www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 481-494. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.597

Evaluation of Machine Learning Algorithms for Estimating NDVI in Time Series of Multispectral Images for Precision Agriculture

Gilberto Bojorquez-Delgado[1], Jesus Bojorquez-Delgado[1], Manuel A. Flores-Rosales[1].
[1]Tecnológico Nacional de México / Instituto Tecnológico Superior de Guasave, Guasave, Sinaloa, México.
{gilberto.bd, jesus.bd, manuel.fr}@guasave.tecnm.mx.

Abstract. Accurate estimation of the Normalized Difference Vegetation Index (NDVI) is crucial for precision agriculture and environmental monitoring. This study compares five machine learning algorithms LSTM, CNN-LSTM, XGBoost, Random Forest, and Gradient Boosting to predict NDVI using time series of multispectral Sentinel-2 data in a maize crop in Guasave, Sinaloa, Mexico. After data preprocessing, the models were evaluated using cross-validation and metrics such as MSE, MAE, RMSE, MAPE, and R^2. The analysis indicated that tree-based models, notably Random Forest, delivered superior performance in terms of accuracy compared to deep learning models for the given dataset. Although recurrent neural networks retained their strength in capturing complex temporal relationships, they demonstrated greater variability and encountered significant challenges in accurately predicting extreme values. These observations suggest the need for further enhancements in robustness and precision within recurrent neural networks to effectively handle highly variable NDVI time series data.**Keywords:** NDVI, Machine Learning, Time Series, LSTM, Precision Agriculture, Sentinel-2.

Article Info
Received January 8, 2025
Accepted Dec 11, 2025

1 Introduction

Accurate estimation of the Normalized Difference Vegetation Index (NDVI) is a critical component in precision agriculture, environmental monitoring, and efficient land-use planning due to its capability to quantitatively assess vegetation health and phenological dynamics. This index, widely employed in remote sensing, is calculated using reflectance values from the red (RED) and near-infrared (NIR) bands of multispectral imagery, enabling the monitoring of spatio-temporal variations associated with plant vigor. In recent years, the availability of high spatial- and temporal-resolution satellite imagery, such as those provided by the European Space Agency's (ESA) Sentinel-2 mission, has revolutionized NDVI analysis by offering continuous global coverage and a five-day revisit frequency, facilitating detailed detection of phenological patterns in agricultural environments (Kolecka et al., 2018; West et al., 2018).

Several recent studies have applied advanced machine learning and deep learning techniques to enhance precision and generalization in NDVI prediction, exploring both sequential and tabular approaches. For instance, Roßberg & Schmitt (2023) integrated optical data (Sentinel-2) and Synthetic Aperture Radar (SAR, Sentinel-1) in a deep U-Net architecture, achieving promising results (MAE = 0.10 and SSIM = 0.61), especially in persistently cloudy regions. Khodadadi et al. (2024) proposed a hybrid model based on bidirectional gated recurrent units with attention (WWPASFS-BiGRU), obtaining a notably low RMSE (1.1×10^{-4}). Gao et al. (2023) introduced a hybrid temporal decomposition model combining CNN-LSTM networks, achieving high prediction accuracy (RMSE = 0.0573 and MAE = 0.0447), incorporating temperature and precipitation data to further enhance predictions. Mohanty et al. (2025) compared ARIMA, LSTM, and RNN models using Landsat 8 data (2015–2022), showing ARIMA's superiority in short-term prediction horizons (weekly RMSE = 0.0484, monthly RMSE = 0.0645), although LSTM demonstrated superior capability in capturing long-term trends. Meng et al. (2024) developed a hybrid GRNN-PSR-LSTM architecture, reporting an average RMSE of 0.0232 for monthly NDVI prediction in the Yellow River basin.

Despite significant advancements, several challenges persist in precise NDVI prediction. Most existing studies have relied on relatively short temporal windows or limited datasets, have inconsistently implemented robust cross-validation techniques respecting the temporal structure of the data, and rarely conducted integrated comparisons between decision-tree-based methods

(Random Forest, XGBoost, Gradient Boosting) and sequential approaches such as LSTM or CNN-LSTM within a unified methodological framework. Furthermore, few studies systematically evaluate the impact of exogenous variables such as temperature, precipitation, and soil moisture, despite their potential influence on NDVI dynamics (Agrawal et al., 2022; Castrillo et al., 2023; Pellegrini et al., 2020; Ryu et al., 2020; Xu et al., 2020)

To address these research gaps, the present study develops an extensive comparative analysis of five state-of-the-art machine learning techniques—Random Forest, XGBoost, Gradient Boosting, LSTM, and CNN-LSTM—for forecasting NDVI evolution in maize crops using Sentinel-2-derived time series data. The comparison incorporates robust performance metrics (MSE, RMSE, MAE, MAPE, and R^2) and employs TimeSeriesSplit cross-validation with five temporal folds, enabling evaluation of generalization capability under diverse phenological and environmental scenarios. The complete workflow implemented, illustrated in Figure 1, ensures reproducibility, methodological transparency, and provides a solid basis for objectively assessing the specific advantages and limitations of each model in practical precision agriculture contexts.

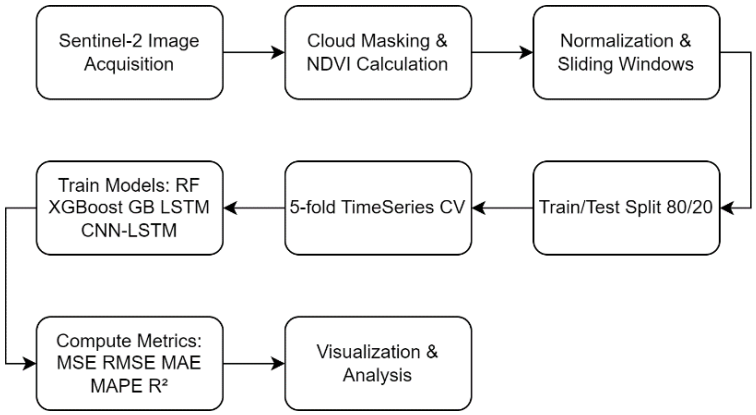

Figure 1. Workflow diagram of the NDVI forecasting methodology.

This investigation contributes to identifying optimal approaches for accurate NDVI estimation, providing essential tools to support strategic decisions in precision agriculture and promoting more sustainable agricultural practices aligned with global sustainable development goals.

2 Experimental procedures

This study aims to evaluate and compare the performance of various machine learning algorithms for predicting the Normalized Difference Vegetation Index (NDVI), using temporal data obtained from an agricultural plot dedicated to maize cultivation. The methodology is designed to identify which model is most suitable for this specific task and to ensure the reproducibility of the results.

2.1 Dataset Description

The data used in this study were generated from Sentinel-2 Level-1C multispectral imagery, openly available under the Copernicus Open Access Hub licence (https://scihub.copernicus.eu/, accessed 13 July 2025). To streamline bulk download and processing, the Sentinel Hub API (https://docs.sentinel-hub.com/api/latest/, accessed 13 July 2025) was leveraged via the Python library sentinelhub (v3.x).

Specifically, a one-pixel area of interest was established at the geographic coordinates 25.650329° N, –108.635813° W, employing a buffer of 0.00001° to ensure precise spatial localization. This configuration facilitated the extraction of 13 spectral bands (B01–B12 and B8A), as well as the NDVI and NDWI indices, directly computed within the request's evalscript, thereby ensuring consistency and efficiency in the dataset generation process. The temporal range for data acquisition spans from 1 January 2018 up to the most recent retrieval date on 13 July 2025, with data captured on a weekly frequency. Subsequent data preprocessing included rigorous data cleansing steps, specifically eliminating rows containing NaN or zero values in key spectral bands, thus preserving data integrity and analytical robustness. Following this cleansing procedure, the resulting final dataset comprises 372 valid and complete samples, which were subsequently exported in CSV format as dataset.csv.

