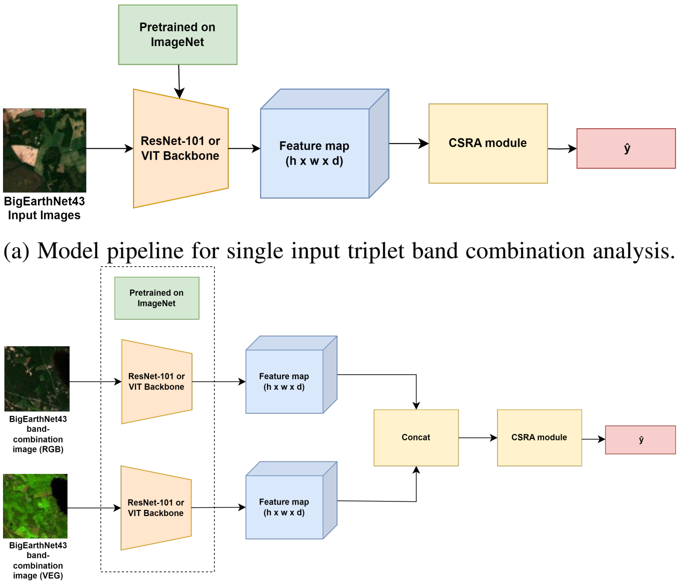
Multi-band Satellite Image Analysis for Multi-label Classification

Sarah Shamina Abdul Rauf, Mas Ira Syafila Mohd Hilmi Tan, and Yuen Peng Loh
Multimedia University, Malaysia
E-mail: {1201102479@student, masirasyafila@, yploh@}.mmu.edu.my

***Abstract*—Multispectral satellite imagery captures rich information of the Earth surface from an extensive range of wavelengths that enhances the understanding and discrimination of Earth objects. When integrated with multi-label classification, it has great potential to provide perceptual comprehension of land cover that exceeds the details observable from the visible light spectrum (RGB). However, the variety of spectral bands each capture a different aspect of the earth, thus can be challenging to identify the optimal spectral band, and their combinations for a classification task. To overcome these issues, this study leverages on deep learning by exploring various band combinations and neural network architectures for multi-label classification. Specifically, the performance of triplet band combinations was explored and compared against standard RGB imagery, using ResNet101 as the backbone and the incorporation of the Class Specific Residual Attention (CSRA) mechanism. We then propose the multi-stream triplet band fusion model with Vision Transformer (ViT) backbone and CSRA for multi-label satellite image classification. We found that in the single triple band input approach, the ShortWave InfraRed (SWIR1 or SWIR2) combinations is able to improve F1 scores by 1.25% to 2.02% compared to models using only RGB bands. More notably, our proposed multi-stream approach that fuses the triplets of RGB and Vegetation bands outperformed all other models with an F1 score of 0.7484, consistently surpassing the ResNet-101 backbone baseline with similar configurations. These findings reveal the potential of band combination approaches to enhance the Earth object discrimination and land cover analysis.**


Pretrained on ImageNet
BigEarthNet43 Input Images
ResNet-101 or VIT Backbone
Feature map (h x w x d)
CSRA module
ŷ
Pretrained on ImageNet
BigEarthNet43 band-combination image (RGB)
ResNet-101 or VIT Backbone
Feature map (h x w x d)
Concat
CSRA module
ŷ
BigEarthNet43 band-combination image (VEG)
ResNet-101 or VIT Backbone
Feature map (h x w x d)

(a) Model pipeline for single input triplet band combination analysis.

(b) Proposed multistream model with CSRA for two triplet inputs for improved multi-label classification.

Fig. 1: Multi-band multi-label satellite image classification frameworks.

I. Introduction

Multispectral satellite imagery, which captures a wide range of spectral bands beyond the visible spectrum, has become an invaluable tool for detailed Earth observation. Unlike traditional RGB images, multispectral images offer enhanced discrimination of various land cover types as they provide additional spectral information that can differentiate objects appearing homogeneous in RGB imagery [1], [2]. This added spectral resolution allows for more precise analysis and classification of land cover, which is important for applications like environmental monitoring, agriculture, and urban planning [2].

