ForestViT: A Vision Transformer Network for Convolution-free Multi-label Image Classification in Deforestation Analysis

Maria Kaselimi [1] **Athanasios Voulodimos** [2] **Ioannis Daskalopoulos** [2] **Nikolaos Doulamis** [1] **Anastasios Doulamis** [1]

Abstract

Understanding the dynamics of deforestation as well as land uses of neighboring areas is of vital importance for the design and development of appropriate forest conservation and management policies. In this paper, we approach deforestation as a multi-label classification problem in an endeavor to capture the various relevant land uses from satellite images. To this end, we propose a multi-label vision transformer model, ForestViT, which leverages the benefits of self-attention mechanism, obviating any convolution operations involved in commonly used deep learning models utilized for deforestation detection.

1. Introduction

Deforestation/land-use changes have impact in greenhouse gas emissions and are major drivers of regional climate change (de Bem et al., 2020). Human activities are among the main causes of global deforestation. The expansion of agriculture is a major driver of deforestation, with the construction of infrastructures such as roads or dams, together with mining activities and urbanization, constitute the main causes of deforestation. Identifying these driving forces of deforestation (agriculture, urbanization, infrastructures, etc.) of primary forest loss using satellite images is challenging, mainly due to the heterogeneity of the various drivers within images. Land uses located nearby a forest often act as driving forces of deforestation for these remaining forests. Understanding the dynamics of these changes can assist planning future conservation actions to prevent or mitigate adverse impacts. In this work, we formulate deforestation as a multi-label classification problem attempting to capture the various land uses related to deforestation.

Transformers (Vaswani et al., 2017) have recently demonstrated very good performance in a wide range of time-dependent applications. Transformer architectures are based on a self-attention mechanism that learns *long-range temporal dependencies* between elements of a sequence in the 1D space. Thus, the self-attention layers consider causality in a given sequence by learning the relationships between the token set elements. In the 1D space, transformers replace successfully the recurrent operations that process one local neighborhood at a time (Wang et al., 2018) and search for dependence (locally) at its previous time-related element. Moving from time (1D) to 2D space, the recently proposed vision transformer (Dosovitskiy et al., 2020) is an interesting attempt to showcase how (convolution-free) transformers can replace standard convolutions in deep neural networks in a similar manner transformers replace recurrent neural networks in 1D. There, attention mechanisms detect non-localized patterns and long-range pixel inter-dependencies (long-range spatial dependencies) (Cordonnier et al., 2019), (Wang et al., 2021). Vision transformers are applied on large-scale computer vision datasets, forming a CNN-free image classification model, able to capture *long-range spatial dependencies*.

1.1. Paper contribution

We propose, design and train a vision transformer model to identify the driving forces of deforestation of primary forest loss using satellite imagery in Amazon rainforest. This task is challenging to automate due to the heterogeneity of drivers within images and driver classes and the rapid evolution and changing of landscapes. Our model significantly benefits from multi-label image classification that simultaneously assigns multiple labels related to drivers of deforestation in near-the-forest areas in an image. Furthermore, vision transformer is exploited here as an efficient and scalable structure (Dosovitskiy et al., 2020) with multi-head attention mechanisms that derive long-range contextual/spatial relation between different areas in images (Bazi

[1] National Technical University of Athens, School of Rural, Surveying and Geoinformatics Engineering, Greece [2] University of West Attica, Department of Informatics and Computer Engineering, Greece. Correspondence to: Maria Kaselimi <mkaselimi@mail.ntua.gr>.

Tackling Climate Change with Machine Learning ICML 2021 Workshop. Financial support has been provided by the European Health and Digital Executive Agency (HADEA) under the powers delegated by the European Commission through the Horizon 2020 program "HEAlthier Cities through Blue-Green Regenerative Technologies: the HEART Approach", Grant Agreement number 945105, Copyright 2021 by the author(s).

