iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VIII Month of publication: August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84643

XGBoost-LSTM Hybrid Model for Predicting Vegetation Cover in Kenya's North Rift Region: Leveraging LULC and Environmental Variables

Eve Khatundi Wamasebu[1], Betty Mayeku[2], Jonathan Mutonyi[3], Nelly Masayi[4]
[1, 2]*Department of Computer Science, School of Computing and Informatics (SCAI), Kibabii University*
[3]*Department of Agriculture and Veterinary Sciences, Faculty of Science, Kibabii University*
[4]*Department of Geography, Faculty of Arts and Social Sciences (FASS), Kibabii University*

Abstract: In Kenya's North Rift region, rapid Land Use Land Cover (LULC) transformations have resulted from agricultural activities, population growth, and infrastructure development. While these changes support economic growth and food security, they impact terrestrial vegetation health and ecosystem stability. Accurate mapping and prediction of these changes is crucial for sustainable land use strategies.

A key vegetation health metric is the Normalized Difference Vegetation Index (NDVI). However, predicting NDVI from LULC is challenging due to their complex non-linear relationships and environmental influences. Additionally, satellite-based NDVI predictions face limitations from atmospheric distortions and sensor noise complicating spatial and temporal trend analysis. To enhance NDVI prediction, this study explored a hybrid approach that integrates spatial analysis with temporal forecasting for NDVI prediction. It employed ensemble methods for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal dependencies. First, LULC changes and their relationship with NDVI were analyzed, followed by a comparison of ensemble methods to identify the better model for capturing non-linear interactions between LULC and environmental variables. The superior ensemble model was then integrated with LSTM to develop a hybrid model. Comparing the performance of ensemble models, the results showed that XGBoost performed superiorly. Therefore, based on XGBoost's superior performance, an XGBoost-LSTM hybrid model was developed, improving prediction accuracy with an RMSE of 0.085 and an R^2 score of 0.92. The XGBoost-LSTM hybrid model surpassed individual model performance, enhancing predictive accuracy, scalability, and robustness. This study provides a framework for large-scale vegetation health monitoring with practical applications in sustainable land management and climate adaptation.

Keywords: Land Use and Land Cover (LULC); Normalized Difference Vegetation Index (NDVI); Machine Learning; eXtreme Gradient Boosting (XGBoost); Random Forest (RF); Long Short-Term Memory (LSTM).

I. INTRODUCTION

A. Background of the Study

Land Use and Land Cover (LULC) refers to the physical characteristics of the Earth's surface, shaped by natural and human activities. Changes in LULC, driven by urbanization, agricultural expansion, deforestation, and climate variability, profoundly affect ecosystems, biodiversity, and sustainable development [10]. Monitoring these changes is essential for environmental conservation, land use planning, and policy formulation. Kenya's North Rift region has undergone rapid LULC transformations due to increasing agricultural activities, population growth, and infrastructure development. While these changes contribute to economic growth and food security, they also lead to land degradation, deforestation, and biodiversity loss, affecting terrestrial vegetation health and ecosystem stability [7]. Accurate mapping and prediction of these changes are crucial for sustainable land management and mitigating negative environmental impacts [9]. The Normalized Difference Vegetation Index (NDVI) is a widely used metric for assessing vegetation health. Derived from satellite imagery, NDVI quantifies vegetation cover by measuring the difference between near-infrared and red light reflectance. High NDVI values indicate healthy vegetation, while low values suggest sparse or degraded cover. However, NDVI is dynamic and fluctuates in response to LULC changes. For instance, deforestation reduces NDVI, while agricultural expansion may initially increase it before declining due to soil degradation. Urbanization causes a permanent drop in NDVI as vegetation is replaced by impervious surfaces [14].

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

The relationship between LULC and NDVI is complex, as land cover types interact with climatic and soil conditions to influence vegetation health. Forests and wetlands maintain high NDVI values due to dense vegetation and moisture retention, while agricultural lands exhibit seasonal fluctuations. Grasslands and shrublands vary based on grazing intensity, fire disturbances, and rainfall variability, whereas urban areas consistently show low NDVI values. These diverse interactions make NDVI prediction challenging, requiring advanced computational models to capture spatial and temporal trends effectively [15].

This study attempts to explore a hybrid approach that aims to effectively integrate spatial analysis with temporal trend forecasting for precise NDVI predictions. This study specifically (a) analyzed LULC changes and their relationship with NDVI trends; (b) compared assembled models (RF and XGBoost) to identify the model that was best suited for capturing non-linear relationships and high-dimensional interactions between environmental variables and LULC classes; (c) developed a hybrid model that models both spatial patterns and temporal trends by integrating the better-performing ensemble model (from (b)) with LSTM. By combining the spatial capabilities of ensemble models with LSTM's temporal strengths, this hybrid approach is aimed at effectively capturing both spatial and temporal trends, offering a comprehensive framework for large-scale vegetation monitoring and addressing the challenges of rapid LULC transitions.

The remainder of the paper is structured as follows: Section II reviews related work on LULC classification, NDVI prediction, and hybrid modeling techniques. Section III details the methodology, including the study area, datasets, preprocessing steps, and model development. Section IV presents the experimental results and performance evaluation of the models. Section V discusses key findings, limitations, and implications for land management and climate adaptation. Section VI concludes the paper and suggests future research directions for improving hybrid modeling in NDVI prediction.

II. RELATED WORK

Several studies have explored different approaches to the classification and prediction of vegetation cover. Historically traditional LULC classification methods, such as field surveys and aerial photograph interpretation, were labor intensive and lacked spatial-temporal efficiency [5], [6]. The introduction of remote sensing and automated classification using sensors such as Landsat, MODIS, and Sentinel-2 significantly improved the mapping accuracy [17], [20]. However, challenges such as spectral confusion, where similar types of land cover exhibit overlapping spectral signatures, continue to hinder the precision of the classification [21], as illustrated in Fig. 1 and Fig. 2. Fig. 1 shows the spatial variations of the normalized difference vegetation index (NDVI) for (a) 1995, (b) 2000, (c) 2005, (d) 2011, (e) 2016, and (f) 2020 over the North Rift counties. Fig. 2 also shows the spatial variations of LULC for (a) 1995, (b) 2000, (c) 2005, (d) 2011, (e) 2016, (f) 2020, and (g) 2023 over the North Rift counties. As noted above, it is difficult to identify all the different spatial variations that have taken place from 1995 to 2023 using the naked eye.

