Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Evaluation of Machine Learning Algorithms for Irrigated Area Mapping in the Upper Blue Nile Basin

Abdurezak Arage Hamza

abdurezak.arage@haramaya.edu.et

Addis Ababa University

Zeleke Agide Dejen

Addis Ababa University

Sirak Tekleab Gebrekristos

Addis Ababa University

Haileyesus Belay Lakew

Addis Ababa University

Tewele Gebretsadkan Haile

Adigrat University

Research Article

Keywords: Irrigated area, Machine Learning, Random Forest, Gradient Tree Boosting, Remote Sensing, Upper Blue Nile Basin

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11524145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparative Evaluation of Machine Learning Algorithms for Irrigated Area Mapping in the Upper Blue Nile Basin

Abdurezak Arage Hamza [1,3], Zeleke Agide Dejen [2], Sirak Tekleab Gebrekristos [2], Haileyesus Belay Lakew [2], Tewele Gebretsadkan Haile [4]

[1] Africa Center of Excellence for Water Management, Addis Ababa University, Addis Ababa, Ethiopia

[2] Center for Water Research, Institute of Water, Environment and Climate Research, Addis Ababa University, Addis Ababa, Ethiopia

[3] School of Water Resource and Environmental Engineering, Haramaya Institute of Technology, Haramaya University, Haramaya, Ethiopia

[4] College of Engineering and Technology, Adigrat University, Adigrat, Ethiopia

Email: abdurezakhwre@gmail.com

Abstract

Reliable mapping of irrigated agriculture is important for sustainable water resource management, allocation and agricultural productivity, particularly in data-scarce river basins such as Upper Blue Nile Basin. Despite the basin's significant irrigation potential, accurate and spatially explicit irrigated area maps remain limited. This study evaluated and compared machine learning classifiers including Random Forest (RF), Support Vector Machine (SVM), Gradient Tree Boosting (GTB), and Bidirectional Long Short-Term Memory (BiLSTM) for mapping irrigated areas in the Upper Blue Nile Basin, Ethiopia during the December 2025 to April 2026 irrigation season using multi source remote sensing datasets. The classifiers used two complementary data groups: multi-source remote-sensing and ancillary predictor variables, including Sentinel-2 multispectral imagery, Sentinel-1 synthetic aperture radar backscatter, CHIRPS rainfall estimates, and NASADEM-derived topographic variables; and ground-reference samples of irrigated and rainfed fields for model training and validation. The model performance was evaluated using the performance metrics overall accuracy, kappa coefficient, F1-score, precision, recall, and AUC-ROC, supported by nested spatial cross-validation with five outer spatial folds. Among the four classifiers, GTB scored the highest overall performance with accuracy, Kappa coefficient, F1-score, with values of 0.96, 0.91, and 0.96, respectively. RF ranked second and scored the highest AUC-ROC value of 0.99, followed by BiLSTM and SVM. Analysis showed that February

Sentinel-2 spectral indices were the most influential predictors and February was the most important acquisition month for GTB and RF. Multi model ensemble was used to produce a conservative consensus map. The consensus map classified 588,545 ha as irrigated agriculture in the Upper Blue Nile Basin.

Keywords: Irrigated area, Machine Learning, Random Forest, Gradient Tree Boosting, Remote Sensing, Upper Blue Nile Basin

1. Introduction

Globally, Irrigated agriculture uses approximately 70% of the available freshwater withdrawals [1] and up to 90% in some least income countries [2]. Due to population growth and climate variability efficient irrigation system management and monitoring have become critical to ensure food security and sustainable water resource utilization [3]. To understand the impact of consumptive water use by agriculture development of high-resolution irrigated area maps is essential [4]. Recent technological advancement in irrigated agriculture significantly improves water productivity and reduces environmental impacts [1].

Challenges in food security and increasing water scarcity need accurate irrigation mapping for sustainable agricultural water management [5]. Traditional ground-based irrigation mapping is inadequate for large scale irrigation assessment due to lack of spatial and temporal extent and high associated cost [6]. As a result, remote sensing technologies have become reliable tools for providing spatially comprehensive and temporally dynamic information on irrigated agriculture [7]. However, in heterogeneous smallholder farms, distinguishing irrigated areas from rainfed agriculture areas pose significant challenge [8].

Upper Blue Nile is a transboundary basin that is one of the essential water resources feeding into the Nile River and supporting around 250 million people. The basin experiences limited research on irrigated areas and water demands due to lack of data and faces significant environmental pressures due to poor land management practices [9]. Recent studies in the basin indicate dramatic land use changes especially increasing the cultivated lands [10]. The basin's irrigation potential is largely untapped, except small fractions of suitable land being currently under irrigation [11].

Although irrigation is expanding in the basin, its actual extent remains uncertain because many smallholder irrigated fields are scattered and fragmented across the basin. This makes it challenging to conduct irrigation planning, monitor the growth of irrigation, assess agricultural water demands, and analyze the hydrological impacts of irrigation.

Remote sensing has become a powerful tool for monitoring irrigated agriculture across various climatic conditions. Over the past decades advancement of remote sensing technologies established vegetation indices as the reliable source of identifying irrigation fields. Studies show that multi-temporal remote sensing analysis are effective for irrigated area mapping rather than single date satellite imagery that enable general irrigation classification [7]. Higher spatial resolution satellite missions such as Sentinel 1 and Sentinel 2 have transformed field scale agricultural analysis, which is especially important in smallholder and fragmented agricultural areas. High resolution imagery provided by Sentinel 2 is the highest spatial resolution 10 m from the open-source sensors that is considerable improvement for agricultural landscapes [12].

The integration of remote sensing data with machine learning algorithms enhances the accuracy and robustness of irrigated area classification [13]. Random Forest has been implemented in satellite remote sensing-based land classification studies for classification of agricultural land uses and suitable to handle high dimensional data [4]. Support Vector Machine has shown strong performance for large scale irrigation mapping [14]. Recent studies in Ethiopia shows that application of machine learning is effective for accurately mapping irrigation areas in different watersheds [15]. Advanced machine learning approaches, particularly ensemble methods and deep learning are essential for crop mapping because they can capture by nonlinear relationships among predictor variables [16].

Therefore, the aim of this study is to identify irrigated areas in the Upper Blue Nile Basin during the December 2025–April 2026 irrigation season by comparatively evaluating Bidirectional Long Short-Term Memory algorithms, Random Forest, Support Vector Machine, and Gradient Tree Boosting models.

2. Materials and Methods

2.1. Study Area

The upper Blue Nile Basin (UBNB) is part of the Blue Nile Basin contained within the Ethiopian national boundary. The main river originates in Lake Tana and join the White Nile at Khartoum, Sudan. The basin covers a large part of the Ethiopian Highlands, from an altitude of less than 500 meter above mean sea level (m amsl) at the Sudanese border to more than 4200 m amsl in the Central and the Eastern escarpment of the Ethiopian Highlands [17]. This steep elevation gradient defines its rugged topography of escarpments, high plateaus, and deep gorges, which strongly influence hydrological responses and sediment transport processes [18]. The climate of the basin is governed by the seasonal migration of the Intertropical Convergence Zone (ITCZ) and associated monsoon systems. The main rainy season (Kiremt) extends from June to September and contributes roughly 70–80% of annual rainfall and streamflow, followed by the dry season (Bega) from October to January. A smaller rainy season (Belg) occurs from February to May. Annual rainfall varies greatly, from 800 mm in the drier lowlands to more than 2,200 mm in the highlands [19]. Temperature also varies with elevation, ranging from 15–25°C in highland zones to over 35°C in the lowlands during peak dry months [20] . The Hydrology of the Upper Blue Nile River basin exhibits large seasonal and interannual variations. The average annual discharge at the Sudanese outlet has been estimated 48-52 billion cubic meters with more than two-thirds of the discharge during the rainy(Kiremt) season [21]. Recently, the Blue Nile River Basin has been experiencing changes in the regimes of rainfall and temperatures influenced by climate change. Ensemble projections show increase in rainfall by 10-25% in some scenarios, while the variability of the distribution of the latter will raise uncertainties of water availability [22]. The rise in temperatures will increase evapotranspiration demand with implications for crop water demand and for the operation of the reservoirs [23].

