UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Drone Aerial Image Identification of Tropical Forest Tree Species using the Mask R-CNN

Robiah Hamzah[1], Mohammad Faizuddin Md. Noor[2]
Malaysian Institute of Information Technology
Universiti Kuala Lumpur
Kuala Lumpur, Malaysia
Email: robiah@unikl.edu.my[1]* , mfaizuddin@unikl.edu.my[2]

Submitted: 26/8/2022. Revised edition: 29/10/2022. Accepted: 30/10/2022. Published online: 20/11/2022
DOI: https://doi.org/10.11113/ijic.v12n2.381

Abstract—Tropical forests have a wide variety of species and support environmental activities. The drone's image resolution is 90% more accurate than satellite data. It boosted productivity, safety, and the capacity to make better decisions by comparing archived and prospective images. Labeling tree species in heavily forested locations is labor-intensive, time-consuming, and expensive. This research seeks to design a new model for classifying tree species based on drone imagery, then test and assess its effectiveness. This study shows that drone technology can diminish productivity per hectare compared to conventional ground approaches. The study shows drones are more productive than ground approaches. The approach is feasible since it targets commercial timber species in the forest's higher stratum. Drones are cheaper than satellite data, therefore they're being used more in forest management and deep learning. Drones allow flexible, high-resolution data collection. This research uses Mask R-CNN to recognize and segment trees. This study uses high-resolution RGB images of tropical forests. The mAP, recall, and precision all performed well. Our suggested method yields a solid prediction model for detecting tree species, validated by 75% of ground truth data. This strategy can help plan and execute forest inventory, as shown. This initiative's success may lead to the first phase of a forest inventory, affecting the region's logging and forest management.

Keywords: **Mask RCNN, tree crown species detection, tropical forest, RGB drone images**

I. INTRODUCTION

Modern forest management in Malaysia is dependent on the maintenance and monitoring of resources, particularly for timber output [1]. Because remote sensing employs a large dataset that might enhance their studies, satellite data such as Light Detection and Ranging (LiDAR) have traditionally been used to identify tree species in tropical forests [2]. Remote sensing is increasingly being used to study modern forest management practices [3]. Pre-inventory data from drones was planned to be used as an alternative, low-cost remote sensing method to evaluate forest activity. Drone technology is currently the most popular because it can improve forest management planning by providing data on vital factors such as volume, tree height, and the overall condition of forest activities, and because it is a cost-effective tool based on time and cost savings as well as increased operational efficiency. Much current research such as [4], [5], [6] indicates that the tropical canopy diversity and complexity can be distinguished from tree tops using spectral and textural information from aerial images and still uses classical feature engineering methods, machine learning techniques such as Bayesian classifiers, support vector machines (SVMs), and clustering to identify tree species. Although these schemes are highly computationally efficient, their accuracy is limited in challenging conditions, such as crop variability, multiple crop detection, lighting changes, and occlusion issues, among others. The canopy of each tropical forest image often overlaps at a high ratio due to the presence of many tree species. Delineating the crown of the tropical forest is very challenging because it has a complex structure and is rich in biodiversity and species composition. Many previous studies have proposed the segmentation and classification of tree species. The method of detectree2 and Mask RCNN is used in Hickman, Ball, Jackson *et al*. [7] but the result showed it is only able to delineate a single object classification class using high-resolution satellite images and RGB images. K. Yu, Z. Hao, C. Post, *et al*. [8], described in their study, the Mask R-CNN model achieves the best accuracy of 94.68% for tree delineation (ITD) using the RGB band combination as opposed to the local maxima algorithm (LM). The drone capable of capturing photographs with a better temporal resolution is believed to facilitate and enhance fieldwork tasks. This may facilitate improved monitoring of trees of different types in the landscape. Deep learning has flourished in recent years, addressing a wide range of computer vision problems. CNNs are the most popular deep learning method, especially for vegetation remote sensing. Mask R-CNN is the newest CNN advancement for picture object instantiation. It is determined that it could be utilized efficiently in forest

management. In the case of forests, the acquired image might calculate the number of commercial trees and then assess whether a ground inventory is necessary [9], [10]. This study will offer the authority preliminary and relevant information that can be utilized to provide authorization for logging and determine the trees for logging in a particular area.

