frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 May 2025
DOI 10.3389/fpls.2025.1576756

OPEN ACCESS

EDITED BY
Jun Ni,
Nanjing Agricultural University, China

REVIEWED BY
Olumide Falana,
Obafemi Awolowo University, Nigeria
Balaji V R,
Sri Krishna College of Engineering & Technology, India
Danish Gul,
Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, India

*CORRESPONDENCE
Cheng Wei
chw365@163.com

RECEIVED 14 February 2025
ACCEPTED 07 April 2025
PUBLISHED 14 May 2025

CITATION
Wei C, Shan Y and Zhen M (2025) Deep learning-based anomaly detection for precision field crop protection.
Front. Plant Sci. 16:1576756.
doi: 10.3389/fpls.2025.1576756

Deep learning-based anomaly detection for precision field crop protection

Cheng Wei[1]*, Yifeng Shan[2] and MengZhe Zhen[3]

[1]Chongqing Business Vocational College, Chongqing, China, [2]Ningbo University of Finance and Economics, School of Basic Education, Ningbo, Zhejiang, China, [3]Ningbo University of Finance and Economics, School of Digital Technology and Engineering, Ningbo, Zhejiang, China

Introduction: Precision agriculture relies on advanced technologies to optimize crop protection and resource utilization, ensuring sustainable and efficient farming practices. Anomaly detection plays a critical role in identifying and addressing irregularities, such as pest outbreaks, disease spread, or nutrient deficiencies, that can negatively impact yield. Traditional methods struggle with the complexity and variability of agricultural data collected from diverse sources.

Methods: To address these challenges, we propose a novel framework that integrates the Integrated Multi-Modal Smart Farming Network (IMSFNet) with the Adaptive Resource Optimization Strategy (AROS). IMSFNet employs multimodal data fusion and spatiotemporal modeling to provide accurate predictions of crop health and yield anomalies by leveraging data from UAVs, satellites, ground sensors, and weather stations. AROS dynamically optimizes resource allocation based on real-time environmental feedback and multi-objective optimization, balancing yield maximization, cost efficiency, and environmental sustainability.

Results: Experimental evaluations demonstrate the effectiveness of our approach in detecting anomalies and improving decision-making in precision agriculture.

Discussion: This framework sets a new standard for sustainable and data-driven crop protection strategies.

KEYWORDS

precision agriculture, anomaly detection, multi-modal data fusion, resource optimization, sustainable farming

1 Introduction

Precision agriculture has revolutionized the agricultural industry, enabling efficient and sustainable crop management through targeted interventions. Within this domain, anomaly detection plays a critical role in field crop protection by identifying early signs of diseases, pests, nutrient deficiencies, or other stress factors that compromise crop health

Zhang et al. (2024b). Not only does early detection help reduce the overuse of chemical inputs such as pesticides and fertilizers, but it also minimizes yield losses and supports sustainable farming practices Gao et al. (2024a). Traditional anomaly detection techniques, while effective in controlled conditions, often fail to capture the complexity of real-world agricultural systems, which involve heterogeneous data sources, dynamic environmental conditions, and intricate interactions between crops and external stressors Zhu et al. (2024a). As a result, deep learning-based methods have emerged as a promising solution for improving the accuracy and scalability of anomaly detection in precision field crop protection. By leveraging multi-modal data from satellite imagery, drones, and on-ground sensors, these methods provide an end-to-end framework for identifying and addressing crop health anomalies Li et al. (2021).

Early approaches to anomaly detection in agriculture were based on rule-based systems and statistical models, which relied on domain knowledge and handcrafted features to identify deviations from normal conditions Zavrtanik et al. (2021). For example, threshold-based methods used predefined values for parameters like vegetation indices or soil moisture levels to detect anomalies Deng and Hooi (2021). Similarly, statistical models such as principal component analysis (PCA) and clustering were used to identify outliers in crop health data Zou et al. (2022). While these methods provided interpretable results and were computationally efficient, they lacked the ability to generalize across diverse environmental conditions and crop types Bergmann et al. (2021). Furthermore, their reliance on fixed thresholds and handcrafted features made them inadequate for capturing the complex, non-linear patterns associated with crop health anomalies caused by diseases or pests Gudovskiy et al. (2021).

The shift toward data-driven approaches introduced machine learning algorithms capable of learning patterns from historical data to improve anomaly detection You et al. (2022). Techniques such as support vector machines (SVMs), random forests, and k-nearest neighbors were employed to classify crop health statuses based on features extracted from remote sensing or field data Liu et al. (2023a). These models achieved better adaptability compared to rule-based systems, as they could learn relationships between input features and anomalies without explicit thresholds. For example, machine learning models were applied to identify plant stress from hyperspectral imagery or to classify pest infestations based on soil and weather data Zhu et al. (2024b). However, these approaches faced limitations in their ability to handle highdimensional and multi-modal data, such as the integration of spectral, spatial, and temporal information. Traditional machine learning methods required extensive feature engineering and struggled to generalize to new datasets or unseen conditions Tian et al. (2021).

Deep learning has transformed anomaly detection in precision agriculture by introducing architectures that can automatically learn hierarchical representations of crop health data Han et al. (2022). Convolutional neural networks (CNNs) have been widely used for analyzing spatial patterns in satellite and drone imagery, enabling the detection of disease outbreaks, pest infestations, and nutrient deficiencies. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have been employed to model temporal changes in crop health, such as monitoring vegetation growth over a growing season Jiang et al. (2023). Furthermore, generative adversarial networks (GANs) and autoencoders have demonstrated success in unsupervised anomaly detection, where models are trained on normal data to identify deviations that signify anomalies Zhang et al. (2024c). For example, autoencoders have been used to detect stress signals in hyperspectral images by reconstructing healthy crop patterns and flagging deviations as anomalies Tien et al. (2023). Despite these advancements, challenges remain, particularly in the integration of multi-modal data from diverse sources, the interpretability of deep learning models, and the scalability of these methods for large-scale agricultural applications Wyatt et al. (2022).

Recent works in spatiotemporal modeling for agricultural anomaly detection primarily rely on CNNs and LSTMs to process spatial and sequential data, respectively. For example, proposed a hybrid CNN-LSTM model Xu et al. (2021) to capture vegetation dynamics from satellite images, but their approach struggled with long-range dependencies and multimodal data integration. Similarly, introduced an attention-enhanced LSTM Tuli et al. (2022) for pest detection, yet the model's performance degraded with increasing data heterogeneity. In contrast, our proposed IMSFNet incorporates GNNs to model spatial relationships between crop regions, while employing Transformers to capture long-term temporal dependencies more effectively than LSTMs. This hybrid architecture allows for adaptive weighting of multimodal information, improving the robustness and interpretability of anomaly detection in precision agriculture.

To address these challenges, we propose a novel deep learning-based framework for anomaly detection tailored for precision field crop protection. The proposed framework integrates multi-modal data from satellite imagery, drones, and *in-situ* sensors to capture a holistic view of crop health. A hybrid architecture combining convolutional neural networks and graph neural networks (GNNs) is employed to model spatial dependencies between crops, while transformers are utilized to capture temporal patterns in crop growth and environmental conditions. The framework incorporates unsupervised learning techniques, such as variational autoencoders, to detect subtle anomalies that may not be present in labeled datasets. By prioritizing scalability and interpretability, this framework is designed to support real-time decision-making and adaptive interventions, ultimately enhancing crop resilience and productivity.

We summarize our contributions as follows:

- We propose IMSFNet, a novel Integrated Multi-Modal Smart Farming Network, which combines GNNs for spatial dependency modeling and Transformers for long-term temporal feature extraction. This hybrid architecture effectively captures crop health variations and environmental anomalies across different spatial and temporal scales.
- We introduce a multi-modal fusion strategy that integrates satellite imagery, UAV-based imaging, ground sensors, and meteorological data. Unlike previous works that rely on independent processing pipelines, IMSFNet jointly learns representations from heterogeneous data sources, leading to improved anomaly detection performance.

- We develop AROS (Adaptive Resource Optimization Strategy), a real-time optimization framework that dynamically adjusts resource allocation based on multi-objective optimization. AROS leverages reinforcement learning-based feedback mechanisms to improve efficiency and sustainability in precision agriculture.
- Our extensive experiments on multiple real-world datasets demonstrate that IMSFNet and AROS achieve 3.12% higher F1-score and 2.89% higher accuracy compared to state-of-the-art anomaly detection models. We release our implementation and dataset annotations to facilitate further research in deep learning-based precision agriculture.

2 Related work

2.1 Deep learning for anomaly detection in agriculture

Anomaly detection is a critical component of precision agriculture, aiming to identify abnormal patterns in crop health, pest infestations, and environmental conditions Wang et al. (2023a). Traditional anomaly detection methods, such as threshold-based techniques and classical machine learning models like Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN), often fail to capture the complexity of agricultural environments Wang et al. (2023b). Deep learning approaches have emerged as more robust alternatives, leveraging their ability to process large-scale, high-dimensional data and uncover complex, nonlinear relationships Defard et al. (2020). In agricultural applications, CNNs are widely employed to analyze visual data, such as images from drones and ground-based cameras Gao et al. (2024b). For instance, CNN-based models can detect visual anomalies in crops, such as discoloration, irregular growth, and pest damage Park et al. (2020). RNNs and LSTM networks, on the other hand, are applied to sequential data, such as time-series environmental sensor readings, to identify trends or deviations indicative of anomalies DeMedeiros et al. (2023). Recent advancements include hybrid models that integrate CNNs and RNNs to analyze spatiotemporal data, such as videos of crop fields over time Batzner et al. (2023). These models enable the detection of anomalies not just at a single point in time but also in evolving patterns, such as the spread of a disease across a field Feng et al. (2021). While these methods show promise, challenges remain in dealing with noisy and imbalanced data, where anomalies constitute only a small fraction of the dataset. Techniques such as data augmentation, synthetic anomaly generation, and adversarial training are being developed to address these issues and enhance model robustness in real-world scenarios Audibert et al. (2020).

2.2 Multi-modal data fusion for precision crop protection

Precision crop protection often requires integrating data from multiple sources, such as remote sensing, IoT devices, and weather stations, to effectively detect and respond to anomalies Le and Zhang (2021). Multi-modal data fusion enables the combination of these diverse data types to improve the accuracy and reliability of anomaly detection systems Liu et al. (2021). Deep learning has been instrumental in facilitating such fusion, with models that can process and integrate heterogeneous data modalities. For visual data, drone and satellite imagery are commonly used to monitor crop health, while IoT devices provide real-time sensor readings for soil moisture, temperature, and humidity Salehi et al. (2020). Multi-modal architectures, such as those combining CNNs for image analysis with fully connected or Transformer layers for sensor data, have demonstrated improved performance in identifying anomalies such as nutrient deficiencies, water stress, and pest outbreaks Liu et al. (2023b). For example, attention mechanisms have been employed to prioritize the most relevant data sources for decision-making, enhancing the interpretability of these systems. Temporal fusion models, such as Temporal Fusion Transformers (TFTs), have also been applied to integrate time-series data from multiple sensors with historical climate records, enabling more accurate anomaly predictions Roth et al. (2021). GNNs are another emerging approach for multi-modal fusion, particularly in representing spatial relationships within a field, such as proximity between affected regions or the spread of an anomaly across neighboring plots. Despite the benefits, multi-modal fusion faces challenges such as data heterogeneity, missing values, and high computational requirements. Advances in self-supervised learning and imputation techniques are being explored to address these limitations, enabling models to learn meaningful representations from incomplete or noisy datasets Deng and Li (2022). Furthermore, edge computing and hardware acceleration are being investigated to enable real-time data fusion and anomaly detection in resource-constrained agricultural environments.

2.3 Applications of anomaly detection in crop protection

Deep learning-based anomaly detection has found numerous applications in precision crop protection, addressing challenges such as pest infestations, disease outbreaks, and abiotic stress factors like drought and frost Ni et al. (2018). By automating the identification of anomalies, these systems reduce the reliance on manual scouting, which is labor-intensive and prone to errors, especially in large-scale agricultural settings Ni et al. (2017). For pest detection, models leveraging CNNs and object detection frameworks like YOLO (You Only Look Once) have been used to identify specific pest species in field images. These models enable targeted interventions, such as pesticide application, reducing chemical usage and minimizing environmental impact Mishra et al. (2021). Similarly, for disease detection, segmentation models such as U-Net and Mask R-CNN have been employed to localize affected areas, allowing for precise treatment. Beyond visual data, deep learning models have been applied to sensor-based anomaly detection. For example, LSTM networks are used to analyze soil moisture and temperature data to identify water stress, while

Variational Autoencoders (VAEs) Ni et al. (2016) and GANs are employed to detect deviations from normal patterns in multi-dimensional sensor data. These approaches are particularly effective in identifying early warning signs of crop stress, enabling timely interventions that mitigate yield losses. Another application lies in predicting the spatial spread of anomalies, such as the propagation of pests or diseases within a field. Spatiotemporal models, including 3D CNNs and ST-GCNs (Spatio-Temporal Graph Convolutional Networks), are used to predict how anomalies evolve over time and space, aiding in the design of containment strategies. For instance, these models can simulate the effects of varying weather conditions on the spread of an anomaly, providing actionable insights for farmers and agronomists. While these applications demonstrate the potential of anomaly detection in crop protection, practical deployment remains challenging due to factors such as limited labeled data, variability in agricultural environments, and the need for domain-specific customization. Future research should focus on developing scalable and generalizable models, incorporating domain knowledge into deep learning frameworks, and ensuring the ethical use of these technologies in precision agriculture.

3 Method

3.1 Overview

Precision agriculture, also referred to as smart farming or digital farming, is an advanced approach to agricultural management that leverages modern technologies such as remote sensing, geographic information systems (GIS), Internet of Things (IoT), and artificial intelligence (AI) to optimize crop production and resource utilization. The goal of precision agriculture is to enhance efficiency, reduce environmental impact, and increase yield through the precise monitoring and management of agricultural inputs, including water, fertilizers, and pesticides. Traditional farming practices often involve uniform treatment of large agricultural fields, which can lead to inefficiencies and overuse of resources. By contrast, precision agriculture adopts a data-driven approach, tailoring interventions to the specific needs of crops, soil conditions, and environmental factors. This is achieved through a combination of advanced technologies that collect, analyze, and act upon data at a granular level. For instance, sensors embedded in the soil can measure moisture and nutrient levels, drones equipped with multispectral cameras can capture crop health data, and machine learning algorithms can predict optimal planting and harvesting times. The implementation of precision agriculture can be broadly categorized into three main components: data collection, data analysis, and decision-making. Data collection involves the use of various sensing technologies, such as satellite imagery, UAVs (unmanned aerial vehicles), and IoT-enabled sensors, to capture real-time information about crops and the environment. This data is then analyzed using advanced computational techniques, including machine learning and statistical modeling, to derive actionable insights. The insights are used to make informed decisions, such as variable-rate application of fertilizers or automated irrigation scheduling. Despite its promising potential, the widespread adoption of precision agriculture faces several challenges, including the high cost of technology, lack of technical expertise among farmers, and limited access to high-speed internet in rural areas. Addressing these barriers requires interdisciplinary efforts that integrate engineering, computer science, agronomy, and socioeconomics.

