
Design Choice Sensitivity in Remote Sensing Multi-Label Classification

Maryam Gholami Shiri^{1,2} Eva Tuba^{3,4,5} Sašo Džeroski¹ Tome Eftimov^{2,3} Ana Nikolij^{2,3}

¹Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

²Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

³Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia

⁴Trinity University, San Antonio, TX, USA

⁵Singidunum University, Belgrade, Serbia

Abstract Benchmarking deep learning (DL) models for multi-label classification (MLC) of remote sensing images (RSI) typically yields rankings that do not generalize beyond the evaluated datasets. We move beyond rankings by using functional ANOVA (fANOVA) to quantify how much individual design choices and their interactions contribute to performance variability. Across 48 DL models and seven MLC RSI datasets spanning architecture, fine-tuning, and learning-strategy choices, we build dataset meta-representations of design-choice sensitivity and cluster them. Datasets group not by performance level but by *which* choices matter: fine-tuning dominates for large-scale datasets, architecture becomes decisive in data-limited regimes, and their interaction governs intermediate regimes. These regimes align with intrinsic dataset properties (scale, resolution, label complexity), showing that benchmarking conclusions are dataset-conditional rather than universal.

1 Introduction

Remote sensing image (RSI) analysis supports monitoring of phenomena in agriculture (Chlingaryan et al., 2018) and urban planning (Huang et al., 2018). Since a single image typically carries multiple semantic labels, multi-label classification (MLC) (Bogatinovski et al., 2022a) is a core task, and deep learning (DL) is now its dominant paradigm (Chen et al., 2021). Model performance depends on architecture, fine-tuning strategy (Yosinski et al., 2014), and learning strategy, yet benchmarking studies typically report only rankings, which do not reveal *why* one configuration wins or whether it will generalize to a new dataset (Dimitrovski et al., 2023). Dataset meta-features such as scale, resolution, and label complexity are known to affect method performance (Bogatinovski et al., 2022b), but their link to *design-choice sensitivity* specifically is rarely made explicit.

Contributions: (i) we apply fANOVA to an MLC benchmarking scenario of 48 DL models across seven RSI datasets, quantifying design-choice and interaction contributions to performance variability; (ii) we build dataset meta-representations from fANOVA scores and show, via hierarchical clustering, that datasets group by design-choice sensitivity rather than by performance; (iii) we show the dominant factor shifts with dataset regime; and (iv) we reframe conclusions from an established benchmarking study as dataset-conditional. Code and data: <https://zenodo.org/records/19630021>.

2 Related Work

MLC learns $f : X \rightarrow 2^C$ over a label set $C = \{c_1, \dots, c_L\}$, where labels may co-occur (Tsoumakas et al., 2010; Zhang and Zhou, 2014). fANOVA (Sobol', 1993) decomposes a model's output variance into contributions from individual features and their interactions via marginal effects, typically estimated with a random forest surrogate (Hutter et al., 2014). Originally used to identify important hyperparameters in algorithm configuration (Hutter et al., 2014), it was later shown that

Table 1: Design choices describing the DL models.

Module	Options
Architecture	VGG-16, VGG-19 (Simonyan and Zisserman, 2015), ResNet-34/50/152 (He et al., 2016), EfficientNet-B0/B1/B2 (Tan and Le, 2019)
Fine-tuning	pre-trained (last layer only), fine-tuned (all layers)
Learning strategy	end-to-end, Random Forest (Breiman, 2001), Extra Trees (Geurts et al., 2006)

hyperparameter importance varies across datasets (van Rijn and Hutter, 2018), and the method has since been integrated into tools such as Optuna (Akiba et al., 2019) and applied to architecture and black-box optimization analysis (Klein et al., 2018; Watanabe et al., 2023). On the meta-feature side, Bogatinovski et al. (2022b) showed that label-space meta-features explain MLC performance differences well, motivating our use of dataset meta-features to interpret design-choice sensitivity clusters.

3 Methodology

We apply the extended fANOVA framework of Nikolikj and Eftimov (2025). For a set of datasets $\{D_k\}_{k=1}^K$ and models $\{M_i\}_{i=1}^m$, each M_i is described by a design-choice vector $\mathbf{x}_i$; evaluating it on D_k gives $D_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. fANOVA decomposes performance variance on each D_k into main and interaction effects, aggregated into a meta-representation ϕ_k . We then cluster $\{\phi_k\}_{k=1}^K$ to find datasets with similar *sensitivity* profiles (not similar performance) and relate the clusters to dataset meta-features. We choose fANOVA over local attribution methods (e.g., SHAP (Lundberg and Lee, 2017)) because it gives a globally consistent decomposition over the whole configuration space, explicitly captures higher-order interactions, and its random-forest surrogate is tractable at $m \leq 48$ (Hutter et al., 2014).