This in turn made machine learning come up as a powerful tool to address such issues. A wide range of studies have explored machine learning (ML) and deep learning (DL) for LULC classification. These approaches address challenges in accurately mapping LULC changes and predicting vegetation health across diverse ecological settings. This section reviews existing literature, focusing on the strengths, limitations, and applications of these techniques, with particular emphasis on hybrid models that integrate ML and DL for improved spatial-temporal predictions.

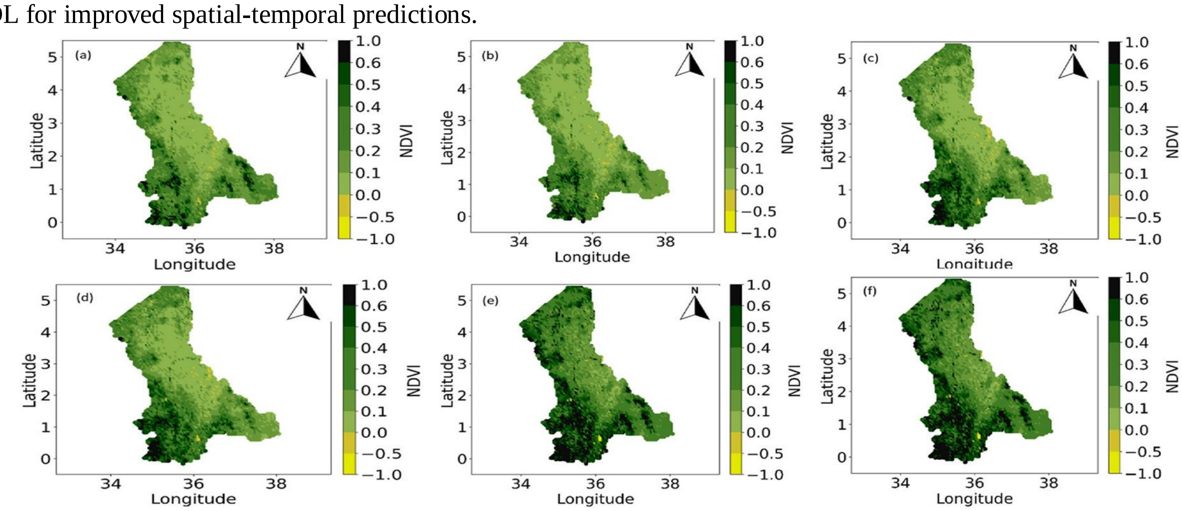


Fig. 1: Spatial variations of Normalized Difference Vegetation Index (NDVI) over the North Rift counties for (a) 1995, (b) 2000, (c) 2005, (d) 2011, (e) 2016, and (f) 2020.

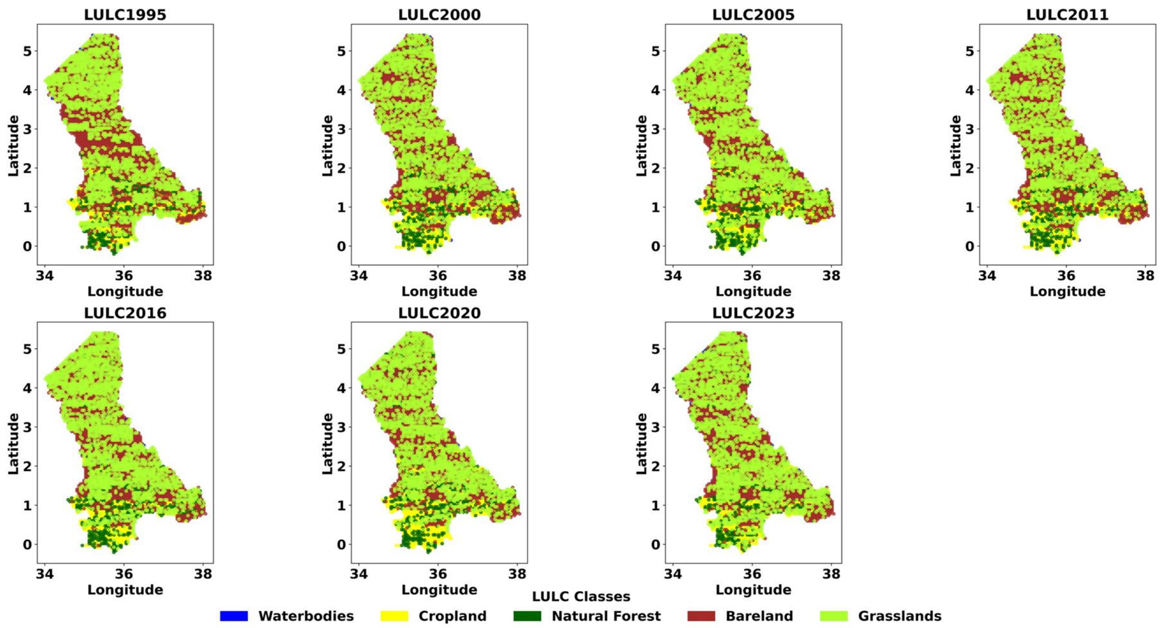


Fig. 2: Spatial variations of Land Use Land Cover (LULC) over the North Rift counties for (a) 1995, (b) 2000, (c) 2005, (d) 2011, (e) 2016, (f) 2020, and (g) 2023.

A. Machine Learning in NDVI Prediction

ML algorithm techniques have emerged as powerful tools for predicting NDVI from LULC data due to their ability to model complex, non-linear spatial-temporal relationships. Unlike traditional manual interpretation methods, ML models process multi-source remote sensing data, enabling automation and improved classification accuracy [20]. Common algorithms such as Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) significantly enhance NDVI prediction efficiency [2], [17], [20].

