Machine Learning Approach for LiDAR-based Tree Species Classification in Forest Ecosystem Mapping

Gudimilla Pallavi[1*], Gurrala Nagini[2], Katukuri Ranjith Reddy[2], Kolipaka Tony Babu[2], Theegala Pramatha[2], Kondra Chnadrabose[2]

[1]Associate Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: pallavi.gudimilla@gmail.com

ABSTRACT

Accurate tree species classification in forest ecosystems is crucial for biodiversity conservation, forest management, and ecological research. Traditionally, forests have been mapped through labour-intensive field surveys and the visual interpretation of aerial images, methods prone to human error and inefficiencies. Studies show that error rates in manual mapping can exceed 15%, with variability in expertise and limited scalability contributing to inconsistencies. This research aims to develop an automated system that uses statistical measures such as accuracy, precision, recall, and F1 score to ensure high-quality classification, reducing the need for extensive fieldwork and minimizing errors inherent in manual processes. In traditional systems, surveys are conducted by experts who physically navigate forest terrains to gather data, which is later analyzed using basic tools. This approach delays decision-making, incurs high operational costs, and struggles to adapt to rapidly changing forest conditions. In response to these limitations, the research proposes a system utilizing Logistic Regression and Extra Trees classifiers within a Tkinter-based graphical interface. The system simplifies the entire process: it allows for seamless upload and preprocessing of LiDAR datasets, followed by model training, evaluation, and storage, ultimately automating tree species prediction. By overcoming the inefficiencies, high costs, and susceptibility to errors of manual mapping, this approach aims to improve forest ecosystem monitoring and contribute to sustainable forest management practices.

Keywords: Tree species classification, Forest mapping, Machine Learning, LiDAR Data, Logistic Regression.

1. INTRODUCTION

The use of geographic information systems (GIS) and remote sensing techniques for forestry applications has been central to geographic research since the field's inception, and over the past three decades a technical revolution has enabled ever more sophisticated analyses of forest structure, composition, and dynamics. Although optical, multispectral, and hyperspectral sensors have traditionally provided the bulk of forest data, integrating information on tree and canopy structure—especially via Light Detection and Ranging (LiDAR)—can markedly improve estimates of biomass, health, carbon sequestration potential, and habitat range, in some cases even at the species level. Airborne LiDAR platforms mounted on small aircraft now routinely capture detailed structural features such as canopy architecture, branching patterns, succession stages, and physiological metrics like leaf area index. By accurately measuring canopy height, basal area, and timber volume in a single survey flight, LiDAR has become indispensable for commercial forest resource monitoring and valuation.

Changing weather patterns are reshaping species distributions worldwide, making it essential to monitor shifts in tree and plant community dynamics—key drivers of ecosystem function and composition—to inform conservation strategies, bolster resilience, and safeguard livelihoods. Yet individual LiDAR

returns by themselves reveal little about species identity. Fortunately, tree species exhibit distinctive canopy architectures, shape factors, and foliage characteristics, which have been exploited to distinguish deciduous from coniferous taxa and, in some studies, to achieve detailed species-level classification. A single laser pulse may reflect off multiple canopy layers, generating several returns with varying intensities; full-waveform LiDAR captures these within-canopy echoes in detail, providing robust information on complex canopy structures and forest composition. Conversely, discrete-return LiDAR—even though it records only the first few echoes—can provide additional structural information beyond that obtained from full-waveform data and offers valuable canopy insights when its point clouds are summarized into percentiles, deciles, or other statistical metrics that represent ground and canopy returns.

