ADVANCEMENTS IN FOREST COVER TYPE CLASSIFICATION: MACHINE LEARNING APPROACHES FOR PREDICTION

Mohammed Fayazuddin,
U G Student,
Department of CSE,
St. Martin's engineering college,
Secunderabad, Telangana, India
Mohdfayazuddin86@gmail.com

Mr. J. Raja,
Assistant Professor,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India
jrajacse@smec.ac.in

Abstract- The classification of different forest cover types plays a crucial role in monitoring the environment, managing forests, and conserving biodiversity. It's essential to accurately identify and map these forest types to understand ecological patterns, assess forest health, and make well-informed decisions about sustainable land use. In the past, forest cover type classification heavily relied on manual interpretation of remote sensing data and the expertise of forestry professionals. These traditional methods were time-consuming and labour-intensive, limiting their accuracy and scalability. Experts used handcrafted rules and decision trees to differentiate forest cover types based on specific spectral or textural features extracted from satellite or aerial imagery. However, as the demand for accurate forest cover information grew, traditional methods faced challenges in handling large-scale datasets and adapting to varying forest conditions. This created a need for more efficient and automated approaches that could enhance the accuracy, scalability, and adaptability of forest cover type classification. To address these challenges, machine learning (ML) techniques emerged as a promising solution. ML offered the potential for more accurate and efficient predictions. As a result, this project explores recent developments and innovations in using ML techniques such as support vector machine classifier, and random forest classifier for forest cover type classification. By embracing these advancements, we can better support informed decision-making for sustainable forest management and environmental conservation. ML techniques have the potential to revolutionize forest cover type classification, making it faster, more accurate, and capable of handling large datasets. This, in turn, will contribute to more effective environmental monitoring, smarter forest management, and the preservation of our valuable biodiversity.

I. Introduction

Every nation relies heavily on its natural resources, and among these, forests stand out as one of the most critical. Forests are more than just vast expanses of trees; they play a pivotal role in maintaining essential geo-chemical processes and bio-climatic functions. These functions are not limited to providing oxygen and absorbing carbon dioxide; forests are also responsible for maintaining the water cycle, providing habitats for a vast variety of species, and influencing regional weather patterns. Detailed knowledge about the composition of forests can serve multiple purposes. Primarily, it aids in ensuring the health and vitality of these significant ecosystems. This information is not only crucial for environmental reasons but also has a significant impact on sectors like agriculture, which depend on predictable weather patterns and healthy soil, both influenced by the presence and condition of nearby forests. One of the fundamental metrics used to understand and measure forests is the Forest Cover (FC) type. This metric provides valuable insights into the type and density of vegetation present in a specific area. For those involved in natural resource planning, such data is invaluable. Crafting effective ecosystem governance strategies, ensuring sustainable resource usage, and anticipating environmental challenges all require detailed and accurate forest cover data. Approaches to Gathering FC-type Data: Field Collection: This method involves sending field personnel to gather data firsthand. They would visit the forested areas, conduct surveys, and collect samples to understand the composition of the forest better. However, this method can be labor-intensive, timeconsuming, and, in some cases, might be an expensive endeavor, especially if the region is vast or challenging to navigate. Remote Sensing: Advancements in technology have enabled the collection of data without being physically present in the location. Instruments like the Landsat satellite imagery, hyper spectral airborne data, radar data, and other geographic data tools can capture detailed information about an area remotely. These methods, though efficient, can sometimes also be costly, depending on the tools and technologies used. The inventory details for adjoining areas that situated beyond the natural resource 2 planner's jurisdiction are helpful, but it is economically and lawfully not feasible to gather the data. An alternate solution to this problem can be predictive models for obtaining such data. The machine learning (ML) based modeling methods can be efficiently used and incur a cheaper solution. These methods are founded either by applying statistical modeling or by using ML approaches.

II. Need of the Project

The need for advancements in forest cover type classification using machine learning arises from several factors such as: • Environmental Conservation: Accurate classification of forest cover types helps in monitoring and conserving ecosystems. It allows for the identification of areas at risk of deforestation or degradation, aiding in conservation efforts.

