UDC 004
DOI: 10.15587/1729-4061.2022.265807

IMPLEMENTATION OF DEEP LEARNING BASED SEMANTIC SEGMENTATION METHOD TO DETERMINE VEGETATION DENSITY

Yuslena Sari
Master of Computer Science, Head of Department
Department of Information Technology*
Yudi Arifin
Doctor of Silviculture and Forest Ecology, Professor, Vice Rector for Planning, Cooperation and Public Relations
Department of Forestry*
Novitasari
Doctor of Civil Engineering, Head of Hydraulic Laboratory
Department of Civil Engineering*
Mohammad Faisal
Corresponding author
Doctor of Bioinformatic, Secretary of Department
Department of Computer Science*
E-mail: reza.faisal@ulm.ac.id
*Universitas Lambung Mangkurat
Jl. Hasan Basry, Kayutangi, Banjarmasin, Indonesia, 70123

The dryness of peatlands is influenced by the density of vegetation. If peatlands are dry, they become vulnerable to a fire risk. To calculate the drought index, professionals must conduct a vegetation density analysis. However, field analysis requires vast amounts of resources. Moreover, the accuracy of the analysis based on satellite data is not adequate. Therefore, this research presents drone-captured two-dimensional image data. The object of this research is The Liang Anggang Protection Forest Block I in Banjarbaru, South Kalimantan, Indonesia. It is surveyed for information on its vegetation cover. Afterwards, There are 300 images of vegetation cover collected and utilized in total. The method of deep learning with semantic segmentation will be used to compare the results of determining methods with expert results as ground truth. The contribution of this study is to determine the optimal performance of deep learning model used for classifying vegetation density into three categories: bare/ungrazed, lightly grazed, and heavily grazed. Performance is evaluated based on correctness and intersection over union (IoU). Obtaining the proper parameters for the classification model using deep learning techniques and comparing the results of the best segmentation model are the objectives of the following contribution. From experimental studies conducted, the optimal momentum parameter value for MobileNetV2, Xception, and Inception-ResNet-v2 is 0.9, and the optimal accuracy performance is 82.69 percent on average. The most appropriate momentum for ResNet 18 architecture is 0.1. The result of semantic segmentation using the DeepLabV3 model with Inception-ResNet-v2 architecture is the optimal model for estimating vegetation density compared to U-Net model

Keywords: vegetation density, deep learning, semantic segmentation, classification model, two-dimensional image data

Received date 22.07.2022
Accepted date 27.09.2022
Published date 30.10.2022

***How to Cite**: Sari, Y., Arifin, Y., Novitasari, Faisal, M. (2022). Implementation of deep learning based semantic segmentation method to determine vegetation density. Eastern-European Journal of Enterprise Technologies, 5 (2 (119)), 42–54. doi: https://doi.org/10.15587/1729-4061.2022.265807*

1. Introduction

The state of Indonesia has 149,056 km^2 of peatland located on three islands such as Kalimantan, Sumatra and Papua [1]. Peatlands in Indonesia are the second largest of all peatlands in the world [2]. The land provides many important services to local communities, including maintaining air and water quality, and supporting fish populations. But apart from this, the rate of forest loss in Indonesia is very large [3]. This is due to agricultural expansion such as oil palm, pulpwood, as well as timber harvesting, mining and fires. Whether deliberate or not, human action is the main cause of fires on peatlands and caused many operations to be disrupted in Indonesia as well as nearby nations [4, 5]. Additionally, the characteristics of the current drought have an impact on the circumstances that allow peatland fires to start [6].

Drought index can be calculated using the Keetch and Byram Drought Index models, which are based on ground water level and vegetation cover [7, 8]. Numerous researchers have concentrated on quantifying vegetation density; their measurements are based on wetlands [9], peatlands [10], and common land regions [11]. Nearly all of this research uses machine learning techniques that combine statistical features, NDVI's temporal features, near-infrared (B8) and red (B4) bands, and spatial data from the NDVI, where NVDI is abbreviation of Normalized Difference Vegetation Index, is a most well-known index to detect vegetation and their condition in an area by using bands of remote sensing data. However, it is challenging to get data from satellites.

Therefore, this research proposes a novel method that uses camera sensors flown by drones to cover the needed area to close this research gap. By using deep learning-based computer vision technology to segment the image. Deep learning uses data from automatically extracted images and does not require particular features; instead, the program learns from its errors. With the convolution neural network (CNN) technique, researchers have used hyper- and

multispectral images to identify satellite image data [12]. Also carried out in [13], which applies a deep learning method for time series classification that is based on the recurrent neural network (RNN) architectural model Long Short-Term Memory (LSTM), where LSTM is a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

Due to technological advancements that enable electronic devices to monitor environmental conditions, this research is primarily necessary for modern states. As a result, this research can be used to issue early warnings in regions with dry characteristics to prevent fires from occurring in peatlands. Therefore, developing and implementing deep learning-based semantic segmentation of vegetation density in 2-dimensional images are relevant.

2. Literature review and problem statement

Prior studies on the classification of general land areas, wetlands, and peatlands have been conducted. Supervised classification is appropriate in all circumstances, in the form of a study of the post-classification change detection approach employing the supervised maximum likelihood classifier (MLC) [11].

Paper [14] adopted the supervised classification strategy in their research, according to [11]. They employed dual temporal (planting and non-growing seasons) from Landsat image data and an acceptable index for topography and geomorphology. The median color of each super pixel region serves as a feature for the classification method's input throughout the data segmentation process utilizing a multi-resolution segmentation methodology. This research used techniques like random forest, support vector machine (SVM), decision trees, and artificial neural network (ANN) in the interim for training and classification. The results show that RF produced the overall best accurate maps. Similar to this, after evaluating several researches, Paper [15] concluded that the majority of them employ supervised classification techniques in this work. Previous researchers looked at information about common plain areas using satellite data from several of these studies [11, 14, 15]. According to [14], the proposed classification method necessitates a feature extraction step. In this study, the features were obtained using a combination of bands 1 to band 7 and band 9, which include the beach, blue, green, red, near infrared (NIR), short wave infrared 1 (SWIR 1), SWIR 2, and cirrus bands. Researches [11, 15] are the outcomes of a survey of other studies that also make use of satellite data.

Research using Landsat data was done in [9]. They used the clustering approach based on the incorrect colour composition to discriminate between binary categories of wetland and non-wetland. Paper [16] used Landsat data from numerous years combined with the MLP classification algorithm. The classification results were computed based on the segmentation approach, and the Markov chain was used to examine them. Paper [17] also completed the work. Tuning is done to get the best performance utilizing Landsat data that has been categorized using the random forest approach. The characteristics are determined by the colors present in the data. By employing the Matthew's correlation coefficient (MCC) evaluation and the classification approach in this study applies color characteristics. In order to categorize changes in the time of data retrieval, this study uses the geographical properties of satellite images based on times series. Multi-resolution-based segmentation based on the classification decision tree is utilized to accomplish the same task [18]. This approach is used to get change detection and land cover categorization. This research tracks yearly variations in land cover; however, it doesn't describe how the methodology performed. The same thing was investigated in [13] using radiometric correlation, atmospheric correlation, and RPC orthorectification. The pre-processing results were sampled, and optimization was carried out using the Z matrix coefficient and combining it with the LRR method to create the proposed feature. A spectral spatial based kernel filter was used to classify the collected data according to nonlinearity. Despite a performance improvement of 5.33 % over earlier research, it was concluded from this investigation that the categorization for highways was quite poor. The entire study discusses data from the wetlands area using satellite data [9, 16–19].

Paper [14] conducted a study on peatlands utilizing 4 types derived from Landsat data and categorized using the Random Forest classification approach. The strategy is coupled to give the performance that improves accuracy by 6 to 9 %. Paper [20] classified the data using unsupervised K-Means cluster analysis, a single polarization-based land cover model, and Sentinel-1 data. High accuracy and kappa value performance are shown in the results. With a non-metric multidimensional scaling ordination axis and cluster based on k-medoid fuzzy, previous research tracking plant functional kinds using satellite data and the Random Forest Regression approach [21]. Based on all of the studies done in the peatland region with using satellite data [10, 20, 21].

