SWIFTT

Satellites for Wilderness Inspection and Forest Threat Track

WP1 Model Development & Improvements

D1.4 Intermediary Report with model results checked on the ground

Version: 1.0

Date: 31/07/2025

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Document control

Project title	Satellites for Wilderness Inspection and Forest Threat Tracking
Project acronym	SWIFTT
Call identifier	HORIZON-EUSPA-2021-SPACE
Grant agreement	101082732
Starting date	01/11/2022
Duration	42 months
Project URL	http://swiftt.eu
Work Package	WP1 Model development and improvement
Deliverable	D1.4 Intermediary report with model results checked on the ground
Contractual Delivery Date	M32
Actual Delivery Date	M32
Nature[1]	R
Dissemination level[2]	PU
Lead Beneficiary	ERS - Ecosystem Restoration Standard (ERS)
Editor(s)	Gauthier Masson
Contributor(s)	Annalisa Appice (UNIBA), Pasquale Ardimento (UNIBA), Nicola Boffoli (UNIBA), Donato Malerba (UNIBA), Sven Ysker (LUH), Quentin Voituron (AXAC)
Reviewer(s)	Annalisa Appice, Andrii Shelestov, Hanna Yailymova
Document description	The document is an intermediary technical report for the SWIFTT project, detailing the development, validation, and on-ground testing of machine learning models to detect forest threats—specifically bark beetle outbreaks, windthrow, and fire risk—using Sentinel satellite imagery.

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Version control

Version[3]	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.1	Gauthier Masson	01/07/2025	Table of Content
0.3	Annalisa Appice (editor), Donato Malerba, (contributor) Pasquale Ardimento (contributor), Nicola Boffoli (contributor)	18/07/2025	Section 2 on “Bark beetle model update”
0.5	Quentin Voituron	24/07/2025	Section: “Windthrow model development update”
0.6	Hanna Yailymova	27/07/2025	Document Review
0.7	Gauthier Masson	28/07/2025	Section Introduction
0.9	Sven Ysker	29/07/2025	Section: Fire Risk Model Development Update
1.0	Gauthier Masson	30/07/2025	Document ready for publication

[3]0.1 – TOC proposed by editor; 0.2 – TOC approved by reviewer; 0.4 – Intermediate document proposed by editor; 0.5 – Intermediate document approved by reviewer; 0.8 – Document finished by editor; 0.85 – Document reviewed by reviewer; 0.9 – Document revised by editor; 0.98 – Document approved by reviewer; 1.0 – Document released by Project Coordinator.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Abstract

Forests are essential to life on Earth. They provide habitats for thousands of creatures and combat climate change through carbon sequestration. However, our forests are threatened by insect outbreaks, fires, windthrow and droughts. Notably, insect outbreaks are one of the leading causes of forest loss globally, destroying 85 Mha of forest worth €15B annually. At the same time, wildfires destroy 400 Mha annually on a global scale, according to the European Space Agency. The wind is also a significant forest disturbance agent in the temperate forests of France, Germany, and most of Europe.

Climate change affects forests, causing insects to breed more frequently. It also provides more dry fuel for global wildfires. The dry conditions increase the length of the fire season and the size of areas affected by the fire. In addition, both the frequency and the severity of large storms causing windthrow can be attributed to climate change. As a result, countless habitats are lost, and CO2 sequestered yearly decreases by over 4850M tons.

Our solution, SWIFTT, will provide a scientifically sound and technically feasible way to help monitor and manage forest risks: windthrow, insect outbreaks, and forest fires. SWIFTT will enable forest managers to adapt to climate change with affordable, simple and effective remote sensing tools backed up by powerful machine learning models. Our solution will offer a monthly health monitoring service using Copernicus satellite imagery to detect and map the various risks to which forests and their managers are exposed. Early threat detection aids timely intervention. SWIFTT will be tested in real conditions by several end-users from the forest industry, which include Fürstliches Forstamt, Groupe Coopération Forestière and the Rigas Mezi. We anticipate monitoring and protecting up to 40 Mha of global forests by 2030, saving foresters over €468M in monitoring costs and creating over 50 direct jobs.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Disclaimer

This document does not represent the opinion of the European Union or European Union Agency for the Space Programme (EUSPA), and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

Neither the SWIFTT consortium as a whole, nor a certain party of the SWIFTT consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

Acknowledgement

This document is a deliverable of SWIFTT project. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement Nº 101082732.

Table of Contents

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

List of abbreviations

WP	Work package
ESA	European Space Agency
SWIFTT	Satellites for Wilderness Inspection and Forest Threat Tracking
RF	Random Forest
XGB	XGBoost
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
S2	Sentinel-2
GEE	Google Earth Engine
SVI	Spectral Vegetation Index
SS	Spectral-spatial
CVA	Change Vector Analysis
CNN	Convolution Neural Network
XAI	eXplainable Artificial Intelligence
LLM	Large Language Model
AI	Artificial Intelligence
IoU	Intersection over Union
R	Recall
P	Precision
F	Fscore
OA	Overall Accuracy
PCA	Principal Component Analysis
SCL	Scene Classification Map
DE	Deep Embedding
FWI	Fire Weather Index

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

List of tables

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

List of figures

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Introduction

Overview of the SWIFTT Project

The SWIFTT project empowers forest managers in monitoring and mitigating risks such as windthrow, insect outbreaks, and forest fires, which are increasingly intensified by climate change. By leveraging Copernicus satellite imagery and advanced machine learning models, SWIFTT's goal it to offer an affordable platform using sophisticated remote sensing tools for comprehensive monthly forest health assessments. The project's emphasis on damage detection and evaluation to facilitate timely interventions, thereby safeguarding forest ecosystems effectively.

Scope of Work Package 1

The focus of Work Package 1 is to develop and enhance the models that foresters will use on the SWIFTT platform. This involves a series of tasks aimed at establishing the modeling infrastructure and refining the models before final selection. The detailed list of tasks associated with WP1 is as follows:

- 1.1 Establish a modeling infrastructure
- 1.2 Set up data flows to fetch and preprocess satellite imagery
- 1.3 Create a high-fidelity, up-to-date forest basemap and masks
- 1.4 Train models for each risk
- 1.5 Retrain the best-performing models with field data

Objective of Task 1.5

This deliverable is connected to the field results section of task 1.5 in Work Package 1: "Retrain best performing models with field data." As part of this task, the consortium must:

- Test the best performing models using current data and collaborate with field partners to evaluate the results. [Done]
- Gather feedback from field partners on the models within the app. [Done]
- Calculate model statistics and retrain the models with updated data. [To-do]

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Bark Beetle Model Development Update

In this Section we first describe new research results achieved by the team UNIBA developing Machine Learning and Deep Learning methods for training models for bark beetle detection in Sentinel-2 images of forest areas. We analyse the accuracy performance of new models trained and evaluated using a Sentinel-2 imagery dataset created according to information regarding historical benchmark bark beetle outbreaks observed in Czeck Republic, in September 2020, and recorded in the DEFID2 database (Forzieri et al 2022). In addition, we present the methodology used to perform the back-testing analysis of the method developed by UNIBA and, finally, selected for the integration in the SWIFTT platform. With this regard we introduce the selection of testers and describe the testing phase execution conducted by UNIBA using the on-ground data already acquired within the data collection of the project. Finally, we illustrate the results of the back-testing phase by reporting an interpretation of results, illustrating limits, learned lessons and recommendation for future testing.