RF has been successfully applied in Romania for LULC classification and NDVI prediction, leveraging ensemble learning for higher accuracy [2]. Similarly, an SVM-based study in Kabul predicted LULC trends using Landsat images [19]. However, these conventional models struggle with high-dimensional data and require manual feature engineering, limiting their ability to fully capture spatial-temporal dynamics [16]. To overcome these challenges, deep learning (DL) and hybrid models are increasingly adopted for better scalability and accuracy [18].

B. Deep Learning for NDVI Prediction

Deep Learning (DL) methods, a more complex subset of Machine Learning, involve neural networks with many layers. These methods have significantly advanced the field of Land System Science by enabling more sophisticated image classification, object detection, and semantic segmentation tasks [12]. DL's ability to automatically learn hierarchical features from data makes it especially powerful for analyzing complex land system dynamics. Deep learning (DL) techniques, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have shown great promise in predicting NDVI from LULC data. They learn directly from data, eliminating the need for manual feature engineering and efficiently handling high-dimensional data [22], [8]. For instance, a study in Sierra del Cando (Galicia, northwest Spain) using Convolutional Neural Network (CNN), revealed that CNNs prove to be a powerful classification tool with training and test accuracy of 91% and 88%, respectively, on ten different LULC classes [1]. Similarly, a study by [4] reviewed the application of deep learning in multi-temporal remote sensing image classification, demonstrating their effectiveness in predicting vegetation trends over time. The study concluded that RNNs, especially Long Short-Term Memory (LSTM) networks, are well-suited for time series data as they can capture sequential patterns and predict future vegetation trends based on past observations. Furthermore, [8] demonstrated that RNNs can effectively model extreme vegetation responses to climate drivers, underscoring their potential for complex environmental interactions. However, the main limitation of DL models is the requirement for extensive computational resources and large datasets.

C. Hybrid Approaches

To address the limitations of single-method models, various studies have explored hybrid approaches by integrating ML and DL techniques to enhance prediction accuracy and robustness. For instance, a study in China [27] introduced a spatiotemporal NDVI fusion model using residual CNNs, which significantly improved NDVI prediction accuracy compared to traditional linear regression methods. Similarly, studies in India and Brazil integrated classification and regression tasks or employed Genetic Algorithm–Artificial Neural Network (GA-ANN) models to mitigate data inconsistencies [3], [23]. A study in Guna Tana Watershed of Abay Basin, Ethiopia combined SVM and Cellular Automata-based Artificial Neural Network (CA-ANN) models to detect and predict LULC changes in the Guna Tana watershed [5]. Sahoo and Chakraborty developed a hybrid model combining 3D Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) networks for hyperspectral image classification. This model effectively captured spatial-spectral features but required a large number of parameters and significant computational resources, potentially limiting its scalability [13].

Despite the success of these hybrid models, they exhibit several limitations. CNN-based hybrids (e.g., CNN-LSTM and 3D CNN-BiLSTM) require extensive computational resources, limiting their scalability for large-scale environmental monitoring. GA-ANN and BiLSTM-based models rely on extensive training data, which is often unavailable in regions with limited ground-truth data like sub-Saharan Africa, where this research study was done. Furthermore, models dependent on remote sensing imagery (e.g., CNN-based models) struggle with cloud cover, sensor inconsistencies, and missing data. Approaches such as GA-ANN and CA-ANN require intricate hyperparameter adjustments, making them challenging to optimize. While SVM-CA-ANN effectively models spatial relationships, it lacks robust temporal forecasting capabilities.

To address these challenges, this study compares Random Forest (RF) and Extreme Gradient Boosting (XGBoost), two ensemble techniques known for their robustness in handling high-dimensional data and minimizing overfitting by combining multiple decision trees. LSTM is also chosen specifically for its ability to model long-term dependencies in sequential data, making it well-suited for NDVI trend forecasting. Thus, by combining the spatial feature extraction strengths of ensemble methods (RF/XGBoost) with LSTM's temporal learning capabilities, this hybrid approach provides a more efficient, scalable, and interpretable solution for NDVI prediction than existing CNN-heavy architectures.

D. Problem Statement

Predicting NDVI from LULC remains challenging due to non-linear interactions between environmental variables such as soil properties, precipitation, temperature, and human activities [4]. Additionally, satellite-based NDVI predictions continue to face challenges related to atmospheric distortions and sensor noise [8]. Despite deep learning's success in handling some of these limitations, DL models require vast labeled datasets, posing a significant challenge in regions with limited ground-truth data, such as sub-Saharan Africa. Furthermore, despite hybrid models having demonstrated significant improvements in prediction performance, as highlighted in the literature, most existing models primarily focus on either spatial or temporal patterns, rather than effectively integrating both aspects.

To effectively capture both spatial and temporal trends, this study employed a novel hybrid approach that integrates ML-based ensemble models with DL techniques. Specifically, the study explored ensemble methods (RF and XGBoost) for spatial feature extraction, combined with LSTM networks for temporal sequence modeling. Unlike CNN-based models, which heavily rely on image data, this study leverages structured environmental datasets (temperature, precipitation, Bare Soil Index (BSI), and Moisture Stress Index), reducing dependency on high-resolution satellite imagery and mitigating sensor limitations like noise issues.

The aim of employing this approach was to attempt to improve prediction accuracy, reduce computational complexity, and enhance model interpretability, making it more scalable for large-scale vegetation monitoring. Additionally, it aimed at minimizing sensor dependencies, ensuring greater resilience to missing data and atmospheric distortions.

III. METHODOLOGY

This study highlights the importance of integrating ML and DL techniques for environmental analysis, positioning the XGBoost-LSTM hybrid model as a viable solution for monitoring vegetation dynamics, particularly in regions experiencing rapid LULC transitions and data scarcity. This section outlines how this was achieved.

