EFFECTS OF CLASS IMBALANCE IN MULTI-LABEL CLASSIFICATION SPLITS: A STUDY CASE ON BIGEARTHNET

Matthieu Verlynde, Ammar Mian, Yajing Yan

Université Savoie Mont Blanc, LISTIC, Annecy France

ABSTRACT

Class imbalance is a common issue for multi-label classification tasks, and land-use land-cover classification tasks in remote sensing are, inherently, often multi-label problems. Training deep learning models on benchmark datasets that address this issue is then a key component to ensure high performances. This paper proposes the application of an iterative stratified sampling method to rebalance a benchmark dataset. The BigEarthNet dataset [1] is used as a study case to show the improved training performances of convolutional neural networks trained on the resampled data compared to those trained on the original randomly split data.

Index Terms— Multi-label classification, Iterative stratified sampling, Deep learning, Land-use land-cover

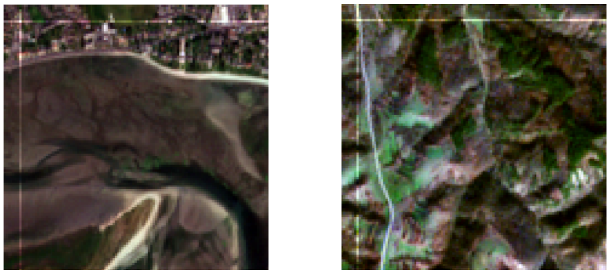
1. INTRODUCTION

Among the common machine learning tasks in remote sensing, classification is one of the most studied topics in the literature. For supervised classification tasks, assigning land use and land cover (LULC) classes to images plays a pivotal role across a broad spectrum of applications from wildfire [2] or crop yield predictions [3] to environmental studies [4, 5]. Changes in land cover are also a key component in the assessment of climate change [6]. However, due to the nature of the data, multiple land covers are often present within the same image making this task a multi-label classification problem.

In supervised classification, deep learning tools, particularly convolutional neural networks (CNN), have gained prominence for their strong performance on such task [7]. Their performance is directly linked to the training data used. Therefore, the demand for high-quality benchmark datasets for transfer learning approaches is rapidly increasing to effectively calibrate such models. In this context, several datasets have been published such as Sen-2 LULC [8], EuroSAT [9], SEN12MS [10] and BigEarthNet [1].

One of the key challenges in multi-label classification is class imbalance, which refers to significantly different proportions of each label within the dataset. This problem is extensively studied in the literature [11, 12], as class imbalance can significantly affect model performance. Deep learning models can focus their learning task on the major labels and discard the least represented labels, resulting in overfitting. Furthermore, class imbalance between the training and validation splits can result in poor performance evaluation during the training phase. To address these issues, common practices include using undersampling [13] or oversampling [14] to balance the different spits. Using a composite approach to take into account class imbalance also emerges as a solution. Liu X. et al. [13] and Tahir M.A. et al. [15] both carry out multiple undersampling and combine the results of several classifiers trained on their splits. While these different approaches may be effective for multi-label tasks [13], they can be challenging to operate for users who want to train simple deep learning models.

In this paper, we investigate the influence of class imbalance in multi-label classification performance, in particular on the BigEarthNet dataset [1], a benchmark dataset of multilabelled images with LULC classes. After identifying the issue within the dataset, we propose a solution to rebalance the training splits and evaluate the performance of both a simple model and a computationally intensive model on the original and rebalanced datasets.



(a) Discontinuous urban fabric ; Beaches, dunes, sands ; Salt marches ; Intertidal flats ; Estuaries

(b) Pastures ; Broad-leaved forest ; Mixed forest ; Natural grassland ; Transitional woodland/shrub

Fig. 1: Examples of multi-labelled RGB images from the BigEarthNet dataset [16].
The image IDs are S2A_MSIL2A_20170717T113321_76_67 (a) and S2A_MSIL2A_20171208T093351_56_86 (b)

This work has been done thanks to the facilities offered by the Univ. Savoie Mont Blanc - CNRS/IN2P3 MUST computing center

2. THE BIGEARTHNET DATASET

The BigEarthNet dataset [1] contains 590,326 pairs of Sentinel-1 and Sentinel-2 images that correspond to 125 Sentinel-2 level-1C tiles. The Sentinel-2 images are present within 3 formats (120 × 120 pixels with a 10 m resolution, 60 × 60 pixels with a 20 m resolution and 20 × 20 pixels with a 60 m resolution) for the 12 bands from the multi-spectral sensor in optical imagery. These images are acquired from June 2017 to May 2018 within 10 European countries. Every patch—group of images at the same location—is multi-labelled according to 43 LULC classes of the Corine Land Cover 2018 dataset. Two examples of Sentinel-2 red, green and blue bands with their corresponding labels are presented in Fig. 1.