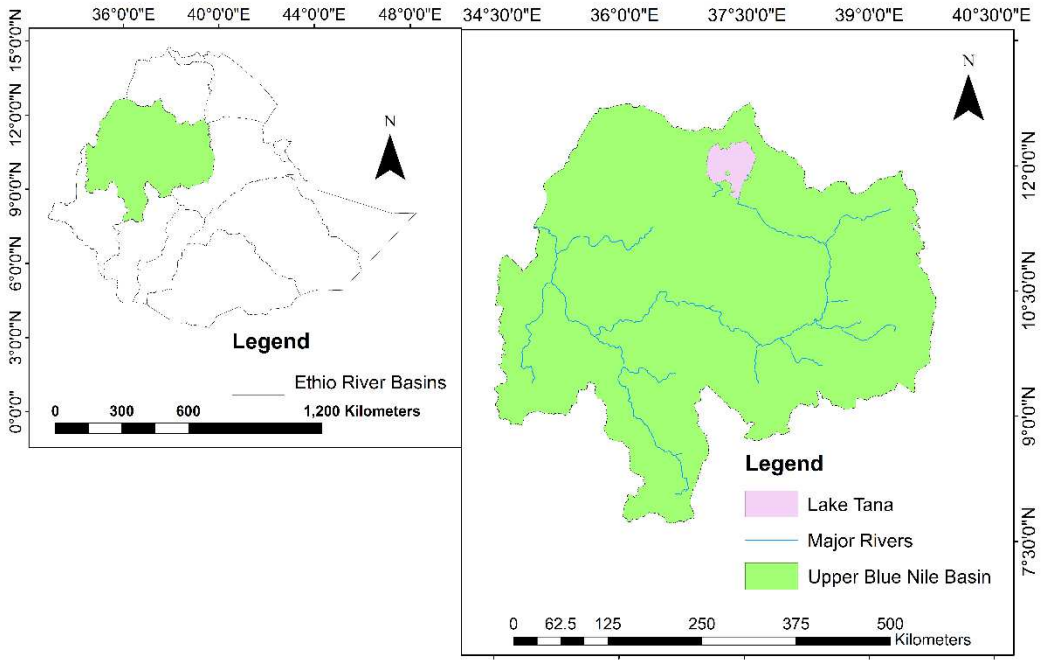


Figure 1: Location map of the Upper Blue Nile Basin study area

2.2. Data Sources

Various static and time series data were used for the machine learning classification models. Multi source geospatial datasets including topographic, remote sensing rainfall data, and remote sensing imagery products are integrated to map irrigated area in the upper Blue Nile basin.

2.2.1. Remote Sensing Rainfall Data

The accuracy of irrigation mapping during the irrigation season is affected by the amount of rainfall. Satellite derived rainfall estimates provide precipitation data at different temporal and spatial scales and play a key role for hydrological applications. CHIRPS provides high-resolution precipitation estimates (0.05° or about 5 km) daily, pentadal, and monthly time steps from 1981 to the current year. CHIRPS outperforms other satellite rainfall estimates indicating strong regional preference in Upper Blue Nile Basin [24].

2.2.2. Remote Sensing Imagery Products

Satellite imagery products provide spatially explicit information about land cover, vegetation, and moisture status. They are typically static or derived from repeated acquisitions and are essential for irrigated area mapping and suitability analysis. Sentinel-2 MSI has the capability to provide multi-spectral images with resolutions of 10 to 20 m since 2015. The main applications include irrigated area mapping, vegetation index computation, and vegetation growth and health status. Sentinel-1C-band Synthetic Aperture Radar Ground Range Detected (SAR GRD) which is calibrated and terrain corrected was obtained from Google Earth Engine. SAR provide dual polarization backscatter, Vertical Transmit Vertical Receive (VV) and Vertical Transmit Horizontal Receive (VH). VV is mainly sensitive to soil moisture condition and VH is sensitive to crop canopy characteristics. The combination used of the two polarizations, VV and VH, improves the detection of agricultural dynamics.

2.2.3. Topography

Topography data was obtained from the NASA Digital Elevation Model (NASADEM) through Google Earth Engine (GEE). Digital Elevation Model influence surface hydrology, agricultural suitability and soil moisture distribution [25]. Slope is a critical factor for irrigation feasibility with flat terrains are more suitable for irrigation development.

2.2.4. Reference Data

The reference dataset included 482 polygons of field in the basin agricultural areas. Polygon delineation was performed using visual interpretation of very high-resolution imagery from Google Earth using the field boundary, the vegetation condition during dry season, water-related structures and agriculture landscape pattern. Irrigated fields showed stable greenness in dry season and patterns of irrigation, and rainfed fields showed poor greenness during dry season and absence of irrigation structure. Predictor values were extracted from pixels within each polygon boundary on the analysis grid. Samples were partitioned by whole polygons and spatial blocks rather than individual pixels, ensuring that all observations from a given field were assigned exclusively to either the model-development or validation set.

2.2.5. Land Use Land cover

A 2024 land use land cover map of Upper Blue Nile was used as ancillary dataset to identify agricultural land. The purpose of the LULC map was to mask nonagricultural land cover such as

forests, shrubland, water bodies, and other classes before model training and prediction. The agricultural land derived from the LULC map was used to filter the predictor variables and reference samples before model development.

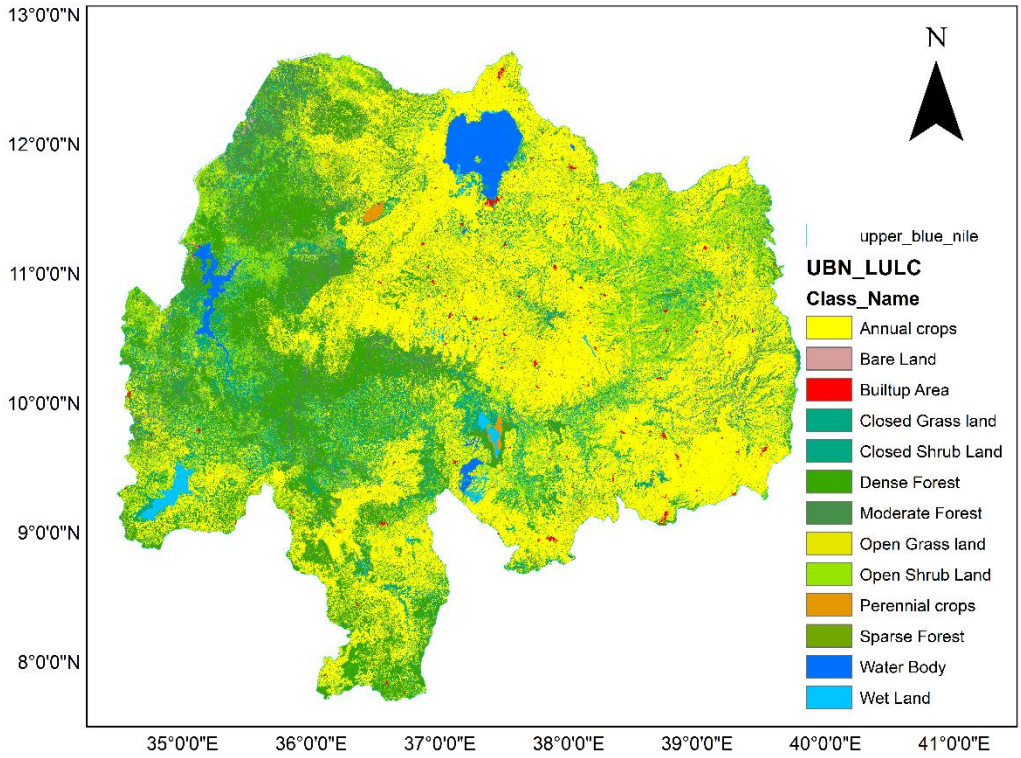


Figure 2: Land use/land cover map of the Upper Blue Nile Basin (2024) used to derive the agricultural mask applied in irrigation classification

2.3. Preprocessing of Remote Sensing and Ancillary Data

Sentinel-1C-band Synthetic Aperture Radar Ground Range Detected (SAR GRD) is suitable for all-whether agricultural monitoring because SAR observations are largely unaffected by cloud cover [26] . To reduce the skewness in the data distribution Radar backscatter were converted to decibel scale using logarithmic transformation. To improve the class separability between rainfed and irrigated the VH/VV ratio was computed. Sentinel 2 surface reflectance data were atmospherically corrected and radiometrically consistent from the Sentinel 2A and Sentinel 2B satellites [27]. Cloud masking was applied using the quality assessment (QA60) band. Based on bitwise filtering, the cloud and cirrus pixels were masked. After cloud masking, vegetation and moisture indices derived from multispectral satellite data were used to identify irrigated areas.

Irrigated croplands generally maintain higher vegetation vigor and sustained moisture conditions compared to rainfed systems, particularly during dry periods [28]. Therefore, integrating multiple spectral indices and their temporal dynamics improves the reliability of irrigation mapping.

The Normalized Difference Vegetation Index (NDVI) was derived from Sentinel-2 images to characterize vegetation greenness and crop development. NDVI was included because irrigated fields usually maintain higher vegetation activity than rain-fed fields during the dry season. The index was calculated from red and near-infrared reflectance values.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

However, NDVI can be influenced by soil background effects, particularly in sparsely vegetated areas. To address this limitation, the **Soil Adjusted Vegetation Index (SAVI)** is incorporated, which minimizes soil brightness effects and improves vegetation detection under varying soil conditions:

$$SAVI = \frac{NIR - RED}{(NIR + RED + L)} * (1 + L) \tag{2}$$

where L is a soil adjustment factor (commonly 0.5).