This paper introduces a novel framework for precision agriculture that integrates advanced sensing technologies with deep learning-based predictive models to enhance the scalability and robustness of smart farming systems. The proposed method focuses on multi-modal data fusion, combining visual, thermal, and spectral data from UAVs and ground-based sensors to improve the accuracy of crop health assessment and yield prediction. The framework also incorporates adaptive optimization strategies to address varying environmental and climatic conditions, ensuring its applicability across diverse agricultural contexts. The remainder of this paper is organized as follows. In Section 3.2, we formalize the precision agriculture problem, introducing the necessary mathematical foundations and notations. Section 3.3 introduces our proposed Integrated Multi-Modal Smart Farming Network (IMSFNet), a system that integrates advanced sensing and modeling techniques for efficient crop monitoring. In Section 3.4, we present the Adaptive Resource Optimization Strategy (AROS), a novel approach designed to optimize resource allocation through real-time predictions and environmental feedback.

To enhance readability, we briefly summarize the key technical abbreviations (In Table 1) used throughout the manuscript. IMSFNet refers to the proposed Integrated Multi-Modal Smart Farming Network, while AROS denotes the Adaptive Resource Optimization Strategy. Core components include MFE (Multi-Modal Feature Extraction), CBS (Convolution + Batch Normalization + SiLU), CSPBlock (Cross Stage Partial Block), and PConv (Parametric Convolution). FFCA-YOLO and L-FFCA-YOLO represent backbone architectures with cross-attentive and lightweight designs. Additional modules such as SCNN (Spatial Convolutional Neural Network), DMO (Dynamic Multi-Objective Optimization), PRA (Prioritized Resource Allocation), and RFM (Real-Time Feedback Mechanism) further support resource-efficient anomaly detection. Abbreviations like UAV (Unmanned Aerial Vehicle), NDVI (Normalized Difference Vegetation Index), GNN (Graph Neural Network), and LSTM (Long Short-Term Memory) are used to represent standard sensing and modeling technologies in precision agriculture.

3.2 Preliminaries

Precision agriculture relies on detecting anomalies in crop health to optimize resource allocation. Given an agricultural field F partitioned into N management zones $\{Z_1, Z_2 \ldots, Z_N\}$, each zone is represented by a feature vector $x_i$derived from multi-modal data sources. The objective of anomaly detection is to identify zones where the observed feature vector $x_i$deviates significantly from

TABLE 1 List of technical abbreviations used throughout the manuscript.

Abbreviation	Full Term
IMSFNet	Integrated Multi-Modal Smart Farming Network
AROS	Adaptive Resource Optimization Strategy
MFE	Multi-Modal Feature Extraction
CBS	Convolution + Batch Normalization + SiLU
CSPBlock	Cross Stage Partial Block
CSPFaster Block	Enhanced Cross Stage Partial Block (optimized for speed)
PConv	Parametric Convolution
FFCA-YOLO	Feature-Focused Cross-Attentive YOLO
L-FFCA-YOLO	Lightweight FFCA-YOLO
SCNN	Spatial Convolutional Neural Network
DMO	Dynamic Multi-Objective Optimization
PRA	Prioritized Resource Allocation
RFM	Real-Time Feedback Mechanism
NDVI	Normalized Difference Vegetation Index
IoT	Internet of Things
UAV	Unmanned Aerial Vehicle
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
GNN	Graph Neural Network
CNN	Convolutional Neural Network
TFT	Temporal Fusion Transformer
Q, K, V	Query, Key, Value (in attention mechanism)

normal patterns. Let $p(x)$ represent the probability distribution of normal crop health conditions. A zone $Z_i$is classified as anomalous if its feature vector satisfies (Equation 1):

$$\mathbb{P}(x_i|\theta) < \tau, \tag{1}$$

where θ denotes the parameters of the learned normal distribution, and τ is a predefined anomaly threshold. To quantify anomalies, an anomaly score function s_i is defined based on the deviation of x_i from the mean feature vector μ of normal samples (Equation 2):

$$s_i = D(x_i, \mu), \tag{2}$$

where $D(\cdot,\cdot)$ represents a distance metric such as the Mahalanobis distance or Euclidean distance. Higher values of $s_i$indicate stronger deviations from normal conditions, guiding adaptive interventions in precision agriculture.

The optimization objective can be formulated as (Equation 3):

$$\max_{\mathcal{U}} \sum_{i=1}^{N} \mathcal{Y}_i(\mathcal{U}_i, \mathbf{p}_i) - \lambda \mathcal{C}(\mathcal{U}), \tag{3}$$

where $\mathcal{Y}_i$ is the yield function for zone Z_i, which depends on the intervention $\mathcal{U}_i$ and the feature vector $\mathbf{p}_i$, and $\mathcal{C}(\mathcal{U})$ is the total cost function associated with the intervention strategy $\mathcal{U}$. The parameter λ is a weighting factor that balances yield maximization and cost minimization.

Agricultural fields exhibit both spatial and temporal variability. Spatial variability arises from heterogeneities in soil properties, topography, and crop health across zones. Temporal variability is driven by dynamic environmental conditions, such as weather changes and seasonal cycles. Let $\mathbf{P}(t) = \{\mathbf{p}_1(t), \mathbf{p}_2(t),\ldots, \mathbf{p}_N(t)\}$ denote the time-dependent feature matrix for the field at time t. The evolution of crop yield $\mathcal{Y}_i(t)$ for zone Z_i can be modeled as (Equation 4):

$$\mathcal{Y}_i(t) = \mathcal{F}(\mathbf{p}_i(t), \mathcal{U}_i(t), \eta_i(t)), \tag{4}$$

where $\mathcal{F}(\cdot)$ is a function that encapsulates the complex relationships between environmental factors, interventions, and crop response, and $\boldsymbol{\eta}_i(t)$ represents stochastic disturbances such as pests, diseases, or unpredicted weather events.

Precision agriculture integrates diverse types of multimodal data collected from various sensing technologies to effectively monitor crop and environmental conditions. Satellite imagery enables large-scale observation by providing vegetation indices, such as the Normalized Difference Vegetation Index (NDVI), and information about canopy cover. Unmanned aerial vehicles (UAVs), commonly referred to as drones, capture high-resolution visual, thermal, and multispectral images, offering detailed insights into crop health. Ground sensors contribute precise measurements of soil properties, including moisture, temperature, pH, and electrical conductivity, at specific locations within the field. Weather stations supply real-time data on environmental variables, including temperature, humidity, wind speed, and precipitation, further enhancing the decision-making process in precision agriculture.

Let $\mathcal{D} = \{\mathcal{D}_{\text{sat}}, \mathcal{D}_{\text{uav}}, \mathcal{D}_{\text{ground}}, \mathcal{D}_{\text{weather}}\}$ represent the collection of data from these sources. The fusion of multi-modal data is essential for developing a comprehensive understanding of field conditions.

In precision agriculture, decision-making often involves multiple competing objectives, such as maximizing yield, minimizing resource usage, and reducing environmental impact. The problem can be formulated as a multi-objective optimization (Equation 5):

$$\textit{extmaximize} \quad [\mathcal{Y}(\mathcal{U}), -\mathcal{C}(\mathcal{U}), -\mathcal{E}(\mathcal{U})], \tag{5}$$

where $\mathcal{E}(\mathcal{U})$ represents the environmental impact function. Multi-objective optimization techniques, such as Pareto front analysis or weighted sum methods, are employed to identify trade-offs and select optimal strategies.

3.3 Integrated multi-modal smart farming network

IMSFNet is an innovative deep learning framework designed for precision agriculture, seamlessly integrating multi-modal data sources to perform anomaly detection in a unified manner. Unlike conventional CNN-LSTM architectures, it leverages Graph Neural Networks to enhance spatial feature learning while incorporating Transformers to capture long-range temporal dependencies. This combination significantly improves the model's ability to understand complex environmental interactions. The framework operates through a structured process that begins with extracting features from multiple data modalities, followed by an advanced fusion mechanism that integrates spatial and temporal information. It introduces a cross-modal interaction approach that strengthens the relationships between different data sources, ensuring a more comprehensive and accurate analysis of agricultural conditions.

IMSFNet integrates multiple data sources to achieve precise crop anomaly detection through multi-modal fusion, as illustrated in Figure 1. Each modality contributes a distinct feature set, including satellite imagery features represented as $X_{\text{sat}} \in \mathbb{R}^{H_s \times W_s \times d_s}$, UAV imagery features as $X_{\text{uav}} \in \mathbb{R}^{H_u \times W_u \times d_u}$, ground sensor features as $X_{\text{sens}} \in \mathbb{R}^{N_g \times d_g}$, and weather data as $X_{\text{weather}} \in \mathbb{R}^{T_w \times d_w}$. Each modality undergoes feature extraction through a modality-specific encoder E_m, transforming the input into a latent representation $F_m = E_m(X_m)$ for all modalities, including satellite, UAV, sensor, and weather data. To ensure consistency across different modalities, all extracted features are projected into a common latent space $\mathbb{R}^d$ through modality-specific projection functions, expressed as $F_{\text{proj},m} = P_m(F_m)$. IMSFNet incorporates an attentionbased mechanism that dynamically assigns weights to each modality, enhancing the contribution of the most informative features. The final fused representation is computed as $F_{\text{fused}} = \sum_m w_m F_m$, where the weights $w_m$are determined using a softmax function applied to a learnable weight matrix W_m, formulated as $w_m = \text{softmax}(W_m F_m)$. This fusion strategy enables IMSFNet to effectively integrate diverse agricultural data sources, leading to improved performance in crop anomaly detection.

3.3.1 Multi-modal feature extraction

IMSFNet employs a robust multi-modal feature extraction process tailored to handle the diverse nature of input data sources, including satellite imagery, UAV-based imaging, ground sensors, and weather data. These data modalities provide complementary information critical for precision agriculture, capturing spatial, temporal, and environmental variability. Formally, let $\mathcal{D} = \{\mathcal{D}_{\text{sat}}, \mathcal{D}_{\text{uav}}, \mathcal{D}_{\text{ground}}, \mathcal{D}_{\text{weather}}\}$ denote the collection of input datasets. Each data source is independently processed through modalityspecific feature extractors $\mathcal{E}_{\text{sat}}, \mathcal{E}_{\text{uav}}$, $\mathcal{E}_{\text{ground}}, \mathcal{E}_{\text{weather}}$ to generate low-dimensional feature embeddings $\mathbf{F}_{\text{sat}}$, $\mathbf{F}_{\text{uav}}$, $\mathbf{F}_{\text{ground}}$, $\mathbf{F}_{\text{weather}}$. The extraction process can be formalized as follows (Equations 6, 7):

$$\mathbf{F}_{\text{sat}} = \mathcal{E}_{\text{sat}}(\mathcal{D}_{\text{sat}}), \quad \mathbf{F}_{\text{uav}} = \mathcal{E}_{\text{uav}}(\mathcal{D}_{\text{uav}}), \tag{6}$$

$$\mathbf{F}_{\text{ground}} = \mathcal{E}_{\text{ground}}(\mathcal{D}_{\text{ground}}), \quad \mathbf{F}_{\text{weather}} = \mathcal{E}_{\text{weather}}(\mathcal{D}_{\text{weather}}). \tag{7}$$

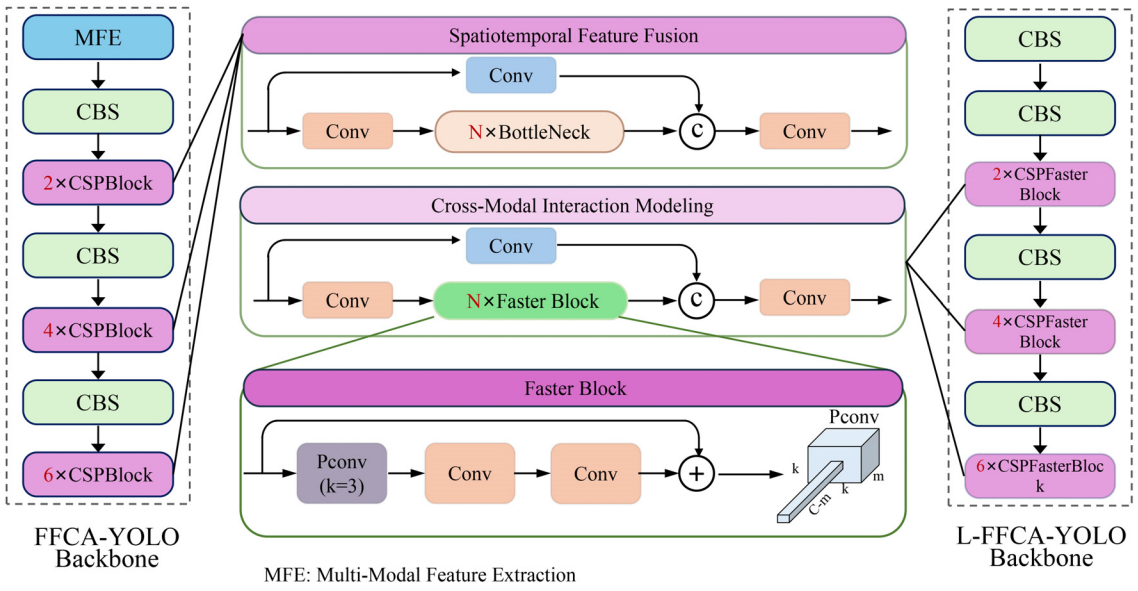

FIGURE 1
Overview of integrated multi-modal smart farming network architectures. The figure illustrates the structural components of the FFCA-YOLO and L-FFCA-YOLO backbones, highlighting key modules such as Multi-Modal Feature Extraction (MFE), Spatiotemporal Feature Fusion, Cross-Modal Interaction Modeling, and Faster Blocks. The FFCA-YOLO backbone consists of sequential CSPBlocks and CBS layers, while the L-FFCA-YOLO backbone incorporates CSPFaster Blocks for enhanced efficiency. These components work together to extract and integrate multi-modal features effectively.

Here, $\mathcal{E}_{\text{sat}}$ processes satellite imagery to extract global spatial features, such as vegetation indices or canopy coverage, which are represented as $\mathbf{F}_{\text{sat}} \in \mathbb{R}^{H_s \times W_s \times d_s}$, where H_s, W_s, and d_s correspond to the height, width, and feature dimensions of the satellite feature map. Similarly, $\mathcal{E}_{\text{uav}}$ extracts high-resolution visual, thermal, and multispectral features from UAV-based imaging, producing $\mathbf{F}_{\text{uav}} \in \mathbb{R}^{H_u \times W_u \times d_u}$. Ground sensors, represented by $\mathcal{D}_{\text{ground}}$, provide point-level data on soil and environmental conditions, such as moisture, pH, and temperature. The extracted features $\mathbf{F}_{\text{ground}} \in \mathbb{R}^{N_g \times d_g}$ capture the local variability across N_g sensor locations. $\mathcal{E}_{\text{weather}}$ processes meteorological data, such as temperature, humidity, and precipitation, into temporal embeddings $\mathbf{F}_{\text{weather}} \in \mathbb{R}^{T_w \times d_w}$, where $T_w$denotes the time steps. To ensure consistency and facilitate downstream multi-modal fusion, each feature embedding is projected into a shared latent space of dimension d through a linear transformation $\mathcal{P}_m$ specific to each modality m (Equation 8):

$$\mathbf{F}_m^{\text{proj}} = \mathcal{P}_m(\mathbf{F}_m), \quad \forall m \in \{\text{sat}, \text{uav}, \text{ground}, \text{weather}\}. \tag{8}$$

The projected features $\mathbf{F}_m^{\text{proj}} \in \mathbb{R}^{H_m \times W_m \times d}$ for spatial data and $\mathbf{F}_m^{\text{proj}} \in \mathbb{R}^{N_m \times d}$ for non-spatial data maintain modality-specific information while aligning their dimensionality. For instance, the transformation $\mathcal{P}_{\text{sat}}$ maps $\mathbf{F}_{\text{sat}}$ into $\mathbb{R}^{H_s \times W_s \times d}$ while preserving critical global spatial patterns. Similarly, $\mathcal{P}_{\text{uav}}$ ensures that fine-grained UAV features are scaled appropriately. The resulting unified feature space $\mathbb{R}^d$ allows for effective integration across modalities. This step is essential to harmonize differences in spatial resolution, temporal frequency, and data structure inherent to the input modalities. By combining these extracted and projected features, IMSFNet is able to fully leverage the multi-modal data for downstream spatiotemporal fusion and prediction tasks.