• **Resource Management:** Forests are sources of valuable resources like timber and biodiversity. Effective classification helps in sustainable resource management and prevents overexploitation.

• **Climate Change Analysis:** Forests play a critical role in climate regulation. Understanding changes in forest cover types contributes to climate change research.

• **Efficiency:** Machine learning models can process and analyze vast datasets more efficiently than manual methods, saving time and resources. Significance The significance of this research lies in its potential to enhance our ability to:

• **Preserve Biodiversity:** Accurate classification aids in identifying diverse forest ecosystems, allowing for targeted conservation efforts to protect endangered species.

• **Promote Sustainability:** By optimizing resource management, the project contributes to sustainable forestry practices.

• Inform Policy: The insights gained from machine learning-based forest cover classification can inform policy decisions related to land use and conservation.

• **Advance Science:** The project aligns with the broader goal of advancing environmental science and contributing to a deeper understanding of forest ecosystems.

• **Problem Definition:** The core problem addressed by this research is the accurate classification of forest cover types based on a variety of input data sources, which may include satellite imagery, 3 geographical information, and environmental variables. Specifically, the project seeks to develop and enhance machine learning models capable of classifying different forest cover types (e.g., coniferous, deciduous, mixed, etc.) accurately and providing interpretable insights into the factors influencing forest cover and changes. Additionally, this work aims to address challenges related to data preprocessing, feature selection, model optimization, and scalability, all of which are crucial for building effective and efficient forest cover classification systems. Ultimately, the success of this research is measured by the quality of its classification results, its contributions to environmental research, and its potential impact on forest conservation and management efforts.

III. Literature Survey

“Applying machine learning models to identify forest cover”- 2018 This paper presents an application of several classification techniques on forested lands data. More specifically, testing the efficiency of each classifier in its ability to identify a specific forest cover type. Four classifiers were used in the study, and testing was performed with bothunscaled data and data scaled via two different methods. The study completes successfully witha stand-out algorithm that easily exceeded its peers in the given task; the Random Forest classifier. It concludes with speculation on how this algorithm, and the study itself, can be built-upon in the future.

“Tunability: Importance of Hyperparameters of Machine Learning Algorithms”- 2018 This paper provides concise and intuitive definitions for optimal defaults of ML algorithms andthe impact of tuning them either jointly, tuning individual parameters or combinations, all based on the general concept of surrogate empirical performance models. Tunability values as defined in our framework are easily and directly interpretable as how much performance can be gainedby tuning this hyperparameter? This allows direct comparability of the tunability values acrossdifferent algorithms. The framework is based on the concept of default hyperparameter values,which can be seen both as an advantage (default values are a valuable output of the approach) and as an inconvenience (the determination of the default values is an additional analysis step and needed as a reference point for most of our measures).

“A Review on Unbalanced Data Classification”-2022 For the classification of an unbalanced data-set, different machine-learning techniques are presented by various researchers. In this paper, an attempt is made to summarize popular ML classification techniques to handle an unbalanced data set. This paper classifies the existing techniques into three groups: data level approach, algorithm level approach, and classifier’s ensemble. This paper also discusses the brief technical details, advantages and disadvantages of these methods. Finally, some of the popular unbalanced data sets available on the UCI repository are also summarized.

IV. Existing System

The existing system for forest cover type classification using machine learning approaches for prediction is typically based on the Support Vector Machine (SVM) algorithm. SVMs are a powerful machine learning algorithm that can be used to classify data into different categories. However, SVMs have some limitations, such as sensitivity to data scaling, the choice of kernel, and scalability.