II. MATERIALS AND METHODS

A. Study Site

Fig. 1 shows the location map of the study area located at compartment 17 at the Cherul concession area of Terengganu Timber Collection Group (KPKKT) in Kuala Dungun, Terengganu. Most of the sites are dense canopy forests with mixed understory vegetation. This activity was assisted by tree-marking experts from the KPKKT team. This fieldwork covers most of the activities to get the actual data of each tree, such as the variety of tree species, and also to know how conventional methods for identifying tree species are applied.

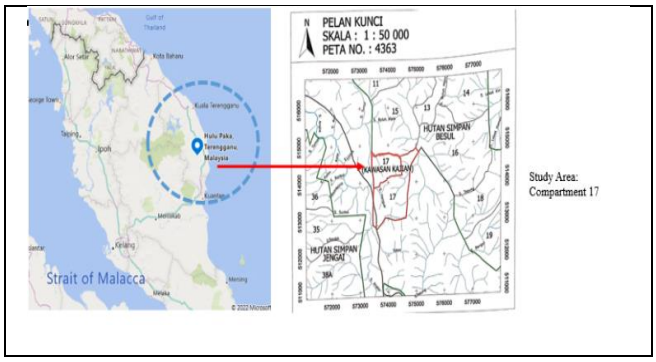

Fig.1. Location map of the study area

B. Data Collection and Pre-Processing

For data collection activity, a DJI Phantom 4 Pro drone was flown at altitudes ranging from 80 to 320 meters to capture images for use as datasets. A number of researchers are now incorporating drones into their trial evaluation programs as the technology has improved, become less expensive, and become easier to operate [11]. Table 1 summarizes the information related to drone features and image specification being used in this research. Supported by [11], a global image footprint is defined by the field of view (FOV) and altitude parameters.

TABLE I. DRONE AND IMAGE SPECIFICATION

UAV flight Parameters	Value
Camera Model	DJI FC6310
Flight pattern	Double grid
The angle of the camera	80 degree
Front overlap	90%
Side overlap	80%
Flight height	80m-320m
Image ID	Dimensions : 5472(w) x3648(h) pixels
Color representation	RGB
Focal length	9 mm
Exposure time	1/120 sec
Metering Mode	Center Weighted Average
GPS	Latitude, Longitude, and Altitude

For deep learning, narrow FOVs generate crisp images that are better suited to fine-grained details. In forestry applications where objects have a significant height or area, like tree crops, the narrower FOV reduces the geometric distortion made when rendering 3D objects in 2D. In most cases [11], [12], a smaller FOV is better for object detection. By using multispectral sensors, some interesting traits can be assessed, but the most intriguing and impactful traits are typically extracted from RGB imaging captured with high resolution. In this study, the data acquisition was conducted for covering the area. The pixel count plays a significant role in the segmentation of images. In a tropical forest, tree species can be detected and counted using a higher resolution with more pixels. High-resolution drones with cameras can fly at greater altitudes and reduce flight times. Since flying time for drones has a limitation of 30 minutes to 60 minutes per mission, there are a total of 64 images were captured during this session. New and previously inaccessible traits are now possible to measure using drones, including wheat headcount, fruit counting and classification, as well as tree species detection [13]. As shown in Fig. 2, several images were captured during the data collection process and it illustrates that one image can contain multiple trees. By labeling any suitable objects in drone images, the system is able to distinguish between selected species by identifying all contours in each image. The system calculates the total label of species based on the JSON file after species have been labeled.