Model Description

Literature review update

The literature review update includes remote sensing studies exploring Convolutional Neural Networks (CNN) in semantic segmentation problems, Data-Centric AI studies exploring the re-use of foundation deep neural models in remote sensing downstream tasks, Artificial Intelligence (AI) studies exploring the use of Large Language Models (LLM) in remote sensing, and unsupervised learning methods for change detection in remote sensing.

CNN. Both Turkulainen et al. (2023) and Zwieback et al. (2024) have recently analysed the performance of CNNs trained for mapping bark beetle outbreaks in several types of satellite images, including Sentinel-2 images. However, both studies neither use the attention mechanism, nor investigate how to explain decisions.

Foundation model-reuse. Fine-tuning is a few-shot learning strategy that is used to adapt a deep neural model to new tasks or case studies without the need to restart training from scratch on the set of newly accumulated data (Peng and Wang, 2020). Fine-tuning of deep neural models has been recently investigated in multiple fields (e.g., natural language processing (Raiaan et al., 2024). Several computer vision studies explore the performance achieved by fine-tuning several large-scale, pre-trained computer vision models showing great potential for the adaptability of powerful representational abilities of vision foundation models across various downstream vision tasks. Although most computer vision studies use fine-tuning with colour images obtained from photographers (Kirillov et al., 2023), there is some recent background on fine-tuning with Sentinel-2 images. For example, Andresini et al. (2023a) has recently considered fine-tuning in combination with active learning as a component of a change detection pipeline tailored for pairs of Sentinel-2 images. In addition, a few recent studies have started to make available foundation deep neural models, pre-trained with big volumes of Sentinel-2 images mainly acquired in general-purpose tasks of land cover classification and segmentation. Wang et al. (2022) describe several pre-trained models, based on ResNet50 and ViT-S/16 architectures, that are obtained from the SSL4EO-S12 dataset – a large-scale multitemporal dataset in Earth observation. Similarly, Sumbul et al. (2021) describe several pre-trained models, based on K-Branch CNN, VGG16, VGG19,

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

ResNet50, ResNet101, and ResNet152, that are obtained from the BigEarthNet Sentinel-2 dataset with the original CORINE Land Cover (CLC) Level-3 class nomenclature of 43 classes for deep learning applications. Finally, Cai et al. (2023) present a pre-trained model, that is obtained for the land cover segmentation of 19 classes in the multi-temporal sentinel-2 PASTIS dataset.

LLM. Although LLMs have recently attracted some attention in the remote sensing domain, the scientific literature on the implementation of LLMs for Earth observation remains limited. Preliminary remote sensing studies have explored LLMs mainly in image captioning and visual question answering (Li et al., 2024). These studies commonly consider multi-model approaches that use LLMs to process textual information (e.g., imagery captions) collected in addition to imagery data, to provide a comprehensive understanding of the scene. The potential of combining LLMs' capabilities and visual image analysis through Vision Language Models has been recently analysed within Visual ChatGPT for generating textual descriptions of images, performing edge and straight-line detection, and conducting image segmentation by using text-based guidance (Osco et al., 2023). Recent studies have started the investigation of several text-supervised semantic segmentation methods that use text descriptions for guidance in image segmentation tasks. These studies employ a text encoder to calculate embeddings of descriptive input labels in conjunction with a transformer-based image encoder (Sun et al., 2024). These approaches are mainly used to mitigate the need for extensive manual labelling to fuel accurate supervision. However, vision language studies consider colour images, neglecting multi-spectral images recorded with the Sentinel-2 technology. In addition, to the best of our knowledge, this is the first study that starts the exploration of how a foundation LLM can be adapted for mapping bark beetle outbreaks based on imagery information of forest areas.

Unsupervised learning for change detection. In the unsupervised machine learning paradigm (Bruzzone & Pireto, 2000), changes are commonly detected by resorting to the Change Vector Analysis (CVA) strategy that computes a measure of similarity (or distance) between co-located pixels of a couple of images and uses a threshold-based approach to identify a distance threshold to separate the changed pixels from the unchanged background. Various similarity (or distance) measures have been investigated for CVA methods (Appice et al., 2020). The threshold to detect the changes is estimated by resorting to the spectral data, i.e. in a data-driven manner (Lu et al., 2010) by leveraging probabilistic information extracted from the distribution of the (distance or similarity) measure among the pixels. A well-known approach commonly used for the threshold determination is Otsu's method (Sahoo et al., 1988). Lopez-Fandino et al. (2019) evaluate the performance of the Otsu's algorithm in combination with SAM and Watershed algorithm. Alternatively, a clustering algorithm is adopted by Appice et al. (2020), to separate distances (or similarities) of changed pixels from the unchanged background.

Algebra-based methods, similarity-based methods, as well as distance-based methods belong to the threshold-based family of change detection approaches. These methods use mathematical operations (such as image differencing or image ratio) on images taken at different times to generate a change matrix output (Ilsever & Unsalan, 201). Similarity-based change detection methods resort to the computation of a similarity measure (e.g., correlation measure) between a pair of spectral vectors (Choi et al., 2010). Distance-based CD methods are founded on a spectral distance measure (e.g. SAM, Z-score Information Divergence) computed between the spectral vectors on corresponding pixels (Choi et al., 2010). Instead,

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

the study of Appice et al. (2020) adopts a spectral-spatial distance for CVA. In addition, it introduces an iterative upgrade of the traditional distance-based approach by accounting for a representation of the possible change iteratively learned through classification. Due to the lack of ground change information, the classification step is supervised with pseudo-labels yielded on the spectral-spatial distance information via clustering.

A different unsupervised perspective performs the change detection on image combination or transformation. Deng et al. (2008) use the Principal Component Analysis (PCA) to extract the difference between two images suppressing correlated information and highlighting variance in multi-temporal data. The change is identified in the second component, while the first component is assumed to be the sum of the common information. Gao et al. (2016) use the PCA as a convolutional filter to determine the representative neighbourhood features from each pixel and generate change matrices with less noise spots. Gabor wavelets and Fuzzy c-means are utilized to select bi-temporal pixels that have a high probability of being changed or unchanged. Then, new image patches centred at selected pixels are generated and a PCANet model--a deep learning network with its convolution filter banks chosen from PCA filters--is trained using these patches. Finally, pixels of bi-temporal images are classified by the trained PCANet model.

Andresini et al., 2022 describe a method that trains an autoencoder artificial neural network on the primary image. Then the reconstruction of both images is restored via the trained autoencoder so that the spectral angle distance can be computed pixelwise on the reconstructed data vectors. Finally, a threshold algorithm is used to automatically separate the foreground changed pixels from the unchanged background. Kalinicheva et al. (2019) train a convolution autoencoder on the patches of a time series of images. The reconstruction error of each patch is analysed to discriminate changed pixels from the background. Finally, Andresini et al. (2023) propose a change detection method based on a Siamese network, which takes advantage of both Transfer Learning and Active Learning to handle the constraint of limited supervision.