The varied reflectance behaviors of different Earth objects across multiple wavelengths enable the experimentation with different band combinations to improve object differentiation. However, utilising multispectral imagery for classification poses significant challenges. For instance, using all available bands for the classification of remote sensing images can lead to the curse of dimensionality [3], where the increase in noise worsens classification performance [4]. Reducing the number of bands is a way to avoid this problem, but the selection of optimal bands can be difficult due to the sheer number of spectral bands available [5]. Additionally, in multi-label datasets, the multiple labels can co-occur in complex patterns, where a single image may contain multiple classes such as water, vegetation, and urban areas, which can appear simultaneously but with varying spatial distributions. This makes it difficult for a model to correctly identify and classify all relevant labels in a single representation.

In this study, we aim to explore the effect of different combinations of multispectral bands for multi-label satellite imagery classification. We utilized the BigEarthNet43 [6] multispectral dataset for experimentation with various band combinations and trained various multi-label classifiers by adopting the ResNet101 [7] as the baseline backbone and the Class Specific Attention (CSRA) module [8] for multi-label classification. Lastly, we designed a multi-stream model with ViT backbone, that takes in two band triplets as “images” and concatenates their feature maps for the final multi-label classification by the CSRA module as shown in Fig. 1.

The contributions of the work can be summarised as follows:

- A comparative study on multispectral band combinations for enhancing multi-label satellite image classification. To the best of our knowledge, there has yet to be any work that explores such band combinations for multi-label classification using multispectral data.
- An empirical analysis on band combinations and their contributions to specific classes in multi-label classification. We found that combinations containing ShortWave InfraRed (SWIR1 or SWIR2) improved F1-scores by 1.25-2.02% over RGB-only models, consistently.
- A novel multi-stream approach using the Vision Transformer (ViT) backbone, fusing RGB and vegetation (VEG) band triplets with CSRA module for multi-label satellite image classification. The model achieved an F1 score of 0.7484, outperforming all other tested models.

II. Related Work

With the recent rise in the number of large scale multi-label satellite image datasets, this opens up more avenues for exploring multi-label satellite image classification. Despite this, the research in multi-label classification specifically for multispectral satellite datasets with a large number of samples and class labels is still lacking. Moreover, in the image classification domain, RGB images are most widely used.

For instance, [9] introduced a CNN-based model for hierarchical multi-label annotation, incorporating attention modules and skip-layer connections to enhance discriminative capabilities. Using Inception-ResNet-v2 as a feature extractor pretrained with ImageNet weights, the model integrates attention modules and skip-layer connections in the classification branch, and employs embedding learning to preserve scene-level semantic similarity. It achieved significant improvements in the harmonic mean of label-based and example-based F1 scores (H-F1) when evaluated on datasets like RGB UC-Merced, Ankara, and RGB Multi-label AID.

Alternatively, optimal band combinations had been explored for Landsat-8 in land use and land cover (LULC) classification using Support Vector Machine (SVM) [3]. Addressing the 'curse of dimensionality', the study compared various three- and four-channel band combinations to using all bands, selecting combinations through correlation analysis. The SVM model was specifically trained and evaluated on Landsat-8 with 4 LULC classes in single label classification. Similarly, [5] used a Mask-RCNN to explore optimal band combinations for mapping permafrost tundra landforms, targeting tussock and non-tussock sedge vegetation. The WorldView-02 satellite data was evaluated with five three-band combinations. The Mask-RCNN model, with a ResNet-101 backbone pre-trained on the COCO dataset and retrained on 40,000 annotated ice-wedge polygons, processed 200x200 pixel image tiles. They found that triplets of band combinations were most effective.

As a whole, while there have been advancements in multi-label satellite image classification and optimal band combinations, there is considerable potential for further exploration in multispectral datasets with extensive samples and class labels. Thus, our study aims to address this gap by investigating novel band combinations and advanced models to enhance classification accuracy and leverage the full potential of multispectral data.