Disadvantages of Existing System

• Limiting the accuracy and scalability

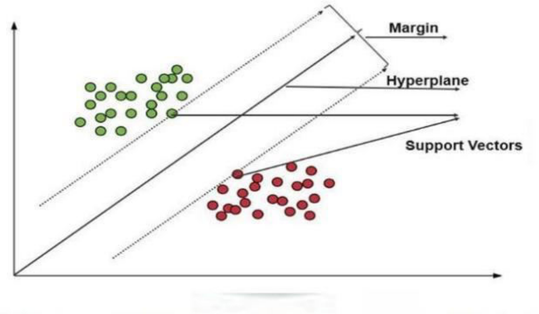
V. Proposed system

The proposed system for forest cover type classification using machine learning approaches for prediction is based on the Random Forest algorithm. Random Forests are an ensemble learning method that uses multiple decision trees to make predictions. Random Forests are less sensitive to data scaling and kernel choice than SVMs, and they can be more scalable for large datasets.

Advantages of Proposed System:

• More accuracy and less sensitive to scaling

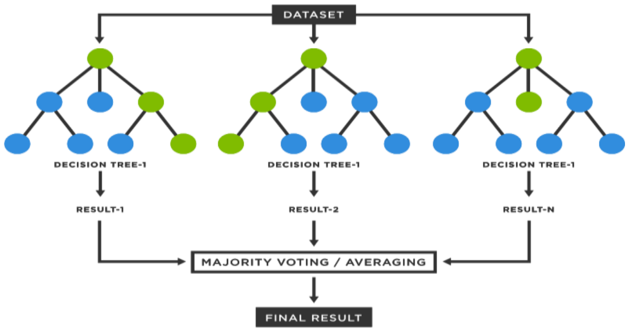
SVM Algorithm: Support vector classifier (SVC) is the most common type of support vector machine (SVM) used for classification tasks. It works by finding a hyperplane in the feature space that separatesthe data points into two classes with the maximum margin. The margin is the distance betweenthe hyperplane and the closest data points from each class, also known as support vectors. SVCis a powerful classifier that can handle both linear and non-linear data. It is also known for its robustness to noise and outliers. However, SVC can be computationally expensive to train, especially for large datasets. Consider the below diagramin which there are two different categoriesthat are classified usinga decision boundary or hyperplane:



The SVM algorithm works by finding the hyperplane in the feature space that separates the twoclasses with the maximum margin. This ensures that the SVM classifier is robust to noise and outliers. The image shows a two-dimensional example of an SVM classifier. The red and bluedata points represent the two classes that the SVM classifier istrying to separate. The black lineis the hyperplane that the SVM algorithm found. The red and blue circles are the support vectors. For example, if we have a new data point that is closer to the hyperplane than the red support vectors, the classifier will predict that the new data point is blue

Random Forest Algorithm:

The Random Forest algorithm is a powerful ensemble learning method widely used in machinelearning for both classification and regression tasks. It is based on the concept of decision treesbut offers enhanced performance, robustness, and reduced overfitting. Random Forest is an ensemble method that combines the predictions of multiple individual decision trees to make more accurate and reliable predictions. By leveraging the wisdom of the crowd, it often outperforms single decision trees. The key innovation in Random Forest is the introduction of randomness during both the data and feature selection processes. It builds multiple decision trees by randomly selecting a subset of the training data (bootstrapping) and a subset of the features for each tree. This randomness helps in decorrelating the trees and reducing overfitting. The below diagram explains the working of the Random Forest algorithm:



The classifier requires only two parameters namely the number of decision trees and the numberof prediction variables to be used in each node of the tree, to produce a prediction model. Increasing the number of decision trees leads to the convergence of generalization error and produces more reliable predictions whereas the model's accuracy rises even if the number of predictive variables is reduced. Therefore, to minimize the error, both parameters need to be optimized.

Modules

1. Upload Dataset
2. Pre-processing
3. Model selection and Training
4. Model Evaluation
5. Prediction
6. Data Visualization

Upload Dataset The first step in any machine learning project is to upload the dataset. It is important to inspectit to make sure that it is in the correct format and that there are no errors. Collected the dataset from a public cloud and divided the dataset from training and testing.

