2025. Volume 9, Number 1 *Pages 62-74*

UOT: 004.8:004.93:528.8:528.9
DOI: https://doi.org/10.30546/09090.2025.210.016

LAND USE AND LAND COVER MAPPING IN THE KARABAKH REGION USING SENTINEL-2 MULTI-SPECTRAL INDICES AND MACHINE LEARNING

ARTUGHRUL GAYIBOV[1]
agayibov@beu.edu.az
https://orcid.org/0009-0009-7349-0286

VAGIF GASIMOV[1]
vaqasimov@beu.edu.az
https://orcid.org/0000-0003-3192-4225
Baku Engineering University
Baku, Azerbaijan

ARTICLE INFO

Article history:
Received:2025-07-02
Received in revised form:2025-07-02
Accepted:2025-10-22
Available online

Keywords:
Land Use/Land Cover Classification;
Machine Learning;
Sentinel-2 Satellite Imagery;
Random Forest;
Remote Sensing;
Post-conflict Reconstruction

2010 Mathematics Subject Classifications: 8U10 → 62H35; 68T10; 86A30; 62P12.

ABSTRACT

Planning for reconstruction, resource management, and environmental monitoring all depend on Land Use/Land Cover (LULC) mapping, especially in areas that have experienced conflict, such as Karabakh, Azerbaijan. Five machine learning algorithms for LULC classification using high-resolution Sentinel-2 satellite imagery are thoroughly compared in this study. We implemented and assessed the Random Forest (RF), Classification and Regression Trees (CART), Gradient Tree Boosting (GTB), k-Nearest Neighbors (k-NN), and Gaussian Naïve Bayes classifiers using the Google Earth Engine (GEE) platform. In order to improve the separability of seven LULC classes—Water, Trees, Grass, Flooded Vegetation, Crops, Built Area, and Bare Ground—a rich feature set consisting of eight multispectral bands and thirty derived spectral indices was added to the classification. The ESA WorldCover 2020 dataset was used to generate the training and validation data. The Kappa coefficient and overall accuracy were used to quantitatively evaluate the findings. Gradient Tree Boosting (Kappa ≈ 0.71) and Random Forest (Kappa = 0.697) outperformed k-NN (Kappa = 0.637), CART (Kappa = 0.599), and Naïve Bayes (Kappa = 0.173) by a significant margin. Ensemble methods also showed superior performance. The results demonstrate how well ensemble classifiers handle high-dimensional remote sensing data and offer a methodological framework for quick and precise LULC mapping to aid in the region's post-conflict recovery and sustainable development initiatives.

1. INTRODUCTION

A key component of sustainable resource management, strategic regional planning, and efficient environmental governance is the monitoring of land use and cover (LULC). The need for precise, timely, and high-resolution LULC data becomes even more critical in post-conflict areas. A crucial case study is the Karabakh region of Azerbaijan, which has just recently recovered from a protracted conflict. Although decades of occupation and the direct effects of conflict have

permanently altered the landscape, the end of hostilities has ushered in a new era of reconstruction and resettlement. In order to inform national policy on ecological restoration, agricultural revitalization, water resource management, and urban and rural development, it is essential to have a thorough understanding of the current condition of the region's forests, agricultural lands, water bodies, and infrastructure. Armed conflicts are known to be potent catalysts for environmental change, frequently leading to unchecked urbanization or the destruction of built-up areas, deforestation, agricultural land abandonment, and land degradation (Akar & Güngör, 2015). To quantify the environmental legacy of the conflict and set a course for a resilient and sustainable future, it is imperative that a thorough and accurate LULC baseline be established first.

Remote sensing has changed dramatically with the introduction of publicly accessible satellite data from initiatives like the European Space Agency's (ESA) Copernicus. One mission that is particularly useful for LULC classification is Sentinel-2. Its unique ability to map the Earth's surface in fine detail is made possible by its combination of high temporal resolution (5-day revisit time), rich multispectral sensor array (13 bands), and high spatial resolution (10-20 meters) (Griffiths et al., 2019). This is particularly relevant for landscapes that are heterogeneous, such as Karabakh, where intricate mosaics of various land cover types coexist in close proximity. The capacity to extract a wide range of spectral indices is a significant benefit of Sentinel-2 data. Spectral indices, which are mathematical combinations of various spectral bands, are designed to improve particular biophysical characteristics of the surface, whereas a basic color image can differentiate between broad categories. For example, the Normalized Difference Water Index (NDWI) efficiently delineates water bodies, the Normalized Difference Vegetation Index (NDVI) is extremely sensitive to vegetation health, and the Normalized Difference Built-up Index (NDBI) aids in identifying urban areas. We can greatly enhance machine learning algorithms' capacity to discriminate between spectrally similar but functionally distinct LULC classes by constructing a rich "feature space" that includes both raw spectral bands and dozens of these derived indices (Zhang & Xie, 2019).