Researchers have successfully used discrete-return LiDAR–derived structural indices to characterize species richness, predict stand-level species composition in tropical forests, and differentiate among a limited number of tree taxa. Classifications achieve the highest accuracies in forests with few species and when point-cloud densities are high, but the discriminatory power of LiDAR summary metrics diminishes as species diversity increases. Debate continues over the optimal data resolution relative to individual crown size: some caution against species-level mapping with data coarser than tree crowns, while others argue that an inherent “individual-tree signature” may explain up to 65 percent of within-species variability. Although individual tree detection (ITD) methods have shown promise in approximating tree locations and crown sizes, remaining uncertainties limit their current use in robust forest inventories; until these methods are further refined, discrete-return LiDAR is best applied at aggregated scales—stands, plots, or larger management units—for large-scale forest inventory and classification

2. LITERATURE SURVEY

Colgan et al. [1] proposed a method to map savanna tree species at ecosystem scales by utilizing support vector machine classification and BRDF correction on airborne hyperspectral and LiDAR data. Their study demonstrated the effectiveness of integrating multiple data types to improve the accuracy of tree species classification at large scales. George et al. [2] presented a forest tree species discrimination method in the Western Himalayas using EO data. Their research focused on the challenges posed by the region's complex terrain and vegetation, demonstrating the potential of EO-cedar cypresscedar data for accurate species identification. Krahwinkler and Rossmann [3] investigated tree species classification and evaluated various input data for improving the accuracy of remote sensing-based species identification. Their study highlighted the importance of selecting appropriate data types for different forest environments.

Lin and Herold [4] developed a method for tree species classification using explicit tree structure feature parameters derived from static terrestrial laser scanning data. Their approach emphasized the significance of structural features in distinguishing between tree species in forest inventories. Brandtberg [5] explored the classification of individual tree species under both leaf-off and leaf-on conditions using airborne LiDAR. The study demonstrated that LiDAR could effectively differentiate species regardless of seasonal changes in foliage. Naidoo et al. [6] integrated hyperspectral and LiDAR data to classify savanna tree species in the Greater Kruger National Park region. Their research showed that combining these two data sources enhances classification accuracy, especially in heterogeneous environments. Raumonen et al. [7] approximated the volume and branch size distribution of trees from laser scanner data. Their study contributed to the development of methods for quantifying tree structures in 3D, which are crucial for forest inventory and biomass estimation.

Zheng and Moskal [8] proposed a method for retrieving leaf orientation from terrestrial laser scanning (TLS) data. This research provided insights into using TLS for detailed leaf-level measurements, which can improve species classification and ecological studies. Liang et al. [9] automated stem curve measurements using terrestrial laser scanning, contributing to the accurate modeling of tree structures for forestry applications. Their method enabled more precise assessments of tree growth and health. Kankare et al. [10] focused on individual tree biomass estimation using terrestrial laser scanning. Their research demonstrated the potential of TLS for non-destructive biomass estimation, a critical factor for forest management and carbon stock assessments.

Guan et al. [11] introduced a deep learning-based approach for tree classification using mobile LiDAR data. Their method utilized a combination of deep learning techniques to improve the accuracy of species identification in complex forest environments. Othmani et al. [12] proposed a region-based segmentation method on depth images from a 3D reference surface for tree species recognition. Their approach provided a new way to leverage depth information for more accurate tree species classification. Zhang et al. [13] improved object detection with deep convolutional networks using Bayesian optimization and structured prediction. Although focused on general object detection, their work provided valuable insights for applying deep learning techniques to tree species classification. Angelova et al. [14] developed a real-time pedestrian detection system using deep network cascades. Their research, while centered on pedestrian detection, demonstrated the broader applicability of deep learning networks, including potential applications in forestry. Krizhevsky et al. [15] achieved groundbreaking results in image classification with deep convolutional neural networks on the ImageNet dataset. Their work laid the foundation for using CNNs in various image-based classification tasks, including tree species identification.