Pre-processing Data preprocessing is the process of cleaning and preparing the dataset for training. This may involve tasks such as removing outliers, handling missing values, and scaling the data. It is important to note that data preprocessing is a crucial step in the machine learning process, as the quality of the data will directly impact the performance of the model. Handling all the missing values, outliers and errors in the dataset. Normalized numerical features to bring them to a consistent scale. Considered only the features that are important in making the classification. **Model selection and Training** Once the dataset has been preprocessed, the next step is to select a machine learning model andtrain it. There are many different types of machine learning models available, each with its ownstrengths and weaknesses. The best model to choose will depend on the specific problem that you are trying to solve. Once a model has been selected, it needs to be trained on the dataset. This involves feeding themodel the data and allowing it to learn the patterns in the data. The training process can take some time, depending on the size and complexity of the dataset. 22 We selected Support Vector Machine classifier and Random Forest and trained the model with80% of data. Hyperparameters for SVM is

Linear kernel and for Random Forest the number ofdecision trees is 100 and random state is 42.

Model Evaluation Once the model has been trained, it is important to evaluate its performance on a heldout test set. This will help to ensure that the model is not overfitting the training data. If the model performs well on the test set, then it can be deployed to production. Assess the final model's performance on the test set to estimate its real-world accuracy. Calculated precision, recall, F1 score, support and confusion matrix for both the models.

Prediction Once the model has been deployed, it can be used to make predictions on new data. The modelis given new, unseen data that is 20% for testing and predict the cover type.

Data Visualization Data visualization is the process of creating visual representations of data. This can be a usefulway to explore and understand data, as well as to communicate findings to others. There are many different data visualization tools available, such as Matplotlib, Seaborn, and Tableau. Creating visualizations to explain model predictions and feature relationships.

VI. EXPERIMENTAL RESULTS

Dataset description

The dataset used in this research is related to land cover classification, specifically for forested areas. Each row in the dataset represents a location or sample within a forested region, and the dataset consists of various features (attributes) that describe these locations. Below is a description of each column in the dataset:

• **Id**: A unique identifier for each data point or location in the dataset.

• **Elevation:** The elevation of the land surface at the location. This can be considered as the height above sea level.

• **Aspect**: Aspect represents the compass direction (in degrees) that the slope of the land surface faces. It indicates which way the slope is oriented (e.g., north, south, east, west).

• **Slope**: The steepness or incline of the land surface at the location, measured in degrees.

• **Horizontal_Distance_To_Hydrology:** The horizontal distance from the location to the nearest surface water (e.g., river or stream).

• **Vertical_Distance_To_Hydrology:** The vertical distance from the location to the nearest surface water. It can be positive (above the water) or negative (below the water surface).

• **Horizontal_Distance_To_Roadways:** The horizontal distance from the location to the nearest roadway or road.

• **Hillshade_9am:** Hillshade index at 9 am, which represents the amount of shadow cast by hills and terrain at that time.

• **Hillshade_Noon:** Hillshade index at noon, indicating shadow patterns around noon.

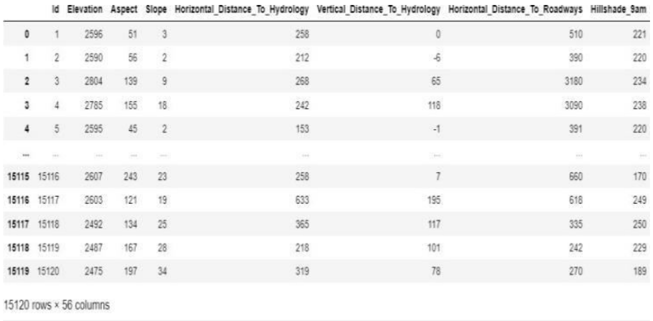
• **Hillshade_3pm:** Hillshade index at 3 pm, representing the shadow patterns in the afternoon.

• **Horizontal_Distance_To_Fire_Points:** The horizontal distance from the location to the nearest point of ignition (e.g., a fire lookout tower).

• **Wilderness_Area1 to Wilderness_Area4:** Binary indicators representing the presence 46 or absence of certain wilderness areas. Each wilderness area may have unique characteristics.

