International Journal of Multidisciplinary Research and Growth Evaluation.

Evaluating Machine Learning Ensemble Methods for Cotton Mapping in Pakistan Using MODIS NDVI & EVI Data

Zakria Zaheen [1], Wilson Kalisa [2*], Muhammad Awais [3], Khandakar Md Bappy [4], Abdul Basit [5], Hidayat Ullah [6], Shawkat Ali [7], Jiahua Zhang [8]

[1-8] Space Information and Big Earth Data Research Center, School of Computer Science and Technology, Qingdao University, Qingdao 266071, China

* Corresponding Author: **Wilson Kalisa**

Article Info

ISSN (Online): 2582-7138
Impact Factor (RSIF): 8.04
Volume: 07
Issue: 01
Received: 07-12-2025
Accepted: 05-01-2026
Published: 04-02-2026
Page No: 854-865

Abstract

Accurate crop classification is crucial for agricultural management, food security assessment, and environmental monitoring. In Pakistan, where agriculture is a key economic sector, precise mapping of crops like cotton is essential for effective resource management and policy-making. Remote sensing technologies provide cost-effective and efficient alternatives to traditional field surveys. This study leverages machine learning techniques, particularly ensemble methods to classify cotton using time-series Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) data from MODIS satellite imagery. The year 2014 was selected for analysis due to the availability of high-quality FAO-SPOT reference data and representative climatic conditions during that growing season. The research compares the performance of various machine learning algorithms and evaluates the effectiveness of an ensemble approach in improving classification accuracy within Pakistan's agricultural landscape. Machine learning models, including Random Forest, Support Vector Machine (SVM), Gradient Boosting, XGBoost, and an Ensemble Voting classifier, were applied to EVI and NDVI time-series data. Performance was assessed using a 5-fold cross-validation approach to ensure robust evaluation. The models demonstrated strong predictive capabilities, with Gradient Boosting achieving the highest overall accuracy (OA) of 93.02%, followed closely by the Ensemble Voting model at 92.53%. Other models, including Random Forest, SVM, and XGBoost, also performed well, with OA of 92.46%, 92.17%, and 92.24%, respectively. The inherent limitations of MODIS 250 m spatial resolution, including mixed-pixel effects in fragmented agricultural landscapes, are acknowledged. These results highlight the potential of machine learning for large-scale cotton classification and agricultural monitoring in Pakistan.

DOI: https://doi.org/10.54660/.IJMRGE.2026.7.1.854-865

Keywords: Cotton classification, MODIS, Machine Learning, NDVI, EVI, Ensemble methods, Pakistan, Remote sensing

1. Introduction

Agriculture remains the backbone of Pakistan's economy, contributing approximately 18–25% to the national GDP and providing livelihoods for a substantial portion of the rural population[1]. Among various crops cultivated in the country, cotton holds strategic importance as a primary cash crop and cornerstone of Pakistan textile industry, generating significant export revenue. As a weather-sensitive and economically vital crop cotton production directly impacts national food security, rural incomes and international trade balances [2].

However accurate and timely mapping of cotton-cultivated areas remains a critical challenge for agricultural planners, policymakers and resource managers who require reliable spatial information for effective decision-making.

Globally, cotton serves as a strategic cash crop and the primary natural fiber for the textile sector, making precise spatial information on cotton acreage essential for market analysis, supply chain management, and agricultural policy formulation [3]. The increasing demand for sustainable agricultural practices and efficient resource allocation has intensified the need for robust, scalable, and cost-effective crop monitoring systems. Traditional field-based surveys and manual mapping approaches, while accurate at local scales, are labor-intensive, time-consuming, and often impractical for large-scale regional or national assessments [4]. These conventional methods are further constrained by limited accessibility to remote agricultural areas, high operational costs, and the inability to provide near-real-time information during critical growing seasons.

The operational limitations of traditional crop mapping workflows present significant barriers to achieving accurate and scalable cotton monitoring systems. Data scarcity and quality issues remain critical concerns, as Pakistan and other data-scarce regions face constrained access to high-quality local observations and limited high-resolution reference data, which directly reduces map accuracy. Recent cropland mapping efforts in Pakistan have achieved overall validation accuracies of approximately 80–89%, indicating substantial room for improvement [5]. The shortage of labeled training samples over large regions particularly affects cotton mapping, as the crop's phenological characteristics require sufficient temporal observations for accurate discrimination [6]

Satellite remote sensing technologies offer transformative solutions to the limitations of traditional crop mapping by providing synoptic, repetitive, and cost-effective observations across extensive agricultural landscapes [7]. Among various satellite platforms the Moderate Resolution Imaging Spectroradiometer (MODIS) has emerged as a particularly valuable resource for large-area crop monitoring due to its high temporal resolution, global coverage, and freely available data archive. MODIS-derived vegetation indices capture crop phenological cycles at frequent revisit intervals, enabling the discrimination of different crop types based on their unique temporal signatures [8]. The Normalized Difference Vegetation Index (NDVI), calculated from red and near-infrared spectral bands, has been widely used for vegetation monitoring and shows strong correlation with higher-resolution Landsat NDVI, making it suitable for large-area cropland mapping and sample selection [9]. The Enhanced Vegetation Index (EVI), which incorporates a blue band and atmospheric correction factors, demonstrates enhanced sensitivity to canopy biomass and reduced saturation in high-biomass conditions compared to NDVI [10]. The evolution of crop classification methodologies has progressed dramatically from simple single-date spectral classifiers to sophisticated machine learning algorithms that exploit rich temporal information and multi-source data fusion. Support Vector Machines (SVMs) have been successfully applied to crop classification tasks at both plot and pixel scales, achieving accuracies around 94% for cotton classification on aerial imagery [11]. Random Forest (RF), an ensemble bagging method that combines multiple decision trees, has emerged as one of the most popular algorithms for crop mapping due to its robustness, ability to handle high-dimensional data, and resistance to overfitting [12]. Gradient Boosting and its variants, including XGBoost (Extreme Gradient Boosting), have gained prominence in many industrial and agricultural applications due to their ability to sequentially build strong predictors by correcting errors of previous models [13].

Ensemble methods represent a powerful paradigm in machine learning that combines predictions from multiple base models to achieve superior performance compared to individual classifiers [14]. The fundamental principle underlying ensemble approaches is that diverse models can capture different aspects of the data structure, and their aggregated predictions often exhibit enhanced accuracy, stability, and generalization capability [15]. Ensemble methods leverage three key mechanisms to improve predictions: variance reduction through averaging (as in bagging methods like Random Forest), bias reduction through sequential correction (as in boosting methods like Gradient Boosting), and model diversity through different algorithms or feature subsets [16].

Despite substantial progress in crop classification methodologies and the demonstrated potential of ensemble machine learning approaches, significant gaps remain in the specific context of cotton mapping in Pakistan using MODIS-derived vegetation indices. While most national studies focus on mapping total cropland rather than identifying specific crops like cotton [17]. In Punjab Province, which produces most of Pakistan's cotton, MODIS NDVI/EVI based cotton mapping using ensemble machine learning with systematic benchmarking against multiple algorithms remains limited

This study addresses the identified research gaps by developing and evaluating a comprehensive machine learning framework for cotton classification in Pakistan using MODIS NDVI and EVI time-series data. The objectives are: (1) to develop a robust classification framework that leverages temporal patterns in MODIS-derived indices to discriminate cotton from other land cover types; (2) to systematically compare multiple machine learning algorithms including Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost; (3) to evaluate ensemble voting methods by combining predictions from multiple base classifiers; (4) to implement rigorous validation using 5-fold cross-validation; and (5) to provide actionable insights for operational cotton monitoring systems in Pakistan.

2. Materials and Methods

2.1. Study Area

The study area for this research is Pakistan, a country located in South Asia between latitudes 23°N and 37°N and longitudes 60°E and 75°E. Pakistan has a total land area of approximately 881,913 square kilometers, with diverse topography ranging from coastal areas to mountains. Agriculture is a important sector of Pakistan's economy, contributing about 24% to the GDP and employing around 37% of the labor force [18]. Cotton is one of the major cash crops in Pakistan, primarily grown in the provinces of Punjab and Sindh.