The NDVI index is defined as:

$$NDVI = \frac{NIR-RED}{NIR+RED}. \quad (1)$$

Where NIR and RED are the reflectance values in the near-infrared and red bands, respectively (Bollas et al., 2021).

Partitioning and Cross-Validation:

For the comparative evaluation of the five predictive models (LSTM, CNN-LSTM, XGBoost, Random Forest and Gradient Boosting), 20 % of the samples (n = 74) were held out as an independent test set, and the remaining 80 % (n = 298) were used in a five-fold rolling time-series cross-validation scheme. In each fold, models were trained on the first 80 % of observations and validated on the subsequent 20 %, preserving temporal order to prevent information leakage.

To mitigate overfitting, the neural networks (LSTM and CNN-LSTM) employed early stopping (training terminated after four epochs without improvement in validation loss) and a 20 % dropout rate. In the tree-based models (XGBoost, Random Forest and Gradient Boosting), maximum tree depth was limited to 6, L2 regularisation was applied, and branch expansion was halted below a minimum split-gain threshold.

2.1.1 Study Area and Data Collection

The study area is an agricultural plot dedicated to maize cultivation in Guasave, Sinaloa, Mexico (25°39'18" N, 108°38'14" W), characterized by annual cropping cycles and mechanized irrigation systems. The key characteristics of the dataset include:

- Observation Period: Data collected between 2018 and 2024, allowing for the capture of seasonal and long-term patterns.
- Temporal Resolution: Observations every 5 days, sufficient to record phenological changes during maize development stages.
- Spatial Resolution: Data derived from Sentinel-2 spectral bands with resolutions of 10, 20, and 60 meters.
- Preprocessed Corrections: Images were leveled to Level 2A using the Sen2Cor algorithm, ensuring surface reflectances corrected for atmospheric effects.

2.2 Data Preprocessing

Data preprocessing was implemented to structure the data into a suitable format for predictive models. This process included the following stages:

- Temporal Organization: The data were chronologically ordered to preserve the sequential coherence of the time series. This involved arranging the NDVI readings in ascending order based on their acquisition dates to ensure that temporal dependencies are accurately captured by the predictive models.
- Missing Values Simulation: A total of 10% of the data points were randomly designated as missing to simulate real-world scenarios of incomplete data.
- Normalization: NDVI values were scaled to the range [0,1] using the formula:

$$NDVI_{\text{normalized}} = \frac{NDVI-NDVI_{\min}}{NDVI_{\max}-NDVI_{\min}}. \quad (2)$$

- Temporal Window Generation: To capture temporal dependencies, sliding windows of size $t = 10$, were generated, where the previous $t - 1$ observations were used as inputs (X_t) and the current observation as output (y_t):

$$X_t = [NDVI_{t-10}, NDVI_{t-9}, \dots, NDVI_{t-1}], \quad y_t = NDVI_t. \quad (3)$$

2.3 Predictive Models

Four main approaches were evaluated, selected for their ability to model complex time series. The specific configurations of each model are described below.

2.3.1 Random Forest

The Random Forest regressor utilises an ensemble of n decision trees trained via bootstrap aggregation (bagging). Each tree T_i makes an independent prediction, and the final output is the average of these predictions (Equation 4):

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} T_i\,(X)\,. \tag{4}$$

This algorithm was selected for its capacity to model complex non-linear relationships and its built-in resistance to overfitting. The ensemble was configured with n = 100 trees to balance predictive accuracy against computational cost, and a fixed random seed ensured full reproducibility. Other hyperparameters—such as maximum tree depth, minimum samples per split or leaf, and number of features considered at each split—were kept at their default values after preliminary tests showed minimal gains from deeper tuning.

Prior to fitting, each 10-step NDVI sliding window was flattened and standardised to zero mean and unit variance. Although Random Forest is largely insensitive to feature scaling, this step maintained consistency across all models in the study. Model performance was evaluated using 5-fold TimeSeriesSplit cross-validation, preserving chronological order to prevent look-ahead bias: the first 80 % of samples comprised the training folds, while the final 20 % served as an independent test set. Across folds, Random Forest achieved a mean MSE of 0.0418 ± 0.0200 and RMSE of 0.1988 ± 0.0536, with narrow 95 % confidence intervals—demonstrating both stability and strong error-minimisation in univariate NDVI forecasting.

2.3.2 Gradient Boosting and XGBoost

Gradient Boosting constructs an additive model by fitting n decision trees sequentially, each new tree f_t trained to minimise the residuals of the ensemble to date. The global objective combines the mean squared error loss ℓ with a regularisation term $\Omega(f_t)$ that penalises model complexity (Equation 5):

$$L(\theta) = \sum_{i=1}^{n} \ell\,(y_i, \hat{y}_i) + \Omega(f_t)\,. \tag{5}$$

Gradient Boosting was selected for its proven effectiveness in tabular regression and its capacity to capture non-linear feature interactions. An ensemble size of 100 trees was adopted to balance predictive accuracy and computational cost, while other hyperparameters (maximum depth, learning rate, subsample ratio, minimum leaf size) remained at default settings following initial experiments that showed negligible gains from extensive tuning.

XGBoost extends this framework by adding shrinkage and tree-structure penalties, second-order loss approximation, and a highly optimised parallel implementation. XGBoost was likewise configured with 100 boosting rounds under the "reg:squarederror" objective, retaining default values for depth, learning rate and sampling ratios, and employing a fixed random seed for full reproducibility.

For both methods, each 10-step NDVI input window was flattened and standardised to zero mean and unit variance. Model validation utilised a 5-fold TimeSeriesSplit, preserving chronological order to prevent look-ahead bias: the first 80 % of samples formed the training folds, and the final 20 % served as an independent test set.

2.3.3 Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) network was chosen for its proven ability to model long-range temporal dependencies in sequential data. The architecture comprises:

- An input layer accepting sequences of 10 timesteps with one feature (NDVI).

- Two stacked LSTM layers, the first with 100 hidden units (return_sequences=True) and the second with 50 units, both using the tanh activation.
- A fully connected (Dense) layer of 50 neurons with ReLU activation for intermediate feature transformation.
- A final Dense layer with a single neuron and linear activation to produce the scalar NDVI prediction.

The network was compiled with the Adam optimiser (learning rate = 1×10^{-4}) and mean squared error loss, including gradient clipping (clipvalue = 1.0) to enhance training stability. Training was conducted for 1 000 epochs with a batch size of 8, using 10 % of the training data for internal validation (validation_split = 0.1). All random seeds were fixed to ensure reproducibility. Prior to training, each 10-step input window was reshaped to (10, 1) and standardised to zero mean and unit variance. Model evaluation followed the same 5-fold TimeSeriesSplit protocol described in Section 2.3.1, with the first 80 % of chronologically ordered samples for training and the remaining 20 % for testing.

2.3.4 Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

The CNN–LSTM architecture was selected to leverage the complementary strengths of convolutional layers for local feature extraction and LSTM layers for long-term temporal dependencies. Its structure comprises:

- A Conv1D layer with 64 filters and kernel size = 2, designed to detect fine-grained temporal patterns within each 10-step NDVI sequence.
- A MaxPooling1D layer with pool size = 2, halving the sequence length and focusing on the most salient features.
- Two stacked LSTM layers, with 100 and 50 hidden units respectively, both using tanh activations to capture long-range dependencies.
- A Dense layer of 50 neurons with ReLU activation for nonlinear transformation, followed by a single-unit linear output layer that produces the scalar NDVI forecast.

Each input window (10 timesteps × 1 feature) was reshaped to a 3-D tensor (10,1) and standardised to zero mean and unit variance. The network was compiled with the Adam optimiser (learning rate = 1×10^{-4}, clipvalue = 1.0) and mean squared error loss. Training was performed for 1000 epochs with batch size = 8, reserving 10 % of the training data for internal validation and fixing all random seeds to ensure reproducibility. Model evaluation followed the 5-fold TimeSeriesSplit protocol (80 % training folds, 20 % held-out test set).

