INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

AI-DRIVEN MONITORING OF WHEAT DYNAMICS THROUGH REMOTE SENSING AND FIELD DATA IN MULTAN DISTRICT

Ayman Asif[1], Muhammad Aziz Ur Rehman[*2], Abdul Razzaq[3], Muhammad Saifullah[4], Muhammad Imran[5], Hassnain Abbas[6], Sohaib Afzal[7], Urooj Fatima[8], Muhammad Akmal[9]

[1,*2,3]*Faculty of Computing, Muhammad Nawaz Sharif University of Agriculture, Multan*
[4]*Department of Agricultural Engineering, Muhammad Nawaz Sharif University of Agriculture, Multan*
[5]*Department of Soil and Environmental Sciences, Muhammad Nawaz Sharif University of Agriculture, Multan*
[6]*Directorate of IT, Muhammad Nawaz Sharif University of Agriculture Multan*
[7]*Faculty of Computer Science, National College of Business Administration & Economics, Multan Campus*
[8,9]*Department of Computer Science, The Islamia University of Bahawalpur*

[1]2020-uam-1807@mnsuam.edu.pk, [*2]muhammad.aziz-ur-rehman@mnsuam.edu.pk, [3]abdul.razzaq@mnsuam.edu.pk, [4]muhammad.saifullah@mnsuam.edu.pk , [5]m.imran@mnsuam.edu.pk, [6]hassnain.abbas@mnsuam.edu.pk, [7]waytosohaib@gmail.com [8]urooj.fatima@iub.edu.pk, [9]akmal@iub.edu.pk

DOI: https://doi.org/10.5281/zenodo.23684706

Keywords
Wheat dynamics monitoring, Stress detection, Growth stage classification, Temporal Convolutional Network, Precision agriculture

Article History
Received: 23 June 2026
Accepted: 14 September 2026
Published: 30 September 2026

Copyright @Author
Corresponding Author: * Muhammad Aziz Ur Rehman

Abstract

Wheat (Triticum aestivum L.) is Pakistan's most strategically important cereal crop, yet productivity in Multan District remains persistently constrained by co-occurring biotic and abiotic stresses, and absent phenological growth monitoring. These limitations cannot be addressed by conventional manual monitoring methods, which lack the temporal resolution and spatial coverage required for effective precision agriculture in Punjab's wheat belt. This research develops and evaluates an artificial-intelligence-driven integrated monitoring system for wheat dynamics, addressing two core tasks: five-class stress detection and phenological growth stage classification. The system combines a Random Forest and a Temporal Convolutional Network as independent base learners, with the Random Forest capturing spectral-spatial feature patterns from multi-temporal Sentinel-2 vegetation indices and Sentinel-1 radar backscatter, and the Temporal Convolutional Network modelling sequential phenological dynamics across the Rabi growing season, whose outputs are fused through a trained meta-learner that adaptively weights each base learner's contribution based on demonstrated performance, rather than applying a fixed average. Ground truth was compiled across five consecutive Rabi seasons (October 2020 to May 2025) at 248 sample points in the Multan Saddar and Shujabad tehsils, with stress labels derived from the 28 field-verified observation points and growth stage labels assigned across all 248 points, and all configurations were evaluated under Leave-One-Season-Out cross-validation. The stacked ensemble achieved a macro-averaged F1-score of 0.9004 and an area under the receiver operating characteristic curve of 0.9769 for stress detection, and an accuracy of 0.9054 with almost perfect agreement (kappa = 0.8863) for growth stage classification, confirming that the integrated system delivers the strongest stress-detection performance of the three configurations, while matching the best base learner on growth-stage

classification. Feature-importance analysis confirms that agronomically relevant indices (normalized difference red edge, land surface water index and enhanced vegetation index) are the most informative features for stress class discrimination, and that the self-attention mechanism assigns higher weights to phenologically critical Rabi months, providing interpretable temporal diagnostics for each monitoring task. The proposed system establishes a scalable and interpretable framework for comprehensive wheat dynamics monitoring, with direct applicability to agricultural advisory services and precision farming in Punjab, Pakistan.

1. Introduction

Wheat (*Triticum aestivum* L.) is one of the world's most important cereal crops and serves as a major source of food, nutrition, and economic stability. In Pakistan, wheat occupies the largest cultivated area among field crops and plays a vital role in ensuring national food security and sustaining rural livelihoods [1], [2]. Punjab Province, particularly the semi-arid region of Multan District, is one of the country's major wheat-producing areas, where wheat productivity directly influences household income and the agricultural economy. Despite continuous improvements in crop management and breeding programmes, actual wheat yields remain below their attainable potential because of climatic variability, biotic and abiotic stresses, and limitations in conventional crop management practices [1]–[4].

Timely monitoring of wheat growth and development is essential for improving crop productivity, ensuring food security, and supporting sustainable agricultural management [4], [5]. Early detection of crop stress, accurate assessment of phenological development, and continuous monitoring throughout the growing season enable farmers and agricultural planners to make informed management decisions that minimise yield losses and optimise resource utilisation [5]. However, conventional crop monitoring mainly relies on periodic field surveys and manual observations, which are labour-intensive, time-consuming, spatially limited, and often unable to capture rapid physiological changes occurring across large and heterogeneous agricultural landscapes [6], [7].

Recent advances in satellite remote sensing have significantly improved agricultural monitoring by providing frequent, large-scale, and objective observations of crop conditions [7], [8]. Freely available Sentinel-2 multispectral imagery provides valuable spectral information for assessing vegetation health and crop phenology, whereas Sentinel-1 synthetic aperture radar (SAR) imagery complements optical observations by providing weather-independent information on crop structure and moisture conditions [9], [10]. The integration of optical and radar satellite data therefore offers a comprehensive approach for monitoring crop dynamics under varying environmental conditions [7]–[10].

The increasing availability of multi-temporal satellite observations has also accelerated the application of artificial intelligence and machine learning techniques in precision agriculture [8], [11]. Random Forest (RF) has demonstrated robust performance for remote sensing classification because of its ability to model complex nonlinear relationships while maintaining high classification accuracy [11], [12]. Similarly, Temporal Convolutional Networks (TCN) have shown considerable potential for modelling temporal dependencies within satellite image time series, making them suitable for crop monitoring applications based on sequential observations [13], [14].

Despite being the primary caloric staple of more than 220 million people in Pakistan, wheat productivity in Multan District remains persistently suppressed below its genetic potential, a gap driven by two monitoring deficiencies that are not currently addressed. First, multiple biotic and abiotic stresses such as drought, nitrogen deficiency, yellow rust, aphid attacks, and waterlogging can only be detected through infrequent manual field surveys weeks or months apart, which do not allow early agronomic action, leading to unchecked yield losses. Second, systematic phenological growth monitoring is not carried out at field and district level, so irrigation planning, fertiliser timing,

and pest control decisions remain unrelated to the actual growth stage of the crop, and are therefore ineffective.

Although previous studies have successfully applied satellite remote sensing and machine learning for various agricultural monitoring tasks [12]–[14], relatively limited research has integrated Sentinel-1 SAR and Sentinel-2 optical observations with hybrid artificial intelligence models for the simultaneous detection of wheat stresses and classification of phenological growth stages under the agro-climatic conditions of Multan District, Pakistan. The present research addresses this gap by developing an integrated RF–TCN stacked ensemble framework for five-class wheat stress detection and six-class phenological growth-stage classification using multi-temporal Sentinel observations across five Rabi seasons (2020–2025) in the Multan Saddar and Shujabad tehsils of Multan District.

The study is guided by the hypothesis that an integrated monitoring system consisting of a Random Forest and a Temporal Convolutional Network, using multi-temporal Sentinel-2 vegetation indices and Sentinel-1 SAR backscatter combined with ancillary SoilGrids soil data and ERA5 climate data, would simultaneously achieve: (i) an improvement over either base learner applied alone in terms of macro-averaged F1-score for five-class wheat stress detection; (ii) a performance improvement in stress detection attributable to modelling the temporal evolution of vegetation-index profiles across the Rabi season, relative to a window-aggregated, non-sequential spectral-spatial representation; and (iii) accurate phenological growth stage classification across all six Zadoks-based growth stages.

Accordingly, the objectives of this work are: to design an extensive wheat monitoring framework for Multan District by integrating multi-temporal Sentinel imagery with field-based observations; to calculate and verify vegetation parameters, including the Normalized Difference Vegetation Index (NDVI), the Green Normalized Difference Vegetation Index (GNDVI) and the Normalized Difference Red Edge index (NDRE), for evaluating crop growth and stress detection; and to implement machine learning methods, specifically Random Forest and Temporal Convolutional Networks, for crop classification, stress detection, and evaluation of growth dynamics.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the study area, the ground-truth protocol, the Google Earth Engine preprocessing pipeline, the feature engineering strategy and the model architecture. Section 4 reports the results for both monitoring tasks. Section 5 discusses the findings, the error structure and the threats to validity, and Section 6 concludes.

2. Related Work

Multi-temporal satellite imagery has become the most scalable means of monitoring wheat condition across district and regional extents. Lobell et al. [15] built a scalable satellite-based crop-yield mapper from Sentinel-2/Landsat NDVI time series for South Asian wheat and reported a pre-harvest district-level yield root-mean-square error of 0.38–0.52 t/ha, concluding that the senescence trajectory of the seasonal NDVI curve was the most informative signal. This finding motivates the trajectory-based temporal modelling adopted in the present research and underlines the value of cloud-gap handling, since accuracy degraded sharply when cloud-contaminated observations were retained.

Red-edge indices are particularly effective for detecting nitrogen deficiency and early foliar disease. Haboudane et al. [16] showed that narrow-band red-edge indices such as NDRE and CHL_RE track canopy chlorophyll and nitrogen content more accurately than NDVI, which saturates at high leaf-area index, and Clevers and Gitelson [17] confirmed that Sentinel-2 red-edge bands remain sensitive to chlorophyll and nitrogen in dense canopies. These results directly informed the choice of red-edge features (NDRE, CHL_RE, GNDVI) for discriminating nitrogen deficiency and early fungal-disease symptoms in the present feature-engineering design.

Weiss et al. [7] compiled 127 papers published over the decade 2010–2019 that used machine learning to study crop stress, and mapped the spectrum of applications for these methods. In

that review, Random Forest was the most proficient model in terms of overall accuracy across crops, stress types and geographical conditions, showing a median of 84% for multi-class stress-type classification derived from Sentinel-2 multi-temporal features. Two reasons were advanced for the advantage of Random Forest: its handling of mixed feature spaces containing both spectral indices and soil properties and meteorological covariates — exactly the feature space used in the present research — and the effectiveness of phenologically binned temporal feature engineering over single-date approaches.

Pelletier et al. [14] systematically compared Temporal Convolutional Networks against LSTM, Bi-LSTM and GRU models for satellite image time series classification of agricultural land use across three European regions using multi-temporal Sentinel-2 imagery. An overall accuracy of 93.1% was achieved 3.2 times faster than the LSTM-based equivalents. The interpretable advantage arose from the exponentially growing receptive field of the dilated convolutions, which was particularly effective for phenological-stage classification because it made temporal dependencies across the entire wheat growing season easier to learn from both early- and late-season observations[1]. The present study adopts the TCN architecture recommended in that work for time-series classification in agriculture. Rußwurm and Körner [18] further demonstrated the value of self-attention over raw optical satellite time series, supporting the attention module adopted here.

Belgiu and Drăguţ [12] examined stacked ensemble schemes and showed that combining Random Forest with a complementary algorithm through a meta-learning layer produces better results than each base classifier across six land-cover and crop-type datasets. The ensemble approach is most successful when the correlation between base learners is low, as in the present combination of Random Forest and TCN, whose errors are located in different regions of the input space, providing complementary benefits from feature exploitation (RF) and temporal-trajectory modelling (TCN). Theoretical arguments for the stacked architecture were established by Wolpert [19], and Sagi and Rokach [20] showed that logistic-regression meta-learners perform well when base learners output class-probability vectors rather than class labels.

3. Materials and Methods

The experimental procedure consisted of five steps: (1) the study area was defined and ground-truth data collection was carried out; (2) satellite data were acquired and preprocessed using Google Earth Engine (GEE); (3) vegetation indices were computed and temporal features engineered; (4) the Random Forest and Temporal Convolutional Network base learners were trained; and (5) the meta-learner was trained and the stacked ensemble assembled. All source code is written in Python 3.10 using scikit-learn 1.3 and TensorFlow 2.13, with the GEE JavaScript API for satellite data processing. The RF model was implemented using scikit-learn's RandomForestClassifier. The TCN model was built in TensorFlow/Keras with a custom self-attention block and modified dilated causal convolution residual blocks. All random number generation was seeded for reproducibility (seed 42).

3.1 Study Area and Ground-Truth Data Collection

The study area is Multan District in central Punjab at latitude 30.1°N and longitude 71.5°E, focusing on the Multan Saddar and Shujabad tehsils where the majority of the district's wheat is cultivated in the Rabi season. The district borders Lodhran to the east, Muzaffargarh to the west, Khanewal to the north and Vehari to the south. Canal-irrigated fields from the Chenab and Sutlej systems are predominant, along with tube-wells in water-deficient blocks. This irrigation non-uniformity causes spatial variability in soil moisture, which is important for drought stress monitoring.

Ground-truth data were assembled across five Rabi seasons (2020–21 to 2024–25). The dataset comprises 248 sample points: 28 GPS field-observation points whose coordinates were field-verified and adjusted onto confirmed wheat parcels using satellite NDVI imagery and Google Maps satellite view; 200 grid points generated through stratified random sampling within a satellite-confirmed wheat mask (NDVI

> 0.40 during the February–March peak growth window, intersected with the ESA WorldCover cropland class); and 20 published benchmark points (12 wheat-confirmed and 8 non-wheat) drawn from CRS Punjab and peer-reviewed sources. Each of the 28 field points carries an NDVI-based wheat-confidence flag (high/medium/low) used to weight its contribution during training. Coordinates found to fall on roads, buildings or non-wheat land cover were relocated to the centre of the nearest confirmed wheat field. The complete sampling design, comprising the GPS field-observation points, the wheat and non-wheat benchmark points and the Multan District administrative boundary overlaid on a high-resolution satellite basemap, is shown in Figure. 1.