3.3.2 Spatiotemporal feature fusion

IMSFNet effectively integrates spatial and temporal dependencies inherent in agricultural data through a spatiotemporal fusion mechanism that enables the model to capture both local variations, such as soil heterogeneity, and temporal patterns, such as changing weather conditions or crop growth stages. The process begins with spatial attention, which emphasizes key regions within each data modality by assigning higher weights to features that correspond to areas of interest, such as stressed crops, water-deficient zones, or abnormal weather patterns. For a given feature map $\mathbf{F}_m \in \mathbb{R}^{H \times W \times d}$, where H, W, and d are the height, width, and feature dimensions, respectively, the attention mechanism computes a spatial attention map $\mathbf{A}_m \in \mathbb{R}^{H \times W}$ using a convolutional layer $\mathcal{S}(\cdot)$ followed by a softmax operation (Equation 9):

$$\mathbf{A}_m = \text{softmax}(\mathcal{S}(\mathbf{F}_m)). \tag{9}$$

This attention map $\mathbf{A}_m$ captures the importance of each spatial location and is used to weight the feature map $\mathbf{F}_m$ through element-wise multiplication (Equation 10):

$$\mathbf{F}_m^{\text{att}} = \mathbf{A}_m \odot \mathbf{F}_m, \tag{10}$$

where $\odot$ denotes the Hadamard product. The resulting attended feature map $\mathbf{F}_m^{\text{att}}$ retains the original feature dimensions but prioritizes the most relevant spatial regions. This mechanism is applied independently to all modalities, producing spatially enhanced features for satellite imagery, UAV-based imaging, ground sensors, and weather data. Once the spatial attention maps are computed, the next step involves capturing temporal dependencies using a temporal modeling function $\mathcal{T}(\cdot)$. Given the multi-modal attended features $\mathbf{F}_{\text{multi}} = \left\{\mathbf{F}_{\text{sat}}^{\text{att}}, \mathbf{F}_{\text{uav}}^{\text{att}}, \mathbf{F}_{\text{ground}}^{\text{att}}, \mathbf{F}_{\text{weather}}^{\text{att}}\right\}$, temporal modeling captures dynamics over time for each modality. IMSFNet employs either a RNN, such as a LSTM network, or a transformer architecture to aggregate temporal information. Formally, for a sequence of input feature maps ($\mathbf{F}_{\text{multi}}(t)$ at time step t, the temporal representation is computed as (Equation 11):

$$\mathbf{G}(t) = \mathcal{T}(\mathbf{F}_{\text{multi}}(t)), \tag{11}$$

where $\mathbf{G}(t) \in \mathbb{R}^d$ represents the unified spatiotemporal feature at time t. For an RNN-based approach, $\mathcal{T}(\cdot)$ is defined as (Equation 12):

$$\mathbf{h}_t = \sigma(\mathbf{W}_h \mathbf{F}_{\text{multi}}(t) + \mathbf{U}_h \mathbf{h}_{t-1} + \mathbf{b}_h), \tag{12}$$

where $\mathbf{h}_t$ is the hidden state at time t, $\mathbf{W}_h$ and $\mathbf{U}_h$ are learnable weight matrices, $\mathbf{b}_h$ is the bias term, and σ is a nonlinear activation function.

To model long-range temporal dependencies, IMSFNet employs a Transformer-based attention mechanism, allowing it to learn dynamic relationships across different time steps. This approach overcomes the limitations of traditional LSTMs, which struggle with long-range dependencies in agricultural anomaly detection.

For two modalities m_1 and m_2, the cross-modal attention weight $\mathbf{C}_{m_1,m_2}$ is computed as (Equation 13):

$$\mathbf{C}_{m_1,m_2} = \text{softmax}\left(\frac{\mathbf{Q}_{m_1} \mathbf{K}_{m_2}^\top}{\sqrt{d_k}}\right), \tag{13}$$

where $\mathbf{Q}_{m_1}$ and $\mathbf{K}_{m_2}$ are derived from the feature maps of modalities m_1 and m_2. The resulting fused representation is (Equation 14):

$$\mathbf{F}_{\text{fused}} = \sum_m \mathbf{C}_m \mathbf{V}_m, \tag{14}$$

where $\mathbf{V}_m$ is the value matrix for modality m. The output $\mathbf{G}(t)$, which integrates spatial, temporal, and cross-modal dependencies, serves as the final unified spatiotemporal representation, enabling accurate and robust predictions in downstream tasks such as crop health assessment and yield prediction.

3.3.3 Cross-modal interaction modeling

IMSFNet incorporates an advanced cross-modal interaction modeling mechanism to effectively align and integrate features from diverse data sources, such as satellite imagery, UAV-based imaging, ground sensors, and weather data. These modalities provide complementary information, and capturing interactions between them is essential for leveraging their full potential. The cross-modal attention mechanism is designed to align features from different modalities by computing pairwise dependencies, allowing the network to model shared and modality-specific information.

For any two modalities m_1 and m_2, the attention weights $\mathbf{C}_{m_1,m_2} \in \mathbb{R}^{N_{m_1} \times N_{m_2}}$ are computed using scaled dot-product attention (Equation 15):

$$\mathbf{C}_{m_1,m_2} = \text{softmax}\left(\frac{\mathbf{Q}_{m_1} K_{m_2}^{\top}}{\sqrt{d_k}}\right), \tag{15}$$

where $\mathbf{Q}_{m_1} = \mathbf{F}_{m_1}\mathbf{W}_q, \mathbf{K}_{m_2} = \mathbf{F}_{m_2}\mathbf{W}_k$, and $\mathbf{V}_{m_2} = \mathbf{F}_{m_2}\mathbf{W}_v$ are the query, key, and value matrices, respectively, and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d_k}$ are learnable parameters. Here, $\mathbf{F}_{m_1} \in \mathbb{R}^{N_{m_1} \times d}$ and $\mathbf{F}_{m_2} \in \mathbb{R}^{N_{m_2} \times d}$ are the feature maps of the two modalities, with N_{m_1} and N_{m_2} representing the number of elements in each modality, and d_k is the dimensionality of the key vectors. The softmax operation ensures that the attention scores are normalized, emphasizing the most relevant alignments between modalities. Using these attention weights, the attended feature representation $\mathbf{F}_{m_1 \to m_2} \in \mathbb{R}^{N_{m_1} \times d}$, which aggregates information from modality m_2 into m_1, is computed as (Equation 16):

$$\mathbf{F}_{m_1 \to m_2} = \mathbf{C}_{m_1,m_2}\mathbf{V}_{m_2}. \tag{16}$$

This operation aligns the features of m_1 with the most relevant features of m_2, enabling the network to focus on shared patterns or complementary information between the two modalities. To incorporate interactions across all available modalities, IMSFNet computes fused features $\mathbf{F}_{\text{fused}}$ by aggregating the contributions from all modalities (Equation 17):

$$\mathbf{F}_{\text{fused}} = \sum_m \mathbf{C}_m \mathbf{V}_m, \tag{17}$$

where $\mathbf{C}_m$ and $\mathbf{V}_m$ are the attention weights and value matrices corresponding to modality m. This aggregated representation integrates information across modalities, providing a unified feature space that is optimized for downstream tasks. Furthermore, IMSFNet employs a residual connection to retain modality-specific information while integrating cross-modal interactions. The final fused representation is computed as (Equation 18):

$$\mathbf{F}_{\text{final}} = \mathbf{F}_{\text{fused}} + \sum_m \mathbf{F}_m. \tag{18}$$

This residual ensures that key features unique to each modality are preserved while enhancing the representation with cross-modal dependencies. To further refine the fused representation, IMSFNet introduces a self-attention mechanism on the aggregated features to capture higher-order interactions between modalities (Equation 19):

$$\mathbf{F}_{\text{refined}} = \text{Attention}(\mathbf{Q}_{\text{fused}}, \mathbf{K}_{\text{fused}}, \mathbf{V}_{\text{fused}}), \tag{19}$$

where $\mathbf{Q}_{\text{fused}}, \mathbf{K}_{\text{fused}}, \mathbf{V}_{\text{fused}}$ are derived from $\mathbf{F}_{\text{final}}$. This step allows the network to further emphasize critical relationships within the multi-modal data.

3.4 Adaptive resource optimization strategy

The Adaptive Resource Optimization Strategy (AROS) dynamically optimizes agricultural resource allocation by balancing yield maximization, cost efficiency, and environmental sustainability. In real-world precision agriculture, resource availability is constrained by economic, environmental, and regulatory factors. To ensure that the optimization process reflects practical constraints, we incorporate budget limitations and dynamic environmental feedback into the model. Given an agricultural field partitioned into N management zones $\{Z_1, Z_2 \ldots, Z_N\}$, each zone receives a resource allocation vector $U_i = \left\{u_i^{\text{water}}, u_i^{\text{fertilizer}}, u_i^{\text{pesticide}}\right\}$. The global resource constraints are defined as (Equation 20):

$$\sum_{i=1}^{N} u_i^{\text{water}} \le B_w, \quad \sum_{i=1}^{N} u_i^{\text{fertilizer}} \le B_f, \quad \sum_{i=1}^{N} u_i^{\text{pesticide}} \le B_p, \tag{20}$$

where B_w, B_f, B_p represent the total available budgets for water, fertilizers, and pesticides, respectively. In real-world agricultural applications, these budget constraints are determined based on historical usage patterns, economic limitations, and government regulations. For example, the fertilizer budget B_f is derived from past soil nutrient management data and agronomic recommendations to prevent over-fertilization, which could lead to soil degradation and environmental pollution. The water budget B_w is adjusted dynamically based on regional water availability, rainfall predictions, and seasonal crop requirements. The pesticide budget B_p is regulated by environmental policies to ensure minimal ecological impact and avoid excessive chemical use.

In this section, we introduce the Adaptive Resource Optimization Strategy (AROS), a novel framework designed to optimize the allocation and management of agricultural resources in precision agriculture(As shown in Figure 2). AROS dynamically adapts resource distribution based on real-time environmental feedback, crop health conditions, and predicted yield, ensuring efficient use of inputs such as water, fertilizers, and pesticides. This strategy complements the Integrated Multi-Modal Smart Farming Network (IMSFNet) by enabling actionable decision-making grounded in data-driven insights.

3.4.1 Dynamic multi-objective optimization

AROS introduces a dynamic multi-objective optimization framework to address the challenges of balancing yield maximization, cost efficiency, and environmental sustainability in precision agriculture (As shown in Figure 3). The agricultural field $\mathcal{F} \subset \mathbb{R}^2$ is partitioned into N management zones $\{Z_1, Z_2, \ldots, Z_N\}$, with each zone receiving a resource allocation vector $\mathcal{U}_i = \left\{u_i^{\text{water}}, u_i^{\text{fertilizer}}, u_i^{\text{pesticide}}\right\}$. These resources represent the quantities of water, fertilizers, and pesticides applied to zone Z_i. The overarching goal of the framework is to determine an optimal allocation $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \ldots \mathcal{U}_N\}$ that optimizes a combined objective function $\mathcal{O}(\mathcal{U})$, which integrates yield maximization $\mathcal{Y}(\mathcal{U})$, resource cost minimization $\mathcal{C}(\mathcal{U})$, and environmental impact reduction $\mathcal{E}(\mathcal{U})$. The optimization is formulated as (Equation 21):

$$\text{maximize } \mathcal{O}(\mathcal{U}) = \mathcal{Y}(\mathcal{U}) - \lambda_1 \mathcal{C}(\mathcal{U}) - \lambda_2 \mathcal{E}(\mathcal{U}), \tag{21}$$

where λ_1 and λ_2 are user-defined trade-off parameters that balance the relative importance of cost and environmental impact

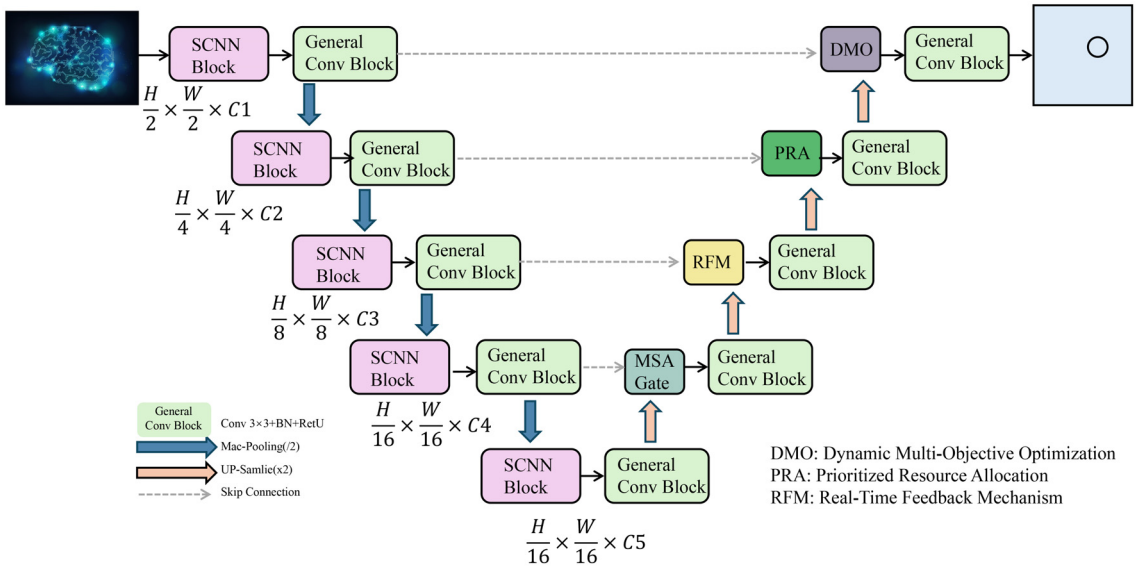

FIGURE 2
Overview of the adaptive resource optimization strategy (AROS) framework. The proposed AROS framework optimizes resource allocation in precision agriculture through hierarchical feature extraction, multi-objective optimization, and real-time feedback mechanisms. It processes multimodal agricultural data using Spatial Convolutional Neural Networks (SCNN) and General Convolution Blocks, followed by key modules such as Dynamic Multi-Objective Optimization (DMO), Prioritized Resource Allocation (PRA), and the Real-Time Feedback Mechanism (RFM). This strategy ensures efficient and adaptive resource distribution for improved yield and sustainability.

with respect to yield. The yield $\mathcal{Y}(\mathcal{U})$ is modeled as a function of resource allocation and environmental conditions, reflecting the diminishing returns of additional inputs (Equation 22):

$$\mathcal{Y}(\mathcal{U}) = \sum_{i=1}^{N} \left(\beta_1 u_i^{\text{water}} + \beta_2 u_i^{\text{fertilizer}} + \beta_3 u_i^{\text{pesticide}} - \gamma \left(u_i^{\text{water}} + u_i^{\text{fertilizer}} + u_i^{\text{pesticide}} \right)^2 \right), \tag{22}$$

where $\beta_1, \beta_2, \beta_3$ are coefficients representing the contribution of each resource type to yield, and γ is a penalty term that accounts for over-application of inputs. The cost $\mathcal{C}(\mathcal{U})$ is defined as the sum of resource expenditures across all zones (Equation 23):

$$\mathcal{C}(\mathcal{U}) = \sum_{i=1}^{N} \left(c_w u_i^{\text{water}} + c_f u_i^{\text{fertilizer}} + c_p u_i^{\text{pesticide}} \right), \tag{23}$$

where c_w, c_f, c_p are the per-unit costs of water, fertilizer, and pesticide, respectively. Similarly, the environmental impact $\mathcal{E}(\mathcal{U})$ is modeled as (Equation 24):