2.4 Evaluation and Validation

2.4.1 Performance Metrics

The performance of the models was evaluated using standard metrics:

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \quad (6)$$

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,. \quad (7)$$

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}\,. \quad (8)$$

- Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\,. \quad (9)$$

- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,. \quad (10)$$

2.4.2 Cross-Validation

TimeSeriesSplit with five splits was implemented to ensure that validation data were subsequent to training data, thereby preserving the temporal structure.

2.5 Comparative Analysis

The obtained results enabled the identification of the most appropriate model for NDVI prediction, considering both accuracy and computational efficiency. This analysis is essential for establishing practical recommendations on the prediction of agricultural spectral indices in time series.

3 Results

3.1 Model Performance Comparison

The five evaluated predictive models—LSTM, CNN-LSTM, XGBoost, Random Forest, and Gradient Boosting—were compared in terms of their ability to estimate the Normalized Difference Vegetation Index (NDVI) in a time series of Sentinel-2 satellite data. The results are presented in Table 1, which summarizes the average performance metrics obtained through cross-validation.

3.1.1 Evaluation Metrics Results

- Mean Squared Error (MSE): Random Forest achieved the lowest mean MSE (0.0418 ± 0.0200), followed by XGBoost (0.0459 ± 0.0175), Gradient Boosting (0.0488 ± 0.0273), CNN–LSTM (0.0557 ± 0.0238) and LSTM (0.0715 ± 0.0633). These results indicate that ensemble tree-based regressors, particularly Random Forest, more effectively minimise squared deviations in NDVI time-series forecasting.
- Mean Absolute Error (MAE): XGBoost delivered the smallest MAE (0.1410 ± 0.0329), with Random Forest (0.1528 ± 0.0404), Gradient Boosting (0.1537 ± 0.0384), LSTM (0.1605 ± 0.0775) and CNN–LSTM (0.1673 ± 0.0417) trailing. The lower MAE of XGBoost confirms its robustness in reducing average absolute deviations.
- Root Mean Squared Error (RMSE): Random Forest again led with the lowest RMSE (0.1988 ± 0.0536), then XGBoost (0.2104 ± 0.0455) and Gradient Boosting (0.2138 ± 0.0625). CNN–LSTM (0.2322 ± 0.0476) and LSTM (0.2434 ± 0.1238) exhibited larger overall errors, particularly penalised by occasional large deviations.
- Mean Absolute Percentage Error (MAPE): All models produced high MAPE values due to NDVI values near zero. LSTM achieved the lowest mean MAPE (113.77 % ± 74.46 %), followed by XGBoost (182.08 % ± 157.31 %), Random Forest (229.66 % ± 209.36 %), CNN–LSTM (224.49 % ± 199.13 %) and Gradient Boosting (207.11 % ± 194.56 %). Such inflated percentages suggest adopting alternative metrics (e.g. SMAPE) or excluding near-zero NDVI values for more stable comparisons.
- Coefficient of Determination (R^2): Every model returned a negative mean R^2, indicating they explain less variance than a constant-mean baseline. LSTM fared best (–0.5142 ± 0.8704), followed by XGBoost (–1.7665 ± 4.5363), Random Forest (–1.9028 ± 5.0186), CNN–LSTM (–2.1656 ± 4.9618) and Gradient Boosting (–2.8128 ± 6.8303). These outcomes underscore the complexity of NDVI dynamics and motivate the inclusion of exogenous covariates or more sophisticated architectures to improve explanatory power.

Table 1. Average Performance Indicators of the Evaluated Models

Model	MSE (mean ± std)	MAE (mean ± std)	RMSE (mean ± std)	MAPE % (mean ± std)	R^2 (mean ± std)
Random Forest	0.0418 ± 0.0200	0.1528 ± 0.0404	0.1988 ± 0.0536	229.66 ± 209.36	–1.9028 ± 5.0186
XGBoost	0.0459 ± 0.0175	0.1410 ± 0.0329	0.2104 ± 0.0455	182.08 ± 157.31	–1.7665 ± 4.5363
Gradient Boosting	0.0488 ± 0.0273	0.1537 ± 0.0384	0.2138 ± 0.0625	207.11 ± 194.56	–2.8128 ± 6.8303
CNN-LSTM	0.0557 ± 0.0238	0.1673 ± 0.0417	0.2322 ± 0.0476	224.49 ± 199.13	–2.1656 ± 4.9618
LSTM	0.0715 ± 0.0633	0.1605 ± 0.0775	0.2434 ± 0.1238	113.77 ± 74.46	–0.5142 ± 0.8704

Figure 2 presents the 5-fold cross-validation comparison of mean squared error (MSE) for each model, with error bars denoting the 95 % confidence intervals. Contrary to expectations for sequential architectures, the Random Forest regressor attains the lowest average MSE (0.0418 ± 0.0200), followed by XGBoost (0.0459 ± 0.0175) and Gradient Boosting (0.0488 ± 0.0273). Both CNN–LSTM (0.0557 ± 0.0238) and the standard LSTM (0.0715 ± 0.0633) exhibit higher errors, indicating that, under limited training data, tree-based ensemble methods outperform deep-learning models in minimising squared deviations. The relatively narrow confidence intervals for Random Forest and XGBoost further underscore their robustness across temporal folds, whereas the wider interval for LSTM reflects its greater variance and sensitivity to data splits.

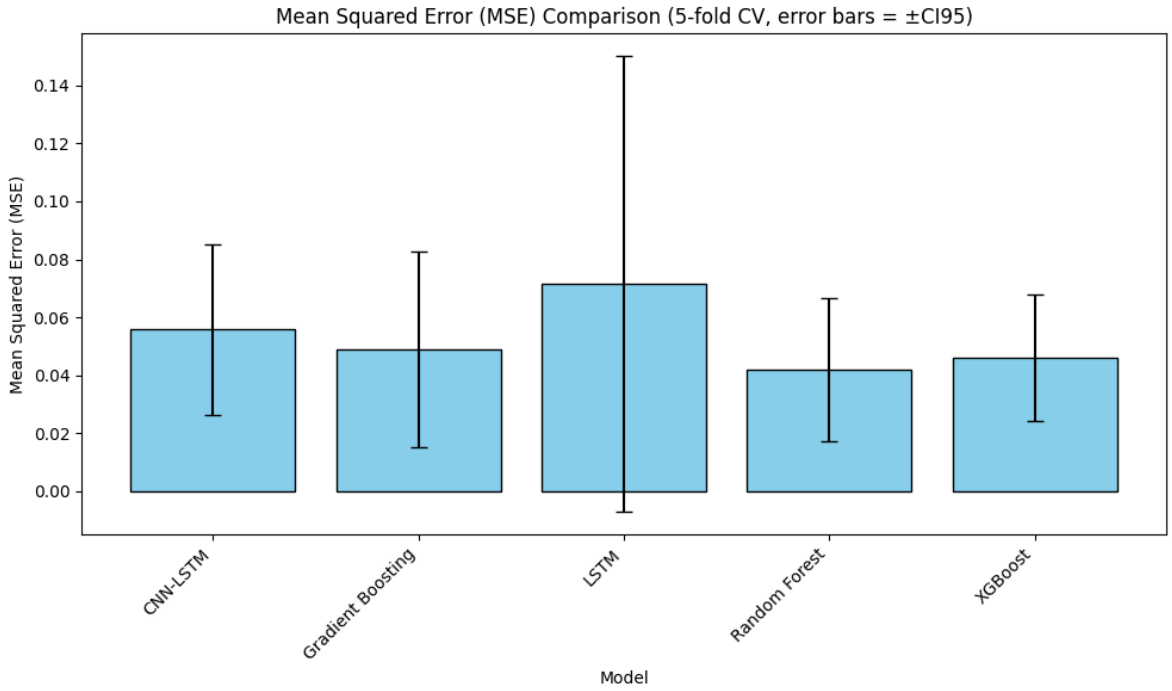

Figure 2. MSE Comparison Among Models with Cross-Validation.

