AUTHOR(S)
Zelioli, L., Farahnakian, F., Farahnakian, F., Middleton, M., & Heikkonen, J.

TITLE
Enhancing Peatland Classification using Sentinel-1 and Sentinel-2 Fusion with Encoder-Decoder Architecture.

YEAR
2024

DOI
10.23919/fusion59988.2024.10706276

CITATION
Zelioli, L., Farahnakian, F., Farahnakian, F., Middleton, M., & Heikkonen, J. (2024). Enhancing Peatland Classification using Sentinel-1 and Sentinel-2 Fusion with Encoder-Decoder Architecture. *2024 27th International Conference on Information Fusion (FUSION)*, 1957, 1–7. https://doi.org/10.23919/fusion59988.2024.10706276

VERSION
Accepted Manuscript

Enhancing Peatland Classification using Sentinel-1 and Sentinel-2 Fusion with Encoder-Decoder Architecture

Luca Zelioli[1], Fahimeh Farahnakian[1,2], Farshad Farahnakian[1], Maarit Middleton[2], and Jukka Heikkonen[1]
[1]Department of Computing, University of Turku, Finland
[2]Geological Survey of Finland (GTK)
Email: {luzeli, fahfar, farfar, jukhei}@utu.fi, {maarit.middleton}@gtk.fi

***Abstract*—Peatland classification provides valuable information for greenhouse gas inventory and biodiversity protection. In this paper, we proposed an encoder-decoder-based architecture for peatland classification that fuses two open-source satellite data, Sentinel-1 and Sentinel-2. We show the effect of fusion by comparing the multi-modal fusion architecture with uni-modals which are trained only based on one input data source. We also investigate the influence of skip connections as the main component of the encoder-decoder to recover fine-grained details that are lost during the downsampling process. The experimental results are acquired on a study area in Finland which covers a variety minerotrophic aapa mire peatlands. The results demonstrate that multi-modal architecture consistently outperforms uni-modal architectures for peatland classification. In addition, the fusion architecture with one skip connection achieved a total accuracy of 57.44%. This shows 8.51% accuracy improvement compared with the model without skip connections.**

***Index Terms*—encoder-decoder, data fusion, peatland classification, skip connections, remote sensing, land cover classification**

I. Introduction

Peatland classification is a fundamental subproblem of land cover classification that aims at automatically labeling a pixel in a raster dataset to a specific category of peatlands. It can help us to understand where different types of peatlands occur, which can be important for the purpose of conservation and greenhouse gas inventory. In addition, peatlands perform a variety of important functions for the environment, such as storing carbon, filtering water, and providing habitat for wildlife [1], [2]. Remote sensing offers several advantages over traditional ground-based methods for peatland classification by collecting data from satellites and airborne sensors [3].

In recent years, Deep Learning (DL) models have advanced considerably in the field of remote sensing. This advancement can be attributed to the ability of Convolutional Neural Networks (CNNs) and encoder-decoder to process data in the form of multiple layers, which applies to processing multiband remote sensing data. Recent networks such as FCN [4], DeepLab [5], SegNet [6], and U-Net [7] demonstrate the ability of CNNs for pixel-wise classification. In addition, CNN outperforms traditional machine learning methods such as Random Forest (RF), and Support Vector Machine (SVM) for wetland classification [8]. However, applying DL models for multi- and hyperspectral images is very challenging with high input data dimensionality and few available labeled data in real applications. Another challenge for peatland classification is the ecological peatland classes are not clearly defined instead they form gradients transforming from one class to another and combined with human-induced changes thus forming highly fragmented landscapes.

To address these problems, we presented a CNN-based network to perform peatland segmentation in our previous work [9] where CNN was trained on different data including optical and radar satellite remote sensing, airborne laser scanning data, and multi-source forest inventory GIS datasets. In conclusion, the evidence presented in this paper [10] clearly demonstrates the importance of Sentinel-1 and Sentinel-2 compared to other data for peatland fertility classification. For this reason, we used the same data for this paper. This paper presents an extension of our previous works [9], [10] on peatland classification using DL. In our previous papers, we used a traditional CNN's structures for fusion satellite images to classify peatlands. Here, we proposed an Encode-Decoder (ED) architecture for the same problem. Compared to CNN, the ED architecture is better able to capture spatial relationships and hierarchical patterns in remote-sensing images. This is because ED consists of two main components: an encoder that extracts features from the input image, and a decoder that reconstructs the image from these features. The ED structure allows the model to learn a more compact representation of the image, which in turn leads to better classification performance. However, as an ED architecture goes deeper, it becomes increasingly likely that fine-grained details will be lost during the downsampling process. To address this issue, skip connections are introduced to recover fine-grained details that may be lost during the downsampling process [11]. By directly connecting the output of an encoder layer to the input of a corresponding decoder layer, skip connections allow the decoder to access information from earlier stages of the encoder, which can help reconstruct the original input [7].