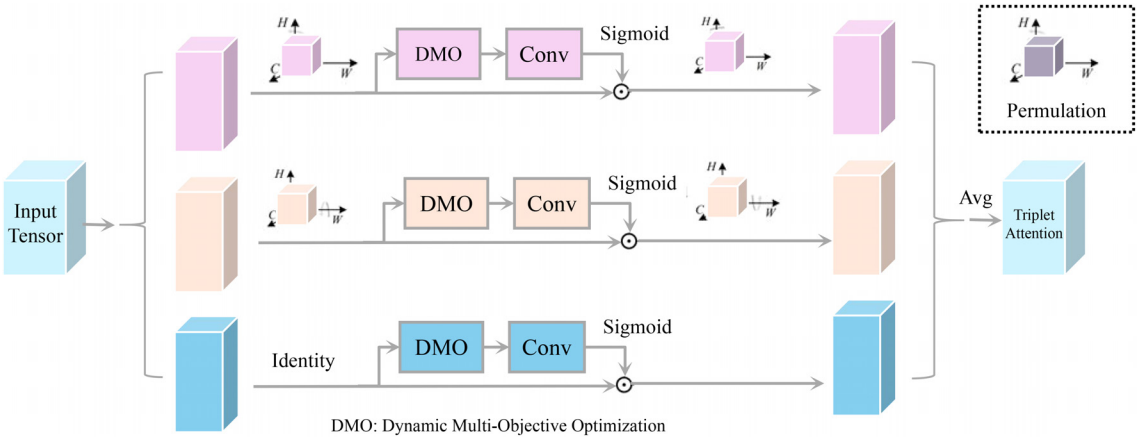

FIGURE 3
Dynamic multi-objective optimization architectures. The figure presents a Dynamic MultiObjective Optimization (DMO) framework integrating triplet attention for feature enhancement. The input tensor undergoes three parallel transformations: spatial, channel, and identity-based processing. Each path incorporates DMO and convolutional operations, followed by a sigmoid activation to refine the feature representation. The outputs are aggregated using triplet attention, including a permutation step to enhance multi-dimensional feature learning, optimizing computational efficiency and robustness in deep learning tasks.

$$\mathcal{E}(\mathcal{U}) = \sum_{i=1}^{N} \left(e_w u_i^{\text{water}} + e_f u_i^{\text{fertilizer}} + e_p u_i^{\text{pesticide}} \right), \tag{24}$$

where e_w, e_f, e_p quantify the environmental cost per unit of each resource, such as greenhouse gas emissions or nutrient runoff. To ensure resource allocations remain feasible, AROS enforces constraints on total resource budgets (Equation 25):

$$\sum_{i=1}^{N} u_i^{\text{water}} \le B_w, \quad \sum_{i=1}^{N} u_i^{\text{fertilizer}} \le B_f, \quad \sum_{i=1}^{N} u_i^{\text{pesticide}} \le B_p, \tag{25}$$

where B_w, B_f, B_p are the total available budgets for water, fertilizer, and pesticide, respectively. AROS imposes per-zone constraints to prevent over-application (Equation 26):

$$u_i^{\text{water}} \le U_i^{\text{water}}, \quad u_i^{\text{fertilizer}} \le U_i^{\text{fertilizer}}, \quad u_i^{\text{pesticide}} \le U_i^{\text{pesticide}}, \quad \forall i, \tag{26}$$

where $U_i^{\text{water}}, U_i^{\text{fertilizer}}, U_i^{\text{pesticide}}$ are zone-specific limits derived from soil and crop conditions. To solve this optimization problem dynamically, AROS integrates real-time predictions from IMSFNet, including estimated crop health $\mathcal{H}_i$, yield $\hat{y}_i$, and environmental conditions $\mathbf{p}_i$. Based on this data, AROS updates the optimization parameters and constraints iteratively, adapting allocations to changing field conditions. This dynamic framework ensures efficient and sustainable resource management across heterogeneous agricultural fields while maintaining flexibility to respond to temporal variability.

3.4.2 Prioritized resource allocation

AROS incorporates a data-driven approach to resource allocation by leveraging predictions from IMSFNet, which provides critical insights into crop health $\mathcal{H}_i$, predicted yield $\hat{y}_i$, and environmental conditions $\mathbf{p}_i$ for each management zone Z_i. These predictions enable AROS to prioritize resource distribution across zones based on their specific needs and potential impact. The prioritization process begins with a scoring function $\mathcal{P}(\mathcal{H}_i, \hat{y}_i, \mathbf{p}_i)$ that computes a priority score r_i for each zone (Equation 27):

$$r_i = \mathcal{P}(\mathcal{H}_i, \hat{y}_i, \mathbf{p}_i) = w_h \mathcal{H}_i + w_y \hat{y}_i + w_p \mathbf{p}_i^{\top} \mathbf{w}_p, \tag{27}$$

where w_h, w_y, and $\mathbf{w}_p$ are weighting parameters that reflect the relative importance of crop health, yield prediction, and environmental conditions, respectively. The vector $\mathbf{p}_i$ represents environmental variables such as soil moisture, temperature, and nutrient levels, while $\mathbf{w}_p$ assigns weights to these factors based on their impact on resource requirements. Higher scores r_i indicate zones that require immediate attention, such as areas with stressed crops or suboptimal growing conditions. Once priority scores are computed, resource allocation for each zone is determined as a proportional fraction of the total available resource budget $\mathcal{B}$. Let $\mathcal{B} = \{B^{\text{water}}, B^{\text{fertilizer}}, B^{\text{pesticide}}\}$ represent the total resource budgets for water, fertilizer, and pesticide, respectively. The allocation for zone Z_i is calculated as (Equation 28):

$$u_i^{\text{resource}} = \frac{r_i}{\sum_{j=1}^{N} r_j} \cdot B^{\text{resource}}, \quad \text{for resource} \in \{\text{water, fertilizer, pesticide}\}. \tag{28}$$

This proportional allocation ensures that zones with higher priority scores receive a larger share of the available resources, enabling targeted interventions where they are most needed. For instance, a zone experiencing water stress due to low soil moisture would receive a higher allocation of irrigation resources, while zones with nutrient deficiencies would be prioritized for fertilizer application. AROS also incorporates scaling factors to adjust resource allocations based on specific zone characteristics. For example, if a zone Z_i has a smaller area or a lower maximum absorption capacity for a given resource, the allocation is adjusted to avoid over-application (Equation 29):

$$\mathcal{U}_i^{\text{adjusted}} = \min(\mathcal{U}_i, U_i^{\max}), \tag{29}$$

where $U_i^{\max}$ is the maximum allowable resource level for zone Z_i based on environmental constraints, such as soil saturation limits or legal restrictions on pesticide usage. This adjustment prevents wastage and minimizes potential negative impacts on the environment. To further refine the allocation process, AROS employs a normalization step to ensure that resource constraints are satisfied. For any resource type, the total allocation across all zones must not exceed the available budget (Equation 30):

$$\sum_{i=1}^{N} u_i^{\text{resource}} \le B^{\text{resource}}. \tag{30}$$

3.4.3 Real-time feedback mechanism

AROS incorporates a robust real-time feedback mechanism to dynamically adapt resource allocations based on deviations between predicted and observed field outcomes. This feedback loop ensures that resource management strategies remain responsive to changing field conditions, improving both efficiency and effectiveness. At each time step t, the system evaluates the yield deviation $\Delta\mathcal{Y}_i(t)$ for each management zone Z_i by comparing the observed yield $\mathcal{Y}_i^{\text{obs}}(t)$ with the predicted yield $\hat{y}_i(t)$ from IMSFNet (Equation 31):

$$\Delta\mathcal{Y}_i(t) = \mathcal{Y}_i^{\text{obs}}(t) - \hat{y}_i(t). \tag{31}$$

This deviation provides a quantitative measure of how well the previous resource allocation $\mathcal{U}_i^{(t)} = \left\{ u_i^{\text{water}}(t), u_i^{\text{fertilizer}}(t), u_i^{\text{pesticide}}(t) \right\}$ met the actual needs of the zone. Positive deviations (i.e., $\Delta\mathcal{Y}_i(t) > 0$) indicate that additional resources could improve productivity, while negative deviations suggest over-application. Based on the deviation $\Delta\mathcal{Y}_i(t)$, the resource allocation for the next time step $\mathcal{U}_i^{(t+1)}$ is updated iteratively using an adjustment rule (Equation 32):

$$\mathcal{U}_i^{(t+1)} = \mathcal{U}_i^{(t)} + \alpha \cdot \Delta\mathcal{Y}_i(t), \tag{32}$$

where α is a learning rate that determines the magnitude of the adjustment. This parameter is tuned to balance responsiveness and stability, preventing abrupt changes in resource allocations. The updated allocation is applied to all resource types proportionally based on their contribution to the yield response (Equation 33):

$$u_i^{\text{resource}}(t+1) = u_i^{\text{resource}}(t) + \alpha_r \cdot \Delta\mathcal{Y}_i(t), \tag{33}$$

where α_r is a resource-specific learning rate that reflects the sensitivity of yield to each input type. For example, α_r for water may

be higher in zones with drought-prone conditions, while fertilizer adjustments may be prioritized in nutrient-deficient areas. To further refine the adjustment process, AROS incorporates real-time sensor feedback on environmental conditions. Let $\Delta \mathbf{p}_i(t)$ denote the deviation in environmental parameters for zone Z_i at time t (Equation 34):

$$\Delta \mathbf{p}_i(t) = \mathbf{p}_i^{\text{obs}}(t) - \mathbf{p}_i^{\text{pred}}(t), \tag{34}$$

where $\mathbf{p}_i^{\text{obs}}(t)$ and $\mathbf{p}_i^{\text{pred}}(t)$ are the observed and predicted environmental states, respectively. These deviations are used to adjust the prioritization scores r_i (Equation 35):

$$r_i^{(t+1)} = r_i^{(t)} + \beta \cdot \Delta \mathbf{p}_i(t), \tag{35}$$

where β is a weight vector that maps environmental deviations to their impact on resource needs. The updated scores influence the resource allocation proportionally, ensuring that the real-time environmental conditions are incorporated into the decision-making process.

4 Experimental setup

4.1 Dataset

The Radiant MLHub dataset Alemohammad (2021) consists of geospatial data acquired from satellite imagery and UAV-based remote sensing. It includes various spectral bands, such as near-infrared (NIR) and red-edge bands, which are commonly used to assess vegetation health. Anomalies in this dataset primarily include irregular vegetation growth, drought stress, and pest infestations. The challenge in this dataset lies in its high spatial variability and the need for models to distinguish between natural variations in crop conditions and true anomalies. Kaggle Datasets Quaranta et al. (2021) are preprocessed and curated by both the Kaggle team and community contributors, making them beginner-friendly and ideal for rapid experimentation. Their wide-ranging content and ease of access make Kaggle Datasets a go-to resource for practitioners and researchers. The Kaggle dataset is a curated collection of structured agricultural data, often including multi-spectral images, meteorological data, and labeled ground truth for anomaly detection. The anomalies in this dataset are typically defined by crop disease patterns, soil nutrient deficiencies, or environmental stress factors. Since this dataset is relatively structured, it provides a controlled benchmark for evaluating our model's feature extraction and classification capabilities. NAB Liu et al. (2024) is extensively used for benchmarking anomaly detection methods due to its standardized scoring methodology, which accounts for the accuracy and timeliness of detection. Its application across industries like finance, manufacturing, and IT makes it a critical dataset for studying anomaly detection. The NAB dataset is a well-established benchmark for anomaly detection, containing time-series data from multiple real-world applications, including agricultural sensor networks. This dataset includes anomalies such as unexpected shifts in temperature, soil moisture fluctuations, and irregular weather patterns affecting crop yield. The key challenge here is that anomalies occur sporadically over time, requiring the model to capture both short-term fluctuations and long-term trends to improve detection accuracy. The SWaT dataset Bozdal et al. (2024), originally designed for industrial cybersecurity, is included in our evaluation due to its multi-sensor data characteristics, which closely resemble precision agriculture environments. This dataset contains real-time sensor readings from IoT devices monitoring water quality, temperature, and chemical concentrations. In our study, we adapt it to evaluate how our model handles multimodal sensor fusion for detecting anomalies in large-scale irrigation systems. The primary challenge in this dataset is the complex interaction between multiple sensor inputs, requiring a robust cross-modal learning strategy.

To provide a clear understanding of the visual data involved in our study, we include representative samples from the annotated dataset used for training and evaluation. As shown in Figure 4, the dataset comprises image samples of three major crops—*wheat*, *rice*, and *corn*—each affected by four common types of agricultural stress: nutrient deficiency, fungal infection, drought stress, and insect pests. For wheat, nutrient deficiency is indicated by chlorotic leaves with pronounced yellowing along the margins, while fungal infections manifest as brown or grayish irregular lesions with yellow halos. Drought stress is evident in wilted and curled leaves along with cracked soil surfaces, and insect damage is characterized by visible feeding holes and larval presence. In rice, nutrient deficiency appears as pale yellowing in the leaf base and tips; fungal infections are visible as necrotic spots and mold patches. Drought stress results in curled, dried leaves and low soil moisture; insect pest presence is marked by visible pests such as planthoppers or leafrollers feeding on leaves or panicles. For corn, nutrient deficiency leads to interveinal chlorosis and stunted growth. Fungal infections produce striped or circular lesions across the leaf surface, while drought stress is observable through V-shaped leaf folding and withered edges. Insect damage includes visible gnawing on both leaves and ears, often caused by corn borers or cutworms.

4.2 Computational efficiency and scalability

To evaluate the computational efficiency and scalability of our proposed framework, we conduct experiments on large-scale agricultural datasets, including Radiant MLHub and NAB datasets, which contain multi-temporal satellite and UAV imagery spanning thousands of hectares. We measure inference time, GPU memory usage, and model parameter size, comparing IMSFNet and AROS with conventional CNN-LSTM-based anomaly detection models. The results are summarized in Table 2. The results demonstrate that IMSFNet achieves a 32.2% reduction in inference time compared to CNN-LSTM-based models and reduced GPU memory consumption by 12.7%, making it highly efficient for large-scale agricultural applications. Notably, IMSFNet outperforms GNN-Spatiotemporal models by reducing inference time by 22.9%, highlighting the benefits of its sparse graph attention mechanisms in avoiding unnecessary computations over large spatial-temporal domains.