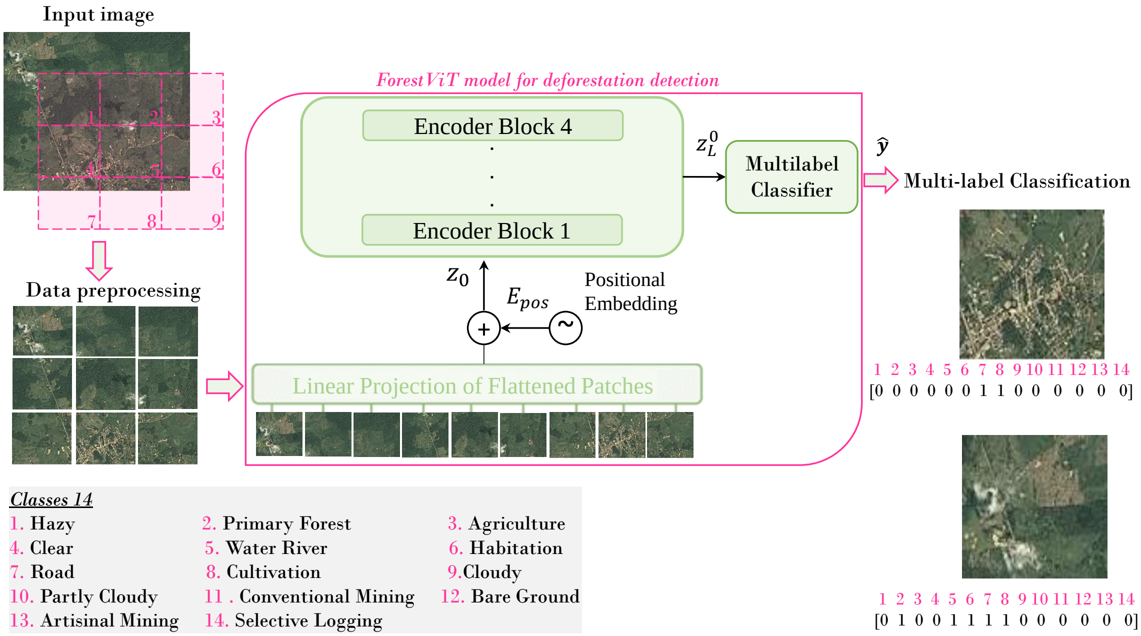

Figure 1. ForestViT is inspired by the vision transformer idea from (Dosovitskiy et al., 2020) and the encoder part of the NLP Transformer.