Even though Sentinel-2 LULC studies have proliferated, many of these studies only use one classification algorithm—Random Forest, most frequently—without methodically weighing other options. According to Maxwell et al. (2018), the "No Free Lunch" theorem in machine learning asserts that no single algorithm is best suited for all problems. The type of data, the intricacy of the terrain, and the particular classes being mapped all have a significant impact on a classifier's performance. To determine the most reliable and accurate algorithmic approach for a particular situation, a thorough comparison of various approaches is necessary (Li et al., 2014). By performing a thorough comparative analysis of five different machine learning classifiers integrated into the potent Google Earth Engine (GEE) cloud computing platform, this study seeks to close this gap (Gorelick et al., 2017). We assess a probabilistic classifier (Gaussian Naïve Bayes), an instance-based learner (k-Nearest Neighbors), a classic decision tree algorithm (CART), and two popular ensemble techniques (Random Forest and Gradient Tree Boosting). Finding the best classification method for creating a high-resolution LULC map of the Karabakh region is the main goal of this study. By doing this, we provide a transparent and repeatable methodological framework that can be modified for quick and efficient environmental evaluation in other post-conflict or data-poor areas across the globe, ultimately serving as a vital scientific instrument to promote sustainable development and evidence-based reconstruction.

2. LITERATURE REVIEW

Over the past 50 years, the field of LULC classification using satellite remote sensing has undergone significant change due to developments in sensor technology, computing power, and analytical techniques. Using Landsat MSS data, early research in the 1970s and 1980s mostly employed supervised techniques based on statistical measures like maximum likelihood classification (MLC) or unsupervised techniques like ISODATA. Although revolutionary at the time, these techniques were sensitive to atmospheric conditions and frequently had trouble with spectral confusion between various LULC classes (Phiri & Morgenroth, 2017). An important turning point was the introduction of sensors with greater resolution and the creation of increasingly complex algorithms.

The 1990s and 2000s saw the rise of machine learning, which provided strong new classification tools. By assuming less about the underlying data distribution, algorithms such as Support Vector Machines (SVM) and single Decision Trees (like CART) outperformed conventional statistical techniques (Maxwell et al., 2018). But it was the widespread adoption of ensemble learning methods, especially Random Forest (RF), that brought about the real revolution. Because of its high accuracy, resilience to noise and overfitting, and capacity to handle high-dimensional data without the need for intricate feature selection, radio frequency (RF) rapidly emerged as the de facto standard in remote sensing after being introduced by Breiman (2001). Using a variety of satellite platforms, numerous studies have shown that RF is superior for LULC mapping across diverse ecosystems. One of the first thorough assessments, for example, was given by Pal (2005), who demonstrated that RF performed better at classifying Landsat data than SVM and a single decision tree. In a more recent review, Belgiu and Drăguţ (2016) confirmed the popularity and efficacy of radio frequency (RF) in remote sensing, pointing to its robust performance and ease of use as the main drivers of its widespread adoption.

Research was further accelerated by the 2015 launch of the Sentinel-2 constellation, which made worldwide high-resolution multispectral data freely and publicly available. Sentinel-2's distinctive spectral bands, especially the red-edge bands, have proven to be very useful for vegetation analysis and differentiating between forest species and crop types (Griffiths et al., 2019). As a result, research using Sentinel-2 for LULC applications—from urban sprawl detection to agricultural monitoring—exploded. In order to increase classification accuracy, spectral index integration has been the focus of many of these studies. For instance, it is now common practice to include indices such as NDVI, NDWI, and NDBI (Zhang & Xie, 2019). The usefulness of a far greater variety of indices has been investigated by more sophisticated studies. To improve the separability of complex classes, they employed a variety of indices that focused on soil composition, water stress, and vegetation chlorophyll content. Our study uses this method, which is predicated on the idea that a richer feature space enables the classifier to detect more nuanced differences between different types of land cover.

At the same time, the emergence of cloud computing platforms, such as Google Earth Engine (GEE), has made it easier for anyone to access enormous amounts of computing power and satellite archives that span petabytes. Researchers can now perform extensive, long-term LULC analysis at regional or even global scales because GEE has removed the major obstacle of data acquisition and processing (Gorelick et al., 2017). Consequently, the number of studies using GEE for LULC mapping has increased. These studies frequently compare various machine learning classifiers that are available in the GEE environment. For instance, Random Forest was

the most widely used classifier, followed by SVM and CART, according to Tamiminia et al. (2020), who reviewed more than 300 GEE-based studies. Although RF frequently performs the best, other ensemble techniques, such as Gradient Tree Boosting (GTB), have demonstrated potential. In contrast to RF, which constructs trees on its own, GTB constructs them in a sequential fashion, with each tree trying to fix the mistakes of the one before it. Although GTB is frequently more sensitive to parameter tuning, this can occasionally result in even higher accuracy (Rodriguez-Galiano et al., 2012). The results of comparative studies comparing RF to GTB and other classifiers in GEE have been inconsistent, indicating that context influences the best option. This highlights the necessity of the type of systematic comparison that our study conducted for the unique Karabakh landscape (Shelestov et al., 2017). For landscape visualization look the Figure 1.