As a multi-labelled classification dataset, the BigEarthNet dataset poses a challenge for users in balancing training and test samples for classification tasks. Class imbalance is a significant issue in the BigEarthNet dataset [16] as the distribution of LULC labels is unevenly distributed across patches. For example, 217,119 Sentinel-2 patches are labelled with the Mixed Forest label, whereas only 328 are labelled with Burnt areas. The rarest labels within the dataset are then likely to be under-represented within splits with random sampling. To address this issue, a specific sampling strategy is necessary to split the data into balanced training samples.

3. ITERATIVE STRATIFIED SAMPLING

The main issue with multi-label data sampling is handling rare labels. Splitting data without focusing on these labels could lead to an uneven distribution of these labels between splits. Balancing the label distributions between the training, validation and test samples, using an iterative stratification sampling strategy [17] allows the data to be divided into several folds. These folds are used in [17] to carry out a k-fold cross-validation process. In this paper, the k folds with similar label distributions are combined to create three folds without replacement forming new training, validation, and test samples.

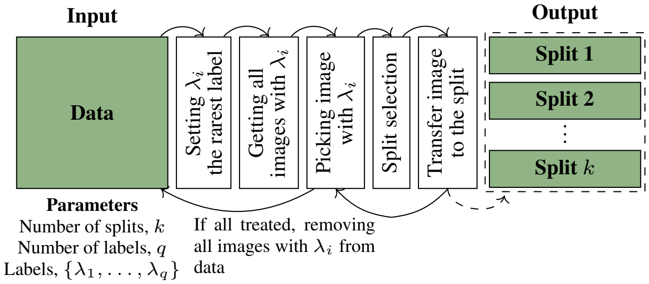

Input
Output
Data
Setting λ_i the rarest label
Getting all images with λ_i
Picking image with λ_i
Split selection
Transfer image to the split
Split 1
Split 2
⋮
Split k
Parameters
Number of splits, k
Number of labels, q
Labels, $\{\lambda_1, \dots, \lambda_q\}$
If all treated, removing all images with λ_i from data

Fig. 2: Iterative stratified sampling algorithm [17] workflow

Iterative stratified sampling consists of adding images to the different splits by using the different label frequencies. At the initialization step, the number of images c_j for each desired split $j \in [\![1, k]\!]$ is calculated, as well as the number of desired images $c_j^{\lambda_i}$ for each label $\lambda_1, \dots, \lambda_q$. Then, the algorithm iteratively adds images to each split until the data set is empty. For each iteration, the rarest label λ_i is selected within the remaining data set. Then, iteratively, each image with this label is added to the split j with the highest number of images $c_j^{\lambda_i}$ needed to be added to this split for this label. If two splits require the same number of images with this label, the split requiring the highest total number of images, c_j, is selected. If, again, multiple splits require the same number of images, j is randomly selected. After each sub-iteration, c_j and $c_j^{\lambda_i}$ are updated.

Applying this strategy to the BigEarthNet dataset produces training splits with similar label distributions. The results of this sampling is presented in the next section.

4. EXPERIMENTS AND RESULTS

Experiment setup: The training was done using Python 3.8.19, PyTorch 2.0.1 and Tensorflow 2.13.1. The experiments were carried out on an NVIDIA RTX 4500 Ada 24GB and three NVIDIA A100 80 GB GPUs. The package iterative-stratification[1] was used to carry out the iterative stratified sampling on the BigEarthNet datasets. All the codes are available on GitHub[2].

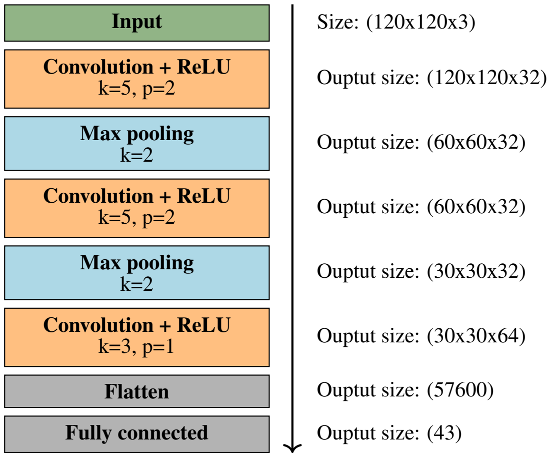


Fig. 3: Structure of the ShortCNN model [16]
The parameters k, s and p are respectively the kernel size, the stride and the padding.

