



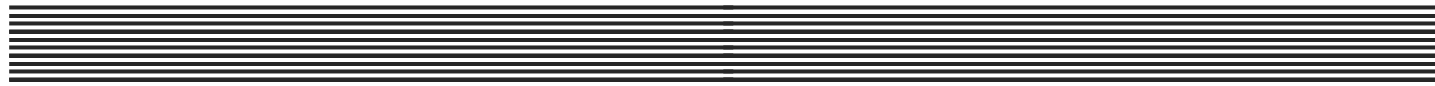
National Technical University of Athens (NTUA)



Thirty-eighth **International Conference on Machine Learning**



ICML 2021 Workshop: Tackling Climate Change with Machine Learning



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

Maria Kaselimi, Athanasios Voulodimos, Ioannis Daskalopoulos, Nikolaos Doulamis, Anastasios Doulamis



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

- **Deforestation** has impact in **greenhouse gas emissions** and land-use changes are major drivers of **regional climate change**.
- **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 actions** to prevent or mitigate adverse impacts.



Land uses act as driving forces of deforestation

Virgin
Forest

Virgin Forest



Bare Ground



"Artisinal" Mining



Conventional Mining



Land
uses

Water River



Selective Logging



Slash and Burn



Road



Land
uses

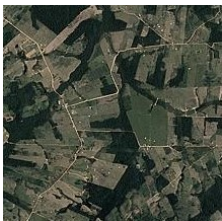
Habitation



Cultivation



Agriculture





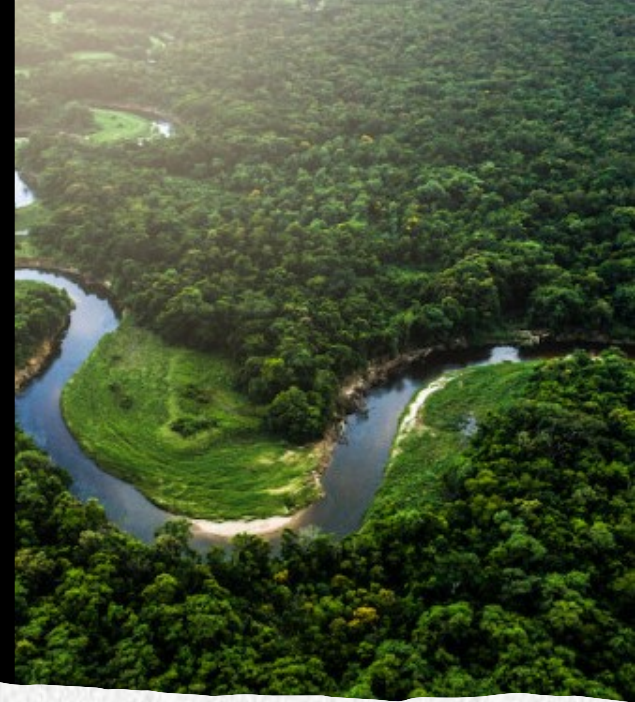
Agriculture+
Road+
Primary



Agriculture+
selective
logging



Shifting
cultivation+
primary

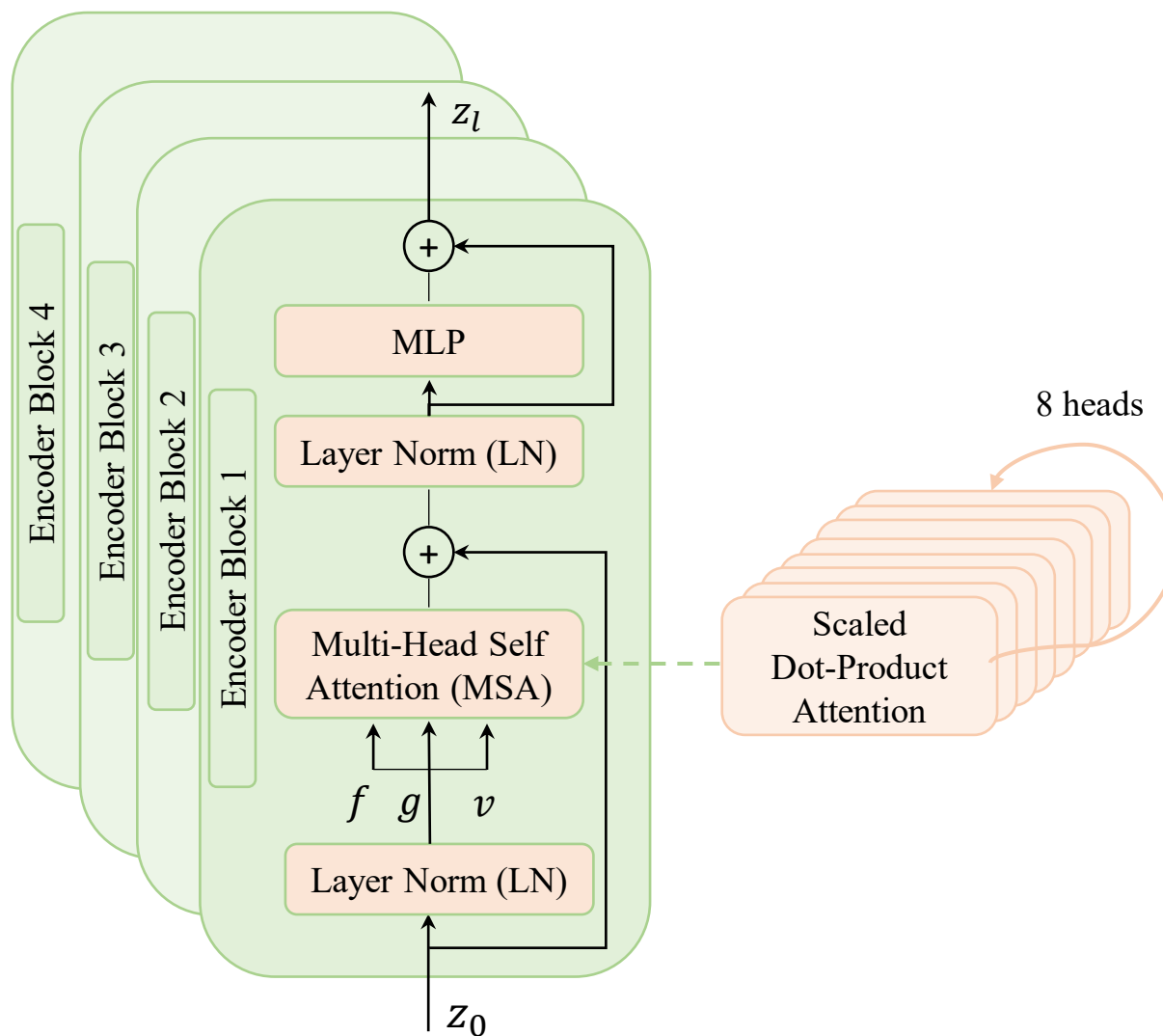


Understanding the Amazon Rainforest with Multi-Label Classification

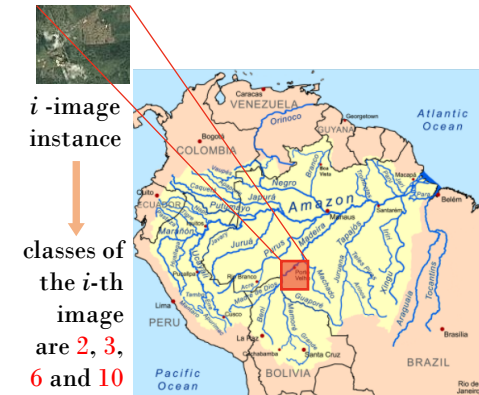
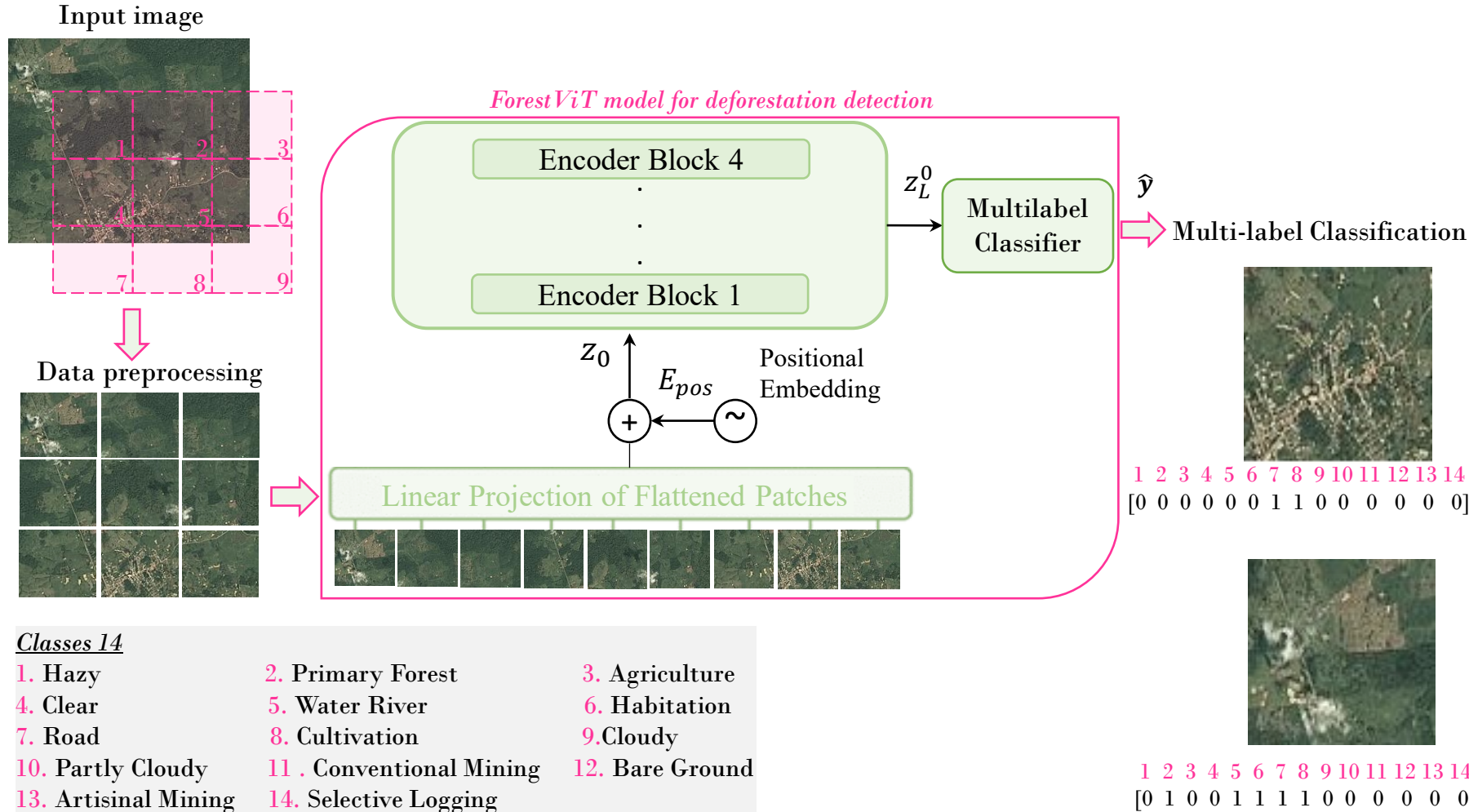
The Amazon Rainforest Case Study

The effect of attention mechanism

attention mechanisms
detect non-localized
patterns and long-range
pixel inter-dependencies
(long-range spatial
dependencies)



ForestViT: a vision transformer for multi-label classification applied to deforestation



- Classes 14**
1. Hazy
 2. Primary Forest
 3. Agriculture
 4. Clear
 5. Water River
 6. Habitation
 7. Road
 8. Cultivation
 9. Cloudy
 10. Partly Cloudy
 11. Conventional Mining
 12. Bare Ground
 13. Artisanal Mining
 14. Selective Logging

Per Class Analysis

Accuracy:

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

Techniques	Classes													
	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