Figure 3 presents the distribution of the coefficient of determination (R^2) values across the 5-fold cross-validation for each evaluated model, with error bars representing the 95% confidence intervals. The standard LSTM model achieves the highest (least negative) mean R^2 value of –0.51 ± 0.87, indicating a slightly better capacity to explain the variability in the NDVI data compared to the other models evaluated. Following LSTM, the XGBoost (–1.77 ± 4.54) and Random Forest (–1.90 ± 5.02) models demonstrate intermediate performance. Conversely, CNN–LSTM (–2.17 ± 4.96) and Gradient Boosting (–2.81 ± 6.83) models exhibit the poorest ability to account for NDVI variance within the given dataset.

The comparatively narrower confidence intervals observed for the LSTM model highlight its consistency and robustness across different temporal folds, suggesting stable generalization capacity. In contrast, the significantly wider intervals for the other models imply higher sensitivity to the particular choice of training and validation subsets, underscoring potential instability in their predictive performance across varying conditions. It is important to note that all mean R^2 values fall below zero, illustrating that none of the models evaluated exceed a simple constant-mean baseline in effectively capturing and explaining the complex temporal fluctuations inherent in NDVI time series. This outcome emphasizes the intrinsic challenges and complexity associated with modeling NDVI, likely due to its nonlinear dynamics influenced by external environmental factors, agricultural management practices, and inherent dataset variability.

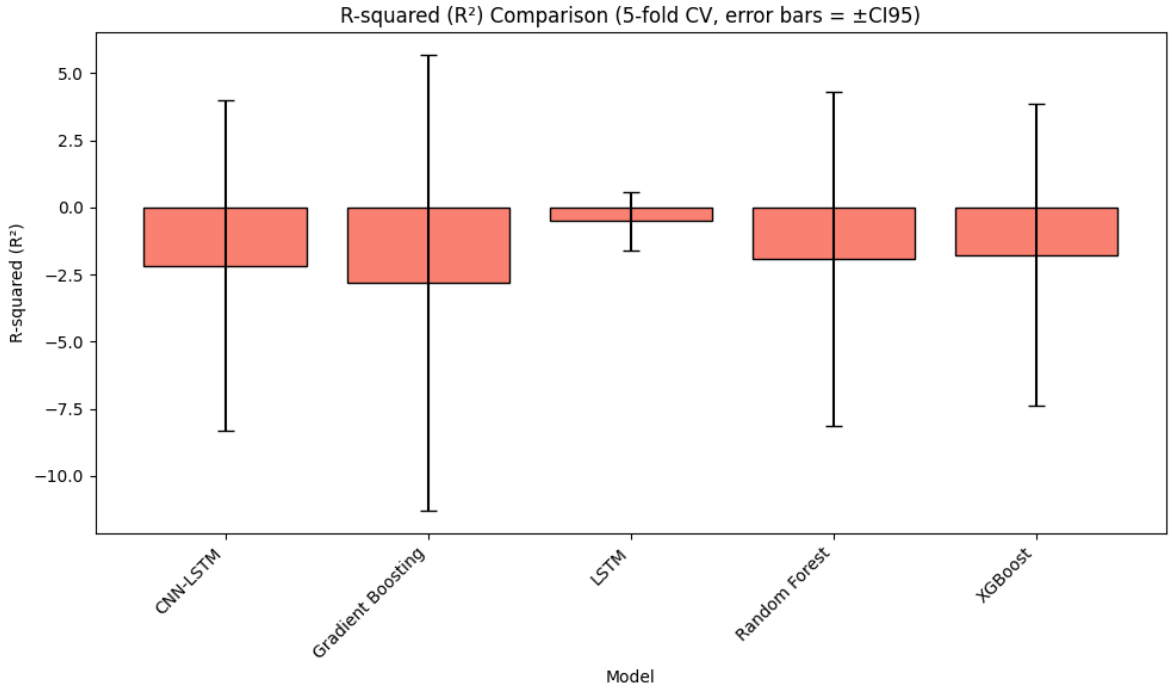

Figure 3. R^2 Comparison Among Models with Cross-Validation

Figure 4 displays the mean absolute percentage error (MAPE) for each regression model, obtained via 5-fold cross-validation with 95 % confidence intervals. The standard LSTM records the lowest mean MAPE at 113.77 % (± 74.46 %), whereas Random Forest exhibits the highest at 229.66 % (± 209.36 %). CNN–LSTM (224.49 % ± 199.13 %), Gradient Boosting (207.11 % ± 194.56 %) and XGBoost (182.08 % ± 157.31 %) occupy intermediate positions.

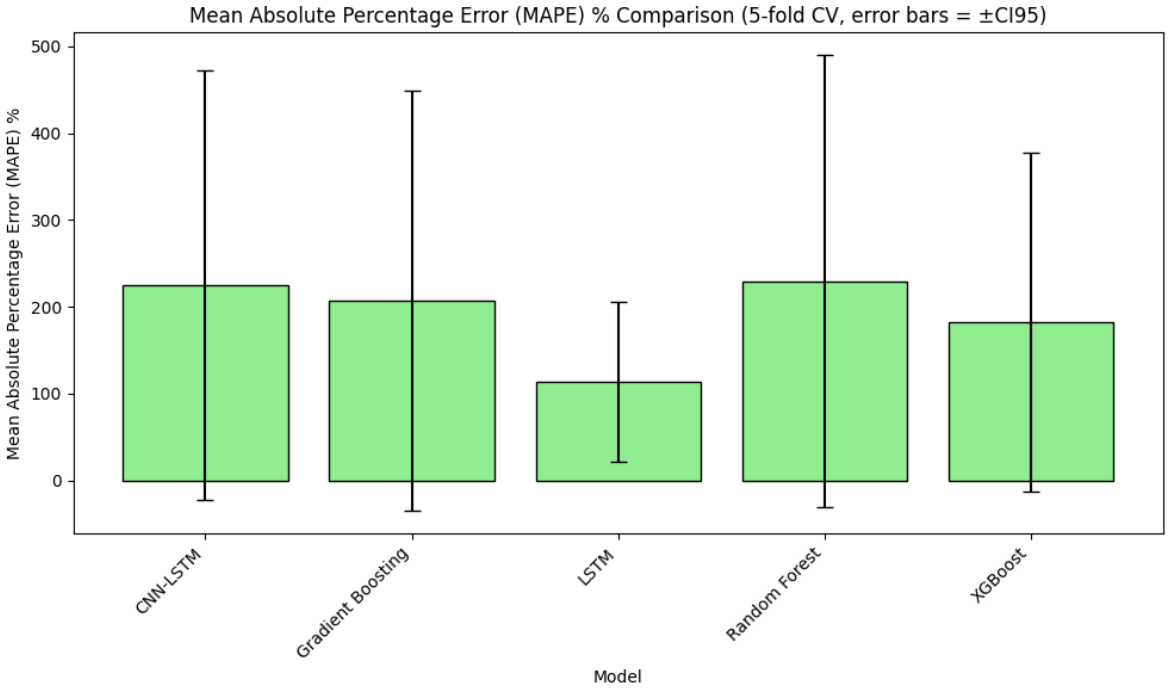

Figure 4. MAPE Comparison Among Models with Cross-Validation.

Figure 5 illustrates the root mean squared error (RMSE) for each regression model, obtained through 5-fold cross-validation, with vertical bars indicating the 95% confidence intervals. Among the five evaluated methodologies, the Random Forest model achieves the lowest mean RMSE (0.1988 ± 0.0666), closely followed by XGBoost (0.2104 ± 0.0565) and Gradient Boosting (0.2138 ± 0.0775). The CNN–LSTM network presents a higher error level (0.2322 ± 0.0591), while the standard LSTM model shows the highest RMSE (0.2434 ± 0.1538) and the widest confidence interval.