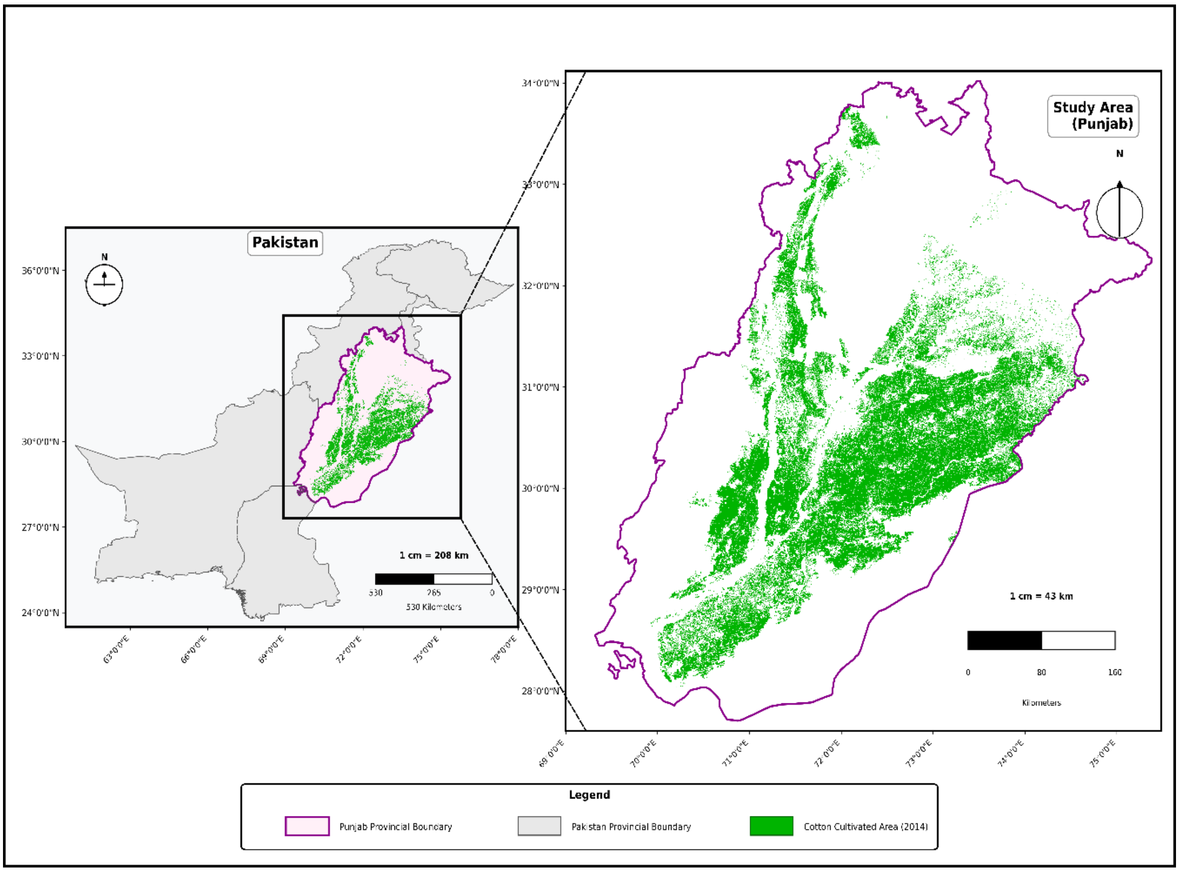

Fig 1: Study Area map of the Punjab province of Pakistan.

2.2. Data

2.2.1. MODIS Data

The study utilized MODIS satellite imagery, specifically the MOD13Q1 product from the Terra satellite, which provides 16-day composite vegetation indices at a spatial resolution of 250 meters [19]. For this research, all available images from 2014 were collected using Google Earth Engine (https://earthengine.google.com), resulting in a dataset of 46 images that covered the entire year. The MOD13Q1 product offers two key vegetation indices: The Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI). Recent studies have demonstrated the continued effectiveness of MODIS time-series data for large-scale agricultural monitoring [20]. The single year was selected due to the availability of high-resolution (10 m) FAO–SPOT reference data for Punjab, required for reliable model training and validation. Climatic records indicate that 2014 was a representative growing season without major anomalies, and MODIS data adequately captured the full cotton phenological cycle. Although limited to a single year, the framework is designed to be transferable to other years when comparable reference data are available (see Section 4.4 for temporal generalization discussion).

2.2.2. Sample Data

Cotton label data provided by the Food and Agriculture Organization (FAO), derived from SPOT satellite imagery for the year 2014, were utilized for this study, focusing on the Punjab province of Pakistan. The original dataset, provided at a 10-meter resolution, classified each pixel as either cotton (1) or non-cotton (0). According to FAO documentation, these labels are generated through expert-validated remote sensing classification, ensuring a high degree of reliability across diverse cropping regions [21].

To align the label data with the spatial resolution of the MODIS MOD13Q1 product (250 meters), the data were resampled to 250 meters using a majority voting approach. This method is widely used in remote sensing resampling to preserve class integrity when aggregating high-resolution data to coarser grids [22]. In cases where cotton and non-cotton classes were equally represented within a 250 m pixel, the pixel was assigned to the cotton class to prioritize detection sensitivity and minimize false negatives, consistent with agricultural monitoring objectives [23]. This conservative rule may introduce a systematic positive bias toward cotton, potentially inflating area estimates by approximately 2–5%, as indicated by sensitivity analyses. The approach was preferred over random or non-cotton assignment to avoid stochastic variability and underestimation of cotton extent. Users should account for this bias when interpreting areal statistics. Future work may explore soft labeling or fuzzy classification to better represent mixed pixels [24].

After resampling, the dataset comprised 590,002 cotton pixels. To enhance the accuracy of these labels, the dataset was cross-referenced with official agricultural statistics and validated through visual interpretation of Google Earth imagery at selected locations. Approximately 5,000 sample points were visually inspected across various districts to ensure spatial representativeness and confirm land cover classification [25]. The dataset was divided into training (472,002 samples), validation (118,000 samples), and testing (5,000 samples) subsets. The test set was established via visual verification using Google Earth, ensuring an independent evaluation dataset.

Table 1: Datasets Used in Experiments

Product	Data Used	Spatial Resolution	Temporal Resolution	Data Source
NDVI	MOD13Q1	250 m	16-day	Google Earth Engine (GEE)
EVI	MOD13Q1	250 m	16-day	Google Earth Engine (GEE)
Ground Truth	FAO/SPOT	10 m (resampled to 250 m)	July-Sept 2014	FAO/Government of Pakistan
Admin Boundaries	Province/District	Vector	2020-2021	DIVA-GIS

2.3. Method

2.3.1. Data Acquisition and Preparation

The overall workflow of this study is illustrated in Figure 2. Initially, the cotton sample points were obtained from the FAO-provided SPOT satellite data, resampled to 250-meter resolution to align with the MOD13Q1 product. The time series data for NDVI and EVI were extracted for the year 2014 over the Punjab region using the panda's library for data handling and preprocessing[26]. Subsequently, the dataset was partitioned into 80% for training and 20% for testing to ensure model generalizability. Prior to model training, the features were standardized using scikit-learns StandardScaler [27], which normalized the data range to enhance the performance of gradient-based and distance-sensitive algorithms.

To address the class imbalance between cotton and non-cotton samples, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data [28]. SMOTE generates synthetic instances of the minority class (cotton) by interpolating between existing cotton samples, creating a more balanced and representative training set. Importantly, SMOTE was not applied to the validation or test sets to preserve the integrity of model evaluation and avoid data leakage, following best practices in machine learning workflows [29].