4 Experimental Design

We study the benchmarking setting of Stoimchev et al. (2023): end-to-end CNN vs. CNN feature extraction with tree ensembles. Models vary along three modules: architecture, fine-tuning strategy, and learning strategy (Table 1), yielding 48 configurations evaluated on seven RSI datasets via ranking loss (fraction of incorrectly ordered label pairs, $[0, 1]$; lower is better). BigEarthNet-19/43 and MLRSNet are large-scale (100k–590k samples); UCM, AID, and DFC-15 are medium-scale, high-resolution; Ankara is data-limited (216 samples). With three modules, the meta-representation has dimensionality $3 + \binom{3}{2} + \binom{3}{3} = 7$ (main effects, pairwise, and three-way interaction). Datasets are clustered with hierarchical clustering (cosine distance, average linkage) (Müllner, 2011), evaluated via silhouette coefficient and cophenetic correlation.

5 Results

Raw performance variability. Figure 1 shows kernel density estimates of ranking loss per dataset across all 48 configurations. Ankara and AID show tight, low-loss distributions, indicating limited sensitivity to design choices overall, whereas UCM, DFC-15, MLRSNet, and both BigEarthNet variants spread across a much wider range of the $[0, 1]$ interval. A first hint that different datasets react very differently to the same set of design choices, motivating a finer-grained, per-choice analysis.

Individual effects dominate. The fANOVA surrogate fits well (median $R^2 = 0.966$). As Figure 2 shows, individual design choices explain most of the variance (a single choice explains over 80%

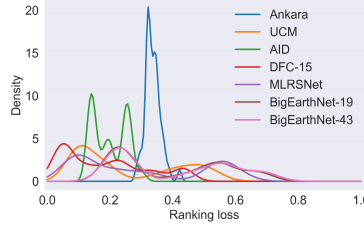


Figure 1: Kernel density estimates of ranking loss (lower is better) across all 48 model configurations, per dataset.

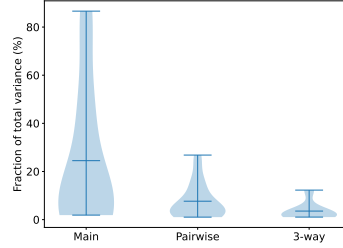


Figure 2: Variance explained by main effects, pairwise, and three-way interactions across the seven datasets.

in some datasets), while pairwise interactions contribute less and three-way interactions are near zero, modules largely act independently, so we can meaningfully speak of “the” dominant factor per dataset.

Datasets cluster by sensitivity, not performance. Hierarchical clustering of the fANOVA meta-representations (cosine distance, average linkage; silhouette = 0.7571, cophenetic correlation = 0.9674) yields three clusters, visualized in Figure 3.

Cluster 1 – large-scale (BigEarthNet-19/43, MLRSNet). Fine-tuning dominates: adapting pre-trained weights to the target data is what matters most for good performance on large-scale datasets. Figure 4 illustrates this for BigEarthNet-43, fine-tuned models consistently achieve much lower ranking loss than pre-trained-only ones (panel a), and the architecture×fine-tuning heatmap (panel b) is nearly uniform across architectures once conditioned on fine-tuning status, showing that the benefit of fine-tuning holds regardless of which backbone is used. The learning-strategy module behaves similarly on this cluster.

Cluster 2 – medium-scale (UCM, AID). Fine-tuning’s importance drops relative to Cluster 1, while architecture and learning strategy become jointly more influential. Performance here is better explained by the *combination* of choices than by any single dominant one. Fine-tuned models still outperform pre-trained-only ones, but once fine-tuning is applied, the learning-strategy choice has little further effect; without fine-tuning, however, end-to-end training clearly beats using the network as a frozen feature extractor for Random Forest or Extra Trees.

Cluster 3 – data-limited / high-resolution (Ankara, DFC-15). Architecture becomes the decisive factor, and its interaction with fine-tuning strengthens. Figure 5 shows this for Ankara: simpler, shallower architectures such as VGG-16 outperform much deeper ResNet variants (panel a). With only 216 training samples, model capacity that exceeds the dataset’s scale appears to hurt rather than help. The architecture×learning-strategy heatmap (panel b) further shows VGG-16 combined with an end-to-end learning strategy achieves the lowest ranking loss, suggesting joint, task-specific representation learning compensates for a simpler backbone. For DFC-15, the pattern flips: its high spatial resolution rewards deeper, modern architectures (EfficientNet, ResNet) over VGG, and fine-tuning helps some architectures far more than others.