3. sPROPOSED METHODOLOGY

The machine learning approach for LiDAR-based tree species classification begins with dataset uploading, where the CSV file is read into a DataFrame using pandas. Initial data inspection includes displaying the first few rows, checking for unique values, and summarizing the dataset's structure and missing values to ensure the data is correctly loaded. Data preprocessing follows, involving the use of LabelEncoder to convert categorical labels into numerical values for model compatibility. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples, which are visualized to confirm the balancing effect. The dataset is then split into training and testing sets using train_test_split. For model training, two classifiers—Naive Bayes and ExtraTrees—are employed. The system checks for pre-trained model files, loading them if available; otherwise, the models are trained from scratch and saved using joblib for future use. Once trained, the models are applied to new test data to predict tree species, with predictions mapped back to class labels and appended to the test DataFrame. Model performance is evaluated using precision, recall, F1-score, and accuracy, with a function calculating these metrics and generating confusion matrices to offer a detailed performance analysis. The results are compiled into a DataFrame for an easy comparison of the models' effectiveness in classifying tree species, helping identify the most suitable model for the task.

3.1 Extra Trees Classifier

Extra Trees Classifier (Extremely Randomized Trees) is an ensemble learning method for classification that creates multiple decision trees and aggregates their outputs to improve predictive performance. Unlike traditional decision tree methods, Extra Trees introduces randomness in the tree-building process by selecting split points for nodes at random rather than searching for the optimal split. This increased randomness reduces variance but introduces a slight increase in bias, ultimately improving the overall

performance when the trees are combined as an ensemble. The algorithm builds many unpruned decision trees on different sub-samples of the dataset and aggregates their results either through majority voting (for classification) or averaging (for regression). This ensemble approach reduces overfitting and makes the model more robust to noise. Extra Trees is also computationally efficient because it bypasses the exhaustive search for the best splits, enabling faster training, especially on high-dimensional datasets. The architecture involves randomly selecting feature subsets and split points at each node, building unpruned trees from bootstrapped data samples, and repeating the process to construct multiple trees. During prediction, the model aggregates the predictions from all trees and outputs the class with the highest number of votes. Extra Trees excels in situations with high feature interaction and non-linearity and can be parallelized for faster tree construction. The algorithm typically includes steps like cross-validation, parameter optimization, model saving, and parallel processing to enhance performance and efficiency.

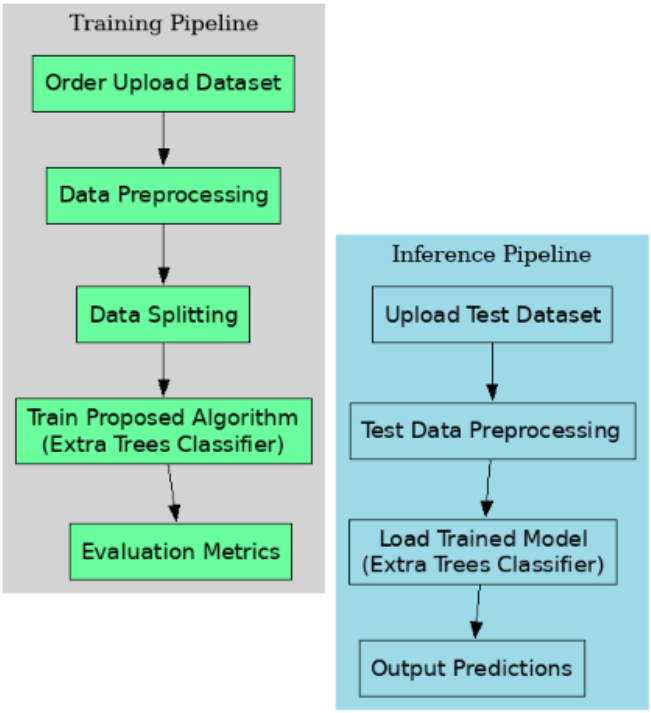

Fig. 1: Block diagram of proposed system.