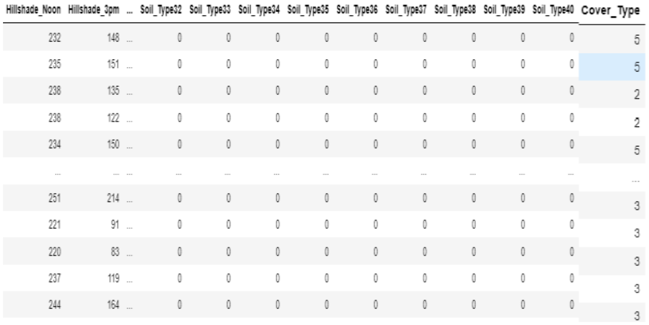
• **Soil_Type1 to Soil_Type40:** Binary indicators representing the presence or absence of specific soil types. Each soil type may have different properties, such as texture and fertility.

• **Cover_Type:** The target variable or class label, which indicates the forest cover type at the location. This is what the machine learning model aims to predict. It may have multiple classes (e.g., coniferous, deciduous, mixed, etc.).



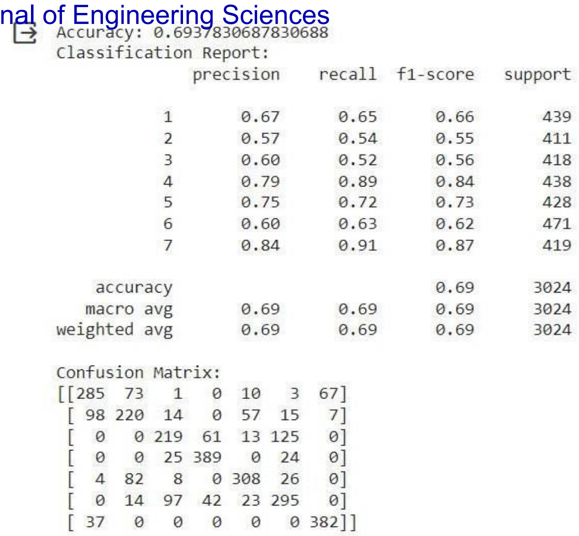
	Id	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am
0	1	2596	51	3	258	0	510	221
1	2	2590	56	2	212	-6	390	220
2	3	2804	139	9	268	65	3180	234
3	4	2785	155	18	242	118	3090	238
4	5	2595	45	2	153	-1	391	220
...	...	...	...	...	...	...	...	...
15115	15116	2607	243	23	258	7	660	170
15116	15117	2603	121	19	633	195	618	249
15117	15118	2492	134	25	365	117	335	250
15118	15119	2487	167	28	218	101	242	229
15119	15120	2475	197	34	319	78	270	189

15120 rows × 56 columns



Hillshade_Noon	Hillshade_3pm	...	Soil_Type32	Soil_Type33	Soil_Type34	Soil_Type35	Soil_Type36	Soil_Type37	Soil_Type38	Soil_Type39	Soil_Type40	Cover_Type
232	148	...	0	0	0	0	0	0	0	0	0	5
235	151	...	0	0	0	0	0	0	0	0	0	5
238	135	...	0	0	0	0	0	0	0	0	0	2
238	122	...	0	0	0	0	0	0	0	0	0	2
234	150	...	0	0	0	0	0	0	0	0	0	5
...	...	...	...	...	...	...	...	...	...	...	...	...
251	214	...	0	0	0	0	0	0	0	0	0	3
221	91	...	0	0	0	0	0	0	0	0	0	3
220	83	...	0	0	0	0	0	0	0	0	0	3
237	119	...	0	0	0	0	0	0	0	0	0	3
244	164	...	0	0	0	0	0	0	0	0	0	3

represents a visual overview or summary of the dataset used in the project. It shows that the dataset contains 15,120 rows, which correspond to individual data samples or observations. There are 56 columns in the dataset, which are used to represent various features or attributes associated with each data sample. It provides a high-level view of the dataset's structure, helping to understand its size and dimensionality.



