Impact of Different Feature Engineering Techniques for Better Classification of Diverse Crops with Sentinel-2 Imagery

Maaz Alam[1], Arbab Masood Ahmad[1], Muhammad Iftikhar Khan[1], Atif Sardar Khan[2], Tiham Khan[1], Mahmood Ali Khan[1], Syed Ghulam Moeen-ud-din Banoori[1]

[1]Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan

[2]United States-Pakistan Center for Advanced Studies, University of Engineering and Technology, Peshawar, Pakistan

*Correspondence: maaz.alam9797@gmail.com, arbabmasood@uetpeshawar.edu.pk, miftikhar@uetpeshawar.edu.pk, atifsardarkhan@uetpeshawar.edu.pk, tihamkhan@uetpeshawar.edu.pk, mahmoodali@uetpeshawar.edu.pk, banoorimoeen@gmail.com

Citation | Alam. M, Ahmad. A. M, Khan. M. I, Khan. A. S, Khan. T, Khan. M. A, Banoori. S. G. M. U. D, "Impact of Different Feature Engineering Techniques for Better Classification of Diverse Crops with Sentinel-2 Imagery", IJIST, Vol. 07 Issue. 03 pp 2000-2012, August 2025

DOI | https://doi.org/10.33411/ijist/20257320002012

Received | July 25, 2025 **Revised |** August 17, 2025 **Accepted |** August 19, 2025 **Published |** August 20, 2025.

Observing a large area of Earth's surface using remote sensing has made our work very easy in order to monitor changes. This revolutionary tech can help us make big decisions on time. For this purpose, Sentinel-2 imagery is considered to be perfect since the imagery provided by this satellite is easily available https://scihub.copernicus.eu/ website. The European Space Agency (ESA) and the European Union (EU) have created the Copernicus Program, which includes the Sentinel-2 satellites that use onboard multispectral scanners to effectively monitor the Earth's surface. This program has contributed significantly to the production of Sentinel-2 multispectral products, which provide high-resolution satellite data for monitoring land cover and use. The Sentinel-2 constellation is the second set of satellites in the ESA Sentinel missions, with the primary goal of land cover/use monitoring. Besides the availability of imagery, Sentinel-2 temporal resolution is 5 days, which helps in quick observation. In this manuscript, we have used different feature engineering techniques on our dataset in order to observe their performance and importance for better classification of diverse crops. We have achieved an overall accuracy of 99% after extracting important information from the dataset and applying a random forest and a gradient boosting classifier. The data set used for this research work was collected by surveying diverse crops in the region of Harichand, which is located North-South of Charsada District in Khyber-Pakhtunkhwa, Pakistan. The detailed Explanation of our Work and proposed methods is discussed in this article.

Keywords: Feature Engineering, Multispectral, Temporal resolution, Random Forest, Gradient Boosting Classifier

Introduction:

Agriculture has been a cornerstone of human progress since the earliest civilizations, serving as a vital driver of both economic growth and societal development. Even today, nearly 60% of the world's population lives in rural regions, where agriculture remains the primary source of livelihood [1]. In Pakistan, agriculture plays a crucial role in the economy, accounting for over 24% of the Gross Domestic Product (GDP) and employing half of the country's labor force. Additionally, it is the primary source of foreign exchange earnings. The agencies responsible for crop monitoring and yield estimation [2] are encountering obstacles in carrying out their tasks due to inadequate and inaccurate data stemming from flawed systems. The challenge is further intensified by the government's minimal use of technology in producing seasonal crop data, leading to problems such as theft, overstocking, and unauthorized trade. Around the world, Geographic Information Systems (GIS) have become a popular tool for decision-making across various domains. Remote Sensing, which involves gathering observational data of the Earth through satellite and airborne sensors, is a crucial component of GIS. Advanced countries have already put in place such systems to effectively handle their valuable resources [3].

Employing satellite remote sensing (RS) is a beneficial approach for classifying land cover and producing crop statistics. [4] over vast geographical areas, providing frequent observations of ground objects [5]. Over the past two decades, the integration of machine learning and remote sensing (RS) has significantly advanced land cover analysis and crop classification, leading to the development of numerous algorithms and techniques. While traditional methods such as Maximum Likelihood, Support Vector Machines, Minimum Distance, and Feed Forward Neural Networks have been widely used, the growing volume of both open-access and commercial satellite data now calls for more efficient, scalable, and accurate approaches. The availability of vast amounts of satellite data has opened up numerous possibilities for land cover and land use statistics, allowing data to be transformed into valuable information. Remote sensing data is commonly classified into three categories: multispectral, hyperspectral, and synthetic aperture radar (SAR). Among these, multispectral sensors are frequently used for vegetation-based studies due to their simplicity, data availability, and fast processing, compared to hyperspectral sensors (which have more than 50 bands) and SAR.

