Spectral-Temporal Feature Analysis for Vegetation Classification using Sentinel-2: A Comparison of XGBoost, Random Forest, and Neural Network

Konstantinos Christofi[a,b], Charalambos Chrysostomou[a], Charalampos Kontoes[c], Diofantos G. Hadjimitsis[a,b], and Michalis Mavrovouniotis[a]

[a]ERATOSTHENES Centre of Excellence, 82 Franklin Roosevelt, 3012, Limassol, Cyprus
[b]Dept. of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol, Cyprus
[c]Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece

ABSTRACT

Vegetation classification plays a critical role in land use management, ecological monitoring, and biodiversity conservation. This study investigates the use of Sentinel-2 multi-spectral time series data and feature importance analysis for grassland type classification, focusing on meadows and orchards. We compare three supervised learning algorithms—Random Forest, XGBoost, and neural network—across four classification tasks: intra-class (standard vs. wet meadows; traditional vs. intensive orchards), binary inter-class (meadows vs. orchards), and multiclass classification. All three models are trained on 156-dimensional pixel-level spectral time series from 2017-2018 and evaluated using 10-fold cross-validation. Our results show that XGBoost achieves the highest performance in binary and multiclass settings, while Random Forest shows high effectiveness for intra-class classification. Additionally, band-wise importance analysis reveals that the NIR band contributes most in intra-class tasks, whereas all bands are more evenly relevant in inter-class tasks. These findings support interpretable, fine-grained vegetation mapping using time series remote sensing data.

Keywords: Meadows vs Orchards classification, Sentinel-2, Feature Importance Analysis, Neural Network, XGBoost, Random Forest, Remote sensing

1. INTRODUCTION

Vegetation classification is an essential task in Remote Sensing applications, covering a range of uses from land use management, biodiversity monitoring, agricultural planning and ecological restoration activities.[1,2] In particular, the capacity to discriminate between vegetation types such as meadows and orchards is essential due to their differing roles in ecosystem services, biodiversity support, and land management practices. While meadows are dynamic habitats shaped by grazing or mowing, orchards—especially traditional—-combine tree cover and grassland elements, offering mosaic habitats critical for pollinators, birds, and other wildlife. Sentinel-2 imagery has become a core resource for large-scale vegetation classification due to its spatial resolution, frequent revisit time, and rich spectral range across visible, near-infrared (NIR), and red-edge bands.[2,3] The temporal and spectral richness of Sentinel-2 enables detailed monitoring of phenological cycles, vegetation health, and structural variation, which are key factors in differentiating complex vegetation types.[4]

Despite this potential, vegetation classification remains challenging with current approaches. Many models rely on static or single-date imagery, which fails to capture temporal dynamics that are crucial for separating visually similar classes such as wet vs. standard meadows or traditional vs. intensive orchards.[3] Furthermore, existing work often treats models as "black boxes", with limited interpretability regarding which spectral bands or time periods contribute most to classification.[4] Detailed, band-wise analysis and model transparency are

Further author information: (Send correspondence to C.C.)
C.C.: E-mail: charalambos.chrysostomou@eratosthenes.org.cy

essential, especially when results inform policy or conservation outcomes. Finally, most prior studies focus on binary classification or artificially balanced scenarios, without testing models in realistic, multi-class ecological contexts. Recent studies confirm that analyzing feature importance in Sentinel-2 time series improves vegetation classification.[5] Identifying the most influential features makes models more robust and easier to interpret.

To fill these gaps, this study explores the application of feature importance analysis and model interpretability techniques for classifying meadows and orchards using Sentinel-2 time series data. We compare the performance of three machine learning models—Random Forest, XGBoost, and a neural network—across four tasks: intra-class classification of meadows and orchards, binary classification between vegetation types, and a full multiclass classification. We also evaluate the spectral contribution of each band over time based on model-derived feature importance scores. Our results demonstrate both the predictive power and interpretability advantages of combining tree-based models with temporal multispectral data for fine-grained vegetation classification.