Methodology: To compare performances between training on original and resampled splits, two CNNs for multi-label classifications were used. The first model was a short CNN, described in [16] and shown in Fig. 3. This model is referred to as ShortCNN in the rest of this paper and has been selected

[1] `https://pypi.org/project/iterative-stratification/`

[2] `https://github.com/MattVerlynde/balanced-bigearthnet`

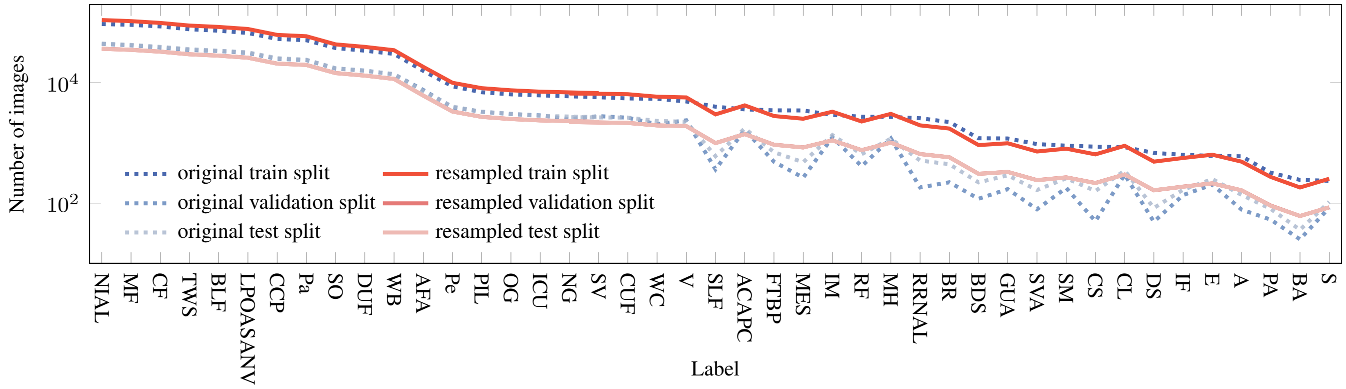
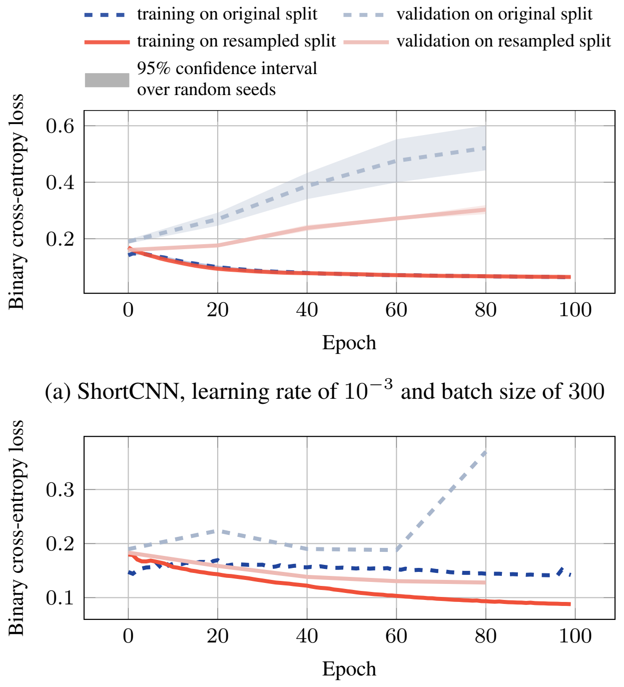


Fig. 4: Unbalanced label distribution within splits in BigEarthNet (original splits and resampled splits). The lines for the resampled validation split and the resampled test split are merged on this graph. The mapping of classes labels and acronyms is available on GitHub.



(a) ShortCNN, learning rate of 10^{-3} and batch size of 300

(b) InceptionV3, learning rates of 10^{-2} and batch size of 200

Fig. 5: Trainings on original and resampled splits of BigEarthNet

to study the performance of a simple model on the classification task, one that can easily be reproduced by other users. The InceptionV3 model [18] used in the latest version of BigEarthNet, reBEN [19], and pre-trained on the ImageNet (ILSVRC) 2012 dataset [20] was also utilized to demonstrate the application of a large model for such classification tasks. Trainings and evaluations were repeated 10 times for the ShortCNN model for each set of parameters with different random seeds for statistical significance. Training on the InceptionV3 model was performed only once due to its long training time.