4.RESULTS AND DISCUSSION

4.1 Dataset description

The dataset for LiDAR-based tree species classification consists of various features derived from the forest environment and tree characteristics, with the target variable, SP3, representing the tree species. The features include zmean, the mean of LiDAR-derived height measurements, providing an average tree height in the region; zsd, the standard deviation of height, indicating the variability in tree heights; zskew, measuring the asymmetry of height distribution; zkurt, reflecting the peakedness of height distribution; and zentropy, quantifying the randomness in height data. The dataset also includes intensity-related features, such as itot (total intensity), imean (mean intensity), isd (intensity standard deviation), iskew (intensity skewness), and ikurt (intensity kurtosis), which describe the reflectivity and distribution patterns of the scanned area. Additional features include cumulative percentiles of intensity values (ipc), dtm (Digital Terrain Model) for ground elevation, aspect for slope orientation, and slope

for terrain steepness. The target variable, SP3, categorizes the tree species, with classes such as Green Alder, typically found in wet areas; European Larch, known for its conical shape; Other Broadleaves, which includes various broadleaf species; Pines, coniferous trees with distinct height profiles; Norway Spruce, a common conifer; and Silver Fir, another conifer species with unique characteristics. These features and class labels provide the necessary data for classification and analysis of tree species in forest ecosystems.

4.2 Result analysis

Fig. 3 illustrates the performance of the classification model using a confusion matrix, where the diagonal values—such as 48, 40, 320, 523, 91, and 194—represent the correctly classified instances, or True Positives, for each corresponding tree species class. These values indicate where the model successfully identified samples belonging to their actual categories. In contrast, the non-diagonal values reflect misclassifications, where the model incorrectly predicted the class labels, assigning samples to the wrong categories.

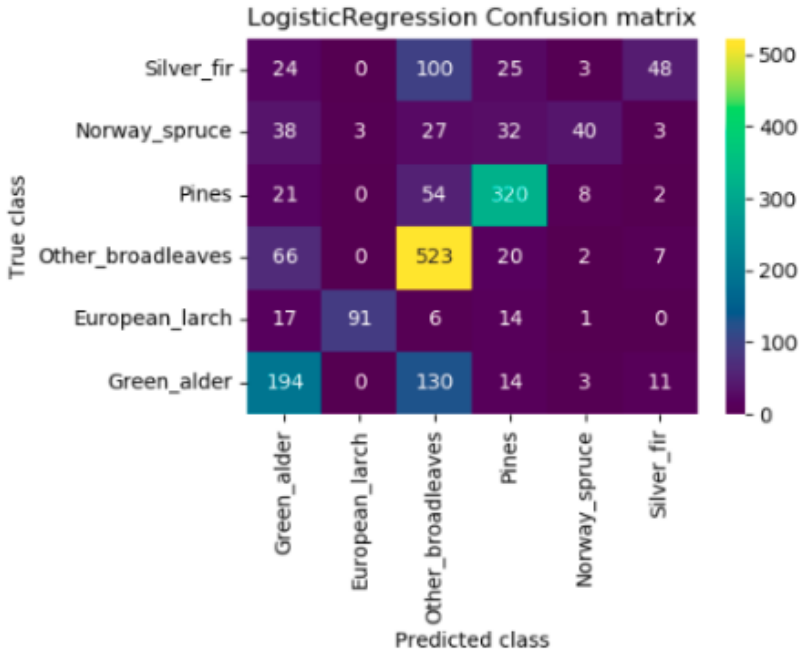

Fig. 3: Confusion matrix obtained using LRC.