Our results show that the proposed ED architecture achieves an accuracy of 57.44%, which is significantly higher than the accuracy of the CNN architecture (44.36%). This improvement

in accuracy is due to the encode-decoder architecture's ability to better capture the spatial relationships and hierarchical patterns in the data. We also investigate the effect of skip connection on the performance of our proposed architecture. The collected results indicate the skip connections can improve the accuracy of our fusion architecture since they can transfer low-resolution information from the encoder part to high-resolution information from the decoder.

The remainder of the work is organized as follows. Section II discusses some of the most important related works. We describe the study area and data in Section III. The proposed architecture is described in Section IV followed by the evaluation in Section VI. Finally, conclusions are presented in Section VII.

II. Related Work

Peatlands, a significant carbon (C) source compared to other terrestrial ecosystems [12], have undergone extensive drainage in Finland since the 1950s for various purposes such as forestry, agriculture, energy production, road construction, and peat extraction. Additionally, development activities like water reservoir creation have further impacted them. This widespread human intervention has raised concerns about the health and sustainability of peatlands, prompting researchers to explore various methodologies for analyzing them. Generally, the segmentation methods using satellite imagery can be categorized into three main groups: (1) pixel-based segmentation, (2) object-based image analysis (OBIA), and (3) DL-based segmentation [13]. All segmentation methods have been summarized and visualized in Fig. 1. In this section, we only review existing studies that used segmentation or pixel-wise classification methods for remote sensing monitoring.

Traditional machine-learning approaches rely on hand-crafted features, such as texture, color, and shape, extracted by human experts. These features are then fed into algorithms like Random Forest (RF) and Support Vector Machine (SVM) for peatland classification. For example, studies in [14], [15] used RF and SVM to classify tropical peatlands and various peatland vegetation types, respectively, using multi-sensor satellite imagery. While these traditional methods offer varying levels of accuracy, data type compatibility, and analysis complexity, their reliance on hand-crafted features limits their ability to capture the full complexity and rich semantic information present in remote sensing data.

DL algorithms remove the need for manual feature extraction by automatically discovering relevant and more complex features within the data. One of the widely-used DL techniques for segmentation tasks is the ED architecture which has demonstrated significant effectiveness in various applications, including peatland classification through satellite data fusion. This approach has been explored in different contexts, aiming to leverage the capabilities of multi-modal satellite data sources for environmental monitoring and land management. In [10], the authors proposed a CNN-based methodology for boreal peatland fertility classification by fusing Sentinel-1 and Sentinel-2 imagery, which aligns with our approach

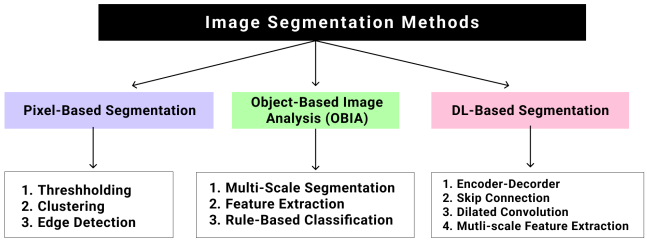

Fig. 1: The categories of image segmentation methods.

of combining Sentinel-1 and Sentinel-2 data for peatland classification. In another article [16], the use of encoder-decoder architecture was further explored in their development of SegNet, a deep convolutional encoder-decoder architecture for image segmentation.

In [17], the authors discussed and tested the fusion of Sentinel-1 and Sentinel-2 image time series models for permanent and temporary surface water mapping, showcasing the potential of multi-source satellite data fusion in environmental analysis. Similarly, in another article [18], the effectiveness of Sentinel-1 and Sentinel-2 data fusion was examined for urban change detection using a dual-stream U-Net, underscoring the advantages of combining synthetic aperture radar (SAR) and optical data in capturing dynamic urban changes.