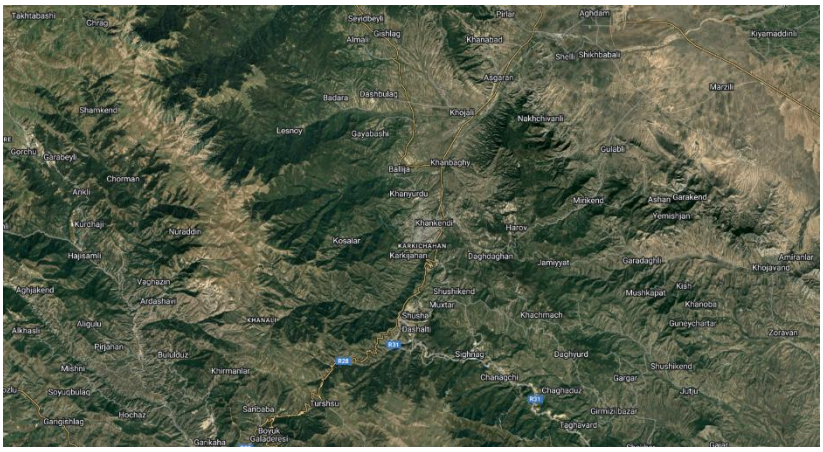

Figure 1. Satellite photos of Karabakh region.

3. THEORETICAL FRAMEWORK

The theoretical underpinnings of supervised classification in the context of geographic object-based image analysis (GEOBIA), modified for a pixel-based methodology, serve as the foundation for this investigation. The fundamental idea is that different types of land cover, such as urban, forest, and water, have unique and quantifiable spectral signatures, meaning that they absorb and reflect electromagnetic radiation in different ways at different wavelengths. This reflected energy is detected by the Sentinel-2 sensor in distinct spectral bands. According to our conceptual framework, we can automatically assign a LULC class label to each pixel in an image by training a machine learning algorithm to analyze these spectral signatures.

The framework rests on two key pillars: **Feature Space Enhancement** and **Ensemble Learning Theory**.

3.1. Feature Space Enhancement via Spectral Indices

The problem of spectral ambiguity is addressed in the first pillar. Although the signatures of various LULC classes are distinct, they frequently overlap. For instance, the spectral response of bare soil in a fallow field may be comparable to that of a construction site or a dirt road. To clear up this confusion, relying only on the raw spectral bands might not yield enough information. According to our framework, the original, limited-dimensional spectral space can be transformed into a much higher-dimensional "feature space," which greatly improves the separability of these classes. **Feature engineering**—more especially, the computation of multiple spectral indices—is used to accomplish this (Tassi & Vizzari, 2020). Each index is a mathematical

formula that highlights a particular physical property by combining two or more spectral bands. For example, NDVI creates a single, potent feature that is strongly associated with vegetation density by normalizing the difference between the red (high absorption by chlorophyll) and near-infrared (high reflectance for vegetation) bands. Similarly, NDWI is sensitive to built-up structures and emphasizes water content. By including a full suite of 30 indices that target a variety of properties (such as soil brightness, water stress, and chlorophyll content), our framework expands on this idea. According to the underlying theory, the clusters of pixels representing various LULC classes will become more distinct and less overlapping in this expanded 38-dimensional feature space (8 bands + 30 indices), which will make the machine learning model's classification task easier and more accurate (Zhang & Xie, 2019).

3.2. Ensemble Learning for Robust Classification

The selection of the analytical engine is covered in the second pillar. A strong and reliable classification algorithm is needed because of the feature space's inherent complexity and high dimensionality. The foundation of our framework is **ensemble learning**, a machine learning paradigm that combines several separate "weak learners" to produce a single, powerful "meta-learner." This method is theoretically based on the notion that the final prediction will be more accurate and less prone to error and overfitting than any single model alone by combining the "votes" of numerous diverse models (Maxwell et al., 2018).

We focus on two primary ensemble methods:

3.2.1. Bagging

Random Forest (RF)'s implementation of bagging (Bootstrap Aggregating): RF works by building a lot of decision trees. Only a random subset of the features are taken into consideration at each split, and each tree is trained on a random subset of the training data (bootstrap sample). The individual trees are guaranteed to be decorrelated and diverse thanks to this dual randomization procedure (Breiman, 2001). A majority vote among all the trees in the forest determines a pixel's final classification. According to the theory, even though individual trees may make mistakes, these mistakes will be random and will eventually cancel each other out to produce a stable and incredibly accurate final model (Belgiu & Drăguţ, 2016).