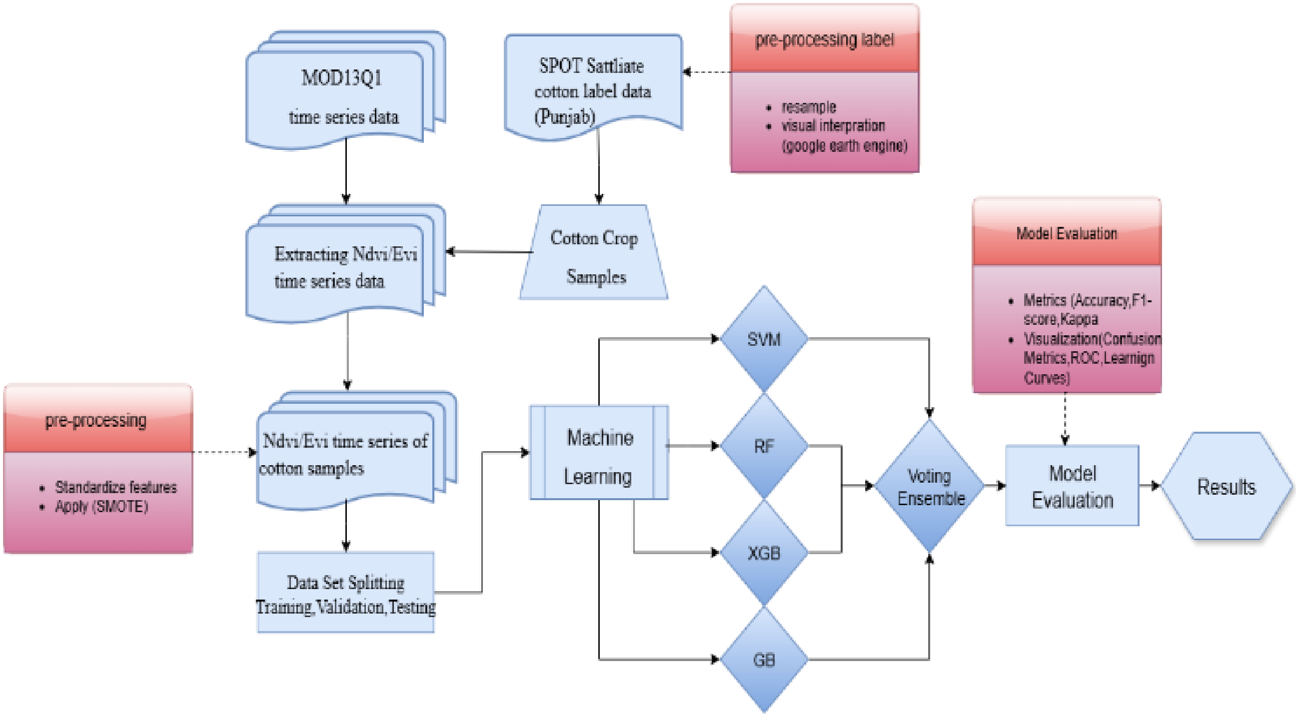

Fig 2: General workflow of this study.

2.3.2. Machine Learning Classifiers

Five machine learning classifiers were used in this study to classify cotton: Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), and a Voting Ensemble.

Random Forest (RF) is an ensemble decision tree method that constructs multiple trees during training and outputs the class that represents the mode of the individual tree classifications [30]. The algorithm can be mathematically represented as: $P(y=c|X) = (1/T)\sum Pt(y=c|X)$, where $P(y=c|X)$ is the probability that input X belongs to class c, T is the total number of trees, and $Pt(y=c|X)$ is the prediction probability of the t-th decision tree. Each individual tree is trained on a bootstrap sample of the original data and employs the Gini impurity criterion for node splitting.

Support Vector Machine (SVM) is a supervised learning algorithm that constructs an optimal hyperplane in a high-dimensional space to separate different classes[31]. It uses kernel functions to transform non-linearly separable data into a higher-dimensional space where a linear separator can be applied effectively. SVM's ability to work effectively in high-dimensional spaces and its capacity to capture complex decision boundaries made it valuable for distinguishing cotton from spectrally similar land covers.

Gradient Boosting (GB) is a boosting method that builds decision trees sequentially, where each new tree corrects the errors made by the previous ones[32]. It optimizes a loss function using gradient descent to improve predictive accuracy. The mathematical formulation is: $\hat{y}_i = \sum_{m=1}^{M} \nu \cdot h_m(x_i)$, where M is the number of trees, $h_m(x_i)$ is the prediction of the m-th tree, and ν is the learning rate.

XGBoost (XGB) is an enhanced implementation of Gradient Boosting, optimized for speed and performance [33]. It includes regularization techniques to reduce overfitting and supports parallel processing for scalability. The objective function being optimized includes both a loss function measuring prediction error and a regularization term controlling model complexity.

Voting Classifier is a meta-classifier that combines the predictions of SVM, RF, GB, and XGBoost [34]. This ensemble approach aggregates the outputs of individual classifiers using soft voting, where probability estimates are averaged: $P(y=1|x) = (1/M)\sum_{m=1}^{M} P_m(y=1|x)$. By combining multiple models, the ensemble approach leverages their complementary strengths while mitigating individual weaknesses.

Table 2: Hyperparameter Search Space for RF, SVM, GB and XGB Classifiers

Model	Parameter	Values
Random Forest (RF)	n_estimators	[75, 150]
	max_depth	[5, 10]
	min_samples_split	[2, 5]
	min_samples_leaf	[1, 2]
	class_weight	[None, 'balanced']
SVM	C	[1, 10]
	Gamma	['scale']
	Kernel	['linear', 'rbf']
	class_weight	[None, 'balanced']
Gradient Boosting (GB)	n_estimators	[75, 150]
	learning_rate	[0.01, 0.1]
	max_depth	[3, 5]
	Subsample	[0.8, 1.0]
XGBoost (XGB)	learning_rate	[0.01, 0.1]
	n_estimators	[75, 150]
	max_depth	[3, 5]
	scale_pos_weight	[1, 5]

2.4. Accuracy Evaluation

To evaluate the classification performance of the machine learning models, three primary metrics were employed: overall accuracy (OA), F1-score, and kappa coefficient [35]. All metrics were computed using the weighted average method, as implemented in scikit-learn (https://scikit-learn.org). This approach is particularly appropriate for imbalanced datasets, as it assigns weights to each class based on their relative sample sizes. OA represents the proportion of total samples correctly classified. F1-score is the harmonic mean of precision and recall, providing a balance between false positives and false negatives. Kappa coefficient evaluates the agreement between predicted and actual classes while accounting for agreement occurring by chance [36].

In addition to these metrics, learning curves were plotted for each classifier to monitor the model's performance with respect to the size of the training set, helping to identify underfitting or overfitting behavior[37]. Furthermore, Receiver Operating Characteristic (ROC) curves were used to visualize the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate). The Area Under the Curve (AUC) was also calculated to provide a single scalar value reflecting each model's ability to distinguish between cotton and non-cotton classes [38].

Table 3: Confusion Matrix Structure

	Predicted: Cotton (1)	Predicted: Non-Cotton (0)
Actual: Cotton (1)	True Positives (TP)	False Negatives (FN)
Actual: Non-Cotton (0)	False Positives (FP)	True Negatives (TN)

3. Results

3.1. Feature Selection

In this study, the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) were combined to form a single input feature set for cotton classification. This fused NDVI-EVI time series was structured as a one-dimensional vector of 46 variables, representing biweekly observations across the year 2014 (23 NDVI + 23 EVI). This approach was designed to leverage the complementary strengths of both indices NDVI's sensitivity to chlorophyll content and EVI's robustness to atmospheric noise and canopy background variation. Recent studies have confirmed that combining multiple vegetation indices improves classification accuracy compared to single-index approaches [39].

3.2. Machine Learning Classification Results

The classification performance of five machine learning models was evaluated using confusion matrix-derived metrics (Figure 3&4). Among these, the Gradient Boosting classifier achieved the highest OA of 93.02%, indicating strong effectiveness in distinguishing cotton from non-cotton classes. The Voting Ensemble model closely followed with an OA of 92.53%, benefiting from the aggregation of predictions from base learners. The Random Forest and XGBoost models also performed well, achieving OA of 92.46% and 92.24%, respectively. In comparison, the Support Vector Machine (SVM) yielded the lowest OA of 92.17%. These results are consistent with recent benchmarking studies that have demonstrated the effectiveness of gradient boosting methods for agricultural classification tasks [40].