Fig. 2. Example images captured from the drones

C. Instance Segmentation with Mask R CNN

The concept of instance segmentation refers to the process of identifying and locating objects of interest within an image at the pixel level. In the field of computer vision, this is one of the most challenging tasks. It is suitable to use for object detection in complex images such as medical imaging and forest applications. Mask RCNN (Regional Convolutional Neural Network) is a state-of-the-art model that addresses this challenge and it architecture is a good instance segmentation model. The process of detecting, segmenting, and categorizing every individual item contained inside an image is referred to as instance segmentation. Object

detection, which also included categorization, and semantic segmentation are the two primary components that make up instance segmentation. In other words, it merely performs object recognition as the initial step and then employs a semantic segmentation model within each rectangle's enclosing box. Instance segmentation is one step further than semantic segmentation in that, in addition to classification at the pixel level, it anticipates that the computer should characterize each occurrence of a class independently. In other words, instance segmentation is one step ahead of semantic segmentation. Based on [14], mentioned the Mask R-CNN was developed specifically to outline the precise contour of objects, which can be difficult to do in the environments such as tropical forests. There is a wide range of possible variations in the crown size and shape of tropical trees. These variations are caused by several factors, including the location of the tree within the forest canopy and the proximity of other individuals. It is believed that the region-based CNN technique (R-CNN) provides the framework for object detection, with a unique network structure that enables segmentation of objects in a logical manner [13], [8]. Faster R-CNN and Mask R-CNN architectures use the region-based CNN (R-CNN) detection technology to segment objects within images using an object identification framework [5], [15]. The two most prevalent R-CNN techniques are Faster R-CNN and Mask R-CNN. Faster R-CNN has a feature extractor as its backbone and two task-specific branches as classification and regression for classifying regions and localizing them properly. The Faster R-CNN provides high performance without specifying the size of the input image, hence improving the model's computational efficiency [16]. However, the Faster R-CNN model can only identify the object's approximate area and cannot detect the precise location of each target pixel [11].

D. Propose Research Pipeline

A pipeline depicted in Fig. 3 was proposed to generate accurate predictions. Based on the experiments, the collected data is then processed to produce a new custom data set to be used as a preprocessing pipeline to build a good model for object detection[17]. A total of 549 annotations were found on 64 images that were captured by the drones in the course of the project. Two tree species have been manually labeled as custom datasets in order to facilitate their identification (see Fig. 4). They are classified as species B and C, respectively.

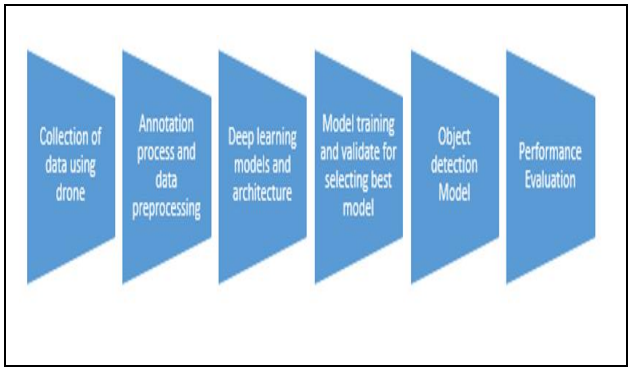

Fig. 3. Proposed pipeline

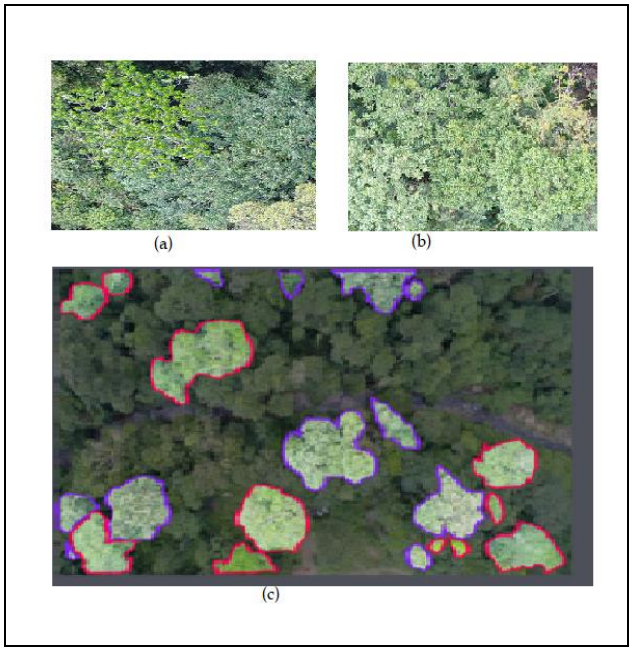

Fig. 4. Three sample classes and numbers of annotations in the dataset: (a) speciesB (234 boxes); (b) speciesC(315 boxes) and (c); Sample of training data annotation. The image annotation is done using labelMe and roboflow application