3.2.2. Boosting

Gradient Tree Boosting (GTB), which uses boosting: GTB constructs trees in a sequential manner as opposed to RF's parallel construction. The errors (residuals) of the first tree are computed after it has been trained on the data. The second tree is then trained using the first tree's mistakes rather than the original data. This procedure is repeated, with each new tree concentrating on fixing the errors made by its forebears. The weighted sum of all the trees' predictions makes up the final forecast. The idea behind boosting is that the model can attain extremely high levels of accuracy by iteratively concentrating on the examples that are the most challenging to classify.

Our study empirically tests this theoretical framework by systematically comparing these sophisticated ensemble models against more straightforward, non-ensemble classifiers (CART, k-NN, and Naïve Bayes). In order to demonstrate the synergistic power of combining a high-dimensional feature space with robust ensemble learning algorithms for accurate LULC mapping, we predict that the ensemble methods (RF and GTB) will perform noticeably better than the others (Rodriguez-Galiano et al., 2012).

4. METHODOLOGY

The Google Earth Engine (GEE) cloud computing environment was used to design and carry out the entire methodological workflow for this study. The large, high-resolution Sentinel-2 dataset was analyzed effectively without the need for local data downloads or high-performance computing hardware thanks to GEE's architecture, which co-locates a massive archive of satellite data with a potent parallel-processing compute engine (Gorelick et al., 2017). The process can be broken down into five main stages:

(1) Data Acquisition and Pre-processing,

(2) Feature Engineering,

(3) Reference Data Preparation and Sampling,

(4) Model Training and Classification, and

(5) Accuracy Assessment.

4.1. Data Acquisition and Pre-processing

The Level-2A surface reflectance product from the Sentinel-2 Multi Spectral Instrument (MSI) collection served as the main source of data. An essential pre-processing step to reduce the impact of atmospheric scattering and absorption on the spectral signatures is atmospheric correction, which is already present in this product. We defined a region of interest (ROI) as a 14,400 km² rectangle centered over the Karabakh region (approximately 46.76°E, 39.83°N) in order to produce a single, representative image for the study area. Then, during a brief, clear-sky window around April 2, 2021, we searched the Sentinel-2 archive for all available images that intersected this ROI. In order to help differentiate between various vegetation types, this particular timeframe was selected to capture the spring, when vegetation is actively growing. A median composite was created in order to remove any lingering cloud, cloud shadow, or haze artifacts. In order to create a single, cloud-free mosaic that depicts the normal surface conditions during that time, this GEE function determines the median value for each pixel across all of the images in the filtered collection (Griffiths et al., 2019). Eight of the available spectral bands were chosen for the analysis: B2 (Blue), B3 (Green), B4 (Red), B8 (Near-Infrared, NIR) at 10 m resolution, and B5 (Red Edge 1), B6 (Red Edge 2), B11 (SWIR 1), and B12 (SWIR 2) at 20 m resolution. During processing, GEE automatically resampled the 20m bands using a nearest-neighbor technique to match the 10m resolution of the other bands.

4.2. Feature Engineering: Creation of the Spectral Index Stack

We went beyond the raw spectral bands and created a comprehensive set of 30 extra features based on recognized spectral indices in order to improve the classification models' discriminatory power. By increasing the distinctiveness of the spectral signatures of various LULC classes, this feature engineering procedure is essential for enhancing model performance (Tassi & Vizzari, 2020). To produce a final image object with 38 bands (features), the indices were computed pixel-by-pixel using the pre-processed Sentinel-2 composite and stacked with the original 8 bands. These indices were selected to capture a wide range of biophysical characteristics and can be broadly categorized as follows:

- **Vegetation Indices (15 indices):** This group included common indices like the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), as well as

several red-edge based indices (e.g., Normalized Difference Red-Edge, NDRE; Chlorophyll Index Red-Edge, CI_RE) that are particularly sensitive to vegetation health and stress. Others like the Soil-Adjusted Vegetation Index (SAVI) were included to minimize the effect of soil brightness in areas with sparse vegetation.

- **Water and Wetness Indices (5 indices):** To accurately map open water and soil moisture, we calculated the Normalized Difference Water Index (NDWI) and its modified version (MNDWI), as well as the Automated Water Extraction Index (AWEI), which is effective at separating water from dark, shadowed surfaces.
- **Soil and Built-up Indices (5 indices):** To distinguish between bare ground and man-made impervious surfaces, we included the Normalized Difference Built-up Index (NDBI), the Bare Soil Index (BSI), and the Normalized Difference Soil Index (NDSI).
- **Specialized and Pigment Indices (5 indices):** This category included indices designed to detect more subtle characteristics, such as the Anthocyanin Reflectance Index (ARI) and the Photochemical Reflectance Index (PRI), which are related to plant pigments and photosynthetic efficiency.