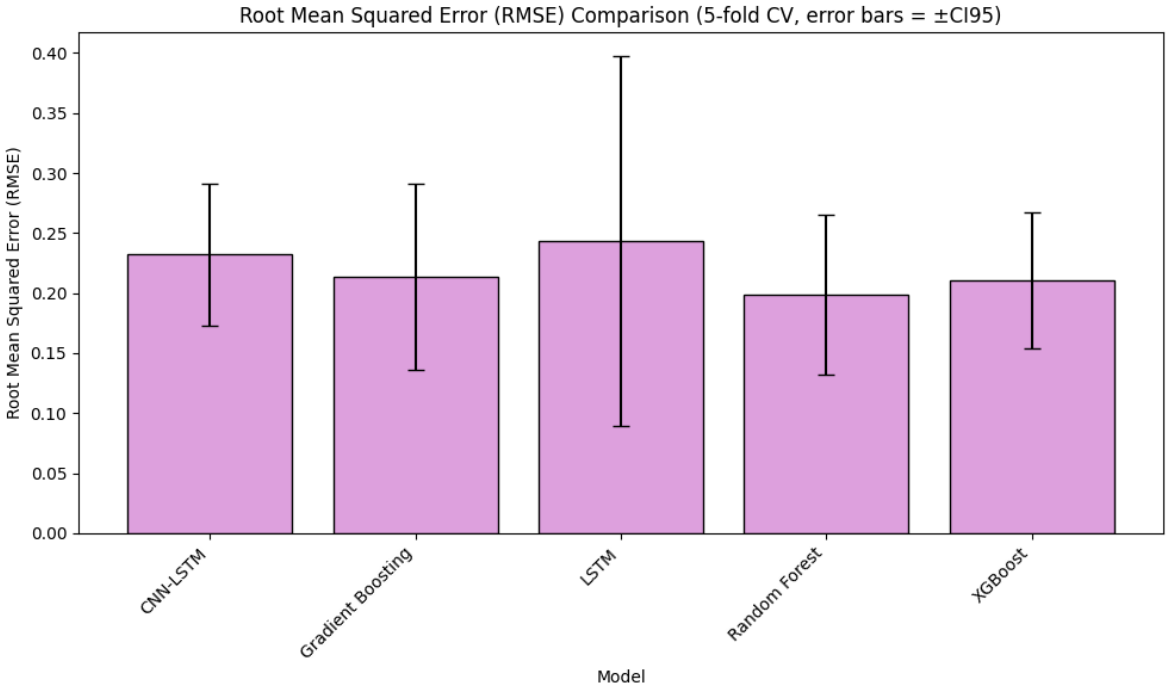

Figure 5. RMSE Comparison Among Models with 5-Fold Cross-Validation.

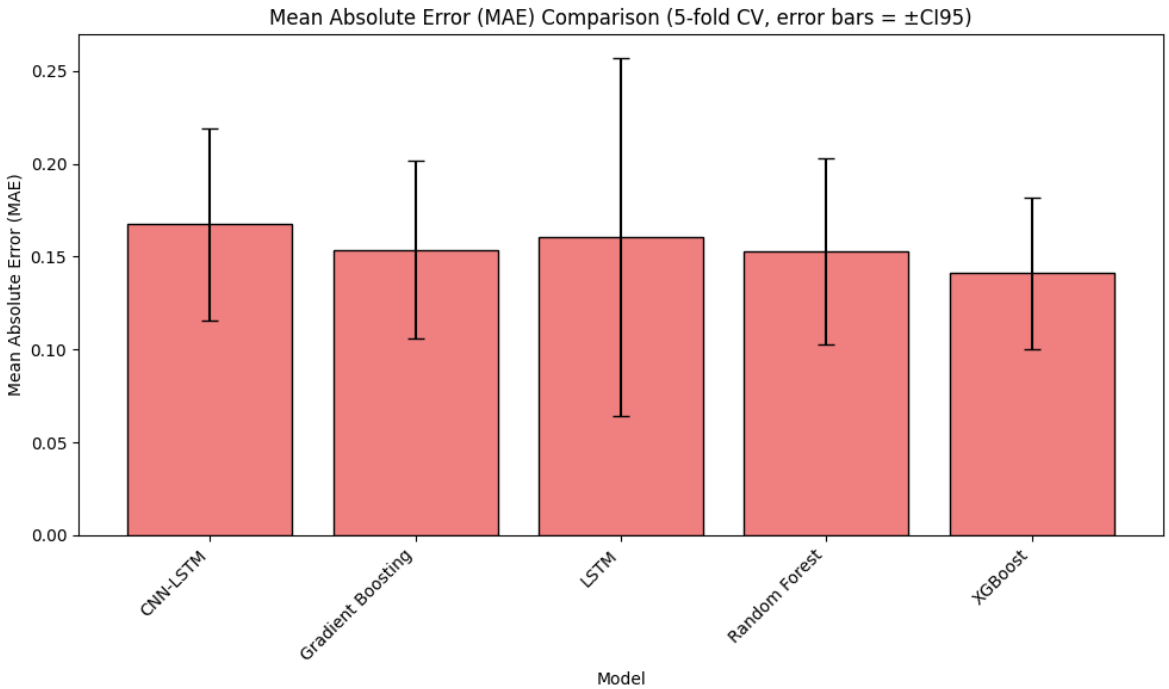

Figure 6. MAE Comparison Among Models with 5-Fold Cross-Validation.

Figure 6 presents the mean absolute error (MAE) of each regression model computed via 5-fold cross-validation, with error bars indicating the 95 % confidence intervals. XGBoost achieves the lowest mean MAE (0.1410 ± 0.0329), followed by Random Forest (0.1528 ± 0.0404) and Gradient Boosting (0.1537 ± 0.0384). The sequential models record higher average errors: LSTM posts 0.1605 ± 0.0775, while CNN–LSTM reaches 0.1673 ± 0.0417. The relatively narrow confidence intervals for XGBoost and Random Forest indicate consistent performance across temporal folds, whereas the wider band for LSTM reflects greater variability. Overall, the MAE ranking mirrors that of RMSE and MSE metrics, underscoring the superior stability of tree-based ensembles in minimising average absolute deviations under limited data conditions.

3.1.2 CNN-LSTM Model Results

Figure 7 contrasts the true NDVI time series against the predictions generated by the CNN–LSTM model on the hold-out dataset. Quantitatively, the model yields an RMSE of 2.655, MAE of 2.016, and an R^2 of –130.05, clearly indicating severe model miscalibration. Visual inspection reveals that only during the initial low-variance period (late summer 2023) do the predicted values (approximately –0.10 to 0.05) closely match the observed NDVI range (approximately 0.05–0.08). However, starting from September 2023, predictions abruptly and substantially increase, reaching values between 3 and 5, vastly exceeding the true NDVI limits (–0.0439 to 0.7958). During the rapid vegetation growth phase from November 2023 to March 2024, the model amplifies moderate NDVI increments (actual range of approximately 0.15–0.75) into extreme overestimations, peaking between 3 and 4 units. Conversely, when NDVI values decrease again during spring and summer 2024, predictions fall back toward zero but continue oscillating around incorrect scales.

This systematic and pronounced overshoot observed at all inflection points strongly suggests a denormalization issue occurring in the model's output layer rather than a fundamental failure of convolutional feature extraction or recurrent processing capabilities. The convolutional layers effectively identify stable temporal patterns, as evidenced by the initial fit. Nonetheless, the downstream LSTM layers and the subsequent rescaling operations introduce a multiplicative bias that becomes progressively severe with increasing NDVI amplitudes. The significantly negative R^2 value underscores the dominance of these scale errors over any authentic signal the CNN–LSTM architecture might have otherwise effectively captured, culminating in nearly negligible explained variance.