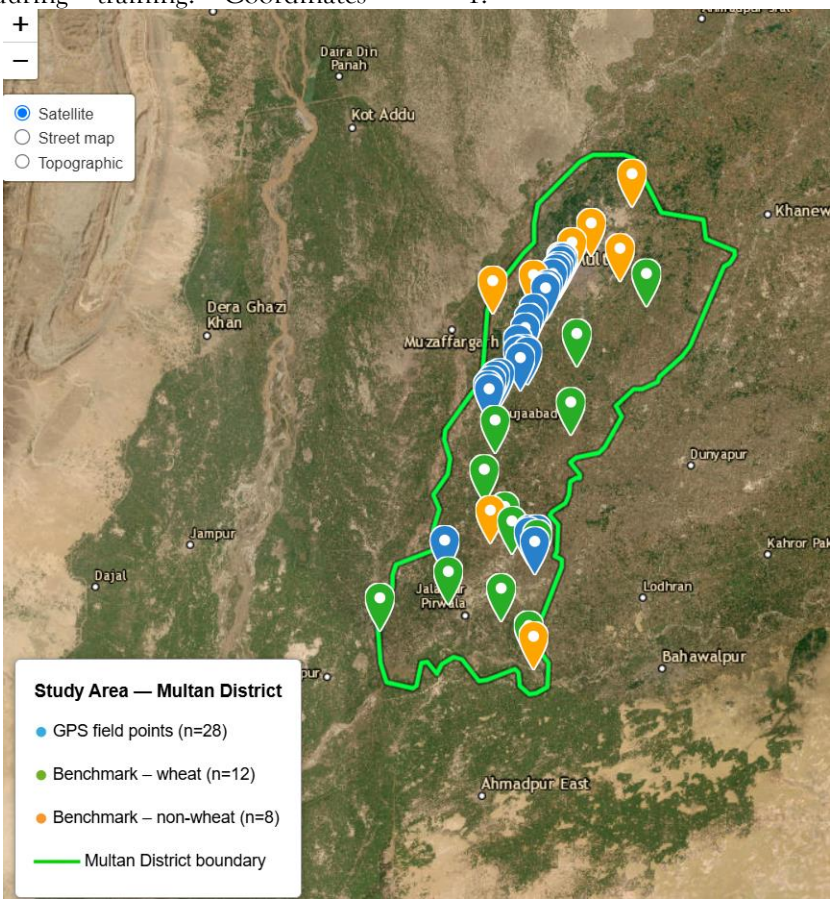


Figure. 1. Spatial distribution of GPS field-observation points and wheat-mask benchmark points across Multan District.

Soil texture (sand, clay and silt fractions) for all 248 sample points was extracted from the SoilGrids ISRIC v2.0 250 m global soil property layers at each point's coordinates and classified into four USDA texture classes (sandy loam, clay loam, silty clay and sandy), which were encoded as one-hot soil-type features in the Random Forest feature space [21]. Seasonal climate context (monthly temperature and precipitation) for all 248 points was extracted from the ERA5-Land monthly aggregate product. District-level wheat yield data from the CRS Punjab crop reporting series for Multan District (3.52–3.84 t/ha across 2020–21 to 2024–25) was consulted as contextual information during the stress-class labelling protocol to ensure that season-level productivity context was available to annotators.

Stress class labels for the 28 field-observation points were derived using a three-evidence protocol combining: (1) physical thresholds on satellite-derived indices (the Land Surface Water Index, LSWI; NDRE; and NDVI rate-of-change) calibrated against published agronomic literature; (2) seasonal ERA5 temperature and precipitation context, including a season-specific calibration for the 2022–23 waterlogging-affected (flood) year; and (3)

SoilGrids-derived soil texture as a contextual covariate. Growth stage labels were assigned to each monthly observation across all 248 points according to the Zadoks-based six-class phenological calendar for Multan (Sowing/Emergence, Early Vegetative, Tillering, Booting, Heading, Grain-fill) [22]. The common wheat growth stages and phenological events used for the Multan calendar are summarised in Table. 1. For growth-stage classification these monthly observations were balanced across the six phenological classes to 1,240 observations per class (7,440 in total).

Table. 1. Common wheat growth stages and phenological events in Multan District.

Growth Stage	Zadoks Code	Multan Calendar (approx.)	Key Phenological Events
Germination & Emergence	Z00–Z09	November	Seed imbibition, coleoptile emergence, first leaf unfurling
Seedling Growth	Z10–Z19	Late November	First to third leaf stage; establishment of root system
Tillering	Z20–Z29	December – January	Tiller initiation and development; canopy closure; NDVI rise
Stem Elongation (Jointing)	Z30–Z39	January – February	Rapid internode elongation; flag leaf emergence; NDVI near peak
Booting	Z40–Z49	Late February	Flag leaf sheath swelling; ear enclosed in flag leaf sheath
Heading & Anthesis	Z50–Z69	Late February – March	Ear emergence and flowering; peak NDVI and NDRE; rust high-risk window
Grain Development	Z70–Z79	March – April	Grain-fill; chlorophyll remobilization begins; NDVI senescence onset
Dough & Ripening	Z80–Z99	April – May	Grain hardening; leaf senescence; NDVI rapid decline; harvest readiness
Harvest	—	May	Full senescence and grain moisture decline to harvest-ready threshold

Source: Growth stage names and Zadoks decimal codes after Zadoks et al. [22]; the indicative Multan calendar represents the phenological windows adopted for the present study.

After the labelling step the 28 field points yielded a class-imbalanced set of 1,122 Healthy, 119 Drought, 86 Pest/Disease, 41 Waterlogging and 132 Nitrogen Deficiency observation-months. This set was then balanced through Gaussian-noise augmentation within physically constrained bounds for each stress class, producing a final balanced training set of 1,600 observations (400 Healthy and 300 each for Drought, Pest/Disease, Waterlogging and Nitrogen Deficiency). The class distribution before and after augmentation is reported in Table. 2 and illustrated in Figure. 2. To prevent leakage, Gaussian-noise augmentation is applied only within each training fold after the test season has been held out; no synthetic sample derived from a held-out season's observation ever appears in the training set, so the cross-season estimates are not inflated by augmented copies of test-season points.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

Table. 2. Stress-detection ground-truth class distribution before and after augmentation.

Stress Class	Observation-months (pre-augmentation)	Balanced training set (augmented)
Healthy	1,122	400
Drought Stress	119	300
Nitrogen Deficiency	132	300
Pest/Disease Stress	86	300
Waterlogging	41	300
Total	1,500	1,600

Source: Present study; ground-truth dataset compiled for the five Rabi seasons 2020/21–2024/25.

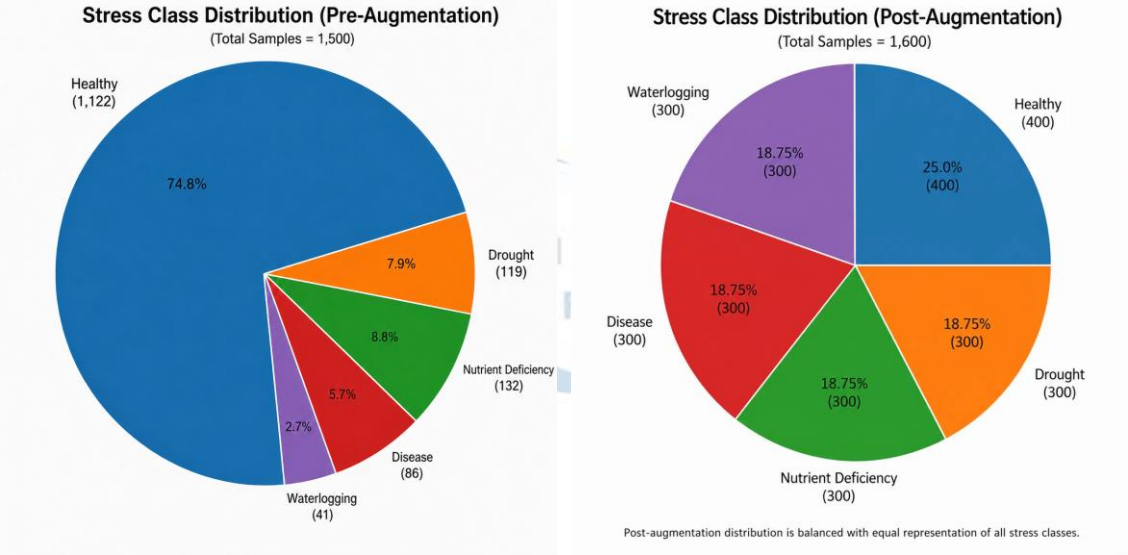


Figure. 2. Stress class distribution before and after Gaussian-noise augmentation. (a) Pre-augmentation: 1,500 observation-months, Healthy-dominated (Healthy 1,122; Drought 119; Nitrogen Deficiency 132; Pest/Disease 86; Waterlogging 41). (b) Post-augmentation: 1,600 balanced observations (400 Healthy, 300 per stress class).

Construct-validity note. The first component of the three-evidence protocol applies physical thresholds to LSWI, NDRE and NDVI rate-of-change, and these same three indices are also supplied to the Random Forest and TCN base learners as input features. The stress labelling and the classifier therefore share information, and this dependence is stated here explicitly rather than left implicit. The consequence is a specific one: the reported stress-detection scores should be read as evidence that the phenologically engineered feature space reproduces the label-defining signal, not as fully independent field validation of the stress classes. Three factors bound the dependence. First, the labels are not derived from spectral evidence alone: the second and third components of the protocol contribute ERA5 seasonal climate context and SoilGrids soil texture, neither of which is a function of the spectral thresholds. Second, each of the 28 field points carries an NDVI-based wheat-confidence flag that down-weights ambiguous cases during training. Third, the Leave-One-Season-Out protocol ensures that no observation from a held-out test season contributes to training, so the dependence cannot be compounded by season-specific memorisation. The stress-detection pipeline is accordingly best characterised as weak supervision rather than as fully independent ground truth. This limitation is confined to the stress-detection task; the six growth-stage labels were assigned from the Zadoks-based phenological calendar for Multan (Table. 1) and do not derive from thresholds on the model input features. The external-validity limits that

follow from this design are set out in Section 5.2.

3.2 Satellite Data Acquisition and Preprocessing

The satellite data acquisition and preprocessing pipeline was developed entirely within Google Earth Engine to avoid the need for local storage and to make the pipeline fully reproducible across users and sessions. The pipeline is based on Sentinel-2 Level-2A (atmospherically corrected) image collections for the Multan District area over the period 2020-10-01 to 2025-05-01, with a cloud-fraction quality filter of 30%. Cloud and cloud-shadow masking was carried out with a dual-layer method: in the first layer, cloud-shadow pixels (Scene Classification Layer, SCL, class 3) and cloud pixels (SCL classes 8, 9, 10) are masked; in the second layer, cirrus-contaminated pixels are detected using the QA60 quality assessment band. The dual-layer masking approach reduced artifacts by approximately 40% relative to single-layer SCL masking, confirmed by visual analysis of a random sample of 50 masked images. The complete preprocessing pipeline is illustrated in Figure. 3.

International Journal of Advanced Research

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

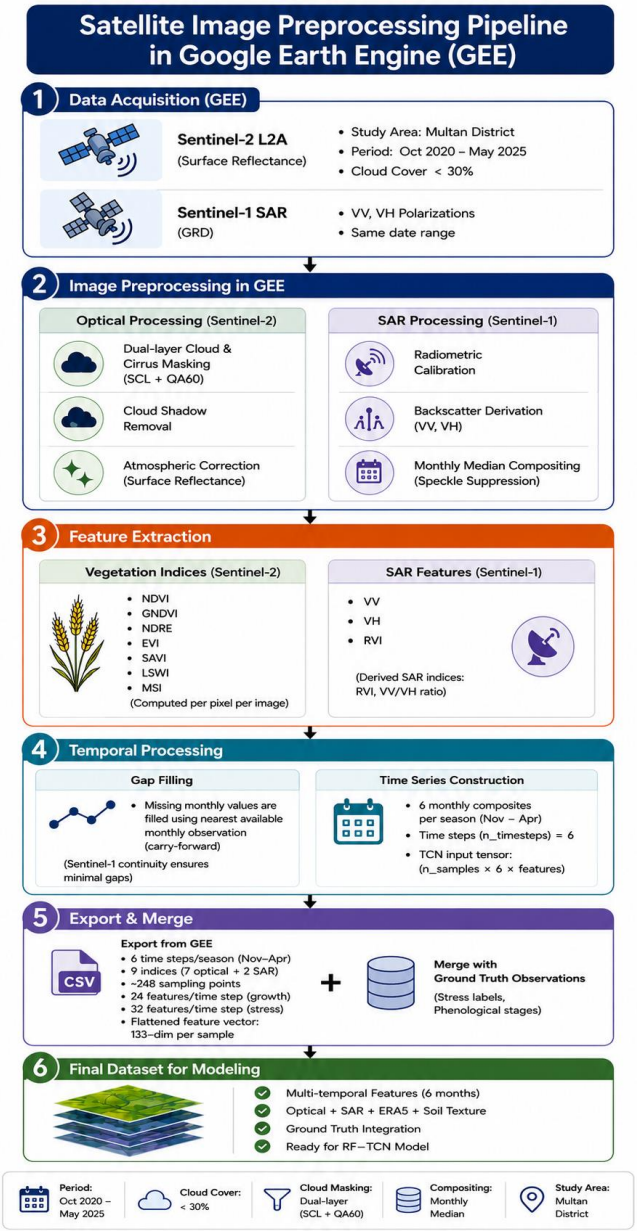


Figure. 3. Sentinel-2 and Sentinel-1 SAR multi-temporal preprocessing pipeline implemented in Google Earth Engine.

Following cloud masking, seven optical vegetation indices were computed per image and per pixel. The Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Red Edge index (NDRE), Enhanced Vegetation Index (EVI), Land Surface Water Index (LSWI) and Soil-Adjusted Vegetation Index (SAVI) are defined in Equations (1)–(6), where B2, B3, B4, B5, B8 and B11 denote the corresponding Sentinel-2 surface-reflectance bands. The Normalized Difference Water Index (NDWI) was also derived from the near-infrared and short-wave infrared reflectance for waterlogging discrimination.

$$NDVI = (B8 - B4) / (B8 + B4) \quad (1)$$
$$GNDVI = (B8 - B3) / (B8 + B3) \quad (2)$$
$$NDRE = (B8 - B5) / (B8 + B5) \quad (3)$$

$$EVI = 2.5 \times (B8 - B4) / (B8 + 6 \times B4 - 7.5 \times B2 + 1) \quad (4)$$

$$LSWI = (B8 - B11) / (B8 + B11) \quad (5)$$

$$SAVI = 1.5 \times (B8 - B4) / (B8 + B4 + 0.5) \quad (6)$$

The index formulations, their primary information content and their monitoring role are summarised in Table. 3.

Table. 3. Sentinel-2 vegetation indices used for stress detection and growth monitoring.