For both models, the stochastic gradient descent and binary cross-entropy with logits function were used as optimizer and loss function, and the red, green and blue bands of BigEarthNet optical images were taken as input. The evolution of the loss function value during training was used to select the best parameter. Then the evaluation is performed on the test data.

Resampling: The label distributions between the training, validation and test samples appear to be heterogenous. As seen in Fig. 4, the most represented labels within the initial BigEarthNet dataset are split according to the splitting ratios of the three samples (60%, 20% and 20% for the training, validation and test samples respectively). This is not the case for the least represented labels, such as BA (Burnt areas) for which the number of images between the validation and test samples are not equivalent. This visualization shows the random splitting method used by [1] favours the most represented labels for multi-labelled data. Applying the iterative stratified sampling strategy helps mitigating this imbalance, as the new samples exhibit a similar distribution for each label according to the intended splitting ratios.

Results on BigEarthNet: During the training of both the ShortCNN and the InceptionV3 models, the evolution of the loss value for the validation split shows overfitting when training is carried out on the original samples of [16], as shown in Fig. 5. Learning rates showing the best evolution profiles of the loss function during training were selected (10^{-3} for ShortCNN and 10^{-2} for InceptionV3). overfitting during the training of the ShortCNN model could be explained by the relatively modest size of the convolutional model containing 2,523,403 trainable parameters. In relation with the number of images within the original and resampled datasets and the

Model	Samples	Macro average				Micro average			
		Precision	**Recall**	**F1-score**	**F2-score**	**Precision**	**Recall**	**F1-score**	**F2-score**
ShortCNN	original	0.148	0.133	0.130	0.130	0.378	0.378	0.378	0.378
	resampled	0.377	**0.251**	**0.287**	**0.263**	0.567	0.503	0.533	0.514
InceptionV3	original	0.021	0.004	0.005	0.004	0.422	0.016	0.031	0.020
	resampled	**0.405**	0.182	0.207	0.189	**0.712**	**0.506**	**0.591**	**0.537**

Table 1: Performances on test samples after training on BigEarthNet data. The results were averaged over 10 seeds for the ShortCNN model.

complexity of the task, a multi-label classification with 43 labels likely induces a lack of capacity to learn complex aspects of both datasets. For the InceptionV3 model, overfitting occurs after a significantly higher number of epochs—60 over the first epoch of ShortCNN. InceptionV3 contains 25,200,371 trainable parameters and this higher size likely permits a better learning of the label characteristics. For both models, the iterative stratified resampling approach tends to reduce overfitting. Fig. 5 shows that overfitting occurs after 20 epochs for the ShortCNN model, and this is not observed within the 100 epochs of training of the InceptionV3 model. Furthermore, the loss value is significantly lower during training of both models on the resampled validation dataset, indicating a better training process for the multi-label classification.

The performances of both models with the test samples after training are shown in Table 1. Being a multi-label classification problem, micro- and macro-average approaches are considered to take into account the imbalance in label distribution within the test samples. The differences in their calculations for a performance score s are presented in Fig. 6 with TP_{λ_i}, FP_{λ_i}, TN_{λ_i} and FN_{λ_i} being respectively the counts of true positives, false positives, true negatives, and false negatives in the binary classification for the label λ_i.

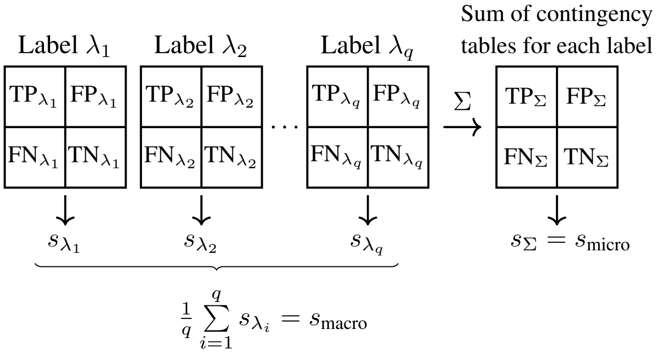

Label λ_1
Label λ_2
Label λ_q
Sum of contingency tables for each label
TP_{λ_1}
FP_{λ_1}
FN_{λ_1}
TN_{λ_1}
TP_{λ_2}
FP_{λ_2}
FN_{λ_2}
TN_{λ_2}
$\cdots$
TP_{λ_q}
FP_{λ_q}
FN_{λ_q}
TN_{λ_q}
Σ
TP_{Σ}
FP_{Σ}
FN_{Σ}
TN_{Σ}
s_{λ_1}
s_{λ_2}
s_{λ_q}
$s_{\Sigma} = s_{\text{micro}}$
$\frac{1}{q}\sum_{i=1}^{q} s_{\lambda_i} = s_{\text{macro}}$

Fig. 6: Macro- and micro- average appraoches for classification performance metrics.