2. DATA

The dataset used in this research is derived from a publicly available Sentinel-2[6] pixel time series dataset for vegetation classification near Strasbourg in northeastern France.[7] The dataset contains annotated pixels representing four vegetation types: traditional orchards, intensive orchards, standard meadows, and wet meadows. Each pixel is associated with a 156-dimensional time series derived from 39 acquisition dates between 2017 and 2018, with four spectral bands per date: Red, Green, Blue, and Near-Infrared (NIR). All bands are resampled to a spatial resolution of 10 meters. Cloudy observations—occurring in up to 50% of the time steps—were addressed by the dataset authors through linear temporal interpolation. No further atmospheric corrections or vegetation index calculations were performed in this study; all models were trained on the raw surface reflectance values directly. The full dataset consists of **311,739 labelled samples**, distributed across four land cover classes. The majority belong to *Standard Meadow (class 211)* with **236,846 samples**, followed by *Wet Meadow (212)* with **27,892**, *Intensive Orchard (221)* with **25,255**, and *Traditional Orchard (220)* with **21,746 samples**. A detailed breakdown is provided in Table 1.

Table 1: Class distribution in the dataset, including class labels, class names, and number of labeled samples.

Class Label	Class Name	Number of Samples
211	Standard Meadow	236,846
212	Wet Meadow	27,892
220	Traditional Orchard	21,746
221	Intensive Orchard	25,255
Total		**311,739**

To support multiple classification scenarios, we considered:

- **Binary Classification** for meadows (211 vs. 212),
- **Binary Classification** for orchards (220 vs. 221),
- **Binary Classification** between meadows and orchards (211 and 212 mapped as 0, 220 and 221 mapped as 1),
- **Multiclass Classification** across all four vegetation types.

We merged all data into one dataset and applied 10-fold stratified cross-validation to evaluate all models consistently. Columns of metadata such as polygon ID were ignored during model training and validation so that we would only focus on pixel-level spectral features.

3. METHODOLOGY

3.1 Overview

This study evaluates the effectiveness of feature importance analysis for vegetation type classification using multi-temporal Sentinel-2 data. Three supervised learning models—Neural Networks (NN), XGBoost (Extreme Gradient Boosting), and Random Forests (RF)—are designed and compared in binary and multiclass classification problems of meadows and orchards. All of the models are trained and validated through 10-fold stratified cross-validation to ensure robustness. Every data sample corresponds to a single pixel, represented as a multivariate time series with 156 input features. These features consist of surface reflectance values for four spectral bands (Red, Green, Blue, and NIR) collected over 39 acquisition dates from 2017 to 2018. All models are trained on these raw reflectance values, with no vegetation indices or engineered features.

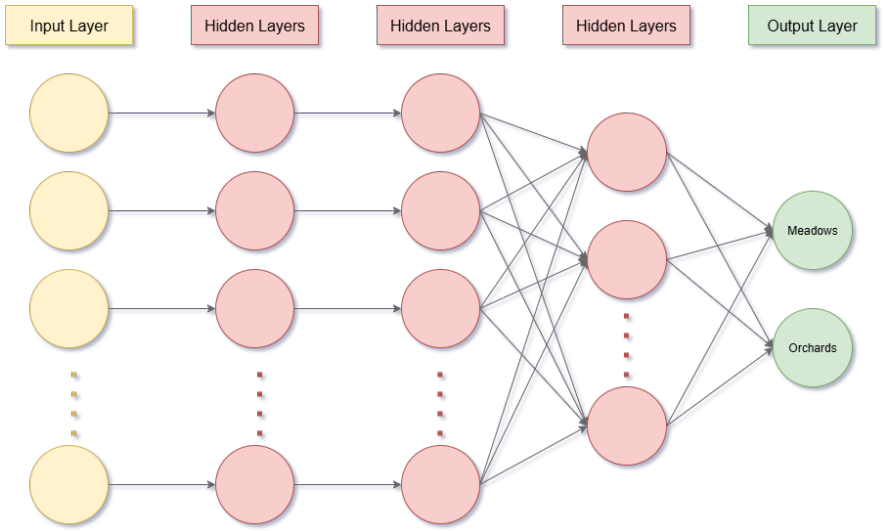


Figure 1: Proposed neural network architecture with an integrated feature selection layer for dynamic input weighting, followed by hidden layers and an output layer for land cover classification (e.g., meadows, orchards)

3.2 Model Architectures

A feedforward fully connected network is employed for the neural network[8] model. The input layer consists of 156 nodes, representing the complete feature set. The architecture has four hidden layers with 256, 128, 64, and 32 neurons, respectively, all with ReLU activation functions[9] (Fig. 1). Dropout of probability 0.3 is used for regularization. The output layer consists of two or four neurons, depending on whether the classification task is multiclass or binary.