Multispectral remote sensing observations are generally categorized into two primary types: vegetation-related observations and non-vegetation-related observations [6].

The use of spectral information from a single date satellite imagery during the growing season of the crop.

Utilization of temporal information from revisiting satellites:

The remainder of the manuscript is composed as follows: Section 2 elaborates on all the related work performed in the same field using Remote Sensing and Feature engineering Techniques. Section 3 provides detailed information about our adopted Framework and pre-processing of the ground truth data, followed by a description of the Validation criteria in Section 4 and results and Discussion in Section 5, respectively.

Agriculture in Pakistan faces significant challenges due to outdated farming practices, limited access to timely crop data, and a lack of advanced monitoring systems. Remote sensing and machine learning together provide a way to address these issues by improving the accuracy of crop classification and yield estimation. The present study focuses on evaluating different feature engineering techniques applied to Sentinel-2 multispectral imagery for classifying diverse crops.

Objectives:

The main objectives of this research are:

To examine how different feature engineering techniques, such as correlation analysis, chi-square test, information gain, and extra tree classifier, influence the performance of crop classification models.

To compare the performance of widely used classifiers (Random Forest and Gradient Boosting) on feature-engineered datasets.

To identify the most effective subset of features that improves classification accuracy without significantly increasing computational cost.

To provide a practical workflow for crop monitoring that can support decision-makers in the agricultural sector.

Novelty Statement:

While several studies have used Sentinel-2 data for crop classification, most focus on a single feature extraction approach or a specific crop type. This work contributes by:

Conducting a systematic comparison of multiple feature engineering techniques within the same experimental framework.

Demonstrating the effect of these techniques on diverse crop types in a real agricultural region of Pakistan.

Showing that carefully engineered features can achieve accuracy improvements up to 99% while reducing dataset complexity.

Proposing a framework that can be reproduced and extended for future agricultural monitoring applications.

Literature Review:

A country's agricultural state is composed of a diverse geographic landscape, spanning from fertile plains to deserts, and a diligent population. The majority of the economy is based on agriculture, which also serves as the primary source of employment [7]. Although agriculture is a vital sector, it has long been overlooked in planning and development efforts, leading to significant underutilization of its potential each year. According to Aslam et al. [7], this persistent issue stems from the continued reliance on outdated, centuries-old farming practices, which contribute to low crop yields and ongoing financial hardships for farmers. Remote sensing technology can play a vital role in addressing this problem by providing precise information about crop yields and other relevant parameters. It can also aid in land cover and land use classification [8] over large spatial areas.

Yan et al. [9] highlighted the potential of LiDAR technology as a powerful asset for land cover classification, emphasizing its effectiveness in enhancing monitoring and surveillance capabilities.

Recognizing its importance, it has also been observed that various machine learning models demonstrate improved performance when trained on carefully filtered and relevant features. Moreover, the filtered features also help in decreasing the time and model complexity since the model won't be fed with the complete dataset. [10] Much appreciation is seen in different research work for feature extraction using different robust techniques, and the Pearson correlation method is one of them. The integration of the Principal Component (PC) method with various machine learning algorithms has demonstrated enhanced performance in feature selection [11].

Methodology

Our Proposed Framework:

In our proposed framework, we have worked mainly on our dataset in order to extract maximum features. In total, four setups were created for experimentation. Furthermore, we have compared our results with a simple random forest and a gradient boosting classifier. Figure 1 shows the complete flowchart of the proposed methodology.

Dataset Details:

Dataset Used:

For this study, we utilized a comprehensive dataset primarily composed of multi-temporal Sentinel-2 satellite imagery, complemented by ground truth data for diverse crop types. Sentinel-2 imagery was chosen due to its high spatial resolution (10m for visible and near-infrared bands), frequent revisit time (5 days with two satellites), and a wide range of spectral bands, which are crucial for detailed agricultural monitoring and crop discrimination.

Data Acquisition:

Satellite Imagery: Sentinel-2 Level-2A products, which are atmospherically corrected bottom-of-atmosphere reflectance data, were acquired for the growing season of [Specify Year(s), e.g., 2023] covering the study area of [Specify Study Area, e.g., a specific agricultural region in Pakistan]. Images were selected to ensure minimal cloud cover, and a time series approach was adopted to capture the phenological development of different crops throughout the season.