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

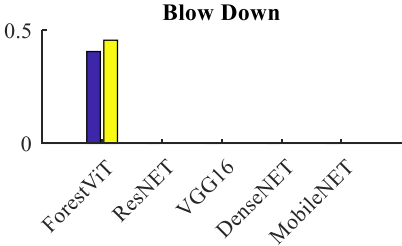
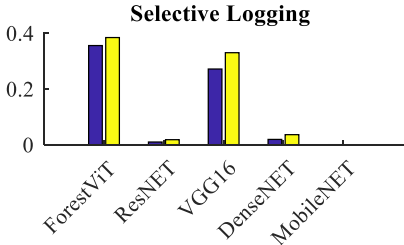
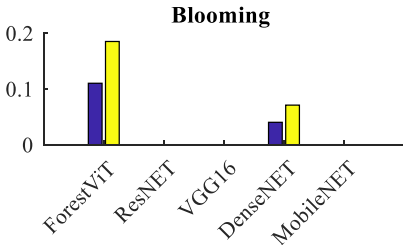
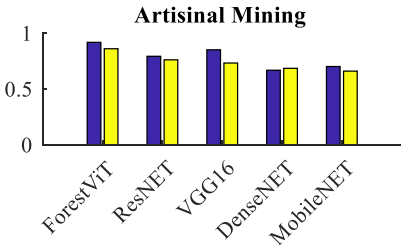
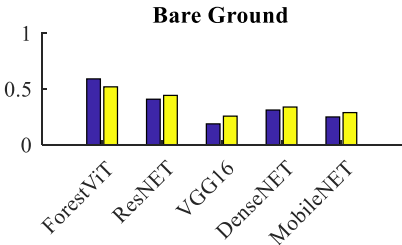
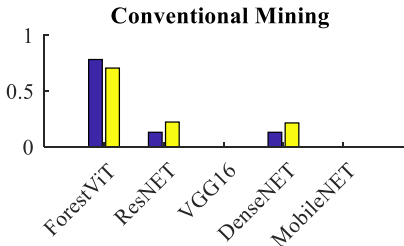
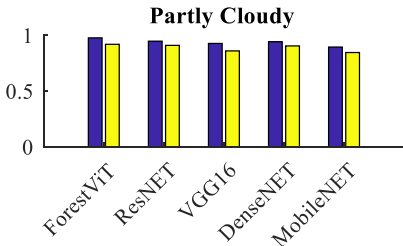
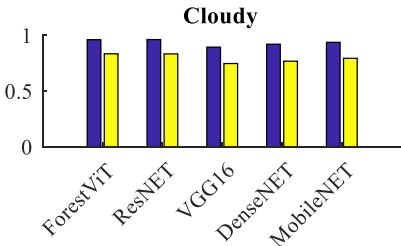
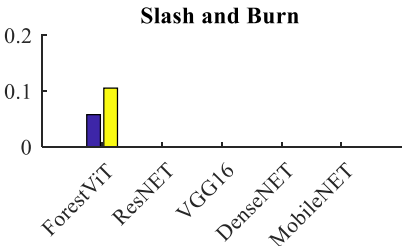
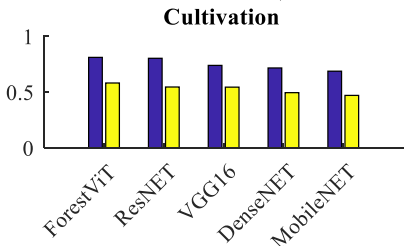
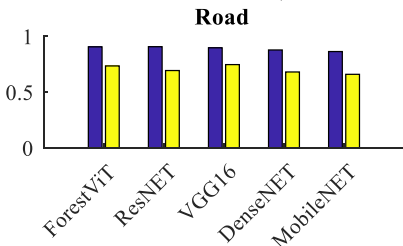
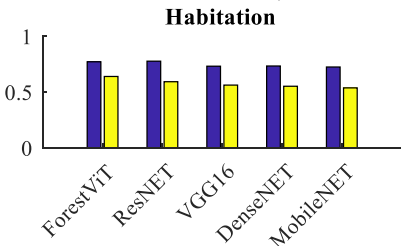
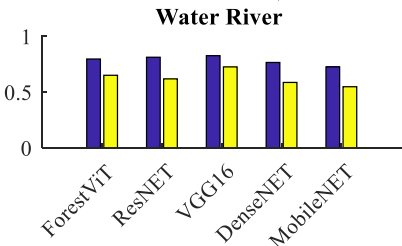
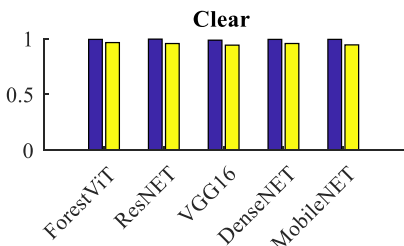
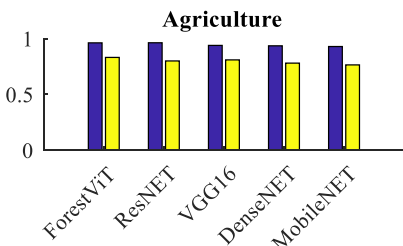
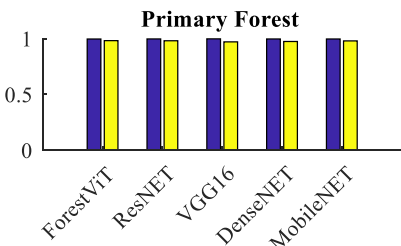
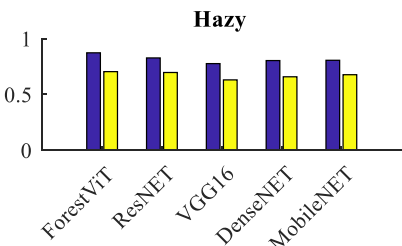
Per Class Analysis