The acquired data is then processed based on the results of the tests to create a new custom data set that will be utilized as a pre-processing pipeline to construct an appropriate model for object detection. The free annotation software tool labelMe is used to quickly annotate 64 photos acquired by drones. Each annotation is a JSON-formatted file that stores the coordinates of all polygons and associated tags. It is a time-consuming process, but a correctly annotated dataset is thought to be essential for training a model with high prediction accuracy.

E. Preparing the Training, Validation, and Testing Data

After manually labeling the images using the Labelme tool, the custom dataset of field trees was randomly split into 70% for training, 20% for validation, and 10% for testing. Roboflow is a framework that helps people who work on computer vision find better ways to collect, process, and train models. Adding notes can be done in different ways. There are steps like resizing, reorienting, comparing and adding more data when pre-processing data. It makes it much easier to go from raw images to a computer vision model that has been trained and is ready to be used.

F. Transfer Learning using Pre-trained Models

There has been a recent release of Detectron2 from Facebook's AI research team, which is built on Python and designed for object detection in real-time. PyTorch is used to implement the algorithm, which is based on the mask RCNN benchmark. Models in the Detectron2 model zoo can be trained to perform a variety of tasks, including bounding box detection, instance segmentation, keypoint detection, and dense pose detection[8], [13]. This transfer learning is a

machine learning approach where a model developed for one task becomes the starting point for a model of a different task [18]. This technique works well when the available dataset is not large enough, and also the model converges faster. Therefore, we trained COCO Detection faster RCNN with transfer learning using weights from a model trained with the detectron2 model zoo dataset. The Detectron2 system allows you to plug in custom state-of-the-art computer vision technologies into your workflow. There are several ways to train a model using transfer learning. The preceding feature map's classifier, a dense layer with 1,000 neurons, was eliminated after loading the pre-trained model and discarding the last layer[8]. When a layer is discarded, the other layers become untrainable, preventing them from updating their weights. Then, a thick layer containing the types of tree species was applied. [19]. The advantage of this approach is it is able to retain all the characteristics of the detectron2 COCO model zoo that was acquired. It can be reused for the identification of tree species in the forest. The network was trained for an additional 20 epochs after the initial 1500 epochs, during which the 100th to 155th layers were marked as trainable. To fine-tune the model, the learning rate adjusted to 1/10 of the base learning rate this time. As a result, weights have to be calibrated from generic feature maps to features that are explicitly related to our dataset, and this leads to the calibration of weights.

G. Augmentation Dataset

When dealing with deep learning models, it is most often the case that they do best when they have a lot of data to work with. In general, the better the model will work, the more data we have. In the absence of enough data, the deep learning model might not be able to learn any patterns or functions from the data, so it might not be able to perform well as well as it should [20]. Using image augmentation, it is possible to make more data for elaborating a model by changing the existing data that already exists. It is the process of altering the existing data in order to create more data so that a model can be trained based on the altered data. In other words, it is the process of making the dataset used to train a deep learning model bigger than it is naturally. In this experiment, the roboflow train was utilized, which automatically trains a model regardless of GPU configuration or model architecture type. A comparison of this particular custom dataset with and without augmentation has been carried out for purposes of data analysis. A custom dataset without augmentation was used in this experiment, which has mAP performance of 13.14 %, 21.10 % precision, and 19.17 % recall. As a baseline, this performance result will be used to determine if significant improvements can be made on a larger dataset. Following this, additional tests will be conducted with the custom dataset with data augmentation. Due to the variable levels of sunlight, aerial data can be captured under a variety of conditions. Furthermore, the image may be taken very close to the camera, and sunlight changes can affect the colors of objects. On the basis of the results of the test set, Table 2 demonstrates how the proposed model enhanced performance accuracy using dataset augmentation by a significant margin compared to the original Mask R-CNN dataset. This method improves object detection by 19.17% at 50% IoU when compared to the original dataset. There is not much improvement in the condition of tropical forests due to irregular borders that are difficult to discern with the naked eye.