Similar studies were conducted using satellite data, the convolution neural network (CNN) technique, and the recurrent neural network (RNN) technique [12, 13]. However, they used this technique to classify single image and time series data. It's just that nearly all of this research uses machine learning techniques that combine statistical features, NDVI's temporal features, near-infrared (B8) and red (B4) bands, and spatial data from the NDVI. However, the drawback of this research is getting data from satellites. This problem was reportedly resolved in [22], who is researching the vegetation of wetlands. This work contrasts SVM, CMER, and SCG-MLP with other classification methods.

However, they do not categorize pictures based on CNN-based segmentation of the image region; rather, they do it based on the characteristics of the features they extract from the image. Therefore, let's suggest deep learning-based picture segmentation to set our study proposal apart from previous investigations. The deep learning approach to semantic picture segmentation is examined in this work. DeepLabv3+ is one method that may be applied. A variety of designs are available for DeepLabv3+, which is based on the CNN architecture and offers competitive performance.

3. The aim and objectives of the study

The aim of the study is developing a deep learning-based semantic segmentation approach in order to increase the quality of image segmentation and calculate vegetation density levels using 2-dimensional images.

To achieve this aim, the following objectives are accomplished:

– to segment the vegetation density level based on 2-dimensional images;

– to choose the appropriate parameters to achieve the best performance results in the classification model based on deep learning technique with inception resnetv2 architecture;

– to compare the outcomes of the best segmentation model to determine vegetation density.

4. Materials and methods

4. 1. Object and hypothesis of the study

The deep learning technique for semantic image segmentation is tested in this work. DeepLabv3+ is a method that may be applied. Based on the CNN architecture, DeepLabv3+ features a variety of designs that enable it to perform competitively.

The object of this research is the Liang Anggang Protected Forest area, Banjarbaru City block 1 area with targeted data collection for a month. The location of the research object on map can is shown in Fig. 1. The selection of this research location was based on observations and survey of the block 1 area which in the landuse pattern of the block 1 area based on the South Kalimantan Provincial Forestry Service in 2017, an area of 479 hectares of block 1 area which is filled with land such as agriculture, plantations, roads and settlements and an area of 494 hectares full of weeds. In addition, the location is also categorized as peat land. The research location, especially in block 1, fulfills the characteristics and suitability of the need for data collection for land cover classification in terms of the type of vegetation density (bare, medium and high) that can be seen with the human eye during observations and surveys.

In contrast to prior research that used classification-based approaches, this study offered a segmentation method for calculating land density. The study approach was carried out in line with Fig. 2, as the theoretical method employed according to the description produced. The suggested model's experimental design is provided in the experimental design portion.

Fig. 2 displays the processes of data collection, image processing, segmentation, and assessment, with the stages and supporting theories provided concurrently as shown below. In order to ensure the success of this research, the following tools are used:

– hardware: Computer using Intel® Core™ i9-10980XE CPU @ 3.00 GHz, 31Gb, with GPU using RTX 3090;

– software: Windows 10, MATLAB R2022a Student Version;

– data collection using the DJI Mavic Pro Drone.

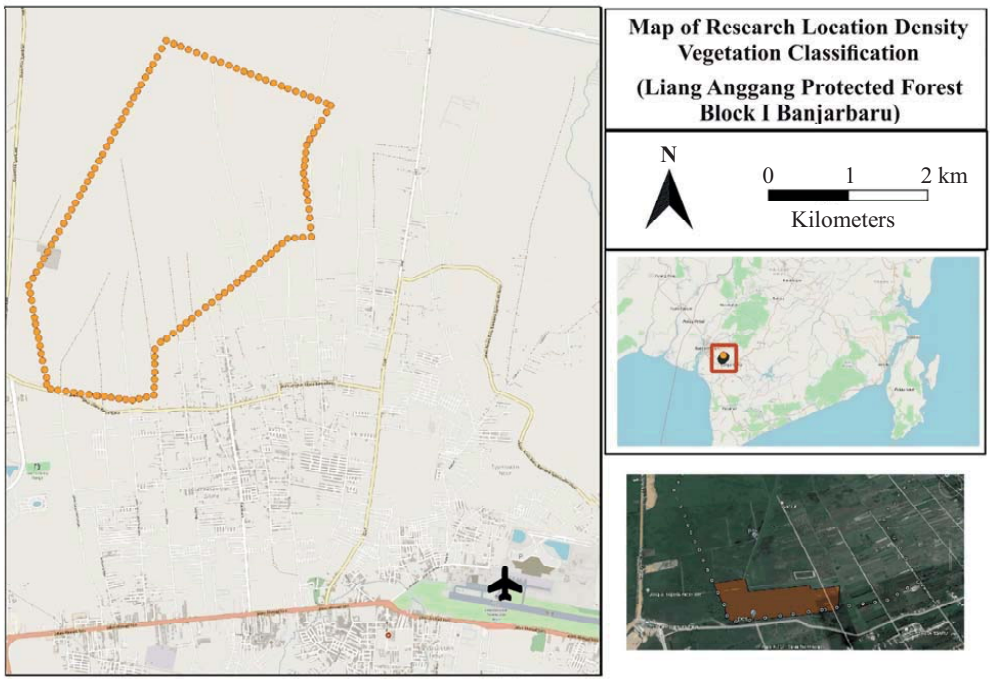


Fig. 1. Research object location on map

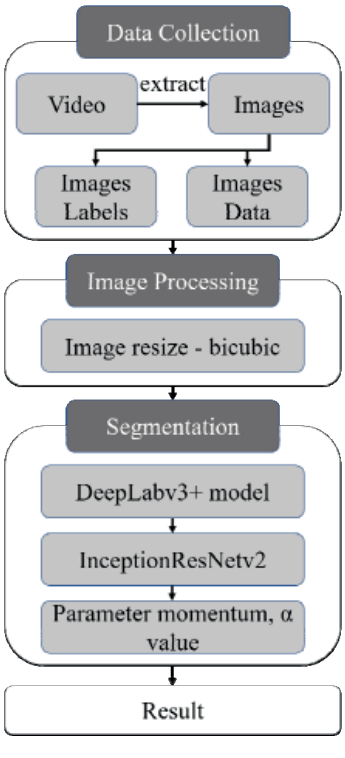


Fig. 2. Proposed method

4. 2. Data collection

The dataset utilized was created from 2-dimensional imagery captured by a drone in July during a clear midday period between 10:00 and 14:00 Eastern Indonesian Time. The drone hovers between 20 and 25 meters in the air. The drone is positioned 20 meters away from the object's outer surface. The drone camera angle setting is combined with the height setting. Fig. 3 depicts an illustration of the drone's

images capturing angle. The drone's image capture angle is set at 600 to 900 degrees.

Fig. 3. An example of a drone's shooting angle

300 images total, with three different densities-high density (heavily grazed), medium density (softly grazed), and low density (bare/un-grazed) – make up the quantity of data used in this study. There are different qualities for each density. Bare/un-grazed typically consists of dry shrubs, soil, water, and settlements, while heavily grazed typically consists of vegetation that has a density between 15 and 30 meters. Meanwhile, softly grazed typically consists of shrubs that are 5 to 6 meters apart and fresh shrubs up to a maximum of 3 meters. According to Fig. 3, an expert performs the initial segmentation manually and uses it as a guide for classifying the image per pixel.

As seen in Table 1, each image has multiple labels and categories. Unlike gently grazed, which has two groups, bushes and shrubs with RGB color compositions of [255 0 102] and [102 0 102], extensively grazed only has one, the vegetative category, which is identified by an RGB colon composition of [255 0 0].

Table 1

Preview image original and image label

Labels	Image Original	Image Label
Bare/un-grazed		
Heavily grazed		
Softly grazed		

At low concentrations, the "bare" category comprises the dry bush, soil, water, and settlement categories. This category includes the RGB values [255 102 0], [102 51 0], [1 110 192], and [255 255 0].