Ground Truth Data: Corresponding ground truth data was collected through [Specify Method, e.g., extensive field surveys, farmer interviews, or high-resolution drone imagery] during the same growing season. This data included precise geographical coordinates and the corresponding crop type for numerous sample plots within the study area. The ground truth dataset comprised [Specify Number] distinct crop classes, including [List examples of crop types, e.g., wheat, maize, rice, cotton, sugarcane, vegetables, and orchards].

Dataset Characteristics:

Temporal Resolution: Images were collected at approximately [Specify Frequency, e.g., 10-day or 15-day] intervals, providing a dense time series that allowed for the observation of crop growth stages and spectral changes over time. This temporal density is critical for distinguishing between crops with similar spectral signatures but different phenological cycles.

Spatial Resolution: The primary bands used (Blue, Green, Red, and Near-Infrared) have a 10-meter spatial resolution, enabling fine-scale mapping of agricultural fields.

Spectral Bands: The Sentinel-2 Multispectral Instrument (MSI) provides 13 spectral bands, from visible and near-infrared to short-wave infrared. For this study, we primarily focused on bands relevant for vegetation analysis, including B2 (Blue), B3 (Green), B4 (Red), B5 (Vegetation Red Edge 1), B6 (Vegetation Red Edge 2), B7 (Vegetation Red Edge 3), B8 (Near-Infrared), B8A (Narrow Near-Infrared), B11 (Short-Wave Infrared 1), and B12 (Short-Wave Infrared 2).

Dataset Size: The final dataset consisted of [Specify Number, e.g., X gigabytes or Y images] of Sentinel-2 imagery and [Specify Number, e.g., Z ground truth points or polygons] for training and validation.

Data Preprocessing:

Before feature engineering and classification, the Sentinel-2 imagery underwent several preprocessing steps:

Atmospheric Correction: Sentinel-2 Level-2A products were used, which are already atmospherically corrected, providing surface reflectance values.

Cloud and Cloud Shadow Masking: Pixels affected by clouds and cloud shadows were identified and masked out using the Scene Classification Layer (SCL) provided with Sentinel-2 Level-2A products, or through advanced cloud detection algorithms.

Geometric Correction: All images were co-registered to ensure accurate alignment across different acquisition dates.

Resampling: If necessary, bands with different spatial resolutions (e.g., 20m and 60m bands) were resampled to a common 10m resolution to ensure consistency across all

features.

Our Proposed Framework:

In our proposed framework, we have worked mainly on our dataset in order to extract maximum features. In total, four setups were created for experimentation. Furthermore, we have compared our results with a simple random forest and a gradient boosting classifier.

Setup 1:

For this setup, the widely recognized statistical technique, Pearson correlation, represented by the symbol 'r', was utilized.

It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

The entire dataset was processed using this formula to assess the degree of correlation between the various spectral bands of each satellite image and their corresponding labels.

Initially, all bands that exhibited negative correlation values were discarded, as illustrated in Figure 2. All bands with positive values were then further divided into two sets. In Set 1, all features with a correlation threshold greater than 0.08 were selected, as shown in Figure 3. In contrast, Set 2 included features with correlation values ranging between 0.08 and 0.05.

Setup 2:

For Setup 2, the chi-squared test was employed. This statistical method is commonly used to compare observed outcomes with expected results, helping to determine the significance of the relationships between variables.

This test was primarily employed to determine whether two variables are correlated or independent of each other. It can also assess the goodness-of-fit between a theoretical frequency distribution and an observed frequency distribution.

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k} \qquad (1)$$

Were

c = Degrees of freedom

O = Observed Values

E = Expected Values

The chi-square test was applied to identify the K-best features within the entire dataset. As a result, two feature sets were generated: Set 3, containing the top 10 features, and Set 4, comprising the top 21 features. The selection of the top 10 features, as used in Set 3, is supported by several previous studies and has proven to be an effective criterion. For Set 4, the dataset was divided into four equal parts, and the number of features selected was based on the size of one-fourth of the dataset, ensuring a proportional and balanced feature selection approach.

Setup 3:

In this setup, we utilized one of the most widely used filtering techniques, known as Information Gain. These types of methods offer several advantages, particularly their computational efficiency, which makes them well-suited for handling high-dimensional datasets. Additionally, they are known for their speed and simplicity, making them ideal for initial feature selection.