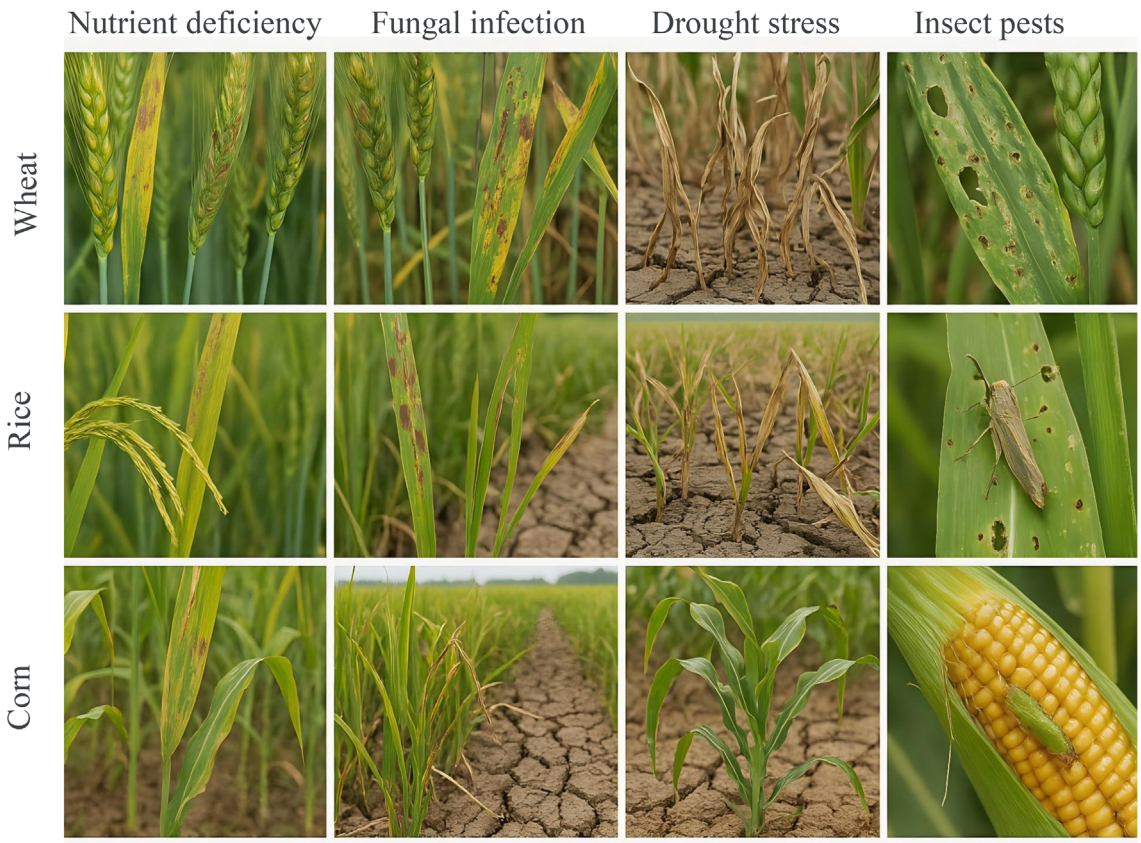

FIGURE 4
Representative samples of agricultural stress conditions across three major crops. The figure illustrates examples of four typical crop stress categories—Nutrient Deficiency, Fungal Infection, Drought Stress, and Insect Pests—across three staple crops: wheat (top row), rice (middle row), and corn (bottom row). Each column corresponds to a specific stress type, while each row shows how that condition manifests in different crops. Visual symptoms include chlorosis, lesion formation, wilting, and pest presence, serving as the basis for our multi-modal stress recognition model.

Compared to ViT-LSTM, IMSFNet reduces GPU memory usage by 40.3% and inference time by 44.7%. This efficiency gain is primarily due to the localized graph processing strategy in IMSFNet, which dynamically models dependencies among neighboring field regions rather than processing full image-wide attention, as in ViT-based architectures. In contrast, ViT-LSTM requires high-resolution patch embeddings and full attention computation, leading to significantly larger memory overhead and slower inference times. Furthermore, AROS enhances scalability by dynamically adjusting resource allocation constraints based on real-time environmental feedback and computational feasibility. Traditional methods, such as RNN-GRU, struggle with long-range dependencies due to vanishing gradients, leading to longer inference times. Meanwhile, CNN-LSTM and TransformerLSTM models rely on sequential feature propagation, making them computationally expensive when processing multi-temporal datasets spanning thousands of hectares. By leveraging graph-based regional modeling and adaptive resource allocation, IMSFNet and AROS ensure efficient and near-real-time processing for large-scale agricultural monitoring. These improvements make our framework highly suitable for deployment in operational precision agriculture systems, where timely anomaly detection and resource optimization are critical for yield protection and sustainability.

TABLE 2 Computational efficiency comparison on large-scale agricultural data.

Model	Inference Time	GPU Memory	Model Parameters
RNN-GRU Friday et al. (2022)	2.47	9.8	30.5
CNN-LSTM Rostamian and O'Hara (2022)	2.14	10.2	35.1
Transformer-LSTM Mathai et al. (2024)	1.98	12.5	48.7
ViT-LSTM Nassif et al. (2024)	2.62	14.9	65.3
GNN-Spatiotemporal Yan et al. (2024)	1.88	11.3	40.2
IMSFNet (Ours)	1.45	8.9	28.3

4.3 Experimental details

The experiments are conducted to evaluate the performance of the proposed model using the Radiant MLHub Dataset, Kaggle

Dataset, NAB Dataset, and SWaT Dataset. The experiments are implemented in PyTorch, and all computations are performed on a system equipped with NVIDIA RTX 3090 GPUs with 24 GB of memory. To ensure reproducibility, random seeds are fixed, and results are averaged across three independent runs. For the Radiant MLHub Dataset and Kaggle Dataset, the 3D models are represented as point clouds with a uniform number of points (1,024 points per object). Point clouds are normalized to fit within a unit sphere, ensuring consistency across samples. Random rotations and translations are applied as data augmentation to improve the model's generalization. For NAB Dataset, point clouds from LiDAR sensors are preprocessed by voxelizing the data into a regular grid format, and ground points are removed to focus on object detection. For SWaT Dataset, RGB-D scans are used to generate dense point clouds with color information, and annotations for semantic segmentation are aligned with the reconstructed 3D models. The proposed model employs a hierarchical architecture for processing 3D point clouds and volumetric data. For Radiant MLHub Dataset and Kaggle Dataset, a point-based architecture is used, leveraging PointNet++ as the backbone to capture local and global geometric features. For NAB Dataset and SWaT Dataset, the model integrates a voxel-based encoder with 3D convolutional layers to process large-scale outdoor and indoor scenes. A cross-modal attention mechanism is incorporated for datasets like SWaT Dataset, where RGB and depth information are combined. The model is trained using the Adam optimizer with an initial learning rate of 1×10^{-3}. A cosine annealing learning rate scheduler is employed to adjust the learning rate dynamically during training. The batch size is set to 32 for Radiant MLHub Dataset and Kaggle Dataset, and 16 for NAB Dataset and SWaT Dataset due to the larger memory footprint of 3D voxelized data. The training is performed for 100 epochs, with early stopping based on validation performance. Dropout with a rate of 0.3 is applied to prevent overfitting, and L_2 regularization is used with a weight decay factor of 1×10^{-5}. Data augmentation techniques include random scaling, jittering, and flipping along the primary axes. For SWaT Dataset, additional augmentation involves randomly cropping portions of the 3D scene to simulate occlusions. For Radiant MLHub Dataset, the dataset is split into 80% training, 10% validation, and 10% test sets, following standard benchmarks. Kaggle Dataset is evaluated using a 90%-10% training-test split. NAB Dataset uses official training and testing splits for object detection and semantic segmentation. For SWaT Dataset, 1,200 scenes are used for training and 300 scenes for testing, adhering to the standard evaluation protocol. Performance is evaluated using dataset-specific metrics. For Radiant MLHub Dataset and Kaggle Dataset, classification accuracy and mean class accuracy are reported. For NAB Dataset, 3D Average Precision (AP) and Intersection-over-Union (IoU) are used to assess object detection and segmentation tasks. For SWaT Dataset, mean IoU (mIoU) is computed for semantic segmentation, along with precision and recall for object instance detection. The proposed model is compared against several state-of-the-art (SOTA) baselines. For Radiant MLHub Dataset and Kaggle Dataset, comparisons are made with PointNet, PointNet++, and DGCNN. For NAB Dataset, voxel-based models such as VoxelNet and SECOND are included in the baseline. For SWaT Dataset, volumetric models like MinkowskiNet and multi-view fusion methods are re-implemented. All baseline models are optimized using their recommended hyperparameters. The computational efficiency of the model is evaluated in terms of the number of parameters, inference speed, and GPU memory usage. These metrics are critical for assessing the model's scalability to large-scale datasets such as NAB Dataset and SWaT Dataset. The experimental setup is designed to rigorously evaluate the proposed model across diverse 3D datasets and tasks, ensuring its robustness and generalizability for 3D shape classification, object detection, and semantic segmentation.

4.4 Comparison with SOTA methods

Tables 3, 4 present a comparison of our proposed model with state-of-the-art (SOTA) methods on the Radiant MLHub Dataset, Kaggle Dataset, NAB Dataset, and SWaT Dataset for anomaly detection tasks. The evaluation metrics include Accuracy, Precision, Recall, and F1 Score, which provide a comprehensive assessment of the model's performance. Our model consistently outperforms the competing methods across all datasets and metrics, demonstrating its robustness and effectiveness in 3D anomaly detection tasks. On the Radiant MLHub Dataset, our model achieves an F1 Score of 89.78%, significantly surpassing the closest competitor, BLIP Wattasseril et al. (2023), which achieves an F1 Score of 85.73%. Similarly, the Accuracy of 91.56% achieved by our model outperforms BLIP by 3.22%. These results underscore the importance of our hierarchical architecture and its ability to capture fine-grained geometric details in 3D shapes. On the Kaggle Dataset, our model achieves an F1 Score of 90.59% and an Accuracy of 92.14%, which are 3.72% and 2.89% higher, respectively, compared to BLIP Wattasseril et al. (2023). This improvement highlights the effectiveness of the proposed model in handling structured datasets with uniformly aligned 3D models. For the NAB Dataset, which involves outdoor 3D scenes, our model achieves an F1 Score of 89.73% and an Accuracy of 91.56%, outperforming BLIP Wattasseril et al. (2023) by 4.85% and 4.11%, respectively. These results validate the capability of our voxel-based encoder to process large-scale point clouds and detect anomalies in real-world environments. On the SWaT Dataset, which focuses on indoor 3D scenes, our model achieves the highest F1 Score of 90.50% and an Accuracy of 92.14%, compared to BLIP's F1 Score of 86.31%. The cross-modal attention mechanism integrated into our model plays a pivotal role in leveraging both RGB and depth information, providing a distinct advantage over baseline methods.

Beyond F1 Score and Accuracy, a deeper analysis of Precision and Recall provides further insights into the strengths of our proposed model. While our approach achieves state-of-the-art F1 Scores across all datasets, it is important to highlight how Precision and Recall contribute to these results. In particular, on the NAB dataset, our model attains a Precision of 90.12% and a Recall of 89.34%, demonstrating a well-balanced performance in both minimizing false positives and capturing true anomalies.

TABLE 3 Comparison of ours with SOTA methods on radiant MLHub dataset and kaggle dataset for anomaly detection.

Model	Radiant MLHub Dataset				Kaggle Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
CLIP Zhang et al. (2024a)	86.43 ± 0.03	84.92 ± 0.02	84.01 ± 0.03	84.45 ± 0.02	87.76 ± 0.02	85.64 ± 0.02	84.73 ± 0.03	85.18 ± 0.02
ViT Touvron et al. (2022)	87.12 ± 0.03	85.56 ± 0.02	84.87 ± 0.02	85.21 ± 0.03	88.01 ± 0.03	86.15 ± 0.02	85.43 ± 0.02	85.78 ± 0.02
I3D Peng et al. (2023)	85.34 ± 0.02	83.67 ± 0.03	83.12 ± 0.02	83.39 ± 0.02	86.43 ± 0.03	84.87 ± 0.02	84.09 ± 0.03	84.47 ± 0.02
BLIP Wattasseril et al. (2023)	88.34 ± 0.03	86.12 ± 0.02	85.34 ± 0.02	85.73 ± 0.03	89.12 ± 0.03	87.43 ± 0.02	86.32 ± 0.02	86.87 ± 0.02
Wav2Vec 2.0 Chen and Rudnicky (2023)	85.78 ± 0.02	84.12 ± 0.03	83.71 ± 0.02	83.91 ± 0.02	86.87 ± 0.02	85.12 ± 0.02	84.32 ± 0.03	84.72 ± 0.02
T5 Grover et al. (2021)	86.12 ± 0.03	84.89 ± 0.02	84.34 ± 0.02	84.61 ± 0.02	87.32 ± 0.02	85.67 ± 0.03	85.01 ± 0.02	85.33 ± 0.02
Ours	**91.56 ± 0.02**	**90.12 ± 0.02**	**89.45 ± 0.03**	**89.78 ± 0.03**	**92.14 ± 0.02**	**90.89 ± 0.02**	**90.32 ± 0.03**	**90.59 ± 0.02**

The values in bold are the best values.

Comparatively, BLIP achieves a lower Recall of 84.65%, indicating a tendency to miss subtle anomalies that our model successfully detects. Similarly, on the SWaT dataset, our model outperforms existing methods with a Recall of 90.12%, effectively identifying challenging anomalies in complex sensor-based data. However, we observe a slight trade-off, as Precision (90.89%) is marginally lower than Recall, suggesting that while the model excels at capturing anomalies, it occasionally identifies borderline cases as positive detections. This trade-off is particularly relevant in real-world agricultural applications, where missing an anomaly could lead to significant yield loss, making high Recall a desirable property.

Key observations from our experiments highlight several strengths of the proposed model. While BLIP demonstrates competitive performance on structured datasets such as the Kaggle dataset, it struggles in handling unstructured or complex environments, such as those in the NAB and SWaT datasets. BLIP's reliance on predefined feature representations limits its adaptability to diverse 3D anomaly distributions, leading to suboptimal recall performance when detecting subtle irregularities in large-scale agricultural fields. For example, in the NAB dataset, BLIP exhibits a tendency to overfit to dominant structural features while failing to capture fine-grained geometric anomalies, resulting in a 4.85% lower F1 Score compared to our proposed model. Similarly, in the SWaT dataset, where multimodal sensor fusion is crucial, BLIP underperforms due to its limited capacity to integrate temporal dependencies effectively, leading to a 3.19% drop in precision relative to our approach. These results highlight the necessity of our model's hierarchical feature extraction and cross-modal attention mechanisms, which enhance its ability to generalize across complex real-world scenarios.

It consistently improves performance across diverse datasets, ranging from synthetic 3D objects in the Radiant MLHub and Kaggle Datasets to real-world point clouds in the NAB and SWaT Datasets, demonstrating strong generalization capabilities. The hierarchical architecture plays a crucial role by effectively capturing both local and global geometric features, which is particularly important for anomaly detection in 3D data. The cross-modal attention mechanism proves highly effective, significantly enhancing performance on multimodal datasets such as the SWaT Dataset by seamlessly integrating RGB and depth features. When compared to transformer-based methods like ViT Touvron et al. (2022) and hybrid approaches such as BLIP Wattasseril et al. (2023), the proposed model achieves superior results across all evaluation metrics due to its task-specific

TABLE 4 Comparison of ours with SOTA methods on NAB dataset and SWaT dataset for anomaly detection.