4. 3. Image Processing

At this point, the data is changed to the same size, 256×256, in accordance with the layer›s thickness. Additionally, translation and rotation are widely used in image processing to retrieve data attributes. For shifts in different directions, the translation range is calculated using a value of 10 pixels for the x and y coordinates. When in rotation, multiples of 10 until 40 are used.

4. 4. DeepLabv3+ model

One of the best models for semantic segmentation is the DeepLabv3+ network, but it has certain drawbacks [23]. After removing the network with hollow convolutional features, DeepLabv3+ connects to the Atrous Spatial Pyramid Pooling (ASPP) structure to improve the ability to segment multi-scale targets. Deeplabv3+ upgraded resNet101 to Xception and reconfigured the primary network on the original base [24]. As seen in Fig. 4, the structure comprises global average pooling operations and 3×3 hole convolutions with expansion ratios of 1, 6, 12, and 18, respectively. The target characteristics of the image edge cannot be accurately extracted at a high expansion rate.

There is a hole phenomenon in the segmentation of the large-scale target because it cannot fully replicate the relationship between the local elements of the large-scale target. These findings decreased the large-scale targets and the DeepLabv3+ network segmentation accuracy of remote sensing image edge targets. To overcome this problem, a model is proposed by utilizing the inceptionresnetv2 architecture to upgrade the Xception architecture so that performance can be more reliable and has lower computational complexity.

Atrous convolution is potent technique generalizes the ordinary convolution operation and enables to explicitly regulate the resolution of features calculated by deep convolutional neural networks as well as the field-of-view of the filter to collect multi-scale information.

Significantly reducing processing complexity is depthwise separable convolution, which factors a conventional convolution into a depthwise convolution followed by a pointwise convolution (i. e., 1×1 convolution). In particular, the pointwise convolution is used to aggregate the output from the depthwise convolution, which conducts a spatial convolution individually for each input channel.

In order to retrieve the features calculated by deep convolutional neural networks at any resolution, DeepLabv3 uses Atrous convolution as its encoder. When comparing the final output resolution to the spatial resolution of the input image, the output stride is used (before global pooling or fully-connected layer). The output stride for the image classification task is 32 since the final feature maps' spatial resolution is typically 32 times lower than the resolution of the input images. By removing the striding in the final one (or two) blocks and applying the appropriate Atrous convolution, one can use output stride of 16 (or 8) for denser feature extraction when performing semantic segmentation (for example, let's apply rate of 2 and rate of 4 to the final two blocks, respectively for output stride of 8).

The encoder features from DeepLabv3 are usually computed with output stride of 16. The features are bilinearly upsampled by a factor of 16, which could be considered a naive decoder module. However, this naive decoder module may not successfully recover object segmentation details.

Based on that, the encoder features are first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level features from the network backbone with the same spatial resolution, which uses the ResNet-101 architecture. Let's propose a model using inceptionresnetv2 architecture to replace Xception architecture which is more reliable and has a lower computational complexity.

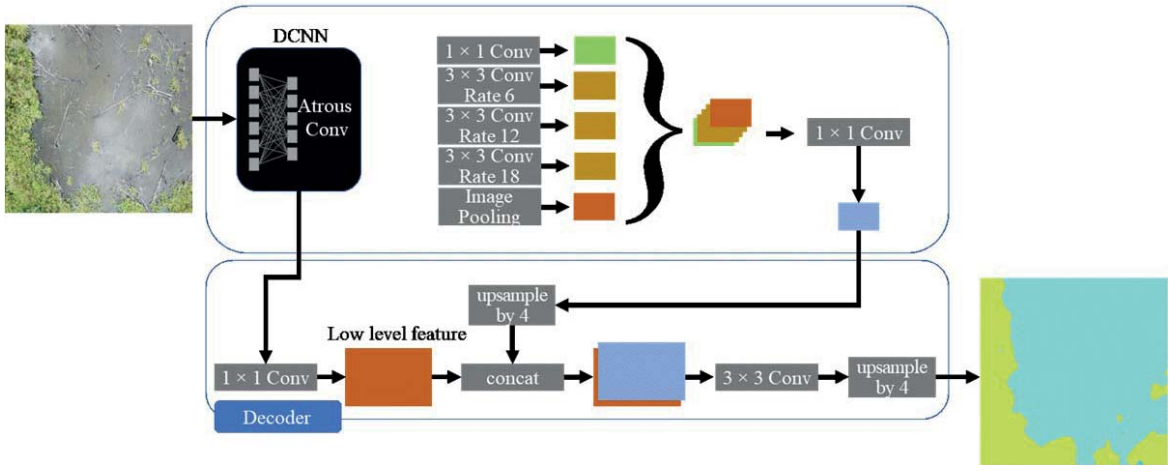


Fig. 4. DeepLabv3+Model

4. 5. Inception-ResNet-v2

In order to accommodate the entire model in memory, older Inception models used to be trained in a partitioned manner, where each replica was divided into a number of sub-networks [25]. The number of filters in the various layers may, however, be changed in a variety of ways with the Inception design without affecting the effectiveness of the fully trained network. The extraneous baggage has been removed in inceptionv4, and the inception blocks have been uniformly chosen for each grid size.

The less expensive inception blocks than the original Inception are used for the residual versions of the Inception networks. A filter-expansion layer (1×1 convolution without activation) is used to scale up the dimensionality of the filter bank before the addition to match the depth of the input after each Inception block to match the depth of the information. This is essential to compensate for the dimensionality loss brought on by the Inception block.

Only two of the various leftover forms of inception are described here. The computational cost of the first one, inceptionresnetv1, is comparable to that of inceptionv3. Inceptionresnetv2 is comparable in terms of raw cost to the recently released inceptionv4 network. However, in fact, the step time for inceptionv4 turned out to be far slower, most likely as a result of the higher number of layers. The overall schematic of the inceptionresnetv2 architecture is shown in Fig. 5.

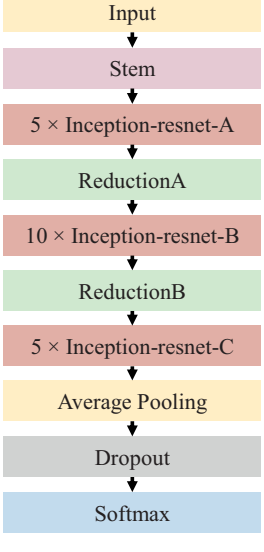


Fig. 5. Overall schema of the Inceptionresnetv2 network

The scheme's specifics are depicted in Fig. 6 through 11. The stem scheme, which is the second stage of the Inceptionresnetv2 architecture, is shown in Fig. 6.

The Inceptionresnetv2 front module is a stem layer that uses five 3×3 convolutions, one 3×3 max set, and one 1×1 convolution from 299×299×3 image inputs to produce a 35×35×256 feature map. Stem layers are included in the inceptionresnetv2 architecture to enhance the performance of the original neural network. Afterwards, there are three inception modules used to extract the image features namely, inceptionresnetA, inceptionresnetB, and inceptionresnetC schemas as shown in Fig. 7–9, respectively. Each inception module functions as a number of convolution filters, such as the 1×1 and 3×3 filters, which are intended for effective extraction of visual features.