A. Study Area

This study was conducted in the North Rift region of Kenya, spanning latitudes 1°N to 5°N and longitudes 34°E to 38°E (see Fig. 3).

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

The region includes the counties of Trans-Nzoia, Uasin Gishu, Elgeyo-Marakwet, Nandi, Samburu, Turkana, Baringo, and West Pokot. The area's diverse climate is influenced by its topography and atmospheric systems. Rapid urbanization and agricultural expansion have caused significant vegetation changes, making it vital for LULC monitoring and prediction. Vegetation cover ranges from forests and savannas to wetlands, shrublands, and grasslands. Highland counties like Uasin Gishu, Nandi, and Elgeyo-Marakwet experience temperate conditions (10°C to 25°C) and are known for agriculture, particularly maize, wheat, and tea cultivation. In contrast, lowland counties such as Turkana, West Pokot, and Samburu exhibit semi-arid to arid climates, where pastoralism and food crop farming are predominant, making climatic factors critical for land productivity [11].

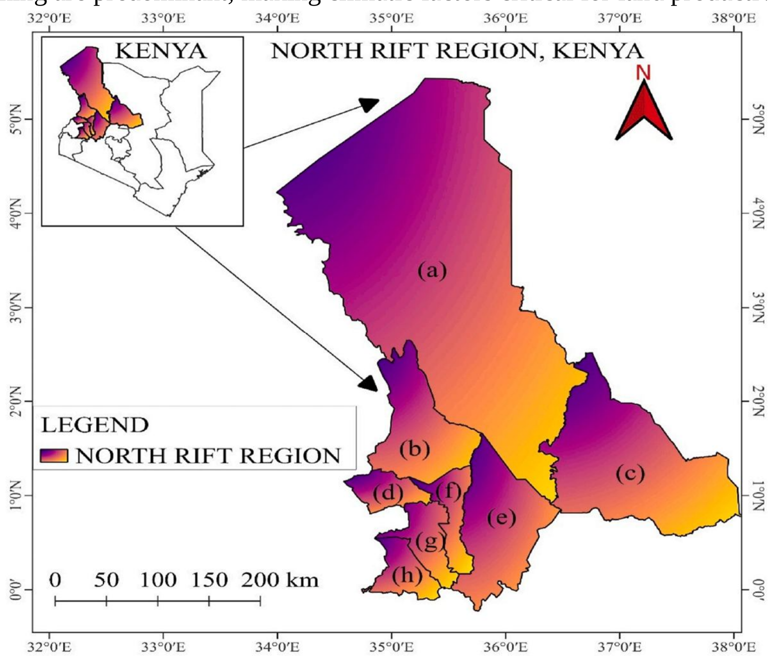


Fig. 3: Map of North Rift Region, Kenya, highlighting Turkana, West Pokot, Trans Nzoia, Elgeyo-Marakwet, Uasin Gishu, Nandi, Baringo, and Samburu counties.

B. Datasets

Multiple datasets were used to assess LULC changes and predict NDVI across Kenya's North Rift region. Satellite imagery from Landsat (USGS Earth Explorer) and Sentinel-2 (Copernicus Open Access Hub) spanned from 1990 to 2022 at two-year intervals. Landsat 4-5 Thematic Mapper (1990–2013) and Landsat 7 ETM+ (2014–2021) provided 30-meter resolution imagery, while Sentinel-2 (2022) offered 10-meter resolution data. Band combinations included Bands 4, 3, and 2 for Landsat images and Bands 8, 4, and 3 for Sentinel imagery. Environmental variables like temperature, precipitation, Bare Soil Index (BSI), and Soil Moisture Index (SMI) were incorporated to assess environmental influences on NDVI. Temperature and precipitation data were sourced from CHIRPS and TerraClimate, respectively, while BSI and SMI were obtained from SoilGrids.

TABLE I

DATASETS USED FOR LULC CLASSIFICATION AND NDVI ANALYSIS

Dataset	Source	Timeframe	Resolution
Landsat 4-5 TM	USGS Earth Explorer	1990–2013	30 m
Landsat 7 ETM+	USGS Earth Explorer	2014–2021	30 m
Sentinel-2	Copernicus Open Access Hub	2022	10 m
Precipitation	CHIRPS	1990–2022	~5 km
Temperature	TerraClimate	1990–2022	~4 km
Bare Soil Index (BSI)	SoilGrids	1990–2022	–
Soil Moisture Index (SMI)	SoilGrids	1990–2022	–

C. Methods

This study employed ensemble methods—Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Long Short-Term Memory (LSTM) networks—for LULC classification and NDVI prediction. RF and XGBoost, known for handling high-dimensional data and reducing overfitting, captured spatial relationships, while LSTM modeled temporal dependencies, enhancing NDVI trend forecasting. This hybrid approach integrated structured environmental variables (temperature, precipitation, soil moisture index (SMI), and bare soil index (BSI)) to improve scalability and mitigate sensor-related issues.

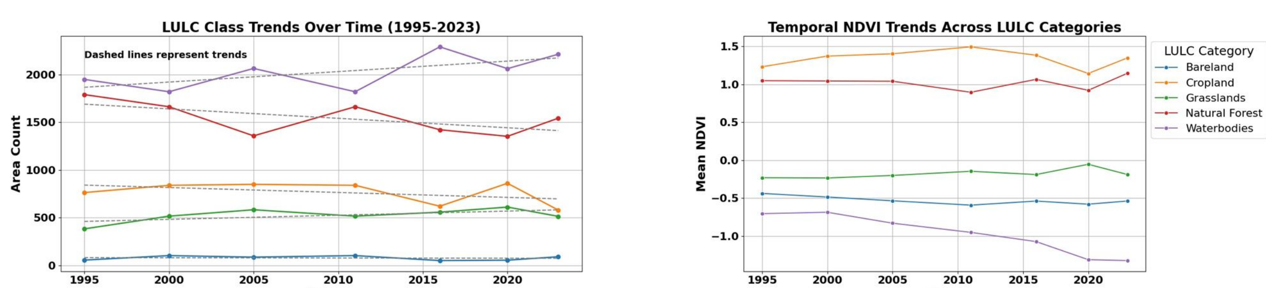
First, XGBoost and RF were compared for LULC classification to identify the better model for capturing non-linear environmental interactions. The superior model was then integrated with LSTM, combining ensemble-based spatial feature extraction with temporal sequence modeling. The purpose of this fusion was to effectively capture both spatial and temporal trends to improve predictive accuracy, scalability, and robustness for large-scale vegetation monitoring. The ultimate goal is to improve vegetation health monitoring and sustainable land management.