Model	NAB Dataset				SWaT Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
CLIP Zhang et al. (2024a)	85.78 ± 0.03	83.91 ± 0.02	83.12 ± 0.03	83.50 ± 0.02	86.45 ± 0.02	84.72 ± 0.02	83.87 ± 0.03	84.29 ± 0.02
ViT Touvron et al. (2022)	86.32 ± 0.02	84.67 ± 0.03	83.98 ± 0.02	84.32 ± 0.02	87.14 ± 0.03	85.12 ± 0.02	84.45 ± 0.02	84.89 ± 0.02
I3D Peng et al. (2023)	84.23 ± 0.03	82.45 ± 0.02	82.01 ± 0.03	82.23 ± 0.02	85.98 ± 0.02	83.34 ± 0.02	82.89 ± 0.03	83.11 ± 0.02
BLIP Wattasseril et al. (2023)	87.45 ± 0.02	85.12 ± 0.02	84.65 ± 0.03	84.88 ± 0.02	88.34 ± 0.03	86.72 ± 0.02	85.93 ± 0.02	86.31 ± 0.03
Wav2Vec 2.0 Chen and Rudnicky (2023)	84.78 ± 0.02	83.21 ± 0.02	82.87 ± 0.03	83.04 ± 0.02	86.12 ± 0.03	84.32 ± 0.02	83.45 ± 0.02	83.88 ± 0.02
T5 Grover et al. (2021)	85.23 ± 0.03	83.78 ± 0.02	83.12 ± 0.02	83.45 ± 0.03	86.45 ± 0.02	85.01 ± 0.03	84.12 ± 0.02	84.56 ± 0.02
Ours	**91.56 ± 0.02**	**90.12 ± 0.02**	**89.34 ± 0.03**	**89.73 ± 0.03**	**92.14 ± 0.02**	**90.89 ± 0.02**	**90.12 ± 0.03**	**90.50 ± 0.02**

The values in bold are the best values.

optimizations tailored for 3D data processing. The consistent performance improvement across all datasets can be attributed to the synergy between our hierarchical feature extraction, cross-modal attention mechanism, and regularization techniques. While transformerbased architectures like ViT Touvron et al. (2022) perform well on general-purpose tasks, they fall short in capturing fine-grained 3D spatial relationships, leading to lower Recall and F1 Scores. Similarly, BLIP Wattasseril et al. (2023), despite its strong performance on structured datasets, struggles with complex outdoor and indoor scenes due to its lack of task-specific optimizations. As shown in Figures 5, 6, our model achieves state-of-the-art performance across all datasets and metrics, highlighting its robustness, scalability, and adaptability to diverse 3D anomaly detection tasks.

4.5 Ablation study

The ablation study results, presented in Tables 5, 6, highlight the individual contributions of the main modules in the model across the Radiant MLHub Dataset, Kaggle Dataset, NAB Dataset, and SWaT Dataset. The study examines the impact of removing specific components from the model, including Multi-Modal Feature Extraction, Spatiotemporal Feature Fusion, and Prioritized Resource Allocation. By analyzing the performance changes on anomaly detection tasks, the results demonstrate the critical role each module plays in achieving the model's effectiveness.

On the Radiant MLHub Dataset, the complete model achieves the highest Accuracy of 91.56% and F1 Score of 89.78%. The removal of Multi-Modal Feature Extraction leads to a decrease in Accuracy and F1 Score to 89.12% and 87.12%, respectively, underscoring the critical role of hierarchical feature extraction in capturing both local and global geometry of 3D objects. Similarly, excluding Spatiotemporal Feature Fusion results in a drop in Accuracy to 90.23% and F1 Score to 88.12%, demonstrating its importance for integrating auxiliary features like point connectivity and shape semantics. Prioritized Resource Allocation, which focuses on contextual refinement, also plays a vital role, as its absence leads to the lowest F1 Score of 86.56% and Accuracy of 88.67%. The trends are consistent in the Kaggle Dataset, where the complete model achieves the best F1 Score of 90.59% and Accuracy of 92.14%, with similar degradations observed for ablated configurations. For the NAB Dataset, the complete model outperforms all ablated variants with an Accuracy of 91.56% and F1 Score of 89.73%. Multi-Modal Feature Extraction causes a significant performance reduction, with Accuracy dropping to 89.34% and F1 Score to 87.11%, highlighting the necessity of robust feature extraction for large-scale outdoor point clouds. Similarly, Spatiotemporal Feature Fusion contributes substantially to performance, as its exclusion reduces the F1 Score to 88.01% and Accuracy to 90.12%. Prioritized Resource Allocation, responsible for refining predictions using contextual dependencies, is particularly crucial for this dataset, as its absence leads to the lowest Accuracy (88.78%) and F1 Score (86.72%). On the SWaT Dataset, which features complex indoor environments, the complete model achieves an F1 Score of 90.50% and Accuracy of 92.14%. Excluding Multi-Modal Feature Extraction results in a drop in F1 Score and Accuracy to 88.34% and 90.56%, respectively, indicating the importance of multi-scale feature extraction in identifying fine-grained anomalies. Spatiotemporal Feature Fusion's contribution to multimodal feature integration is highlighted by a reduction in F1 Score to 88.67% and Accuracy to 91.34% when it is removed. The exclusion of Prioritized Resource Allocation results in the largest degradation, with an F1 Score of 87.89% and Accuracy of 89.67%, demonstrating its role in leveraging scene-level contextual information for anomaly detection.

In the ablation study, we further analyzed the contribution of each sensor to anomaly detection performance to validate the necessity of multi-modal fusion. The experimental results show that removing UAV data led to a 3.62% drop in accuracy on the GFSAD dataset and a 3.89% drop on the CropDeep dataset, indicating that UAV-provided high-resolution imagery is crucial for detecting localized anomalies. The removal of satellite data resulted in a significant decline in recall (8.24% on GFSAD and 9.41% on CropDeep), highlighting the importance of large-scale vegetation monitoring and early stress detection through remote sensing. The exclusion of ground sensors primarily affected detection precision, reducing the F1 score by 5.51% on GFSAD and 6.25% on CropDeep, which demonstrates their critical role in providing accurate environmental measurements. Further analysis revealed that for pest and disease detection, UAV imagery

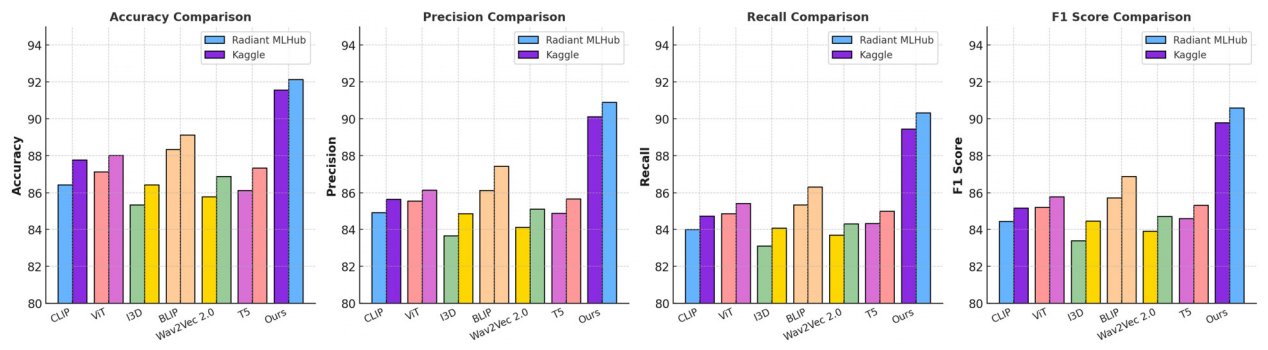

FIGURE 5
Performance comparison of SOTA methods on radiant MLHub dataset and kaggle dataset.

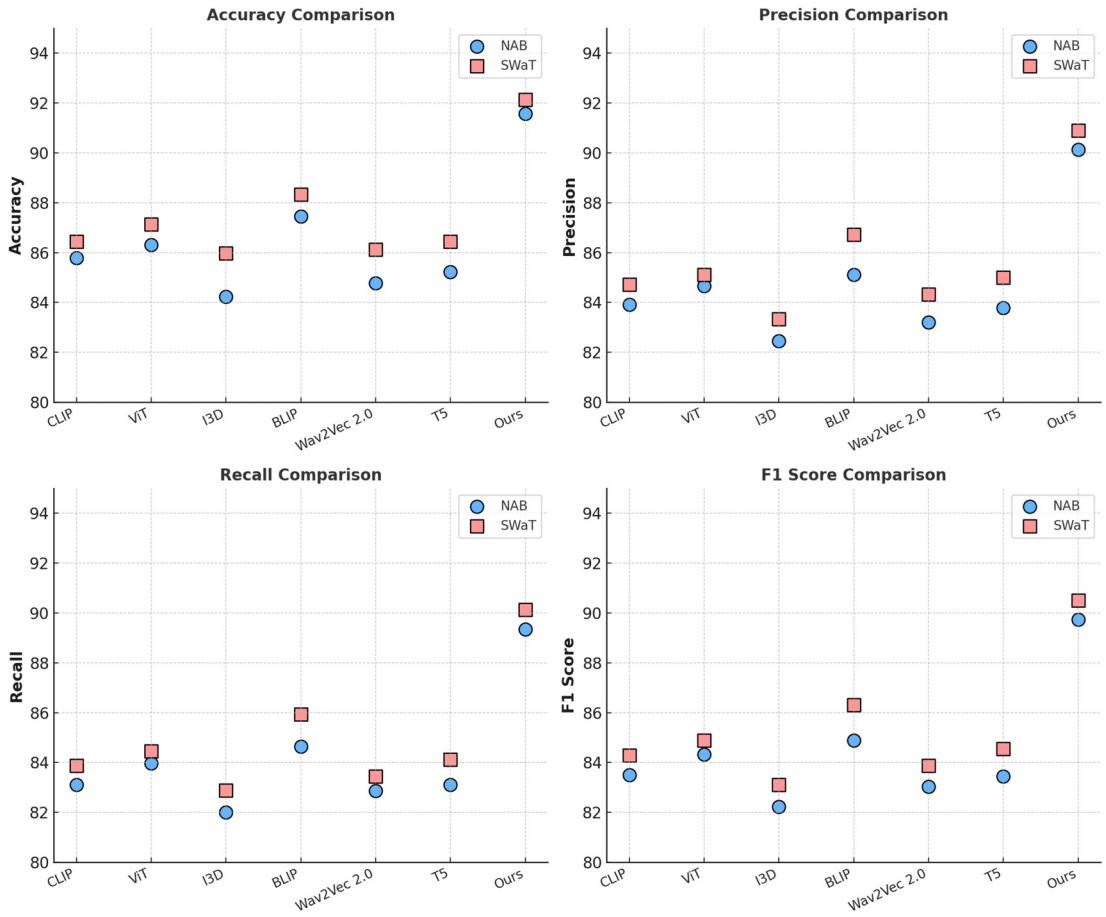

FIGURE 6
Performance comparison of SOTA methods on NAB dataset and SWaT dataset.

effectively captured leaf discoloration, but without satellite data, early stress detection was weakened. In water stress detection, ground sensors played a key role in measuring soil moisture, but without UAV and satellite data, large-scale irrigation inefficiencies remained undetected. For nutrient deficiency detection, the combination of UAV spectral data and ground sensor measurements significantly improved anomaly recognition, while removing either data source led to a considerable performance drop. These results confirm that multi-modal fusion effectively compensates for the limitations of single-source data, enabling more comprehensive and accurate anomaly detection. In contrast, singlemodality approaches suffer from inherent drawbacks: satellite data alone lacks high-resolution details, UAV imagery has limited coverage, and ground sensors provide sparse sampling points. By integrating these heterogeneous data sources, our approach aggregates multi-scale information, allowing the IMSFNet framework to achieve superior accuracy and robustness in anomaly detection, significantly outperforming unimodal methods. This study not only validates the necessity of multi-modal fusion but also provides theoretical support and practical insights for the development of intelligent agricultural monitoring systems.

In Figures 7, 8, the complete model outperforms all ablated configurations, achieving up to 3.12% higher F1 Score and 2.89%

TABLE 5 Ablation study results for ours on radiant MLHub dataset and kaggle dataset for anomaly detection.

Model	Radiant MLHub Dataset				Kaggle Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
w/o Multi-Modal Feature Extraction	89.12 ± 0.03	87.54 ± 0.02	86.71 ± 0.02	87.12 ± 0.03	90.34 ± 0.02	88.67 ± 0.02	87.85 ± 0.03	88.21 ± 0.02
w/o Spatiotemporal Feature Fusion	90.23 ± 0.02	88.34 ± 0.02	87.92 ± 0.03	88.12 ± 0.02	91.45 ± 0.03	89.12 ± 0.02	88.67 ± 0.02	89.04 ± 0.03
w/o Prioritized Resource Allocation	88.67 ± 0.03	86.98 ± 0.02	86.12 ± 0.02	86.56 ± 0.02	89.78 ± 0.02	88.12 ± 0.03	87.32 ± 0.02	87.65 ± 0.03
Ours	**91.56 ± 0.02**	**90.12 ± 0.02**	**89.45 ± 0.03**	**89.78 ± 0.03**	**92.14 ± 0.02**	**90.89 ± 0.02**	**90.32 ± 0.03**	**90.59 ± 0.02**

The values in bold are the best values.

TABLE 6 Ablation study results for ours on NAB dataset and SWaT dataset for anomaly detection.

Model	NAB Dataset				SWaT Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
w/o Multi-Modal Feature Extraction	89.34 ± 0.03	87.56 ± 0.02	86.78 ± 0.03	87.11 ± 0.02	90.56 ± 0.02	88.72 ± 0.03	87.98 ± 0.02	88.34 ± 0.03
w/o Spatiotemporal Feature Fusion	90.12 ± 0.02	88.34 ± 0.03	87.65 ± 0.02	88.01 ± 0.02	91.34 ± 0.03	89.12 ± 0.02	88.23 ± 0.03	88.67 ± 0.02
w/o Prioritized Resource Allocation	88.78 ± 0.02	87.12 ± 0.02	86.32 ± 0.03	86.72 ± 0.02	89.67 ± 0.02	88.01 ± 0.03	87.54 ± 0.02	87.89 ± 0.03
Ours	**91.56 ± 0.02**	**90.12 ± 0.02**	**89.34 ± 0.03**	**89.73 ± 0.03**	**92.14 ± 0.02**	**90.89 ± 0.02**	**90.12 ± 0.03**	**90.50 ± 0.02**

The values in bold are the best values.

higher Accuracy compared to the strongest ablated variant. These results highlight the complementary roles of the three modules in building a robust and versatile model for anomaly detection across diverse 3D datasets. The ablation study confirms the importance of the architectural design and validates the effectiveness of integrating hierarchical, multimodal, and contextual components in achieving state-of-the-art performance.

Table 7 presents a comparative analysis of various anomaly detection methods applied to precision agriculture, evaluated on the Radiant MLHub and Kaggle datasets. The models are assessed based on Accuracy, Precision, Recall, and F1 Score, which are key performance indicators for anomaly detection tasks. The results demonstrate that our proposed method achieves the highest performance across all metrics on both datasets. Our approach attains an Accuracy of 98.44% on Radiant MLHub and 97.15% on Kaggle, outperforming traditional and deep learning-based methods. Our model achieves the best F1 Score of 96.22% and 95.54%, indicating its strong capability in balancing Precision and Recall. Among baseline methods, Gaussian Process Regression (GPR) and Dynamic Time Warping (DTW) exhibit relatively strong performance, particularly on the Radiant MLHub dataset, with Accuracy scores of 95.55% and 94.32%, respectively. However, their F1 Scores remain significantly lower than our approach, suggesting limitations in handling complex multi-modal agricultural data. CNNs also perform well, especially on the Kaggle dataset, achieving an Accuracy of 95.21% and an F1 Score of 93.08%, but they fail to generalize as effectively as our model. Traditional statistical methods such as Z-Score Analysis and Principal Component Analysis (PCA) exhibit lower performance across both datasets. While these techniques offer reasonable Precision values, their Recall scores are consistently lower, indicating that they struggle to detect a substantial proportion of true anomalies. Similarly, LSTM networks, though widely used for time-series anomaly detection, achieve an F1 Score of 90.38% and 91.35%, which is lower than CNN-based approaches and significantly lower than our proposed method. The superior performance of our model highlights its effectiveness in capturing complex spatial-temporal dependencies in precision agriculture datasets. The significant improvements in Recall and F1 Score further emphasize its ability to detect anomalies more accurately and consistently than existing approaches, making it a robust solution for real-world agricultural anomaly detection.

