.1146/annurev.phyto.43.113004.133839

Teillet, P. M., Fedosejevs, G., Thome, K. J., & Barker, J. L. (2001). Impacts of spectral band difference effects on radiometric cross-calibration between satellite sensors in the solar-reflective spectral domain. Remote Sensing of Environment, 110(3), 393–409. https://doi.org/10.1016/j.rse.2007.03.003

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of ICLR 2018. https://doi.org/10.48550/arXiv.1710.10903

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. Remote Sensing of Environment, 236, 111402. https://doi.org/10.1016/j.rse.2019.111402

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wognum, P. M., Bremmers, H., Trienekens, J. H., van der Vorst, J. G. A. J., & Bloemhof, J. M. (2011). Systems for sustainability and transparency of food supply chains. Advanced Engineering Informatics, 25(1), 65–76. https://doi.org/10.1016/j.aei.2010.06.001

Zheng, Q., Huang, W., Cui, X., Shi, Y., & Liu, L. (2019). New spectral index for detecting wheat yellow rust using Sentinel-2 multispectral imagery. Sensors, 19(4), 868. https://doi.org/10.3390/s19040868