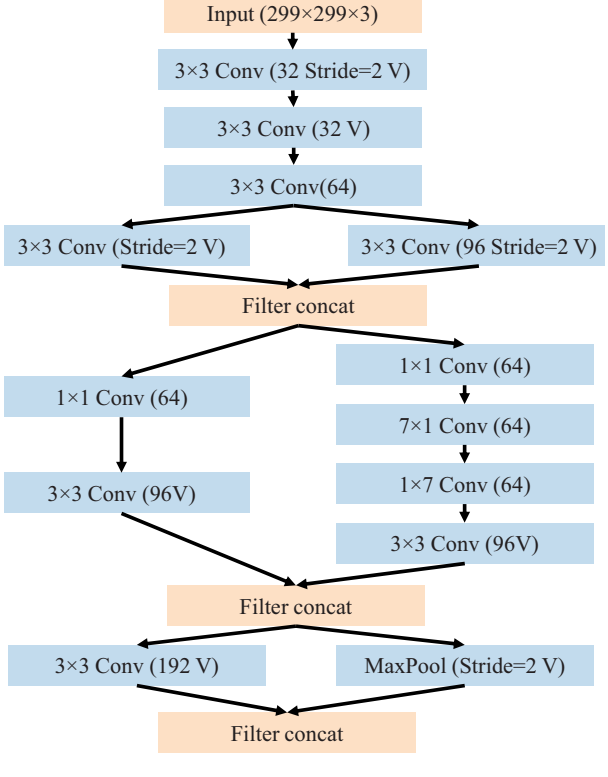


Fig. 6. The schema for stem

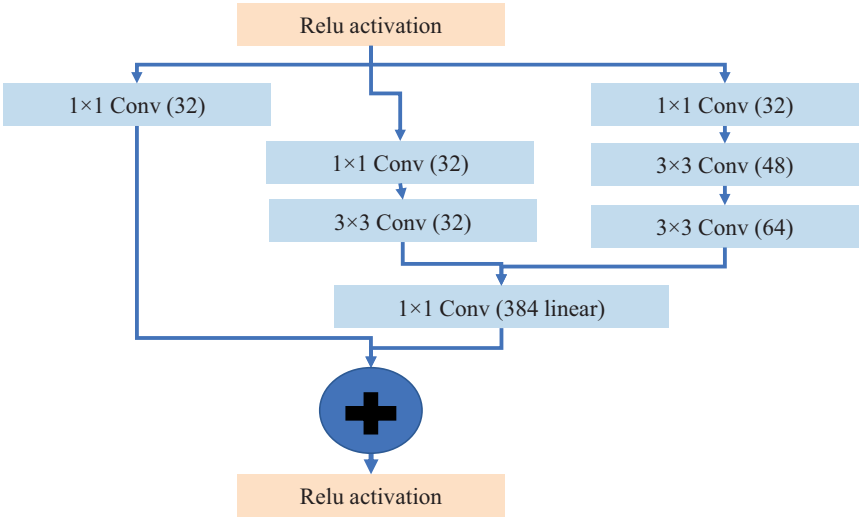


Fig. 7. Inception-resnet-A block schema

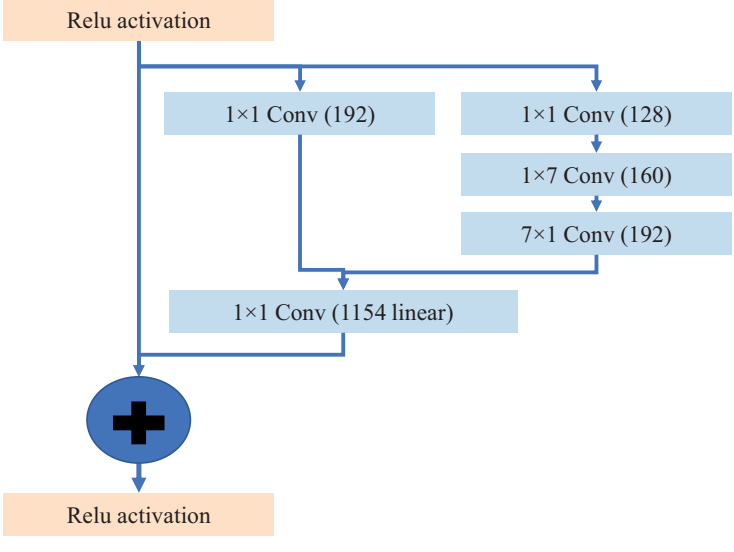


Fig. 8. Inception-resnet-B block schema

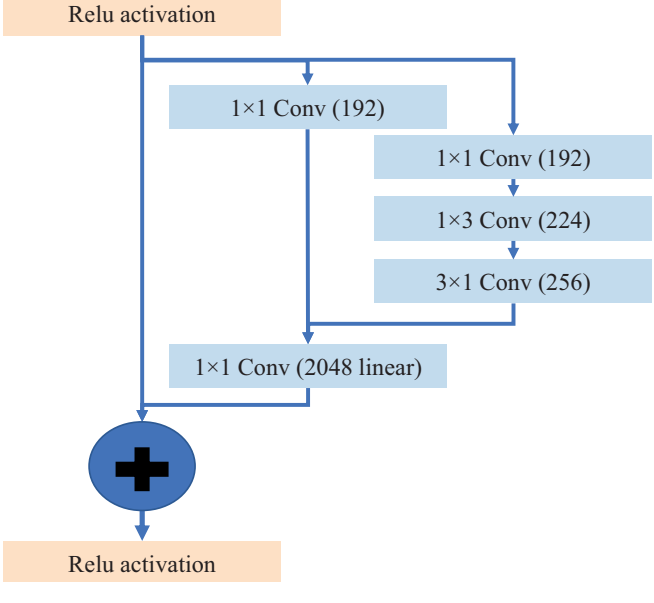


Fig. 9. Inception-resnet-C block schema

While Fig. 10, 11 depict the reduction technique. The 35×35×256 feature map is reduced by the reduction-A module to a 17×17×896 feature map, and by the reduction-B module to an 8×8×1792 feature map. The characteristics from the previous layer are passed to the subsequent layer by these two modules, which minimize their size.

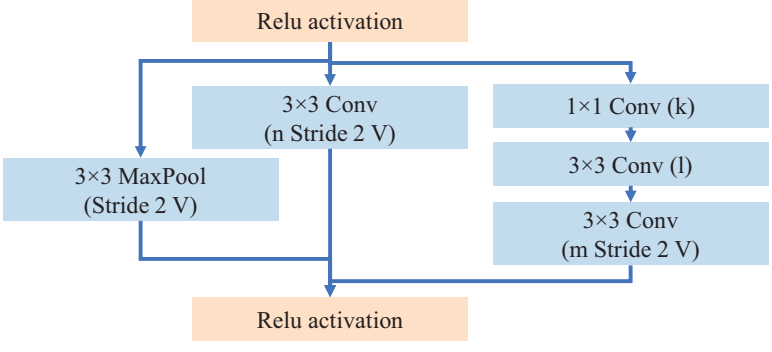


Fig. 10. Reduction module A, with the *k, l, m, n* number represent filter bank sizes

As seen in Fig. 10, the reduction technique used parallel convolutional model. As for the reduction module A, Relu activation function is used as the base start. Then, the first parallel model used 3×3 Max Pool (Stride 2 V), the second parallel model used 3×3 Conv (n Stride 2 V) and the third parallel model used 1×1 Conv (k), 3×3 Conv (l) and 3×3 Conv (m Stride 2 V).

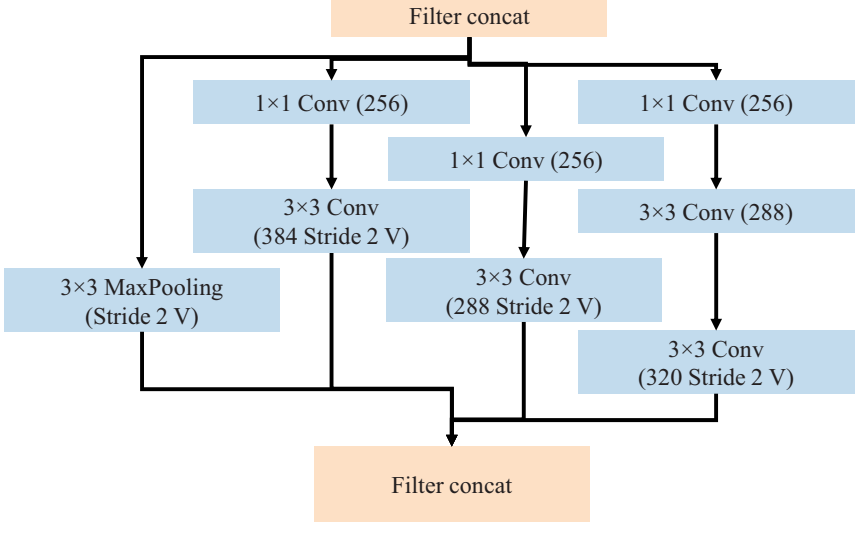


Fig. 11. Reduction module B

In Fig. 11, the reduction module B used Filter concat as the base function. Then, the data processed into four parallel model. These parallel models used for both reduction technique so that a deep neural network can be trained on less powerful GPUs. Also, Parallel convolutions model seem to be learning better representations of the data. Essentially, the correlation between kernel filters of different filter groups is usually quite less, which implies that, each filter group is learning a unique representation of the data [26].