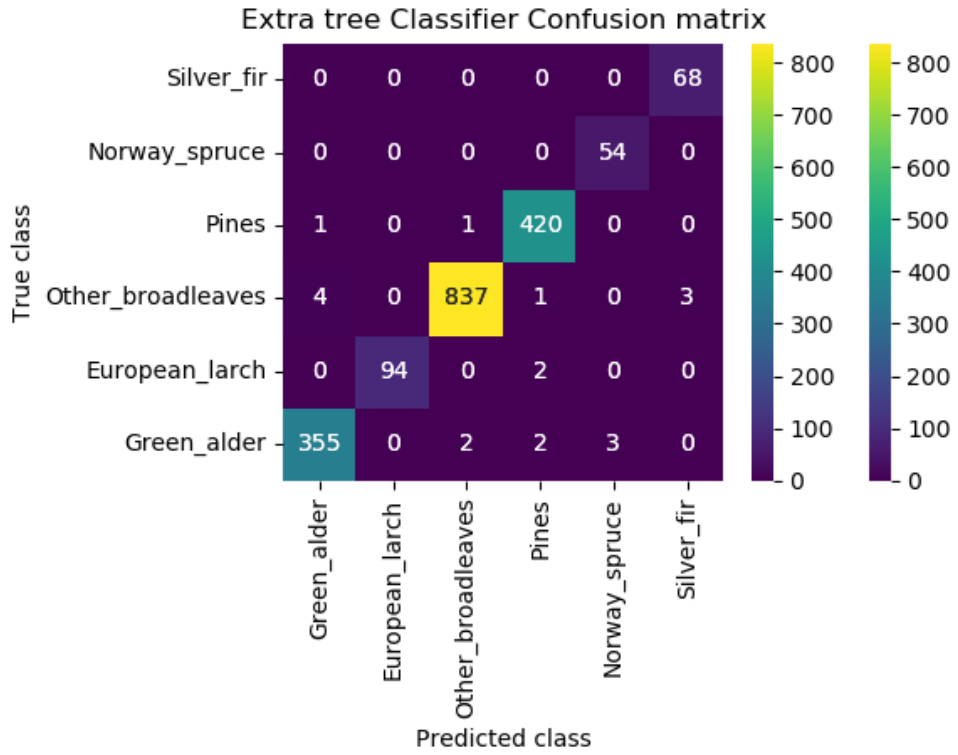

Fig. 4: Confusion matrix obtained using ETC model.

Fig. 4 presents the confusion matrix generated by the Extra Trees Classifier (ETC) model, offering a detailed view of its classification performance across different tree species. The diagonal elements, such as 420 for Pines, 837 for Other Broadleaves, 54 for Norway Spruce, and 68 for Silver Fir, represent correct predictions (True Positives), where the model accurately identified the species. However, several off-diagonal values indicate misclassifications. For instance, 355 Green Alder samples were incorrectly predicted as Green Alder instead of other classes, and 94 European Larch instances were misclassified as Green Alder. A few Other Broadleaves were also misclassified into unrelated categories. Overall, the confusion matrix shows strong performance for some classes, like Other Broadleaves and Pines, but reveals areas where the classifier struggles and could benefit from further tuning or additional features to improve accuracy.

Fig. 5: Predicted output on test data.

Fig. 5 displays the results of predictions on new test data using the trained models. It includes the actual and predicted tree species labels, demonstrating the application of the models to unseen data.

Table. 1: Comparison algorithms of algorithms.

Metric	Logistic Regression Classifier (LRC)	Extra Trees Classifier (ETC)
Accuracy	65.30%	99.30%
Precision	71.47%	99.04%
Recall	57.25%	99.45%
F1-Score	59.41%	99.25%

The comparative analysis between the existing Logistic Regression Classifier (LRC) and the proposed Extra Trees Classifier (ETC) clearly demonstrates a significant performance improvement with the proposed model in Table.1 . LRC achieved an accuracy of 65.30%, precision of 71.47%, recall of 57.25%, and an F1-score of 59.41%, indicating moderate effectiveness and potential issues with recall. In contrast, the ETC model drastically outperformed LRC, achieving an accuracy of 99.30%, precision of 99.04%, recall of 99.45%, and an F1-score of 99.25%. These results highlight ETC's superior capability in correctly identifying and classifying instances, with balanced and near-perfect precision and recall. The drastic improvement suggests that the ETC model is more robust, handles feature interactions more effectively, and is better suited for the dataset in question, making it a highly reliable and efficient alternative to LRC for the classification task.