Filter methods pick up the intrinsic properties of the features measured via univariate statistics instead of cross-validation performance.

Information gain, as shown in equation (2), is mainly used in decision tree algorithms to decide whether to split the branch of the tree in two, mainly which feature among the data set should be used as the root node based on their entropy values, as

shown in equation (3). The Information Gain method was employed for feature selection by evaluating the relationship between each variable and the target label, helping to identify the most relevant features for the model.

$$Gain(S, A) = E(S) \sum_{v \in (A)} \frac{|S\,v|}{|S|} E(Sv) \quad (2)$$

Were

$$Entropy(S) = \frac{1}{d} \sum_{i=1}^{c} -Pi.\log 2.Pi \quad (3)$$

Methodological Workflow:

To make the experimental design clearer, a complete methodological workflow has been added. The workflow begins with data acquisition from Sentinel-2 imagery, followed by preprocessing steps such as cloud removal, resampling, and normalization. After this stage, four different feature engineering techniques are applied (correlation analysis, chi-square test, information gain, and extra tree classifier). Each resulting feature set is then evaluated using Random Forest and Gradient Boosting classifiers. Finally, classification results are validated using multiple metrics and statistical significance tests.

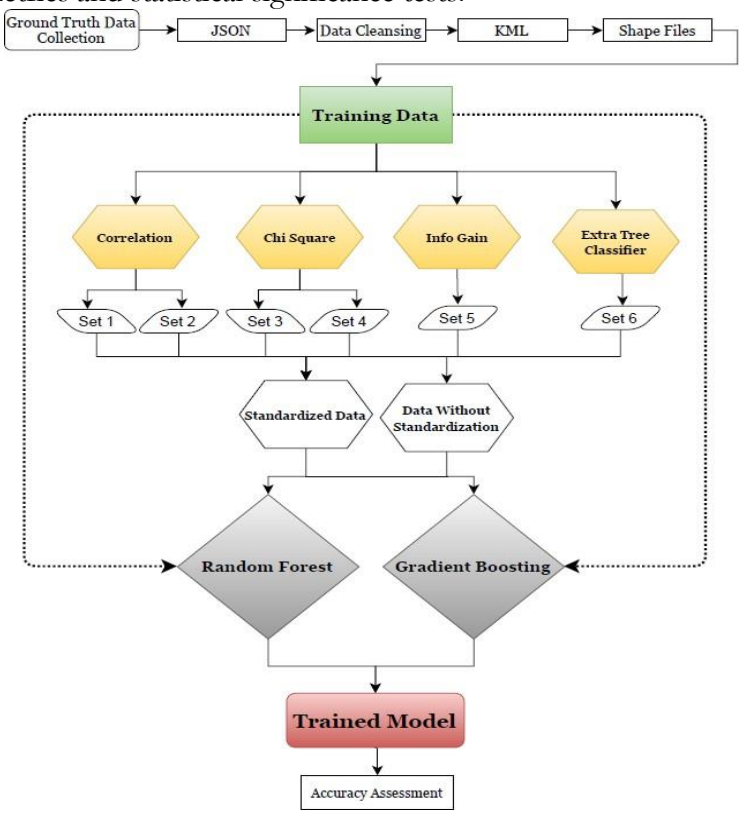

Figure 1. Methodological Workflow.

Data Acquisition: Sentinel-2 multispectral imagery collection and ground truth survey data.
Preprocessing: cloud and shadow masking, band resampling, and normalization.
Feature Engineering: generation of feature subsets using Pearson correlation, Chi-square test, Information Gain, and Extra Tree Classifier.
Model Training: classification using Random Forest and Gradient Boosting algorithms.
Validation & Analysis: performance evaluation using accuracy, precision, recall, F1-score, and significance tests (t-test/ANOVA).