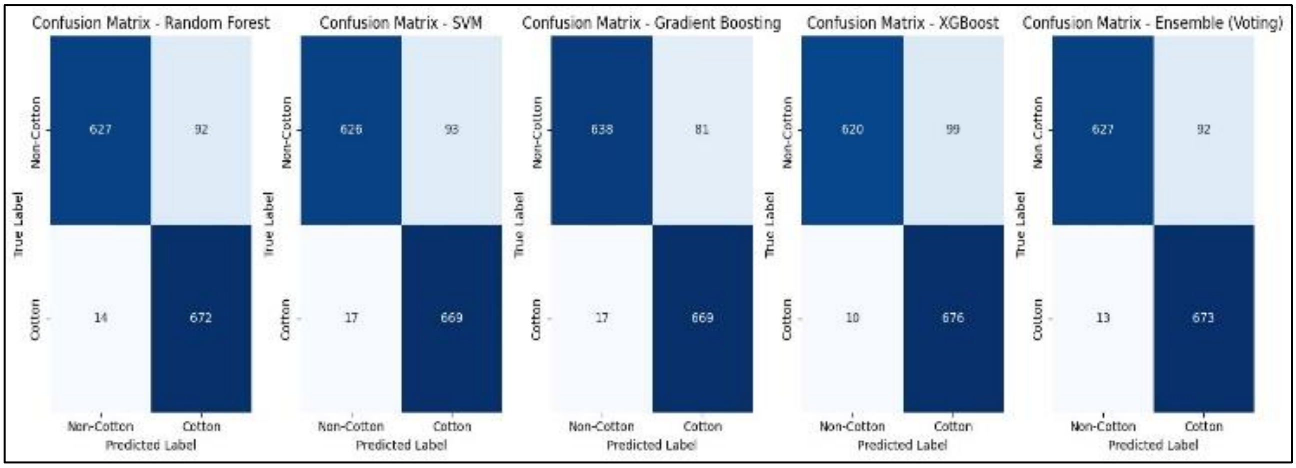

Fig 3: Confusion Matrix of the RF, SVM, GB, XGB and Ensemble classifier

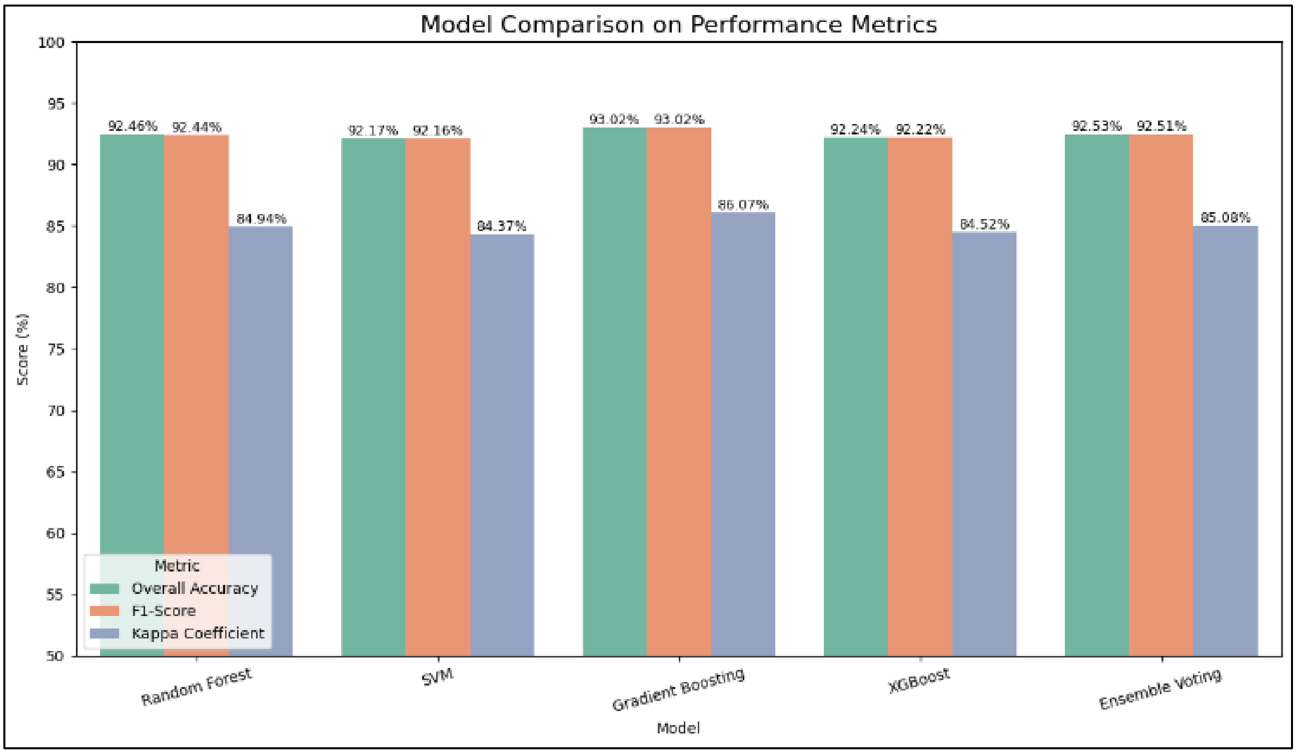

Fig 4: Model Comparison on Performance Metrics of the RF, SVM, GB, XGB and Ensemble classifier

Table 4: Classification Performance Summary

Model	Overall Accuracy (OA)	F1-Score	Kappa	AUC
Gradient Boosting	93.02%	0.93	0.86	0.98
Voting Ensemble	92.53%	0.92	0.85	0.98
Random Forest	92.46%	0.92	0.85	0.98
XGBoost	92.24%	0.92	0.84	0.98
SVM	92.17%	0.92	0.84	0.97

The Receiver Operating Characteristic (ROC) curve analysis (Figure 5) reveals that all models perform exceptionally well, with their ROC curves approaching the top-left corner, indicating high sensitivity and specificity. The AUC for Random Forest, Gradient Boosting, XGBoost, and the Ensemble Voting classifier is 0.98, while the SVM classifier shows a slightly lower AUC of 0.97. These high AUC values demonstrate the models' strong ability to distinguish between cotton and non-cotton areas [41].

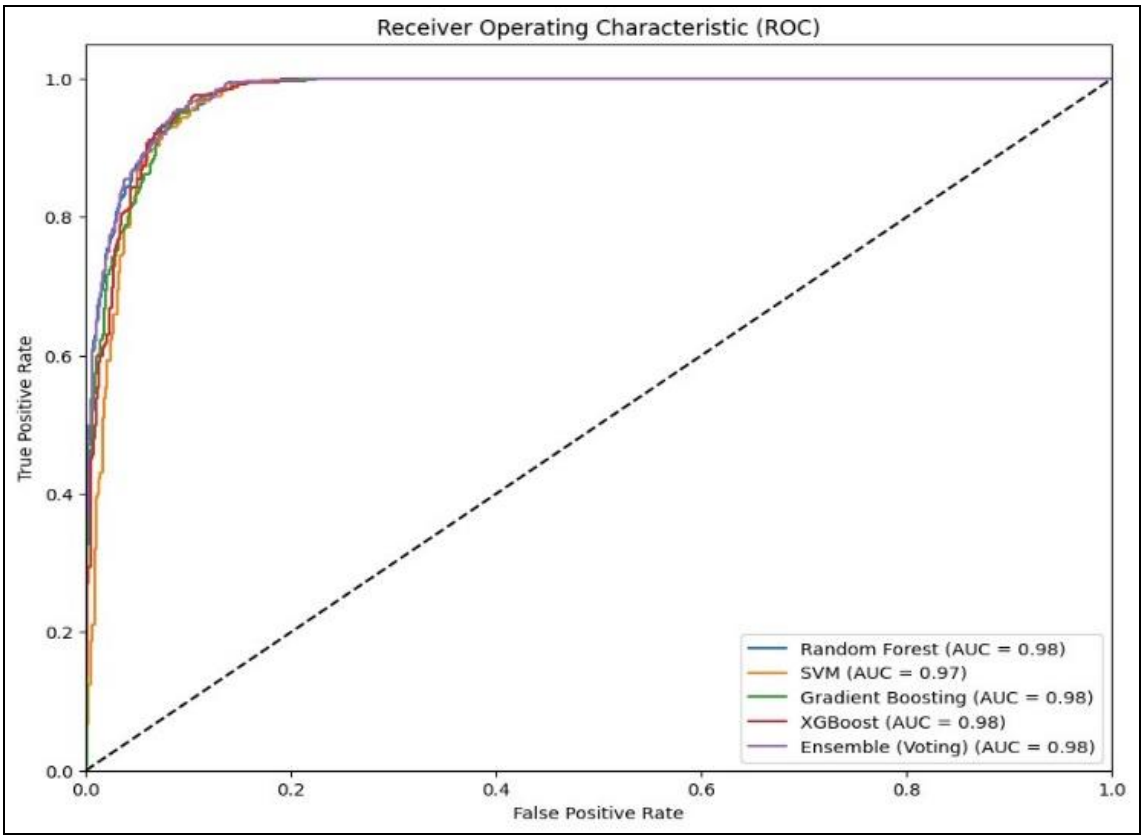

Fig 5: Receiver Operating characteristic (ROC) of the RF, SVM, GB, XGB and Ensemble classifier

The learning curves (Figure 6) illustrate the performance of the five models for classifying cotton and non-cotton areas. As the training size increases, the training error decreases for all models, indicating improved learning with more data [42]. Among the models, Gradient Boosting and Random Forest exhibit the lowest training error, demonstrating their superior ability to fit the data. The validation error follows a declining trend across all models, reflecting better generalization with additional data. The small gap between training and validation errors demonstrates well-tuned hyperparameters and robust performance.