III. Methodology

A. Band Combinations

In multispectral satellite imagery, combining and rearranging spectral bands into RGB channels is a commonly used technique [1]. This technique produces false colour composites images, which make discerning certain classes easier as it is able to highlight various difference Earth objects.

Table I shows the types of spectral bands captured by the Sentinel-2 satellite that is used to collect the BigEarthNet43 dataset. Spatial resolution for the Sentinel-2 ranges from 10m to 60m, where the 10m bands are standard satellite image bands used for land cover classification, and 20m bands are used for vegetation monitoring, whereas bands relating to atmospheric correction have a resolution of 60m [10].

TABLE I: Sentinel-2 Spectral Bands [11].

Band	Description	Resolution (m)	Center (nm)	Band (nm)
B01	Coastal aerosol	60	443	20
B02	Blue	10	490	65
B03	Green	10	560	35
B04	Red	10	665	30
B05	Vegetation red edge	20	705	15
B06	Vegetation red edge	20	740	15
B07	Vegetation red edge	20	783	20
B08	NIR (Near Infrared)	10	842	115
B09	Water vapor	60	940	20
B11	SWIR1	20	1610	90
B12	SWIR2	20	2190	180
B8A	Narrow NIR (Near Infrared)	20	865	20

The main hypothesis of this exploration is the positive impact of band combinations on classification performance, where specific combinations can result in significant improvements in accuracy [12]. As such, various triplet sets of band combinations were formed based on their use in satellite imagery analysis for both Sentinel-2 and Landsat satellites, where each band combination serves to emphasise distinct features, as described in [13]. Table II shows the specific combinations for our investigations, and Fig. 2 illustrates the visualization of color composites created from these band combinations showing different resolutions and compositions that my not be as intuitive for human observation.

To elaborate, the RGB composite closely mirrors human vision, making it effective for general analysis, where healthy vegetation appears green, while unhealthy vegetation looks brown or yellow. It is particularly useful for analysing water bodies but often appears low in contrast due to atmospheric scattering [13]. False Color (Urban) composite, is used for distinguishing urban areas from vegetation, displaying water bodies as blue-black and urban areas as grey-purple [13]. The Color Infrared composite, is used in differentiating vegetation types and health, showing healthy vegetation in red hues [12], [13]. The Agriculture composite is used for crop health assessment, with vibrant greens indicating healthy crops [13]. The Land/Water composite is often used for distinguishing land from water, useful for highlighting flooded areas [13]. The Natural (Atmospheric Removal) composite provides a clearer view by reducing atmospheric interference, helping in agricultural and post-fire analysis [13]. The Shortwave Infrared composite is useful in vegetation studies, highlighting healthy vegetation in deep red [13]. Lastly, the Vegetation (VEG) Analysis composite is effective for soil moisture and vegetation health analysis, making it suitable for detecting plant vigor and water bodies [13].

TABLE II: Band combinations

Colour-composite	Abbreviation	Sentinel-2 Band combination (RGB channel)
RGB Colour	RGB	4, 3, 2
False Colour (Urban)	FC	12, 11, 4
Colour Infrared	IR	8, 4, 3
Agriculture	AGRI	11, 8, 3
Land/Water	LW	8, 11, 4
Natural (Atmospheric Removal)	NWAR	12, 8, 3
Shortwave Infrared	SWIR	12, 8, 4
Vegetation Analysis	VEG	11, 8, 4

B. Multi-label Classification

To address the challenge of classifying images with multiple labels consisting varying object locations, and sizes, we explore both CNN and ViT backbones for feature extraction and adopted the Class-Specific Residual Attention (CSRA) module [8] that enhances spatial attention for each object class for the task, as shown in Fig. 1.

Particularly, the pipeline begins with an image input into the backbone, which extracts image features and outputs a feature tensor of dimensions $d \times h \times w$. This tensor is then fed into a classifier composed of a fully connected layer with a kernel size of 1×1, producing score tensors of shape $C \times h \times w$, where C is the number of classes. These tensors provide class-specific attention scores for each location within the feature tensor, representing the probability of each class being present at specific locations. Residual attention is applied to these score tensors using either a single-head or multi-head attention mechanism. Single-head attention computes one set of class-specific attention scores for the entire tensor, whereas multi-head attention generates multiple sets of scores, resulting in various logits combined to form the final prediction..