TABLE 2. COMPARISON BETWEEN TWO TYPES OF DATASETS

Dataset	mAP	precision	recall
Custom dataset without augmentation	13.14	21.10	19.17
Custom dataset with data augmentation	19.17	67.9	16.9

As shown by Table 2, deep learning neural network models that have been trained on more data can provide more skilled models when the author used the augmentation approach, and image augmentation techniques can produce variants of images that can increase the fit models' ability to generalize what they have learned to new images when they are trained on more data. The results indicate that data augmentation can produce better result based on accuracy in significant improvements in neural network performance. [21].

H. Metrics Prediction

In order to evaluate computer vision objects and identification models, mean average precision (mAP) is used. The mAP of our forecasting model is measured at some level of confidence, such as mAP at 0.5 or mAP at 0.95. An intersection over union (IoU) of 1.0 indicates that forecast boxes and ground truth labels are perfectly overlapping.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

where;

TP = True positive
TN =True negative
FP = False positive
FN = False negative

Precision is found by dividing the number of true positives by the number of things that were predicted to be positives. Recall, also called the true positive rate, is found by dividing the number of true positives by the number of things that should have been predicted as positives [22]. It is also capable of testing whether all positives will be located. Three outputs can be obtained from Mask R-CNN: a bounding box, a mask, and a confidence score for the projected class [23]. Using two samples, Fig. 5 shows instance segmentation resulting in an average precision (AP) averaging over all IoU levels from 0.5 to 0.95. The step size for this segmentation is 0.05, and the precision is 50 for the two classes Species B and C at IoU 0:5. The bounding boxes show the confidence levels (%).

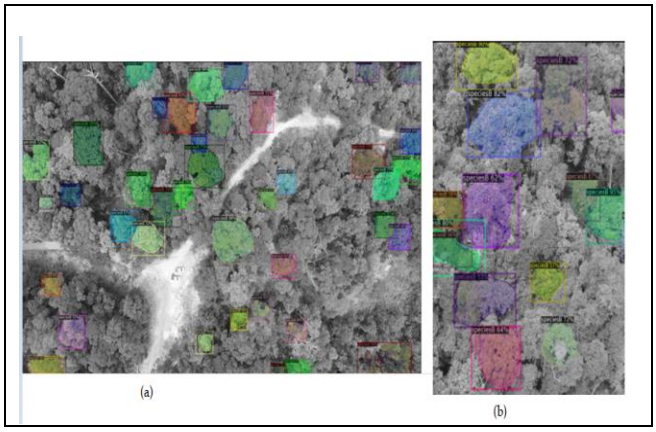

Fig. 5. (a) A full visual scene sample taken from the collection of tree species, tested using Mask RCNN RestNet50-FPN as the backbone in Detectron 2. (b) The zoom of the output, along with the box and mask prediction, as well as the confidence levels percentages, are displayed on the bounding boxes as the output.

It is important to note that the multi-task loss function of Mask R-CNN combines the loss of classification, localization, and segmentation mask.

$$-\frac{1}{m^2}\sum_{1<i,j\leq m}\left[yij\log y_{ij}^{k}+\left(1-y_{ij}\right)\log\left(1-y_{ij}^{k}\right)\right]$$

where;

$$L = Lcls + Lbox + Lmask \quad (1)$$

Lcls and Lbox are the same as in Faster R-CNN. Mask branches generate masks of dimension m x m for each ROI and class; K classes are generated in total. For this reason, there exists no competition among classes when it comes to generating masks since the model attempts to learn a mask for each class. In this example, Lmask represents the average cross-entropy loss for each region, including only the k-th mask if the region is associated with the ground truth class k. Yij can be obtained by multiplying the label of a cell (i, j) in the true mask for the region of size m x m by the predicted value of the same cell in the mask learned for the class k which is the ground truth.