Moreover, the impact of multi-modal data fusion and the significance of ED architectures have been demonstrated across various studies focusing on different environmental and land management applications. In [19], [20], the authors worked on early crop classification through multi-modal satellite data fusion, and the investigation into the classification and monitoring of corn nitrogen concentration from Sentinel-1 and Sentinel-2 data fusion.

Unlike traditional pixel-based and object-based methods, which often struggle with the high variability and complexity of satellite data, our methodology leverages the complementary strengths of both Sentinel-1 and Sentinel-2 imagery, enabling more accurate and robust semantic segmentation across diverse environmental conditions and landscapes. Furthermore, by employing ED models with the skip connection technique, our work achieves superior segmentation accuracy and detail recovering fine-grained details lost during the downsampling process.

III. Data and Methods

A. Study area

The Keminmaa study area, as depicted in Fig. 2, lies within the southern region of the *aapa* mire zone in northern Finland [21]. These *aapa* mires are characterized as minerotrophic vegetation-ecological peatland complexes with flat lawn-level vegetation and concave surface topography. They receive additional nutrients from their immediate surroundings, as elucidated by Eurola [22]. Within this study area, there exists a diversity of peatland site types across

TABLE I: Fertility level and number of samples and their abbreviations.

Fertility level (classes)	Description of fertility class	No. samples	
		Drained	Undrained
FL1	Herb rich type	422	58
FL2	Vaccinium myrtillus type	378	374
FL3	Vaccinium vitis-idaea type	358	118
FL4	Dwarf shrub type	69	37
FL5	Cladina type	68	99
Abandoned field on peat soils (AFOPS)	Abandoned field on peat soils	24	20
Organic soil agricultural fields (OSAF)	Organic soil agricultural fields	40	0
		1359	**706**

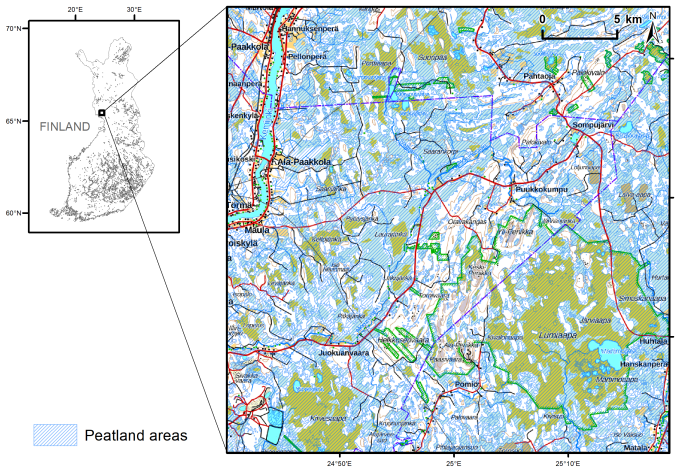

Fig. 2: Location of the Keminmaa study area in northern Finland. Contains data from the National Land Survey of Finland Topographic Database 2023.

various fertility classes, attributed to the assortment of local lithologies and predominantly, glacial till subgrade. These peatlands possess an average depth of 1.1 meters, covering approximately 50% (42800 hectares) of the total land area. To refine the classification process and mitigate potential errors, the study area was stratified into distinct subareas labeled as 'drained' and 'undrained', distinguishing between pristine regions and those previously subjected to artificial forestry drainage. Subsequently, these subareas were modeled and assessed independently. The classification system followed the Finnish peatland fertility level classification with five classes: Herb rich type/Oxalis-Myrtillus type(FL1), Vaccinium myrtillus type I (FL2), Vaccinium vitis-idaea type II(FL3), Dwarf shrub type (FL4) and Cladina type (FL5). Table I shows the ground truth of the main five fertility levels and two land use classes used in this paper. The total number of labeled samples is 1359 and 706 for drained and undrained, respectively. The labeled dataset is limited with an imbalanced organization of the classes.

B. Dataset

Table II shows the characteristics of the data used in this paper. Synthetic Aperture Radar (SAR) satellite data comprises two distinct types of Sentinel-1 (S1) imagery. 1) The intensity images (i.e., ground range detected products) including both VV and VH polarizations (V=vertical, H=horizontal)

TABLE II: Description and number of inputs in the proposed study area.