IV. Experiments

In our experiments, we conducted comprehensive evaluations using the BigEarthNet43 dataset [6], consisting 519,284 multispectral images across 43 class labels with diverse environmental and land-use categories. As the dataset features significant imbalanced class distributions, we randomly subsampled 118,065 images for our study. Our experimental setup included a consistent 70/15/15 train-validation-test split, resulting in 82,645 training images, 17,710 validation images, and 17,710 testing images, and all images were normalized to the range [0, 1].

A. Implementation

1) Single Input Triplet Model: The baseline for our experimentation involves using the single-headed CSRA module as a classifier, combined with either the CNN (ResNet-101) or ViT backbones. Two baseline models will be created and trained with BigEarthNet43 RGB images: one using a Resnet-101 backbone and the other using the ViT backbone. Both models were trained for 100 epochs, with a learning rate of 0.01 and a batch size of 16. Furthermore, the Resnet-101-based model used an input size of 120×120, whereas ViT-based model used 224×224 to support a reasonable resolution for the patch-splitting mechanism. The pipeline of the baseline model is shown in Fig. 1a.

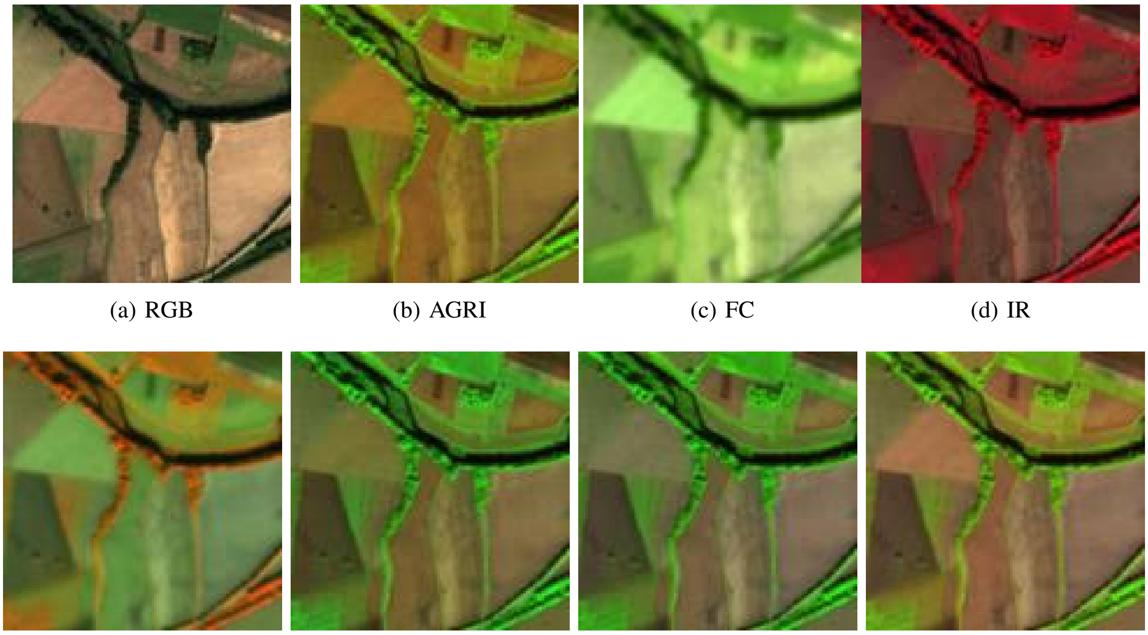
2) Proposed Model: In our investigation, we aim to explore two aspects: first, identifying the optimal band combination that yields the best performance for the BigEarthNet43 image dataset, and second, examining the impact on performance when two different color-composite images are input into the model.