Normalized Difference Water Index (NDWI) was used as a measure of surface moisture conditions. Higher NDWI values are commonly found in irrigated fields due to the presence of water during the irrigation season.

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \tag{3}$$

The Normalized Difference Moisture Index (NDMI) was used to represent vegetation moisture status. As irrigated crops generally exhibit higher canopy moisture than rainfed crops, NDMI provides useful information for distinguishing irrigated areas.

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR} \tag{4}$$

Monthly and seasonal composites of NDVI, NDWI, NDMI, and SAVI were generated for the irrigation season. Summary statistics including mean, maximum, and standard deviation were calculated to capture temporal variations in vegetation and moisture conditions. These variables were subsequently used as predictor features in the machine learning models.

2.4. Machine Learning Classifications Models

Four machine learning classification algorithms were evaluated for irrigated area mapping in the Upper Blue Nile Basin. Random Forest (RF), Support Vector Machine (SVM), Gradient Tree Boosting (GTB), and Bidirectional Long Short-Term Memory (BiLSTM). These methods represent different learning approaches, including ensemble learning, margin-based classification, boosting, and deep learning. Random Forest is ensembles learning method that builds multiple decision trees for training, and it combines its prediction through majority voting for classification [29]. The Random Forest algorithm is widely used in remote sensing applications because it can handle high-dimensional data and nonlinear relationships among predictor variables [30], [4]. Support Vector Machine identifies an optimal hyperplane that increases the margin between classes in multidimensional feature space [31]. The method has been successfully applied in irrigation mapping and other remote sensing classification studies because of its ability to perform well with complex datasets and limited training samples [14]

Gradient Tree Boosting is an ensemble machine learning algorithm that sequentially incorporates multiple weak learners, typically low decision trees to create strong predictive models. This algorithm iteratively trains new trees to minimize errors by optimizing a differentiable loss function through gradient descent [32]. The step-by-step learning strategy of GTB enable to capture complex nonlinear relationships for effective classification [33]. Bidirectional Long Short-Term Memory is a deep learning architecture derived from Recurrent Neural Network and specially designed for temporal and sequential data modeling. It incorporates forward and backward parallel recurrent processing simultaneously to learn the model temporal dependencies from both past and future observations [34]. BiLSTM models are effective for time series analysis because they can collect seasonal and temporal variations [35].

The same predictor dataset was used to develop the four classifiers and included Sentinel-1 SAR variables, Sentinel-2 spectral indices, CHIRPS rainfall data, topographic variables, and seasonal metrics. The models were evaluated with nested spatial cross-validation with five outer folds. The

performance was compared by overall accuracy, Kappa coefficient, precision, recall, F1-score and AUC-ROC. This comparative framework allowed us to evaluate the relative performance of ensemble-learning, margin-based and deep-learning approaches for mapping of irrigated areas in the Upper Blue Nile Basin. For RF and GTB, predictor importance was obtained from the fitted tree-based models. For SVM, permutation importance was computed on the full reference dataset using the final tuned models as the mean decrease in accuracy over 10 permutations per predictor; values were not averaged across outer folds.

2.5. Evaluation Metrics and Model Selection

Model performance was assessed using multiple evaluation metrics to ensure a multi-dimensional comparison of the four classifier models. Overall accuracy is the proportion of correctly classified samples compared to the total number of samples. Precision, recall, F1-score were calculated for potential class imbalance, where F1-score served as balance between precision and recall avoiding misleading conclusion that overall accuracy may produce in imbalanced datasets. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates each model's discriminative ability across all possible classification thresholds. Cohen's Kappa coefficient was used to quantify the degree of agreement between model predictions and reference data. The performance evaluation metrics were evaluated with nested spatial cross validation with five outer spatial folds to ensure that the model performance reflects true spatial generalizability rather than overfitting spatially autocorrelated training samples [36]. The final model selection was based on combination of performance across all metrics. The evaluation metrics were calculated using Equations (5)-(9)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

$$k = \frac{P_o - P_e}{1 - P_e} \quad (9)$$

where TP, TN, FP, FN, P_o, and P_e denote true positives, true negatives, false positives, and false negatives, observed agreement, agreement expected by chance respectively.

2.6. Multi Model Ensemble Method

To enhance the individual model irrigated area map in the Upper Blue Nile Basin, a Multi Model Ensemble Mapping approach was developed by integrating predictions from the four machine learning models. The ensemble applied a conservative 3 of 4 voting rules to retain areas supported by at least three models and minimize the misclassification errors from the individual models. Its performance was evaluated using pooled out of fold predictions from the five spatial cross validation folds. Akbar et al., (2025) show combining output of multiple machine learning classifiers significantly improve individual models in irrigation area mapping compared to ground truth accuracy. Ensemble mapping ensures the irrigation map classification was made on those locations where strong cross model consensus exists.

3. Results and Discussions

3.1. Model Performance and Statistical Evaluation

The comparison of the four machine learning algorithms for December 2025 to April 2026 irrigation season showed clear differences in descriptive performance metrics. GTB scored the highest average performance of 0.96, which indicates it correctly classified 96% in the evaluation set. Followed by Random Forest within an accuracy of 0.95. BiLSTM attaining an accuracy of 0.95 and support vector machine scored 0.91. The kappa coefficient adjusts the possibility of agreement occurring by chance and it is robust indicator of model reliability. The kappa coefficient result shows notable drop from the top GTB and RF models. The AUC-ROC result indicates highly uniform for all models with RF and GTB scored 0.99, respectively. BiLSTM scored AUC-ROC of 0.98, and SVM recorded 0.97. AUC-ROC values near to 1 indicate strong ability to discriminate between negative and positive classes across all classification threshold for all models. The F1-score indicates the harmonic mean of precision and recall that balance false positives and false negatives. The result confirmed the superiority of GTB and RF. This indicates RF and GTB

maintained strong balance between the false negative and false positive to minimize the classification errors. The performance metrics of the four classifiers are summarized in Figure 3.

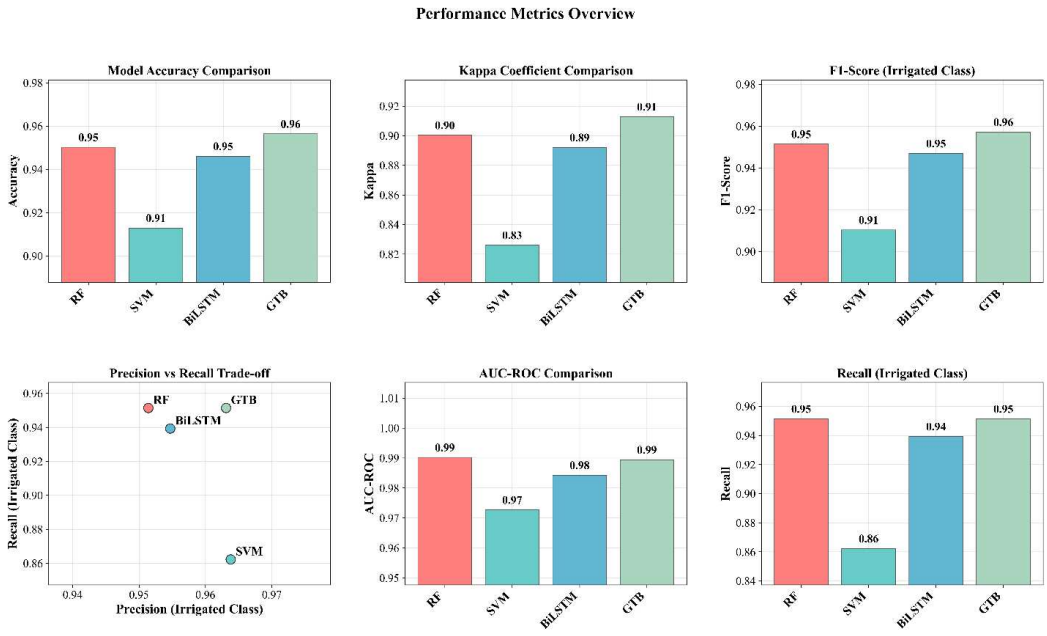


Figure 3: Performance metrics of the evaluated machine learning models

The GTB consistently scores high performance across all performance evaluation metrics, and it is the most reliable general-purpose classifier in this study. This model's precision score shows as key finding, suggests that its learned decision boundary generated positives predictions when the input features provided convincing evidence. GTB scored a recall of 0.95, indicating that approximately 5% of reference irrigated samples were classified as non-irrigated. This precision-recall tradeoff is a fundamental challenge in classification. From an architectural perspective, its powerful performance is consistent with its ability to capture complex nonlinear relationships through sequentially constructed decision trees.