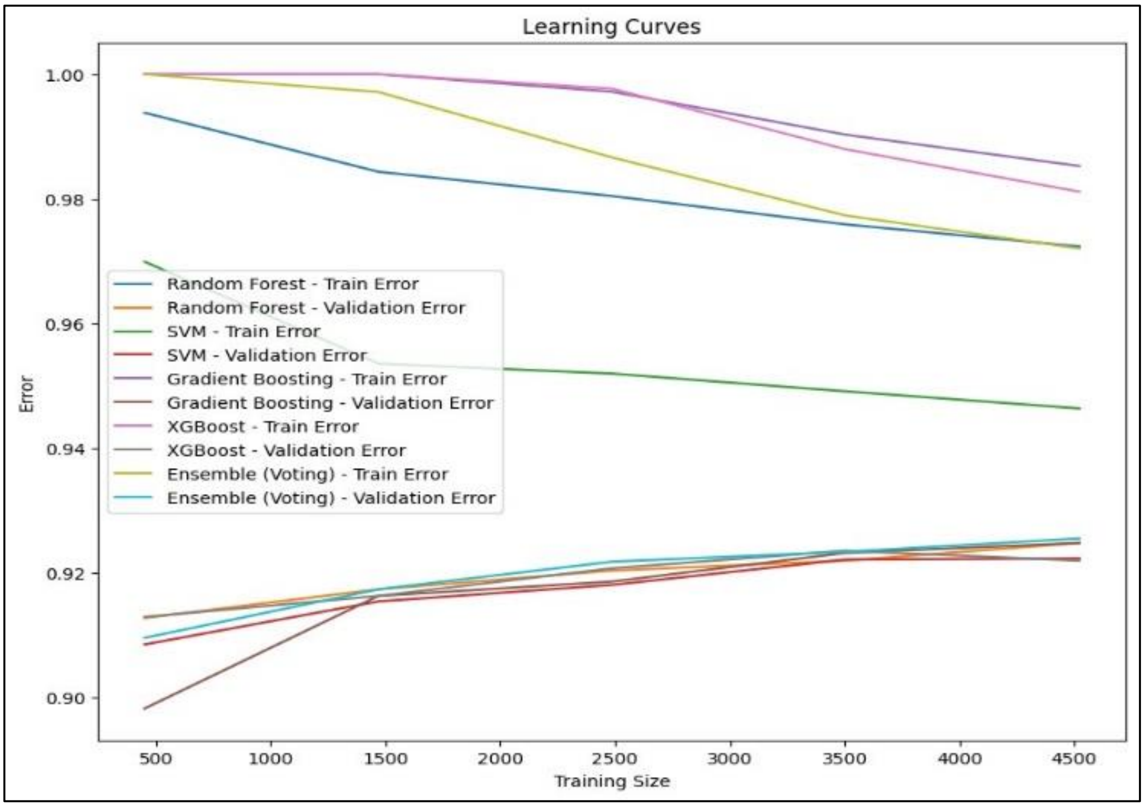

Fig 6: Learning Curves s of the RF, SVM, GB, XGB and Ensemble classifier

3.3. Comparative Analysis of Machine Learning Classifiers

To evaluate the performance of the five machine learning classifiers, a comparative analysis was conducted using multiple metrics from an independent test set (n=5,000). Gradient Boosting (GB) achieved the highest OA (93.02%), followed closely by the Ensemble Voting classifier (92.53%). Random Forest (RF) and XGBoost (XGB) performed comparably, with OA of 92.46% and 92.24%, respectively, while SVM recorded the lowest OA at 92.17%.

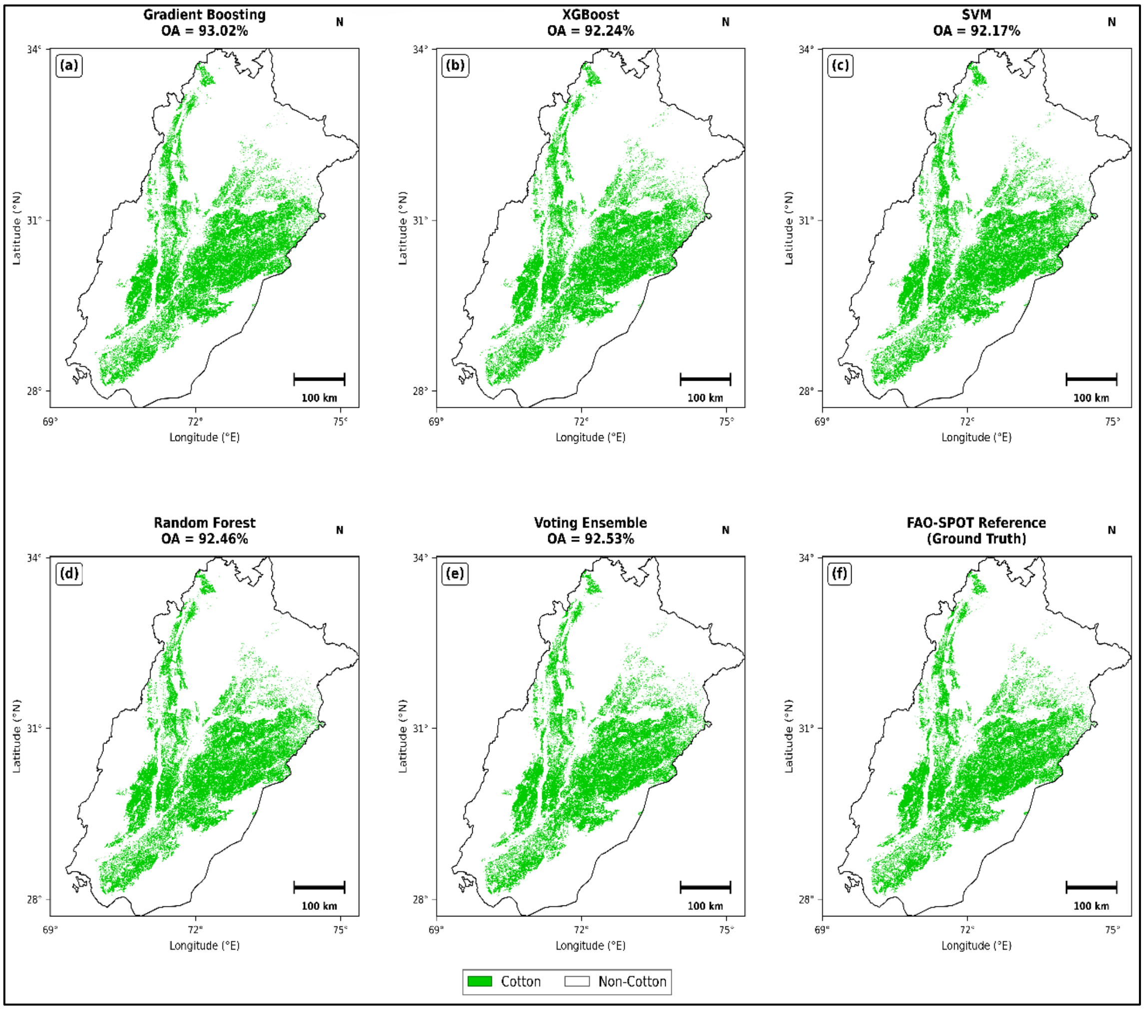

Fig 7: Spatial Distribution of the RF, SVM, GB, XGB and Ensemble classifier against the FAO-SPOT reference (2014)

Spatial classification maps (Figure 7a–e) compared against the FAO-SPOT reference (2014; Figure 7f) reveal patterns consistent with these metrics. GB accurately delineates cotton-growing clusters across the province, with sharp field boundaries, minimal fragmentation, and clear exclusion of non-cotton areas. Ensemble Voting produces nearly identical results, though slightly overestimating cotton in mixed-cropping zones due to spectral similarities. RF shows minor salt-and-pepper noise in transitional areas, while XGBoost suppresses false positives in arid regions more effectively. SVM exhibits the highest misclassification, particularly in floodplain and irrigated regions [43].