Machine Learning and Deep Learning method development update

In this Section we illustrate new achievements of UNIBA in developing and evaluating *supervised* semantic segmentation models for bark beetle outbreak detection. Specifically, four new deep learning approaches were developed and evaluated in the supervised setting:

- AVALON (Attention-based conVolutional neurAL network fOrest tree dieback in seNtinel-2 images) -- an approach that reformulates the semantic segmentation task as an imagery classification task and trains an Attention-based CNN for imagery classification (Recchia et al., 2024).
- EMF (Exchanger+Mask2Former) – an approach that reuses a sophisticated semantic segmentation deep neural model developed for land cover segmentation with a large amount of accurately annotated multi-temporal Sentinel-2 data (Andresini et al. 2024a).
- GANDALF (a larGe lANguage moDel-based approach for semAntic segmentation of sentineL-2 images of Forest areas) – an approach that uses a narrative to transform Sentinel-2 spectral data into semantic text stories and fine-tunes a foundation LLM do the downstream sematic segmentation task (Pasquadibisceglie et al., 2025).
- BigEarth-UNet -- an approach that uses a foundation ResNet model, pretrained on a big volume of Sentinel-2 images (BigEarth BigEarthNet v2.0 dataset), as encoder of

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

an UNet architecture. The model is updated on the Sentinel-2 data of the considered downstream task using a fine-tuning strategy (Recchia et al., 2025).

The accuracy performance of methods listed above was investigated by using Sentinel-2 images of scenes that were annotated with bark beetle outbreaks observed in September 2020 in the Czech Republic and recorded in the DEFID2 database (Forzieri et al 2022). We considered the dataset created according to the pipeline using the operator BEST described in the D1.3 intermediary report. For each scene, this pipeline selected the best Sentinel-2 image in the query period according to cloud and noise index. As described in deliverable D1.3, this index was computed based on the output of the Scene Classification Level (SCL) algorithm. Specifically, it was computed as the percentage of imagery pixels that the SCL algorithm recognizes as noise, defective, dark, cloud, cloud shadow or thin cirrus. In the BEST pipeline, the Sentinel-2 image, that in the query period, achieved the lowest cloud and noise index was selected. In this way, we were able to compare the performances of these four new models with the models based on Random Forest (RF), XGBoost, Support Vector Machines (SVM) and UNet (TITANIA and DIAMANTE) described in the D1.3 intermediary report.

In addition, a new pipeline to prepare Sentinel-2 imagery datasets for the development and evaluation of models for bark beetle outbreak detection is described. This new pipeline uses the MEDIAN operator in place of the BEST operator to obtain Sentinel-2 images of scenes in a specific period. Hence, we evaluated the performance of the two models trained for bark beetle outbreak detection by considering images of Czech Republic scenes obtained with the BEST operator and the MEDIAN operator, respectively.

Finally, we illustrate an *unsupervised* method that was evaluated for change detection in a bark beetle outbreak problem. The developed method includes a combination of CVA, Spectral Vegetation Indices (SVIs) and Deep Neural Embedding. OBJ

Supervised Classification and Semantic Segmentation methods

AVALON

This is a semantic segmentation approach that trains a CNN model for the semantic segmentation of Sentinel-2 images of forest areas. It integrates a pixel attention mechanism that allows the adaptive selection of imagery pixels where the network “sees” the most important spectral information for the local decisions. This mechanism can mitigate local disturbances introduced with useless or noise data (especially along the boundary between patches labelled with opposite semantic labels). In addition, the attention provides insight into how the CNN model achieves its decisions by contributing to the enhancement of the explainability of how a pixel neighbourhood contributes to the decision on each semantic label.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

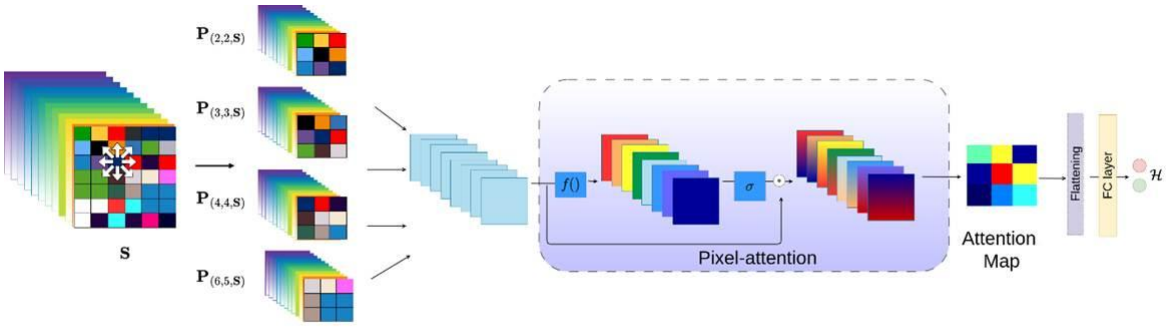


Figure 1: Schema of AVALON. Abbreviations: FC=Fully Connected

Specifically, AVALON takes ***S*** – a collection of labelled Sentinel-2 images – as input. Each Sentinel-2 image $\mathbf{S} \in \boldsymbol{S}$ is a tensor with shape $W_S \times H_S \times C$, which covers an Earth scene spanned across $W_S \times H_S \times C$ pixels that are observed across C spectral bands. Let I(i,j) denote the spectral vector with size C recorded at pixel (i,j) in **S**. Every Sentinel-2 image $\mathbf{S} \in \boldsymbol{S}$ is one-to-one associated with a label mask **M** with shape $W_S \times H_S$ so that every pixel **S**(i,j) is assigned to a binary semantic label **M**(i,j)) that assumes the values: "healthy" (0) or "damaged" (1). AVALON learns a semantic segmentation model to predict the unknown label mask of any new Sentinel-2 image. This is done by reformulating and addressing the semantic segmentation problem as an image classification problem. To this aim, we first transform every Sentinel-2 imagery pixel of the input collection into an image that shows the pixel within its pixel neighbourhood. Then, we train a CNN with Attention for the image classification task. The pipeline of AVALON is shown in Figure 1.

To achieve a formulation of the semantic segmentation problem as an image classification problem, we transform ***S*** into a new imagery dataset ***P***. Formally, given a target pixel (i,j) recorded into a Sentinel-2 image $\mathbf{S} \in \boldsymbol{S}$, this is mapped into a pixel image $\mathbf{P}_{(i,j,\mathbf{S})}$ with shape $W \times H \times C$, which shows pixel (i,j) within its squared pixel neighbourhood in **S**. Let us consider H=W=2k+1 with k as a user-defined, positive, integer parameter to define the neighbourhood size. $\mathbf{P}_{(i,j,\mathbf{S})}$ covers $(2k+1)^2$ pixels of **S** spanned across a pixel neighbourhood N(i,j,k) defined as follows:

$$N(i,j,k) = \{(i',j') \in \mathbf{S} \mid i-k \leq i' \leq i+k,\ j-k \leq j' \leq j+k\}.$$

In the case of boundary pixels, we add padding pixels populated with the average spectral values calculated on the remaining neighbour pixels for each spectral band. Every image $\mathbf{P}_{(i,j,\mathbf{S})} \in \mathbf{P}$ is associated with a binary label (0 or 1) that is the label recorded in **M**(i,j). A CNN with Attention model is trained from ***P*** to learn a function for image classification.