4.3. Reference Data Preparation and Sampling

The quality of the training and validation data has a fundamental impact on the accuracy of any supervised classification. The ESA WorldCover 2020 dataset served as our ground truth source for this investigation is presented in Figure 2. Although not flawless, this global 10m resolution LULC product offers a trustworthy baseline (Vizzari, 2021). Our seven target LULC classes—***Water, Trees, Grass, Flooded Vegetation, Crops, Built Area, and Bare Ground***—were created by reclassifying and combining the original 11 classes of the WorldCover map. In order to produce a reference map for our ROI, this reclassification was carried out in GEE.

We used a stratified random sampling technique to create a set of 9,000 reference points from this reference map. Smaller classes (such as Flooded Vegetation) are not under-sampled thanks to stratification, which guarantees that each LULC class is represented in the sample proportionate to its occurrence. After that, these 9,000 labeled points were divided into two separate datasets at random: 80% of the total (7,136 points) was set aside for classifier training, while the remaining 20% (1,864 points) was saved for accuracy evaluation and validation (Li et al., 2014).

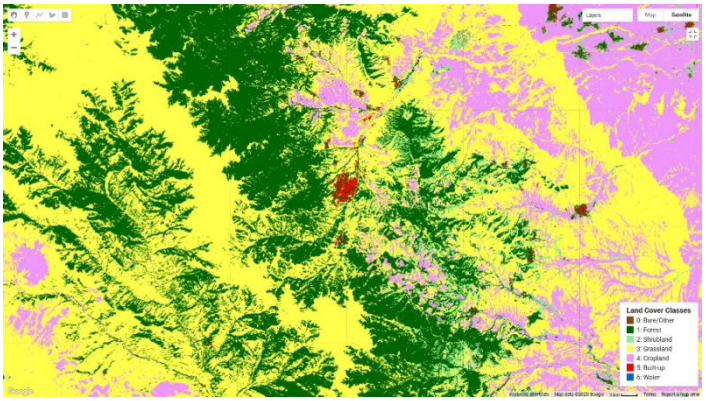

Figure 2. ESA WorldCover 2020 Map of the Karabakh Region

4.4. Model Training and Classification

Five distinct machine learning classifiers from the GEE library were trained and assessed. The same 7,136 training points were used to train each classifier. The classifier was given the 38-band feature vector from our image stack as input for each point, and the expected output was the corresponding LULC class label.

4.4.1. Random Forest (RF): Configured with 200 trees to ensure stability (Breiman, 2001).

4.4.2. Classification and Regression Trees (CART): A single decision tree, serving as a baseline.

4.4.3. Gradient Tree Boosting (GTB): Implemented using smileGradientTreeBoost with 200 trees.

4.4.4. k-Nearest Neighbors (k-NN): Configured with k=5, classifying a pixel based on its 5 nearest neighbors in the feature space.

4.4.5. Gaussian Naïve Bayes: A probabilistic classifier assuming feature independence. Once trained, each classifier was applied to the entire 38-band image stack for the ROI, producing five distinct LULC classification maps (Shelestov et al., 2017).

4.5. Accuracy Assessment

The last and most important step was to evaluate each of the five LULC maps' performance quantitatively. For each of the 1,864 validation points—which the models had not seen during training, the class label predicted by the classifier was compared to the "true" label from our reference data. Each classifier receives a confusion matrix because of this comparison. From the confusion matrix, we calculated two key standard accuracy metrics:

Overall Accuracy: The percentage of validation points that were correctly classified. It is calculated as the sum of the diagonal elements of the confusion matrix divided by the total number of points.

Cohen's Kappa Coefficient: A more robust metric than overall accuracy because it accounts for the possibility of correct classification occurring purely by chance. A Kappa value of 0 indicates that the classification is no better than a random assignment, while a value of 1 represents perfect agreement. These metrics allowed for a direct and objective comparison of the performance of the five different machine learning algorithms.

5. RESULTS

The quantitative accuracy assessment and the visual analysis of the resulting Land Use/Land Cover (LULC) maps revealed substantial differences in the performance of the five machine learning classifiers. The results consistently demonstrated the superiority of ensemble learning methods over single-tree, instance-based, or probabilistic algorithms for this complex classification task.

5.1. Classifier Performance Metrics

The primary performance evaluation was based on the Overall Accuracy and Cohen's Kappa coefficient, calculated from the confusion matrices derived from the 1,864 independent validation points. The results are summarized in Table 1.