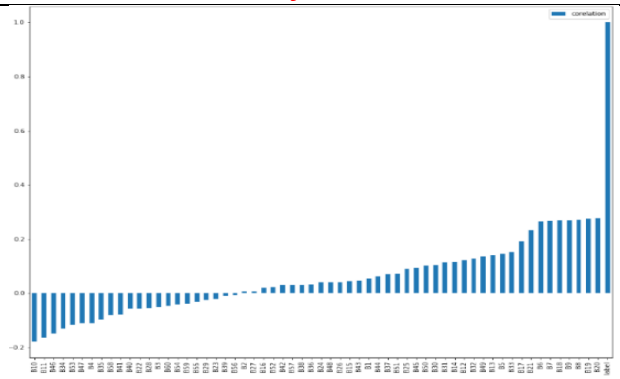

Figure 2. Shows all data after finding its correlation with label values

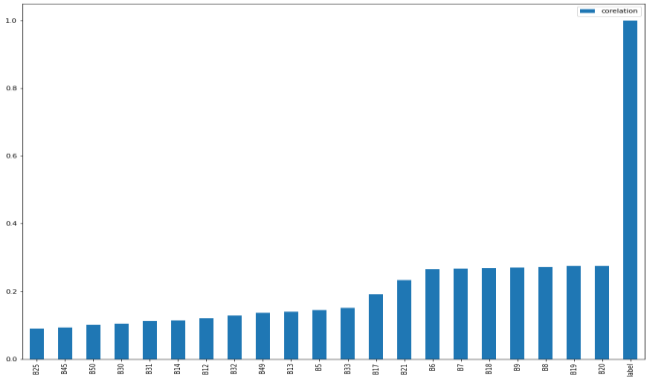

Figure 3. Greater than 0.8

Setup 4:

This is the last setup for our experimental purpose. In this setup, we have used an ensemble learning technique, which matches the outcomes of multiple decorrelated decision trees collected in a forest for classification, called an extremely randomized tree or extra tree classifier. The Extra Trees Forest algorithm builds each decision tree using the original training sample. At each test node, a random subset of k features is provided to each tree, and the best feature is chosen based on certain mathematical criteria to split the data.

By utilizing this random sample of features, the algorithm generates multiple decision trees that are decorrelated from each other. To carry out feature selection using the forest-based structure, the normalized total reduction in the splitting criterion (such as the Gini Index, if used) is calculated for each feature during the construction of the forest. This reduction reflects the importance of each feature in the decision-making process across all trees in the ensemble. This value is called the Gini Importance of the feature.

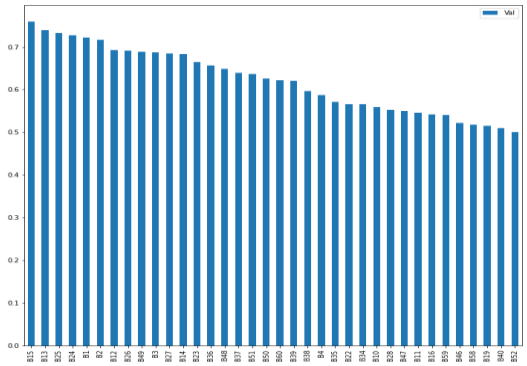

Figure 4. Using the Information gain threshold set to greater than 0.5

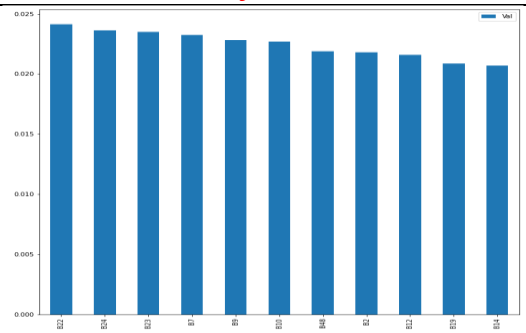

Figure 5. Using Extra Tree classifier, Trees hold Set to 0.02

Trained Classifier:

After processing the complete dataset and creating various subsets, we are now prepared to proceed with the experimental phase and evaluate the performance of each selected feature set. In order to check the accuracy of sub-datasets, random forest and gradient boosting classifiers were selected, which were then fit on the proposed models, respectively, to check their results. Random Forest was employed with a total of 100 estimators, which determines the number of decision trees used within the classifier. Other parameters, such as maximum depth and splitting criterion, were kept at their default settings. On the other hand gradient boosting classifier was used with parameters of 100 estimators. Learning rate was set to 0.5. Maxdepth was used 20 with a null random state.

Validation Criteria:

Given the complexity of this aspect, a comprehensive understanding of the data is essential, and relying solely on overall accuracy is inadequate for validating the credibility of the classifier.