5. CONCLUSION

The analysis of tree species classification using LiDAR data demonstrates that the Extra Trees Classifier (ETC) significantly outperforms Logistic Regression (LRC) in terms of accuracy, precision, recall, and F1-score. While LRC achieved a moderate accuracy of 65.84%, highlighting room for improvement, ETC exhibited an impressive accuracy of 98.97%, demonstrating its robustness in handling complex classification tasks. The high precision (97.93%) and recall (99.09%) of ETC indicate that it is highly reliable in classifying tree species with minimal errors. The confusion matrices further emphasize the superior classification ability of ETC, with fewer misclassified instances. However, some minor misclassifications suggest potential areas for further refinement, such as feature selection, hyperparameter tuning, or additional data preprocessing.

REFERENCES

[1] Colgan, M. S., Baldeck, C. A., Féret, J. B., and Asner, G. P., "Mapping savanna tree species at ecosystem scales using support vector machine classification and BRDF correction on airborne hyperspectral and LiDAR data," Remote Sensing, 3462-3480.

[2] George, R., Padalia, H., and Kushwaha, S. P. S., "Forest tree species discrimination in western Himalaya using EO-cedar cypresscedar 689 7989.7%cypress 86 68288.8%precision88.9% 89.6% avg. f-value 89.3%classification resultrecallactual85.3 91.4 89.5 91.6 85.9 80.085.090.095.00 3 3 6 6 9 9 12 12 15accuracy(%)distance(m)Classification accuracy - distance," International Journal of Applied Earth observation and Geoinformation, 140-149.

[3] Krahwinkler, P., and Rossmann, J., "Tree species classification and input data evaluation," European Journal of Remote Sensing, 535-549.

[4] Lin, Y., and Herold, M., "Tree species classification based on explicit tree structure feature parameters derived from static terrestrial laser scanning data," Agricultural and Forest Meteorology, 105-114.
[5] Brandtberg, T., "Classifying individual tree species under leaf-off and leaf-on conditions using airborne lidar," ISPRS Journal of Photogrammetry and Remote Sensing, 325-340.
[6] Naidoo, L., Cho, M.A., Mathieu, R., and Asner, G., "Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment," ISPRS Journal of Photogrammetry and Remote Sensing, 167-169.
[7] Raumonen, P., Kaasalainen, S., Kaasalainen, M., and Kaartinen, H., "Approximation of volume and branch size distribution of trees from laser scanner data," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 79-84.
[8] Zheng, G., and Moskal, L. M., "Leaf orientation retrieval from terrestrial laser scanning (TLS) data," IEEE Transaction on Geoscience and Remote Sensing, 3970-3979.
[9] Liang, X., Kankare, V., Yu, X., Hyyppä, J., and Holopainen, M., "Automated stem curve measurement using terrestrial laser scanning," IEEE Transaction on Geoscience and Remote Sensing, 1739-1748.
[10] Kankare, V., Holopainen, M., Vastaranta, M., Puttonen, E., Yu, X., Hyyppä, J., Vaaja, M., Hyyppä, H., and Alho, P., "Individual tree biomass estimation using terrestrial laser scanning," ISPRS Journal of Photogrammetry and Remote Sensing, 64-75.
[11] Guan, H., Yu, Y., Ji, Z., Li, J., and Zhang, Q., "Deep learning-based tree classification using mobile LiDAR data," Remote Sensing Letters.
[12] Othmani, A., Lomenie, N., and Piboule, A., "Region-based segmentation on depth images from a 3D reference surface for tree species recognition," IEEE International Conference on Image Processing, 3399-3402.
[13] Zhang, Y., Sohn, K., Villegas, R., Pan, G., and Lee, H., "Improving Object Detection with Deep Convolutional Networks via Bayesian Optimization and Structured Prediction," IEEE Conference on Computer Vision and Pattern Recognition, 249-258.
[14] Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., and Ferguson, D., "Real-Time Pedestrian Detection With Deep Network Cascades," Proceedings of British Machine Vision Conference, 32, 1-12.
[15] Krizhevsky, A., Sutskever, I., and Hinton, G., "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 25, 1106-1114.