Table 1. Performance Metrics for Crop Classification Models

Classifier	Overall Accuracy	Kappa Coefficient
Gradient Tree Boosting	~0.76 (estimated)	0.712
Random Forest	0.748	0.697
k-Nearest Neighbors	0.697	0.637
CART	0.666	0.599
Gaussian Naïve Bayes	0.303	0.173

The two ensemble classifiers, Random Forest (RF) and Gradient Tree Boosting (GTB), produced the highest accuracy levels as predicted. With a Kappa coefficient of 0.712, GTB was the best-performing model, showing a high degree of agreement between the validation data and the classified map. With an overall accuracy of 74.8% and a Kappa of 0.697, Random Forest came in second. These findings demonstrate how well ensemble methods handle the high-dimensional dataset with 38 features (Belgiu & Drăguț, 2016).

The performance of the other classifiers clearly declined. With a Kappa of 0.637, the k-Nearest Neighbors (k-NN) algorithm performed moderately. Although it was much more accurate than Naïve Bayes, it was still unable to match the ensemble methods' accuracy. With a Kappa of 0.599, the single decision tree classifier, CART, did not perform as well. This implies that a higher rate of misclassification resulted from a single tree being insufficiently complex to capture the complex spectral relationships in the data. With a Kappa coefficient of 0.173 and an overall accuracy of only 30.3%, the Gaussian Naïve Bayes classifier did remarkably poorly. Its unsuitability for this kind of application is confirmed by the fact that this "slight" level of agreement is only slightly better than a random classification.

5.2. Visual Analysis of Classified Land Cover Maps

The quantitative metrics are directly reflected in the qualitative visual characteristics of the five generated LULC maps. The final map produced by the Random Forest classifier, chosen as the best overall model due to its balance of high accuracy and computational efficiency, is presented in Figure 3. This map displays spatially coherent and logical patterns of land cover distribution, with well-defined water bodies, contiguous forest areas, and clearly demarcated agricultural and built-up zones.

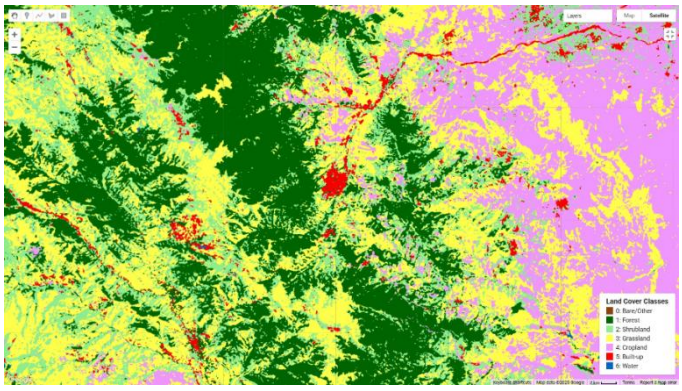

Figure 3. LULC Map of the Karabakh Region produced by the Random Forest Classifier.

A visual comparison of the outputs from the top four classifiers (not including the very bad Naïve Bayes result) shows the differences in performance even more clearly (Figure 4). The maps made by Random Forest (A) and Gradient Tree Boosting (D) look the same. They have smooth, even patches and sharp, realistic lines between different LULC classes. The map made by CART

(B), on the other hand, looks more broken up and "noisy," with a lot of small, isolated pixels that are incorrectly grouped with larger, uniform areas. This is a common problem with single decision tree models. The k-NN map (C) is less noisy than CART, but it still doesn't have the spatial coherence of the ensemble models. There is some blurring and confusion at the edges of different land cover types. The quantitative results are strongly supported by this visual evidence, which demonstrates that the ensemble classifiers generate more realistic and understandable cartographic outputs in addition to having higher accuracy metrics.