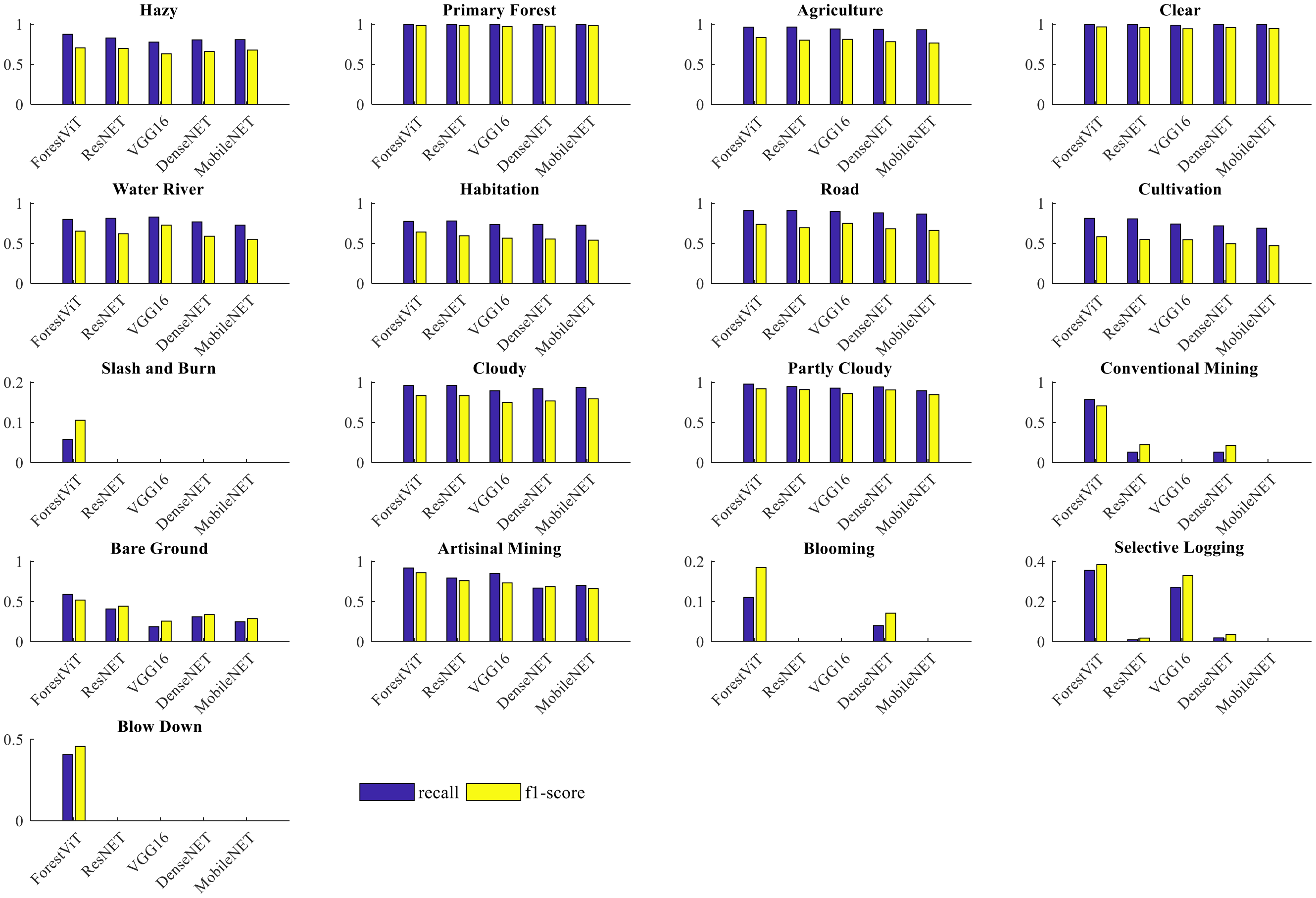

Figure 2. Per-class recall and F1-score evaluation of ForestViT, ResNET, VGG16, DenseNET and MobileNET models.

et al., 2021). Combining the multi-label classification mechanism along with the vision transformer architecture, the ForestViT model exploits the complex dependencies among visual features and labels (Lanchantin et al., 2020) in a satellite image, identifying forest areas at risk of greater levels of deforestation.

2. Multi-label classification deforestation

Multi-label classification (MLC) for deforestation detection in satellite images, refers to the task of assigning multiple labels to satellite images. Let us denote by $I = (I_1, \ldots, I_N) \in \mathbf{I}, \ \forall i = 1, ..., N$ a set of images, then assuming the label vector $\mathbf{y} = (y_1, \ldots, y_M) \in \mathbf{Y} = \{0, 1\}^M, \forall j = 1, ..., M$, where $C = c_1, \ldots, c_M, \ M = |C|$ is a finite set of predefined classes, the purpose is to decided the subset of classes that are found in the $i - th$ image through a learning process. Each label attribute y_j corresponds to the absence (0) or presence (1) of each class c_j. The classes are related to deforestation, land use and mangrove deforestation factors. In contrast to multi-class learning, alternatives are not assumed to be mutually exclusive, such that multiple classes may be associated with a single image (Mencía & Janssen, 2016). The problem at hand is to detect the multiple classes of an image related to deforestation. Let $\mathcal{H} : \mathbf{I} \rightarrow \mathbf{Y}$ be a multi-label classifier that estimates the label subset of $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_M)$ that comprises of various land use classes that appeared in each instance-image $x_i \in X$. Thus: $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_M) = \mathcal{H}(x_i; W)$.

3. ForestViT for deforestation detection

Fig. 1 depicts a schematic overview of the proposed ForestViT. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, we reshape image $I \in R^{h \times w \times b}$ into a sequence of flattened 2D patches $I_p \in R^{\,n \times (p^2 \cdot b)}$, where (h, w) is the spatial resolution of the original image, b is the number of bands/channels, (p, p) is the resolution of each image patch, and $n = \frac{h \cdot w}{p^2}$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. We map the vectorized patches I_p into a latent D-dimensional embedding space using a trainable linear projection. To encode the patch spatial information, we learn specific position embeddings which are added to the patch embeddings to retain positional information as follows (Dosovitskiy et al., 2020): $z_0 = [I_{c_j}; I_p^1 E; I_p^2 E; ...; I_p^n E] + E_{pos}$ where $E \in R^{(p^2 \cdot b) \times D}$ is the patch embedding projection, and $E_{pos} \in R^{(n+1) \times D}$ denotes the position embedding.