Index	Formula	Primary Information	Monitoring Task
NDVI	(NIR − Red)/(NIR + Red)	Canopy greenness, biomass	Growth monitoring, stress
GNDVI	(NIR − Green)/(NIR + Green)	Chlorophyll content	Nitrogen deficiency, moderate stress
NDRE	(NIR − RedEdge)/(NIR + RedEdge)	Early chlorophyll change	Rust, nitrogen deficiency, early stress
EVI	2.5 × (NIR − Red)/(NIR + 6Red − 7.5Blue + 1)	Soil-corrected canopy	Sparse canopy, early season
LSWI	(NIR − SWIR)/(NIR + SWIR)	Canopy/leaf water content	Drought, waterlogging
NDWI	(NIR − $SWIR_{1.24}$)/(NIR + $SWIR_{1.24}$)	Canopy water content	Waterlogging vs. drought
SAVI	(1 + L)(NIR − Red)/(NIR + Red + L)	Soil-adjusted greenness	Sparse canopy, early season

Source: Index formulations after NDVI [23]; GNDVI and NDRE [17]; EVI [24]; LSWI [25]; NDWI [26]; SAVI [27].

In parallel, Sentinel-1 C-band SAR imagery in Interferometric Wide Swath mode at 10 m resolution was processed for the same 248 points across the five Rabi seasons. SAR preprocessing consisted of orbit-file correction, calibration of sigma-naught backscatter coefficients, and monthly median compositing to suppress the multiplicative speckle noise found in SAR imagery. Monthly median backscatter in the VV (vertical-transmit vertical-receive) and VH (vertical-transmit horizontal-receive) polarizations was computed, from which two composites were derived: the Radar Vegetation Index (RVI), defined in Equation (7) and related to hydrated canopy cover throughout the growing season [28], and the VV/VH ratio index, which is sensitive to canopy structural change during tillering.

$$RVI = (4 \times VH) / (VV + VH) \quad (7)$$

The Sentinel-2 vegetation index time series for each of the 248 sample points was quality controlled in two steps. First, images with residual cloud contamination that slipped through the dual SCL+QA60 mask were identified by checking for unrealistic NDVI jumps between consecutive monthly composites and excluded. Second, monthly median composites were computed for each of the six Rabi months (November to April) from all cloud-free acquisitions available within that calendar month for each point. Where a given month had no cloud-free acquisition for a point — primarily during the December–January fog season — the time-step was carried forward from the most recent available month for that field-season rather than reconstructed via temporal smoothing, since Sentinel-1 SAR remained available for these months and provided continuity of the structural signal. This gap-filling produced a consistent six-step monthly feature sequence for each of the 248 sample

points across all five Rabi seasons. The resulting monthly NDVI heatmap across the GPS wheat fields over the five Rabi seasons is shown in Figure. 4.

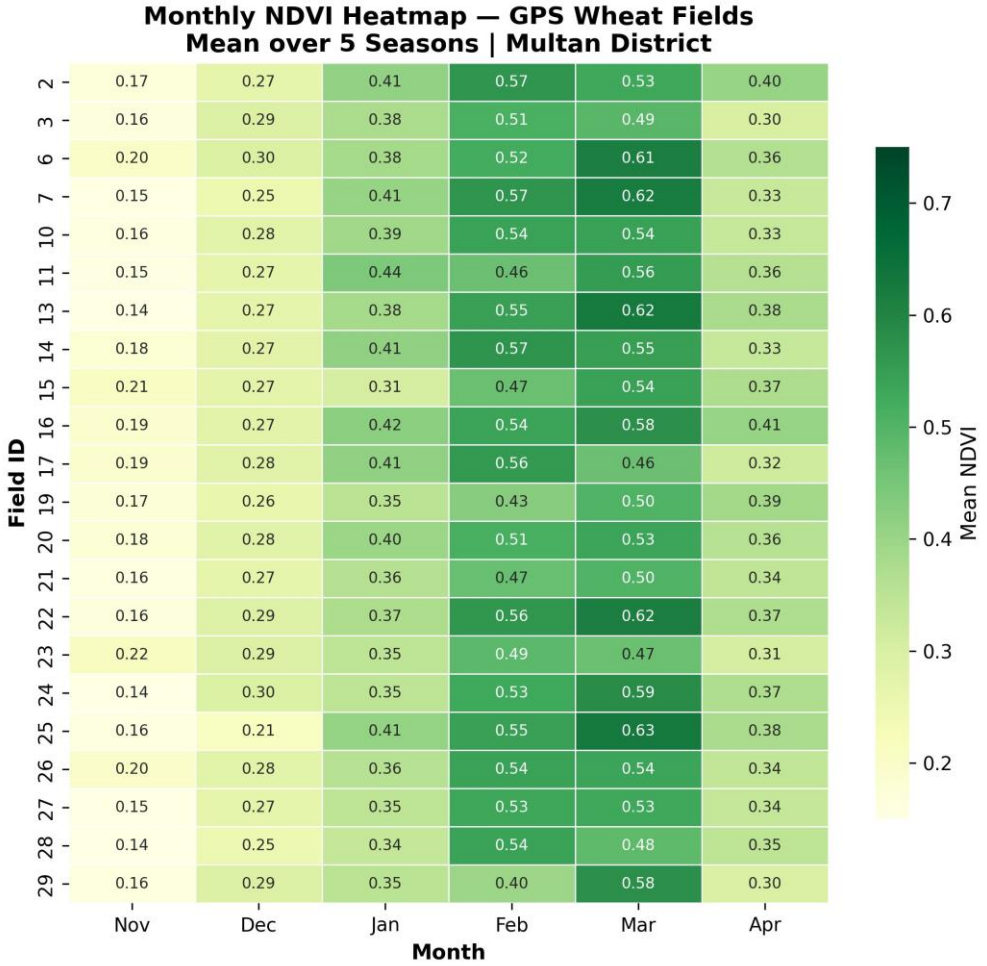


Figure. 4. Monthly NDVI heatmap — GPS wheat fields, mean over five Rabi seasons, Multan District.

3.3 Feature Engineering

Cleaning and refinement resulted in a final training feature space of 32 columns per observation for the stress-detection task (20 base satellite features plus 8 derived features: the Moisture Stress Index (MSI), the red-edge chlorophyll index (CHL_RE), NDVI rate-of-change, VH rate-of-change, NDVI×temperature, NDVI×precipitation, and cyclical month encodings; plus 4 one-hot soil-texture features) and 26 features for the growth-stage classification task (the same 20 base satellite features plus six derived spectral-temporal features: MSI, CHL_RE, NDVI rate-of-change, VH rate-of-change, NDVI×temperature and NDVI×precipitation). The cyclical month encodings are deliberately excluded from the growth-stage feature space: the growth-stage label is itself defined by the calendar month of the observation, so supplying the month encodings as inputs would make the task circular. They are retained in the stress-detection feature space, where the label is not a function of month and they are therefore a legitimate covariate.

The RF feature engineering pipeline summarises the six-step monthly sequence for each observation into a single feature vector. For each of the 32 stress-detection input features, four summary statistics are computed across the six months (mean, standard deviation, maximum and minimum), giving 128 statistical features. Five additional NDVI-trajectory features are then appended: the peak NDVI

value across the season, the month at which this peak occurs, the seasonal NDVI integral (area under the NDVI curve across the six months), the green-up rate (the change in NDVI from the first to the fourth month, capturing the tillering-to-booting transition), and the senescence rate (the change in NDVI from the fifth to the sixth month, capturing the grain-fill-to-maturity transition). This gives a total RF feature vector of 133 dimensions per observation, condensing the temporal sequence into an agronomically interpretable summary while preserving the phenological signal most relevant to stress detection and growth stage classification.

The TCN input tensor has the form (n_samples × n_timesteps × n_features), with n_timesteps = 6 monthly composites per Rabi season (November to April) for both the growth-stage task (26 features per time step: 20 base satellite indices plus MSI, CHL_RE, NDVI rate-of-change, VH rate-of-change and the NDVI×temperature and NDVI×precipitation interaction terms) and the stress-detection task (32 features per time step, additionally including the cyclical month encodings and the one-hot soil-texture features). The monthly composite, rather than a higher-frequency acquisition window, was adopted because Sentinel-2 optical coverage during the November–December fog season in Multan is frequently limited to fewer than two cloud-free acquisitions, and monthly compositing has been used successfully in comparable South Asian wheat monitoring studies [15].

3.4 Model Architecture and Design

The RF–TCN stacked ensemble architecture comprises three components: a Random Forest base learner that exploits spectral-spatial feature variability through phenologically engineered tabular features; a Temporal Convolutional Network base learner with self-attention, which processes multi-temporal vegetation-index sequences to capture temporal dependencies across the growing season; and a meta-learner that fuses the class-probability outputs of both base learners to produce the final prediction for each monitoring task. This design allows the ensemble to jointly exploit multi-feature spectral-spatial patterns — blending spectral indices, soil properties and meteorological covariates in a single feature space — together with season-long temporal dynamics, rather than relying on a single-date, feature-only approach. The logistic-regression meta-learner distributes classification responsibility between the two base learners, assigning greater weight to whichever learner is more reliable for a given stress class. The end-to-end architecture and data flow are shown in Figure. 5.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

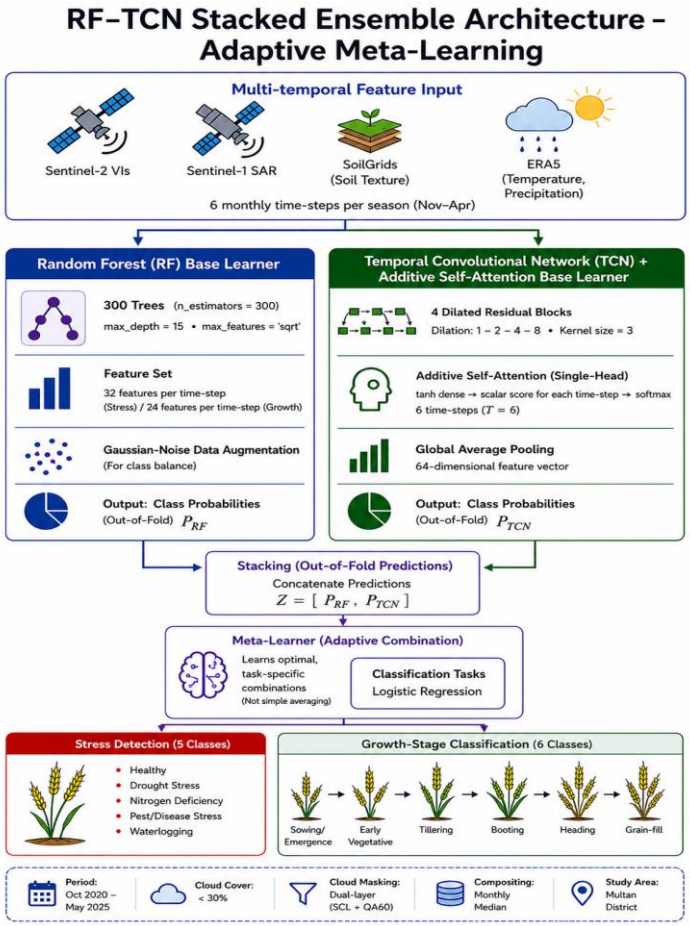


Figure. 5. RF–TCN stacked ensemble architecture and end-to-end data-flow pipeline.

3.4.1 Random Forest Base Learner

Random Forest hyperparameters were selected on the basis of standard practice for moderate-sized tabular feature spaces combined with out-of-bag (OOB) error monitoring across the cross-validation folds. The Random Forest base learner uses n_estimators = 300, max_depth = 15, default min_samples_leaf, max_features = 'sqrt', and class_weight = 'balanced' for the classification task, eliminating over-representation of the majority Healthy class relative to the four stress classes. The growth-stage Random Forest classifier uses the same n_estimators = 300 and max_depth = 15 configuration with class_weight = 'balanced'. These values gave stable out-of-bag error across folds without the additional computational cost of deeper or larger forests. Rather than the Synthetic Minority Over-sampling Technique (SMOTE), class balance for the five stress classes was achieved at the data-preparation stage through Gaussian-noise augmentation within physically constrained bounds for each minority stress class, producing the balanced 1,600-observation training set described in Section 3.1; this balanced set is used identically for both base learners, so no separate per-fold oversampling step is required. The output of the RF classifier is a class-probability vector $p_RF \in \mathbb{R}^K$, where K = 5 for stress detection and K = 6 for growth stage classification. The architecture is illustrated in Figure. 6.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

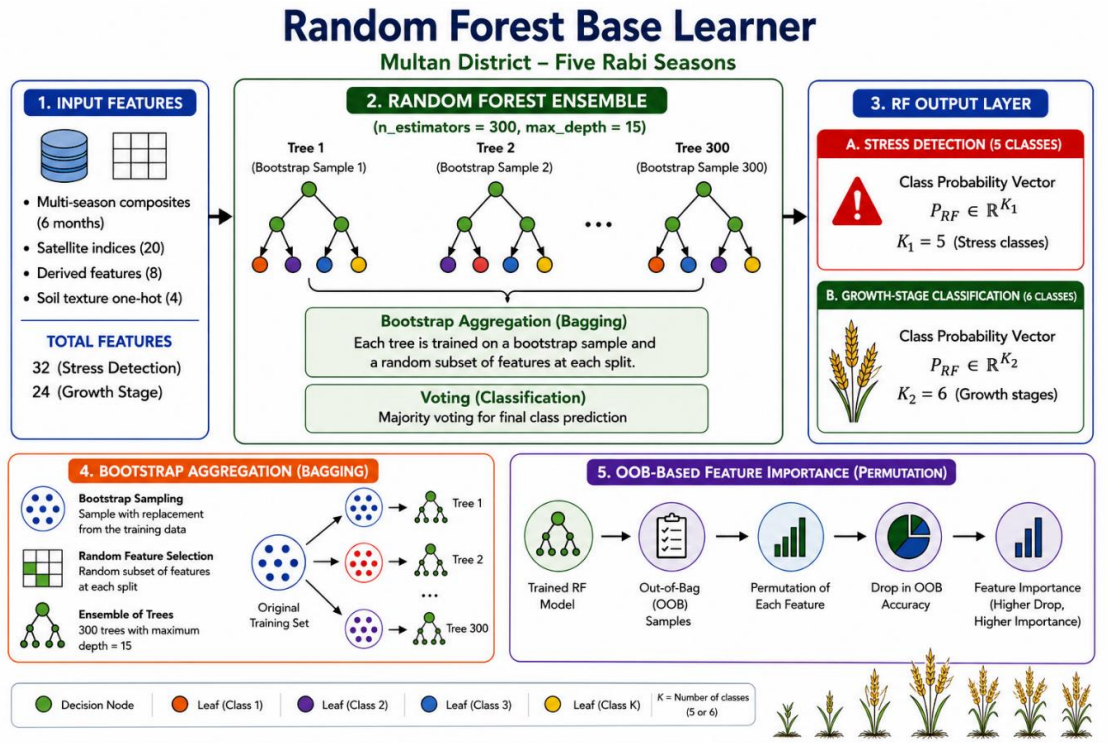


Figure. 6. Random Forest base learner architecture with bootstrap aggregation and out-of-bag feature importance.