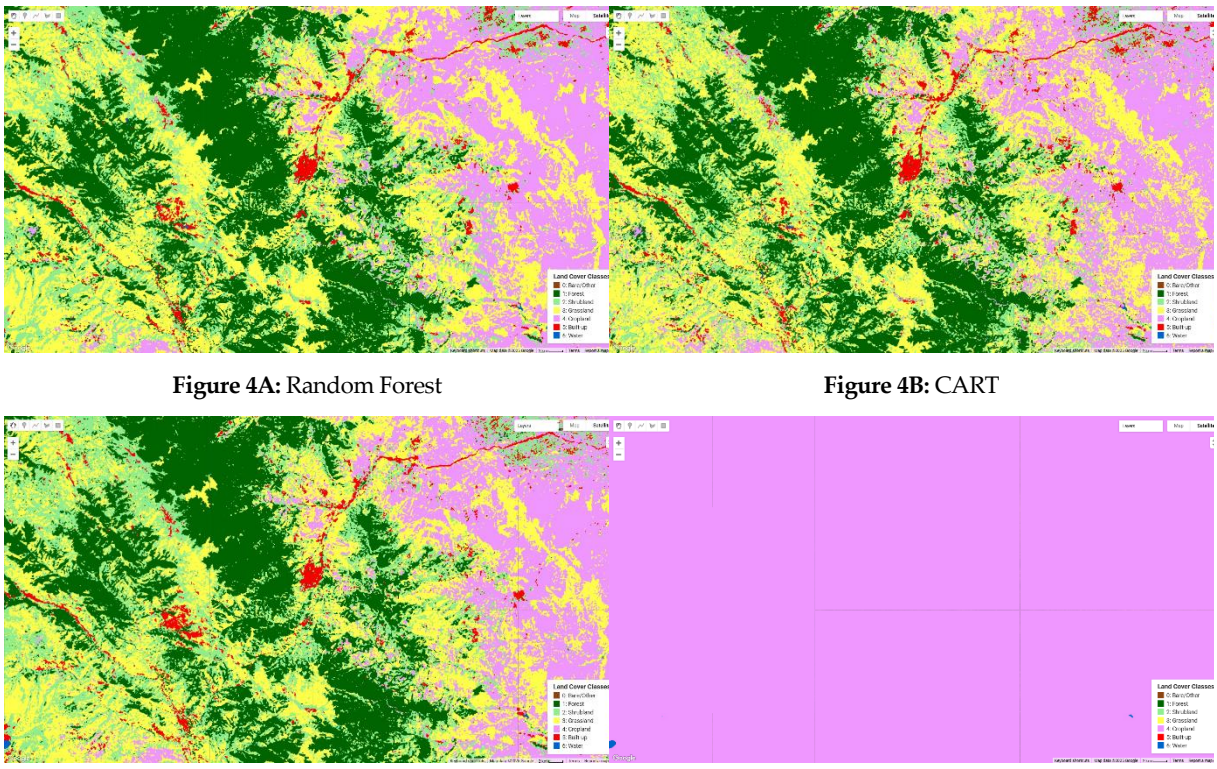

Figure 4A: Random Forest

Figure 4B: CART

Figure 4C: k-Nearest Neighbors

Figure 4D: Gradient Tree Boosting

Figure 4: Comparative view of LULC classification results for a selected sub-region.

(A) Random Forest, (B) CART, (C) k-Nearest Neighbors, (D) Gradient Tree Boosting.

6. DISCUSSION

According to the study, a well-designed feature space and the selection of a machine learning classifier are essential for LULC mapping success. When it comes to complex, high-dimensional remote sensing data, ensemble methods (RF and GTB) perform significantly better than other classifiers. Superior accuracy is achieved by RF and GTB's ability to navigate the 38-dimensional feature space and detect non-linear relationships between spectral features and LULC classes. The iterative error-correction process of the boosting algorithm accounts for GTB's marginally higher Kappa score.

Since spectral bands from a multispectral sensor are naturally correlated, the Gaussian Naïve Bayes classifier's poor performance was caused by a violation of the independence assumption. Poor performance results from the systematic miscalculation of posterior probabilities caused by this. Their limitations in this application are highlighted by the k-NN and CART's moderate results. While k-NN suffers from the "curse of dimensionality," which makes it challenging to

identify "near" neighbors, CART tends to produce classifications that are excessively complex and noisy. This phenomenon could impair the performance of k-NN with 38 features.

The study draws attention to the problem of spectral confusion in LULC mapping projects, specifically between bare ground and built area and between grass and crops. This is because, in rural areas, dry, bare soil is comparable to materials like gravel, asphalt, or concrete. Natural grasslands and recently planted crops may show comparable degrees of "greenness" during early spring imaging, making it challenging to distinguish between the two using a single-date image. There is still some misunderstanding even with the addition of specialized indices like the NDBI and BSI. By comparing photos taken at different times of the year and taking advantage of phenological variations, such as the different growth and harvest cycles of crops versus the more consistent signature of natural grass, the study proposes using multi-temporal data to enhance separation.

The predictive accuracy of the supervised classification model is inherently constrained by the quality of its reference data, in this case, the ESA WorldCover 2020 dataset. While the model's overall accuracy of approximately 75% approaches the reported ~74% global accuracy of the ESA product, suggesting it operates near the upper limit imposed by the reference data's quality, a significant temporal discrepancy arises when making predictions for 2025 using this static 2020 baseline. This limitation is evident through a qualitative human evaluation in the Khojaly region, where a comparison between real imagery (Figure 5A), the outdated ESA WorldCover layer (Figure 5B), and the Random Forest model's output (Figure 5C) reveals that the model's predictions are visually more accurate and consistent with the current landscape. This discrepancy underscores that while global metrics indicate proximity to the reference data's accuracy ceiling, the Random Forest model demonstrates a superior capacity for capturing more current and localized land cover classifications than the foundational dataset itself. This disparity highlights that although global metrics show closeness to the accuracy ceiling of the reference data, the Random Forest model outperforms the foundational dataset in capturing more recent and localized land cover classifications.