Then, the resulting sequence z_0 of embedding entities $x = (x_1, ..., x_n)$ serves as input to the transformer encoder. The transformer encoder has L encoder layers, and each encoder layer is composed of an multi-head self attention layer (MSA) and a feed-forward layer. The MSA layer consists of several attention layers running in parallel (Vaswani et al., 2017). The goal of self-attention is to capture the interaction among all the embedding entities x by encoding each entity in terms of the global contextual information. The output z is normalized using softmax operator to get the attention scores. Each entity then becomes the weighted sum of all entities in the sequence, where weights are given by the attention scores.

In order to encapsulate multiple complex relationships among different elements in the sequence, the multi-head attention mechanism in every l-layer, comprises $h-$ self-attention blocks. Each block has its own set of learnable weight matrices. For an input $-x$, the output z of the $h-$ self-attention blocks in multi-head attention is then concatenated into a single matrix $[z_l^0, z_l^1, ..., z_l^{H-1}]$ where $h = 0, ..., (H - 1)$ and projected onto a weight matrix W. Therefore the output of the l-th multi-head self-attention (MSA) layer is: $z'_l = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1...L$ where $LN(\cdot)$ denotes the layer normalization operator and z_l is the encoded image representation. Then, a fully connected feed-forward dense block follows in every encoder block $z_l = MLP(LN(z'_l)) + z'_l, \quad l = 1...L$ Lastly, a multi-label classifier makes the final predictions $\hat{\mathbf{y}}$ (see Fig. 1). We use a feedforward network (FFN) with two dense layers and a sigmoid activation function $\hat{\mathbf{y}} = FFN(z_L^0)$

4. Experimental evaluation

We compare our proposed convolution-free ForestViT model with traditional deep learning models that have convolutional layers as core structure, such as: VGG16 (Loh & Soo), ResNet50 (Budianto et al., 2017), DenseNET121 (Ching et al., 2019), and MobileNET (Howard et al., 2017), which are widely used models for remote sensing and deforestation applications or used as baseline to evaluate various vision transformers structures. The hyperparameters of these baseline models in our multi-label experiment are the VGG16, ResNet50, DenseNET121, and MobileNET with the image size and batch size hyperparameters equal to 256x256 and 128 respectively.

We utilize a dataset (https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/data) published in a Kaggle competition (by Planet company). All of the images derived from the Amazon basin. In our experiment, the images are classified in 14 classes and the labels are broken into three groups: atmospheric conditions, common land cover/land use phenomena, and rare land cover/land use phenomena. Here, each entry is assigned to one or more classes.

	Classes													
Techniques	1	2	3	4	5	6	7	8	9	10	11	12	13	14
ForestViT	95.1	**96.7**	**88.2**	**95.0**	84.1	**92.1**	86.7	**87.1**	**99.6**	**96.9**	**99.9**	**99.7**	**99.1**	**99.7**
(Budianto et al., 2017)	**95.2**	96.5	85.4	93.7	81.4	90.3	83.8	85.2	99.6	96.6	99.8	99.5	99.1	99.7
(Loh & Soo)	94.0	94.6	86.6	91.5	**88.4**	89.6	**87.7**	86.3	99.6	94.5	99.8	99.4	99.0	99.7
(Ching et al., 2019)	94.5	95.2	84.0	93.7	80.0	89.2	83.3	83.8	99.6	96.4	99.8	99.4	99.1	99.7
(Howard et al., 2017)	94.9	96.3	82.6	91.8	77.8	88.6	81.9	82.9	99.6	94.1	99.8	99.3	99.1	99.7

Table 1. Per-class accuracy evaluation of ForestViT, ResNET, VGG16, DenseNET and MobileNET models.

Techniques	Overall Prec.	Overall Rec.
ForestViT	0.80	0.94
(Budianto et al., 2017)	0.77	0.93
(Loh & Soo)	0.78	0.92
(Ching et al., 2019)	0.75	0.92
(Howard et al., 2017)	0.74	0.91

Table 2. Micro-averaged recall and precision metrics for ForestViT, ResNET, VGG16, DenseNET and MobileNET.

4.1. Implementation details

Image Generator. The dataset was split in $50/20/30$ train/valid/test sets using an image generator. The images with size $[256 \times 256 \times 4]$, where the last channel is the Near Infrared in our case, are re-scaled to $[0, 1]$. Then, each image was divided into patches of size $[16 \times 16 \times 4]$.