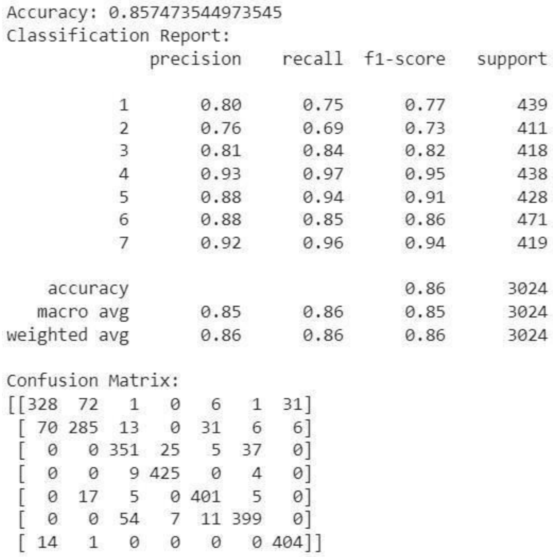
```
Accuracy: 0.6937830687830688
Classification Report:
              precision    recall  f1-score   support

           1       0.67      0.65      0.66       439
           2       0.57      0.54      0.55       411
           3       0.60      0.52      0.56       418
           4       0.79      0.89      0.84       438
           5       0.75      0.72      0.73       428
           6       0.60      0.63      0.62       471
           7       0.84      0.91      0.87       419

    accuracy                           0.69      3024
   macro avg       0.69      0.69      0.69      3024
weighted avg       0.69      0.69      0.69      3024

Confusion Matrix:
[[285  73   1   0  10   3  67]
 [ 98 220  14   0  57  15   7]
 [  0   0 219  61  13 125   0]
 [  0   0  25 389   0  24   0]
 [  4  82   8   0 308  26   0]
 [  0  14  97  42  23 295   0]
 [ 37   0   0   0   0   0 382]]
```

representation of the classification report generated after applying the SVM classifier to the dataset, which includes various performance metrics for each class, such as precision, recall, and F1-score. These metrics quantify how well the model performs in classifying each forest cover type. Displaying these metrics for each class, making it easier to assess the SVM model's performance in terms of classification accuracy and the ability to distinguish between different cover types.

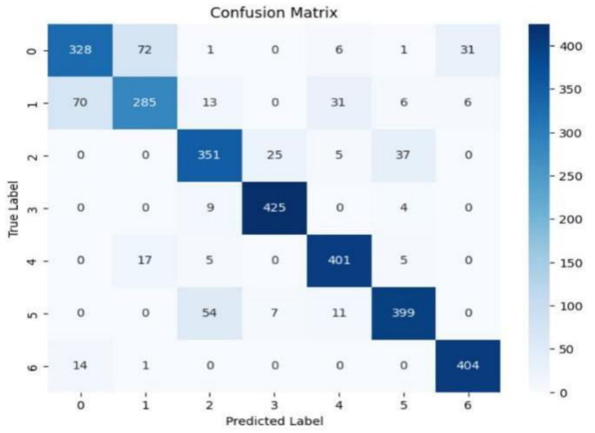


```
Accuracy: 0.8574735449735545
Classification Report:
              precision    recall  f1-score   support

           1       0.80      0.75      0.77       439
           2       0.76      0.69      0.73       411
           3       0.81      0.84      0.82       418
           4       0.93      0.97      0.95       438
           5       0.88      0.94      0.91       428
           6       0.88      0.85      0.86       471
           7       0.92      0.96      0.94       419

    accuracy                           0.86      3024
   macro avg       0.85      0.86      0.85      3024
weighted avg       0.86      0.86      0.86      3024

Confusion Matrix:
[[328  72   1   0   6   1  31]
 [ 70 285  13   0  31   6   6]
 [  0   0 351  25   5  37   0]
 [  0   0   9 425   0   4   0]
 [  0  17   5   0 401   5   0]
 [  0   0  54   7  11 399   0]
 [ 14   1   0   0   0   0 404]]
```



represents a classification report generated after applying a Random Forest classifier to the dataset. This classification report provides performance metrics for each class when using the Random Forest model for forest cover type classification. It helps assess the Random Forest model's accuracy and its ability to classify different cover types. Figure 6.4 is a representation of the confusion matrix specifically obtained from the Random Forest classifier. It visualizes the true positive, true negative, false positive, and false negative counts for each class, offering insights into how well the Random Forest model performed in predicting forest cover types.