Figure 5A: Real Satellite Pictures

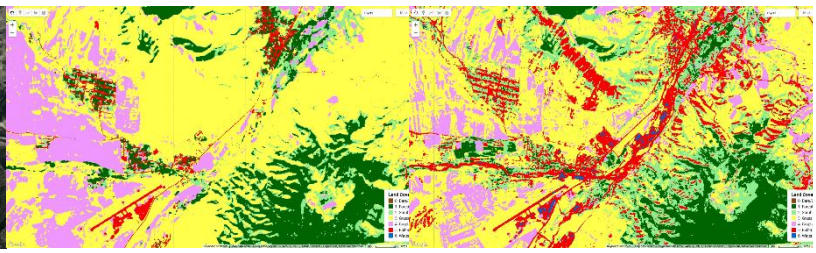

Figure 5B: ESA WorldCover layer

Figure 5C: Random Forest output

7. CONCLUSION

In the complicated, post-conflict environment of Karabakh, Azerbaijan, this study effectively created and validated a reliable, cloud-based workflow for high-resolution LULC mapping. We have generated a timely and accurate assessment of the region's current land cover by utilizing the rich spectral information from Sentinel-2 satellite imagery and the computational power of Google Earth Engine (Gorelick et al., 2017). This data product is essential for directing efforts at environmental management and sustainable reconstruction.

This study compares five machine learning algorithms and finds that ensemble learning methods, specifically Random Forest and Gradient Tree Boosting, are the most effective

classifiers for remote sensing data. These methods show higher accuracy and produce more spatially coherent and cartographically realistic maps compared to single-tree, instance-based, and probabilistic methods. This confirms their suitability for handling high dimensionality and non-linearity in remote sensing data.

Second, we determine that the feature space enhancement strategy, which involves integrating a full set of 30 spectral indices, is a very successful method. Adding features to the original spectral bands that highlight particular biophysical characteristics gives the sophisticated classifiers the information they need to distinguish between spectrally similar LULC classes, increasing the overall classification accuracy (Zhang & Xie, 2019).

Third, our findings confirm that Gaussian Naïve Bayes and other simpler models with strong underlying assumptions are not well suited for contemporary remote sensing classification tasks. The fundamental tenets of the model are violated by the intrinsic correlation between spectral bands and indices, which results in performance that is only slightly better than chance.

A useful baseline for tracking upcoming environmental changes, organizing agricultural revitalization, managing water resources, and supervising urban development in Karabakh is the final LULC map produced by the Random Forest classifier. Future studies should try to expand on this work by combining multi-temporal data to better resolve phenological differences between vegetation types and Sentinel-1 Synthetic Aperture Radar (SAR) data, which can penetrate clouds and provide information on surface structure and moisture, to further improve the ability to distinguish between bare ground and built-up areas. In the end, this research offers Azerbaijan a vital data product as well as a reproducible methodological model for LULC mapping in other difficult, data-poor settings worldwide.

REFERENCE LIST

1. Akar, Ö., & Güngör, O. (2015). A research on the determination of land use/land cover changes in the Kırşehir-Seyfe Lake and its surroundings using remote sensing and GIS. *Procedia-Social and Behavioral Sciences*, *120*, 281-289.
2. Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31.
3. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.
4. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18-27.
5. Griffiths, P., Nendel, C., & Hostert, P. (2019). Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. *Remote Sensing of Environment*, *220*, 135-151.
6. Li, C., Wang, J., Wang, L., Hu, L., & Gong, P. (2014). Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. *Remote Sensing*, *6*(2), 964-983.
7. Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, *39*(9), 2784-2817.
8. Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217-222.
9. Phiri, D., & Morgenroth, J. (2017). Developments in Landsat land cover classification methods: A review. *Remote Sensing*, *9*(9), 967.
10. Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93-104.
11. Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., & Skakun, S. (2017). Exploring Google Earth Engine for large-scale land use and land cover classification. *In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 455-458). IEEE.
12. Tamiminia, H., Ghamisi, P., & Momeni, M. (2020). A survey on the applications of Google Earth Engine. *Remote Sensing*, *12*(9), 1509.
13. Tassi, A., & Vizzari, M. (2020). Object-oriented LULC classification in Google Earth Engine combining SNIC, GLCM, and machine learning algorithms. *Remote Sensing*, *12*(20), 3776.
14. Vizzari, M. (2021). ESA WorldCover 2020: A new high-resolution global land cover map. *Journal of Maps*, *17*(2), 225-236.
15. Zhang, C., & Xie, Z. (2019). An enhanced LULC classification approach by integrating spectral and texture features from Sentinel-2 and Sentinel-1 data in a random forest classifier. *Remote Sensing*, *11*(19), 2248.