1) *Random Forest and XGBoost for LULC Classification:* RF constructs multiple decision trees and outputs the class mode for classification, offering robustness and scalability. Key hyperparameters included estimators (100, 200, 500), maximum depth (10, 15, 20), and the Gini impurity criterion. XGBoost, leveraging gradient boosting with built-in regularization, was optimized with learning rates (0.01, 0.1, 0.3), tree depths (6, 10, 15), and a subsampling ratio (0.8) to prevent overfitting. Both models were evaluated using RMSE and R^2, with the superior model proceeding to hybridization. Hyperparameter tuning optimized performance, varying n_estimators (100, 200, 500), learning rates (0.01, 0.1, 0.3), and tree depth (6, 10, 15). A 0.8 subsampling ratio prevented overfitting, and a multi-class classification objective with a Softmax loss function was used. Model performance was evaluated using Root Mean Squared Error (RMSE) and R^2 score.
2) *Long Short-Term Memory (LSTM) Model for NDVI Prediction:* LSTM captured temporal dependencies in NDVI time-series data, leveraging historical values for future predictions. The model, implemented using TensorFlow and Keras, featured an input layer for sequential NDVI values. Units (50, 64, 128, 256), dropout rates (0.2, 0.3, 0.5), and layers (1, 2, 3) were tested for complexity and overfitting control. The Adam optimizer and Mean Squared Error (MSE) loss function minimized prediction errors. Performance was evaluated using RMSE, Mean Absolute Error (MAE), and R^2 score.
3) *Hybrid Model Development:* The hybrid model integrated the better-performing ensemble model (XGBoost or Random Forest) with LSTM to capture both spatial and temporal patterns in NDVI. Spatial features such as LULC, temperature, precipitation, BSI, and SMI were extracted, and the ensemble model's predictions were fed into LSTM for sequential analysis. The model was trained, validated, and evaluated using RMSE, MAE, and R^2 to ensure accuracy and robustness.

IV. RESULTS

A. Performance of LULC Classification

LULC changes were analyzed between 1995 and 2023 to observe land cover dynamics. Fig. 4(a) shows the trends in each LULC category over time. The “Actual” values represent the observed or recorded data for each LULC class at specific years (1995, 2000, 2005, etc.), shown as solid points connected by lines. The “Trend” values represent a fitted line or curve that shows the overall pattern or direction of change in each LULC class over time; these dashed trend lines help to visualize long-term patterns and smooth out fluctuations in the actual data. Fig. 4(b) shows the trends in NDVI across LULC categories. Analyzing NDVI trends across different LULC categories (e.g., cropland, forests, grasslands) helps reveal how vegetation health changes over time within each land cover type.



(a) LULC trends from 1995–2023. (b) NDVI trends across LULC categories.

Fig. 4: Comparison of NDVI and LULC trends.

The analysis shows that cropland has expanded significantly, particularly between 2016 and 2020, likely driven by increased food demand or agricultural intensification. A declining trend in natural forest cover reflects potential deforestation due to agricultural expansion or urbanization. The sharp fluctuations in waterbodies suggest climate variability impacts or improved water management infrastructure, while increases in bareland signal environmental degradation, possibly due to soil erosion, deforestation, or overgrazing.

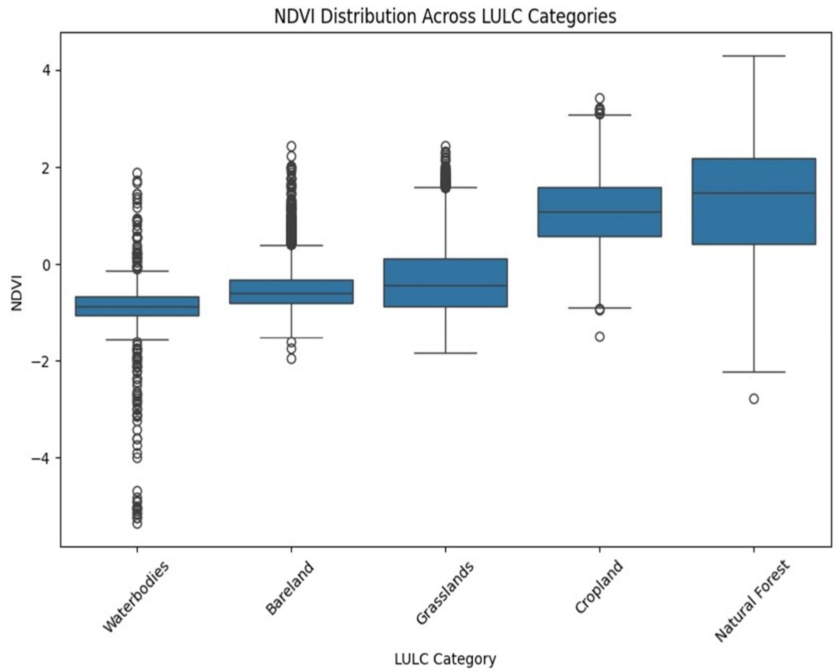


Fig. 5: NDVI distribution among LULC categories.

Fig. 5 shows NDVI distribution among LULC categories. The results show that high NDVI values (close to 1) are represented by the Natural Forest class, indicating the presence of healthy vegetation. Moderate NDVI is indicated by croplands, suggesting the presence of agricultural fields. Grasslands, which have shrub cover, exhibit lower NDVI than croplands due to little to no vegetation. Waterbodies and bare land have negative NDVI, as bare land lacks vegetation and water-specific reflectance affects NDVI for waterbodies.