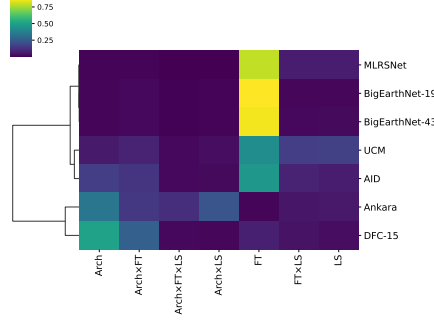


Figure 3: Clustermap of dataset meta-representations: rows are datasets, columns are module effects/interactions, color = variance share (yellow = higher).

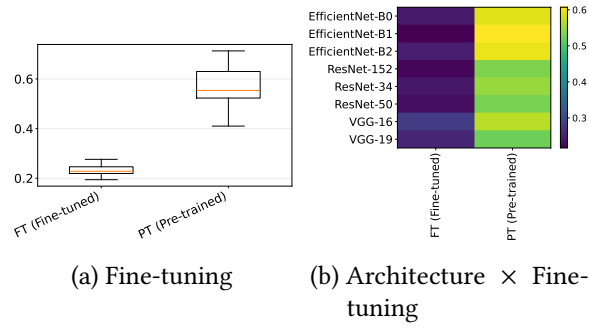


Figure 4: Marginal effects of modules for *BigEarthNet-43* (Cluster 1) .

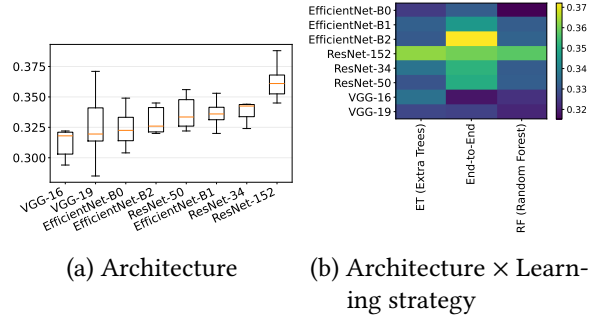


Figure 5: Marginal effects of modules for *Ankara* (Cluster 3).

6 Conclusion

Using fANOVA across 48 models and seven RSI datasets, we showed that design-choice sensitivity, not raw performance, is what naturally separates datasets into large-scale, medium-scale, and data-limited regimes, each governed by a different dominant factor (fine-tuning, combined choices, and architecture, respectively). This dataset-aware view offers more actionable guidance than a single global ranking.

Limitations. Our analysis covers only CNN architectures across three modules, reuses fixed training pipelines from prior work, and clusters just seven datasets. More datasets and modules (e.g., foundation models) would strengthen the statistical robustness of the sensitivity–meta-feature relationship.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.
- Bogatinovski, J., Todorovski, L., Džeroski, S., and Kocev, D. (2022a). Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.
- Bogatinovski, J., Todorovski, L., Džeroski, S., and Kocev, D. (2022b). Explaining the performance of multilabel classification methods with data set properties. *International Journal of Intelligent Systems*, 37(9):6080–6122.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Chen, H., Qi, Z., and Shi, Z. (2021). Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14.
- Chlingaryan, A., Sukkariéh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69.
- Dimitrovski, I., Kitanovski, I., Kocev, D., and Simidjievski, N. (2023). Current trends in deep learning for earth observation: An open-source benchmark arena for image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:18–35.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Huang, X. et al. (2018). Urban land use mapping using remote sensing images and deep learning. *Remote Sensing of Environment*, 204:719–731.
- Hutter, F., Hoos, H., and Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In *International Conference on Machine Learning*, pages 754–762. PMLR.
- Klein, A., Christiansen, E., Murphy, K., and Hutter, F. (2018). Towards reproducible neural architecture and hyperparameter search. In *ICML Workshop on Reproducibility in Machine Learning*.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.
- Nikolikj, A. and Eftimov, T. (2025). Exploring module interactions in modular CMA-ES across problem classes. *Swarm and Evolutionary Computation*, 98:102116.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.
- Sobol’, I. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment*, 1:407.

- Stoimchev, M., Kocev, D., and Džeroski, S. (2023). Deep network architectures as feature extractors for multi-label classification of remote sensing images. *Remote Sensing*, 15(2):538.
- Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114.
- Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.
- van Rijn, J. N. and Hutter, F. (2018). Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2367–2376.
- Watanabe, S., Bansal, A., and Hutter, F. (2023). PED-ANOVA: Efficiently quantifying hyperparameter importance in arbitrary subspaces. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4389–4395.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27.
- Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.