The used CNN integrates a pixel attention mechanism, as described by Zhao et al. (2020). This generates a pixelwise attention map for all pixels in the input data images. In particular, the pixel attention layer takes as input each tensor $\mathbf{X}^l$ of a given convolutional l-th layer with dimension $W^l \times H^l \times C^l$. It convolves a 1×1 convolution layer followed by a sigmoid function. This is to obtain the attention maps of the level l+1 that are then multiplied with the input pixels at l-th layer. Formally, the pixel attention layer is defined as:

$$\mathbf{X}^{l-1} = \sigma\left(\mathrm{d}\left(\mathbf{X}^l\right)\right) \bullet \mathbf{X}^l,$$

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

where d() is a 1×1 convolution layer, σ is the sigmoid function, $\mathbf{X}^l$ and $\mathbf{X}^{l+1}$ are the input tensor at the l-th layer and the resulting output tensor at l+1-th layer. The output of this operation is a new tensor $\mathbf{X}^{l+1}$ with shape $W^{l+1}\times H^{l+1}\times C^{l+1}$ where $W^l=W^{l+1}$, $H^l=H^{l+1}$and $C^l=C^{l+1}$. To produce a single attention map that explains the importance of pixels in the original image, an average layer is finally applied. This layer averages all the channels for each corresponding pixel. More in detail, given the tensor with shape $W^{l+1}\times H^{l+1}\times C^{l+1}$ produced by the pixel attention layer, the average layer produces a single-channel tensor with shape $W^{l+1}\times H^{l+1}\times 1$, for which each attention pixel α_{ij}^{l+1} is equal to:

$$\alpha_{ij}^{l+1} = \frac{1}{C}\sum_{c=1,\dots,C} \alpha_{ijc}^{l},$$

where α_{ijc}^{l} is the attention pixel at position (i,j) in the c-th feature map from the preceding l-1 layer. The average layer explains visually which pixels of the input image received more attention from the model for the image classification.

The prediction is done by minimizing a binary cross-entropy loss function:

$$H = -(y\log\hat{y} + (1-y)\log(1-\hat{y})),$$

where $y \in \{0,1\}$ is the ground truth label and $\hat{y}$ is the CNN output for a single pixel image.

According to the theory reported above, AVALON can use attention values to produce an explanation, named attention scene map, of the semantic segmentation produced for a scene. This map highlights pixels of the scene, which receive attention as part of the bark beetle outbreak. This is done according to Algorithm 1. Specifically, for each pixel (i,j) in **S**, the pixel attention map of the pixel neighbourhood-based image $P_{(i,j,\mathbf{S})}$ sees which pixels in the considered neighbourhood receive more attention from the CNN model to predict the semantic label of (i,j). Based on these premises, all pixel attention maps produced in a scene are combined to yield the overall attention map of the semantic segmentation of **S**. For each pixel in **S**, we sum-up the attention values that are computed for the pixel in every neighbourhood-based image extracted from **S** where the pixel appears as a target pixel or neighbour pixel. Notice that each attention value is a positive, real value. To distinguish between pixels that are seen to predict "healthy" labels and pixels that are seen to predict "damaged" labels, attention values of pixels are multiplied by -1 when they are part of a pixel-neighbourhood image predicted in the class "healthy", while they are considered in their original positive format when they are part of a pixel-based neighbourhood image predicted in the class "damaged". Finally, every accumulated pixel attention value is averaged with respect to the number of single attention values accumulated for the pixel in the scene and scaled between 0 and 1 to be visualised in a heatmap.

*Algorithm 1: Computing the Attention Scene Map **A** of a Sentinel-2 image **S***

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

```
Data: A Sentinel-2 image S with shape Ws × Hs × C; the Attention-CNN model; the neighbourhood size parameter k
Result: The attention map A computed for S with shape Ws × Hs
begin
A= zeros(Ws, Hs)
Counter= zeros(Ws, Hs)
for (i = 1; i ≤ Ws; i + +) do
    for (j = 1; j ≤ Hs; j + +) do
        ŷ(i,j,S) = predict(CNN, P(i,j,S))
        a(i,j,S) = attentionMap(CNN,P(i,j,S))
        if ŷ(i,j,S) == 1 then
            A[i − k : i + k, j − k : j + k]+ = a(i,j,S)
        else
            A[i − k : i + k, j − k : j + k]+ = −1 · a(i,j,S)
        Counter=[i − k : i + k, j − k : j + k]+ = 1
A =minmaxScaler(A/Counter)
return A
```

AVALON was developed in Python 3. The code and parameters of trained models are available at https://github.com/s4rgax/AVALON. The CNN architecture was implemented using the PyTorch library (version 2.0.1). For each dataset, we conducted a Bayesian optimization of the CNN hyperparameters with the tree-structured Parzen estimator algorithm, as implemented in the Optuna library (https://optuna.org/). The implementation of AVALON included the search of the optimozed set-up of the following hyperparameters: mini-batch size in {2^5, 2^6, 2^7, 2^8}, learning rate between 0.0001 and 0.001, number of kernel size in {2, 3, 4}, number of Convolutional layers in {1, 2, 3} and dropout between 0 and 1. The hyperparameter optimisation was performed using 20% of the entire training set (selected with stratified sampling) as a validation set. We selected the configuration of hyperparameters that achieved the highest F1 computed on the validation set by considering the class "damaged" as the positive class. The CNN architecture was defined with a variable number of Convolutional layers, a Pixel Attention layer placed after the last Convolutional layer in the neural network, and two Fully-Connected (FC) layers. A Dropout layer was placed between each pair of Convolutional layers to perform data regularisation and prevent training data overfitting. In all layers, except the final classification layer, the Rectified Linear Unit function (ReLU) was used as the activation function. The SoftMax activation function was used in the classification layer. The gradient-based optimisation was using the Adam update rule. Finally,

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

the CNN model was trained with the maximum number of epochs set equal to 150, and using an early-stopping approach to retain the best model.

EMF

This is a semantic segmentation approach that reuses a foundation deep semantic segmentation model to map bark beetle outbreaks in Sentinel-2 images of forest scenes. The model reuse is one of the main directives of the Data-Centric Artificial Intelligence paradigm that sees the fine-tuning strategy as the engine to power model performance and propel green organisations into the future. In particular, the foundation semantic segmentation model Exchanger+Mask2Former, that was used, was described by Cai et al.(2023). This foundation model was pre-trained from a big volume of Sentinel-2 images by accounting for spatial and temporal dynamics in multi-temporal satellite images. This foundation model was selected according to results illustrated by Cai et al. (2023) showing that it surpasses the previous state-of-the-art results attained on evaluated on the downstream segmentation of the land cover task. Specifically, we considered the foundation Exchanger+Mask2Former model obtained with the semantic segmentation of the land cover semantic labels of the multi-temporal Sentinel-2 dataset PASTIS (https://github.com/VSainteuf/pastis-benchmark). This is a benchmark dataset for semantic segmentation of agricultural parcels from Sentinel-2 time series. It contains 2,433 patches in the French metropolitan territory with semantic labels for each pixel. Each patch is a Sentinel-2 image time series. The dataset, which is 29 GB zipped, contains 2,433 Sentinel-2 time series of scenes with size 128 × 128 at 10 mt pixels, with 38-61 acquisitions per series, and pixelwise labelled in 18 crop types. The pre-trained weights of Exchanger+Mask2Former were used as they were made available by the authors (https://github.com/TotalVariation/Exchanger4SITS). In particular, the Exchanger+Mask2Former architecture of the considered foundation model comprises two components: an Exchanger module that consists of a hierarchical encoder designed to handle multi-temporal satellite images and a Mask2Former module that is used for semantic segmentation. Figure 2 shows the schema of the Exchanger and Mask2Former components of the EMF approach.