VII. Conclusion

In the course of this machine learning project, two classification models, SVM and Random Forest, were examined for their performance on a specific dataset. The results indicate that the Random Forest classifier outperformed the SVM model in several key aspects. Notably, the Random Forest achieved a higher accuracy rate and exhibited superior precision, recall, and F1-scores across various classes when evaluated on the testing dataset. This suggests that, for the given classification task and dataset, the Random Forest model is a more effective choice. Moreover, Random Forest models tend to be more robust compared to SVMs, particularly when dealing with complex or noisy data. The Random Forest's robustness may have contributed to its better performance. Additionally, the Random Forest model is known for its computational efficiency, offering faster training and prediction times, which can be advantageous when dealing with larger datasets. Furthermore, it provides better interpretability by offering feature importance scores, aiding in understanding which features have the most influence on predictions.

VIII. Future Scope

There are several avenues to explore in order to further enhance the Random Forest model's performance and practical application. One potential direction is hyperparameter tuning, where techniques such as grid search or random search can be employed to fine-tune the model's hyperparameters. This can lead to even better predictive performance. Additionally, feature engineering can be investigated to determine whether creating new features or transforming existing ones can improve the model's accuracy. The exploration of various feature engineering techniques is a common practice in machine learning. Ensemble methods represent another promising area for improvement. Techniques like stacking or boosting can be leveraged to combine the strengths of multiple models, potentially yielding more accurate predictions.

IX. References

[1]. Kumar, A., Choudhary T. (2022) A Machine Learning Approach for the LandType Classification.In: Mekhilef S., Favorskaya M., Pandey R.K., Shaw R.N. (eds) Innovations in Electrical and ElectronicEngineering. Lecture Notes in Electrical Engineering, vol 756. Springer, Singapore.

[2]. Kumar, A. (2022). A new fitness function in genetic programming for classification of imbalanced data. Journal of Experimental & Theoretical Artificial Intelligence, 1-13.

[3]. Elaheh Arabmakki, Mehmed Kantardzic, Tegjyot Singh Sethi(2021)-Ensemble Classifier for Imbalanced Streaming Data Using Partial Labeling

[4]. Maurya, P., & Srivastava, N. (2021). Performance Evaluation of the Supervised Machine Learning Algorithms Using R. In Data Engineering and Intelligent Computing (pp. 397-406). Springer, Singapore.

[5]. Kumar, A., & Sinha, N. (2020). Classification of forest cover type using random forests algorithm. In Advances in data and information sciences (pp. 395-402). Springer, Singapore.-

[6]. Trebar, M., & Steele, N.: Application of distributed SVM architectures in classifying forestdata cover types. Computers and Electronics in Agriculture, 63(2), 119-130, (2020)

[7]. Om P Rajora,Alex Mosseler (2020)-Challenges and opportunities for conservation of forest genetic resources.

[8]. David Martin Ward (2020) Powers-Evaluation: From Precision, Recall and FFactor to ROC, Informedness, Markedness & Correlation

[9]. H. Sjöqvist, M. Längqvist and F. Javed, "An analysis of fast learning methods for classifying forest cover types", 2020 Applied Artificial Intelligence, vol. 34, no. 10, pp. 691-709.

[10]. Kumar, A., Kakkar, A., Majumdar, R., & Baghel, A. S. (2019). Spatial data mining: recent trendsand techniques. In 2019 international conference on computer and computational sciences (ICCCS) (pp.39-43). IEEE