Random Forest model ranked second and showed powerful performance across all evaluation metrics. The result indicates a balanced ability to decrease false positive classifications while correctly detecting irrigated fields. Although RF performed slightly below GTB in accuracy, Kappa, and F1-score, its high AUC-ROC shows strong discrimination between irrigated and rainfed fields across the classification thresholds. Lu et al. (2020) achieved an over accuracy of 93.3% and a Kappa coefficient of 0.916 using RF and multispectral imagery, while Magidi et al. (2021) reported an overall accuracy of 88% using Landsat and Sentinel data.

BiLSTM model detects most irrigated fields correctly but omitted approximately 6.07% of the actual irrigated samples. Although performing slightly below RF and GTB, BiLSTM shows strong classification performance. Previous studies show that BiLSTM based models can effectively learn temporal patterns from multispectral remote sensing data. Shi et al. (2026) reported that an LSTM based framework integrating Sentinel-1 and Sentinel-2 observations improved crop classification performance with an overall accuracy of 93.61% and a Kappa coefficient of 91.66%. This finding supports the broader usefulness of LSTM based architecture for detecting temporal agricultural signals.

SVM model demonstrates comparatively lower performance across the evaluation metrics and is ranked fourth among the four models. SVM produced comparatively few false positive irrigated classifications but omitted approximately 14% of the actual irrigation samples. The SVM performance result remains consistent with previous evidence that SVM can effectively classify irrigated and non-irrigated fields using multitemporal remote sensing data. Gao et al. (2018) obtained an overall accuracy of 81.08% using SVM sentinel-1 time series data. Another study using both sentinel-1 and sentinel-2 data obtained an accuracy of 88.4% for classifying irrigation. In this study SVM showed strong classification ability, but its low recall shows that it failed to detect more irrigated fields than other models.

Overall, the performance evaluation shows that model selection should consider multiple metrics rather than overall accuracy alone. The difference in recall, precision, and discriminatory ability shows that each classifier captures distinct aspect of irrigated field characteristics. These harmonious classification behaviors support integrating the four models output through multi model ensemble mapping to reduce errors across individual model's classifiers. A more detailed comparison of model performance is presented in Figure 4.

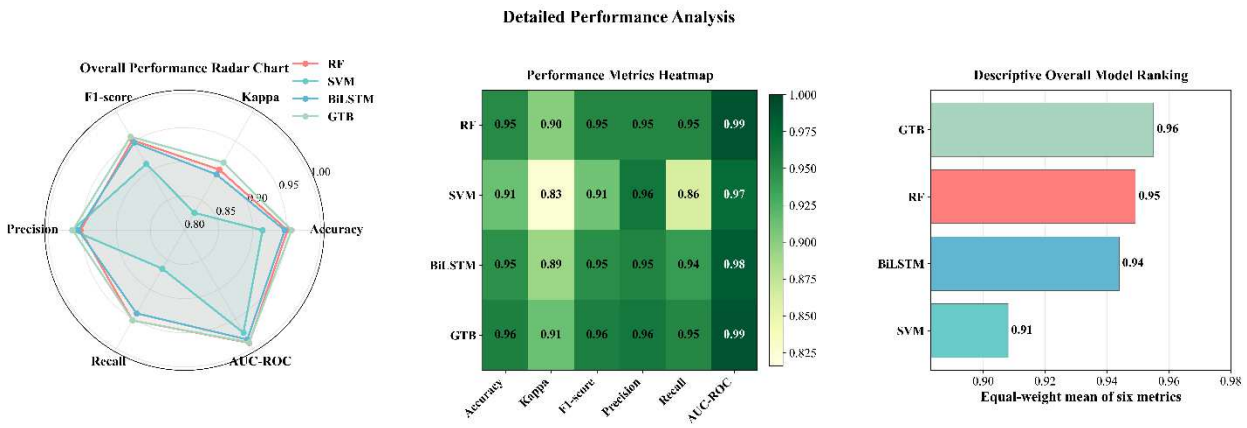


Figure 4: Detail Performance analysis

The radar chart and heatmap show that GTB, RF, and BiLSTM achieved balanced performance across the six-evaluation metrics, while SVM lower recall produced the clearest performance imbalance. The descriptive mean summarizes the overall pattern, but it should be treated as a comparative indicator rather than an independent validation metric.

3.2. Spatial Cross Validation

Model performance was further evaluated using nested spatial cross-validation with five outer spatial folds. Four spatial folds were used in each iteration for model development, while the remaining fold was kept out for independent evaluation. Hyperparameter tuning and model selection were conducted within the outer training data through the inner cross validation loop. Data imputation, scaling, and other fitted preprocessing operations were estimated from the training folds and then applied to the held-out fold. This prevents the outer test folds from affecting model selection and decreases the optimistic performance estimates associated with using the same observation for evaluation and tuning [36]. Spatial Random cross validation may overestimate predictive performance when geographically adjacent training and test samples are spatially dependent [41].

The four-classifiers showed geographically variable performance across the five outer folds. All four classifier models performed particularly well in Fold 4. However, performance was lower in Fold 1 for several models. SVM showed lower scores than the other models in this fold. SVM performance improved in Folds 3 and 4 but declined again in Fold 5, that indicates greater geographic variability than the other classifier models. BiLSTM had a lower mean kappa

coefficient than GTB and RF but showed the smallest fold-to-fold kappa variation, with standard deviation of 0.05. In contrast, SVM had the largest fold to fold Kappa variation with standard deviation of 0.12. GTB scored the highest mean performance with moderate variation, indicating comparatively strong balance between average accuracy and spatial stability. The difference among spatial folds indicates that the classification of irrigated and rainfed fields was not uniform across the study area. Such variation could potentially be associated with geographic differences in crop phenology or irrigation practice. The near perfect scores in Fold 4 show that irrigated and rainfed fields were highly separable within that held out region. The nested spatial cross validation results support GTB as the strongest average classifier, while the comparatively stable performance of BiLSTM and greater geographic sensitivity of SVM. The fold-level performance obtained from nested spatial cross-validation is presented in Figure 5.

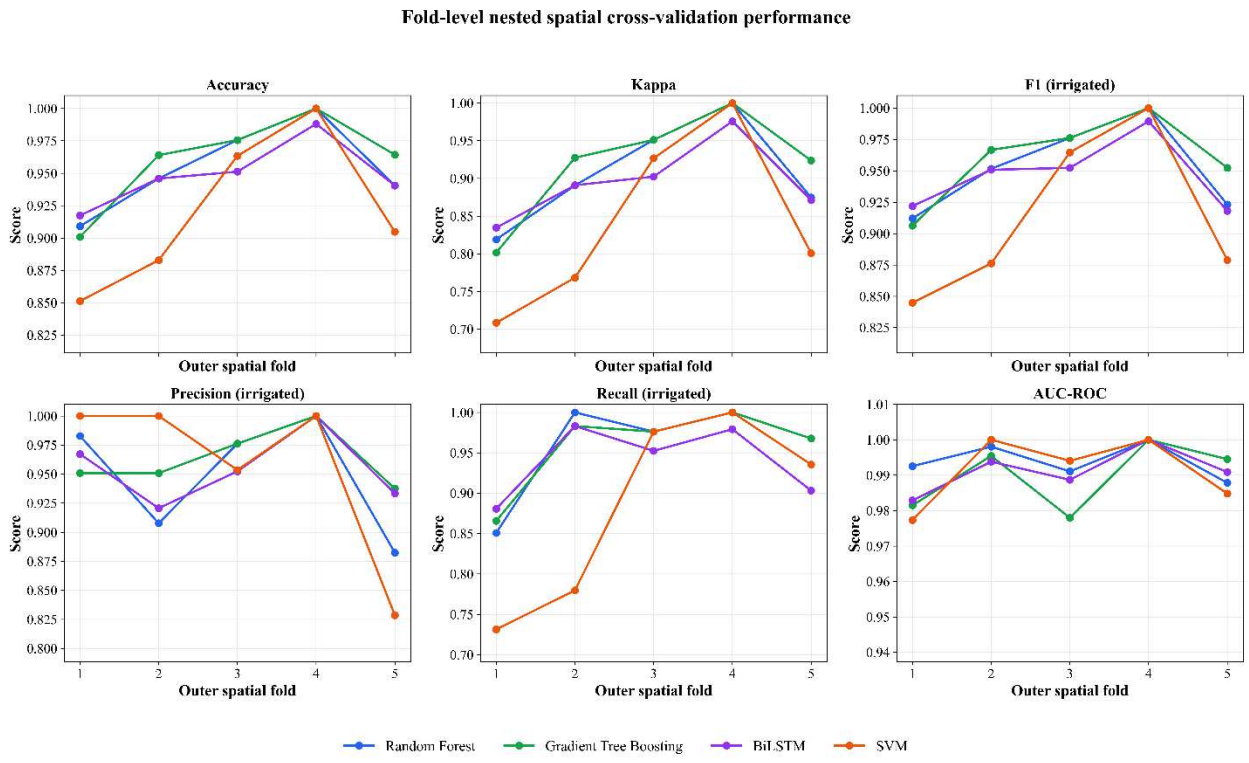


Figure 5: Fold-level performance

Figure 6 summarizes the mean performance and fold-to-fold variability of the evaluated classifiers.