4. 6. Training Methodology

The deep learning network is then trained using a machine learning system on the GTX 1050TI Graphical Processing Unit (GPU) utilizing stochastic gradient descent with momentum (SGDM). In our experiment, let's use momentum multiples ranging from 0.1 to 0.9. Additionally, it employs a minibatchSize of 4, initialLearningRate of 0, and maxEpochs of 10. Additionally, let's employ adaptive moment estimation, and training techniques for root mean squared propagation (RMSProp) in line with the aforementioned parameters adaptive moment estimation (Adam). The SGDM training approach was tested using these two training techniques. A running average of the parameters calculated over time is used to evaluate the model.

4. 7. Evaluation Method

4. 7. 1. Accuracy

The percentage of correctly classified pixels for each class is indicated by accuracy. If you want to see how well each class accurately detects pixels, use the accuracy metric. According to the ground truth, Accuracy is the proportion of correctly identified pixels to all the pixels in a given class. The accuracy value was calculated by applying (1):

$$acc = \frac{TP}{TP + TN}, \tag{1}$$

where *TP* (True Positive) and *TN* (True Negative) is the number of segmented areas classified to the right vegetation level according to ground truth and *TP+TN* is the total number of data used. Mean Accuracy represents the average accuracy across all classes and photos in the aggregate data set. Mean Accuracy represents the average accuracy of all classes in a given image for each image. Although it is a straightforward metric similar to global accuracy, class accuracy can be deceptive.

4. 7. 2. Intersection over union (IoU)

The most used measure is intersection over union (IoU), often known as the Jaccard similarity coefficient. If you want a statistical accuracy assessment that penalizes false positives, use the IoU metric. IoU is the proportion of correctly identified pixels to all ground truth and forecasted pixels for a given class. The formula for the IoU score is shown in (2):

$$IoU = \frac{TP}{(TP + FP + FN)}, \tag{2}$$

where true positives (*TP*) is the number of predicted pixels that overlaps with ground truth box, false positives (*FP*) is the number of predicted pixels that outside of the ground truth box, and false negatives (*FN*) is the

number of pixels of ground truth box that failed to be predicted. Mean IoU represents the average IoU score across all classes in a given image. MeanIoU represents the average IoU score across all classes and photos in the aggregate data set.

4. 8. Experiment Design

By examining the relationship between image processing techniques and the case of vegetation density in 2D image data, this study aims to present the level of vegetation density based on segmentation semantics in 2D images. It also aims to select parameters in the segmentation method and the results of comparisons on the experiments conducted. As a result, studies with semantic segmentation were conducted using a variety of CNN architectures including resnet18, mobilenetv2, exception, and inceptionresnetv2 as a proposed method. The previous architecture is used in the DeepLab model [23], which also uses the exception architecture and compare the xception architecture with inceptionresnetv2 architecture. Each architecture is implemented using various training techniques, including adaptive moment estimation (Adam), root mean squared propagation (RMSprop), and stochastic gradient descent with momentum (SGDM). This study also analyses semantic segmentation techniques based on U-Net [23]. The momentum parameter, which ranges from 0.1 to 0.9 for multiples of 0.1, is used to evaluate all models. The optimal outcomes are displayed in accordance with the accuracy values from (1) and (2).

5. The research result of semantic image segmentation based on deep learning method

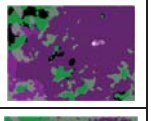
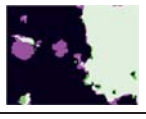
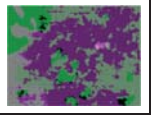
5. 1. Segmentation of the vegetation density level on 2-dimensional image

Based on the best findings of the best accuracy comparison, Table 2 shows the segmentation results. The column in Table 2 is in the form of Bare-density, Medium-density, and High-density labels in the image. Each row consists of original data, segmentation ground truth, segmentation result using inceptionresnetv2 architecture, mobilenetv2 architecture, Xception architecture, and resnet18 architecture, respectively. This result is obtained by implementing the deep learning model for each CNN architecture being observed using matlab student version.

The tables of architecture results demonstrate that the green and magenta sections emphasize places where segmentation results deviate from the fundamental reality anticipated by the expert. The number of green colors obtained by the mobilenetv2 architecture, Xception architecture, and resnet18 architecture are displayed from the prediction results. However, inceptionresnetv2 architecture and mobilenetv2 architecture both produce magenta colors. The segmentation results in magenta and green deviate from the experts' reported "ground truth." It can be seen from the quantity of differences that the inceptionresnetv2 architecture differs slightly from the experts.

Table 2

Visual comparison of several CNN architecture for DeepLabv3+ model

Description	Bare/ungrazed	Softly grazed	Heavily grazed
Original data			
Segmentation ground truth			
Inceptionresnetv2 architecture			
Mobilenetv2 architecture			
Xception architecture			
Resnet18 architecture			

5. 2. Choosing appropriate parameters to achieve the best performance

5. 2. 1. Intersection over union (IoU)

IoU values, or the ratio between correctly categorized pixels and the total number of ground truth and predicted pixels in that class, are compared in Fig. 12. This IoU value obtained by implementing the equation (2) using matlab student version. then it calculates the IoU value for each segmented object detected in images. The IoU number displayed in the inceptionresnetv2 architecture is the second value after xception; it is 0.09 % higher than the inceptionresnetV2 architecture that is suggested. In term of IoU value, xception architecture is better than other CNN architecture observed in this research.

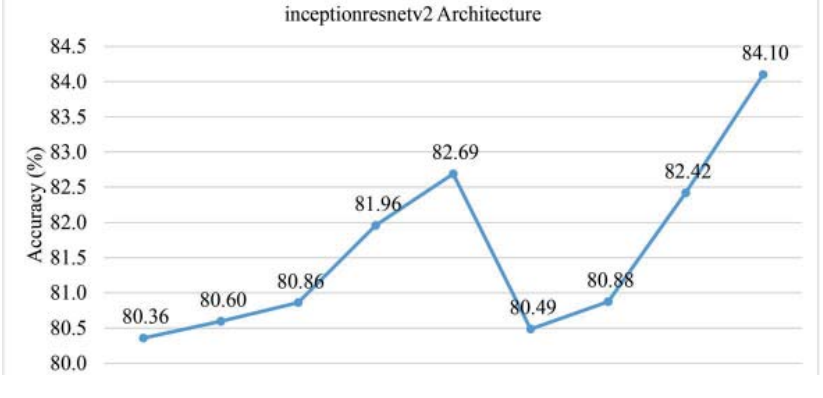


Fig. 12. Performance comparison of Intersection over Union for Convolutional Neural Network architecture

Because IoU in this situation simply calculates the overlap between the ground truth and prediction results, the prediction results from the semantic segmentation approach will be worse the lower the IoU number.

5. 2. 2. Accuracy

Accuracy value obtained after each deep learning model trained and tested. The models are implemented using matlab student version. When momentum is applied between 0.1 and 0.9, performance results in a performance that tends to fluctuate for the proposed inceptionresnetv2 architecture (can been shown in Fig. 13). The value of 0.9 is the point where the features created by the inceptionresnetv2 architecture are in the closest position to the expert's label and acquire a performance of 84.10 %, as seen from these statistics. Furthermore, it demonstrates that the momentum value of 0.9 is the best value for achieving convergence. Momentum 0.5 is the second-best order, and as a result, it performs at 82.69 %.