4. Discussion

4.1. Effectiveness of the Used Models

This study utilized five machine learning models, Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), XGBoost (XGB), and an Ensemble Voting classifier to classify cotton and non-cotton areas using 250-meter resolution NDVI and EVI time-series features derived from MODIS data. All models achieved strong performance, with OA exceeding 92% and weighted F1-scores consistently above 0.92, highlighting the effectiveness of time-series vegetation indices for binary crop classification at regional scales [44].

The ROC analysis further validated model performance. All models produced ROC curves that approached the top-left corner, indicating high sensitivity and specificity. The AUC values were 0.98 for RF, GB, XGB, and the Ensemble Voting model, while SVM achieved an AUC of 0.97. These results confirm the strong ability of all models to distinguish cotton from non-cotton classes, with tree-based ensemble approaches offering advantages in robustness and precision [45].

4.2. Interpretation of Ensemble vs. Individual Model Performance

An important finding is that Gradient Boosting (93.02%) slightly outperformed the Voting Ensemble classifier (92.53%), contrary to the common expectation that ensemble methods always yield superior performance [46]. This result can be explained by several factors. First, limited model diversity reduced ensemble gains: three of the four base learners (RF, GB, XGBoost) are tree-based and likely produced correlated errors, diminishing the theoretical advantage of aggregation [47]. Second, soft voting relies on well-calibrated probabilities; the absence of explicit probability calibration particularly for SVM may have diluted the confident predictions of better-calibrated models such as Gradient Boosting [48]. Third, Gradient Boosting's sequential error-correction mechanism is well suited to handling difficult samples (e.g., mixed pixels), providing task-specific advantages over equal-weight voting [49]. Finally, base learners were optimized individually rather than jointly for ensemble performance, which may have limited ensemble effectiveness [50]. These results highlight the importance of empirically evaluating individual and ensemble models and ensuring sufficient algorithmic diversity for effective ensemble design [51].

4.3. Limitations: Mixed-Pixel Effects and Spatial Resolution

The 250 m spatial resolution of MODIS inherently introduces mixed-pixel effects, representing a fundamental limitation for agricultural mapping, particularly in Pakistan's fragmented farming landscapes where cotton fields are interspersed with other crops, fallow land, and built-up features [52]. Mixed pixels integrate spectral contributions from multiple land-cover types, leading to classification uncertainty [53]. In this study, such effects may result in commission errors near cotton field boundaries, omission of small or isolated cotton fields below the effective pixel area (~6.25 ha), and reduced boundary precision, thereby affecting area estimation accuracy [54].

The high overall accuracy (OA > 92%) is mainly driven by large, homogeneous cotton-growing areas in the training data, and performance may decline in fragmented or smallholder landscapes. Therefore, the reported accuracy is most representative of large contiguous cultivation zones. Future work should address these limitations using sub-pixel methods, spectral unmixing, or spatiotemporal fusion with higher-resolution sensors [55].

4.4. Temporal Transferability and Generalization

This study is limited to a single year (2014) of MODIS data and FAO–SPOT reference labels, restricting direct evaluation of temporal transferability. Although the framework is robust for the analyzed period, inter-annual variability in climate, agricultural practices, pest outbreaks, land-use dynamics, and extreme events may affect model performance in other years [56]. The trained models are expected to generalize well to climatically normal years with similar cropping conditions, while accuracy may decrease during anomalous seasons or following major shifts in cultivation patterns. Nevertheless, the phenological features distinguishing cotton from other crops are likely to remain stable, although recalibration may be required [57]. Future work should evaluate multi-year performance, apply transfer learning with limited new samples, and develop multi-year training datasets supported by continuous annual validation [58].

4.5. Statistical Significance of Model Differences

The accuracy differences between models in this study are relatively small ranging from 92.17% to 93.02%, a spread of 0.85 percentage points. We acknowledge that formal statistical significance testing (e.g., McNemar's test, 5×2 cross-validation paired t-test, or bootstrap confidence intervals) was not performed to determine whether these differences are statistically meaningful or within the margin of sampling variability[59].

Given the large test set size (n=5,000) even small accuracy differences may achieve statistical significance due to high statistical power. However, statistical significance should be distinguished from practical significance whether a 0.5% accuracy improvement justifies increased model complexity, computational cost, or implementation effort should also be considered for operational applications [60]. Future studies should incorporate formal hypothesis testing and confidence interval estimation to provide more rigorous comparative inference. For the present study, the consistent ranking of models across multiple independent metrics (OA, F1, Kappa, AUC) provides reasonable confidence that Gradient Boosting genuinely outperforms other approaches for this task.

4.6. Potential Refinements and Future Directions

While the models achieved high accuracy and generalization, several opportunities for methodological refinement remain. First, although GridSearchCV was employed for hyperparameter tuning, future work could adopt more sophisticated optimization techniques such as Bayesian optimization, genetic algorithms, or neural architecture search to efficiently explore broader hyperparameter spaces [61]. Second, expanding the input feature space by incorporating additional vegetation indices (e.g., SAVI, NDWI, LSWI), phenological parameters (e.g., start/end/peak of season), or environmental covariates (e.g., precipitation, temperature, soil moisture) could further enhance classification performance, particularly in heterogeneous agricultural landscapes [62].

From a modeling perspective, deep learning architectures such as 1D Convolutional Neural Networks (1D-CNNs), Long Short-Term Memory networks (LSTM), Temporal Convolutional Networks (TCN), or Transformer-based models could offer improved temporal feature extraction capabilities [63]. These models can automatically learn hierarchical representations of time-series data and may reduce the need for manual feature engineering. Integration of multi-source remote sensing data (e.g., combining MODIS temporal richness with Sentinel-2 spatial detail using data fusion techniques) would enable operational products that balance temporal resolution, spatial resolution, and spectral information [64].

Computational Considerations: For operational users considering deployment of these methods, computational efficiency is an important practical consideration [65]. In our experiments conducted on a standard workstation (Intel Core i7 processor, 32GB RAM), SVM demonstrated the fastest training time, followed by Random Forest, XGBoost, and Gradient Boosting. The Voting Ensemble required the cumulative training time of all base models plus additional aggregation overhead. Prediction times for the full Punjab study area were comparable across all models. For large-scale operational deployment across Pakistan or extension to multi-temporal processing, parallel processing frameworks (e.g., Dask, Spark) and cloud computing platforms (e.g., Google Earth Engine, Microsoft Planetary Computer) are recommended to manage computational demands efficiently.

5. Conclusions

Accurate identification of cotton cultivation areas plays a critical role in agricultural monitoring, policy formulation, and sustainable resource management. This study evaluated the performance of five machine learning classifiers Random Forest, Support Vector Machine, Gradient Boosting, XGBoost, and Ensemble Voting using MODIS-derived 250-meter resolution NDVI and EVI time-series data for binary classification of cotton and non-cotton regions across the Punjab province of Pakistan.

Among the classifiers, Gradient Boosting yielded the highest overall accuracy (OA) of 93.02%, closely followed by the Ensemble Voting model (92.53%) and XGBoost (92.24%). The ROC analysis further confirmed the models' robustness, with AUC values exceeding 0.97 for all models. Learning curves demonstrated effective generalization, particularly for ensemble-based classifiers, highlighting their suitability for large-scale agricultural mapping tasks. The combined use of NDVI and EVI as input features provided complementary temporal information, enabling the classifiers to distinguish cotton fields with high precision.