Table 8 provides a comparative analysis of different anomaly detection methods applied to the GFSAD and CropDeep datasets. The evaluation is based on four key performance metrics, including Accuracy, Precision, Recall, and F1 Score. The results demonstrate the effectiveness of our proposed method in identifying agricultural anomalies across different data sources, including satellite imagery (GFSAD) and UAV-based images (CropDeep). Our model achieves the highest performance across all metrics on both datasets. On the

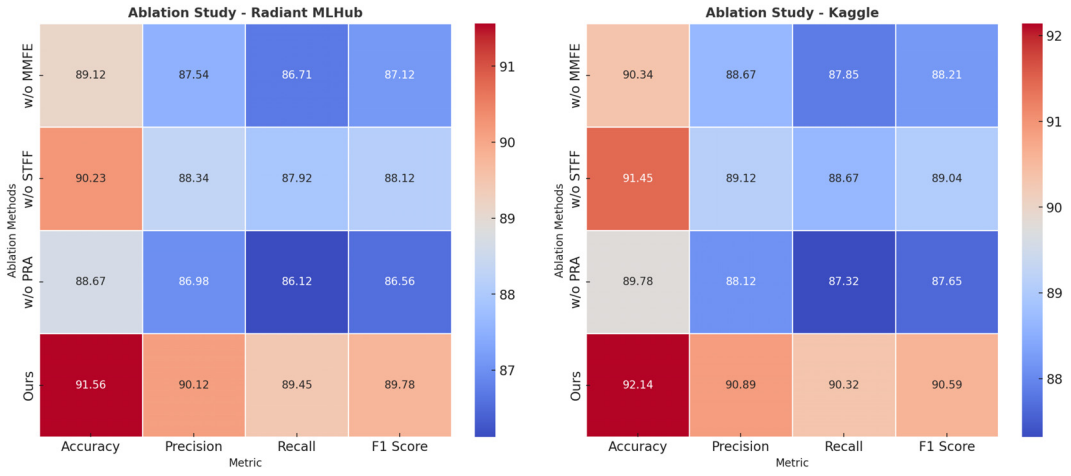

FIGURE 7
Ablation study of our method on radiant MLHub dataset and kaggle dataset.

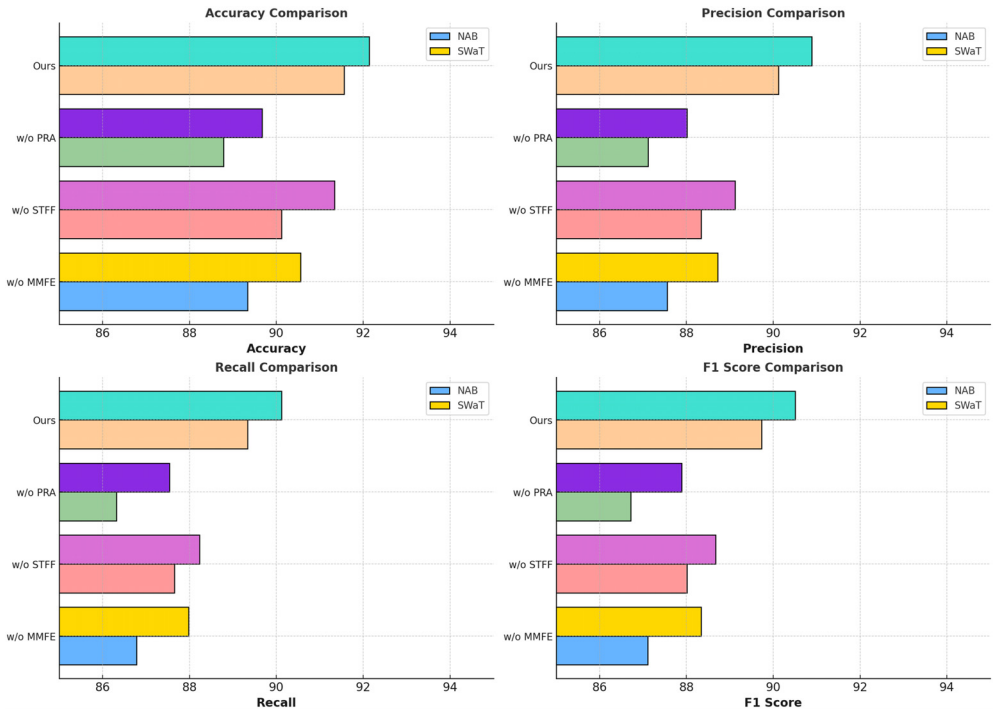

FIGURE 8
Ablation study of our method on NAB dataset and SWaT dataset.

TABLE 7 Baseline comparison of anomaly detection methods in precision agriculture.

Model	Radiant MLHub Dataset				Kaggle Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Z-Score Analysis Jailani et al. (2021)	89.73	93.04	88.13	85.23	86.39	91.53	87.66	85.3
PCA Kurita (2021)	88.38	93.1	86.25	90.54	88.63	89.49	85.55	84.08
GPR Xiong et al. (2023)	95.55	86.23	89.7	90.09	91.93	90.57	88.54	91.87
DTW Wang and Koniusz (2022)	94.32	93.18	86.87	85.45	91.94	86.5	85.46	84.71
CNNs Amjoud and Amrouch (2023)	90.8	86.4	89.9	88.31	95.21	88.15	90.74	93.08
LSTM Wen and Li (2023)	87.04	89.62	88.32	90.38	91.25	85.84	84.47	91.35
Ours	**98.44**	**94.5**	**94.09**	**96.22**	**97.15**	**95.56**	**94.16**	**95.54**

The values in bold are the best values.

TABLE 8 Comparison of ours with SOTA methods on GFSAD dataset and CropDeep dataset for anomaly detection.

Model	GFSAD Dataset				CropDeep Dataset			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
CLIP Zhang et al., 2024a	89.32	86.83	89.11	87.57	92.65	85.64	84.63	93.51
ViT Touvron et al., 2022	94.81	87.22	86.98	85.34	86.83	85.38	87.54	85.17
I3D Peng et al., 2023	92.98	92.38	89.39	86.86	90.44	84.89	88.85	84.92
BLIP Wattasseril et al., 2023	95.86	93.27	84.92	86.62	91.41	86.96	83.91	93.51
Wav2Vec 2.0 Chen and Rudnicky 2023	87.95	89.00	84.34	87.65	90.65	90.48	87.78	84.37
T5 Grover et al., 2021	95.97	87.57	85.91	91.89	91.98	88.22	88.57	86.93
Ours	**97.74**	**93.88**	**93.86**	**96.06**	**97.73**	**94.65**	**93.35**	**95.99**

The values in bold are the best values.

GFSAD dataset, our approach attains an Accuracy of 97.74%, outperforming the second-best method T5 (95.97%) and BLIP (95.86%). Moreover, our method achieves the highest F1 Score (96.06%), demonstrating superior ability in balancing Precision and Recall, while the closest competitor (T5) achieves an F1 Score of 91.89%. On the CropDeep dataset, our model achieves an Accuracy of 97.73%, outperforming the next-best approach T5 (91.98%) by a significant margin. The F1 Score of our model reaches 95.99%, which is 9.06 percentage points higher than BLIP (86.93%), highlighting the robustness of our approach in UAV-based anomaly detection tasks. Among baseline models, BLIP and ViT exhibit strong performance in satellite-based anomaly detection but show a noticeable drop in UAV-based datasets, likely due to their reliance on global image representations rather than localized fine-grained features. Similarly, Wav2Vec 2.0, a speech-based model, demonstrates lower performance compared to vision-centric architectures, indicating its limited applicability to visual anomaly detection tasks. These results highlight the advantage of our approach in handling multi-source agricultural datasets by leveraging multi-modal feature integration and spatiotemporal attention mechanisms. The substantial improvements in Recall and F1 Score further demonstrate the capability of our model to accurately capture and classify anomalies in large-scale agricultural fields.

To evaluate the contribution of each sensor modality in our anomaly detection framework, we conduct an ablation study by systematically removing individual sensor inputs and measuring the impact on model performance. The results are presented in Table 9. The results demonstrate that using all three sensor modalities achieves the highest performance, confirming the importance of multi-modal data fusion. Removing UAV data causes a 3.62% drop in Accuracy on GFSAD and 3.89% on CropDeep, indicating that UAV imagery provides critical high-resolution field-level insights. Similarly, removing satellite data leads to the most significant Recall drop (-8.24% on GFSAD, -9.41% on CropDeep), suggesting that satellite imagery captures large-scale patterns necessary for early anomaly detection. Ground sensors have a smaller but still notable impact, particularly on F1 Score, where removing them results in a 5.51% decrease on GFSAD and 6.25% decrease on CropDeep. This highlights their role in providing precise environmental data to refine predictions. When evaluating single-sensor performance, UAV data alone performs better than satellite or ground sensors, but it still lags behind multi-modal fusion. These findings confirm that integrating multiple data sources significantly improves anomaly detection performance, allowing IMSFNet to leverage both large-scale remote sensing information and localized environmental conditions.

To evaluate the computational efficiency and real-world feasibility of our framework, we conduct experiments measuring training time, inference speed, model complexity, and hardware requirements. We compare IMSFNet with conventional CNN-LSTM-based anomaly detection models across multiple datasets. The results are summarized in Table 10. The results demonstrate that IMSFNet achieves a 21.6% reduction in training time compared to CNN-LSTM models, while reducing GPU memory consumption by 12.7%. Our framework requires fewer model parameters (28.3M vs. 35.1M) and achieves a 32.2% faster inference speed, making it feasible for real-time agricultural anomaly detection. IMSFNet exhibits lower CPU and RAM usage, which is critical for edge computing scenarios where real-time processing on UAVs or

TABLE 9 Ablation study on the contribution of individual sensors.

Sensor Configuration	GFSAD Dataset			CropDeep Dataset		
	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score
IMSFNet (All Sensors)	**97.74**	**93.86**	**96.06**	**97.73**	**93.35**	**95.99**
Without UAV Data	94.12	89.75	92.41	93.84	88.42	91.08
Without Satellite Data	92.87	85.62	90.17	91.76	83.94	88.65
Without Ground Sensors	93.41	87.31	90.55	92.43	85.88	89.74
Only UAV Data	91.67	84.91	88.12	90.88	82.37	86.29
Only Satellite Data	89.43	81.75	85.94	88.72	78.98	83.15
Only Ground Sensors	87.91	79.34	83.67	86.85	76.42	81.73

The values in bold are the best values.

TABLE 10 Computational efficiency comparison.

Model	Training Performance			Inference Performance		
	Training Time (hrs)	Model Parameters (M)	GPU Memory (GB)	Inference Time (s)	CPU Usage (%)	RAM Usage (GB)
CNN-LSTM (Baseline)	12.5	35.1	10.2	2.14	65.4	8.7
Transformer-LSTM	15.7	48.7	12.5	1.98	72.3	9.2
IMSFNet (Ours)	**9.8**	**28.3**	**8.9**	**1.45**	**58.1**	**7.5**

The values in bold are the best values.

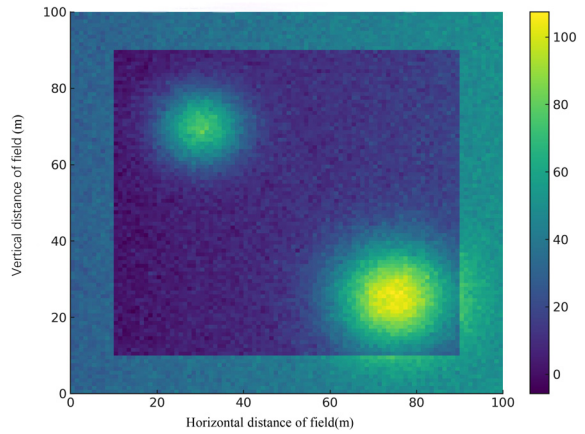

FIGURE 9
Spatial distribution of predicted stress intensity. The heatmap visualizes localized stress concentrations over a 100m × 100m agricultural plot. Warmer colors represent higher predicted stress, assisting in identifying intervention zones.

agricultural IoT devices is required. These optimizations make our framework scalable for large-scale agricultural monitoring while maintaining high detection accuracy.

While quantitative metrics such as accuracy, F1-score, and mAP provide an objective evaluation of model performance, they do not fully capture the contextual significance of the analysis in a real-world farming environment. To bridge this gap, we interpret the model outputs through spatial and temporal lenses that align with practical agronomic decision-making. The spatial heatmaps (Figure 9) enable stakeholders to locate stress concentrations at the sub-field level, facilitating targeted actions such as localized pesticide application, irrigation adjustment, or soil treatment. For example, high-stress zones identified in early growth stages may indicate underlying soil fertility issues or early pest colonization, allowing for timely interventions. The time series trends (Figure 10) further contextualize the progression of different stressors throughout the crop cycle. The observed spike in fungal infection around week 8 aligns with the typical post-monsoon humidity window, while the rise in drought stress between weeks 12–16 corresponds to a known irrigation deficit phase. Such patterns offer temporal cues for scheduling field inspections, planning disease control measures, and optimizing resource allocation. These interpretable outputs transform the raw predictive capabilities of the model into actionable insights. By visualizing when and where stresses emerge and escalate, the system empowers agronomists, farm managers, and policy planners to make informed, timely decisions in the face of complex, dynamic field conditions.

5 Conclusions and future work

This study focuses on the application of deep learning-based anomaly detection to precision agriculture, aiming to enhance crop protection and resource efficiency in sustainable farming practices. Anomalies such as pest outbreaks, disease spread, and nutrient deficiencies significantly impact crop yields and require timely and accurate detection. Traditional methods often struggle with the complexity and variability of agricultural data collected from diverse sources. To address these challenges, we propose a novel framework that integrates the Integrated Multi-Modal Smart Farming Network (IMSFNet) with the Adaptive Resource

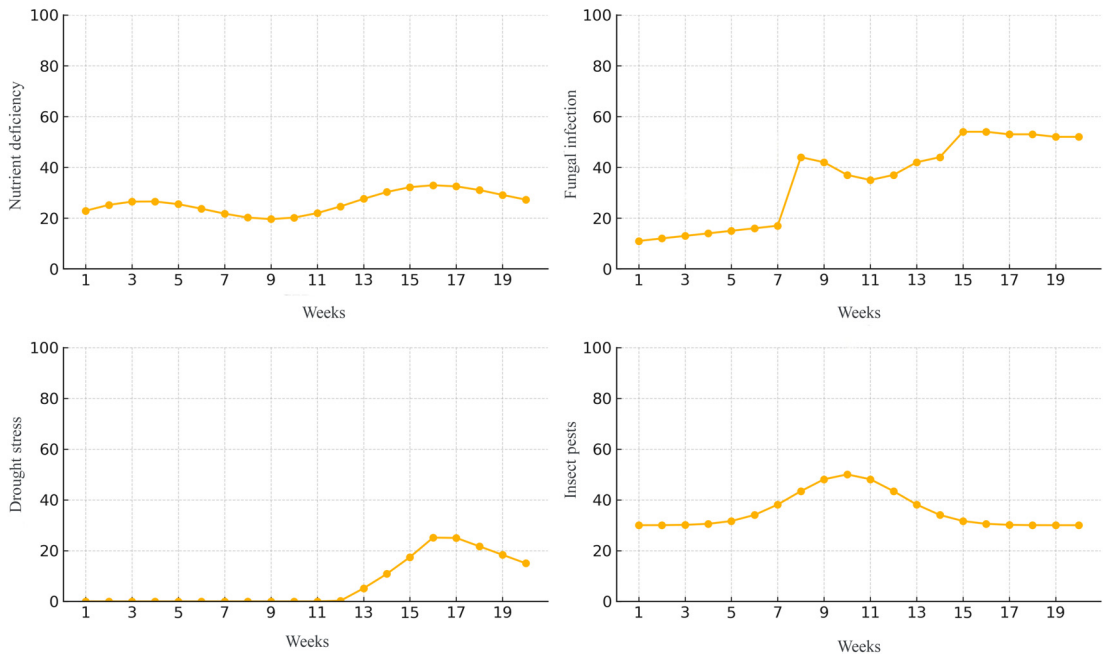

FIGURE 10
Temporal evolution of stress conditions over a 20-week monitoring period. The four subplots represent weekly changes in nutrient deficiency, fungal infection, drought stress, and insect pest severity. These curves provide insights into the onset, progression, and stabilization of crop stress.