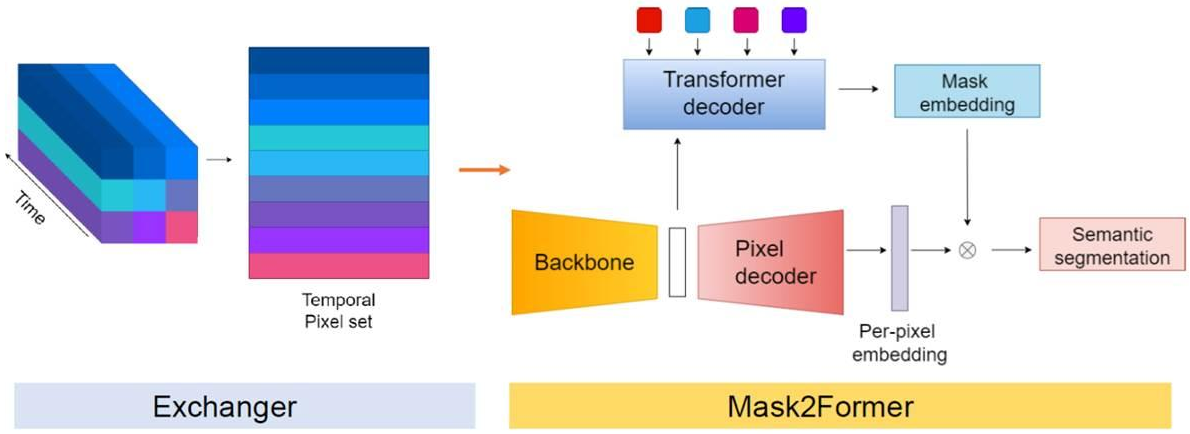


Figure 2: Schema of EMF

The Exchanger module extracts the input feature embedding in a three-stepped process defined *collect-update-distribute*. In the *collect* step, the information is gathered from irregular and asynchronous temporal satellite observations with the creation of a pixel-set to represent

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

each pixel as a set of observations over time. In the *update* step, a query-based transformer decoder processes the pixel-set representation to derive a representation of the spatio-temporal interactions between pixels. Finally, in the *distribute* step, the learned temporal representation of multi-temporal data is propagated to the Mask2Former segmentation architecture (Cheng et al., 2022).

The Mask2Former architecture consists of three components: a *Pixel-level embedding* module that extracts per-pixel embeddings and is used for binary mask predictions; a *Transformer decoder* that computes the low-resolution image features computed into the pixel module and returns the mask embeddings; and a *Segmentation* module, which generates predictions by combining per-pixel embeddings and the mask embeddings. In the Pixel-level embedding module, a backbone network generates low-resolution image feature embeddings that are then used as input to both the transformer encoder and the pixel encoder. The backbone of the Mask2Former is implemented as a Pyramid Vision Transformer (Wang et al., 2021) that generates the low-resolution image feature embeddings by performing a progressive reduction of the size of input data by following a pyramid structure (i.e., the output resolution of the stages progressively shrinks from high to low). This low-resolution image feature is passed to a pixel decoder to generate the per-pixel embeddings, which represent both the local and global contexts of the pixels in the images. In parallel, the low-resolution image feature embeddings are passed to the Transformer decoder that generates a set of mask embeddings, which assign different importance weights to different patches of the image. These mask embeddings define potential segments (i.e., instances of objects) in images that are assigned to the same class-specific instance by performing object class predictions. In the Segmentation module, a binary mask prediction is performed via a dot product between the per-pixel embeddings and the mask embeddings. The dot product is followed by a sigmoid activation and returns a binary mask classification (i.e., the membership of pixels to the object instances). Finally, semantic segmentation predictions are obtained by combining the class prediction with the binary mask prediction using matrix multiplication.

For the downstream task of bark beetle detection, the fine-tuning of the foundation Exchanger+Mask2Former model was performed with the AdamW optimizer. This is a stochastic gradient descent method based on Adam optimizer, which can achieve a better convergence and performance by decoupling weight decay from gradient updates.

GANDALF

This is a semantic segmentation approach that converts Sentinel-2 images of forest scenes into semantic contextual stories and leverages a fine-tuning technique to transfer a foundation LLM model to the downstream task of bark beetle outbreak detection. GANDALF takes a training set that is composed of a collection of Sentinel-2 images of forest scenes and the collection of the masks of the bark beetle outbreaks mapped in the considered scenes. It represents the multispectral data acquired through the Sentinel-2 technology as semantic stories of remotely sensed forest scenes and uses fine-tuning to fit a LLM to these stories, in order to map bark beetle outbreaks. The overall picture of the learning stage of GANDALF approach is shown in Figure 3.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

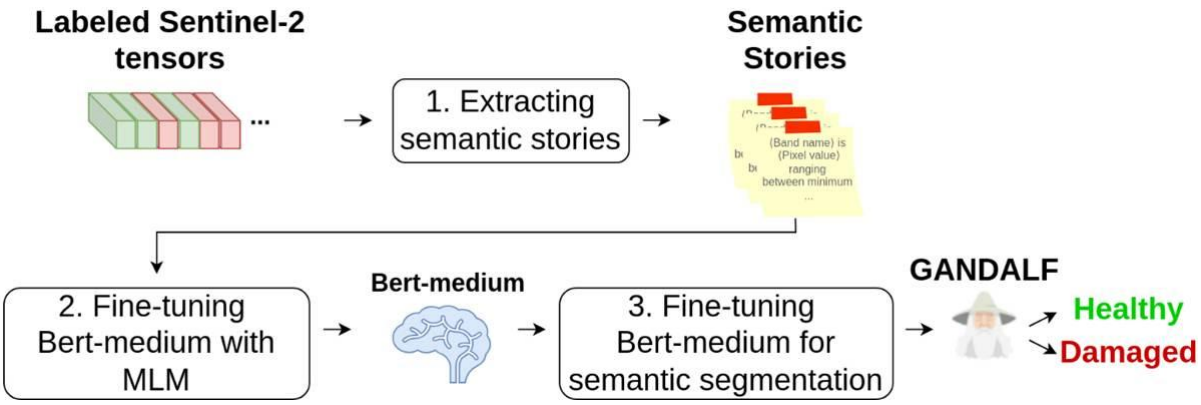


Figure 3: Schema of GANDALF

GANDALF extracts semantic stories from the imagery pixels recorded in the labelled Sentinel-2 tensors of the training imagery collection. It transforms each imagery Sentinel-2 pixel into a semantic story according to a narrative skeleton. This skeleton defines a semantic context of the spectral vector recorded by Sentinel-2 in the given pixel and the spectral-spatial vector obtained aggregating the Sentinel-2 spectral vectors recorded in the pixel neighbourhood. The following narrative skeleton is used to generate the semantic story of a pixel:

<**Band name**> is <**Pixel value**>

ranging between minimum <**Minimum value**> and maximum <**Maximum value**>

with average <**Average value**> median <**Median value**> and deviation <**Deviation value**>

The narrative skeleton reported above includes non-terminal symbols "< >" to denote the name of the Sentinel-2 spectral band, the value of the band measured in the pixel and the values of the band measured in the pixel neighbourhood through the considered aggregate operators. Accordingly, the semantic story of an imagery pixel is the concatenation of the semantic sentences generated with the narrative skeleton reported above and adopted to describe each Sentinel-2 band in the pixel: a semantic sentence is generated for each Sentinel-2 band. Each sentence is obtained from the Sentinel-2 imagery band by replacing each non-terminal symbol of the narrative skeleton with the name of the band, the value of the corresponding band in the imagery pixel and the aggregates measured for the band in the pixel neighbourhood.