Data	Band#	Wavelength	Pixel size (m)
S1 intensity	Microwave band C (VV,VH)	5.5 cm	10
S1 coherence	Microwave band C (VV,VH)	5.5 cm	10
S2	B2 (Blue)	490 nm	10
	B3 (Green)	560 nm	
	B4 (Red)	665 nm	
	B5 (Vegetation Red Edge1)	705 nm	
	B6 (Vegetation Red Edge2)	740 nm	
	B7 (Vegetation Red Edge3)	783 nm	
	B8 (NIR)	842 nm	
	B8A (Vegetation Red Edge4)	865 nm	
	B11 (SWIR1)	1610 nm	
	B12 (SWIR2)	2190 nm	

were downloaded from the Copernicus Open Access Hub[1] and further processed using the ESA Sentinel Applications Platform (SNAP) software. These intensity images represent the backscattered signals detected, multi-looked, and projected into a single image. For analysis, five intensity images were utilized, each corresponding to a different summer month in 2017 (May-September). 2) The coherence images, derived from single-look complex products, capture the similarity of radar reflections between pairs of individual images. Two distinct sets of coherence images were generated. The first set comprises 12 coherence images calculated for each temporally consecutive pair of Sentinel-1A images, spaced at 12-day intervals between May 9 and September 30, 2017, denoted as 'repeated overpass' (RO) in this study. The second set consists of another 12 coherence pairs calculated for the same set of Sentinel-1A images, with the mid image of July 20, 2017, consistently included as a "reference" image in each pair, termed as 'long-term baseline' (LB) coherence.

Sentinel-2 (S2) data were acquired from the Copernicus Open Access Hub. A single image was selected for each summer month (May-September) spanning the years 2018 to 2020. The image chosen for analysis in each month was meticulously evaluated for the least cloud cover or haze, ensuring optimal data quality. Notably, the study area was encompassed within a single S2 frame. All ten bands of the acquired imagery, initially captured at a resolution of 10/20 meters, were utilized as input data. To ensure consistency, the imagery was resampled to a uniform 10-meter raster stack for subsequent analysis. In [23], the authors completely provide more details about the dataset and the pre-processing stage.

IV. Architecture

The proposed architecture is shown in Fig. 3. It is a U-net style encoder-decoder network designed commonly for semantic segmentation and it is adopted for fusion of three input satellite data. Our multi-modal architecture has three branches which are separately extract the features of each input data and finally concatenate the output of each branch to perform pixel-wise peatland classification. Each branch consists of two parts as follows:

[1]https://scihub.copernicus.eu/

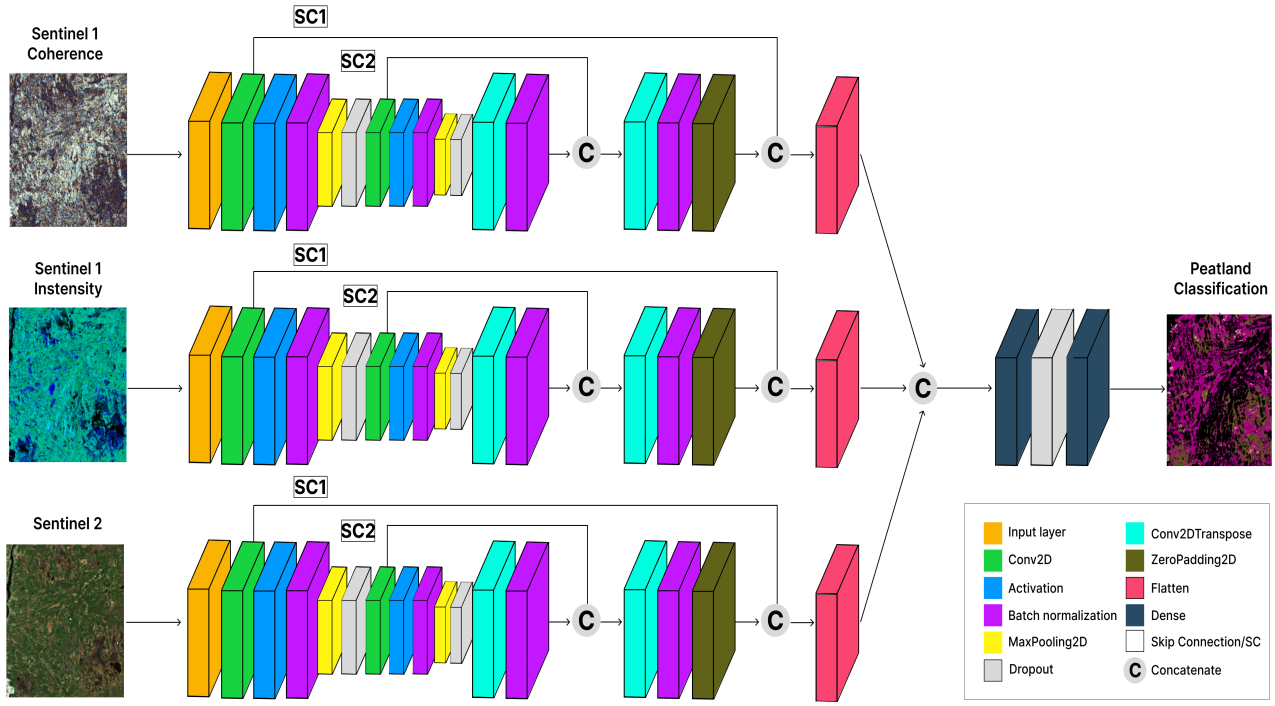

Fig. 3: The proposed multi-modal encoder-decoder fusion architectures for peatland semantic segmentation.

1) Encoder: The encoder part contains the two convolutional layers structured in a list format, sequentially applying 10 and 20 filters. Relu and batch normalization layers are added after each convolutional layer. Then, Max-pooling and down-sampling operations are employed with a pool size and upsampling factor of 2, respectively, aiding in feature extraction and reconstruction. We also implemented dropout layers with a 10% dropout probability after each pooling layer to reduce overfitting.
2) Decoder: The decoder part performs the inverse operation, gradually upsampling the feature maps. Upsampling is achieved using transposed convolutional layers, which increase the spatial dimensions of the feature maps.

After the final flatten layer in the decoder part, the outputs of three separate modalities are concatenated for late fusion. The contribution of proposing the late fusion is our previous work [10] that shows late fusion provides better performance than an early fusion for peatland classification. Finally, additional dense layers, including three dense layers with 16 units, contribute to feature learning and refinement. The final output layer, determined by the number of unique labels in the dataset, utilizes softmax activation for multi-class classification. This layer produces a probability distribution over the classes for each pixel in the input image, effectively segmenting the image into regions corresponding to different fertility classes. In addition, the proposed architecture consists of two Skip Connections (SC1 and SC2) for each modality. They involve simply concatenating the feature maps from the encoder and the decoder. This allows the model to directly combine the low-resolution information from the encoder with the high-resolution information from the decoder.

V. Experimental Setup

To optimize the weights of the network, we used the Adam algorithm [24] and the cross-entropy loss function [25]. Additionally, we used data augmentation techniques, such as random rotation, and vertical and horizontal flips, to generate twice as many training samples. L2 regularization with a rate of change of 0.05 was also applied. Finally, we used 5-fold Stratified cross-validation (SCV) [26] to evaluate the generalization accuracy of the classifier. SCV ensures that each fold of the dataset has the same distribution of the classes in each fold. This helps to address the class imbalance issues. To find the best values of all defined hyper-parameters in the proposed architectures, we have done a hyper-parameter tuning based on grid search.

To reduce computational complexity, the encoder-decoder model is trained on small regions (windows) that are assumed to contain the spatial information around each training pixel. This approach is commonly used in pixel classification tasks [27] and is well-suited for this study because the annotated fertility level classes are determined in the field within a 40-meter radius of each training point location. Each training pixel is centered at a window with a size of 5×5 pixels. The trained ED model is then applied to extracted windows of the

TABLE III: The classification accuracy (%) of ED fusion architectures vs uni-modal architectures with and without skip connection.