Vision Transformer Encoder. The Vision Transformer encoder accepts the images as input and produces a 14×1 tensor containing each label's probabilities as output. The last activation function is sigmoid, so that each distinct probability in the output tensor can take values in $[0, 1]$ regardless of the probabilities of the rest of the labels (multi-label classification problem). The transformer encoder consists of encoder blocks. Each encoder block contains two sublayers: multi-head self-attention and positionwise feed-forward networks, where a residual connection followed by layer normalization is employed around both sublayers. Our implementation contains 4 transformer encoder blocks, each one with an eight-head self attention mechanism.

Training process and optimization. We used Adam optimizer with a learning rate of $1e^{-4}$ and binary cross-entropy loss function. The training process of the model ran on a GTX 1060 6GB on a laptop. The training process took around 42 epochs to finish (max allowed epochs was 50).

4.2. Evaluation of Deep Learning techniques for deforestation detection

Per-class analysis. Table 1 demonstrates the proposed model performance over the unseen (test) data. To verify the performance of our self-attention models, we use the unseen (test) set to assess the model performance to data totally outside the training phase. The results have been obtained using the accuracy objective criterion (see Section 4.2) for the tested set, separately for each category. We can see that high-performance results are obtained. Our convolution-free ForestViT model has slightly better results compared to ResNet VGG16, DenseNET and ModileNet approaches. Diving into the the different classes of the deforestation dataset, the per-class recall and F1-score objective criteria are depicted in Fig. 2.

The 'hazy', 'primary', agriculture', 'clear', 'water river', 'cloudy', 'partly cloudy' classes appear the best performance compared to the 'habitation', 'road', 'cultivation' and 'artisinal mining' ones, that appear lower performance classification. The performance classification for the 'conventional mining' and 'selective logging' classes, that are rarely appeared in the dataset, is lower than the accuracy achieved in above mentioned classes. However, the performance of ForestViT is quit good compared to the performance achieved using most of the state-of-the-art models (e.g., DenseNET, MobileNET, ResNET, and VGG16 networks).

Model's overall accuracy assessment. In Table 2, we report the micro-averaged recall and precision metrics on the test set for ForestViT, ResNET, VGG16, DenseNET, MobileNET networks. Given that the micro-averaged multi-label performance metrics are defined by averaging over both labels and examples, they adequately capture the per-class performance imbalance, also observed in Fig. 2. Thus, the overall precision expressed as micro-averaged precision, is averaged down to 0.80 for ForestViT and < 0.80 for the other models used for comparison.

5. Conclusions

The fact that the human landscape is rapidly evolving emphasizes the need for the analysis of deforestation data, the update of deforestation risk maps and the appropriate adaptation of mitigation strategies. In order to capture co-occurrence patterns among labels, this paper proposes a multi-label vision transformer classifier, ForestViT, to detect dependencies among the output variables. We show that the self-attention between neighboring image patches in ForestViT and without any convolution operations achieves superior performance in multi-label classification of deforestation images compared to state of the art deep learning models.

References

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., and Ajlan, N. A. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3):516, 2021.

Budianto, F., Westman, N., and Ni, B. Understanding the amazon basin from space. 2017.

Ching, D., Li, Y., and Song, G. Understanding the amazon from space. 2019.

Cordonnier, J.-B., Loukas, A., and Jaggi, M. On the relationship between self-attention and convolutional layers. *International Conference on Learning Representations (ICLR) 2020*, 2019.

de Bem, P. P., de Carvalho Junior, O. A., Fontes Guimarães, R., and Trancoso Gomes, R. A. Change detection of deforestation in the brazilian amazon using landsat data and convolutional neural networks. *Remote Sensing*, 12(6):901, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 10 2020.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. General multi-label image classification with transformers. *arXiv preprint arXiv:2011.14027*, 2020.

Loh, A. and Soo, K. Amazing amazon: Detecting deforestation in our largest rainforest.

Mencía, E. L. and Janssen, F. Learning rules for multi-label classification: a stacking and a separate-and-conquer approach. *Machine Learning*, 105(1):77–126, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018. doi: 10.1109/CVPR.2018.00813.

A. Performance metrics

We employ several metrics to comparatively evaluate the proposed and existing methods.