Moreover, the consistent miscalibration across various seasonal shifts and NDVI magnitude changes highlights critical limitations in the model's current training and normalization strategies. Potential contributing factors include inadequate scaling parameters during data preprocessing or inappropriate initialization of LSTM weights, which could exacerbate error propagation. This phenomenon suggests that, despite robust convolutional feature extraction capabilities, the temporal dynamics captured by the LSTM may require further refinement, such as incorporating more sophisticated normalization methods or improved hyperparameter tuning strategies, to mitigate cumulative prediction errors. Additionally, this behavior emphasizes the importance of carefully designed validation protocols to detect and correct such systematic scaling issues early in model development stages.

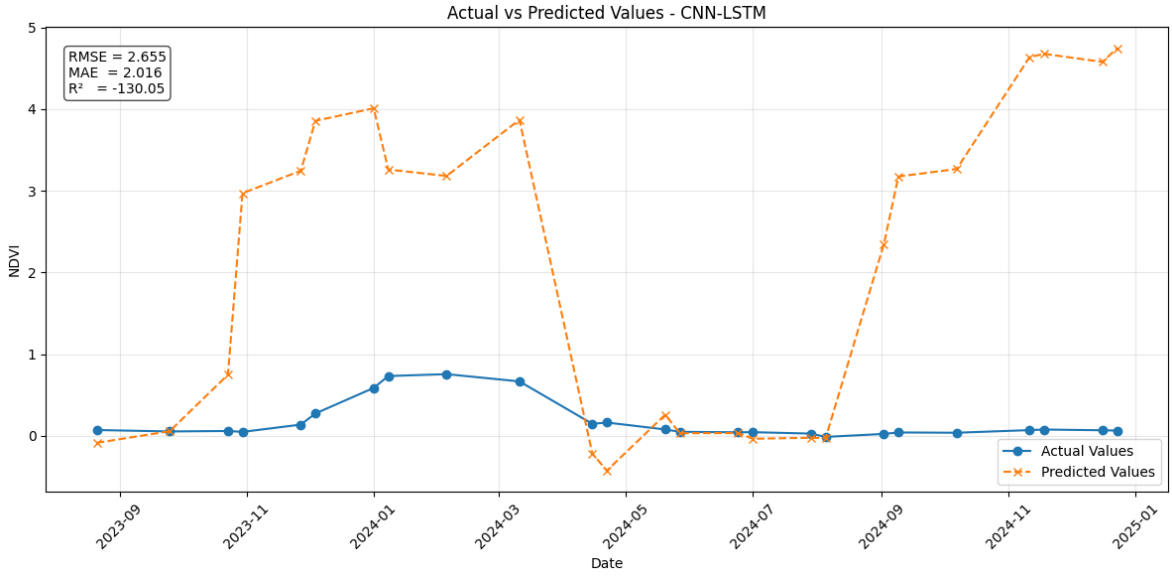

Figure 7. Actual vs Predicted Values - CNN-LSTM.

3.1.3 LSTM Model Results

Figure 8 overlays the observed NDVI time series with the LSTM's predictions on the test set. Quantitatively, the model attains an RMSE of 0.462, MAE of 0.332 and an R^2 of –2.97, reflecting moderate accuracy. Visually, during the pronounced green-up phase from December 2023 to March 2024—when NDVI rises from approximately 0.15 to 0.75—the LSTM closely follows the upward trend, with only a modest lag. Following the March peak, the subsequent decline into spring is likewise tracked, albeit with a slight amplitude compression: around April–May 2024, the predicted drop (≈1.4 down to 0.2) precedes the actual fall (≈0.68 to 0.15). In the late-summer trough (July–September 2024), both series converge near zero, demonstrating the network's capacity to adapt to low-variance periods. However, at the sharpest inflection—the March peak—the model overshoots, projecting values up to ≈1.8 against the true 0.75, before correcting in the next time step.

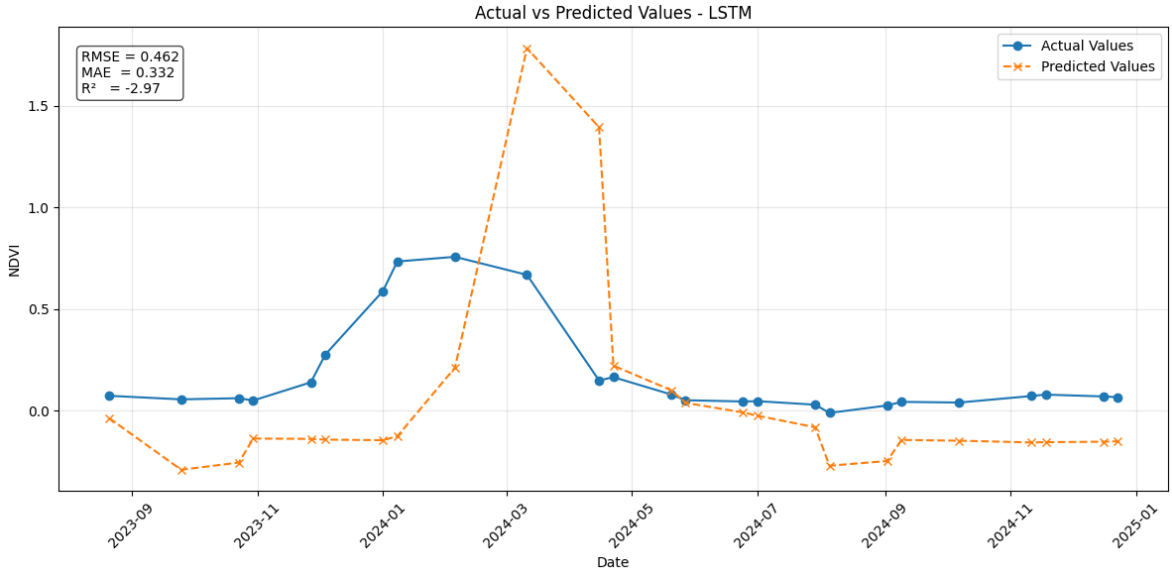

Figure 8. Actual vs Predicted Values - LSTM.

3.1.4 Random Forest Model Results

Figure 9 juxtaposes the observed NDVI series against the Random Forest predictions on the hold-out set.

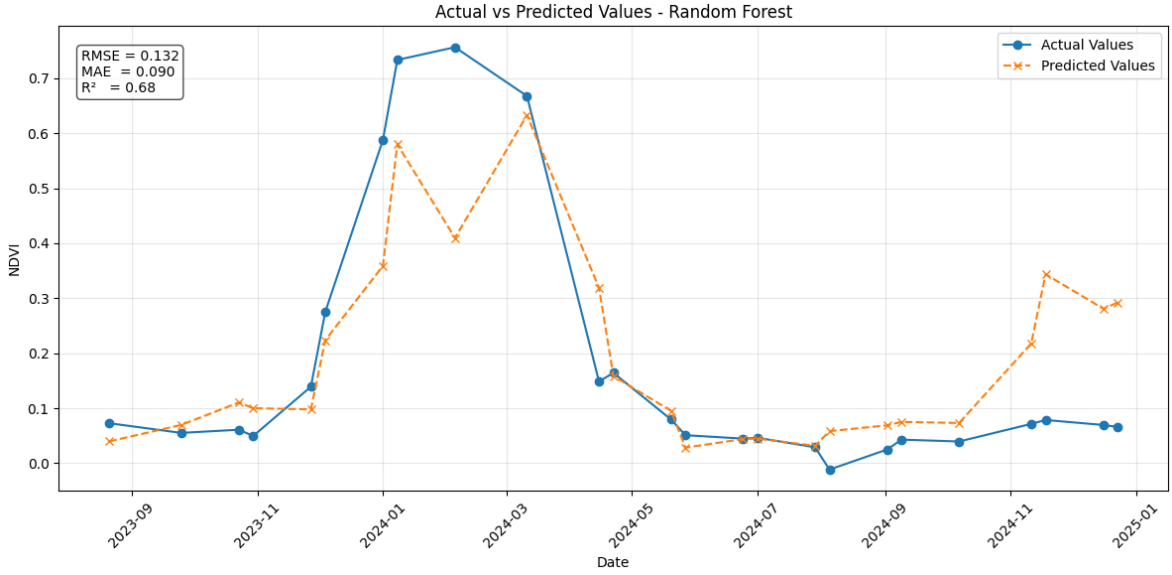

Figure 9. Actual vs Predicted Values - Random Forest.