According to findings in Fig. 14, there is an increase up to epoch 135, but a fluctuating pattern continues until epoch 450. This enables the suggested architecture to update the data continuously. The highest epoch was attained at 315 epochs with an accuracy value of 85.46 percent out of the multiples of every 45 epochs. The performance results are derived using the best momentum value, which is 0.9. This experiment has demonstrated that increasing the number of epochs can enhance performance.

A comparison of the CNN architecture's deployment depending on momentum value is shown in Fig. 15. For each increase in momentum, it is demonstrated that the Xception and mobilenetv2 architectures significantly increase, unlike inceptionresnetV2 and Resnet 18, which offer an up-and-down rhythm. According to these tests, for each momentum parameter, Inceptionresnetv2 performs better than mobilenetv2 and Xception designs. From this, it can be concluded that the outstanding momentum value for every architecture is unique.

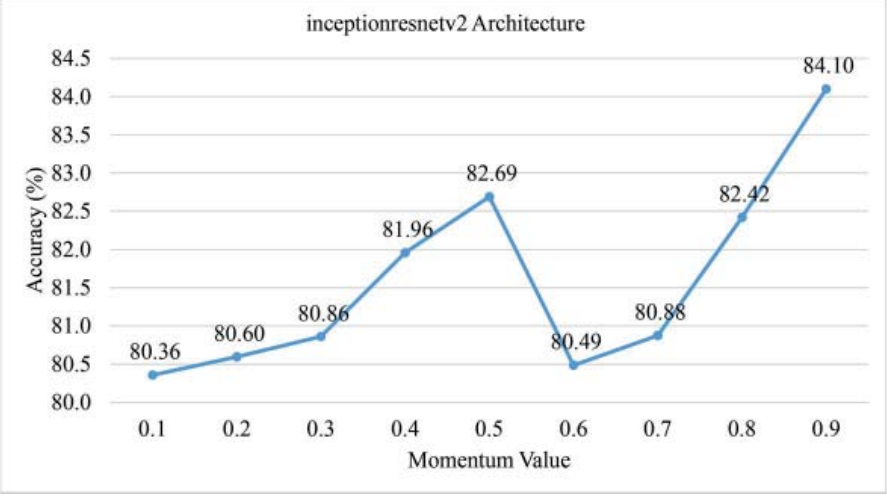


Fig. 13. Average performance accuracy using inceptionresnetv2

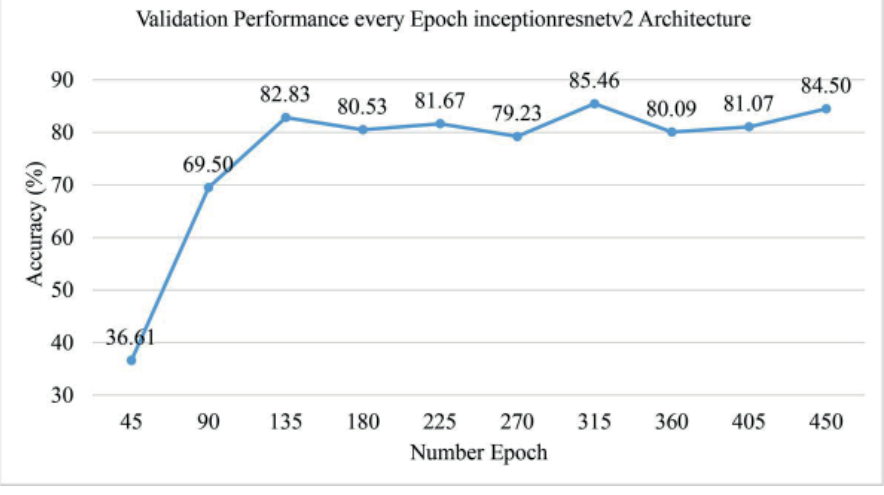


Fig. 14. Validation performance every epoch using momentum=0.9

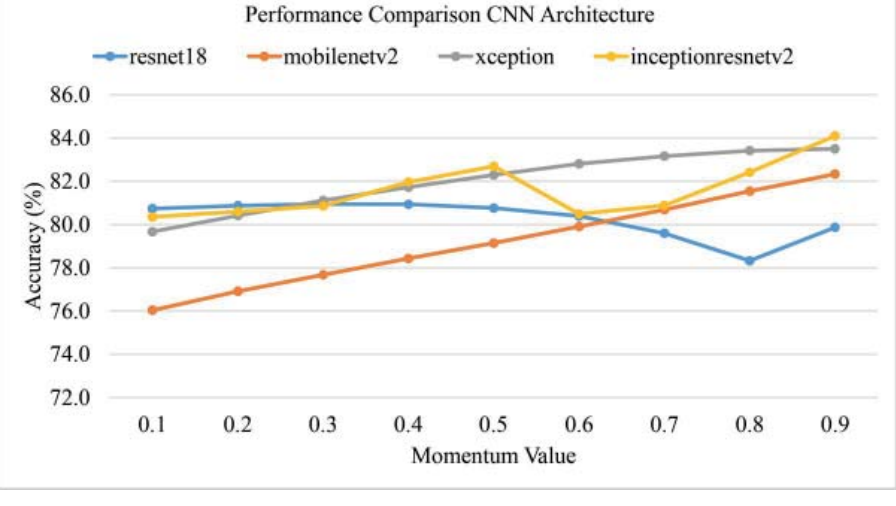


Fig. 15. Performance comparison for Convolutional Neural Network architecture based on momentum value

InceptionresnetV2 achieves superior performance on average than other architectural solutions, according to a comparison of average accuracy in Fig. 16. This backs up one illustration of a result from a visual comparison. The provided number exceeds the resnet18 architecture by 2.2 percent, the mobilenetv2 architecture by 1.8 percent, and the Xception architecture by 0.7 percent. This result demonstrates that the suggested approach performs more effectively when categorizing vegetation density levels based on 2D image.

A comparison of the inceptionresnetv2 architecture training methods is shown in Fig. 17. Stochastic gradient descent with momentum (SGD with momentum), adaptive moment estimation (Adam), and root mean square propagation (RMSprop) were compared (SGDM). The findings demonstrate that utilizing SGDM yields better accuracy outcomes than using Adam or RMSprop. With a performance of 83.23 %, SGDM outperformed RMSProp and Adam by 16.78 and 16.88 % points, respectively.

These training techniques can all lead to various optimal local minimums while incurring the same loss. While momentum might hasten the convergence process, RMSProp slows it down and Adam usually results in a sharper convergence of the minima. Thus, the optimal training techniques suggested for semantic segmentation of vegetation density is Stochastic gradient descent with momentum (SGDM).

5. 3. Comparison of the outcome of the best segmentation model to determine vegetation density

This study uses the same training approach, SGDM, to compare the U-Net method in addition to comparisons based on CNN architecture. The findings demonstrate that DeepLabv3+ with inceptionresnetv2 architecture offers superior performance to U-Net (has been shown in Fig. 18). This is true because the U-Net has problems because it creates a symmetrical expansion path from the up-sampling section of this U-Net approach which contains a lot of feature channels and permits spreading context information to reports with better resolution.

According to the results of these repeated testing, the combination of DeepLabv3 with the inceptionresnetv2 architecture delivers 84.10 % greater performance than U-Net. These findings suggest that it performs well when used to classify vegetation density using semantic segmentation.