Architecture	Input	Encoder-Decoder (without SC)		Encoder-Decoder (with SC1)		Encoder-Decoder (with both SC1 and SC2)		The proposed CNN in [10]	
		Drained	Undrained	Drained	Undrained	Drained	Undrained	Drained	Undrained
Uni-modal	S1 intensity	38.23	39.71	36.74	39.43	37.63	39.89	30.88	35.91
Uni-model	S1 coherence	40.91	40.26	41.17	40.29	43.80	43.42	34.19	35.09
Uni-modal	S2	41.69	46.80	45.22	47.05	**47.51**	**48.93**	37.13	41.65
Multi-modal	S1 intensity + S2	45.95	48.93	50.18	56.73	**54.05**	**57.44**	**50.36**	**44.36**
	S1 intensity + S1 coherence	42.74	44.36	42.64	44.18	46.17	45.68	45.22	45.39
	S2 + S1 coherence	45.74	47.13	47.42	47.51	49.65	57.35	49.44	56.73
	S1 intensity + S1 coherence + S2	43.75	45.62	47.05	46.80	47.79	51.06	45.58	43.26

same size as the ones used for training. These windows are moved across the entire satellite image, allowing the model to classify each pixel into one of the seven defined fertility level classes.

VI. Results and Discussion

The following experiments were designed to reply to three main research questions related to developing an efficient DL model for peatland pixel-wise classification. (1) How much can the accuracy be improved when we fuse S1 and S2 data? (2) How does the classification accuracy change when the proposed ED architectures use a skip connection? (3) Can ED perform better than CNN for peatland fertility classification based on satellite images?

A. Uni-modal vs multi-modal

This section presents a comparison of uni-modal and multi-modal architectures for peatland classification. The results in Table III demonstrate that the multi-modal architecture consistently outperforms uni-modal architectures as it can extract richer features and improve performance. The best multi-modal model achieved an accuracy of 57.44% for undrained, compared to 48.93% for the uni-modal model based on S2 (highlighted with the bold black color). Between, uni-modal architectures, we got the highest accuracy (47.51% and 48.93%) when ED is trained on S2 images for draining for drained and undrained, respectively (highlighted with red color).

B. Effect of skip connection

Skip connection is an efficient way to reduce information loss during the upsampling process in the decoder, leading to more accurate and detailed segmentation results [28]. To investigate this fact, we trained three ED-based uni-modal and multi-modal architectures without and with skip connections. Table III shows the classification accuracy of these architectures. From the results, we can conclude the following observations:

1) Without skip connection: The ED model with no skip connections achieved maximum accuracy of 45.95% and 48.93% with multi-modal based on S1 intensity and S2 for drained and undrained, respectively.
2) With one skip connection (SC1): The same multi-modal model with one skip connection achieved the maximum accuracy of 50.18% and 56.73% for drained and undrained, respectively. This is a significant improvement over the model without skip connections. This suggests that adding one skip connection can help to preserve information that is lost during the downsampling process.
3) With two skip connections (SC1 and SC2): The same model with two skip connections achieved the maximum accuracy of 52.05% and 57.44% for drained and undrained, respectively. This is the best performance of the ED models. This suggests that adding two skip connections can further improve the model's performance.

Overall, the fusion of features from different depths of the network through skip connections leads to a more comprehensive and robust representation of the input image.

C. Encoder-decoder vs CNN

We compared the performance of our proposed encoder-decoder model (ED) with the CNN model in [10]. We trained all models on the same data and training configuration. The results (Table III) show that the ED model achieved superior performance on all architectures compared to the CNN model. For example, the best ED multi-modal accuracy got 3.69% and 13.08% improvement than CNN model for drained and undrained, respectively (highlighted with blue color).

D. Confusion Matrix

Fig.4 illustrates the confusion matrix utilized to determine the accuracy of the architectures employed in achieving the highest accuracy rates for both drained and undrained classifications. The matrix is used for evaluating the accuracy of classification within each fertility level class. In the classification of drained subsections (as depicted in Fig.4(a)), the highest number of correct predictions is observed for the FL1 class, 64% of the actual FL1 instances were correctly predicted as FL1 by the model. Conversely, the AFOPS class shows high misclassification rates, with only 18% correctly predicted and substantial portions being incorrectly predicted as FL1 (42%) and FL3 (20%). This misclassification arises due to the similarity in forest cover, primarily dominated by Scots pine and Norway spruce, making it challenging to differentiate using solely S1 and S2 data. Turning to undrained scenarios (illustrated in Fig.4(b)), the highest number of correct predictions aligns with FL2 (74%). However, a considerable proportion of FL2 samples (18%) are misclassified as FL3 (Vaccinium vitis-idaea type).