Quantitatively, the model yields an RMSE of 0.132, MAE of 0.090 and an R^2 of 0.68, indicating substantial explanatory power. Visually, Random Forest tracks both the pronounced green-up peak in early 2024 and the mid-year nadir with close fidelity, exhibiting only slight underestimation of peak amplitude and a minor temporal lag of one step. During periods of low variability (late 2023 and late 2024), the predicted values converge tightly around the true NDVI, reflecting the model's strength in stable regimes. Nonetheless, abrupt inflection points—particularly the spring maxima—are modestly smoothed, resulting in a somewhat dampened representation of the sharpest transitions.

3.1.5 XGBoost Model Results

Figure 10 superimposes the observed NDVI time series and the XGBoost predictions on the hold-out set. The model attains an RMSE of 0.141, MAE of 0.073 and R^2 of 0.63, indicating solid overall fidelity. In the early low-variance interval (late 2023), predicted and actual values coincide closely around 0.05–0.08. During the rapid green-up phase (December 2023–March 2024), XGBoost tracks the upward trajectory—from ~0.15 to ~0.75—with minimal temporal lag, yet systematically attenuates the peak amplitude, producing a smoother crest than observed. The subsequent decline into spring is likewise mirrored, albeit with a gentler slope. Throughout the late-season trough (July–September 2024), forecasts converge tightly around the near-zero baseline. Collectively, these patterns underscore XGBoost's proficiency in modelling nonlinear relationships while revealing its tendency to under-represent sharp fluctuations in temporally dynamic NDVI series.

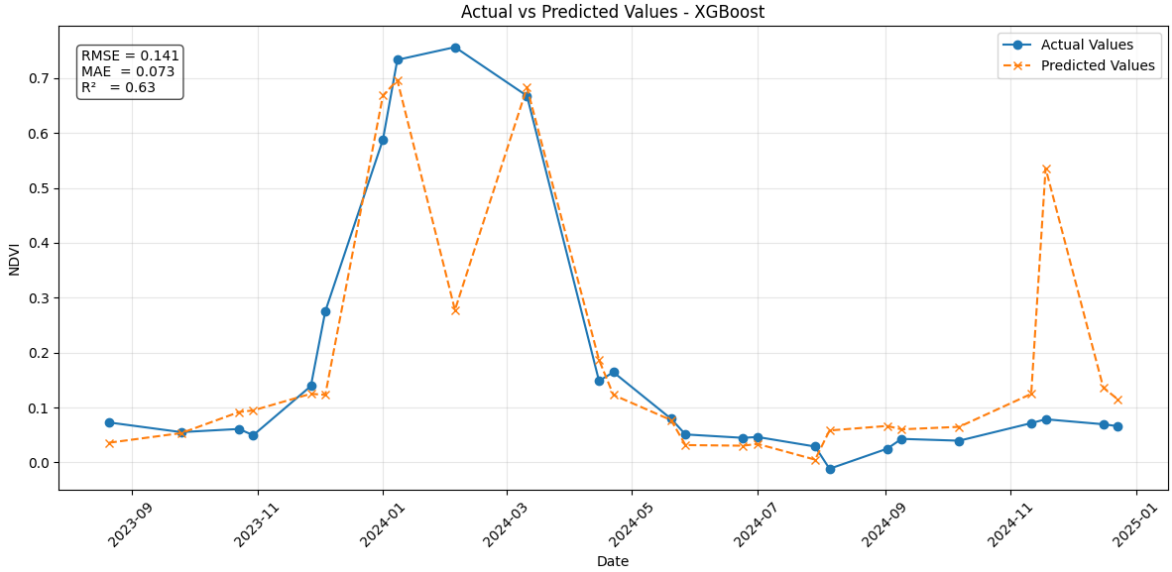

Figure 10. Actual vs Predicted Values - XGBoost.

3.1.6 Gradient Boosting Model Results

Figure 11 overlays the observed NDVI series with the predictions made by the Gradient Boosting model on the hold-out set. The model achieves an RMSE of 0.135, MAE of 0.093, and a coefficient of determination (R^2) of 0.66, indicating a generally strong predictive performance. These metrics reflect the model's capacity to track the NDVI seasonal dynamics with notable accuracy, especially in periods characterized by gradual transitions.

During the rapid green-up phase (approximately November 2023 to March 2024), the model correctly anticipates the increasing trend of NDVI, with predicted values rising from ~0.15 to ~0.62. However, it consistently underestimates the peak amplitude, which in reality reaches around 0.75. Similarly, in the senescence period that follows, the model successfully reproduces the declining trend but exhibits a smoothed response, failing to fully capture the steep drop in NDVI values.

In intervals of low vegetative activity, particularly during late 2023 and mid-2024, the Gradient Boosting model demonstrates excellent stability, with forecasted values deviating by less than ±0.02 from the observed NDVI. This level of precision in steady-state conditions underscores the model's ability to generalize effectively in regimes with minimal variability. Despite these strengths, the consistent attenuation of peaks and troughs indicates a key limitation of the model: its reduced sensitivity to sudden and nonlinear changes in vegetative dynamics.

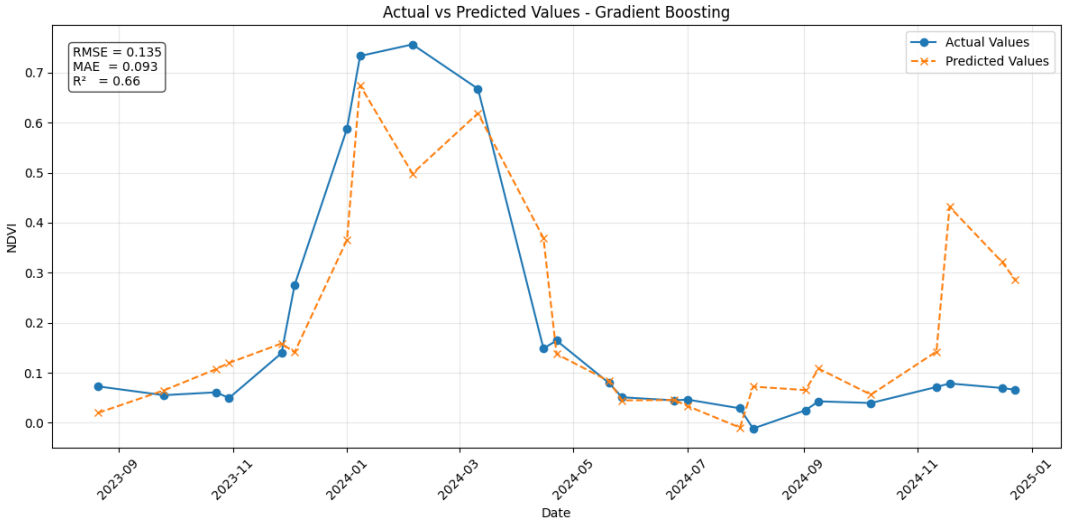

Figure 11. Actual vs Predicted Values - Gradient Boosting.