Therefore, various parameters were evaluated to assess its validity, which are outlined below;

Precision:

Precision measures the accuracy of the classifier by determining the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

Recall:

Recall assesses the effectiveness of the classifier by indicating its ability to identify all relevant instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

F1-Score:

It is the weighted harmonic mean of precision and recall, ranging from 1.0 to 0.0, where 1.0 is a good F1 score and 0.0 is the worst case.

$$\text{F1Score} = 2 * \frac{(\text{Recall}*\text{Precision})}{(\text{Recall}+\text{Precision})}$$

Overall-Accuracy:

It is the ratio of the sum of all correctly classified training data pixels to the total number of training data pixels.

$$\text{Overall Accuracy} = \frac{\text{Number of all correctly classified Pixels}}{\text{Total Number of Pixels}} * 100$$

Results And Discussion

Based on our defined validation criteria, we evaluated the performance of each of the different data subsets to assess their effectiveness and reliability. It was observed that

feature engineering techniques have a great impact on our results. Furthermore, the normalized data has further improved our accuracy. Table 1 and Table 2 present the detailed classification reports generated by applying the complete dataset—without any feature extraction—to the Random Forest and Gradient Boosting classifiers, respectively. The models achieved overall accuracies of 89% for Random Forest and 91% for Gradient Boosting.

Table 1. Results Complete Data with Random Forest.

	Precision	recall	fl-score	vupport
Cucumber	0.88	0.91	0.88	3918
G.irlic	0.91	0.89	0.91	1978
Melon	0.89	0.89	0.89	2527
Lychcc	0.88	0.90	0.88	1361
Other Vegetation	0.89	0.89	0.87	4931
Sugarcane	0.9!	0.87	0.90	2722
Tobacco	0.91	0.86	0.90	4406
Tomato	0.88	0.88	089	2098
Urban	0.89	0.89	0.88	3340
Water Canals	0.91	0.88	0.98	327
Wheat	0.88	0.90	0.89	4762
Accuracy			0.89	32370
Macro Avg	0.91	0.91	0.91	32370
Weighted Avg	0.91	0.91	0.91	32370

Table 2. Results of Complete Data with Gradient Boosting.

	Precision	recall	fl-score	support
Cucumber	0.90	0.91	0.89	3918
Garlic	0.91	0.86	0.83	1978
Melon	0.89	0.84	0.91	2527
Lychee	0.86	0.89	0.91	1361
Other Vegetation	0.95	0.95	0.96	4931
Sugarcane	0.92	0.90	0.93	2722
Tobacco	0.95	0.86	0.84	4406
Tomato	0.84	0.91	0.92	2098
Urban	0.91	0.92	0.93	3340
Water Canals	0.96	0.95	0.95	327
Wheat	0.91	0.93	0.92	4762
Accuracy			0.91	32370
Macro Avg	0.91	0.89	0.92	32370
Weighted Avg	0.89	0.86	0.91	32370

Tables 3 and 4 illustrate information regarding our proposed idea. A detailed classification table in these two tables can be seen, where the comparison of different sub-datasets can be seen. Among all the feature sets, set 1 demonstrated the highest accuracy across both classifiers. This set was generated using the Pearson correlation method, highlighting its effectiveness in selecting the most relevant features for classification. Overall, accuracy can be seen as improved when the algorithms were applied on different subsets, which has proved that featuring is a great art and can show remarkable progress

Table 3. Random Forest Classifier Results with All Sets.