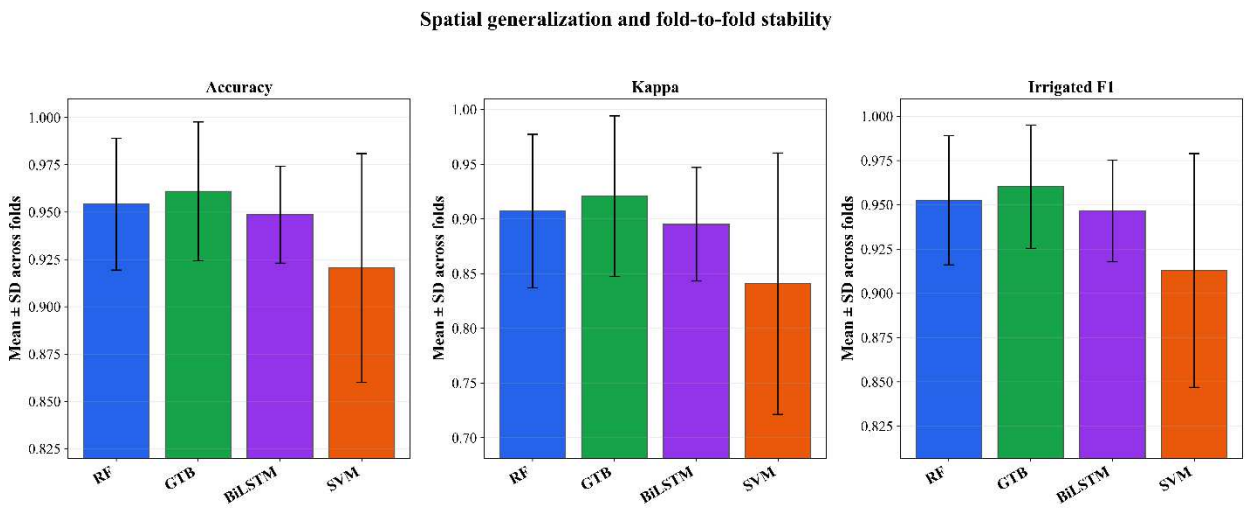


Figure 6: Mean performance and fold to fold standard deviation.

3.3. Statistical Comparison

The pairwise exact McNemar tests indicate that GTB and RF had significantly lower paired classification error than SVM after Holm correction. BiLSTM scored more correct predictions than SVM among the discordant cases. This discrepancy was not significant after adjustment. No statistically significant differences were detected within GTB, RF and BiLSTM, indicating that these models are under statistically comparable leading groups. The pairwise exact McNemar test was used to compare the classification errors of the evaluated models based on paired predictions from the nested spatial cross-validation.

Table 1: Pairwise exact McNemar tests.

Model A	Model B	A+/B-	A-/B+	Exact p-value	Holm p-value	α = 0.05
GTB	SVM	29	8	0.0007	0.0045	Significant
RF	SVM	25	7	0.0021	0.0105	Significant at 5%
BiLSTM	SVM	27	11	0.0138	0.0554	Non-significant
GTB	BiLSTM	9	4	0.2668	0.8005	Non-significant
RF	GTB	4	7	0.5488	1.0000	Non-significant
RF	BiLSTM	9	7	0.8036	1.0000	Non-significant

3.4. Feature Importance

Feature importance score was extracted for SVM, GTB and RF to understand the relative contribution of each feature variable in the classification process. This analysis was used to identify relative contribution of predictor variables. Because the feature importances were obtained from model specific approaches. GTB was dominated by Sentinel-2 NDMI feature followed by February EVI and January NDVI. Random Forest score had a much more evenly distributed feature importance. The top ranked features are February NDMI, February SAVI, February NDVI, and March NDMI. The near equal importance of this February indices shows that Random Forest uses complementary information from multiple spectral indices rather than depending on a single dominant feature. The SVM feature importance depends on with selected Sentinel-1, Sentinel-2, seasonality variable rainfall and topographic variable. This suggests the model leverages on environmental conditioning variables. For SVM, predictor ranking was obtained using permutation feature importance applied to the final tuned model on the reference dataset. From the spectral features SVM uses a broader spectral signal from February to April rather than on concentrating on single temporal window. Figure 7 presents the model-specific predictor rankings.

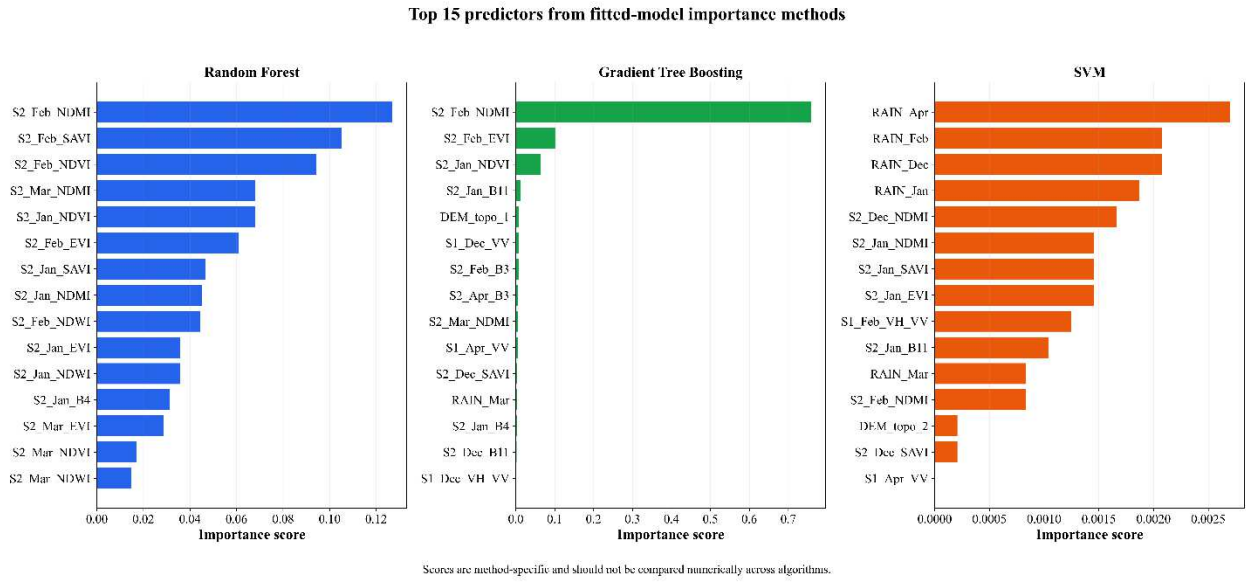


Figure 7: Model-specific predictor importance. RF and GTB use tree-based importance; SVM uses permutation sensitivity

Sentinel-2 contributed approximately 97% of GTB. The other variables Sentinel-1, DEM, and Rainfall made small contributions. This indicates GTB based classification decisions depend on a limited set of optical indicators representing vegetation condition and moisture. Sentinel-2 is also dominant in RF contributing approximately 94% of the grouped importance. Its individual importance was distributed across different variables particularly NDMI, SAVI, NDVI and EVI. Therefore, Random Forest did not depend on one predictor. SVM individual feature importance shows rainfall variables contribute highly rather than the Sentinels imagery. The monthly acquisition result shows February was the most important month for GTB and RF. For GTB and RF approximately 87% and 46% respectively comes from its grouped importance of February dataset. The aggregated importance of predictor variables by data source and acquisition month is presented in Figure 8.