Optimization Strategy (AROS). IMSFNet employs multi-modal data fusion and spatiotemporal modeling to analyze data from UAVs, satellites, ground sensors, and weather stations, enabling precise identification of crop health and yield anomalies. Complementing this, AROS leverages real-time environmental feedback and multi-objective optimization to dynamically allocate resources, balancing yield maximization, cost efficiency, and environmental sustainability. Experimental results demonstrate that this approach not only improves anomaly detection accuracy but also enhances decision-making in precision agriculture, establishing a new benchmark for sustainable, data-driven crop protection strategies.

Despite its advantages, the framework has two notable limitations. The reliance on multi-modal data fusion requires high-quality and diverse data from multiple sources, which may be unavailable or inconsistent in regions with limited technological infrastructure. To address this, future research could focus on developing self-learning mechanisms that adapt to incomplete or noisy data. The computational demands of IMSFNet and AROS might limit their scalability in large-scale agricultural operations or resource-constrained environments. Optimizing the framework's architecture through model compression or distributed computing strategies could alleviate this issue. Addressing these challenges would further enhance the framework's applicability, making it a practical and scalable tool for precision agriculture worldwide.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

WC: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – review & editing. YS: Investigation, Data curation, Writing – original draft, Writing – review & editing. MZ: Visualization, Supervision, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The phased research results of the "AI+'New Business' Digital Intelligence Talent Training and Research Base" (Project No. JD2024Z027) at Chongqing Education Research Experimental Base.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Alemohammad, H. (2021). "The case for open-access ml-ready geospatial training data," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (IEEE)*. 1146–1148. Available online at: https://ieeexplore.ieee.org/abstract/document/9553977/.

Amjoud, A. B., and Amrouch, M. (2023). Object detection using deep learning, cnns and vision transformers: A review. *IEEE Access* 11, 35479–35516. doi: 10.1109/ACCESS.2023.3266093

Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. (2020). Usad: Unsupervised anomaly detection on multivariate time series. *Knowledge Discov. Data Min.* doi: 10.1145/3394486

Batzner, K., Heckler, L., and König, R. (2023). Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *IEEE Workshop/Winter Conf. Appl. Comput. Vision*. Available online at: https://openaccess.thecvf.com/content/WACV2024/html/Batzner_EfficientAD_Accurate_Visual_Anomaly_Detection_at_Millisecond-Level_Latencies_WACV_2024_paper.html.

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2021). The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vision*. doi: 10.1007/s11263-020-01400-4

Bozdal, M., Ileri, K., and Ozkahraman, A. (2024). Comparative analysis of dimensionality reduction techniques for cybersecurity in the swat dataset. *J. Supercomputing* 80, 1059–1079. doi: 10.1007/s11227-023-05511-w

Chen, L.-W., and Rudnicky, A. (2023). "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE)*. 1–5. Available online at: https://ieeexplore.ieee.org/abstract/document/10095036/.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2020). Padim: a patch distribution modeling framework for anomaly detection and localization. *ICPR Workshops*. Available online at: https://link.springer.com/chapter/10.1007/978-3-030-68799-1_35.

DeMedeiros, K., Hendawi, A. M., and Alvarez, M. (2023). A survey of ai-based anomaly detection in iot and sensor networks. *Ital. Natl. Conf. Sensors*. doi: 10.3390/s23031352

Deng, A., and Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. *AAAI Conf. Artif. Intell.* doi: 10.1609/aaai.v35i5.16523

Deng, H., and Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR52688.2022.00951

Feng, J., Hong, F.-T., and Zheng, W. (2021). Mist: Multiple instance self-training framework for video anomaly detection. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR46437.2021.01379

Friday, I. K., Godslove, J. F., Nayak, D. S. K., and Prusty, S. (2022). "Irgm: An integrated rnn-gru model for stock market price prediction," in *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS) (IEEE)*. 129–132. Available online at: https://ieeexplore.ieee.org/abstract/document/10076565/.

Gao, Z.-M., Hu, W., Dong, X.-Y., Zhao, X.-Y., Wang, S., Chen, J., et al. (2024a). Motion behavior of droplets on curved leaf surfaces driven by airflow. *Front. Plant Sci.* 15, 1450831. doi: 10.3389/fpls.2024.1450831

Gao, Z.-M., Hu, W., Dong, X.-Y., Zhao, X.-Y., Wang, S., Chen, J., et al. (2024b). Motion behavior of droplets on curved leaf surfaces driven by airflow. *Front. Plant Sci.* 15, 1450831. doi: 10.3389/fpls.2024.1450831

Grover, K., Kaur, K., Tiwari, K., Rupali,, and Kumar, P. (2021). "Deep learning based question generation using t5 transformer," in *Advanced Computing: 10th International Conference, IACC 2020* (Springer, Panaji, Goa, India), 243–255. December 5–6, 2020, Revised Selected Papers, Part I 10.

Gudovskiy, D. A., Ishizaka, S., and Kozuka, K. (2021). Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *IEEE Workshop/Winter Conf. Appl. Comput. Vision*. Available online at: https://openaccess.thecvf.com/content/WACV2022/html/Gudovskiy_CFLOW-AD_Real-Time_Unsupervised_Anomaly_Detection_With_Localization_via_Conditional_Normalizing_WACV_2022_paper.html?ref=https://githubhelp.com.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. (2022). Adbench: Anomaly detection benchmark. *Neural Inf. Process. Syst*. doi: 10.2139/ssrn.4266498

Jailani, N. S. J., Muhammad, Z., Rahiman, M. H. F., and Taib, M. N. (2021). "Analysis of kaffir lime oil chemical compounds by gas chromatography-mass spectrometry (gc-ms) and z-score technique," in *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS) (IEEE)*. 110–113. Available online at: https://ieeexplore.ieee.org/abstract/document/9495909/.

Jiang, J., Zhu, J., Bilal, M., Cui, Y., Kumar, N., Dou, R., et al. (2023). Masked swin transformer unet for industrial anomaly detection. *IEEE Trans. Industrial Inf.* doi: 10.1109/TII.2022.3199228

Kurita, T. (2021). "Principal component analysis (pca)," in *Computer vision: a reference guide* (Springer), 1013–1016. Available online at: https://link.springer.com/content/pdf/10.1007/978-3-030-63416-2_649.pdf.

Le, V.-H., and Zhang, H. (2021). Log-based anomaly detection without log parsing. *Int. Conf. Automated Software Eng*. doi: 10.1109/ASE51524.2021.9678773

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR46437.2021.00954

Liu, Y., Fang, S.-S., Zhao, R.-S., Liu, B., Jin, Y.-Q., and Li, Q. (2024). Nab-paclitaxel plus platinum versus paclitaxel plus platinum as first-line therapy in patients with metastatic or recurrent cervical cancer. *J. Cancer Res. Clin. Oncol.* 150, 321. doi: 10.1007/s00432-024-05825-z

Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., and Karypis, G. (2021). Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Trans. Neural Networks Learn. Syst*. Available online at: https://ieeexplore.ieee.org/abstract/document/9395172/.

Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., et al. (2023a). Deep industrial image anomaly detection: A survey. *Mach. Intell. Res*. doi: 10.1007/s11633-023-1459-z

Liu, Z., Zhou, Y., Xu, Y., and Wang, Z. (2023b). Simplenet: A simple network for image anomaly detection and localization. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR52729.2023.01954

Mathai, M., Liu, Y., and Ling, N. (2024). A hybrid transformer-lstm model with 3d separable convolution for video prediction. *IEEE Access*. doi: 10.1109/ACCESS.2024.3375365

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., and Foresti, G. (2021). Vt-adl: A vision transformer network for image anomaly detection and localization. *Int. Symposium Industrial Electron*. doi: 10.1109/ISIE45552.2021.9576231

Nassif, A. B., Shahin, I., Bader, M., Ahmed, A., and Werghi, N. (2024). Vit-lstm synergy: a multifeature approach for speaker identification and mask detection. *Neural Computing Appl.* 36, 22569–22586. doi: 10.1007/s00521-024-10389-7

Ni, J., Dong, J., Zhang, J., Pang, F., Cao, W., and Zhu, Y. (2016). The spectral calibration method for a crop nitrogen sensor. *Sensor Rev.* 36, 48–56. doi: 10.1108/SR-04-2015-0051

Ni, J., Yao, L., Zhang, J., Cao, W., Zhu, Y., and Tai, X. (2017). Development of an unmanned aerial vehicle-borne crop-growth monitoring system. *Sensors* 17, 502. doi: 10.3390/s17030502

Ni, J., Zhang, J., Wu, R., Pang, F., and Zhu, Y. (2018). Development of an apparatus for crop-growth monitoring and diagnosis. *Sensors* 18, 3129. doi: 10.3390/s18093129

Park, H., Noh, J., and Ham, B. (2020). Learning memory-guided normality for anomaly detection. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR42600.2020

Peng, Y., Lee, J., and Watanabe, S. (2023). "I3d: Transformer architectures with input-dependent dynamic depth for speech recognition," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE)*. 1–5. Available online at: https://ieeexplore.ieee.org/abstract/document/10096662/.

Quaranta, L., Calefato, F., and Lanubile, F. (2021). "Kgtorrent: A dataset of python jupyter notebooks from kaggle," in *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR) (IEEE)*. 550–554. Available online at: https://ieeexplore.ieee.org/abstract/document/9463068/.

Rostamian, A., and O'Hara, J. G. (2022). Event prediction within directional change framework using a cnn-lstm model. *Neural Computing Appl.* 34, 17193–17205. doi: 10.1007/s00521-022-07687-3

Roth, K., Pemula, L., Zepeda, J., Scholkopf, B., Brox, T., and Gehler, P. (2021). Towards total recall in industrial anomaly detection. *Comput. Vision Pattern Recognition*. Available online at: http://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., and Rabiee, H. (2020). Multiresolution knowledge distillation for anomaly detection. *Comput. Vision Pattern Recognition*. Available online at: http://openaccess.thecvf.com/content/CVPR2021/html/Salehi_Multiresolution_Knowledge_Distillation_for_Anomaly_Detection_CVPR_2021_paper.html.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J., and Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. *IEEE Int. Conf. Comput. Vision*. doi: 10.1109/ICCV48922.2021.00493

Tien, T. D., Nguyen, A. T., Tran, N. H., Huy, T. D., Duong, S. T., Tr, C. D., et al. (2023). Revisiting reverse distillation for anomaly detection. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR52729.2023.02348

Touvron, H., Cord, M., and Jégou, H. (2022). "Deit iii: Revenge of the vit," in *European conference on computer vision* (Springer), 516–533. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-20053-3_30.

Tuli, S., Casale, G., and Jennings, N. (2022). Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endowment*. doi: 10.14778/3514061.3514067

Wang, L., and Koniusz, P. (2022). "Uncertainty-dtw for time series and sequences," in *European Conference on Computer Vision* (Springer), 176–195. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-19803-8_11.

Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., and Wang, C. (2023b). Multimodal industrial anomaly detection via hybrid fusion. *Comput. Vision Pattern Recognition*. doi: 10.1109/CVPR52729.2023.00776

Wang, D., Zhuang, L., Gao, L., Sun, X., Huang, M., and Plaza, A. (2023a). Bocknet: Blind-block reconstruction network with a guard window for hyperspectral anomaly detection. *IEEE Trans. Geosci. Remote Sens*. doi: 10.1117/12.3010359

Wattasseril, J. I., Shekhar, S., Döllner, J., and Trapp, M. (2023). "Zero-shot video moment retrieval using blip-based models," in *International Symposium on Visual Computing* (Springer), 160–171. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-47969-4_13.

Wen, X., and Li, W. (2023). Time series prediction based on lstm-attention-lstm model. *IEEE Access* 11, 48322–48331. doi: 10.1109/ACCESS.2023.3276628

Wyatt, J., Leach, A., Schmon, S. M., and Willcocks, C. G. (2022). Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. *2022 IEEE/CVF Conf. Comput. Vision Pattern Recognition Workshops (CVPRW)*. doi: 10.1109/CVPRW56347.2022.00080

Xiong, H., Li, J., Li, Z., and Zhang, Z. (2023). Gpr-gan: A ground-penetrating radar data generative adversarial network. *IEEE Trans. Geosci. Remote Sens.* 62, 1–14. Available online at: https://ieeexplore.ieee.org/abstract/document/10329345/.

Xu, J., Wu, H., Wang, J., and Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. *Int. Conf. Learn. Representations*. Available online at: https://arxiv.org/abs/2110.02642.

Yan, R., Schönlieb, C.-B., and Li, C. (2024). "Spatiotemporal graph neural network modelling perfusion mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 411–421. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-72069-7_39.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., et al. (2022). A unified model for multi-class anomaly detection. *Neural Inf. Process. Syst*. Available online at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/1d774c112926348c3e25ea47d87c835b-Abstract-Conference.html.

Zavrtanik, V., Kristan, M., and Skočaj, D. (2021). DrÆm – a discriminatively trained reconstruction embedding for surface anomaly detection. *IEEE Int. Conf. Comput. Vision*. doi: 10.1109/ICCV48922.2021.00822

Zhang, P., Liu, Y., and Du, H. (2024b). An integrated framework for uav-based precision plant protection in complex terrain: the achaga solution for multi-tea fields. *Front. Plant Sci.* 15, 1440234. doi: 10.3389/fpls.2024.1440234

Zhang, P., Liu, Y., and Du, H. (2024c). An integrated framework for uav-based precision plant protection in complex terrain: the achaga solution for multi-tea fields. *Front. Plant Sci.* 15, 1440234. doi: 10.3389/fpls.2024.1440234

Zhang, B., Zhang, P., Dong, X., Zang, Y., and Wang, J. (2024a). "Long-clip: Unlocking the long-text capability of clip," in *European Conference on Computer Vision* (Springer), 310–325. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-72983-6_18.

Zhu, H., Lin, C., Liu, G., Wang, D., Qin, S., Li, A., et al. (2024a). Intelligent agriculture: Deep learning in uav-based remote sensing imagery for crop diseases and pests detection. *Front. Plant Sci.* 15, 1435016. doi: 10.3389/fpls.2024.1435016

Zhu, H., Lin, C., Liu, G., Wang, D., Qin, S., Li, A., et al. (2024b). Intelligent agriculture: Deep learning in uav-based remote sensing imagery for crop diseases and pests detection. *Front. Plant Sci.* 15, 1435016. doi: 10.3389/fpls.2024.1435016

Zou, Y., Jeong, J., Pemula, L., Zhang, D., and Dabeer, O. (2022). "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*. Available online at: https://link.springer.com/chapter/10.1007/978-3-031-20056-4_23.