Per-class accuracy. In order to objectively evaluate our results, the metrics of accuracy and recall are considered. Accuracy (ACC_{c_i}) is defined as:

$$ACC_{c_i} = \frac{TP_{c_i} + TN_{c_i}}{TP_{c_i} + TN_{c_i} + FP_{c_i} + FN_{c_i}} \tag{1}$$

where the nominator contains the true positives (TP_{c_i}) and true negatives (TN_{c_i}) samples, while denominator contains the TP_{c_i} and TN_{c_i} and false positives (FP_{c_i}) and false negatives (FN_{c_i}). Precision (PR_{c_i}), recall(REC_{c_i}) and F1-score ($F1_{c_i}$) are given as:

$$PR_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FP_{c_i}},\ REC_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FN_{c_i}} \tag{2}$$

$$F1_{c_i} = \frac{2 \cdot PR \cdot REC}{PR + REC} \tag{3}$$

Overall accuracy. To measure the effectiveness in a multi-label classification problem, averaging metrics is also required. In micro-averaging all $TP_{c_i}, TN_{c_i}, FP_{c_i}$ and FN_{c_i} for each class $c_i, \forall c_i \in C$ are averaged,

$$PR^{micro} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} (TP_{c_i} + FP_{c_i})} \tag{4}$$

$$REC^{micro} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} (TP_{c_i} + FN_{c_i})} \tag{5}$$

Multi-label accuracy. In multi-label classification, a misclassification is no longer a hard wrong or right. A prediction containing a subset of the actual classes should be considered better than a prediction that contains none of them, i.e., predicting two of the three labels correctly is better than predicting no labels at all. Hamming-Loss is the fraction of labels that are incorrectly predicted. Given an i -image input to the model, and assuming an output vector

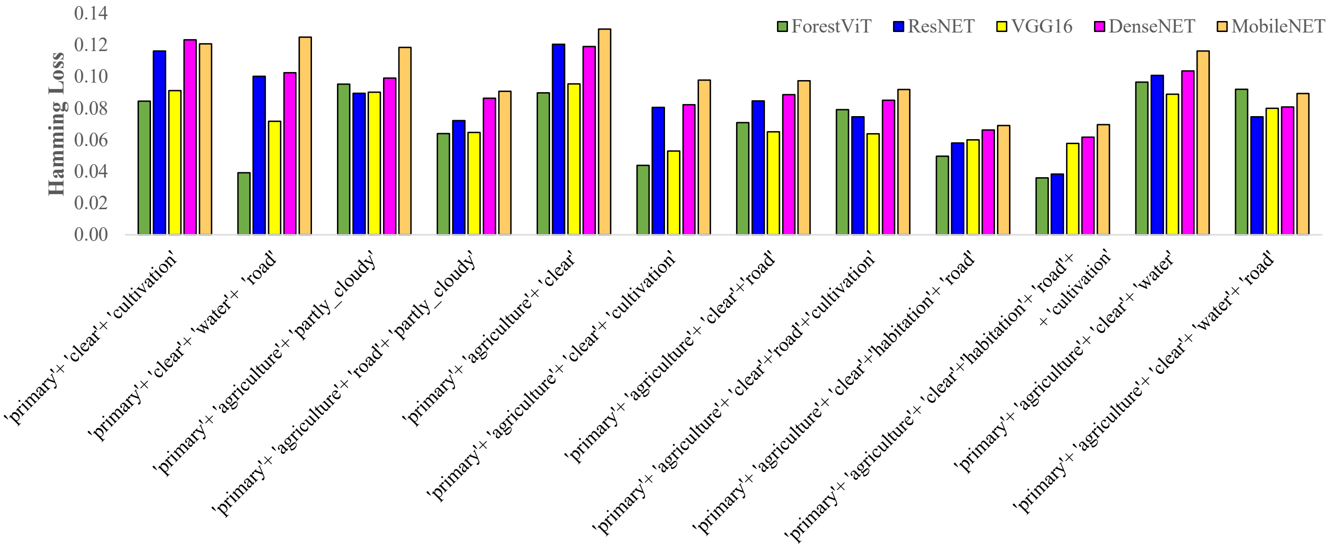

Figure 3. Hamming loss metric for the ForestViT model and benchmark techniques for 12 different multi-label combinations.

with binary elements $\hat{y}_{i,j}$, dimensions $[1 \times N]$ and the corresponding ground truth with element $y_{i,j}$, the hamming loss metric is:

$$HM = \frac{1}{|N|\cdot|M|} \sum_{i=1}^{N} \sum_{j=1}^{M} 1_{y_{i,j} \neq \hat{y}_{i,j}} \qquad (6)$$

B. Deforestation risk analysis enabled by multi-label classification.

We hereby present a deforestation risk analysis as enabled by the described multi-label classification problem. Fig. 3 illustrates the ForestViT model performance in detecting land-use properties acting as drivers for deforestation among with the existence of forest areas (class ”primary’) in the same figure (multi-label classification). Hamming loss metric shows that our proposed ForestViT model has better performance compared to ResNET, VGG16, DenseNET and MobileNET method.