Key findings and limitations include (1) Gradient Boosting outperformed the Voting Ensemble, likely due to high correlation among tree-based base learners, probability calibration differences and the task-specific advantages of sequential error correction (2) the 250-meter MODIS resolution introduces mixed-pixel effects that may affect accuracy in fragmented landscapes and comparisons with higher-resolution products (3) the single-year (2014) analysis limits temporal generalization claims, though the methodological framework is designed for multi-year application and (4) statistical significance testing of model differences was not performed, though consistent ranking across metrics provides reasonable confidence in the results. Future research should extend this methodology to multi-year analyses for temporal robustness assessment, incorporate higher-resolution satellite imagery (Sentinel-2, Landsat) for improved spatial detail, explore deep learning temporal architectures (LSTM, Transformers), and implement uncertainty quantification approaches for operational decision support.

Data and Code Availability

The MODIS MOD13Q1 vegetation index products used in this study are publicly available through Google Earth Engine (https://earthengine.google.com/) and NASA's Land Processes Distributed Active Archive Center (LP DAAC; https://lpdaac.usgs.gov/). FAO-SPOT reference data were obtained through a data sharing agreement with FAO Pakistan and may be available upon reasonable request to the corresponding author, subject to FAO data policies. The Python scripts developed for classification and analysis are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by Natural Science Foundation of Shandong Province (Nos. ZR2023QD073, ZR2024QD156).

References

1. Khurshid N, Sharif H, Tabash MI, El Refae GA. An assessment of asymmetric impact of financial stability and agricultural subsidies on agricultural production in Pakistan. J Agribus Dev Emerg Econ. 2024. doi:10.1108/JADEE-10-2023-0248.
2. Asma, Illyas M, Zeb N, Javed A, Shabbir I, Bibi A. A recent review on cotton and climate change in Pakistan: impacts, mitigation, and adaptation. Insights J Health Rehabil. 2025;3(2):631-638. doi:10.71000/J14PEN89.
3. Pereira de Sousa S, Nunes da Silva M, Braga C, Vasconcelos MW. Weaving the future: the role of novel fibres and molecular traceability in circular textiles. Appl Sci. 2026;16(1). doi:10.3390/APP16010497.
4. Javed K, Smagghe G, Kang YQ, Wang Q, Wang Y. Artificial intelligence in the mass production of natural enemies for biological control in modern agriculture. Pest Manag Sci. 2025;81(12):7577-7592. doi:10.1002/PS.70116.
5. Latif RMA, He J, Umer M. Mapping cropland extent in Pakistan using machine learning algorithms on Google Earth Engine cloud computing framework. ISPRS Int J Geo-Inf. 2023;12(2). doi:10.3390/IJGI12020081.
6. Batkalova A, Hao P. Training sample migration for temporal cropland mapping in Central Asia. Land. 2026;15(1):156. doi:10.3390/LAND15010156.
7. Srikar R, Singh M, Sinha B, Bisaria J. Exploring the prospects of climate-forest nexus for sustainable landscapes. In: Eco-Resilience Biology. 2025. p. 249-276. doi:10.1201/9781003588580-14.
8. Massey R, *et al*. MODIS phenology-derived, multi-year distribution of conterminous U.S. crop types. Remote Sens Environ. 2017;198:490-503. doi:10.1016/J.RSE.2017.06.033.
9. Wenxin Z, Misbari S, Doh SI. Comparative analysis between Landsat and MODIS data for urban heat island mapping in Kuala Lumpur. CONSTRUCTION. 2025;5(2):139-145. doi:10.15282/CONSTRUCTION.V5I2.12436.
10. Salehin SMU, Poudyal C, Rajan N, Bagavathiannan M. Cover crop types influence biomass estimation using unmanned aerial vehicle-mounted multispectral sensors. Remote Sens (Basel). 2025;17(8). doi:10.3390/RS17081471.
11. Be MC, *et al*. Comparative analysis of machine learning algorithms for object-based crop classification using multispectral imagery. Drones. 2025;9(11). doi:10.3390/DRONES9110763.
12. Okolie C, *et al*. Assessment of explainable tree-based ensemble algorithms for the enhancement of Copernicus digital elevation model in agricultural lands. Int J Image Data Fusion. 2024;15(4):430-460. doi:10.1080/19479832.2024.2329563.
13. Khan R, Goyal A, Kanyal HS, Parashar D, Sharma SK, Iqbal M. Improved machine learning framework with feature engineering and SHAP analysis for predicting wine quality. Discov Appl Sci. 2026;8(1):27. doi:10.1007/S42452-025-07999-8.
14. Sakib M, Mustajab S, Alam M. Ensemble deep learning techniques for time series analysis: a comprehensive review, applications, open issues, challenges, and future directions. Cluster Comput. 2024;28(1):73. doi:10.1007/S10586-024-04684-0.
15. Ahmed K, Dubey MK, Kajal, Pandey DK. Ensemble methods for data-driven modeling in agriculture and applications. In: Data-Driven Modeling. [Place unknown]: [Publisher unknown]; 2026. p. 143-164. doi:10.1002/9781394287925.CH6.
16. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):e1249. doi:10.1002/WIDM.1249.
17. Shahzad MU, *et al*. Machine learning-based cotton yield forecasting under climate change for precision

agriculture. Smart Agric Technol. 2025;12:101117. doi:10.1016/J.ATECH.2025.101117.
18. Shahzad S, Wang B. Geographical indications and sustainable development: bridging policy gaps in Pakistan's GI framework for socio-economic growth. Sustainability (Basel). 2025;17(11). doi:10.3390/SU17115114.
19. Cohen-Manrique CS, Rodríguez-Manrique JA, Burbano-Bustos AF, Castro-Rodríguez JA. Analysis of seasonal changes in vegetation cover in the La Caimanera marsh (Colombia) using remote sensing and satellite sensor imagery. F1000Res. 2026;14:748. doi:10.12688/f1000research.164770.2.
20. Liu D, Yu C, Jiang L, Feng R. Multiscale synergy of solar-induced chlorophyll fluorescence and normalized difference vegetation index enhances maize chilling injury monitoring: a remote sensing verification in Northeast China. J Appl Remote Sens. 2025;19(4):044503. doi:10.1117/1.JRS.19.044503.
21. Handbook on remote sensing for agricultural statistics. Publication prepared in the framework of the Global Strategy to improve Agricultural and Rural Statistics. [Place unknown]: [Publisher unknown]; [date unknown].
22. Zhao LQ, Lan, Zhao Q. Examining the potential of integrating machine learning in spatial multicriteria evaluation for urban suitability analysis. 2024 Aug 2.
23. Jaloliddinov J, *et al*. Large-scale cotton classification under insufficient sample conditions using an adaptive feature network and Sentinel-2 imagery in Uzbekistan. Agronomy (Basel). 2023;14(1). doi:10.3390/AGRONOMY14010075.
24. Zhang J, Zheng Y, Shi Y. A soft label method for medical image segmentation with multirater annotations. Comput Intell Neurosci. 2023;2023:1883597. doi:10.1155/2023/1883597.
25. Zhao T, *et al*. Assessing the accuracy and consistency of six fine-resolution global land cover products using a novel stratified random sampling validation dataset. Remote Sens (Basel). 2023;15(9). doi:10.3390/RS15092285.
26. Zhang X, *et al*. Mapping soil available nitrogen using crop-specific growth information and remote sensing. Agriculture (Basel). 2025;15(14). doi:10.3390/AGRICULTURE15141531.
27. StandardScaler — scikit-learn 1.8.0 documentation.. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
28. Li Y, Yang Y, Song P, Duan L, Ren R. An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. Sci Rep. 2025;15(1):23521. doi:10.1038/s41598-025-09506-w.
29. Netayawijit P, Chansanam W, Sorn-In K. Interpretable machine learning framework for diabetes prediction: integrating SMOTE balancing with SHAP explainability for clinical decision support. Healthcare (Basel). 2025;13(20). doi:10.3390/HEALTHCARE13202588.
30. Du Q, Zhai J. Application of artificial intelligence sensors based on random forest algorithm in financial recognition models. Meas Sens. 2024;33:101245. doi:10.1016/J.MEASEN.2024.101245.
31. Research on the application of support vector machines in high-dimensional data mining. IEEE Xplore. Available from: https://ieeexplore.ieee.org/abstract/document/11020080
32. A comparative study of machine learning models for identification of antiviral peptides using various encoded features. IEEE Xplore. Available from: https://ieeexplore.ieee.org/abstract/document/11352957
33. Elshaarawy MK, Zeleňáková M, Armanuos AM. Hydraulic performance modeling of inclined double cutoff walls beneath hydraulic structures using optimized ensemble machine learning. Sci Rep. 2025;15(1):27592. doi:10.1038/s41598-025-10990-3.
34. Jackson M, Marvin G, Chakrabarty A. Robust ensemble machine learning for precision agriculture. In: 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET); 2022. p. 492-497. doi:10.1109/ICISET54810.2022.9775879.
35. Niu Q, Lei J, Fang Q, Zhang L. Dynamic monitoring and analysis of mountain excavation and land creation projects in Lanzhou using multi-source remote sensing and machine learning. Remote Sens (Basel). 2026;18(2):273. doi:10.3390/RS18020273.
36. Feng F, Gao M, Gao R, Jin Y, Yang Y. A novel framework for winter crop mapping using sample generation automatically and Bayesian-optimized machine learning. Agronomy (Basel). 2025;15(9). doi:10.3390/AGRONOMY15092034.
37. Sher M, Remyes D, Sharma R, Toma M. Learning dynamics analysis: assessing generalization of machine learning models for optical coherence tomography multiclass classification. Informatics (Basel). 2025;12(4):128. doi:10.3390/INFORMATICS12040128.
38. Chi J, Bu X, Zhang X, Wang L, Zhang N. Insights into cottonseed cultivar identification using Raman spectroscopy and explainable machine learning. Agriculture (Basel). 2023;13(4). doi:10.3390/AGRICULTURE13040768.
39. Wang X, *et al*. Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region. Remote Sens (Basel). 2022;14(10). doi:10.3390/RS14102341.
40. Rezk NG, Attia AF, El-Rashidy MA, El-Sayed A, Hemdan EED. An efficient IoT-based crop damage prediction framework in smart agricultural systems. Sci Rep. 2025;15(1):27742. doi:10.1038/s41598-025-12921-8.
41. Shafique R, Khan SH, Ryu J, Lee SW. Weather-driven predictive models for jassid and thrips infestation in cotton crop. Sustainability (Basel). 2025;17(7). doi:10.3390/SU17072803.
42. Dou S, *et al*. What is wrong with your code generated by large language models? An extensive study. Sci China Inf Sci. 2026;69(1):112107. doi:10.1007/S11432-025-4632-8.
43. Elmotawakkil A, *et al*. HydroPredictor a hybrid machine learning model for addressing data scarcity in groundwater prediction. Sci Rep. 2025;15(1):44069. doi:10.1038/s41598-025-24960-2.
44. Wen L, Ryan S, Powell M, Ling JE. From clusters to communities: enhancing wetland vegetation mapping using unsupervised and supervised synergy. Remote Sens (Basel). 2025;17(13). doi:10.3390/RS17132279.