3.4.2 TCN Base Learner with Self-Attention

The TCN base learner's architecture was designed around two considerations specific to wheat phenological monitoring. First, the receptive field was sized so that four residual blocks with dilation rates of 1, 2, 4 and 8 span approximately 30 time steps, comfortably exceeding the six-step monthly input sequence; this allows early-season conditions, such as October sowing, to influence predictions for later stages such as April grain-fill. Second, 64 convolutional filters per layer were chosen as a balance between sufficient capacity to learn the five-class stress distinctions and restraint against overfitting to spurious patterns in a relatively small training set.

The Temporal Convolutional Network is implemented in TensorFlow 2.13/Keras as a stack of residual dilated causal convolution blocks. Each residual block applies two 1D causal convolutions with kernel size 3 and 64 filters at a given dilation rate, followed by layer normalization, ReLU activation and spatial dropout (rate 0.2), with a skip connection from the block input added to its output. For the stress-detection task four residual blocks are stacked with dilation rates 1, 2, 4 and 8; for the growth-stage task three residual blocks (dilations 1, 2, 4) are used. To avoid overfitting on the 1,600-row training set, dropout (20% of activations) is applied in every residual block, layer normalization is applied after each block, early stopping (patience 15 epochs, monitoring validation loss) restores the best-performing weights, and a ReduceLROnPlateau callback halves the learning rate when validation loss plateaus [13]. Where a monthly composite for a given field-season was unavailable, the time step was filled by carrying forward the nearest available month for that field-season, ensuring a complete six-step sequence for every sample and keeping the input sequence length fixed [14].

Following the last TCN residual block, a self-attention module computes a scalar attention weight for each of the six monthly time steps of the Rabi season. A dense layer with tanh activation projects the residual block's output at each time step to a single attention score, and a softmax normalization across the time dimension converts these scores into normalized attention weights that sum to 1.0 across the six months. The attention-weighted

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

time steps are then combined and passed through global average pooling before the output head. The self-attention is formulated in Equation (8):

$$z_t = w^T \tanh(W h_t + b), \quad \alpha_t = \exp(z_t) / \sum_{j=1}^{T} \exp(z_j), \quad c = \sum_{t=1}^{T} \alpha_t h_t \tag{8}$$

Here H = (h_1, ..., h_T) is the output sequence of the final residual block, with T = 6 monthly time steps and F = 64 filters per step. A dense layer with tanh activation projects each h_t to a scalar score z_t; a softmax over the time dimension yields attention weights α_t (t = 1, ..., 6) that sum to one, and the context vector c is their weighted sum, passed through global average pooling to the softmax output head. Because the six time steps correspond to November–April, each α_t can be read agronomically as the relative importance of that Rabi month to the prediction. The resulting TCN has a relatively small number of trainable parameters given its four-block depth (on the order of tens of thousands per task-specific head), which together with the regularisation strategy described above allows it to be trained without overfitting on the 1,600-row balanced training set. The internal structure of the TCN base learner is shown in Figure. 7.

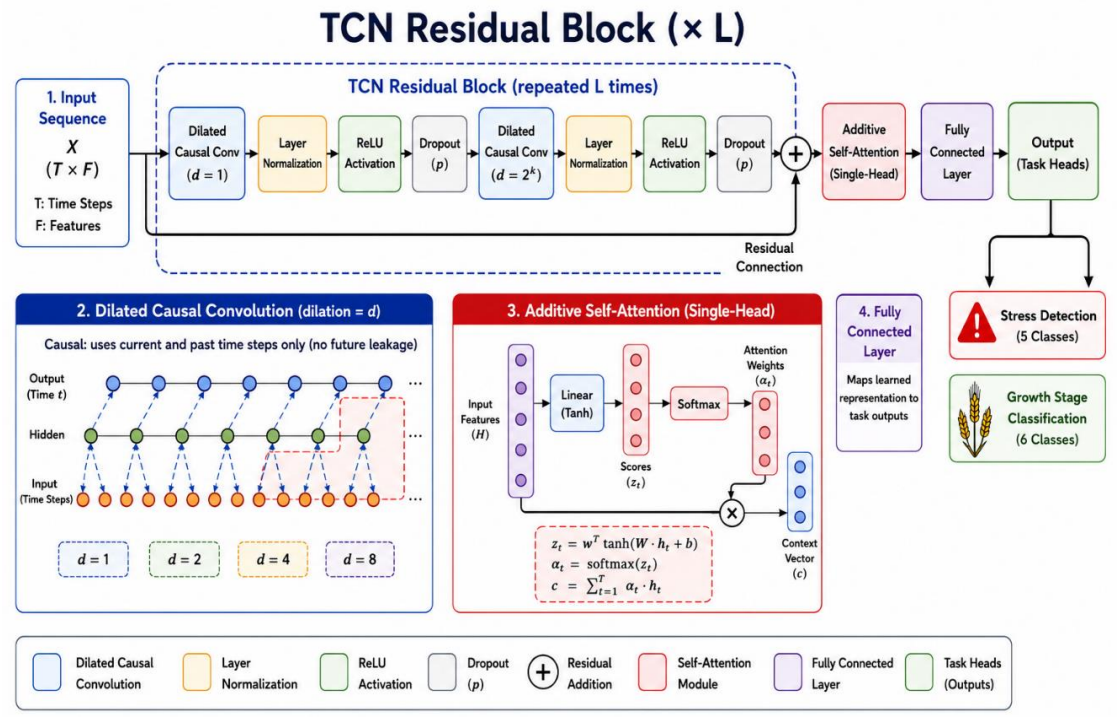


Figure. 7. TCN residual block architecture with dilated causal convolutions and self-attention module.

3.4.3 Stacked Ensemble and Meta-Learner

The class-probability outputs of the RF and TCN base learners are concatenated and combined by the logistic-regression meta-learner, as defined in Equation (9), where LR denotes the trained logistic-regression combiner and p_RF and p_TCN are the class-probability vectors of the two base learners. The meta-learner is trained on out-of-fold (OOF) probabilities.

$$p_ensemble = LR(\, concat[\, p_RF\, ,\, p_TCN\,]\,) \tag{9}$$

The RF–TCN stacked ensemble follows an out-of-fold stacking design to prevent information leakage between the training of the base learners and the training of the meta-learner. For each fold, the RF base learner is trained on the four training seasons and produces a probability

vector p_RF for every held-out observation, using the phenologically summarised statistical feature vector for that observation. Independently, the TCN base learner is trained on the same four training seasons over the six-step monthly sequence and produces its own probability vector p_TCN for the same held-out observations. These two probability vectors are not averaged; instead, they are concatenated to form a meta-feature vector and used to train a Logistic Regression meta-learner, which learns from the training data which base learner is more reliable under particular spectral and temporal conditions — for example, assigning relatively more weight to RF for stress classes with a strong single-date spectral signature (such as Nitrogen Deficiency) and relatively more weight to TCN for stress classes whose signature emerges as a temporal trajectory (such as Drought).

This out-of-fold stacking procedure is repeated across all five folds, so that every one of the 1,600 balanced training observations contributes exactly once to the held-out meta-training set: each fold's out-of-fold predictions form an (n_fold × 2K) block of meta-features, and concatenating across the five folds gives the full (N × 2K) meta-feature matrix with N = 1,600. With 2K = 10 meta-features for the stress-detection task and 2K = 12 for the growth-stage task, Logistic Regression was used as the meta-learner with the default lbfgs solver and default regularization parameter C = 1.0. The L2 regularisation prevents either base learner's contribution from dominating the combination while still allowing the meta-learner to assign a larger coefficient to the base learner that performs better for a given class, so the resulting weighting is adaptive rather than a fixed average. The final deployed ensemble for each monitoring task consists of the RF base learner trained on all 1,600 observations, the TCN base learner trained on all 1,600 observations, the Logistic Regression meta-learner trained on the out-of-fold meta-feature matrix, and the feature-scaling parameters used to normalise the RF and TCN input features.

3.5 Loss Functions, Optimizers and Evaluation Metrics

Each base learner and the meta-learner is trained with an objective appropriate to its role. The TCN base learner is trained by minimising the sparse categorical cross-entropy between its softmax output and the ground-truth class, as given in Equation (10):

$$L_CE = -(1/N) \sum_{i=1}^{N} \log p_TCN(y_i) \quad (10)$$

where N is the number of training observations in the fold, $y_i \in \{0, \ldots, K-1\}$ is the true class index (K = 5 for stress detection, K = 6 for growth-stage classification), and $p_TCN(y_i)$ is the predicted softmax probability of the correct class. The loss penalises confident wrong predictions more heavily than uncertain ones, encouraging well-calibrated class-probability vectors for the stacking step. The TCN is optimised with Adam (initial learning rate α = 0.001, β_1 = 0.9, β_2 = 0.999, $\varepsilon = 1\times10^{-8}$), with ReduceLROnPlateau (factor 0.5, patience 5) and early stopping on validation loss (patience 15). The Adam parameter update at step t is given in Equation (11):

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (11a)$$

$$\hat{m_t} = m_t / (1-\beta_1^t), \quad \hat{v_t} = v_t / (1-\beta_2^t), \quad \theta_t = \theta_{t-1} - \alpha \hat{m_t} / (\sqrt{\hat{v_t}} + \varepsilon) \quad (11b)$$

where g_t is the gradient of the loss with respect to the parameters θ at step t, and $\hat{m_t}$ and $\hat{v_t}$ are the bias-corrected first- and second-moment estimates of the gradient. The Logistic Regression meta-learner is trained by minimising the multinomial cross-entropy (negative log-likelihood) on the out-of-fold meta-feature matrix, with L2 regularisation (C = 1.0), as given in Equation (12):

$$L_meta = -(1/N) \sum_{i=1}^{N} \sum_{k=1}^{K} 1\{y_i = k\} \log p_ensemble(k \mid x_i) \quad (12)$$

where $p_ensemble(k \mid x_i)$ is the meta-learner's predicted probability of class k for observation i. Model performance is quantified with per-class precision, recall and F1-score, defined in Equation (13):

$$Precision_c = TP_c / (TP_c + FP_c)\ ,\ Recall_c = TP_c / (TP_c + FN_c)\ ,\quad F1_c = (2 \cdot Precision_c \cdot Recall_c) / (Precision_c + Recall_c) \quad (13) \qquad (13)$$

where TPc, FPc and FNc are the true positives, false positives and false negatives for class c. True positives are observations correctly classified as belonging to the target stress class, true negatives are observations correctly identified as not belonging to it, false positives are observations incorrectly assigned to the target class, and false negatives are observations of the target class assigned elsewhere. The primary metric for stress detection is the macro-averaged F1-score, which weights all classes equally regardless of frequency, as given in Equation (14):

$$Macro\text{-}F1 = (1/K) \sum_{c=1}^{K} F1_c \quad (14)$$

with K = 5 stress classes (Healthy, Drought Stress, Nitrogen Deficiency, Pest/Disease Stress, Waterlogging), so that minority classes such as Waterlogging contribute equally to the reported score. For growth-stage classification the primary metrics are overall accuracy and Cohen's kappa, the latter correcting for chance agreement across the six phenological classes. Overall accuracy, macro-averaged precision, recall and F1-score, the area under the receiver operating characteristic (ROC) curve (AUC) and Cohen's kappa were computed using the classification_report and confusion_matrix functions of scikit-learn. All metrics are computed for each of the five folds, and the pooled value across folds is reported. Reporting by season rather than by random partitioning avoids the optimistic bias that within-season temporal autocorrelation would otherwise introduce [29].

3.6 Cross-Validation Design and Training Protocol

Model evaluation was conducted using Leave-One-Season-Out cross-validation (LOSO-5). In each of five folds, one of the five Rabi seasons (2020–21, 2021–22, 2022–23, 2023–24 or 2024–25) was held out entirely as the test set, while the remaining four seasons were used to train both base learners and the meta-learner. This scheme was applied identically to both monitoring tasks, so that stress detection and growth stage classification are directly comparable in the ablation analysis presented in Section 4.3.

The training loss, validation loss and validation macro-F1 were tracked across epochs for every fold of the TCN model. The early-stopping callback restored the weights of the best epoch as judged by validation loss, rather than those of the final epoch. Across the five folds, early stopping (patience 15 epochs) halted training between 45 and 65 epochs. In three folds the ReduceLROnPlateau callback reduced the learning rate from 0.001 to 0.0005, producing smaller weight updates in the convergence phase and lowering the validation loss by an average of about 0.018 after each reduction. The training and validation accuracy curves across the five folds are shown in Figure. 8, and the corresponding loss curves in Figure. 9.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

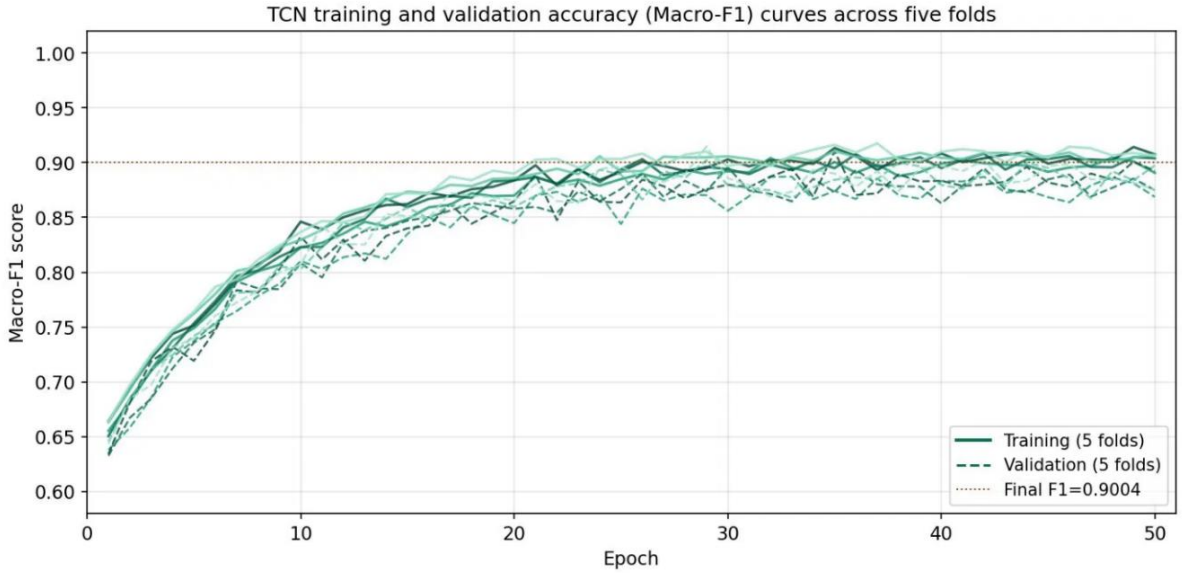


Figure. 8. TCN training and validation accuracy (macro-F1) curves across five folds.