For binary classification tasks using the XGBoost algorithm, the model used a binary logistic regression objective function, and performance was evaluated using logarithmic loss, classification error rate, and Area Under the Curve (AUC).[10] For multiclass classification, a multiclass softmax probability objective function was used, and performance was measured using multiclass logarithmic loss, multiclass error rate, and AUC. Key hyperparameters were configured as follows: a learning rate of 0.1, a maximum tree depth of 6, and sub-sampling and column sampling ratios, both set to 0.8. Training proceeded for 100 boosting rounds, incorporating an early stopping criterion of 10 rounds' patience on performance on a validation set from each fold.

The Random Forest classifier[11] was set up with an ensemble of 100 decision trees. For reproducibility, a fixed random seed of 42 was used. The maximum depth of individual trees was not capped in the implementation, and thus, branches expanded until all leaves were pure or had fewer samples than the minimum needed to perform a split. Performance of the models is evaluated using 10-fold stratified cross-validation with proportional representation of each class in every fold. In each fold, the training data is again divided to obtain a validation set

(10% of the training data). Evaluation metrics include accuracy, balanced accuracy, and Matthews Correlation Coefficient (MCC),[12] which is suitable for imbalanced datasets.

4. RESULTS

4.1 Performance Comparison Across Datasets and Models

To evaluate the effectiveness of various machine learning models in the task of vegetation type classification from multi-temporal Sentinel-2 imagery, we trained and validated Neural Networks (NN), Random Forests (RF), and XGBoost (XGB) classifiers on four classification tasks: Meadows (211 vs. 212), Orchards (220 vs. 221), Meadows vs. Orchards (Binary), and a four-class multiclass scenario. The models were validated with 10-fold stratified cross-validation. Metrics reported include test accuracy, balanced test accuracy, and Matthews Correlation Coefficient (MCC), with results averaged across folds. Tables below (See Tab. 2, 3, 4, 5) provide an overview of performances of the models for the four classification tasks. Overall, **XGBoost** and **Random Forest** consistently outperformed the Neural Network baseline, with performance varying slightly depending on the task.

For the *Meadows* classification task (Table 2), **Random Forest** achieved the highest Matthews Correlation Coefficient (MCC) of 0.650 and the best test accuracy (0.943 ± 0.011), while **XGBoost** followed closely with an MCC of 0.648 and slightly higher balanced accuracy (0.778). The Neural Network though competitive, lagged slightly behind in all three metrics. In the *Orchards* classification task, (Table 3), **Random Forest** again led, with an MCC of 0.753 and a balanced accuracy of 0.877, indicating strong model performance in distinguishing between traditional and intensive orchards. The Neural Network achieved comparable performance (MCC 0.741), while XGBoost trailed slightly in all metrics. For the *Meadows vs. Orchards (Binary)* classification (Table 4), XGBoost once led with an MCC of 0.692 and a balanced test accuracy 0.815 ± 0.074 indicating strong separability between the general vegetation categories. Lastly, for the *Multiclass* classification task (Table 5) between all four vegetation types, XGBoost performed best with an MCC of 0.653. The Neural Network and Random Forest each performed comparatively worse on this task, with decreased balanced accuracies, indicating greater challenge in classifying between classes when sub-classes variability is introduced.

Table 2: Classification results on the Meadows dataset (211 vs. 212). Metrics are averaged over 10 folds and reported as mean ± standard deviation.

Meadows (211 vs. 212)			
Model	**Test Accuracy**	**Balanced Accuracy**	**Test MCC**
Neural Network (NN)	0.927 ± 0.027	0.775 ± 0.061	0.593 ± 0.113
Random Forest (RF)	0.943 ± 0.011	0.763 ± 0.070	0.650 ± 0.086
XGBoost (XGB)	0.941 ± 0.012	0.778 ± 0.063	0.648 ± 0.101

Table 3: Classification results on the Orchards dataset (220 vs. 221). Metrics are averaged over 10 folds and reported as mean ± standard deviation.