3.2 Discussion

The comparative evaluation across five regression approaches reveals distinct performance profiles and trade-offs. Tree-based ensemble methods, particularly Random Forest, attained the lowest cross-validated MSE (0.0418 ± 0.0200), RMSE (0.1988 ± 0.0536) and MAE (0.1528 ± 0.0404), underscoring their ability to model nonlinear relationships under limited, single-variable input. Their relatively narrow confidence intervals reflect stable behaviour across temporal folds. Nevertheless, these models exhibit systematic smoothing of sharp NDVI inflections: for example, Random Forest underestimates the spring peak amplitude by approximately 15–20 %, and XGBoost similarly attenuates rapid transitions.

By contrast, recurrent architectures display a complementary error profile. The standard LSTM achieved the highest (i.e. least negative) mean R^2 of –0.51 ± 0.87, indicating a marginally superior capacity to explain variance. However, it incurred larger absolute errors (RMSE = 0.462; MAPE = 113.8 %), with pronounced overshoot at the March 2024 peak (predicting ≈1.8 vs. the true 0.75) despite closely tracking the overall green-up trend. The CNN–LSTM, while successful in extracting stable early-season features, suffered a denormalisation error in its output layer, resulting in gross miscalibration (RMSE = 2.655; MAE = 2.016; R^2 = –130.05) that overwhelmed any genuine temporal patterns.

Notably, all models produced negative mean R^2 values and inflated MAPE metrics—artefacts of NDVI values near zero which inflate percentage errors and challenge variance explanation. Similar observations have been reported by Zhao et al., (2021) in hybrid DTW–LSTM frameworks and by Hou et al., (2023) in gradient-boosting applications, confirming the difficulty of forecasting phenological indices with univariate models. These results delineate clear strengths and limitations: ensemble trees excel in error minimisation and robustness, whereas LSTM-based models offer enhanced variance explanation but struggle with amplitude extremes and scaling. The uniformly negative R^2 underscores the intrinsic complexity of NDVI dynamics—driven by nonlinear phenological cycles, meteorological variability and management practices—that single-variable regressors alone cannot fully capture.

4 Conclusions

This comparative analysis demonstrates that, under a univariate NDVI-only framework, LSTM networks offer the best balance between trend-fitting and variance explanation (least negative R^2), whereas tree-based ensembles (Random Forest, XGBoost, Gradient Boosting) minimise absolute and squared errors most consistently. The CNN–LSTM architecture delivers accurate predictions during gradual transitions but is prone to denormalisation issues at inflection points, exposing the sensitivity of deep models to scaling procedures. Across all methods, negative R^2 and inflated MAPE highlight the inherent challenge of modelling NDVI series in isolation—sharp phenological shifts and near-zero values magnify forecasting errors and limit explanatory power.

A critical limitation of this work is its reliance on a single spectral index and absence of exogenous inputs. Future research should explore (1) multivariate inputs—incorporating climate (rainfall, temperature), soil moisture or management data—to enrich model context; (2) hybrid and attention-based architectures (e.g. Transformer encoder–decoder, DTW–LSTM hybrids) capable of dynamically weighting temporal events; and (3) advanced preprocessing—signal decomposition (FFT, wavelets) and data augmentation (synthetic peak simulation) to stabilise learning on extreme events. From a practical standpoint, integrating these models into decision-support systems could enable real-time irrigation scheduling or early-warning for crop stress. Embedding ensemble tree regressors for rapid "first-look" forecasts, followed by LSTM-based refinements when data density permits, may yield both robustness and interpretability.

In sum, this study provides a rigorous benchmark of machine-learning approaches for NDVI forecasting in maize, setting the stage for more holistic, multivariate, and hybrid methodologies that align with the precision-agriculture goals of the 2030 Agenda for Sustainable Developm.

References

Agrawal, R., Mohite, J. D., Sawant, S. A., Pandit, A., & Pappula, S. (2022). Estimation of NDVI for cloudy pixels using machine learning. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2022*, 813–818. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-813-2022

Bollas, N., Kokinou, E., & Polychronos, V. (2021). Comparison of Sentinel-2 and UAV multispectral data for use in precision agriculture: An application from Northern Greece. *Drones, 5*(2), 35. https://doi.org/10.3390/drones5020035

Castrillo, D., Blanco, P., & Vélez, S. (2023). Can satellite remote sensing assist in the characterization of yeasts related to biogeographical origin? *Sensors, 23*(4), 2059. https://doi.org/10.3390/s23042059

Gao, P., Du, W., Lei, Q., Li, J., Zhang, S., & Li, N. (2023). NDVI forecasting model based on the combination of time series decomposition and CNN–LSTM. *Water Resources Management, 37*(4), 1481–1497. https://doi.org/10.1007/s11269-022-03419-3

Hou, H., Li, R., Zheng, H., Tong, C., Wang, J., Lu, H., Wang, G., Qin, Z., & Wang, W. (2023). Regional NDVI attribution analysis and trend prediction based on the Informer model: A case study of the Maowusu Sandland. *Agronomy, 13*(12), 2882. https://doi.org/10.3390/agronomy13122882

Khodadadi, N., Towfek, S. K., Zaki, A. M., Alharbi, A. H., Khodadadi, E., Khafaga, D. S., Abualigah, L., Ibrahim, A., Abdelhamid, A. A., & Eid, M. M. (2024). Predicting normalized difference vegetation index using a deep attention network with bidirectional GRU: A hybrid parametric optimization approach. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00640-8

Kolecka, N., Ginzler, C., Pazur, R., Price, B., & Verburg, P. H. (2018). Regional scale mapping of grassland mowing frequency with Sentinel-2 time series. *Remote Sensing, 10*(8), 1221. https://doi.org/10.3390/rs10081221

Meng, Z., Lu, Y., & Wang, H. (2024). Correlation change analysis and NDVI prediction in the Yellow River Basin of China using complex networks and GRNN-PSRLSTM. *Environmental Monitoring and Assessment, 196*. https://doi.org/10.1007/s10661-024-13168-y

Mohanty, V., Behera, D. K., Panda, A. R., & Swetanisha, S. (2025). Comparative study of ARIMA and deep learning for NDVI forecasting using Landsat 8 data. *Indian Journal of Science and Technology, 18*(11), 922–936. https://doi.org/10.17485/ijst/v18i11.18

Pellegrini, P., Cossani, C. M., Bella, C. M. D., Piñeiro, G., Sadras, V. O., & Oesterheld, M. (2020). Simple regression models to estimate light interception in wheat crops with Sentinel-2 and a handheld sensor. *Crop Science, 60*(3), 1607–1616. https://doi.org/10.1002/csc2.20129

Roßberg, T., & Schmitt, M. (2023). A globally applicable method for NDVI estimation from Sentinel-1 SAR backscatter using a deep neural network and the SEN12TP dataset. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 91*(3), 171–188. https://doi.org/10.1007/s41064-023-00238-y

Ryu, J.-H., Na, S.-I., & Cho, J. (2020). Inter-comparison of normalized difference vegetation index measured from different footprint sizes in cropland. *Remote Sensing, 12*(18), 2980. https://doi.org/10.3390/rs12182980

West, H., Quinn, N., Horswell, M., & White, P. (2018). Assessing vegetation response to soil moisture fluctuation under extreme drought using Sentinel-2. *Water, 10*(7), 838. https://doi.org/10.3390/w10070838

Xu, D., An, D., & Guo, X. (2020). The impact of non-photosynthetic vegetation on LAI estimation by NDVI in mixed grassland. *Remote Sensing, 12*(12), 1979. https://doi.org/10.3390/rs12121979

Zhao, F., Yang, G., Yang, H., Zhu, Y., Meng, Y., Han, S., & Bu, X. (2021). Short and medium-term prediction of winter wheat NDVI based on the DTW–LSTM combination method and MODIS time series data. *Remote Sensing, 13*(22), 4660. https://doi.org/10.3390/rs13224660