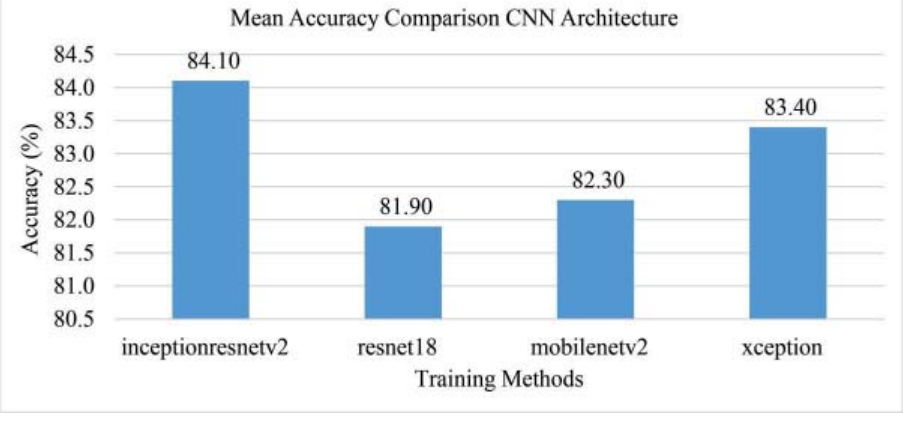


Fig. 16. Performance comparison for Convolutional Neural Network architecture based on momentum value

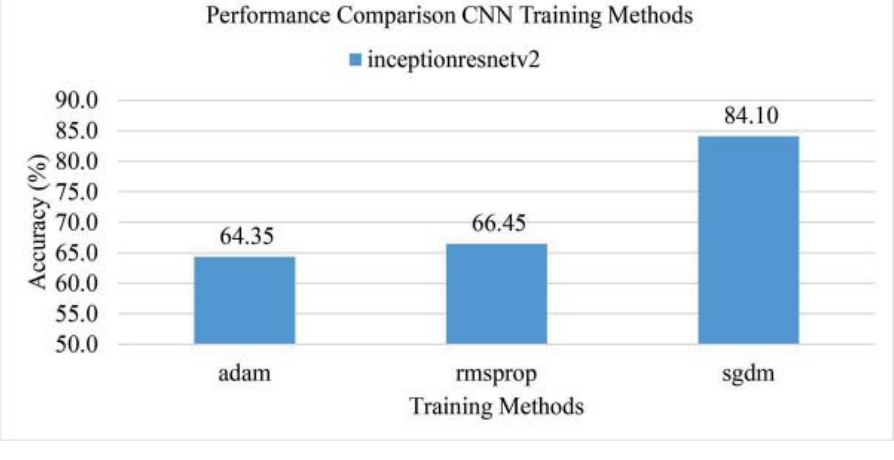


Fig. 17. Performance comparison for inceptionresnetv2 Convolutional Neural Network architecture based on training methods

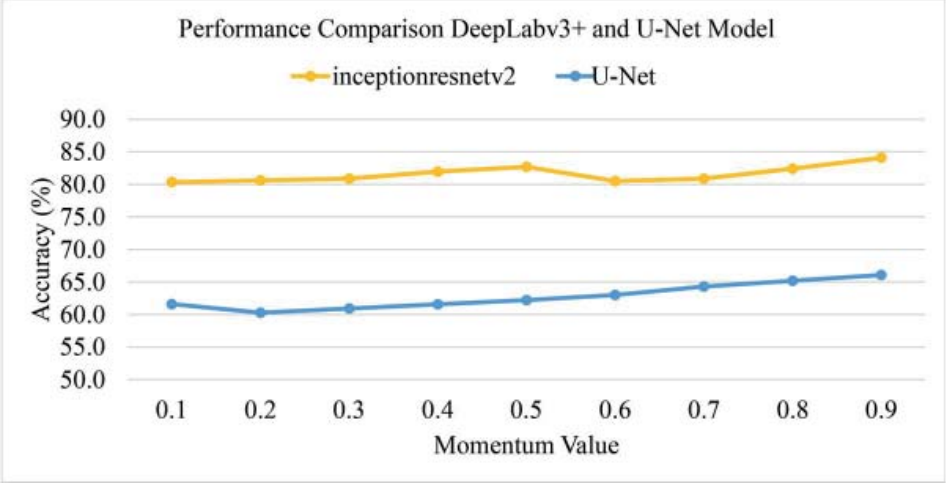


Fig. 18. Performance comparison DeepLabv3+ and U-Net Model

6. Discussion result of semantic image segmentation based on deep learning method

Employing a deep learning-based semantic segmentation method with DeepLabV3+ on a dataset of photos of vegetation, this study investigates the efficacy of using deep learning architecture to semantically segment images. The correctness of the segmentation results is then confirmed by comparison to professional findings. In order to achieve the best performance in the classification model based on deep learning technique with inception resnetv2 architecture, it was determined from that result (Table 2) that there was a relationship between image processing techniques. The appropriate parameters have been shown in Fig. 13, 14 based on this investigation. It validates the use of the appropriate design-related variables.

The experiment employed private data evaluated on multiple semantic segmentation approaches and deep learning architectures. The results reveal that when employing the inceptionresnetv2 architecture, the proposal gives higher performance and approaches ground truth than other suggested architectures such as resnet18 and Xception (Fig. 12, 15). This is consistent with study in [27], which demonstrates that the inceptionresnetv2 design outperforms numerous resnet18 and Xception architectures. This is made feasible by inceptionresnetv2, a convolution network that is modelled after the Inception network but employs residual connections rather than filter circuits. Only over the conventional layer and not over the summation is batch normalization used for the initial block of residuals. It is possible to significantly increase the total number of Inception blocks by dropping batch normalization on top of that layer. The layer size may be adjusted to balance calculations across the model's many sub-networks, and residual scaling can be used to improve training speed. The residue appears to be stabilized by being reduced in size before being added to the prior layer's activation.

Based on the IoU value, it is shown that the better the higher, but in the experiment, it is shown that the Xception architecture gives 0.09 % higher than Inceptionresnetv2 and it is shown that there is a lot of overlap between ground truth and prediction (Fig. 12). These findings indicate that it is reasonable since "extreme inception," or Xception, calls for the use of extreme Inception principles. ReLU non-linearity also follows the Inception architecture but not the Xception architecture. These results show that, because to the same conceptual underpinnings of the Xception and Inceptionresnetv2 designs, the CNN technique may yield competitive performance when applied to those architectures. The results of experiments employing successfully applied semantic segmentation on the level of plant land density based on two-dimensional photos are also impacted by parameter selections. When compared to alternative designs, the drawbacks of the enhanced semantic segmentation approach employing inceptionresnetv2 fall short in terms of required memory utilization. In order to enhance the performance of the model, more study may focus on lowering the colors used for class segmentation. Additionally, to increase the reliability of the data, continuous time series of photographs of land covering are acquired.

According to the experiments carried out for this study and one of its limitations, the dataset collected using a drone is still a portion of the video confined by distance and height. According to data acquired by area in peatland, the number of labels still utilized is limited and does not fully account for situations in densely populated regions. This is possible since each area has different characteristics based on vegetation density. This study is useful, particularly in initiatives to reduce fire-prone locations in densely populated regions.

7. Conclusions

1. Segmentation of the vegetation density level based on 2-dimensional image is conducted. There are 3 level of vegetation density being segmented namely, Bare-density, Medium-density, and High-density. Deep learning architecture used are inceptionresnetv2 architecture, mobilenetv2 architecture, Xception architecture, and resnet18 architecture, respectively. From experimental result, the quantity of segmented region differences that the inceptionresnetv2 architecture produce are differs slightly from the experts.

2. Appropriate parameters value to achieve the best performance is determined. There are three parameters of deep learning architecture used namely, momentum, epoch, and optimizer. In our experiment, we used momentum multiples ranging from 0.1 to 0.9 . The value of 0.9 is the point where the features created by the inceptionresnetv2 architecture

are in the closest position to the expert's label and acquire a performance of 84.10 %, hence momentum value of 0.9 is best. Momentum 0.5 is the second-best order, and as a result, it performs at 82.69 %. The epoch used is ranging from 100 to 450. The highest epoch was attained at 315 epochs with an accuracy value of 85.46 percent out of the multiples of every 45 epochs.

A comparison of the inceptionresnetv2 architecture with different training techniques is conducted. Adaptive moment estimation (Adam), and root mean square propagation (RMSprop) were compared to Stochastic gradient descent with momentum (SGD with momentum) (SGDM). With a performance of 83.23 percent, SGDM outperformed RMSProp and Adam by 16.78 and 16.88 percentage points, respectively.