Orchards (220 vs. 221)			
Model	**Test Accuracy**	**Balanced Accuracy**	**Test MCC**
Neural Network (NN)	0.892 ± 0.062	0.864 ± 0.089	0.741 ± 0.172
Random Forest (RF)	0.888 ± 0.061	0.877 ± 0.062	0.753 ± 0.137
XGBoost (XGB)	0.890 ± 0.061	0.865 ± 0.084	0.740 ± 0.158

4.2 Spectral Band-Wise Feature Importance Analysis

To further interpret our results, we conducted post hoc feature importance analysis on the best-performing model for each classification task: Random Forest for the Meadows and Orchards, and XGBoost for binary and

Table 4: Classification results for Meadows vs. Orchards binary classification. Classes 211 and 212 are mapped to 0, and 220 and 221 to 1.

Meadows vs. Orchards (Binary)			
Model	**Test Accuracy**	**Balanced Accuracy**	**Test MCC**
Neural Network (NN)	0.915 ± 0.031	0.826 ± 0.062	0.651 ± 0.111
Random Forest (RF)	0.929 ± 0.032	0.794 ± 0.070	0.680 ± 0.118
XGBoost (XGB)	0.931 ± 0.031	0.815 ± 0.074	0.692 ± 0.124

Table 5: Classification results for the multiclass classification task (211, 212, 220, 221). Metrics are averaged over 10 folds and reported as mean ± standard deviation.

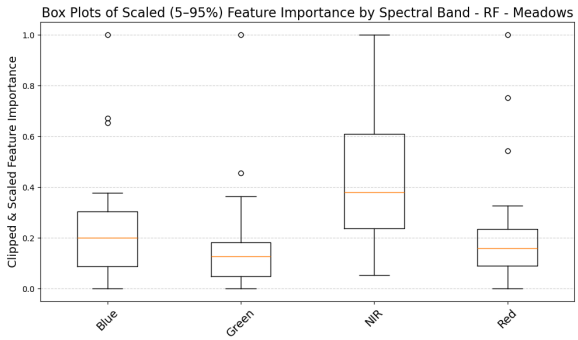
Meadows vs. Orchards (Multiclass)			
Model	**Test Accuracy**	**Balanced Accuracy**	**Test MCC**
Neural Network (NN)	0.842 ± 0.039	0.656 ± 0.081	0.596 ± 0.108
Random Forest (RF)	0.864 ± 0.043	0.604 ± 0.123	0.606 ± 0.154
XGBoost (XGB)	0.876 ± 0.037	0.653 ± 0.101	0.653 ± 0.116

multiclass *Meadows vs. Orchards* tasks. A built-in importance metric is used to quantify the contribution of each spectral-temporal feature. Specifically, the 'gain'-based importance scores from XGBoost and the Mean Decrease in Impurity (MDI) for the Random Forest are used, both of which reflect how much each feature contributes to decision-making during model training. The importance of each input is calculated as `feature_importances_` for Random Forests and `gain`-based importances for XGBoost. In all cases, the full 156-dimensional input was considered, with importance values aggregated per band by first averaging across time steps and then across 10 cross-validation folds. To enhance comparability, all feature importance values were clipped to the 5th-95th percentile range and min-max scaled to the $[0, 1]$ interval.

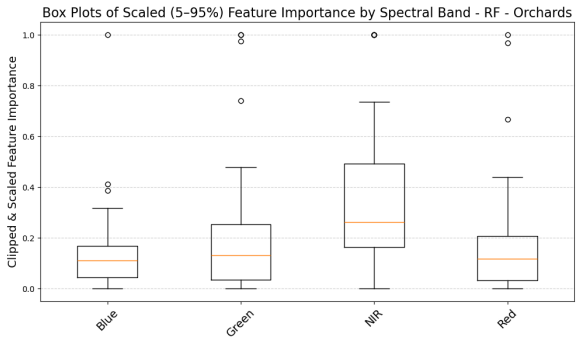
Figure 2 shows the distribution of scaled importances by spectral band for each of the four classification tasks. The results highlight that the NIR band contributes the most to model performance in the *Meadows* (See fig. 2a) and *Orchards* (See fig. 2b) tasks, where the goal is to distinguish between sub-classes within a single vegetation type (See table 1). This suggests that NIR plays a key role in capturing subtle physiological or structural differences within meadows and orchards, consistent with its known sensitivity to biomass and chlorophyll content. In contrast, for the binary and multiclass *Meadows vs. Orchards* (See fig. 2c & 2d) tasks, no single band consistently dominates. Instead, feature importance is more evenly distributed across all four spectral bands. This suggests that when the classification task is to distinguish between higher-level vegetation classes, the model draws on a wider range of spectral information than a particular dominant band. These findings highlight the varying spectral cues leveraged by the model based on whether the classification is intra-class (within vegetation types) or inter-class (between vegetation types).