Precision:

$$PR_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FP_{c_i}}$$

Recall:

$$REC_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FN_{c_i}}$$



recall f1-score

Overall accuracy

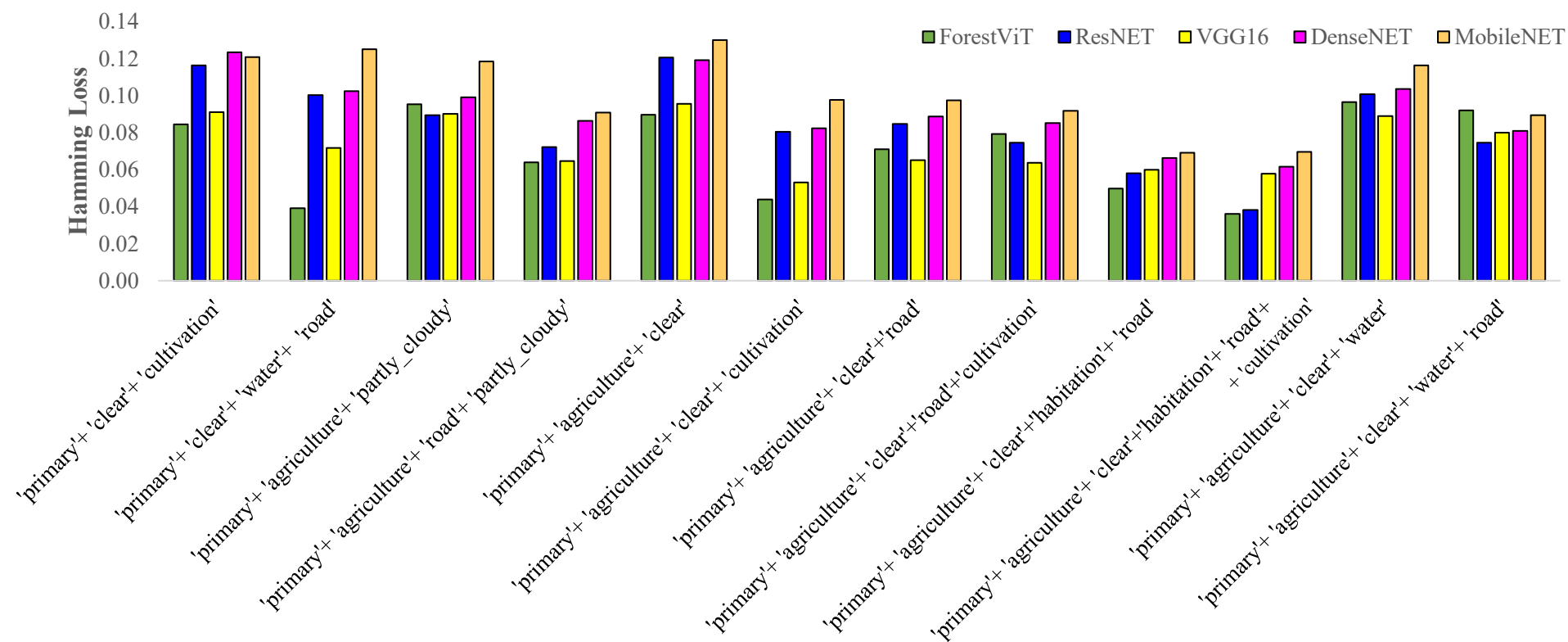
Precision: $PR_{micro} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} (TP_{c_i} + FP_{c_i})}$

Recall: $REC_{micro} = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} (TP_{c_i} + FN_{c_i})}$








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

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. The bigger the hamming loss value is, the worst the performance of the model is.



Multi-label accuracy

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 + 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 + 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 + 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
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 + 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 + 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

In our last scenario, we consider ***seven different cases*** that contain images having at least **two** different labels. The primary (virgin) 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**.