3. Comparison of the outcome of the best segmentation model to another known model is conducted. From experimental study, the best model obtained is inceptionresnetv2 with training approach using SGDM. This model is compared to another known good model namely, U-Net. The combination of DeepLabv3 with the inceptionresnetv2 architecture delivers 20 % greater performance than U-Net with value 84.10 %.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Acknowledgments

This work was supported by the DRTPM grant, National Basic Research Competition scheme, agreement number 113/E5/PG.02.00/PT/2022 and LPDP through the Ministry of Education, Culture, Research, and Technology Indonesia.

Reference

1. Warren, M., Hergoualc'h, K., Kauffman, J. B., Murdiyarso, D., Kolka, R. (2017). An appraisal of Indonesia's immense peat carbon stock using national peatland maps: uncertainties and potential losses from conversion. Carbon Balance and Management, 12 (1). doi: https://doi.org/10.1186/s13021-017-0080-2
2. Gumbricht, T., Román-Cuesta, R. M., Verchot, L. V., Herold, M., Wittmann, F., Householder, E. et. al. (2017). Tropical and Subtropical Wetlands Distribution version 2. Center for International Forestry Research (CIFOR). doi: https://doi.org/10.17528/CIFOR/DATA.00058
3. Margono, B. A., Potapov, P. V., Turubanova, S., Stolle, F., Hansen, M. C. (2014). Primary forest cover loss in Indonesia over 2000–2012. Nature Climate Change, 4 (8), 730–735. doi: https://doi.org/10.1038/nclimate2277
4. Hope, G., Chokkalingam, U., Anwar, S. (2005). The Stratigraphy and Fire History of the Kutai Peatlands, Kalimantan, Indonesia. Quaternary Research, 64 (3), 407–417. doi: https://doi.org/10.1016/j.yqres.2005.08.009
5. Tacconi, L. (2016). Preventing fires and haze in Southeast Asia. Nature Climate Change, 6 (7), 640–643. doi: https://doi.org/10.1038/nclimate3008
6. Sandhyavitri, A., Amri, R., Fermana, D. (2016). Development of Underground Peat Fire Detection. Proceeding of the First International Conference on Technology, Innovation and Society. doi: https://doi.org/10.21063/ictis.2016.1069
7. Garcia-Prats, A., Antonio, D. C., Tarcísio, F. J. G., Antonio, M. J. (2015). Development of a Keetch and Byram – Based drought index sensitive to forest management in Mediterranean conditions. Agricultural and Forest Meteorology, 205, 40–50. doi: https://doi.org/10.1016/j.agrformet.2015.02.009
8. Keetch, J. J., Byram, G. M. (1988). Drought Index. Forest Seruice Research Paper, 36.
9. Abalo, M., Badabate, D., Fousseni, F., Kpérkouma, W., Koffi, A. (2021). Landscape-based analysis of wetlands patterns in the Ogou River basin in Togo (West Africa). Environmental Challenges, 2, 100013. doi: https://doi.org/10.1016/j.envc.2020.100013
10. Karlson, M., Gålfalk, M., Crill, P., Bousquet, P., Saunois, M., Bastviken, D. (2019). Delineating northern peatlands using Sentinel-1 time series and terrain indices from local and regional digital elevation models. Remote Sensing of Environment, 231, 111252. doi: https://doi.org/10.1016/j.rse.2019.111252
11. Chughtai, A. H., Abbasi, H., Karas, I. R. (2021). A review on change detection method and accuracy assessment for land use land cover. Remote Sensing Applications: Society and Environment, 22, 100482. doi: https://doi.org/10.1016/j.rsase.2021.100482
12. Meng, S., Wang, X., Hu, X., Luo, C., Zhong, Y. (2021). Deep learning-based crop mapping in the cloudy season using one-shot hyperspectral satellite imagery. Computers and Electronics in Agriculture, 186, 106188. doi: https://doi.org/10.1016/j.compag.2021.106188
13. Campos-Taberner, M., García-Haro, F. J., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilabert, M. A. (2020). Understanding deep learning in land use classification based on Sentinel-2 time series. Scientific Reports, 10 (1). doi: https://doi.org/10.1038/s41598-020-74215-5
14. Tan, J., Zuo, J., Xie, X., Ding, M., Xu, Z., Zhou, F. (2021). MLAs land cover mapping performance across varying geomorphology with Landsat OLI-8 and minimum human intervention. Ecological Informatics, 61, 101227. doi: https://doi.org/10.1016/j.ecoinf.2021.101227
15. Zaldo-Aubanell, Q., Serra, I., Sardanyés, J., Alsedà, L., Maneja, R. (2021). Reviewing the reliability of Land Use and Land Cover data in studies relating human health to the environment. Environmental Research, 194, 110578. doi: https://doi.org/10.1016/j.envres.2020.110578

16. Bunyangha, J., Majaliwa, Mwanjalolo. J. G., Muthumbi, Agnes. W., Gichuki, Nathan. N., Egeru, A. (2021). Past and future land use/land cover changes from multi-temporal Landsat imagery in Mpologoma catchment, eastern Uganda. The Egyptian Journal of Remote Sensing and Space Science, 24 (3), 675–685. doi: https://doi.org/10.1016/j.ejrs.2021.02.003
17. Magnússon, R. Í., Limpens, J., Kleijn, D., van Huissteden, K., Maximov, T. C., Lobry, S., Heijmans, M. M. P. D. (2021). Shrub decline and expansion of wetland vegetation revealed by very high resolution land cover change detection in the Siberian lowland tundra. Science of The Total Environment, 782, 146877. doi: https://doi.org/10.1016/j.scitotenv.2021.146877
18. Mao, D., Tian, Y., Wang, Z., Jia, M., Du, J., Song, C. (2021). Wetland changes in the Amur River Basin: Differing trends and proximate causes on the Chinese and Russian sides. Journal of Environmental Management, 280, 111670. doi: https://doi.org/10.1016/j.jenvman.2020.111670
19. Su, H., Yao, W., Wu, Z., Zheng, P., Du, Q. (2021). Kernel low-rank representation with elastic net for China coastal wetland land cover classification using GF-5 hyperspectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 171, 238–252. doi: https://doi.org/10.1016/j.isprsjprs.2020.11.018
20. Khakim, M. Y. N., Bama, A. A., Yustian, I., Poerwono, P., Tsuji, T., Matsuoka, T. (2020). Peatland subsidence and vegetation cover degradation as impacts of the 2015 El niño event revealed by Sentinel-1A SAR data. International Journal of Applied Earth Observation and Geoinformation, 84, 101953. doi: https://doi.org/10.1016/j.jag.2019.101953
21. Räsänen, A., Aurela, M., Juutinen, S., Kumpula, T., Lohila, A., Penttilä, T., Virtanen, T. (2019). Detecting northern peatland vegetation patterns at ultra-high spatial resolution. Remote Sensing in Ecology and Conservation, 6 (4), 457–471. doi: https://doi.org/10.1002/rse2.140
22. Lin, P., Lu, Q., Li, D., Chen, Y., Zou, Z., Jiang, S. (2019). Artificial intelligence classification of wetland vegetation morphology based on deep convolutional neural network. Natural Resource Modeling, 33 (1). doi: https://doi.org/10.1111/nrm.12248
23. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y. et. al. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1055–1059. doi: https://doi.org/10.1109/icassp40776.2020.9053405
24. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Lecture Notes in Computer Science, 833–851. doi: https://doi.org/10.1007/978-3-030-01234-2_49
25. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 31 (1), 4278–4284. doi: https://doi.org/10.1609/aaai.v31i1.11231
26. Bianco, S., Cadene, R., Celona, L., Napoletano, P. (2018). Benchmark Analysis of Representative Deep Neural Network Architectures. IEEE Access, 6, 64270–64277. doi: https://doi.org/10.1109/access.2018.2877890
27. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60 (6), 84–90. doi: https://doi.org/10.1145/3065386