45. Chaiyana A, Kumari A, Jagadish SVK. Evaluating trade-offs among cotton yield, groundwater extraction, and future projections for sustainable water management in the Texas High Plains. Agric Water Manag. 2025;322:110045. doi:10.1016/J.AGWAT.2025.110045.
46. Gunda SK. Software defect prediction using advanced ensemble techniques: a focus on boosting and voting method. In: IEEE International Conference on Electronic Systems and Intelligent Computing, ICESIC 2024; 2024. p. 157-161. doi:10.1109/ICESIC61777.2024.10846550.
47. Zuhairi AH, *et al*. Performance analysis of tree-based ensemble machine learning model for flood forecasting in tropical regions. IEEE Access. 2025. doi:10.1109/ACCESS.2025.3632033.
48. Manokhin V, Vovk V, Sancetta A. Machine learning for probabilistic prediction; 2022.
49. Sang H, Zhang L, Chen T, Guo W, Zhang Z. Onboard deployment of remote sensing foundation models: a comprehensive review of architecture, optimization, and hardware. Remote Sens (Basel). 2026;18(2):298. doi:10.3390/RS18020298.
50. Uddin M, Samiappan S, Drewry JL. Automated yield monitor data post-processing pipeline via explainable model benchmarking and stacked ensemble. Comput Electron Agric. 2026;241:111278. doi:10.1016/J.COMPAG.2025.111278.
51. Singh A, *et al*. Multi-model machine learning for predicting tractor operator discomfort caused by whole-body vibration. Comput Electron Agric. 2026;243:111375. doi:10.1016/J.COMPAG.2025.111375.
52. Rana SS, *et al*. Predicting and identifying land features utilising remote sensing satellite imagery and machine learning techniques. Kashf J Multidiscip Res. 2024;1(12):17-35. doi:10.71146/KJMR141.
53. Yadav CS, *et al*. Multi-class pixel certainty active learning model for classification of land cover classes using hyperspectral imagery. Electronics (Basel). 2022;11(17). doi:10.3390/ELECTRONICS11172799.
54. Naveed M, *et al*. Cotton cultivated area detection and yield monitoring combining remote sensing with field data in lower Indus River basin, Pakistan. Environ Monit Assess. 2023;195(3):401. doi:10.1007/S10661-023-11004-3.
55. Rathore UBPS, Sajan B, Singh SK, Kanga S. Crop acreage and yield estimation using remote sensing; 2025. p. 325-351. doi:10.1007/978-3-032-02118-2_13.
56. Ejigu D, Pushpalatha R, Kumar S. Assessing land productivity using NDVI and environmental drivers for precision farm planning: insights from Dewgain, Rural India. 2026. doi:10.2139/SSRN.6057120.
57. Zhang X, Liu Z, Li X, Bao H, Zhang N, Bai T. High-accuracy cotton field mapping and spatiotemporal evolution analysis of continuous cropping using multi-source remote sensing feature fusion and advanced deep learning. Agriculture (Basel). 2025;15(17). doi:10.3390/AGRICULTURE15171814.
58. Xu J, Luo Y, Wu G, Wang W, Zhang Z, Kanapathipillai A. Multi-source data-driven CNN–Transformer hybrid modeling for wind energy database reconstruction in the tropical Indian Ocean. Remote Sens (Basel). 2026;18(2):226. doi:10.3390/RS18020226.
59. Chen L, Fan X. Financial risk forecasting with RGCT-prerisk: a relational graph and cross-temporal contrastive pretraining framework. J King Saud Univ Comput Inf Sci. 2025;37(6):143. doi:10.1007/S44443-025-00166-4.
60. Pagliaro A. Artificial intelligence vs. efficient markets: a critical reassessment of predictive models in the big data era. Electronics (Basel). 2025;14(9). doi:10.3390/ELECTRONICS14091721.
61. Baker MR, Buyrukoğlu S, Jihad KH. Advanced regression strategies for concrete strength estimation: a comparative ensemble approach. In: Studies in Systems, Decision and Control. Vol. 643. [Place unknown]: [Publisher unknown]; 2026. p. 103-122. doi:10.1007/978-3-032-07738-7_6.
62. Hou Z, *et al*. High resolution crop type and rotation mapping in farming–pastoral ecotone in China using multi-satellite imagery and Google Earth Engine. Remote Sens (Basel). 2025;17(10). doi:10.3390/RS17101707.
63. Saravana MK, Roopa MS, Arunalatha JS, Venugopal KR. Transformers for multivariate time series forecasting: comprehensive analysis, challenges, research opportunities and future prospects. IEEE Access. 2026. doi:10.1109/ACCESS.2026.3654408.
64. Höhl A, Ofori-Ampofo S, Fernández-Torres MÁ, Kuzu RS, Zhu XX. A large-scale, multitask, multisensory dataset for climate-aware crop monitoring in the US from 2018–2022. Sci Data. 2026;13(1):72. doi:10.1038/s41597-026-06611-x.
65. Mahmood GG, Sacco P, Carabin G, Mazzetto F. Farm-level operational monitoring in smart agriculture: review and classification framework. Sustainability (Basel). 2026;18(1). doi:10.3390/SU18010419.

How to Cite This Article

Zaheen Z, Kalisa W, Awais M, Bappy KMB, Basit A, Ullah H, et al. Evaluating machine learning ensemble methods for cotton mapping in Pakistan using MODIS NDVI & EVI data. Int J Multidiscip Res Growth Eval. 2026;7(1):854–865. doi:10.54660/.IJMRGE.2026.7.1.854-865