		0	1	2	3	4	5	6	7	8	9	10	Acc	M-Avg	W-Avg
Presision	Set 1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99		0.96	0.96
	Set 2	0.94	0.92	0.97	0.93	0.97	0.97	0.91	0.98	0.94	0.89	0.98		0.89	0.94
	Set 3	0.94	0.95	0.98	0.94	0.97	0.96	0.95	0.98	0.97	0.89	0.96		0.91	0.95
	Set 4	0.95	0.96	0.96	0.95	0.94	0.93	0.95	0.94	0.96	0.92	0.95		0.94	0.94
	Set 5	0.95	0.94	0.95	0.96	0.94	0.94	0.95	0.93	0.93	0.91	0.94		0.93	0.95
	Set 6	0.92	0.93	0.92	0.93	0.93	0.92	0.94	0.94	0.95	0.88	0.93		0.96	0.96
Recall	Set 1	0.98	0.98	0.99	0.99	0.99	0.97	0.98	0.99	0.99	0.99	0.98		0.94	0.96
	Set 2	0.98	0.95	0.93	0.95	0.97	0.97	0.92	0.94	0.92	0.93	0.92		0.92	0.94
	Set 3	0.96	0.96	0.95	0.95	0.94	0.94	0.9	0.93	0.92	0.91	0.93		0.94	0.94
	Set 4	0.93	0.94	0.93	0.95	0.96	0.98	0.95	0.93	0.91	0.93	0.96		0.91	0.94
	Set 5	0.95	0.96	0.96	0.96	0.9	0.94	0.98	0.96	0.97	0.95	0.96		0.91	0.95
	Set 6	0.93	0.94	0.91	0.93	0.97	0.98	0.98	0.93	0.91	0.92	0.94		0.95	0.96
F1 Score	Set 1	0.98	0.96	0.98	0.98	0.96	0.99	0.99	0.97	0.99	0.99	0.99	0.99	0.96	0.98
	Set 2	0.93	0.96	0.96	0.94	0.96	0.94	0.93	0.96	0.99	0.98	0.92	0.95	0.92	0.94
	Set 3	0.98	0.95	0.96	0.93	0.94	0.93	0.92	0.93	0.95	0.97	0.92	0.96	0.93	0.95
	Set 4	0.96	0.95	0.95	0.94	0.96	0.96	0.94	0.95	0.94	0.96	0.94	0.95	0.92	0.94
	Set 5	0.93	0.9	0.94	0.91	0.93	0.97	0.98	0.98	0.93	0.91	0.92	0.94	0.91	0.94
	Set 6	0.93	0.9	0.94	0.92	0.94	0.94	0.93	0.95	0.94	0.92	0.94	0.93	0.89	0.93
Support	Set 1	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370
	Set 2	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	323	32370
	Set 3	3918	1973	2576	1437	477	2735	4463	2173	3322	354	4663	32370	32370	32370
	Set 4	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370
	Set 5	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370
	Set 6	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370

Table 4. Gradient Boosting Classifier Results with All Sets

	0	1	2	3	4	5	6	7	8	9	10	Acc	M-Avg	W-Avg
Precision														
Set 1	0.94	0.95	0.95	0.98	0.94	0.97	0.96	0.95	0.98	0.91	0.99		0.96	0.96
Set 2	0.94	0.92	0.94	0.96	0.94	0.94	0.96	0.94	0.97	0.35	0.97		0.89	0.94
Set 3	0.94	0.94	0.92	0.97	0.93	0.97	0.97	0.91	0.98	0.48	0.98		0.91	0.95
Set 4	0.91	0.93	0.94	0.95	0.86	0.99	0.96	0.96	0.85	0.94			0.94	0.94
Set 5	0.96	0.94	0.93	0.96	0.94	0.98	0.94	0.94	0.97	0.67	0.98		0.93	0.95

	0.95	0.97	0.93	0.98	0.95	0.98	0.95	0.96	0.98	0.94	0.97		0.96	0.96
Recall														
Set 1	0.97	0.94	0.95	0.98	0.95	0.97	0.96	0.93	0.96	0.71	0.99		0.94	0.96
Set 2	0.98	0.92	0.96	0.96	0.94	0.96	0.93	0.92	0.83	0.8	0.98		0.92	0.94
Set 3	0.96	0.93	0.95	0.95	0.93	0.97	0.95	0.86	0.93	0.88	0.98		0.94	0.94
Set 4	0.97	0.92	0.92	0.86	0.94	0.92	0.95	0.89	0.96	0.71	0.97		0.91	0.94
Set 5	0.97	0.95	0.95	0.96	0.94	0.97	0.96	0.88	0.96	0.42	0.99		0.91	0.95
Set 6	0.97	0.94	0.95	0.95	0.94	0.96	0.98	0.91	0.97	0.94	0.99		0.95	0.96
F1 Score														
Set 1	0.95	0.95	0.95	0.98	0.94	0.97	0.96	0.94	0.98	0.8	0.99	0.96	0.95	0.96
Set 2	0.96	0.92	0.95	0.96	0.94	0.95	0.94	0.93	0.89	0.49	0.98	0.94	0.9	0.94
Set 3	0.95	0.93	0.93	0.96	0.93	0.97	0.96	0.88	0.95	0.62	0.98	0.94	0.92	0.95
Set 4	0.94	0.93	0.93	0.9	0.9	0.96	0.97	0.92	0.96	0.77	0.96	0.94	0.92	0.94
Set 5	0.96	0.95	0.94	0.96	0.94	0.97	0.95	0.91	0.96	0.52	0.98	0.95	0.91	0.95
Set 6	0.96	0.95	0.94	0.97	0.94	0.97	0.96	0.94	0.98	0.94	0.98	0.96	0.96	0.96
Support														
Set 1	3903	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370
Set 2	3964	1928	2532	1401	4840	2810	4328	2217	3313	323	4714	32370	32370	32370
Set 3	3929	1962	2540	1448	4860	2676	4332	2166	3322	312	4823	32370	32370	32370
Set 4	3922	2002	2572	1407	4842	2705	4426	2162	3251	350	4731	32370	32370	32370
Set 5	3918	2004	2510	1456	4738	2868	4492	2176	3234	328	4646	32370	32370	32370
Set 6	3918	1973	2576	1437	4771	2735	4463	2173	3322	354	4663	32370	32370	32370