For our first approach, we utilised the same pipeline as in Fig. 1a, with the Resnet-101 based backbone. For the input, we used the band combinations listed in Table II to create eight false colour-composite image. The hyperparameters of this model will follow that of the Resnet-101 baseline, using 120×120 image inputs. For the second approach, we implemented a two-input multi-stream CNN model with feature map concatenation, using the single-headed CSRA classifier with both ResNet-101 and VIT backbones. The images for this model use band combinations (4, 3, 2) and (11, 8, 4), corresponding to the RGB and Vegetation Analysis colour composites respectively. The different colour composites are passed through the backbone separately, and the feature maps generated are concatenated together by the channel dimension. Just before being passed to the CSRA classifier, a dropout layer is used for regularisation. This model will be trained with 100 epochs, a learning rate of 0.001, a batchsize of 16 and a dropout value of 0.5. Images with dimensions 120×120 and 224×224 are used, respectively for the model trained with a ResNet-101 backbone and VIT backbone. Fig. 1b shows the pipeline for the second proposed solution.

B. Evaluation

This section presents an evaluation of both baseline and proposed approaches. The performance metrics used for the evaluation are commonly used in multi-label classification tasks: mean average precision (mAP), overall precision, overall recall and overall F1 score. For clarity, we indicate each different models trained with the respective band combinations based on the convention of {backbone}-{band combination} in the subsequent results and discussions.

As seen in Table III, the ViT-RGB was able to outperform Resnet101-RGB across all metrics. ViT-RGB achieved a higher mAP of 0.5465 compared to 0.4791 for Resnet101-RGB,



(a) RGB (b) AGRI (c) FC (d) IR

(e) LW (f) NWAR (g) SWIR (h) VEG

Fig. 2: Visualisation of colour composites created from band combinations in Table II.

suggesting better overall performance in terms of ranking relevant instances higher. Additionally, ViT-RGB showed higher precision (0.8090), and higher recall (0.6846) which indicates that more accurate and relevant predictions are being returned. The F1 Score, was also higher for ViT-RGB (0.7416 vs. 0.7039). These results suggest that the transformer-based ViT architecture is more effective than the Resnet-101 architecture when combined with the CSRA module for the classification of BigEarthNet43 satellite images.

Table III provides detailed evaluation metrics for the proposed solutions. In the single image models, the most notable observation is that, band combinations containing bands SWIR1 or SWIR2 consistently outperform those without those bands (RGB and IR), with the F1-score being up to 1.250-2.024% higher when compared to RGB and 1.479-2.255% higher when compared to IR. The ResNet101-VEG model achieves the highest performance amongst the single input models, with an mAP of 0.5162 and the top F1 Score of 0.7242. Furthermore, the ViT-RGB+VEG model demonstrates superior performance with an F1 Score of 0.7484, however the Resnet101-RGB+VEG did not demonstrate this same pattern, with lower scores in all metrics, suggesting that the ViT architecture may be better at integrating the diverse information provided by multiple spectral bands compared to the ResNet-101 architecture. When comparing the ViT-RGB+VEG model with ViT-RGB, the former achieved higher scores in all but the recall metric (0.6790 vs 0.6846). The higher mAP and precision indicate that the ViT-RGB+VEG model is better at distinguishing between classes and making accurate classifications but some true positives might be missed.

TABLE III: Evaluation results of multi-label classification with baseline models using RGB bands only, and the proposed multistream models with RGB+VEG fusion.

Models	**Metrics**			
	mAP	**Precision**	**Recall**	F_1 **Score**
Resnet101-RGB	0.4791	0.7857	0.6375	0.7039
ViT-RGB	0.5465	0.8090	0.6846	0.7416
Resnet101-FC	0.5101	0.7816	0.6740	0.7238
Resnet101-SWIR	0.5096	0.7829	0.6686	0.7212
Resnet101-NWAR	0.5124	0.7780	0.6674	0.7185
Resnet101-IR	0.5023	0.7661	0.6472	0.7016
Resnet101-LW	0.5074	0.7806	0.6700	0.7210
Resnet101-AGRI	0.5019	0.7751	0.6660	0.7164
Resnet101-VEG	0.5162	0.7849	0.6721	0.7242
Resnet101-RGB+VEG	0.4667	0.7618	0.5885	0.6640
ViT-RGB+VEG	**0.5942**	**0.8335**	**0.6790**	**0.7484**