As mentioned above, in multi-label classification, a misclassification is no longer a hard wrong or right, given that a prediction containing a subset of the actual classes should be considered better than a prediction that contains none of them. However, in our application the existence of the ’primary’ class in an image among with an additional label related with land use that possibly could act as a driving factor of deforestation, could indicate possible areas with high risk of deforestation. Thus, the consecutive existence of both classes assigned corrected, is of great importance for our application scenario.

In our last scenario, we consider seven different cases that contain images having at least two different labels. The primary forest label is included as the standard label for all the cases and the second label varies and is one of the selected drivers (agriculture, cultivation, mining, road infrastructure, habitation, logging and bare ground) for each case. In this case, we compare the probability to detect the primary forest label in those images with the probability of jointly detecting both the primary forest and the driver respective label.

$$P_{prim} = \frac{n_{prim}}{N_{prim,x}}, \quad P_{prim,x} = \frac{n_{prim,x}}{N_{prim,x}} \qquad (7)$$

where, x stands for the drivers [agr:agriculture, cul:cultivation, min:mining, roa:road, hab:habitation, log:logging, bar:bare ground], $N_{prim,x}$ is the total number of images that include at least the primary forest label and the driver $-x$ label, n_{prim} is the subset from the $N_{prim,x}$ set of mages that correctly identified as primary forest, $n_{prim,x}$ is the subset from the $N_{prim,x}$ set of mages that correctly identified both as primary forest and $x-$ label.

Fig. 4, demonstrates the probability of successfully detecting primary versus the probability of detecting both primary and a driver related class. In particular, we compare P_{prim} and $P_{prim,x}$ values per architecture. As observed, the primary forest class is accurately detected for all the examined cases and deep learning architectures (see also Fig. 2). The cases that include the agriculture and road labels, are identified with high accuracy. Cultivation, habitation and mining labels follow in accuracy performance, whereas logging and

Primary + Agriculture

	P_{prim}	$P_{prim,agr}$
ForestViT	0.99	0.96
ResNET	0.99	0.96
VGG16	0.99	0.93
DenseNET	0.99	0.94
MobileNET	0.99	0.93

Primary + Road

	P_{prim}	$P_{prim,roa}$
ForestViT	0.99	0.90
ResNET	0.99	0.90
VGG16	0.99	0.89
DenseNET	0.99	0.88
MobileNET	0.99	0.86

Primary + Cultivation

	P_{prim}	$P_{prim,cul}$
ForestViT	0.99	0.82
ResNET	0.99	0.80
VGG16	0.99	0.74
DenseNET	0.99	0.72
MobileNET	0.99	0.69

Primary + Habitation

	P_{prim}	$P_{prim,hab}$
ForestViT	0.99	0.76
ResNET	0.99	0.77
VGG16	0.99	0.72
DenseNET	0.99	0.72
MobileNET	0.99	0.71

Primary + Mining

	P_{prim}	$P_{prim,min}$
ForestViT	0.99	0.77
ResNET	0.99	0.77
VGG16	0.99	0.00
DenseNET	0.99	0.77
MobileNET	0.99	0.00

Primary + Logging

	P_{prim}	$P_{prim,log}$
ForestViT	0.99	0.36
ResNET	0.99	0.16
VGG16	0.99	0.27
DenseNET	0.99	0.16
MobileNET	0.99	0.00

Primary + Bare ground

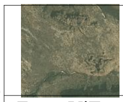	P_{prim}	$P_{prim,bar}$
ForestViT	0.99	0.50
ResNET	0.99	0.31
VGG16	0.99	0.10
DenseNET	0.99	0.21
MobileNET	0.99	0.17

Figure 4. The probability of successfully detecting the probability P_{prim} versus the probability $P_{prim,x}$ of detecting both primary and a driver related class, for the ForestViT, ResNET, VGG16, DenseNET and MobileNET models.

bare ground labels appear the worst performance. Here, we highlight that logging and bare ground classes are rare occurrences (minority class) in dataset, and this explains their low performance.