IV DISCUSSION

It is usually the accuracy of a method or approach that is used as a metric or comparison when presenting new ways, as this indicator tends to reflect the theoretical potential that may be discovered with the approach as a whole. There are some distinct approaches to determining how accurate a measurement is. In this investigation, the term precision refers to an estimation of how accurate your predictions are as well as how many of them were correct, whilst recall refers to a measurement of how many of the correct predictions you made. Training a computer vision model is the process by which a computer learns to detect specific objects within images. The mAP is the measure of our model's ability to correctly forecast bounding boxes at some confidence level of mAP@0.5 or mAP@0.95. If a forecast box perfectly overlaps a ground truth label, the intersection over union (IoU) value is 1. Models with a confidence feature can exchange precision for recall by altering their prediction confidence. Suppose false positives are more important than false negatives. In that instance, the model's confidence criterion might be raised to encourage high-precision forecasts at the cost of coverage. In this experiment, a Mask R-CNN model was trained on a drone image that was obtained in pure RGB format from a drone to simulate a range of appearances. To acquire the ground truth and imagery for this study, UAVs were used. It is focused on recognizing tree species in tropical forest areas using mask RCNN. This mask was created for each cluster of species in the image as part of the labeling process. According to this mask, pixels that were labeled species B and species C would always be referred to as these species. As opposed to labeled pixels, background pixels are pixels that are not labeled. Based on the Mask RCNN method, this study demonstrated that it is possible to detect and segment tree species from drone images using the Mask RCNN method. It is evident from the dataset created, the model trained, and the results provided. According to these results, the Mask RCNN method is capable of detecting and segmenting tree species in drone images. At the moment, there are a number of constraints to consider in order to create a future research project, for example, the limitations of assumptions, as well as the necessity to research the scale effect as well as the overlapping tree, since these factors are common trends in tree identification.

V CONCLUSION

As explained in the result analysis, the fact that drone technology can provide a higher yield than conventional ground methods indicates that drone technology can greatly reduce your cost per hectare. Modern deep learning algorithms have demonstrated efficacy across various applications of computer vision, with deep learning techniques now reaching state-of-the-art performance in both conventional benchmarks for object detection and feature representation. Later, multiple methods in the deep learning model will be investigated for the segmentation and identification of crown images, as well as their species prediction, in order to achieve a more precise result. The project is expected to acquire more images of Malaysian forests and their fauna. Diverse species inhabit Malaysia, and the performance of identification methods for these species could be evaluated. The success of this initiative could have an impact on the sustainable management of tropical forests. The technology can be used to support logging and sustainable forest management at different management levels. It will aid forest authorities such as the Forest Service and timber contractors.

ACKNOWLEDGMENT

Special appreciation to Universiti Kuala Lumpur (UNIKL) and Kumpulan Pengurus Kayu Kayan Terengganu (KPKKT) for the image data set analysis and useful information.

REFERENCES

[1] Z. Roslan, Z. A. Long, M. N. Husen, R. Ismail, and R. Hamzah. (2020). Deep Learning for Tree Crown Detection in Tropical Forest. *Proc. 2020 14th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2020*.

[2] M. S. Iqbal, H. Ali, S. N. Tran, and T. Iqbal. (2021). Coconut Trees Detection and Segmentation in Aerial Imagery using Mask Region-based convolution Neural Network. *IET Comput. Vis.*, 15(6), 428-439.

[3] Y. Yahya and R. Ismail. (2020). Tree-mapping Technique as a Computer System for Sustainable Forest Management. *Proc. 2020 14th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2020*.