GANDALF uses a deep neural network to perform the semantic segmentation of a Sentinel-2 image. This network encloses *Bert-medium* as the initial part of the model and a task-specific head as the final part of the model. Bert-medium (Turc et al., 2019) is a compact pre-trained variant of BERT. The choice of Bert-medium is based on results of the comparative study performed by Turc et al. (2019), which shows that Bert-medium can outperform several LLMs in several tasks (i.e., sentiment classification, natural language inference and textual entailment). However, any other foundation LLM can be evaluated for the scope. Bert-medium includes 8 hidden layers (transformer blocks) and 8 self-attention heads with the hidden embedding size set equal to 512 and the filter size set equal to 2048 for a total of 41.7M parameters. Each transformer block is an encoder block with self-attention layers. In addition, the task-specific head used for the final decision is composed of a dense layer with size equal to 128 and a final layer for the binary classification with the Sigmoid activation function. In GANDALF, the parameters of the foundation Bert-medium model can be fine-tuned on the

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

semantic stories of the Sentinel-2 imagery data based on the gradient computed on the binary cross entropy loss function. However, based on the theory formulated by Howard and Ruder (2018) the fine-tuning of a foundation LLM model may achieve higher accuracy in a decision task by performing a step of domain adaptation before training a task-specific head. Accordingly, we perform a step of domain adaptation to boost the performance of the semantic segmentation task by allowing the Bert-medium model to avoid considering the domain-specific words of the remote sensing corpus as rare tokens. We perform the domain adaptation stage of Bert-medium along the Masked Language Model (MLM) task described by Devlin et al. (2019). This is done by randomly masking some tokens of each semantic story in the training set and adapting the parameters of the foundation Bert-medium to predict the masked tokens. The Bert-medium model with parameters adapted through the domain adaptation step is finally fine-tuned with the task-specific head on the downstream decision task for bark beetle outbreak detection.

Finally, GANDALF uses the Integrated Gradients (IG) technique (Sundararajan et al., 2017) to explain the effect of input dimensions on the LLM decisions. This is a local, XAI technique to compute the gradient value for each input token of the semantic story of a Sentinel-2 pixel assigned to a semantic label. Given a semantic story of a Sentinel-2 pixel, first IG initializes a vector of zeros with one zero for each text token in the semantic story of a pixel. Then, for each token of the pixel semantic story, it computes the gradient of the decision. This gradient value measures the effect of a change of the considered token on the LLM decision. Gradient values are integrated along the network layers from the initial vector to the actual input by resorting to an approximation technique. The higher the integrated gradient value measured for an input token, the more the prediction of the LLM is affected by possible changes in the considered token.

GANDALF was implemented in Python 3.9.18–64 bit version, using Torch 2.1.1 as the back-end. The source code of the proposed approach is available on the GitHub repository (https://github.com/vinspdb/GANDALF). The pre-trained version of Bert-medium used in the evaluation study is available on Hugging Face repository (https://huggingface.co/prajjwal1/bert-medium). For the fine-tuning process, we adopted the AdamW optimizer with learning rate equal to 1e-5 and batch size equal to 1024. The training stage was performed using Accelerate (https://pypi.org/project/accelerate/). The 50\% of semantic stories collected in a training set were used for the domain adaptation stage, while the remaining 50\% of semantic stories were used for the fine-tuning in the downstream task. The percentage of tokens masked during the domain adaptation step was set equal to 15\% of tokens. Finally, the code of GANDALF integrated the LayerIntegratedGradients version of IG that is available in Captum library (https://captum.ai/api/layer.html).

BigEarth-UNet

We defined a UNet architecture that uses the pre-trained ResNet50 deployed in BigEarthNet (Clasen et al., 2024) (https://huggingface.co/BIFOLD-BigEarthNetv2-0) as encoder. The encoder was pretrained for landscape imagery classification considering the Sentinel-2 bands: B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12 and the classes: 'Agro-forestry areas', 'Arable land', 'Beaches, dunes, sands', 'Broad-leaved forest', 'Coastal wetlands', 'Complex cultivation patterns', 'Coniferous forest', 'Industrial or commercial units', 'Inland waters', 'Inland wetlands', 'Land principally occupied by agriculture, with significant areas of natural vegetation', 'Marine waters', 'Mixed forest', 'Moors, heathland and sclerophyllous vegetation', 'Natural grassland

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

and sparsely vegetated areas', 'Pastures', 'Permanent crops', 'Transitional woodland, shrub', 'Urban fabric'. As the ResNet50 was trained on images with size 224×224×10, the padding operation was used to obtain images with size 224×224×10 from scene images with spatial size smaller than 224×224. On the other hand, the tiling operation was performed to obtain images with size 224×224×10 from scene images with spatial size greater than 224×224. The semantic segmentation model was trained by performing the fine-tuning of the encoder branch freezing the encoder layer for 10 epochs and training the decoder starting from random weights. BigEart-UNet was implemented in in Python 3.10–64 bit version, using Torch 2.5.1 as the back-end. The gradient-based optimisation was using the Adam update rule. The Tversky loss was used for the semantic segmentation. The optimization library Optune was used to select batch size between 4 and 64, learning rate between 1e-5 and 1e-2 and Tversky alpha between 0.1 and 0.5 using the validation F1 score (measured on the “damage” class). The model was trained with the maximum number of epochs set equal to 200, and using an early-stopping approach to retain the best model.

Supervised methods evaluation on D1.3 benchmark data

The accuracy performance of AVALON, EMF, GANDALF and BigEarth-UNet was evaluated using the Sentinel-2 imagery dataset created according to the ground truth of bark beetle outbreaks recorded in recorded in the DEFID2 database (Forzieri et al 2022) and observed in September 2020 in Czech Republic. This dataset was already used to evaluate the semantic segmentation models trained for bark beetle outbreak detection described in the intermediary report D1.3.

Dataset. We used the ESA Copernicus open access hub with the Google Earth Engine APIs (to download cloud-free Sentinel-2 images of the considered scenes. For each scene, we selected the Sentinel-2 images acquired in September 2020 which achieved the minimum cloudiness index. This index was computed as the percentage of imagery pixels that the SCL algorithm (Louis et al., 2016) recognizes as noise, defective, dark, cloud, cloud shadow or thin cirrus. If several images achieve the minimum value of the cloudiness index in the considered time interval, then we selected the most recent Sentinel-2 image of this selection. The images were downloaded in the 3857 EPSG system (https://epsg.io/3857). In the obtained Sentinel-2 dataset, the size of images varies from 33×36 to 260×238 pixels at 10 square meters resolution, while the percentage of infested territory per scene varies from 4.14\% to 54.81\% of the scene surface. The total percentage of damaged territory of the entire scene collection is 14.59\%. The semantic segmentation model development and its evaluation were conducted by considering 160 random scenes (covering 1014708 pixels at 10 square meters resolution) as training scene set and 40 left-out scenes (covering 198253 pixels) as testing scene set.

Accuracy metrics. To explore the accuracy of the developed models we considered the Fscore (F) of Precision and Recall computed for the two opposite classes (“damage” and “healthy”), the MacroF score computed as at the average of each F score computed per class (D--“damage” or H—“healthy”) and Intersection over Union (IoU) computed as ratio of the intersected area to the combined area of prediction.