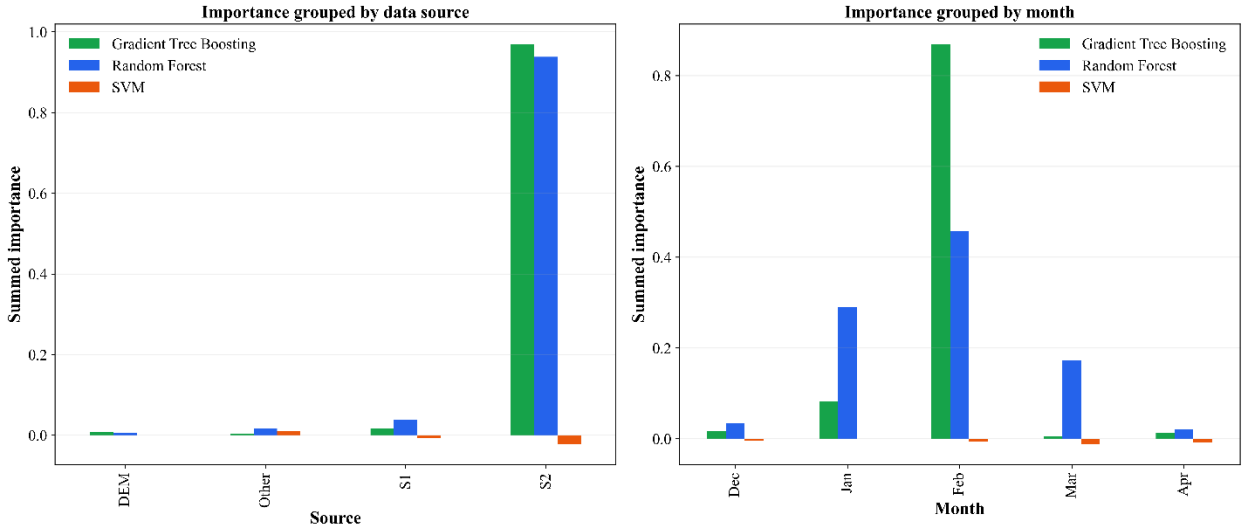


Figure 8: Importance aggregated by source and acquisition month

3.5. Irrigated Area Map of Upper Blue Nile basin

All models consistently identified irrigated agriculture as spatially limited and unevenly distributed across the basin. The irrigated areas were scattered concentrated primarily in the central, northern, southwestern, and southern parts of the Upper Blue Nile. The rainfed agriculture dominated in the northern and central highlands of the basin, indicating high dependence on seasonal rainfall for

agricultural production in the basin. The spatial distribution of irrigated and rainfed agriculture predicted by the four machine learning models is presented in Figure 9.

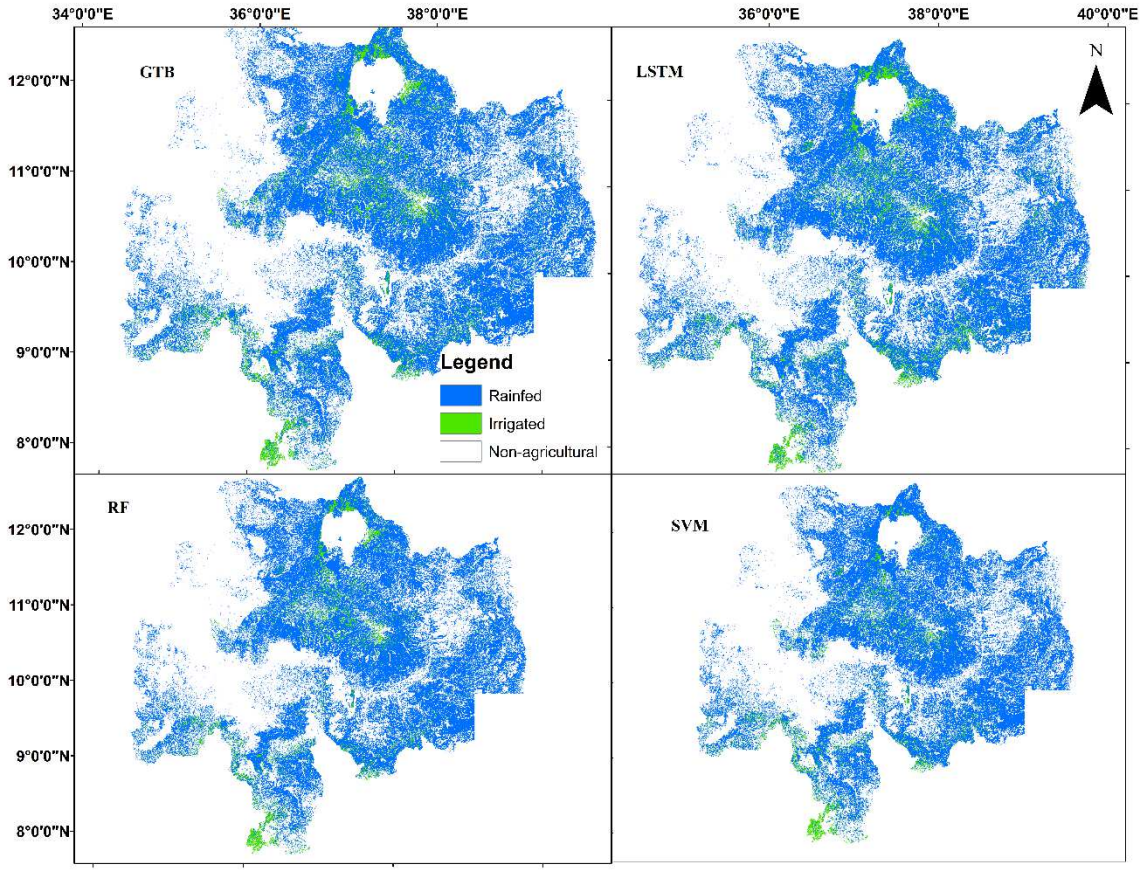


Figure 9: Map of Irrigated and Rainfed agricultural lands

All four classifiers showed the same general spatial pattern, but ranged considerably in estimated irrigated extent. The highest estimation was given by BiLSTM (871,049 ha) followed by GTB (793,809 ha), RF (670,626 ha) and SVM (614,261 ha). The difference between the highest and lowest individual-model estimates was 256,788 ha, indicating that classifier choice can considerably impact basin-scale area estimations, even when traditional accuracy metrics are equal. The GTB model produced extensive and concentrated irrigated clusters, particularly in the central parts of the basin. RF model produced broadly similar patterns but classified some irrigated areas more conservatively. BiLSTM detects many similar clusters, while differences occurred in smaller and spatially scattered areas. SVM has shown the most limited and fragmented irrigated

area distribution, which is consistent with its comparatively low recall and indicates that it ignored more irrigated fields than other models. The spatial differences between the maps reflect variation in how the classifiers used the predictor variables. GTB depended strongly on February Sentinel-2 moisture and vegetation indices, while RF distributed predictors over various spectral indices and months. SVM assigned comparatively greater importance to rainfall variables. Previous studies in the Upper Blue Nile watershed similarly found that monthly radar and vegetation index information improved irrigation classification [42].

GTB obtained the highest overall performance among the evaluated models. However, the comprehensive spatial agreement with GTB, RF, and BiLSTM together with the lack of statistically significant differences in their paired error rates, indicates that the important irrigation patterns detected by these classifiers are strong. The significance of the irrigation map results show irrigation development has limited extent relative to the basin's large agricultural area. Overall, the maps demonstrate consistent irrigation signals and model reliant spatial ambiguity. Areas classified as irrigated by various models have stronger cross model support, while areas predicted only by one model represent higher classification uncertainty. The observed agreement and disagreement indicate the importance of multi model ensemble mapping, which can maintain locations supported by multiple models while decreasing errors related to individual model predictions.

3.6. Model Ensemble Mapping

The application of Multi Model Ensemble Mapping produced a high agreement irrigated area map by combining the predictions of GTB, RF, BiLSTM, and SVM. The pixel level agreement map showed considerable spatial variability in agreement among the classifiers across the basin. The ensemble approach addresses the individual single model output limitation. From the model performance metrics result, there is variation tradeoff between precision and recall in individual models. Quantitative validation using pooled out-of-fold predictions from the five spatial cross-validation folds showed that the 3-of-4 ensemble achieved an overall accuracy of 95.64%, precision of 97.48%, recall of 93.93%, an F1-score of 0.9567, Cohen's Kappa of 0.9129, and an AUC-ROC of 0.9904. The ensemble achieved the highest precision among the evaluated approaches, while its overall accuracy equaled that of GTB. Therefore, it is interpreted as a conservative, high-precision consensus model rather than as universally superior to every individual classifier.