5. DISCUSSION

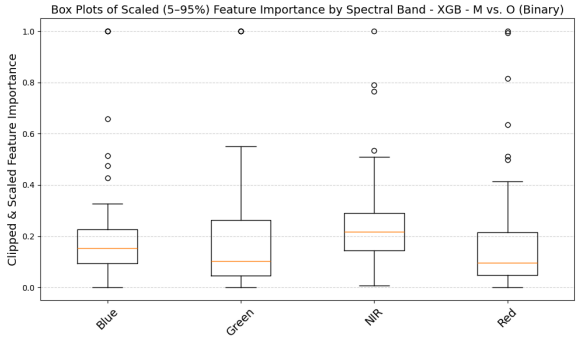
The results of this study provide an informed perspective on how different machine learning models and spectral features contribute to vegetation classification using Sentinel-2 data. With all three models—Random Forest, XGBoost, and Neural Networks performing competitively across the four tasks, some patterns emerged that warrant closer examination. Surprisingly, the Neural Network models, despite their ability to learn non-linear representations, did not consistently outperform the tree-based models.. One possible explanation is that the tabular nature of the input features may favour ensemble models like Random Forest and XGBoost, which are known to handle high-dimensional structured data more effectively. These models also benefit from built-in mechanisms to manage feature redundancy, class imbalance, and noise, which may give them an edge over neural networks in this particular setting. XGBoost achieved the highest performance in both the binary and multiclass *Meadows vs. Orchards* classification tasks. This may be due to its ability to capture complex, non-linear feature interactions more effectively than Random Forest. In contrast, Random Forest slightly outperformed the other



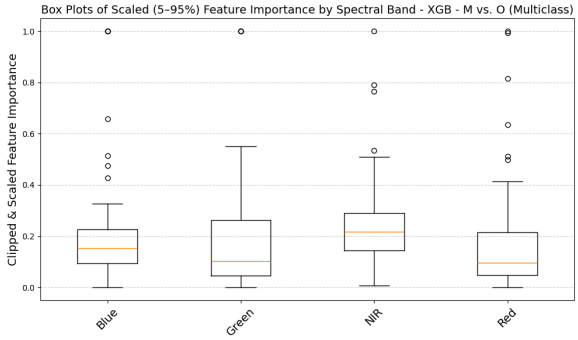
(a) RF – Meadows



(b) RF – Orchards



(c) XGB – Meadows vs. Orchards (Binary)



(d) XGB – Meadows vs. Orchards (Multiclass)

Figure 2: Box plots of clipped and scaled feature importance per spectral band for the best-performing models in each classification task. Feature importances are averaged across time steps and 10 folds.

models on the *Meadows* and *Orchards* tasks, likely benefiting from its robustness to noise and its strength in handling subtle intra-class distinctions through ensemble voting.

The spectral band-wise feature importance analysis provided additional insight. In the *Meadows* and *Orchards* classification tasks—where the goal is to distinguish between subtypes of a single vegetation class—the NIR band was consistently the most informative. This aligns with the established findings in remote sensing, where NIR is highly sensitive to vegetation health, structure, and biomass. Its dominance in these sub-class tasks suggests that it captures fine-grained physiological differences between standard and wet meadows or between traditional and intensive orchards.[13] In contrast, for the inter-class binary and multiclass classification tasks (i.e., distinguishing meadows from orchards), feature importance was more evenly distributed across all four bands. This could suggest that, when distinguishing broader vegetation categories, the model uses a combination of spectral features to make more general distinctions. It is likely that each band provides complementary information based on the spectral signatures of the vegetation types involved. In addition to predictive performance, the feature importance analysis serves a critical interpretability role. By identifying which spectral bands and time steps were most valuable to the model, we gain insight into how temporal and spectral variation contribute to classification performance. This kind of analysis is essential for validating the scientific relevance of machine learning models in remote sensing.