Accuracy Results. Table 1 reports the accuracy metrics measured for AVALON, EMF, GANDALF and BigEarth-UNet on the testing set of the Czech Republic dataset. The model development and evaluation of AVALON, GANDALF and BigEarth-UNet were performed by

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

considering Sentinel-2 images collected in September 2020. The model development and evaluation of EMF were performed by considering the time series of Sentinel-2 images collected every 15 days from April 2020 to September 2020. The best accuracy is achieved with AMF. However, this gain in accuracy is at the cost of a higher amount of data to be downloaded and prepared for both the model development and evaluation. Notably, the RF model still achieves a good trade-off between accuracy and simplicity. Finally, TITANIA and BigEarth-UNet achieve similar performance, although BigEarth-UNet reuses a foundation ResNet encoder.

	F(D)	F(H)	macro F	IoU
AVALON	70.73	93.64	82.19	54.72
EMF	**77.31**	**95.44**	**86.37**	**63.01**
GANDALF	68.63	93.61	81.12	52.24
BigEarth-UNet (fine-tuning)	70.04	93.32	81.68	53.90
RF	70.18	93.31	81.75	54.06
XGB	69.02	92.75	80.88	52.69
SVM	71.15	93.07	82.11	55.23
UNET	67.58	90.78	79.18	51.03
TITANIA (UNET+ATTENTION+SELF-DISTILLATION	70.50	92.92	81.71	54.43
DIAMANTE (MULTI-SENSOR UNET)	69.21	92.21	80.71	52.12

Table 1: Accuracy metrics measured for AVALON, EMF, GANDALF and BigEarth-UNet and the related methods: RF, XGB, SVM, UNET, TITANIA and DIAMANTE. Metrics are measured on the testing set of the Czech Republic dataset (ground truth obtained in September 2020). The related methods are described in the intermediary deliverable. The best results are in bold.

Explainability Results

Both AVALON and GANDALF integrate some explainable mechanisms.

Regarding AVALON, Figure 4 shows the RGB bands, ground truth label masks, predicted semantic segmentation masks and scene attention maps of two Sentinel-2 images selected from the testing set of the Czech Republic imagery dataset. The two scenes were selected as scenes where AVALON achieved high F1(D) (i.e., F1(D)=0.71 in Figures 4a-4d) and low F1(D) (i.e., F1(D) =0.04 in Figures 4e-4h), respectively.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

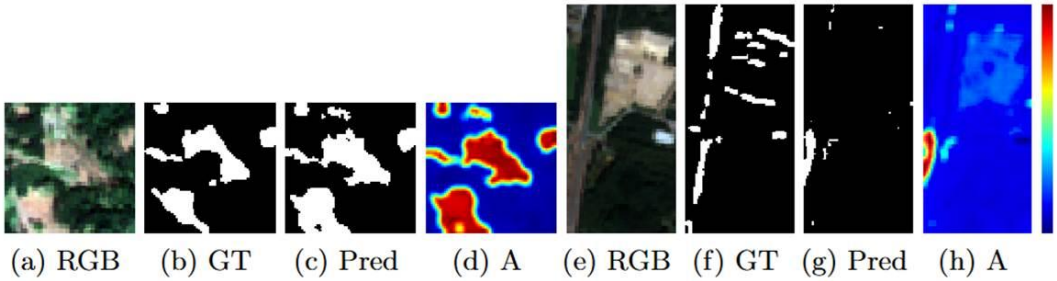


Figure 4: AVALON explanations: RGB images, Ground truth label masks (GT), Predicted semantic segmentation masks (Pred) and Scene Attention maps (A) of two scenes in Czech Republic. Figures a- d refer to a scene for which AVALON achieved F(D)=0.71. Figures e-h refer to a scene for which AVALON achieved F(D)=0.71. Figures e-h refer to a scene where AVALON achieved F(D)=0.04.

The scene attention map plotted in Figure 4.d shows that the zones that received higher attention delimit well the patches where AVALON predicted the presence of a bark beetle outbreaks in Figure 4.c. So, this type of explanation outcomes can be used as a visual explanation for foresters of the certainty of the predictions produced for these scenes. On the other hand, the scene attention map plotted in Figure 4.h shows signals of low certainty for predictions produced from AVALON in this scene and plotted in Figure 4.g. We recall that AVALON achieved low accuracy performance in this scene, so this scene explanation outcome can be seen as an alert for foresters who may decide to integrate inventory results of AVALON in critical zones with the inventory produced with traditional fieldwork. In particular, the predicted mask plotted in Figure 4.g shows that AVALON fails in delimiting outbreak zones in the centre-up of the scene. However, the attention map of the scene reported in Figure 4.h shows some uncertainty in this area. In fact, attention values take on a light shade of blue in the missed outbreak patches (wrongly detected as "healthy"), which contrast with the dark shade of blue we observe in the attention value of patches correctly labelled as "healthy". Further conclusions can be drawn from the scene attention map shown in Figure 4.d in correspondence with the predicted mask shown in Figure 4.c. In this case, AVALON wrongly labels a patch in the centre-up of the scene as "damaged", and the attention values of the patch show that AVALON pays attention to the zone. However, the red extension of the attention (where, anyway, we note a prevalence of yellow and green) is limited compared to other patches correctly predicted as "damaged" in the same scene.

Band	Semantic story
B1	coastal aerosol is 370 ranging between minimum 346 and maximum 440 with average 375 , median 370 and deviation 31 .
B2	blue is 541 ranging between minimum 499 and maximum 618 with average 544 , median 549 and deviation 23 .
B3	green is 736 ranging between minimum 653 and maximum 775 with average 716 , median 715 and deviation 32 .
B4	red is 1004 ranging between minimum 930 and maximum 1116 with average 1005 , median 998 and deviation 63 .
B5	red - edge - 1 is 1414 ranging between minimum 1140 and maximum 1414 with average 1308 , median 1284 and deviation 98 .
B6	red - edge - 2 is 1878 ranging between minimum 1571 and maximum 2156 with average 1827 , median 1878 and deviation 178 .
B7	red - edge - 3 is 2234 ranging between minimum 1854 and maximum 2588 with average 2170 , median 2234 and deviation 219 .
B8	nir is 2630 ranging between minimum 2114 and maximum 2640 with average 2434 , median 2486 and deviation 159 .
B8A	narrow nir is 2669 ranging between minimum 2307 and maximum 2920 with average 2589 , median 2669 and deviation 197 .
B10	water vapor is 2678 ranging between minimum 2280 and maximum 2678 with average 2563 , median 2554 and deviation 141 .
B11	swir - 1 is 3186 ranging between minimum 2649 and maximum 3186 with average 2937 , median 2877 and deviation 213 .
B12	swir - 2 is 2091 ranging between minimum 1624 and maximum 2091 with average 1907 , median 1890 and deviation 153 .

Legend: Negative Neutral Positive

Figure 5: AVALON explanations: IG explanation produced for the semantic story of a “damaged" pixel of a testing scene of the Czech Republic

Regarding GANDALF, Figure 5 shows the semantic story of a “damaged” pixel of a testing scene of the Czech Republic case study, where the IG scores computed for the separate

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

tokens of the story have been highlighted according to their magnitude. Positive scores are in green, neutral scores are in white and negative scores are in red. The plot shows that the LLM model of GANDALF sees the spectral and spectral-spatial values of the B2 (Blue) band, the spectral value of B5 (Red edge 1) and the minimum value of B12 (SWIR 2) as the most positively relevant for the decision.