1) *Feature Importance:* Features with the highest importance and their role in predicting NDVI range from LULC_avg (Land Use Land Cover), which has the highest feature importance, meaning that historical LULC data plays the most significant role in predicting NDVI. This makes sense, as land cover changes directly affect vegetation health. TMIN_avg (Minimum Temperature), Prec_avg (Precipitation), and SMI_avg (Soil Moisture Index) have moderate importance; these variables influence vegetation growth and water availability, impacting NDVI. TMAX_avg (Maximum Temperature) and BSI_avg (Bare Soil Index) have lower importance. While they contribute to NDVI predictions, their impact is relatively smaller than LULC and precipitation-related factors.
2) *Classification Metrics:* Classification precision was assessed using precision, recall, and F1 score for each LULC category. The classification matrix (Table II) illustrates the performance of the model in different types of land cover. The overall classification accuracy achieved by the XGBoost model was 88%, with the highest accuracy recorded for Natural Forest, Bareland, and Grasslands, and the lowest for Cropland.

The overall classification accuracy was 88%, with a macro average precision of 0.93, recall of 0.90, and F1-score of 0.89. The weighted average precision, recall, and F1-score were 0.92, 0.88, and 0.87, respectively. The model performs well overall, with high precision, recall, and F1 scores for most LULC categories. However, the performance for Cropland could be improved, as it has a low recall, indicating that some Cropland instances are being misclassified.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

TABLE II
CLASSIFICATION METRICS FOR LULC CLASS

LULC Class	Precision	Recall	F1-Score
Waterbodies	0.67	1.00	0.80
Cropland	1.00	0.50	0.67
Natural Forest	1.00	1.00	1.00
Bareland	1.00	1.00	1.00
Grasslands	1.00	1.00	1.00
Accuracy		0.88	
Macro Avg	0.93	0.90	0.89
Weighted Avg	0.92	0.88	0.87

3) *Confusion Matrix:* Confusion matrices for Random Forest and XGBoost (Fig. 6) provide insights into their classification performance. Random Forest correctly classifies 703 instances of Natural Forest (class 2) but misclassifies 47 instances as Bareland (class 3), indicating some confusion between these categories. Other classes show moderate accuracy. XGBoost also performs well in class 2, with 704 correct predictions and fewer misclassifications than Random Forest, particularly in handling minority classes. This aligns with its higher R^2 scores and lower RMSE values. XGBoost's ability to assign higher weights to minority classes improves classification performance, making it superior in both overall accuracy and class balance handling.

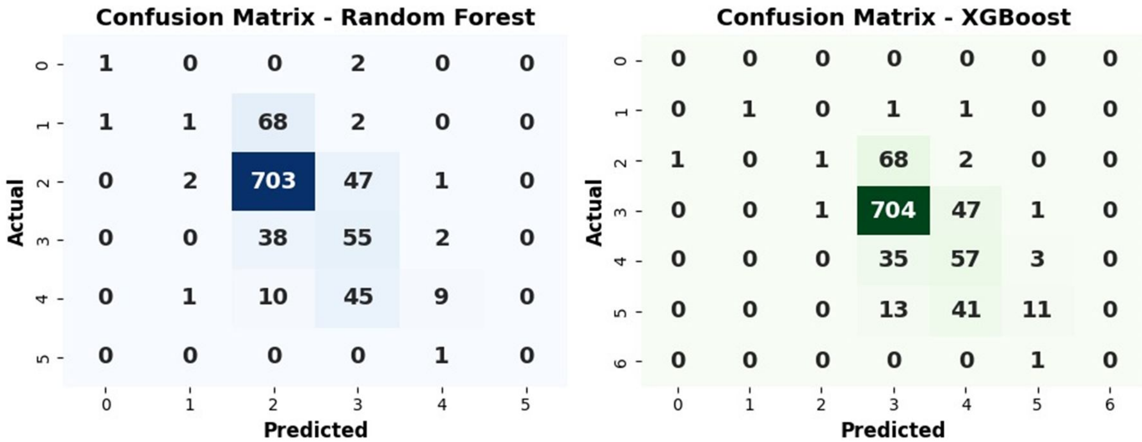


Fig. 6: Confusion matrices for Random Forest and XGBoost classifiers.

B. Model Performance Progression and Accuracy Improvement

1) *Model 1: Random Forest Regression (LULC Only):* The first model used Random Forest Regression with LULC as the only predictor. Despite being a powerful algorithm, its performance was poor, with an RMSE of 0.6948 and an R^2 score of 0.5281. This shows that LULC alone is not enough for accurate NDVI prediction. Adding temperature, precipitation, and soil indices to the Random Forest model improved performance, reducing the RMSE to 0.47 and increasing the R^2 score to 0.78. This confirms that NDVI is influenced by multiple factors, not just LULC. However, the model still had limitations, likely due to irrelevant or less significant features affecting its accuracy.
2) *Model 2: XGBoost with Advanced Features:* The second model used XGBoost with additional features, including lagged NDVI values (NDVI_Lag1, NDVI_Lag2), rolling averages of temperature, precipitation, and NDVI (TMAX_Rolling, TMIN_Rolling, Prec_Rolling), along with hyperparameter tuning. This approach improved performance, achieving an RMSE of 0.5008 and an R^2 score of 0.7553. However, the model may have included redundant features, leading to noise and potential overfitting.