The evaluation of F1-scores across various land use classes, as shown in Table IV reveals that the ViT model with RGB+VEG features outperforms ResNet101 models for many classes, particularly in classes such as coastal lagoons and continuous urban fabric. High F1 scores are observed in classes with high intra-class similarity or high occurrence like 'sea and ocean' (16309 sample images) and 'coniferous forest' (42337 sample images), with the ViT model achieving an F1 score of 0.974 for sea and ocean and 0.869 for coniferous forest. Conversely, certain classes have lower samples or are relating to artificial object tend to have lower intra-class similarity, including airports (183 sample images) and port areas (108 sample images) exhibit notably low or zero scores.

Table V shows sample images and comparison of their respective ground truth labels with predictions from different models for three images. For the first image, ViT-RGB+VEG predicted all labels correctly, while ViT-RGB and Resnet101-RGB+VEG missing some labels. In the second image, ViT-RGB+VEG likely confused 'industrial or commercial unit' or 'airport' with 'sport and leisure facility'. In the third image, ViT-RGB+VEG missed the 'airport' label but correctly identified the other classes. The two images with 'airport' show very little similarity to each other and are not easily visible, particularly in the third image indicating a challenge for models to perform the classification, thus highlights that classes concerning artificial objects is still a challenge.

In overall, different band combinations are able to boost the performance of multi-label classification, specifically in improving the mAP, Recall, and F1-scores irregardless of the backbone feature extractor. Furthermore, the fusion of these band combinations also shows good potential at improving the performance across all metrics, hence, signifying that such multispectral band combinations and fusion have room for further investigation to solve a complex task as multi-label classification of satellite imagery.

V. Conclusion

This study has demonstrated the efficacy of utilising multispectral satellite imagery and advanced deep learning techniques for multi-label classification tasks. Our experiments highlight the importance of band combinations in creating effective false-color composites for satellite image analysis. We found that specific band combinations, such as those incorporating bands SWIR1 or SWIR2, consistently outperformed others in terms of F1 score and mAP metrics. We have also highlighted that the further fusion composite input could help improve the classification performance of different classes in multi-label classification.

We believe that future studies could explore advanced methodologies incorporating such band combination and fusion approach with mechanisms to mitigate class imbalance, such as leveraging techniques like class weighting by assign higher weights to minority classes in the loss function. Additionally, integrating higher-resolution additional spectral bands could provide richer data for more precise classification and analysis as well.

References

[1] F. A. Al-Wassai and N. V. Kalyankar, "Major limitations of satellite images," *CoRR*, vol. abs/1307.2434, 2013. arXiv: 1307.2434. [Online]. Available: http://arxiv.org/abs/1307.2434.

[2] S. Borra, D. Thanki, and N. Dey, *Satellite Image Analysis: Clustering and Classification*. Jan. 2019, ISBN: 9811364230. DOI: 10.1007/978-981-13-6424-2.

[3] Z. Yu, L. Di, R. Yang, *et al.*, "Selection of landsat 8 oli band combinations for land use and land cover classification," Jul. 2019, pp. 1–5. DOI: 10.1109/Agro-Geoinformatics.2019.8820595.

[4] L. P. Osco, J. Marcato Junior, A. P. Marques Ramos, *et al.*, "A review on deep learning in uav remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102 456, 2021, ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2021.102456. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030324342100163X.

[5] M. A. E. Bhuiyan, C. Witharana, A. K. Liljedahl, *et al.*, "Understanding the effects of optimal combination of spectral bands on deep learning model predictions: A case study based on permafrost tundra landform mapping using high resolution multispectral satellite imagery," *Journal of Imaging*, vol. 6, no. 9, 2020, ISSN: 2313-433X. DOI: 10.3390/jimaging6090097. [Online]. Available: https://www.mdpi.com/2313-433X/6/9/97.