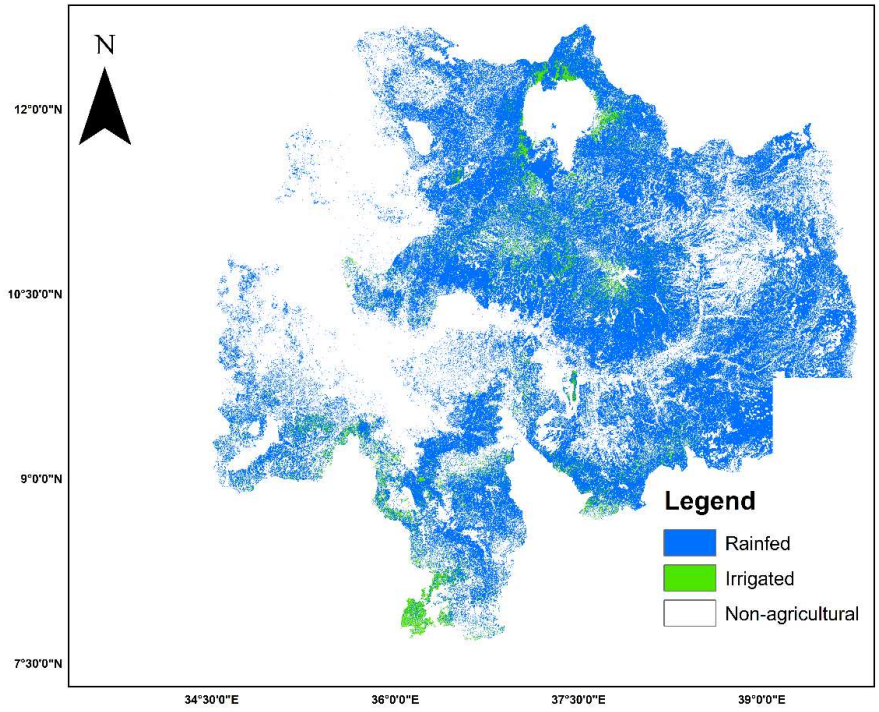


Figure 10: Model ensemble irrigated and rainfed area map in the Upper Blue Nile

The consensus irrigation map using conservative model ensemble found 588,545 ha (5,885.45 km2) for irrigation and around 8,290,763 ha as rain-fed. The minimum and maximum individual irrigation model output was from SVM which had 614,261 ha and BiLSTM which showed 871,049 ha, thus giving a difference of around 256,788 ha. This difference is a result of variability among the model outputs. The consensus irrigation area is smaller than all the individual model irrigation area due to the 3-of-4 majority voting rule. This conservative approach, together with the high overall accuracy (95.64%) and irrigated-class precision (97.48%), supports the reliability of the consensus map.

Gebeyehu, (2018) mapped 310,707 ha of irrigated land in the Upper Blue Nile Basin for 2016/2017 using Sentinel-2 imagery and object-based image analysis. This is considerably lower than the 588,545 ha estimated in the present study. Yimere & Assefa, (2022) reported 738,184 ha, but their estimate covered both existing irrigation and future irrigation potential. The differences in these estimates could be related to the methods and periods considered in each study, as well as

the difficulty of mapping small and scattered irrigated fields. The high classification accuracy and agreement among at least three models suggest that the estimated irrigated area is reliable.

Table 2: Estimated irrigated and rainfed areas classified by each model

Model	Irrigated (ha)	Rainfed (ha)
Random Forest	670,627	8,208,682
GTB	793,810	8,085,499
BiLSTM	871,049	8,008,259
SVM	614,261	8,265,047
Ensemble	588,545	8,290,763

4. Conclusion

Irrigation is important in sustaining agricultural productivity by ensuring food security, especially where water availability fluctuates due to climatic variability in Upper Blue Nile Basin. However, up to date information on irrigation mapping is unavailable which pose challenge in water resource management and allocation in the basin. The aim of this study is to evaluate and compare four machine learning models for irrigated area mapping in Upper Blue Nile Basin using multi source remote sensing datasets. The combination of these data sets allowed a comprehensive characterization of bio-physical conditions that differentiate irrigated land from rainfed agriculture across the basin. Among the four machine learning classifiers models, the GTB model demonstrated the highest performance across all performance evaluation metrics. The result indicates that GTB was strongest individual classifier for irrigated mapping in this study. However, the McNemar tests found no statistically significant differences in paired error rates among GTB, RF, and BiLSTM, implying that these three models formed statistically similar leading group.

The feature-importance analysis results show that Sentinel-2 spectral indices were the most important influential predictors for GTB and RF. February was the dominant acquisition month, highlighting this period as an optimal phenological window for classification of irrigation from rainfed in Upper Blue Nile Basin. The GTB model depends strongly on February moisture and

vegetation indicators, particularly NDMI, whereas RF model feature importance is more balanced and more evenly across multiple spectral indices and months. SVM uniquely depends on rainfall and topographic variables beside spectral features. These distinctive features are important to profile emphasize the value of combining models with different learning strategies in an ensemble approach. The Multi-Model Ensemble Mapping approach had a requirement of consensus from at least three out of the four classifiers before the pixel could be classified as irrigated. It produced relatively conservative consensus value of 588,545 ha.

Mapping irrigated land in the Upper Blue Nile Basin is challenging because the majority of the irrigated fields are small and located in different parts of the basin. And the lack of trustworthy records of irrigation makes it even tougher to tell how much land is actually irrigated. In this study we show that combining machine learning algorithms with satellite data can be used to identify these fields and increase our understanding of where irrigation is occurring. The irrigation map generated in this study can be used to track irrigation expansion and for providing information for estimating irrigation water demand. It can also help water resource managers pinpoint potential problem areas in planning irrigation development and water resource allocation.

Author contributions

A.A.H. contributed to conceptualization, methodology, data collection and processing, model development, data analysis, validation, visualization, and writing the original manuscript. Z.A.D., S.T.G., and H.B.L. supervised the research and contributed to reviewing and editing the manuscript. T.G.H. contributed to data collection and preparation. All authors reviewed and approved the final manuscript.

Funding

This research has been produced with the financial assistance of the European Union (Contract number: AFS2023/444-387). The contents of this document are the sole responsibility of authors and can under no circumstances be regarded as reflecting the position of the European Union.

Data availability

The Sentinel-1 SAR, Sentinel-2 surface reflectance, CHIRPS rainfall, and NASADEM elevation datasets used in this study were accessed through Google Earth Engine and are publicly available through the following links:

https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD

https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED

https://developers.google.com/earth-engine/datasets/catalog/UCSB-CHG_CHIRPS_PENTAD

https://developers.google.com/earth-engine/datasets/catalog/NASA_NASADEM_HGT_001

The processed datasets and irrigation maps generated during this study are available from the corresponding author upon reasonable request.

Declarations

Ethical approval

Not applicable. This study did not involve human participants or animals.

Consent to participate

Not applicable. This study did not involve human participants.

Consent to publish

Not applicable. This manuscript does not contain identifiable personal information requiring consent for publication.

Clinical trial registration

Not applicable. This study did not involve a clinical trial.

Competing interests

The authors declare no competing interests.

References

[1] E. A. Abioye *et al.*, "Precision Irrigation Management Using Machine Learning and Digital Farming Solutions," *AgriEngineering*, vol. 4, no. 1, pp. 70–103, Feb. 2022, doi: 10.3390/agriengineering4010006.

[2] P. Doungmanee, "The nexus of agricultural water use and economic development level," *Kasetsart Journal of Social Sciences*, vol. 37, no. 1, pp. 38–45, Jan. 2016, doi: 10.1016/j.kjss.2016.01.008.

[3] C. Massari *et al.*, "A Review of Irrigation Information Retrievals from Space and Their Utility for Users," *Remote Sens. (Basel).*, vol. 13, no. 20, p. 4112, Oct. 2021, doi: 10.3390/rs13204112.

[4] D. Ketchum, K. Jencso, M. P. Maneta, F. Melton, M. O. Jones, and J. Huntington, "IrrMapper: A Machine Learning Approach for High Resolution Mapping of Irrigated Agriculture Across the Western U.S.," *Remote Sens. (Basel).*, vol. 12, no. 14, p. 2328, Jul. 2020, doi: 10.3390/rs12142328.

[5] D. G. Eshete, B. G. Sinshaw, and K. G. Legese, "Critical review on improving irrigation water use efficiency: Advances, challenges, and opportunities in the Ethiopia context," *Water-Energy Nexus*, vol. 3, pp. 143–154, 2020, doi: 10.1016/j.wen.2020.09.001.

[6] T. Conlon, C. Small, and V. Modi, "A Multiscale Spatiotemporal Approach for Smallholder Irrigation Detection," *Frontiers in Remote Sensing*, vol. 3, Apr. 2022, doi: 10.3389/frsen.2022.871942.

[7] M. Ozdogan, Y. Yang, G. Allez, and C. Cervantes, "Remote Sensing of Irrigated Agriculture: Opportunities and Challenges," *Remote Sens. (Basel).*, vol. 2, no. 9, pp. 2274–2304, Sep. 2010, doi: 10.3390/rs2092274.

[8] A. A. Mohammedshum, B. H. P. Maathuis, C. M. Mannaerts, and D. Teka, "Mapping Small-Scale Irrigation Areas Using Expert Decision Rules and the Random Forest Classifier in Northern Ethiopia," *Remote Sens. (Basel).*, vol. 15, no. 24, p. 5647, Dec. 2023, doi: 10.3390/rs15245647.

[9] Y. T. Dile *et al.*, "Advances in water resources research in the Upper Blue Nile basmin and the way forward: A review," *J. Hydrol. (Amst).*, vol. 560, pp. 407–423, May 2018, doi: 10.1016/j.jhydrol.2018.03.042.