3) *Model 3: Optimized XGBoost with Feature Selection:* To improve generalization, Model 3 focused only on the most relevant variables: LULC, NDVI, temperature, precipitation, BSI, and SMI. This simplified model performed significantly better, with an RMSE of 0.0619 and an R^2 score of 0.9963. This shows that careful feature selection improves accuracy while preventing unnecessary complexity.
4) *LSTM for Time Series Prediction:* The LSTM model was fine-tuned manually and evaluated using RMSE and R^2 scores. It achieved an RMSE of 0.1022, showing low prediction error, and an R^2 score of 0.827, meaning it explained 82.7% of the variation in NDVI. These results highlight LSTM's ability to learn temporal patterns and trends. While slightly less accurate than XGBoost, LSTM is useful for time-series forecasting and can capture long-term vegetation changes.
5) *Hybrid Ensemble Approach:* RMSE and R^2 scores for each model were considered to weigh their contributions to the final prediction. Since XGBoost performed better (lower RMSE and higher R^2), its prediction was weighted more heavily than the LSTM's prediction. The hybrid model performs better than LSTM (RMSE = 0.1022) because it includes feature-based predictions from XGBoost, which has an RMSE of 0.0619. By combining the strengths of both models, the hybrid approach reduces prediction error, leading to a balanced RMSE of 0.085. The RMSE falls between XGBoost and LSTM's values because the hybrid model's performance is essentially an average of the two models' strengths, focusing on minimizing prediction error while maintaining the ability to account for temporal trends. The R^2 score of 0.92 for the hybrid model reflects an improvement over LSTM (R^2 = 0.827), but it does not fully reach XGBoost's near-perfect R^2 (0.9963). This score indicates that 92% of the variance in the NDVI data can be explained by the model, showcasing strong explanatory power. The hybrid model benefits from both XGBoost's feature-based insights (capturing correlations between NDVI and environmental features) and LSTM's temporal insights (capturing trends and patterns over time), making the combination more accurate than using either model alone.

V. DISCUSSION

The model classified Natural Forest, Bareland, and Grasslands with 100% accuracy, suggesting these classes have distinct characteristics that make them easy to identify. However, one Cropland instance was misclassified as Waterbodies, likely due to overlapping spectral properties or transitional land cover. Despite this, the overall classification accuracy was 88%, with Cropland requiring further improvement.

Analysis of NDVI trends across LULC classes reveals clear patterns. Waterbodies showed stable NDVI values over time. Cropland NDVI fluctuated, likely due to changing farming practices and climate variability. Natural Forest NDVI declined from 1995 to 2011, increased from 2011 to 2016, and then declined again from 2016 to 2023. The initial decline may be due to deforestation or land-use changes. The increase between 2011 and 2016 could reflect conservation or reforestation efforts, while the later decline may be linked to renewed deforestation, agricultural expansion, or urbanization.

Bareland NDVI remained consistently low, with minor fluctuations possibly due to land degradation or changing land management. Grasslands, however, showed a steady increase in NDVI, indicating improved vegetation cover or successful restoration initiatives.

A. *Interpretation of Model Performance*

The hybrid model (LSTM + XGBoost) outperformed individual models in NDVI prediction. XGBoost excelled over Random Forest by effectively handling complex feature interactions and reducing errors through sequential learning. Its regularization techniques also minimized overfitting, improving generalization. The hybrid approach combined LSTM's ability to capture temporal patterns with XGBoost's feature-based strengths, enhancing prediction accuracy. However, if the hybrid model's performance is comparable to the individual models, it suggests that the redundancy between the models might limit the potential for improvement. The trade-off between model complexity and performance should be carefully considered when deploying such a hybrid approach.

B. *Analysis of Model Evaluation Metrics*

In terms of evaluation metrics, both RMSE and R^2 scores provide important insights into the model's performance. The RMSE values indicate the average magnitude of error between the predicted and actual NDVI values, with a lower RMSE reflecting better model performance. The R^2 score reveals the proportion of variance in the NDVI data explained by the model, with values closer to 1 suggesting a better model fit. For the hybrid model, the results demonstrate strong predictive power, as evidenced by the optimized RMSE and R^2 values. These metrics show that the hybrid model effectively captures both the temporal and feature-based patterns in the data. Moreover, feature importance analysis in XGBoost can shed light on which variables—such as LULC, temperature, and precipitation—have the greatest impact on NDVI predictions, offering valuable insights into the drivers of vegetation changes in the study area. Table III summarizes the performance metrics of all models evaluated in this study.

TABLE III
PERFORMANCE METRICS FOR DIFFERENT MODELS

Model	RMSE	R^2 Score	Comments
Random Forest (LULC only)	0.6948	0.5281	Limited by absence of environmental variables
Random Forest (LULC + Environmental Variables)	0.47	0.78	Improved with added variables
XGBoost (Advanced Features)	0.5008	0.7553	Performance affected by complexity and noise
XGBoost (Basic Features)	0.0619	0.9963	Best performance with optimized feature set
LSTM (Time Series)	0.1022	0.827	Captures temporal NDVI trends effectively
Hybrid XGBoost-LSTM	0.085	0.92	Combines spatial and temporal strengths

C. Insights and Real-World Implications

The hybrid model offers significant implications for understanding the relationship between LULC and NDVI in the context of vegetation cover prediction. LULC data plays a crucial role in determining vegetation cover by influencing factors such as land availability, agricultural practices, and forest management. By integrating LULC as a key predictor, the hybrid model provides insights into how different land-use categories (e.g., cropland, natural forest, and grasslands) affect NDVI, which is directly linked to vegetation health. The findings of this research offer valuable insights for policymakers and environmental managers. The model can be used to monitor deforestation rates, track agricultural expansion, and guide land-use planning in the North Rift region. Early detection of NDVI declines can inform timely interventions aimed at curbing land degradation and promoting sustainable agricultural practices.

D. Limitations and Challenges

Data gaps and resolution may affect accuracy, especially in sparsely monitored areas. Overfitting remains a concern, though mitigated by hybridization. High computational demands could limit real-time applications, making efficiency a key factor in deployment.

E. Future Work

Future studies could integrate socio-economic factors, such as population density or land tenure data, to improve the explanatory power of the model. Additionally, experimenting with advanced deep learning architectures, such as Transformer-based models, could enhance prediction accuracy and generalizability.