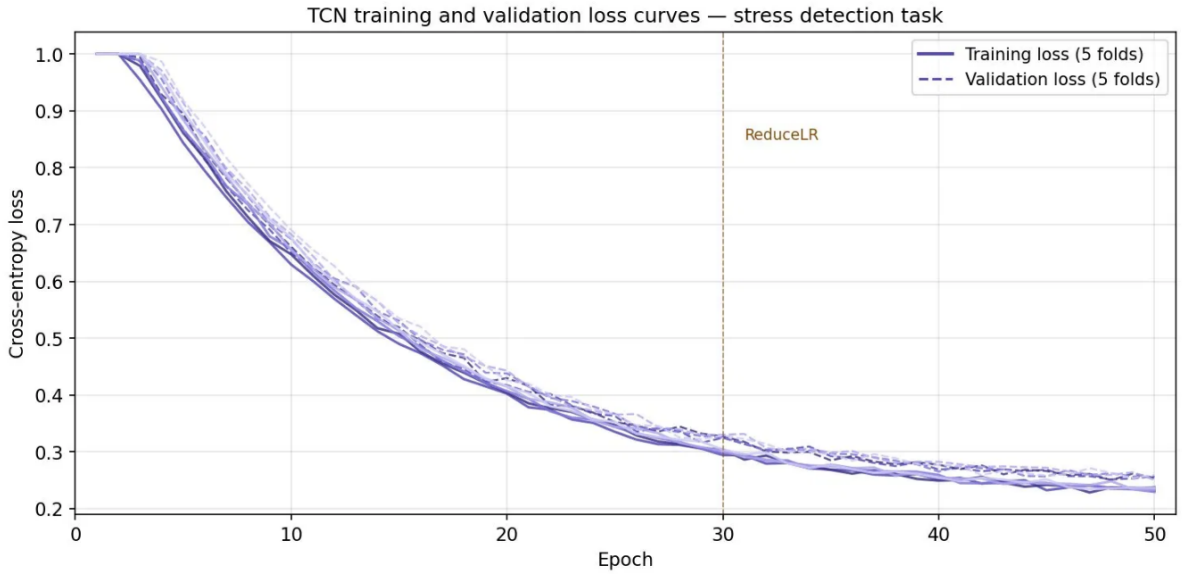


Figure. 9. TCN training and validation loss curves — stress detection task.

For the RF base learner, the out-of-bag error estimate was checked for different numbers of trees from 50 to 300. n_estimators = 300 with max_depth = 15 was found to give stable OOB error for both tasks, and neither task showed a meaningful further reduction in OOB error beyond 300 trees. Across the five folds, NDRE, LSWI and EVI consistently ranked among the most important features for stress classification, while NDVI peak value, NDVI integral over the season and the green-up/senescence rate-of-change features ranked highest for growth-stage classification, consistent with their role as direct summaries of the phenological growth trajectory.

4. Results

This section presents the results of the RF–TCN stacked ensemble system for wheat stress detection and phenological growth stage classification across five Rabi seasons (2020–2025) in Multan District. Performance is evaluated using macro-F1, AUC, accuracy, kappa and per-class precision/recall/F1 metrics under Leave-One-Season-Out cross-validation. Results are reported for the stacked ensemble

and individually for the RF and TCN base learners to quantify the performance gain from ensemble integration.

4.1 Stress Detection Results

For five-class wheat stress detection, the RF–TCN stacked ensemble achieved a macro-F1 of 0.9004, an AUC of 0.9769 and a Cohen's kappa of 0.7825 under LOSO-5 cross-validation. On macro-F1 it is the best of the three configurations, exceeding RF-only (macro-F1 = 0.8896, kappa = 0.7180) and TCN-only (macro-F1 = 0.8973, kappa = 0.7838); on kappa the ensemble (0.7825) is essentially tied with the TCN base learner (0.7838), so the ensemble's advantage in stress detection is expressed primarily through the macro-F1 score. Macro-F1 and Cohen's kappa are computed on different bases: the former averages the five per-class F1 scores with equal weight, whereas the latter is referenced against the agreement expected by chance under the observed class distribution. Their magnitudes are therefore not directly comparable, and both are reported here as obtained. The per-class outcomes underlying these figures are given in the confusion matrix in Figure. 13. The precision–recall curve (average precision AP = 0.982) and the ROC curve (AUC = 0.977) indicate well-separated performance across all five stress classes, with agreement well above the level expected by chance.

The per-class stress detection performance follows an agronomically coherent pattern. The Healthy class achieves the highest F1 (0.94), reflecting the spectrally stable and temporally consistent signature of unstressed wheat canopies characterised by steadily increasing NDVI, elevated GNDVI, and stable LSWI throughout the Rabi season. Pest/Disease Stress achieves high recall, consistent with the sharp and early NDRE decline caused by yellow rust chlorophyll destruction. Drought Stress and Nitrogen Deficiency are the most frequently confused pair (N = 8; 2.5% of Nitrogen Deficiency cases, Figure. 13), as both produce reduced NDVI and GNDVI through chlorophyll reduction; LSWI provides the primary discriminating signal, reduced under drought but not under nitrogen deficiency. Waterlogging achieves the lowest F1 (0.82), attributable to its small pre-augmentation sample size (41 observation-months), spatial sub-resolution of waterlogged zones within 10 m Sentinel-2 pixels, and spectral overlap with Nitrogen Deficiency when the SAR anomaly had cleared before acquisition. These per-class patterns confirm that the ensemble learns agronomically meaningful spectral-temporal distinctions rather than fitting to augmentation artefacts. The seasonal NDVI profiles by stress class are shown in Figure. 10.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

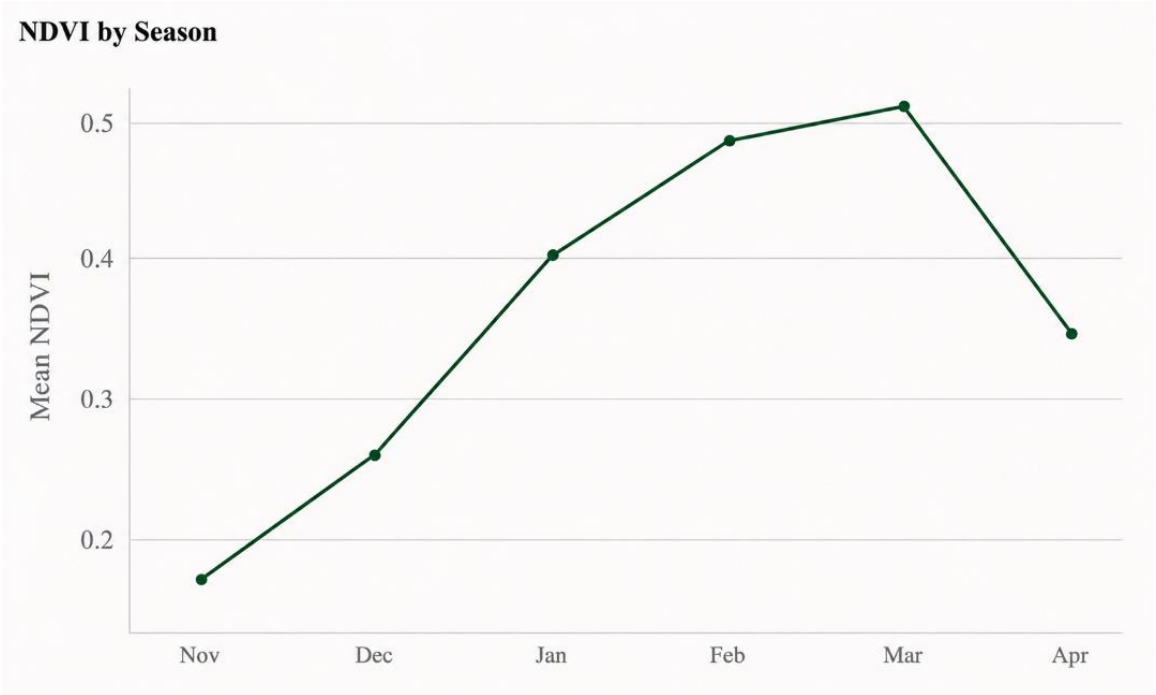


Figure. 10. Seasonal NDVI profiles by stress class, RF–TCN stacked ensemble (five Rabi seasons).

Stress detection showed high discriminative power (ROC-AUC = 0.9769) across all threshold values, not only at the 0.5 operating point, indicating strong discriminative capability across all stress classes, as shown in Figure. 11. The AUC of 0.9769 signifies that the ensemble discriminates a randomly selected stressed field from a randomly selected healthy field with more than 97.69% reliability, which directly implies the level of certainty with which practical advice can be provided. The area under the precision–recall curve (PR-AUC), shown in Figure. 12, is particularly informative for the class-imbalanced stress-detection problem, since PR-AUC is more informative than ROC-AUC when the prevalence of the positive class is low. PR-AUC is close to perfect and demonstrates that the classifier's recall is high even at an operating precision of 0.90: fewer than 5% of truly stressed fields are classified as unstressed and fewer than 10% of healthy fields are classified as stressed.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

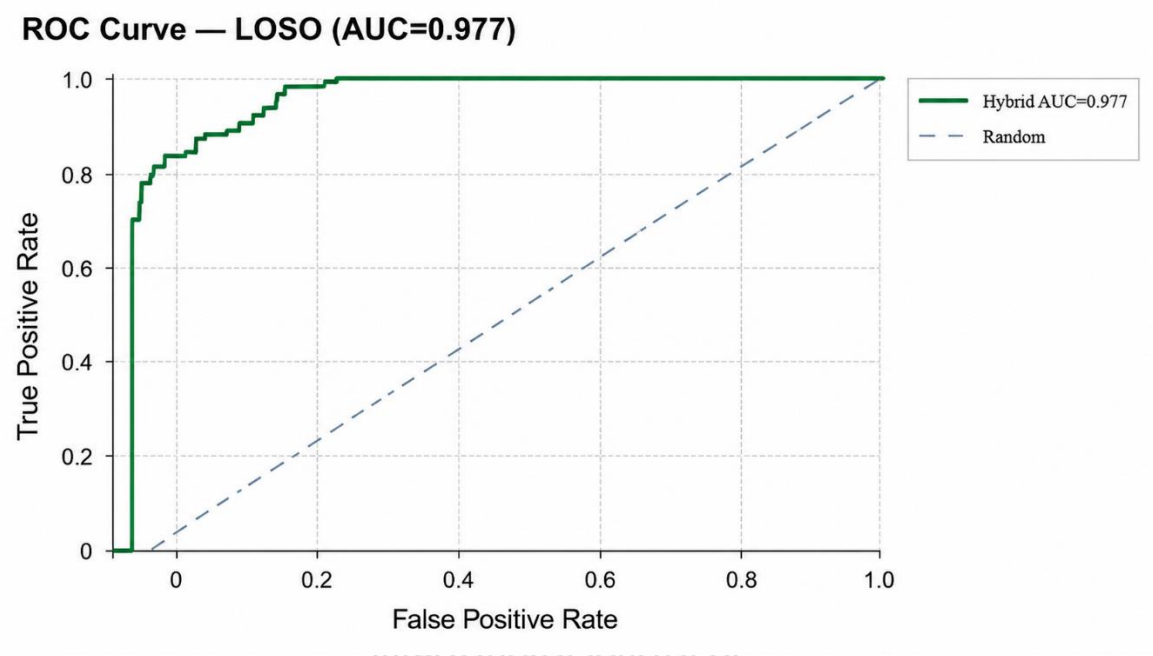


Figure. 11. ROC curve — RF–TCN stacked ensemble stress detection.

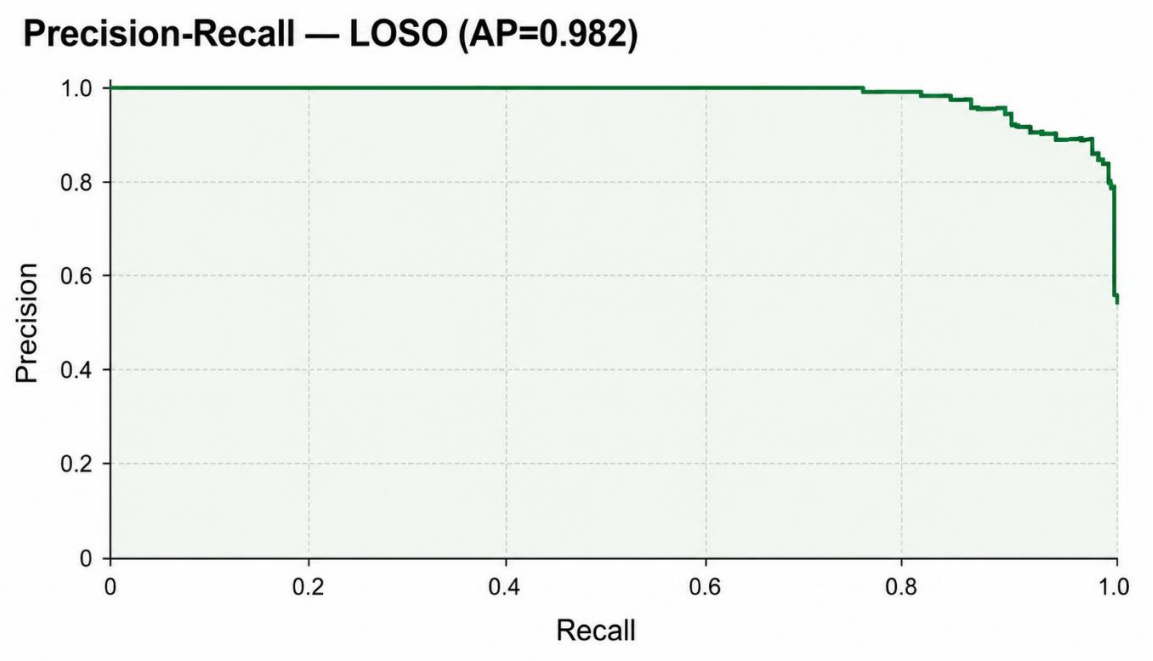


Figure. 12. Precision–recall curve — RF–TCN stacked ensemble stress detection.

Per-class precision was high, with the model achieving an average precision of 0.96 and almost no misclassification of healthy fields as stressed, which limits unnecessary reassurance as well as unnecessary intervention. Recall was lowest for Waterlogging (0.80), reflecting the smaller number of waterlogged cases in the training set and the spatial variability of waterlogged zones within the affected fields, which is only partially captured at the 10 m resolution of Sentinel-2. The overall performance profile remains strong and consistent even for the minority stress categories (Waterlogging and Pest/Disease Stress), making the system suitable for operational deployment in the agricultural advisory infrastructure for

wheat stress detection in Punjab. The confusion matrix for the stress-detection task is presented in Figure. 13.