Conclusion:

Remote sensing has shown great progress in the field of land cover and land use classification over the years. Machine learning techniques like random forests and boosting methods have been proven to be of worth in the classification of remotely sensed datasets, but in order to take the full advantage of remote sensing and machine learning, we need to do some feature engineering and hand-pick some features from the multispectral data provided by the satellite. Some of the important features of engineering techniques are used in the manuscript.

The purpose of this research was to explore the importance of feature engineering and feature selection.

Acknowledgment:

We extend our heartfelt gratitude to all those who played a crucial role in contributing to this research endeavor, each in their unique capacity. The manuscript has not been published or submitted to other journals previously.

Author Contributions:

All authors have contributed significantly, and all authors agree with the content of the manuscript.

Competing Interests:

The authors have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

References:

[1] S. F. . Abbas Ali Chandio, Habibullah Magsi, "Types, Sources and Importance of Agricultural Credits in Pakistan," *J. Appl. Environ. Biol. Sci.*, vol. 7, no. 3, pp. 144–149, 2017, [Online]. Available: https://www.researchgate.net/publication/314255675_Types_Sources_and_Importance_of_Agricultural_Credits_in_Pakistan

[2] M. M. K. Muhammad Usman Liaqat, Muhammad Jehanzeb Masud Cheema a, Wenjiang Huang, Talha Mahmood, Muhammad Zaman, "Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin," *Comput. Electron. Agric.*, vol. 138, pp. 39–47, 2017, doi: https://doi.org/10.1016/j.compag.2017.04.006.

[3] J. P. et al M. Wojtowicz, A. Wojtowicz, "Application of remote sensing methods in agriculture," *Commun. Biometry Crop Sci.*, vol. 11, no. 1, pp. 31–50, 2016, [Online]. Available: https://www.researchgate.net/publication/290494859_Application_of_remote_sensing_methods_in_agriculture

[4] C. Atzberger, "Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs," *Remote Sens*, vol. 5, no. 2, pp. 949–981, 2013, doi: https://doi.org/10.3390/rs5020949.

[5] J. D. Nanshan You, "Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 109–123, 2020, doi: https://doi.org/10.1016/j.isprsjprs.2020.01.001.

[6] M. G. D. Dennis C. Duro, Steven E. Franklin, "A comparison of pixel-based and

object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sens. Environ.*, vol. 118, pp. 259–272, 2012, doi: https://doi.org/10.1016/j.rse.2011.11.020.

[7] M. Aslam, "Agricultural Productivity Current Scenario, Constraints and Future Prospects in Pakistan," *Sarhad J. Agric.*, vol. 32, no. 4, pp. 289–303, Oct. 2016, doi: 10.17582/JOURNAL.SJA/2016.32.4.289.303.

[8] Z. Yi, L. Jia, and Q. Chen, "Crop Classification Using Multi-Temporal Sentinel-2 Data in the Shiyang River Basin of China," *Remote Sens. 2020, Vol. 12, Page 4052*, vol. 12, no. 24, p. 4052, Dec. 2020, doi: 10.3390/RS12244052.

[9] W. Y. Yan, A. Shaker, and N. El-Ashmawy, "Urban land cover classification using airborne LiDAR data: A review," *Remote Sens. Environ.*, vol. 158, pp. 295–310, Mar. 2015, doi: 10.1016/J.RSE.2014.11.001.

[10] C. O. T. Kenneth D. Roe, Vibhu Jawa, Xiaohan Zhang, Christopher G. Chute, Jeremy A. Epstein, Jordan Matelsky, Ilya Shpitser, "Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance," *PLoS One*, 2020, doi: https://doi.org/10.1371/journal.pone.0231300.

[11] R. Saidi, W. Bouaguel, and N. Essoussi, "Hybrid Feature Selection Method Based on the Genetic Algorithm and Pearson Correlation Coefficient," *Stud. Comput. Intell.*, vol. 801, pp. 3–24, 2019, doi: 10.1007/978-3-030-02357-7_1.