VI. CONCLUSION

Model 3 (XGBoost with selected features) proved to be the most effective NDVI predictor, showing that a well-curated feature set outperforms complex models with excessive variables. While Random Forest performed moderately (RMSE = 0.47, $R^2 = 0.78$), it failed to capture intricate relationships as effectively as XGBoost. Model 2 (XGBoost with additional features) did not significantly improve accuracy, likely due to redundant features or overfitting. The hybrid LSTM-XGBoost model combined LSTM's ability to capture temporal trends with XGBoost's feature importance insights, achieving superior accuracy (RMSE = 0.085, $R^2 = 0.92$). This approach enhances NDVI prediction in diverse land-use scenarios. Despite its strengths, the hybrid model faces challenges such as computational complexity, potential overfitting, and data quality issues stemming from inconsistencies in satellite imagery. Nevertheless, it holds significant potential for real-world applications, including environmental monitoring, land-use planning, and climate adaptation strategies. Future research should focus on refining model architecture, improving the quality of input data, and incorporating socio-economic variables to enhance predictive accuracy and scalability across different ecological contexts.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support for implementing this project, which was made possible by the Research Grant (109705-001/002) from the Responsible Artificial Intelligence Network for Climate Action in Africa (RAINCA) consortium. The consortium, comprising WASCAL, RUFORUM, and AKADEMIYA2063, is funded by the International Development Research Centre (IDRC). We appreciate their contribution to advancing climate action and environmental research in Africa.

REFERENCES

[1] Acuña-Alonso, C., García-Ontiyuelo, M., Barba-Barragáns, D., Álvarez, X.: Development of a convolutional neural network to accurately detect land use and land cover. MethodsX 12, 102719 (2024).
[2] Belgiu, M., Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31 (2016).
[3] Chang, C.H., Lin, J., Chang, J.W., Huang, Y.S., Lai, M.H., Chang, Y.J.: Hybrid deep neural networks with multi-tasking for rice yield prediction using remote sensing data. Agriculture 14(4), 513 (2024).
[4] Cheng, X., Sun, Y., Zhang, W., Wang, Y., Cao, X., Wang, Y.: Application of deep learning in multitemporal remote sensing image classification. Remote Sensing 15(15), 3859 (2023).
[5] Kulithalai Shiyam, P., Deka, P.C.: Spatio-temporal classification and prediction of land use and land cover change for the Vembanad Lake system, Kerala. Environmental Science and Pollution Research 29(57), 86220–86236 (2022).
[6] Li, Z., Chen, B., Wu, S., Su, M., Chen, J.M., Xu, B.: Deep learning for urban land use category classification: A review and experimental assessment. Remote Sensing of Environment 311, 114290 (2024).
[7] Liping, C., Yujun, S., Saeed, S.: Monitoring and predicting land use and land cover changes using remote sensing and GIS techniques—a case study of a hilly area, Jiangle, China. PloS One 13(7), e0200493 (2018).
[8] Martinuzzi, F., Mahecha, M.D., Camps-Valls, G., Montero, D., Williams, K.: Learning extreme vegetation response to climate drivers with recurrent neural networks. Nonlinear Processes in Geophysics 31(4), 535–557 (2024).
[9] Mulinge, W., Gicheru, P., Murithi, F., Maingi, P., Kihiu, E., Kirui, O.K., Mirzabaev, A.: Economics of land degradation and improvement in Kenya. Economics of Land Degradation and Improvement–A Global Assessment for Sustainable Development, pp. 471–498 (2016).
[10] Najmuddin, O., Li, Z., Khan, R., Zhuang, W.: Valuation of land-use/land-cover-based ecosystem services in Afghanistan - an assessment of the past and future. Land 11(11), 1906 (2022). https://doi.org/10.3390/land11111906
[11] Nyariki, D.M., Amwata, D.A.: The value of pastoralism in Kenya: Application of total economic value approach. Pastoralism 9(1), 1–13 (2019).
[12] Rangel, A., Terven, J., Cordova-Esparza, D.M., Chavez-Urbiola, E.A.: Land cover image classification. arXiv preprint arXiv:2401.09607 (2024).
[13] Sahoo, A.R., Chakraborty, P.: Hybrid CNN Bi-LSTM neural network for hyperspectral image classification. arXiv preprint arXiv:2402.10026 (2024).
[14] Seifu, T.K., Woldesenbet, T.A., Alemayehu, T., Ayenew, T.: Spatio-temporal change of land use/land cover and vegetation using multi-MODIS satellite data, western Ethiopia. The Scientific World Journal 2023(1), 7454137 (2023).
[15] Singh, S.K., Laari, P.B., Mustak, S., Srivastava, P.K., Szabó, S.: Modelling of land use land cover change using earth observation data-sets of Tons River Basin, Madhya Pradesh, India. Geocarto International 33(11), 1202–1222 (2018).
[16] Solórzano, J.V., Mas, J.F., Gao, Y., Gallardo-Cruz, J.A.: Land use land cover classification with U-Net: Advantages of combining Sentinel-1 and Sentinel-2 imagery. Remote Sensing 13(18), 3600 (2021).
[17] Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.A., Rahman, A.: Land-use land-cover classification by machine learning classifiers for satellite observations—a review. Remote Sensing 12(7), 1135 (2020).
[18] Taye, M.M.: Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. Computers 12(5), 91 (2023).
[19] Ullah, S., Qiao, X., Abbas, M.: Addressing the impact of land use land cover changes on land surface temperature using machine learning algorithms. Scientific Reports 14(1), 18746 (2024).
[20] Wang, J., Bretz, M., Dewan, M.A.A., Delavar, M.A.: Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. Science of The Total Environment 822, 153559 (2022).
[21] Wiatkowska, B., Słodczyk, J., Stokowska, A.: Spatial-temporal land use and land cover changes in urban areas using remote sensing images and GIS analysis: The case study of Opole, Poland. Geosciences 11(8), 312 (2021).
[22] Zhang, X., Zhou, Y., Luo, J.: Deep learning for processing and analysis of remote sensing big data: A technical review. Big Earth Data 6(4), 527–560 (2022).
[23] Zhao, Y., Hou, P., Jiang, J., Zhao, J., Chen, Y., Zhai, J.: High-spatial-resolution NDVI reconstruction with GA-ANN. Sensors 23(4), 2040 (2023).