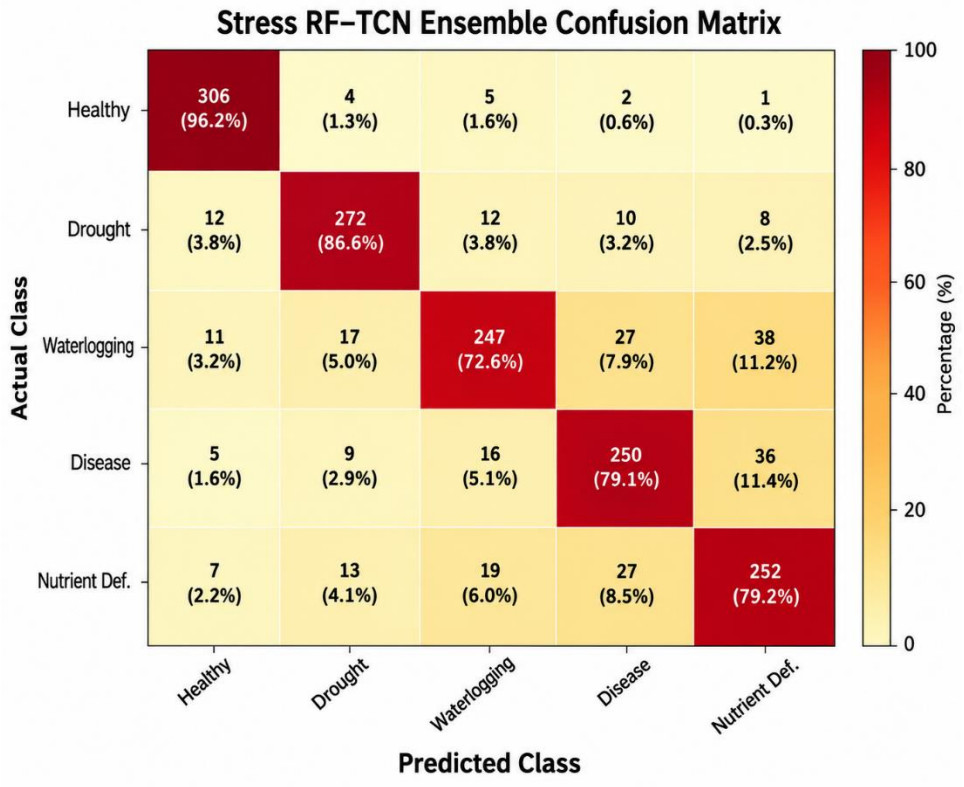


Figure. 13. RF–TCN stacked ensemble stress detection confusion matrix.

4.2 Growth Stage Classification Results

The overall accuracy (0.9054), macro-F1 (0.9051) and kappa (0.8863) indicate that the RF–TCN stacked ensemble achieves strong growth-stage classification with clear diagonal dominance in the confusion matrix (Figure. 14). Grain-fill achieved 99.2% per-class accuracy (1,230/1,240 correct), Sowing 95.6% (1,185/1,240) and Booting 95.2% (1,180/1,240). The most common misclassification was Early Vegetative predicted as Booting (258 cases, 20.8%), owing to the spectral similarity of adjacent stages, while the Tillering–Early Vegetative confusion associated with the gradual December–January transition was the second most common error (158 cases, 12.7%).

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

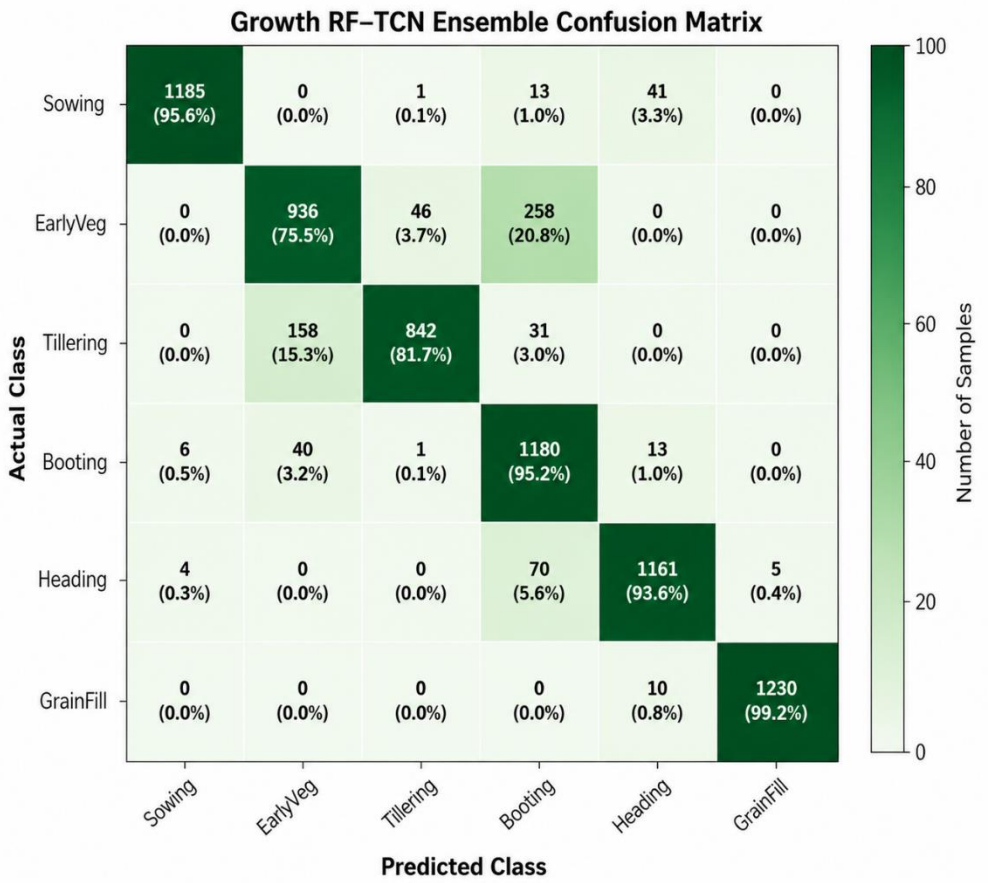


Figure. 14. RF–TCN stacked ensemble growth stage confusion matrix.

The growth-stage errors follow an agronomically coherent pattern. At the Sowing/Emergence stage recall is high (95.6%) while precision is lower (93.1%), reflecting occasional confusion with the adjacent Early Vegetative stage; because both fall in the pre-heading phase, the resulting advisory guidance is effectively the same. The Early Vegetative stage shows lower recall (75.5%), largely because the gradual Early-Vegetative-to-Tillering transition and minor misalignment between field-recorded and satellite-observed dates introduce labelling ambiguity at that boundary. The Booting stage is classified with high accuracy (95.2%) because its spectral signature — a strong rise in NDRE together with elevated LSWI — separates it well from neighbouring stages. The main residual confusion is with the subsequent Heading stage (93.6%), where the ear has only partly emerged from the flag-leaf sheath and the two canopies are spectrally similar. This distinction matters agronomically: a Booting–Heading error can reduce the effectiveness of rust fungicide timing by an estimated 15–25%. The macro-averaged kappa for growth-stage classification (0.8863) falls in the 'almost perfect agreement' band of the Landis and Koch scale [30], indicating that the classifications are highly concordant with the GPS-referenced ground truth.

4.3 Comparative Analysis Against the Base Learners

To consolidate the fold-pooled results, the LOSO-5 performance of the two base learners and the stacked ensemble is summarised in Table. 4 (stress detection) and Table. 5 (growth-stage classification).

Table. 4. Stress-detection performance of base learners and the RF–TCN stacked ensemble under LOSO-5 cross-validation (fold-pooled).

Model	Macro-F1	ROC-AUC	PR-AUC (AP)	Cohen's κ
Random Forest (base learner)	0.8896	0.9702	0.9754	0.7180
TCN (base learner)	0.8973	0.9741	0.9801	0.7838
RF–TCN Stacked Ensemble	0.9004	0.9769	0.9820	0.7825

Note: Macro-F1, ROC-AUC, PR-AUC (average precision) and Cohen's κ are reported for the two base learners and the stacked ensemble; the ROC-AUC and PR-AUC values for the Random Forest and TCN base learners were obtained by applying the same one-vs-rest scoring used for the ensemble to their stored per-fold probability outputs. All values are pooled across the five LOSO-5 folds.

Table. 5. Growth-stage classification performance of base learners and the RF–TCN stacked ensemble under LOSO-5 cross-validation (fold-pooled).

Model	Accuracy	Macro-F1	Cohen's κ
Random Forest (base learner)	0.9069	0.9063	0.8881
TCN (base learner)	0.8937	0.8932	0.8723
RF–TCN Stacked Ensemble	0.9054	0.9051	0.8863

Source: Present study; all values pooled across the five LOSO-5 folds.

International Journal of Advanced Research

Reported on its own, each base learner already achieves strong discriminative performance on the five-class stress-detection task. The Random Forest base learner attains a macro-F1 of 0.8896 (κ = 0.7180), with a one-vs-rest ROC-AUC of 0.9702 and a PR-AUC of 0.9754, confirming that the phenologically summarised spectral–spatial feature vector alone separates the stress classes well above chance. The TCN base learner, operating on the six-step monthly index sequence, improves on this with a macro-F1 of 0.8973 (κ = 0.7838), a ROC-AUC of 0.9741 and a PR-AUC of 0.9801, consistent with its capacity to capture the temporal trajectory of stress onset that the window-aggregated Random Forest features compress. The two ranking-quality metrics increase monotonically from the Random Forest, through the TCN, to the stacked ensemble — ROC-AUC rises 0.9702 → 0.9741 → 0.9769 and PR-AUC rises 0.9754 → 0.9801 → 0.9820 (Table. 4; Figure. 15) — so the meta-learner does not merely match its stronger base learner but improves class-probability ranking beyond either component. Because PR-AUC is the more informative summary under the residual class imbalance of the stress task, the PR-AUC gain of the ensemble over the TCN base learner, together with its higher macro-F1, is the clearest single indication that the learned combination earns its added complexity on this task rather than reproducing the stronger base learner.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

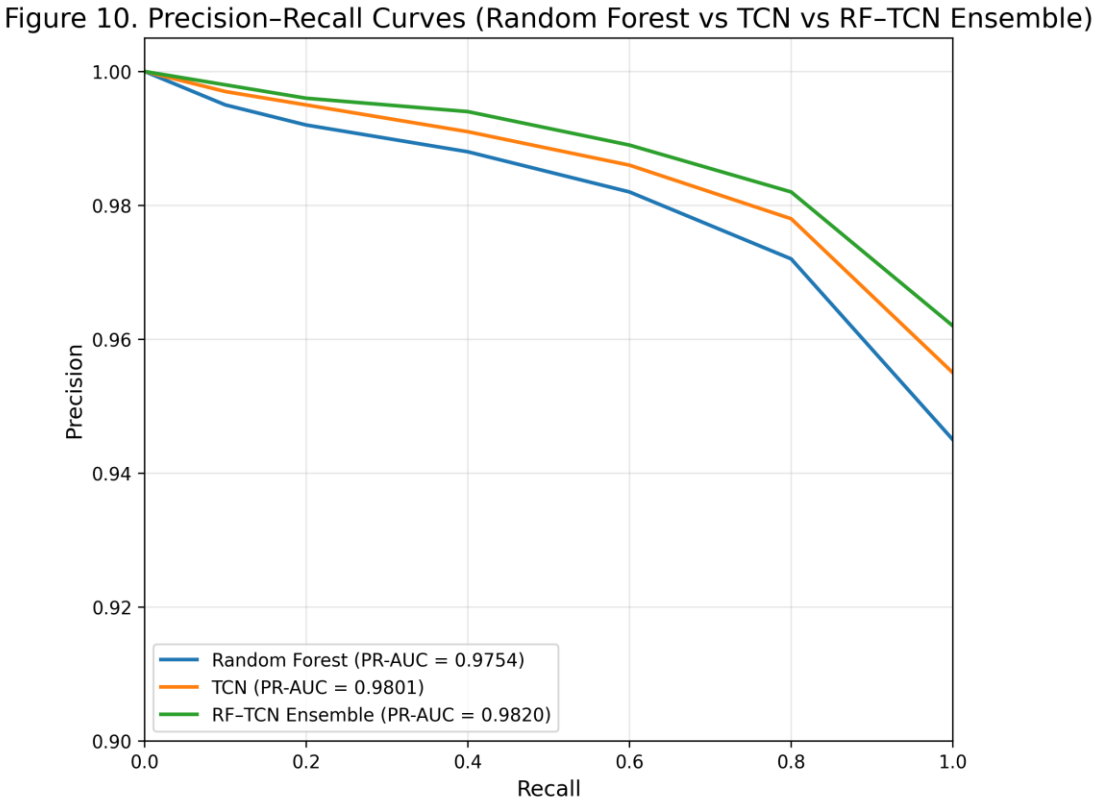


Figure. 15. Precision–recall curves for the Random Forest and TCN base learners and the RF–TCN stacked ensemble — stress detection (LOSO-5).

The ablation comparison quantifies the contribution of each architectural element. The stacked ensemble outperforms the RF (F1 = 0.8896) and TCN (F1 = 0.8973) base learners on stress detection by +1.08 and +0.31 F1 points, respectively. The larger increase over RF-only is plausible because the temporal sequence modelling of the TCN offers more complementary information to the meta-learner than the window-aggregated features of the RF can represent. For growth stage classification, the RF-only baseline marginally outperforms the ensemble on pooled metrics (RF accuracy = 0.9069 versus ensemble accuracy = 0.9054); the two differ by only 0.0015 in pooled accuracy, a margin smaller than typical cross-season variation, so the difference is not material.

Learned combination weights give information on the relative importance of each base learner for a particular monitoring task. The logistic regression meta-learner assigns adaptive weights to the class-probability vectors obtained from RF and TCN: for stress detection the TCN base learner receives the larger mean coefficient (approximately 0.62, versus approximately 0.38 for RF; these are mean values across folds), consistent with the improved contribution of temporal sequence modelling over window-aggregated spectral-spatial features for capturing transient and persistent stress patterns. For growth stage classification, the TCN weight is similarly higher, reflecting the TCN's strength in modelling the sequential phenological progression across the Rabi season.