Macro-averaging consists in computing the average performance of each label treated as a separate binary classification task. In contrast, micro-averaging calculates each performance index based on a global binary contingency table for all labels. Therefore, a macro-average approach gives equal weight to each label, while micro-averaging takes into account class imbalance [21].

The overall performance of both models with the resampled data is higher than with the original data for the BigEarthNet v1.0 dataset. For the original samples, the ShortCNN model shows better test performances than the InceptionV3 model except on the micro-averaged precision. The overfitting during training then affects the test performances of the InceptionV3 model more than the smaller convolutional model. Nevertheless, for the resampled data, the ShortCNN model shows higher macro-averaged recall, F1 and F2-scores than InceptionV3. This is likely due to the influence of the most represented labels within the test samples. The smallest model tends to classify more accurately the most represented labels within the dataset at the cost of the rarest labels, while the InceptionV3 model shows better performances for rare labels.

5. CONCLUSION

In this paper we showed that the imbalanced distribution of labels within the reference BigEarthNet datasets is a source of overfitting during training of the InceptionV3 and ShortCNN models. The solution proposed is an iterative stratified sampling strategy to rebalance training, validation and test splits before training. This method proved to improve the performance of both the ShortCNN and the InceptionV3 models and the new splits appear to be easier to use than the initially randomly sampled splits provided. In further work, we will experiment with learning rate schedulers and classical model-based methods for reducing overfitting, such as drop-out layers and early stopping, and apply this approach to reBEN [19], the latest version of BigEarthNet.

6. REFERENCES

[1] Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begum Demir, and Volker Markl, "Bigearthnet-

mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, Sept. 2021.

[2] Anna Badia, Montserrat Pallares-Barbera, Natàlia Valldeperas, and Meritxell Gisbert, "Wildfires in the wildland-urban interface in catalonia: Vulnerability analysis based on land use and land cover change," *Science of The Total Environment*, vol. 673, pp. 184–196, 2019.

[3] Christopher A Seifert, George Azzari, and David B Lobell, "Satellite detection of cover crops and their effects on crop yield in the midwestern united states," *Environmental Research Letters*, vol. 13, no. 6, pp. 064033, jun 2018.

[4] S. Wang, B. J. Fu, G. Y. Gao, X. L. Yao, and J. Zhou, "Soil moisture and evapotranspiration of different land cover types in the loess plateau, china," *Hydrology and Earth System Sciences*, vol. 16, no. 8, pp. 2883–2892, 2012.

[5] Terefe Tolessa, Feyera Senbeta, and Moges Kidane, "The impact of land use/land cover change on ecosystem services in the central highlands of ethiopia," *Ecosystem Services*, vol. 23, pp. 47–54, 2017.

[6] Johannes J Feddema, Keith W Oleson, Gordon B Bonan, Linda O Mearns, Lawrence E Buja, Gerald A Meehl, and Warren M Washington, "The importance of land-cover change in simulating future climates," *Science*, vol. 310, no. 5754, pp. 1674–1678, 2005.

[7] Junye Wang, Michael Bretz, M. Ali Akber Dewan, and Mojtaba Aghajani Delavar, "Machine learning in modelling land-use and land cover-change (lulcc): Current status, challenges and prospects," *Science of The Total Environment*, vol. 822, pp. 153559, 2022.

[8] Suraj Sawant, Rahul Dev Garg, Vishal Meshram, and Shrayank Mistry, "Sen-2 LULC: Land use land cover dataset for deep learning approaches," *Data in Brief*, vol. 51, pp. 109724, Dec. 2023.

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[10] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "Sen12ms – a curated dataset of georeferenced multispectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, pp. 153–160, 2019.

[11] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, June 2004.

[12] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[13] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[14] José A. Sáez, Bartosz Krawczyk, and Michał Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognition*, vol. 57, pp. 164–178, 2016.

[15] Muhammad Atif Tahir, Josef Kittler, and Fei Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3738–3750, 2012.

[16] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker 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.

[17] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, Eds., Berlin, Heidelberg, 2011, pp. 145–158, Springer Berlin Heidelberg.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[19] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl, "reben: Refined bigearthnet dataset for remote sensing image analysis," July 2024, arXiv, arXiv:2407.03653.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[21] Yiming Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, vol. 1, no. 1, pp. 69–90, Apr. 1999.