While this study focuses exclusively on Sentinel-2 data, incorporating Sentinel-1 (SAR) data or exploring data fusion strategies may further improve classification accuracy and robustness. Additionally, testing the generalizability of the models in geographic regions or for other years would help assess their operational potential and adaptability.

6. CONCLUSION

In this study, we evaluate the effectiveness of feature importance analysis and machine learning models for the classification of vegetation types—meadows and orchards—using Sentinel-2 multispectral time series data. We examine four classification scenarios: sub-class classification (standard vs. wet meadows; traditional vs. intensive orchard), meadows vs. orchards (binary classification, and full multiclass classification across all vegetation subtypes. Three machine learning algorithms—Random Forest, XGBoost, and a neural network—are evaluated on these tasks. Although all models demonstrated strong performance, XGBoost consistently achieved the highest accuracy and MCC in the binary and multiclass classification tasks. Random Forest outperformed the other models in the sub-class settings, likely due to its robustness in handling subtle spectral differences. Neural networks, contrary to expectations, did not outperform the tree-based methods, possibly due to the tabular structure of the input data and limited sample complexity.

In addition to classification performance, we conducted a detailed spectral band importance study based on model-specific feature importance metrics. Our results showed that the Near-Infrared (NIR) band played a dominant role in the sub-class classification tasks, where fine-grained vegetation structure matters most. In contrast, all four bands contributed more evenly to the binary and multi-class classification tasks, suggesting complementary spectral information are required to distinguish broader vegetation classes. Overall, the results show that both spectral-temporal analysis of Sentinel-2 imagery and tree-based models form an effective basis for vegetation classification applications. Data fusion with Sentinel-1 SAR imagery, and model generalizability to new geographical areas or growing seasons can be explored in future research.

7. ACKNOWLEDGMENTS

This work was supported by the European Union's HORIZON Research and Innovation Programme by the 'EXCELSIOR': ERATOSTHENES: Excellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (www.excelsior2020.eu). The 'EXCELSIOR' project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

REFERENCES

[1] Mucina, L., "Classification of vegetation: Past, present and future," *Journal of Vegetation Science* **8**(6), 751–760 (1997).

[2] Christofi, K., Chrysostomou, C., Tsardanidis, I., Mavrovouniotis, M., Guerrisi, G., Kontoes, C., and Hadjimitsis, D. G., "Remote sensing of grasslands: Performance comparison of radar and optical data in machine learning classification," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **48**, 295–300 (2025).

[3] Antoniadis, K., Georgopoulos, N., Katagis, T., Stavrakoudis, D., and Gitas, I. Z., "Classification of seasonal sentinel-2 imagery for mapping vegetation in mediterranean ecosystems," in [*Ninth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2023)*], **12786**, 73–78, SPIE (2023).

[4] Saini, R. and Ghosh, S. K., "Exploring capabilities of sentinel-2 for vegetation mapping using random forest," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **42**, 1499–1502 (2018).

[5] Christofi, K., Chrysostomou, C., Tsardanidis, I., Mavrovouniotis, M., Kontoes, C., and Hadjimitsis, D. G., "Deep learning-based grassland mapping with sentinel-2: Prioritizing key spectral bands and time periods," in [*Eleventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2025)*], SPIE (2025).

[6] Jutz, S. and Milagro-Perez, M., "Copernicus: the european earth observation programme," *Revista de Teledetección* (56), V–XI (2020).

[7] L., B., "Meadows vs orchards," (2022). Accessed: 2025-07-07.

[8] Sze, V., Chen, Y.-H., Yang, T.-J., and Emer, J. S., "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE* **105**(12), 2295–2329 (2017).

[9] LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *nature* **521**(7553), 436–444 (2015).

[10] Huang, J. and Ling, C. X., "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering* **17**(3), 299–310 (2005).

[11] Liu, Y., Wang, Y., and Zhang, J., "New machine learning algorithm: Random forest," in [*International conference on information computing and applications*], 246–252, Springer (2012).

[12] Chicco, D. and Jurman, G., "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics* **21**(1), 6 (2020).

[13] Joseph, P., Pflüger, S., Petersen, M., and Siegmund, A., "Advancing the monitoring of traditional meadow orchards: Current approaches and future directions," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **48**, 133–139 (2025).