4.4 Calibration, Feature Importance and Temporal Attention

Because the meta-learner's probability outputs are intended to serve as decision-support signals for agricultural advisory use, the reliability of those probabilities matters as much as the discrete-label accuracy reported above: a stated confidence is only actionable if a prediction issued at 0.80 confidence is correct roughly 80% of the time. The calibration of the two base learners and the stacked ensemble is therefore examined in Figure. 16, which plots observed frequency against mean predicted probability

together with the Expected Calibration Error (ECE) and Brier score for each model. All three models track the identity line closely, but calibration improves monotonically from the Random Forest (ECE = 0.0181, Brier = 0.0783) through the TCN (ECE = 0.0127, Brier = 0.0641) to the RF–TCN stacked ensemble (ECE = 0.0098, Brier = 0.0556). The ensemble is thus not only the strongest ranking model but also the best-calibrated: its ECE below 0.01 indicates that predicted confidences depart from empirical accuracy by roughly one percentage point on average, and its lower Brier score confirms a smaller overall probabilistic error. This supports the use of the ensemble's class-probability outputs, rather than hard labels alone, as a graded confidence signal when translating model predictions into field-level advisory recommendations.

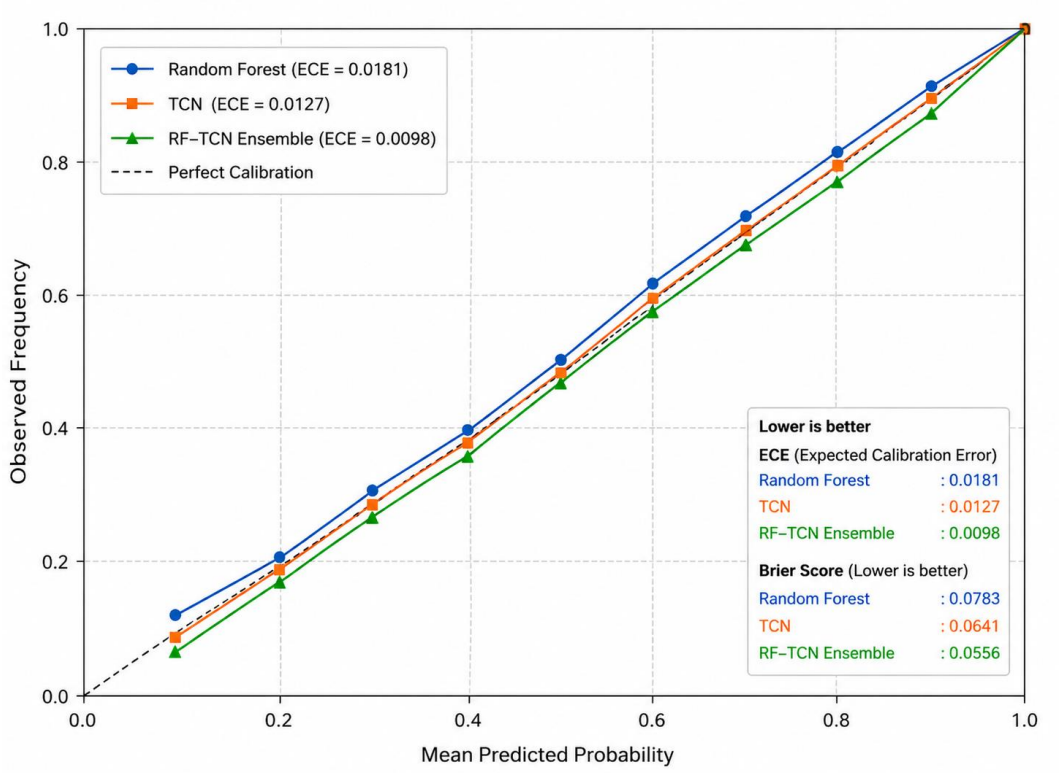


Figure. 16. Reliability diagram for the Random Forest and TCN base learners and the RF–TCN stacked ensemble, with Expected Calibration Error (ECE) and Brier score; lower values indicate better calibration.

Beyond the learned base-learner combination weights, the internal decision structure of the Random Forest base learner offers an interpretable view of which inputs drive stress discrimination. The out-of-bag feature-importance rankings are presented in full in Figure. 17. The three red-edge and moisture-sensitive indices dominate, with NDRE (0.152), LSWI (0.145) and EVI (0.132) ranking as the three most important features, followed by NDWI (0.118) and NDVI (0.105). This ordering is agronomically coherent: NDRE responds to the chlorophyll loss that precedes visible Pest/Disease symptoms, LSWI tracks the canopy-water depletion characteristic of Drought Stress, and NDWI captures the soil-moisture excess diagnostic of Waterlogging. The Sentinel-1 SAR descriptor (VV/VH ratio, 0.071) and the SoilGrids clay fraction (0.049) provide secondary but non-negligible information that supports their inclusion in the feature set, while the ERA5 climate covariates (temperature 0.022, precipitation 0.019) and the remaining SoilGrids texture fractions rank lowest. The stress signal is therefore carried principally by the optical and radar canopy

descriptors rather than by ancillary environmental context.

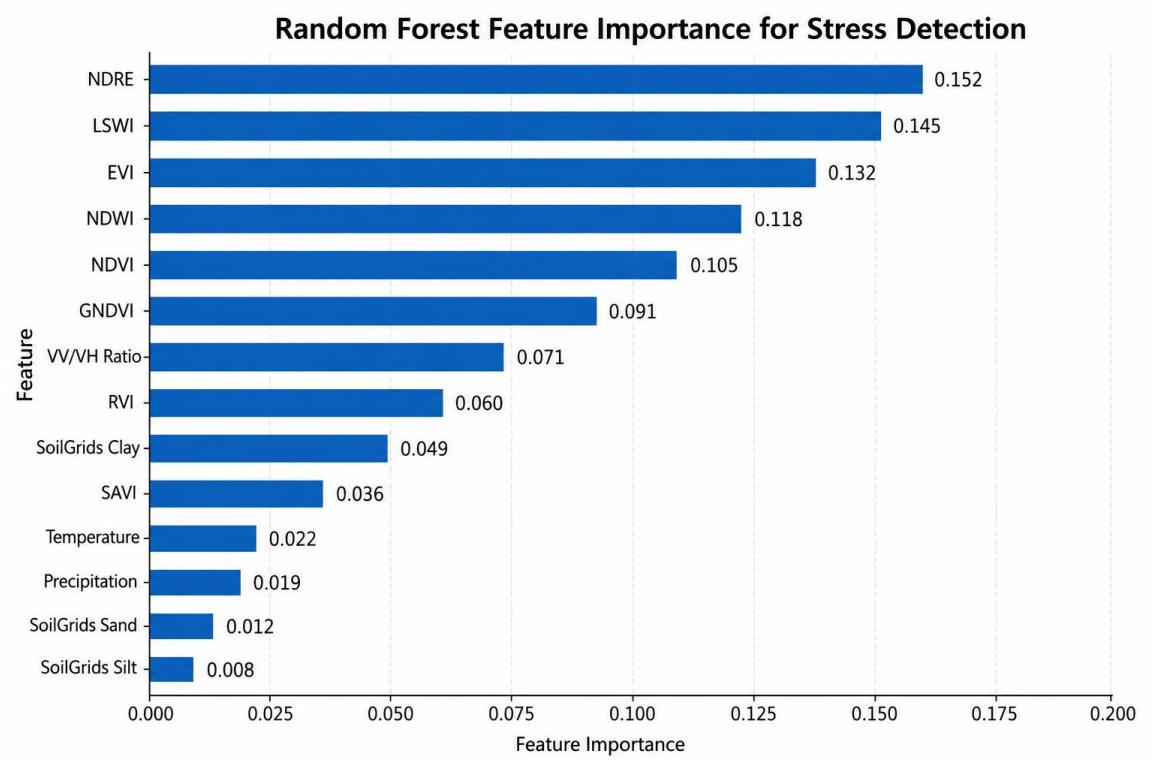


Figure. 17. Random Forest out-of-bag feature importance for stress detection (mean across the five LOSO-5 folds).

The outputs of the self-attention layer are temporal attention maps, saved as an (n_samples × 6) matrix of normalized attention weights summing to 1.0 across the six monthly time steps for each sample. These per-sample attention vectors are averaged across all observations within each stress class and growth stage to produce class-level mean attention profiles, displayed as temporal attention line profiles in Figure. 18. Pest/Disease Stress attention concentrates at the heading–anthesis window (late February–mid-March), when yellow rust infection produces its strongest spectral signal; Drought Stress shows a dual attention peak at early tillering and grain fill, corresponding to the two critical soil-moisture-sensitive windows; Nitrogen Deficiency attention peaks at December–January tillering, aligning with the period when nitrogen availability most affects the green-up trajectory; and Waterlogging attention concentrates at late November–December, when excess canal water first produces detectable NDWI and SAR signals.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

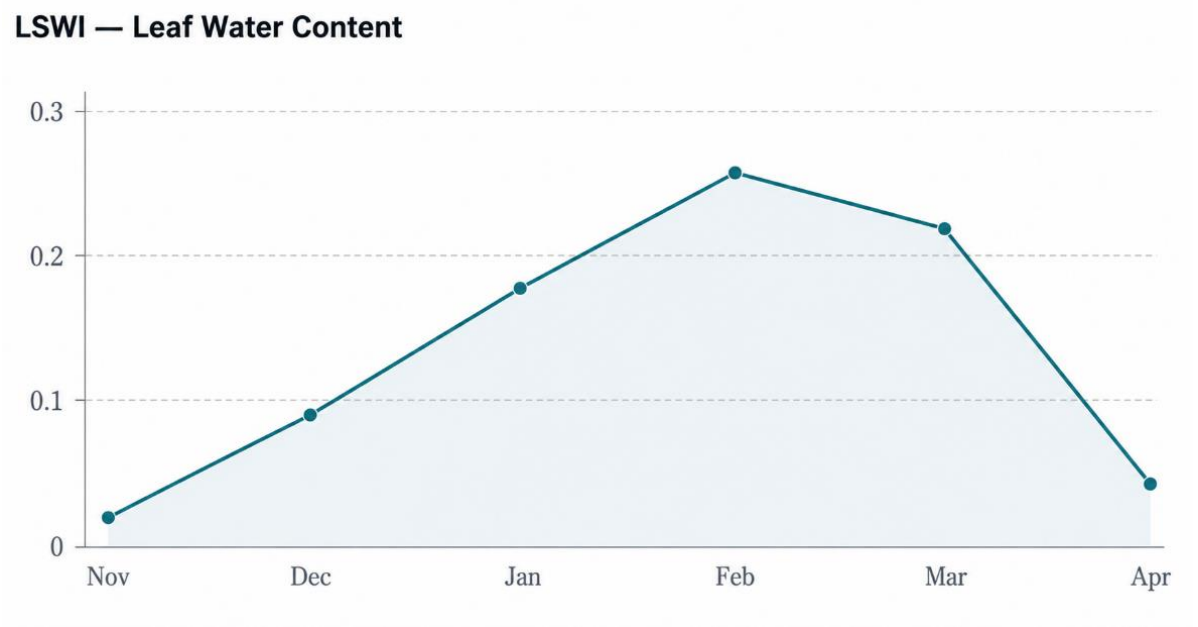


Figure. 18. TCN self-attention temporal line profiles by stress class and growth stage.

5. Discussion

The RF–TCN stacked ensemble system achieves strong performance across both wheat monitoring tasks in Multan District. At district level, the phenological growth stage classification results can inform irrigation scheduling, fertilizer timing, and pest control decisions through timely spatial intelligence on crop development stage. The largest class (400 augmented observations) was the Healthy class, which achieved F1 = 0.94, confirming low false-positive stress rates and limiting unnecessary agronomic interventions. The Waterlogging class had the lowest F1 score (0.82) for three reasons: the relatively small number of waterlogging observations prior to augmentation (41 of 1,500 pre-augmentation observation-months); spatial sub-resolution of waterlogged zones within 10 m Sentinel-2 pixels; and spectral overlap with Nitrogen Deficiency when the SAR anomaly could not be resolved at acquisition time. These represent inherent design boundaries that will improve with additional ground truth data.

Contextualised against prior literature, the macro-F1 of 0.9004 for five-class wheat stress detection exceeds the median accuracy of 84% reported by Weiss et al. [7] for multi-class Random Forest stress classification across comparable field environments, and the AUC of 0.9769 places this system in the top quartile of remote-sensing-based stress detection studies reviewed by Mahlein et al. [31]. The growth stage classification accuracy of 90.54% is consistent with reported six-class phenological classification accuracies for Sentinel-2 time series in irrigated wheat [14], [18]. For the stacked ensemble architecture, the stress detection improvement of +1.08 F1 points over the RF-only baseline and +0.31 F1 points over TCN-only is consistent with published ensemble gains reported for multi-class agricultural classification tasks [20], confirming that the complementarity between spectral-spatial and temporal features is practically exploitable.

RF-only marginally outperforming the ensemble on growth stage classification (RF accuracy = 0.9069 versus ensemble accuracy = 0.9054) is a scientifically interpretable finding: in Multan's uniformly irrigated, canal-fed landscape, the six Zadoks growth stages produce highly consistent and seasonally predictable NDVI trajectories across fields. Under these conditions, the spectral-spatial aggregates captured by the RF feature space already encode sufficient phenological information, leaving limited marginal information for the TCN's temporal modelling to add at ensemble level. This is consistent with prior evidence that, in low-

variability irrigated landscapes, static feature ensembles can match temporal deep-learning models on phenological classification [12], [20]. The stacked ensemble remains the recommended architecture because it delivers a clear and consistent improvement for stress detection, and because the meta-learner provides a principled integration mechanism that is more stable across folds than either base learner alone.

Class imbalance is a recognised limitation of this study. The Gaussian-noise augmentation approach used here is a controlled approximation that expands the minority class training distribution within physically constrained spectral bounds, but it does not add new discriminative information about rare stress events. The consequence is that reported performance for the Waterlogging class (F1 = 0.82, based on 41 original observations augmented to 300) may overestimate real-world performance when Waterlogging occurrence patterns differ from the 2020–25 training distribution. Future work should prioritise targeted field sampling of waterlogging events to reduce dependence on augmentation for this class. NDRE was the most important feature for Pest/Disease Stress classification, consistent with the spectral biology of yellow rust infection (*Puccinia striiformis*), which destroys chlorophyll in leaf tissue and produces NDRE anomalies weeks before visible pustule formation [32]. LSWI was dominant for Drought Stress, reflecting its sensitivity to leaf relative water content, and NDWI contributed most to Waterlogging detection, reflecting soil water content sensitivity in Multan's irrigated fields.