[6] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904. DOI: 10.1109/IGARSS.2019.8900532.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[8] K. Zhu and J. Wu, "Residual attention: A simple but effective method for multi-label recognition," *CoRR*, vol. abs/2108.02456, 2021. arXiv: 2108.02456. [Online]. Available: https://arxiv.org/abs/2108.02456.

[9] P. Zhu, Y. Tan, L. Zhang, *et al.*, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4047–4060, 2020. DOI: 10.1109/TGRS.2019.2960466.

[10] F. Spoto, O. Sy, P. Laberinti, *et al.*, "Overview of sentinel-2," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 1707–1710. DOI: 10.1109/IGARSS.2012.6351195.

[11] M. Drusch, U. Del Bello, S. Carlier, *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012, The Sentinel Missions - New Opportunities for Science, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2011.11.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425712000636.

[12] J. Tamouk, N. Lotfi, and M. Farmanbar, *Satellite image classification methods and landsat 5tm bands*, 2013. arXiv: 1308.1801 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1308.1801.

[13] E. D. Analytics, *Satellite band combinations: Analytical methods for imagery*, eos.com, 2021. [Online]. Available: https://eos.com/make-an-analysis/.

TABLE IV: F1-Scores for land use classes based with respect to the band combinations. [Bold indicates the best band for the class.]