[10] B. G. Tikuye, M. Rusnak, B. R. Manjunatha, and J. Jose, "Land Use and Land Cover Change Detection Using the Random Forest Approach: The Case of The Upper Blue Nile River Basin, Ethiopia," *Global Challenges*, vol. 7, no. 10, Oct. 2023, doi: 10.1002/gch2.202300155.

[11] A. Yimere and E. Assefa, “ASSESSING AND MAPPING IRRIGATION POTENTIAL IN THE ABBAY RIVER BASIN, ETHIOPIA,” *Russ. J. Agric. Socioecon. Sci.*, vol. 114, no. 6, pp. 97–109, Jun. 2021, doi: 10.18551/rjoas.2021-06.11.

[12] M. F. A. Vogels, S. M. de Jong, G. Sterk, H. Douma, and E. A. Addink, “Spatio-Temporal Patterns of Smallholder Irrigated Agriculture in the Horn of Africa Using GEOBIA and Sentinel-2 Imagery,” *Remote Sens. (Basel).*, vol. 11, no. 2, p. 143, Jan. 2019, doi: 10.3390/rs11020143.

[13] J. M. Deines, A. D. Kendall, M. A. Crowley, J. Rapp, J. A. Cardille, and D. W. Hyndman, “Mapping three decades of annual irrigation across the US High Plains Aquifer using Landsat and Google Earth Engine,” *Remote Sens. Environ.*, vol. 233, p. 111400, Nov. 2019, doi: 10.1016/j.rse.2019.111400.

[14] Q. Gao, M. Zribi, M. J. Escorihuela, N. Baghdadi, and P. Q. Segui, “Irrigation Mapping Using Sentinel-1 Time Series at Field Scale,” *Remote Sens. (Basel).*, vol. 10, no. 9, p. 1495, Sep. 2018, doi: 10.3390/rs10091495.

[15] A. K. Yimer, A. T. Haile, S. D. Hatiye, S. Ragettli, and M. T. Taye, “Comparative evaluation of the accuracy of mapping irrigated areas using sentinel 1 images in the Bilate and Gumara watersheds, Ethiopia,” *Cogent Eng.*, vol. 11, no. 1, Dec. 2024, doi: 10.1080/23311916.2024.2357728.

[16] A. Joshi, B. Pradhan, S. Gite, and S. Chakraborty, “Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review,” *Remote Sens. (Basel).*, vol. 15, no. 8, p. 2014, Apr. 2023, doi: 10.3390/rs15082014.

[17] T. Lemann, V. Roth, G. Zeleke, A. Subhatu, T. Kassawmar, and H. Hurni, “Spatial and Temporal Variability in Hydrological Responses of the Upper Blue Nile basin, Ethiopia,” *Water (Basel).*, vol. 11, no. 1, p. 21, Dec. 2018, doi: 10.3390/w11010021.

[18] E. Garzanti, S. Andò, G. Vezzoli, A. Ali Abdel Megid, and A. El Kammar, “Petrology of Nile River sands (Ethiopia and Sudan): Sediment budgets and erosion patterns,” *Earth Planet. Sci. Lett.*, vol. 252, no. 3–4, pp. 327–341, Dec. 2006, doi: 10.1016/j.epsl.2006.10.001.

[19] Z. T. Segele and P. J. Lamb, “Characterization and variability of Kiremt rainy season over Ethiopia,” *Meteorology and Atmospheric Physics*, vol. 89, no. 1–4, pp. 153–180, Jun. 2005, doi: 10.1007/s00703-005-0127-x.

[20] P.-E. Mellander, S. G. Gebrehiwot, A. I. Gärdenäs, W. Bewket, and K. Bishop, “Summer Rains and Dry Seasons in the Upper Blue Nile Basin: The Predictability of Half a Century of Past and Future Spatiotemporal Patterns,” *PLoS One*, vol. 8, no. 7, p. e68461, Jul. 2013, doi: 10.1371/journal.pone.0068461.

[21] G. G. Chakilu, S. Sándor, T. Zoltán, and K. Phinzi, "Climate change and the response of streamflow of watersheds under the high emission scenario in Lake Tana sub-basin, upper Blue Nile basin, Ethiopia," *J. Hydrol. Reg. Stud.*, vol. 42, p. 101175, Aug. 2022, doi: 10.1016/j.ejrh.2022.101175.

[22] U. Kim and J. J. Kaluarachchi, "Climate Change Impacts on Water Resources in the Upper Blue Nile River Basin, Ethiopia [1]," *JAWRA Journal of the American Water Resources Association*, vol. 45, no. 6, pp. 1361–1378, Dec. 2009, doi: 10.1111/j.1752-1688.2009.00369.x.

[23] S. G. Setegn, D. Rayner, A. M. Melesse, B. Dargahi, and R. Srinivasan, "Impact of climate change on the hydroclimatology of Lake Tana Basin, Ethiopia," *Water Resour. Res.*, vol. 47, no. 4, Apr. 2011, doi: 10.1029/2010WR009248.

[24] D. A. Malede, T. A. Agumassie, J. R. Kosgei, T. G. Andualem, and I. Diallo, "Recent Approaches to Climate Change Impacts on Hydrological Extremes in the Upper Blue Nile Basin, Ethiopia," *Earth Systems and Environment*, vol. 6, no. 3, pp. 669–679, Sep. 2022, doi: 10.1007/s41748-021-00287-6.

[25] Y. Deng, J. P. Wilson, and J. C. Gallant, "Terrain Analysis," in *The Handbook of Geographic Information Science*, Wiley, 2007, pp. 417–435. doi: 10.1002/9780470690819.ch23.

[26] E. Beriaux, A. Jago, C. Lucau-Danila, V. Planchon, and P. Defourny, "Sentinel-1 Time Series for Crop Identification in the Framework of the Future CAP Monitoring," *Remote Sens. (Basel).*, vol. 13, no. 14, p. 2785, Jul. 2021, doi: 10.3390/rs13142785.

[27] M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012, doi: 10.1016/j.rse.2011.11.026.

[28] M. Mohammed, B. Birhanu, and F. Abegaz, "Multi-spectral remote sensing for current irrigated area mapping of the Rift Valley Lakes Basin in Ethiopia," *AQUA — Water Infrastructure, Ecosystems and Society*, vol. 72, no. 12, pp. 2396–2407, Dec. 2023, doi: 10.2166/aqua.2023.413.

[29] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[30] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[32] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *The Annals of Statistics*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[33] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[34] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[35] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/j.neunet.2005.06.042.

[36 P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data,” *Ecol. Modell.*, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[37] M. U. Akbar *et al.*, “Multi-model ensemble mapping of irrigated areas using remote sensing, machine learning, and ground truth data,” *Agric. Water Manag.*, vol. 312, p. 109416, May 2025, doi: 10.1016/j.agwat.2025.109416.

[38] Y. Lu, W. Song, Z. Su, J. Lü, Y. Liu, and M. Li, “MAPPING IRRIGATED AREAS USING RANDOM FOREST BASED ON GF-1 MULTI-SPECTRAL DATA,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2020, pp. 697–702, Aug. 2020, doi: 10.5194/isprs-archives-XLIII-B2-2020-697-2020.

[39] J. Magidi, L. Nhamo, S. Mpandeli, and T. Mabhaudhi, “Application of the Random Forest Classifier to Map Irrigated Areas Using Google Earth Engine,” *Remote Sens. (Basel).*, vol. 13, no. 5, p. 876, Feb. 2021, doi: 10.3390/rs13050876.

[40] S. Shi, Q. Liu, and Z. Yan, “Crop Type Mapping in an Irrigation District Using Multi-Source Remote Sensing and LSTM-Based Time Series Analysis,” *Agriculture*, vol. 16, no. 9, p. 920, Apr. 2026, doi: 10.3390/agriculture16090920.

[41] D. R. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, Aug. 2017, doi: 10.1111/ecog.02881.

[42] Y. G. Mekonnen, T. Alamirew, A. D. Chukalla, A. T. Hunegnaw, and D. A. Malede, “Comparison of Google Earth Engine Machine Learning Algorithms for Mapping Smallholder Irrigated Areas in a Mountainous Watershed, Upper Blue Nile Basin,

Ethiopia," *Journal of the Indian Society of Remote Sensing*, vol. 52, no. 3, pp. 569–582, Mar. 2024, doi: 10.1007/s12524-024-01846-w.

[43] Y. Gebeyehu, "Irrigation Potential Mapping using GIS-based Parametric Evaluation Approach in Upper Blue Nile River Basin, Ethiopia ," 2018.

[44] A. Yimere and E. Assefa, "Current and Future Irrigation Water Requirement and Potential in the Abbay River Basin, Ethiopia," *Air, Soil and Water Research*, vol. 15, Jan. 2022, doi: 10.1177/11786221221097929.