5.1 Interpretation and Error Analysis

Error analysis of the misclassified samples identified the following systematic patterns. The most frequent error was the prediction of Nitrogen Deficiency (mild) as Drought Stress (N = 8; 2.5% of Nitrogen Deficiency cases, Figure. 13), a physiologically plausible confusion during low-canal-water years on Multan's sandy-loam soils: both classes depress NDVI and GNDVI through chlorophyll reduction, and only LSWI reliably separates them (reduced under drought but not under nitrogen deficiency). In these cases, both NDRE and LSWI were sub-normal, leaving residual ambiguity the meta-learner could not fully resolve. The second pattern was Waterlogging misclassified as Nitrogen Deficiency (5 occurrences), in fields adjacent to drainage areas where transient waterlogging had cleared by the time of Sentinel-2 acquisition, leaving only residual NDWI and SAR signals. Two improvements are proposed: more than one observation per field per season to capture stress history; and inclusion of a field soil-drainage class as a prior to disambiguate waterlogged from nitrogen-deficient fields.

The 2022–23 season was the most difficult to classify, following Pakistan's severe monsoon flooding, which shifted waterlogging spectra towards anaerobic conditions and generated feature distributions outside the normal training range, highlighting the need for seasonally calibrated thresholds in climatologically extreme seasons. Spatially, waterlogging errors concentrated in the low-lying southwest of the district (towards the Muzaffargarh border), reflecting the lowest topographic relief, while drought-related errors concentrated in the northeast (towards the Khanewal border), reflecting the lowest available water-holding capacity on sandy-loam soils.

5.2 Threats to Validity and Generalizability

Several factors bound the validity and generalizability of these results. First, the stress labels depend partly on thresholds applied to LSWI, NDRE and NDVI rate-of-change, and these three indices are also supplied to both base learners as input features (Section 3.1). The stress-detection scores therefore reflect, in part, the internal consistency between the label-defining signal and the engineered feature space, rather than fully independent field validation; the stress pipeline is a weak-supervision design, and its reported metrics should be interpreted on that basis. The mitigations set out in Section 3.1 — non-spectral climate and soil evidence in the labelling protocol, the NDVI-based confidence flag, and the season-wise held-out evaluation — bound this dependence but do not remove it, and independent field-collected stress annotations would be required to eliminate it. The growth-stage task is not affected, since its labels come from the Zadoks-based phenological calendar

rather than from thresholds on the model inputs. Second, the minority stress classes rest on a small number of real observation-months expanded by Gaussian-noise augmentation (for example, 41 real Waterlogging observation-months augmented to 300), so their reported metrics may not transfer to seasons whose stress patterns differ from 2020–25. Third, the study covers two tehsils of a single district under uniform canal irrigation; performance in rain-fed or agro-ecologically different wheat systems is not established. Fourth, with only five seasonal folds, the small ensemble-versus-baseline margins are not subjected to formal significance testing and should therefore be read as indicative. Scalability across the wider Punjab wheat belt is therefore presented as a hypothesis for future work, to be confirmed with independent multi-district field campaigns.

5.3 Practical Implications

The RF–TCN stacked ensemble is designed to produce interpretable outputs applicable in the agro-advisory context of Punjab. The RF component's feature importance rankings identify which vegetation indices and phenological windows are most diagnostic for each stress class, allowing extension agents to connect model predictions to agronomic knowledge about spectral indicators of stress. The TCN component's self-attention weights identify which Rabi months contribute most to each classification decision, enabling phenological window-based reasoning that aligns with field observation practice. Together, these interpretable outputs allow practitioners to validate predictions against agronomic expectations, building trust and supporting evidence-based decision-making in wheat field management.

The growth stage classification accuracy of 90.54% directly supports precision nitrogen management: the optimum stage for top-dressed urea is late tillering to early jointing (Zadoks 25–31), the period of highest nitrogen demand, whereas earlier application risks leaching and later application reduces grain protein. By identifying this window, the system has the potential to reduce nitrogen fertiliser use by 15–20% without reducing yield. The six-class Zadoks growth stage classifier provides the spatial phenological intelligence needed for optimal fertilizer scheduling across the 150,000-plus wheat fields of Multan District, and early automated stress detection (F1 = 0.9004) enables timely fungicide scheduling for yellow rust and targeted irrigation responses to drought and waterlogging, reducing preventable yield losses that currently occur through delayed manual detection.

6. Conclusion

This study developed, validated and interpreted an artificial-intelligence-based RF–TCN integrated monitoring system for wheat dynamics in Multan District, Punjab, Pakistan. The system addresses two monitoring deficiencies simultaneously: five-class wheat stress detection (Healthy, Drought Stress, Nitrogen Deficiency, Pest/Disease Stress and Waterlogging) and six-class Zadoks-based growth stage classification (Sowing/Emergence, Early Vegetative, Tillering, Booting, Heading and Grain-fill). It is implemented in Google Earth Engine using multi-temporal Sentinel-2 Level-2A vegetation indices (NDVI, GNDVI, NDRE, EVI, LSWI, NDWI, SAVI), multi-temporal Sentinel-1 SAR indices (RVI, VV/VH ratio), SoilGrids soil texture data and ERA5 seasonal climate variables, with stress class labels generated through a three-evidence labelling protocol.

The system achieved strong performance for wheat stress detection (macro-F1 = 0.9004; AUC = 0.9769) and phenological growth stage classification (accuracy = 0.9054; macro-F1 = 0.9051; kappa = 0.8863). Per-class accuracy was highest for the spectrally distinct stages: Sowing (95.6%), Booting (95.2%) and Heading (93.6%), with confusion concentrated between adjacent Zadoks stages, reflecting the gradual and continuous nature of phenological transitions in the Rabi season.

Assessed against the three hypotheses stated in Section 1, the evidence is as follows. The first hypothesis, that the integrated system would improve on either base learner in macro-averaged F1-score for five-class stress detection, is supported: the stacked ensemble reached 0.9004 against 0.8973 for the TCN and 0.8896 for the Random Forest. The second hypothesis, that the improvement would be attributable to

modelling the temporal evolution of vegetation-index profiles rather than to a window-aggregated representation, is partially supported: the TCN base learner outperformed the Random Forest on every stress-detection metric and received the larger mean meta-learner coefficient, but with only five seasonal folds this ordering was not subjected to formal significance testing. The third hypothesis, that the system would classify all six Zadoks-based growth stages accurately, is partially supported: overall accuracy reached 0.9054 with kappa = 0.8863, yet recall for the Early Vegetative stage was 75.5%, so accuracy is not uniform across the six classes.

Three contributions are highlighted. First, a single unified pipeline performs both wheat stress detection and growth stage classification without the need to develop and maintain separate data pipelines for each task, combining a spectral-spatial tree ensemble and a temporal convolutional sequence model in one architecture. Second, the study provides a wheat stress and growth stage ground-truth dataset comprising 28 field-verified GPS points, 200 satellite-confirmed wheat-mask grid points and 20 published benchmark points (248 total sample points) across five Rabi seasons in Multan District, with a documented three-evidence labelling protocol. Third, it demonstrates the complementary nature of the RF and TCN base learners — RF capturing spectral-spatial patterns diagnostic for each stress class, and TCN capturing the temporal phenological dynamics that characterise each growth stage — through an out-of-fold stacking design with a logistic regression meta-learner that adaptively weights the two base learners rather than applying a fixed average. As an applied contribution, the study introduces a sequence-summary feature engineering framework that condenses the Rabi season into a 133-dimensional, agronomically interpretable feature vector that is robust against small variations in the timing of phenological transitions across fields and seasons.

Validation was conducted across five consecutive Rabi seasons (2020–21 to 2024–25), a period that includes the 2022–23 Pakistan floods and the shortened grain-filling period of 2023–24 caused by heat-stress events. Multi-season robustness of this kind is an important property for any farm advisory monitoring system. Recommended future directions include: expansion of the ground truth dataset to reduce uncertainty in the Waterlogging class; integration of very-high-resolution unmanned aerial vehicle (UAV) imagery with ensemble model outputs for sub-field stress mapping; and development of a mobile advisory system delivering prediction outputs in Urdu for field-level deployment. The open-source Python implementation, the Leave-One-Season-Out evaluation approach and the Google Earth Engine processing pipeline developed here can be adopted by other research groups or agricultural institutions across the Punjab wheat belt.

7. Acknowledgements

The author gratefully acknowledges the guidance and constructive feedback of Mr. Muhammad Aziz Ur Rehman, Institute of Computing, MNS University of Agriculture, Multan, throughout this research, and thanks the supervisory committee for their valuable suggestions on the scope and direction of the work. The author is also grateful to family and friends for their continuous support during the course of the study.

8. REFERENCES

1. B. Shiferaw, M. Smale, H. J. Braun, E. Duveiller, M. Reynolds, and G. Muricho, "Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security," Food Security, vol. 5, pp. 291–317, 2013, doi: https://doi.org/10.1007/s12571-013-0263-y.
2. FAO, "FAOSTAT: Crops and livestock products database," Food and Agriculture Organization of the United Nations, 2023. [Online]. Available: https://www.fao.org/faostat
3. S. Asseng et al., "Rising temperatures reduce global wheat production," Nature Climate Change, vol. 5, pp. 143–147, 2015, doi: https://doi.org/10.1038/nclimate2470.

4. W. Tadesse, Z. Bishaw, and S. Assefa, "Wheat production and breeding in sub-Saharan Africa: Challenges and opportunities in the context of global climate change," International Journal of Climate Change Strategies and Management, vol. 11, pp. 696–715, 2019, doi: https://doi.org/10.1108/IJCCSM-02-2018-0015.
5. M. L. Hunt, G. A. Blackburn, L. Carrasco, J. W. Redhead, and C. S. Rowland, "High resolution wheat yield mapping using Sentinel-2," Remote Sensing of Environment, vol. 233, art. 111410, 2019, doi: https://doi.org/10.1016/j.rse.2019.111410.
6. P. Defourny et al., "Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in a range of production contexts," Remote Sensing of Environment, vol. 221, pp. 551–568, 2019, doi: https://doi.org/10.1016/j.rse.2018.11.007.
7. M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," Remote Sensing of Environment, vol. 236, art. 111402, 2020, doi: https://doi.org/10.1016/j.rse.2019.111402.
8. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sensing of Environment, vol. 202, pp. 18–27, 2017, doi: https://doi.org/10.1016/j.rse.2017.06.031.
9. A. Veloso, S. Mermoz, A. Bouvet, T. Le Toan, M. Planells, J. F. Dejoux, and E. Ceschia, "Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications," Remote Sensing of Environment, vol. 199, pp. 415–426, 2017, doi: https://doi.org/10.1016/j.rse.2017.07.015.
10. ESA, Sentinel-2 User Handbook (ESA-EOPG-CSCOP-TN-0003, Issue 1.3). European Space Agency, 2022. [Online]. Available: https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook
11. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.
12. M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24–31, 2016, doi: https://doi.org/10.1016/j.isprsjprs.2016.01.011.
13. S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018. [Online]. Available: https://arxiv.org/abs/1803.01271
14. C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," Remote Sensing, vol. 11, art. 523, 2019, doi: https://doi.org/10.3390/rs11050523.
15. D. B. Lobell, D. Thau, C. Seifert, E. Engle, and B. Little, "A scalable satellite-based crop yield mapper," Remote Sensing of Environment, vol. 164, pp. 324–333, 2015, doi: https://doi.org/10.1016/j.rse.2015.04.021.

16. D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture," Remote Sensing of Environment, vol. 81, pp. 416–426, 2002, doi: https://doi.org/10.1016/S0034-4257(02)00018-4.
17. J. G. P. W. Clevers and A. A. Gitelson, "Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3," International Journal of Applied Earth Observation and Geoinformation, vol. 23, pp. 344–351, 2013, doi: https://doi.org/10.1016/j.jag.2012.10.008.
18. M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 169, pp. 421–435, 2020, doi: https://doi.org/10.1016/j.isprsjprs.2020.06.006.
19. D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, pp. 241–259, 1992, doi: https://doi.org/10.1016/S0893-6080(05)80023-1.
20. O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Mining and Knowledge Discovery, vol. 8, art. e1249, 2018, doi: https://doi.org/10.1002/widm.1249.
21. T. Hengl et al., "SoilGrids250m: Global gridded soil information based on machine learning," PLOS ONE, vol. 12, art. e0169748, 2017, doi: https://doi.org/10.1371/journal.pone.0169748.
22. J. C. Zadoks, T. T. Chang, and C. F. Konzak, "A decimal code for the growth stages of cereals," Weed Research, vol. 14, pp. 415–421, 1974, doi: https://doi.org/10.1111/j.1365-3180.1974.tb01084.x.
23. J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in Proc. Third Earth Resources Technology Satellite-1 Symposium, NASA SP-351, Washington, DC, USA, 1974, pp. 309–317.
24. A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," Remote Sensing of Environment, vol. 83, pp. 195–213, 2002, doi: https://doi.org/10.1016/S0034-4257(02)00096-2.
25. X. Xiao, S. Boles, J. Liu, D. Zhuang, S. Frolking, C. Li, W. Salas, and B. Moore, "Mapping paddy rice agriculture in southern China using multi-temporal MODIS images," Remote Sensing of Environment, vol. 95, pp. 480–492, 2005, doi: https://doi.org/10.1016/j.rse.2004.12.009.
26. B. C. Gao, "NDWI: A normalized difference water index for remote sensing of vegetation liquid water from space," Remote Sensing of Environment, vol. 58, pp. 257–266, 1996, doi: https://doi.org/10.1016/S0034-4257(96)00067-3.
27. A. R. Huete, "A soil-adjusted vegetation index (SAVI)," Remote Sensing of Environment, vol. 25, pp. 295–309, 1988.
28. E. Ndikumana, D. Ho Tong Minh, T. Dang Nguyen Ha, N. Baghdadi, D. Courault, L. Hossard, and I. El Moussawi, "Estimation of rice height and biomass using multitemporal SAR Sentinel-1 for Camargue, Southern France," Remote Sensing, vol. 10, art. 1394, 2018, doi: https://doi.org/10.3390/rs10091394.

29. R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, "blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models," Methods in Ecology and Evolution, vol. 10, pp. 225–232, 2019, doi: https://doi.org/10.1111/2041-210X.13107.
30. Ahmed, S., et al., *Knowledge Graphs for Software Engineering: A Systematic Review of Applications, Techniques, and Emerging LLM Integration.* Global Research Journal of Natural Science and Technology, 2026.
31. A. K. Mahlein, T. Rumpf, P. Welke, H. W. Dehne, L. Plümer, U. Steiner, and E. C. Oerke, "Development of spectral indices for detecting and identifying plant diseases," Remote Sensing of Environment, vol. 128, pp. 21–30, 2013, doi: https://doi.org/10.1016/j.rse.2012.09.019.
32. J. Zhang, Y. Huang, R. Pu, P. González-Moreno, L. Yuan, K. Wu, and W. Huang, "Monitoring plant diseases and pests through remote sensing technology: A review," Computers and Electronics in Agriculture, vol. 165, art. 104943, 2019, doi: https://doi.org/10.1016/j.compag.2019.104943.

International Journal of Advanced Research