Class	**ResNet101**									**ViT**	
	AGRI	**FC**	**LW**	**NWAR**	**SWIR**	**VEG**	**IR**	**RGB**	**RGB+VEG**	**RGB**	**RGB+VEG**
Mixed forest	0.806	0.803	0.796	0.804	0.801	0.803	0.790	0.766	0.781	0.810	**0.821**
Non-irrigated arable land	0.793	0.804	0.816	0.796	0.811	0.817	0.784	0.777	0.729	**0.827**	0.826
Broad-leaved forest	0.715	0.720	0.718	0.711	0.715	0.725	0.678	0.679	0.667	0.724	**0.740**
Complex cultivation patterns	0.587	0.622	0.616	0.607	0.622	0.611	0.558	0.615	0.519	0.662	**0.655**
Water bodies	0.818	0.829	0.820	0.807	0.825	0.826	0.803	0.788	0.789	0.834	**0.846**
Discontinuous urban fabric	0.665	0.704	0.690	0.702	0.702	0.700	0.702	0.705	0.688	0.748	**0.755**
Peatbogs	0.551	0.572	**0.608**	0.538	0.567	0.572	0.578	0.548	0.497	0.594	0.587
Industrial or commercial units	0.438	0.482	0.439	0.480	0.462	0.466	0.505	0.504	0.443	0.506	**0.523**
Olive groves	**0.393**	0.265	0.183	0.252	0.294	0.282	0.244	0.319	0.026	0.296	0.352
Continuous urban fabric	0.327	0.462	0.493	0.359	0.395	0.404	0.350	0.494	0.116	0.617	**0.669**
Vineyards	0.144	0.172	0.267	0.242	0.193	0.185	0.247	**0.339**	0.073	0.253	0.283
Inland marshes	0.170	0.152	**0.185**	0.158	0.161	0.179	0.142	0.063	0.087	0.056	0.066
Sport and leisure facilities	0.122	0.098	0.202	0.074	0.086	**0.167**	0.133	0.025	0.158	0.087	0.134
Mineral extraction sites	0.305	0.323	**0.357**	0.295	0.333	0.299	0.145	0.232	0.272	0.240	0.268
Road and rail networks	0.151	**0.256**	0.233	0.234	0.215	0.234	0.105	0.073	0.172	0.108	0.172
Green urban areas	0.031	0.116	0.031	**0.169**	0.087	0.032	0.085	0.062	0.000	0.000	0.149
Sparsely vegetated areas	0.078	0.172	0.167	0.083	0.151	0.157	0.080	0.172	0.043	0.218	**0.255**
Coastal lagoons	0.549	0.343	0.371	0.568	0.435	0.519	0.480	0.611	0.600	**0.750**	0.711
Estuaries	0.280	0.213	0.226	0.292	0.182	0.200	0.379	**0.438**	0.133	0.353	0.429
Airports	0.000	0.000	**0.059**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Port areas	0.000	0.000	0.000	0.000	**0.118**	0.000	0.000	0.000	0.000	0.000	0.000
Burnt areas	0.000	**0.182**	0.000	**0.182**	**0.182**	0.000	0.000	0.000	0.000	0.000	0.000
Coniferous forest	0.852	0.856	0.852	0.860	0.855	0.859	0.839	0.842	0.806	0.857	**0.869**
Transitional woodland	0.616	0.620	0.605	0.598	0.604	0.610	0.623	0.605	0.568	**0.625**	0.614
LPA-ANV[1]	0.565	0.573	0.569	0.570	0.584	0.558	0.484	0.544	0.350	**0.625**	0.613
Pastures	0.713	0.722	**0.729**	0.709	0.714	0.718	0.704	0.708	0.670	0.721	0.723
Sea and ocean	0.938	0.930	0.928	0.944	0.924	0.925	0.937	0.946	0.928	0.970	**0.974**
Agro-forestry areas	0.739	0.750	0.749	0.738	0.740	0.747	0.751	0.732	0.736	0.771	**0.791**
Permanently irrigated land	0.504	0.480	0.547	0.504	0.474	0.541	0.535	0.478	0.309	0.545	**0.565**
Natural grassland	0.424	0.456	0.431	0.372	0.419	0.438	0.390	**0.467**	0.327	0.432	0.447
Sclerophyllous vegetation	0.441	0.426	0.380	0.390	0.458	0.458	0.341	0.380	0.070	0.540	**0.576**
Water courses	0.698	0.714	0.680	**0.721**	0.696	0.693	0.662	0.584	0.668	0.602	**0.721**
Annual permanent crops	0.405	0.368	0.375	0.386	0.401	0.380	0.348	**0.428**	0.000	0.420	0.399
Moors and heathlands	**0.341**	0.273	0.278	0.310	0.304	0.276	0.271	0.316	0.141	0.306	0.287
Fruit trees and berry plantations	0.126	0.048	0.025	0.134	0.036	0.061	0.085	**0.188**	0.000	0.161	0.153
Rice fields	0.262	0.389	0.365	0.372	0.414	0.382	**0.422**	0.413	0.264	0.327	0.369
Bare rock	0.417	**0.578**	0.421	0.423	0.522	0.489	0.444	0.535	0.385	0.388	0.521
Beaches, dunes, sands	0.611	**0.642**	0.582	0.617	0.585	0.634	0.629	0.641	0.620	0.629	0.640
Salt marshes	0.274	0.194	0.242	0.310	0.187	0.232	0.257	0.282	0.000	0.190	**0.358**
Construction sites	0.000	0.000	0.000	0.000	0.000	0.000	**0.069**	0.000	0.000	0.000	0.000
Intertidal flats	0.154	0.217	0.318	0.195	0.208	0.238	0.150	0.267	0.273	0.341	**0.381**
Dump sites	0.061	**0.063**	0.000	0.057	0.059	0.000	0.000	0.000	0.000	0.000	**0.063**
Salines	0.435	0.560	0.667	0.500	0.615	0.615	0.667	0.417	0.105	0.286	**0.714**

[1] Land principally occupied by agriculture, with significant areas of natural vegetation.

TABLE V: Sample predictions. [Blue: correct predictions; Red: Incorrect predictions.]

Image	Groundtruth	ViT-RGB	ViT-RGB+VEG	Resnet101-RGB+VEG
	Coniferous forest, Mixed forest, Sea and ocean	Mixed forest	Coniferous forest, Mixed forest, Sea and ocean	Coniferous forest, Mixed forest
	Industrial or commercial units, Airports, Non-irrigated arable land, Pastures	Pastures	Non-irrigated arable land, Sport and leisure facilities, Pastures	Non-irrigated arable land, Pastures
	Airports, Coniferous forest, Mixed forest, Transitional woodland/shrub	Coniferous forest, Land principally occupied by agriculture, with significant areas of natural vegetation	Coniferous forest, Mixed forest, Transitional woodland/shrub	Mixed forest, broad-leaved forest, Transitional woodland/shrub