[4] M. Afizzul Misman, H. Omar, S. Yasmin Yaakub, N. Hajar Zamah Shari, N. Ayop, and A. Amira Anuar Musadad. (2021). UAV-Based Hyperspectral Imaging System for Tree Species Identification in Tropical Forest of Malaysia. *J. Adv. Geospatial Sci. Technol.*, 1(1), 145-162.

[5] A. M. M. Kamarulzaman, W. S. W. M. Jaafar, K. N. A. Maulud, S. N. M. Saad, H. Omar, and M. Mohan. (2022). Integrated Segmentation Approach with Machine Learning Classifier in Detecting and Mapping Post Selective Logging Impacts Using UAV Imagery. *Forests*, 13(1).

[6] J. Jamal *et al.* (2022). Dominant Tree Species Classification using Remote Sensing Data and Object-based Image Analysis. *IOP Conf. Ser. Earth Environ. Sci.*, 1019(1), 012018.

[7] S. H. M. Hickman *et al.* (2022). Accurate Tropical Forest Individual Tree Crown Delineation from RGB Imagery using Mask R-CNN.

[8] K. Yu *et al.* (2022). Comparison of Classical Methods and Mask R-CNN for Automatic Tree Detection and Mapping Using UAV Imagery. *Remote Sens.*, 14(2).

[9] M. G. Hethcoat, D. P. Edwards, J. M. B. Carreiras, R. G. Bryant, F. M. França, and S. Quegan. (2019). A Machine Learning Approach to Map Tropical Selective Logging. *Remote Sens. Environ.*, 221(December 2018), 569-582.

[10] F. Ecology and W. Khokthong. (2019). Drone-based Assessments of Crowns , Canopy Cover and Land Use Types in and Around.

[11] S. Bhatnagar, L. Gill, and B. Ghosh. (2020). Drone Image Segmentation using Machine and Deep Learning for Mapping Raised Bog Vegetation Communities. *Remote Sens.*, 12(16).

[12] C. E. Waite, G. M. F. van der Heijden, R. Field, and D. S. Boyd. (2019). A View from Above: Unmanned Aerial Vehicles (UAVs) Provide a New Tool for Assessing Liana Infestation in Tropical Forest Canopies. *J. Appl. Ecol.*, 56(4), 902-912.

[13] M. Yang *et al.* (2022). Detecting and Mapping Tree Crowns based on Convolutional Neural Network and Google Earth Images. *Int. J. Appl. Earth Obs. Geoinf.*, 108(August 2021), 102764.

[14] M. P. Ferreira *et al*. (2020). Individual Tree Detection and Species Classification of Amazonian Palms using UAV Images and Deep Learning. *For. Ecol. Manage.*, 475.

[15] A. Gromova. (2021). Weed Detection in UAV Images of Cereal Crops with Instance Segmentation.

[16] M. Minakshi. (2018). A Machine Learning Framework to Classify Mosquito Species from Smart-phone Images. *ProQuest Diss. Theses*, June, 43.

[17] G. A. Fricker, J. D. Ventura, J. A. Wolf, M. P. North, F. W. Davis, and J. Franklin. (2019). A Convolutional Neural Network Classifier Identifies Tree Species in Mixed-conifer Forest from Hyperspectral Imagery. *Remote Sens.*, 11(19).

[18] Y. Diez, S. Kentsch, M. Fukuda, M. L. L. Caceres, K. Moritake, and M. Cabezas. (2021). Deep Learning in Forestry using Uav-Acquired Rgb Data: A Practical Review. *Remote Sens.*, 13(14), 1-43.

[19] K. N. Tahar *et al.* (2021). Explainable Identification and Mapping of Trees using UAV RGB Image and Deep Learning. *Sci. Rep.*, 11(1), 1-15.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12), 2481-2495.

[21] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. Da Silva. (2021). A Comparative Analysis of Object Detection Metrics with a Companion Open-source Toolkit. *Electron.*, 10(3), 1-28.

[22] S. Alkema. (2019). Aerial Plant Recognition Through Machine Learning, July, 1-40.

[23] M. Akta and H. F. Ate. (2021). Small Object Detection and Tracking from Aerial Imagery, 118.