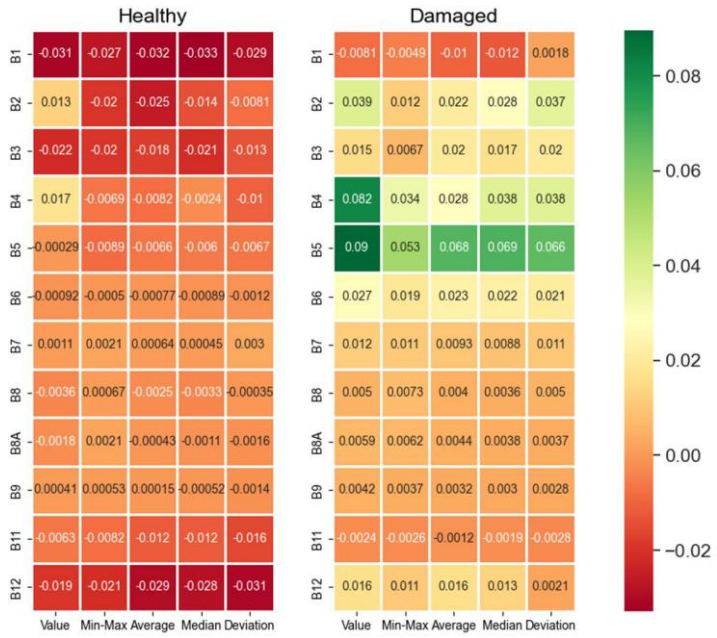


Figure 6: AVALON explanations: Average IG scores per token group (axis X) and spectral band (axis Y) -- "healthy" vs "damaged"

We grouped single tokens of each band-based sentence of the semantic story of a pixel in "token groups" defined as follows:

Group "Value": ⟨Band name⟩ is ⟨Pixel value⟩

Group "Min-Max": ⟨Band name⟩ is ranging between minimum ⟨Minimum value⟩ and maximum ⟨Maximum value⟩

Group "Average": ⟨Band name⟩ is with average ⟨Average value⟩

Group "Median": ⟨Band name⟩ is with median ⟨Median value⟩

Group "Deviation": ⟨Band name⟩ is with deviation ⟨Deviation

value⟩.

Hence, for each band-based sentence of a semantic story of a pixel, for each one of the token groups listed above, we computed the average of the IG scores associated with the tokens of the considered group in the selected sentence. Figure 6 shows the maps of the average IG score computed per spectral band (axis Y) and token group (axis X) for pixels of classes "healthy" and "damaged", respectively. Specifically, the IG scores were computed for the semantic stories of pixels of testing scenes and averaged on the classes: "healthy" and "damaged", separately. In this way, the two maps provide a global explanation of the decision process showing which band-based token groups the LLM of GANDALF sees as the most

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

positively relevant for the decisions regarding the two classes. In particular, the IG results of the considered case study show that the group of tokens (Value) to describe the spectral values of B4 (Red) and B2 (Blue) contain the most relevant information for decisions regarding “healthy” patches, while the ones to describe the spectral values of B4 (Red) and B5 (Red edge 1) contain the most relevant information for decisions regarding “damaged” patches. Notably, only some of the groups of tokens used to describe spectral-spatial values of bands B7 (Red edge 3), B8 (NIR), B8A (NIR-A) and B9 (Water vapor) – specifically, Min-Max, Average, Median and Standard Deviation for B7, Min-Max for B8 and B8A, and Min-Max and Average for B9 – have a positive effect on decisions of “healthy” patches. On the other hand, most of the tokens that describe the spectral and spectral-spatial information of the “damaged” pixels have a positive effect on decisions regarding these pixels. Finally, the B1 (Coastal aerosol) and B11 (SWIR 1) have a negative effect on decisions of both classes. In general, these explanation insights of the decision process of GANDALF highlight that the number of tokens having a positive effect on decisions regarding the “damaged” class is significantly greater than the number of tokens having a positive effect on decisions regarding the “healthy” class. This shows that the decision process concerning the forest damage caused by the bark beetle outbreak is, in general, more complex (in terms of information that drives the decision of the LLM model) than the decision process to delineate the healthy patches. Finally, we note that the B5 information enclosed in all groups of tokens considered in this study is among the most positive relevant for the decisions of the LLM on pixels in the class “damaged”. This supports one of the major conclusions drawn in the seminal study of Abdullah et al. (2019) that the Red-Edge information is one the most sensitive Sentinel-2 band to insect attack

Sentinel-2 image dataset preparation update

Sentinel-2 imagery data used for model development and evaluation are all collected from GEE that provides the API to access to Sentinel-2 images from 2017 to present. The image collection was accessed via Python API calls and downloaded in the 3857 EPSG system[1] . As each query was conducted with a time interval as a selection criterion, it may return a result set of images. The pipeline, initially adopted by team UNIBA and described in the intermediary deliverable D1.3, used the operator BEST to identify, for each scene, the best image that, in the query time, minimizes the cloudiness index computed in the scene according to the output of the Scene Classification Level (SCL) algorithm (Louis et al., 2016). In alternative, team UNIBA defined a revised version of this pipeline that uses the MEDIAN operator in place of the BEST operator. The MEDIAN operator was used to mitigate the risk of selecting cloudy and noisy Sentinel-2 images being less sensitive to outliers. Specifically, the MEDIAN was computed for each band, pixelwise, by ignoring values associated with spectral bands acquired for pixels recognized as noise, defective, dark, cloud, cloud shadow or thin cirrus by the SCL algorithm.

We performed a comparative analysis to evaluate the performance of the updated pipeline synthesized to prepare a Sentinel-2 imagery dataset. This analysis was conducted in the Czech Republic case study, that was already considered for evaluating several supervised AI-based methods developed by UNIBA for bark beetle outbreak detection by processing Sentinel-2 images prepared using the operator BEST. To this aim, we considered the models developed with the machine learning method RF and the deep learning methods EMF and BigEarth-UNet using Sentinel-2 images obtained with BEST, as well as Sentinel-2 images obtained with MEDIAN. The RF was trained considering the Spectral Vegetation Indexes: NGDRI, NMDI, and MCARI identified by Andresini et al. (2023b), and the Spectral-Spatial (SS)

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

features described by Andresini et al (2024b) as described in the intermediary deliverableD1.3. The model development and evaluation phases were conducted by considering 160 random scenes (covering 1014708 pixels at 10 square meters resolution) as training scene set and 40 left-out scenes (covering 198253 pixels) as testing scene set.

Sentinel-2 images were downloaded by considering the query time with duration equal to one month (configuration M) or two weeks (configuration WW) by using either non-overlapping windows (configuration NO) or overlapping windows configuration (O). Table 2 describes the query time intervals considered in the tested configurations.

Duration	Overlap	Time query
M	**NO**	2020-09-01 - 2020-09-30,2020-08-01 - 2020-08-31, 2020-07-01 - 2020-07-31,2020-06-01 - 2020-06-30, 2020-05-01 - 2020-05-31,2020-05-01 - 2020-05-31
M	**O**	2020-09-01 - 2020-09-30,2020-08-15 - 2020-09-14, 2020-08-01 - 2020-08-31,2020-07-15 - 2020-08-14, 2020-07-01 - 2020-07-31,2020-06-15 - 2020-07-14, 2020-06-01 - 2020-06-30,2020-05-15 - 2020-06-14 ,2020-05-01 - 2020-05-31,2020-04-15 - 2020-05-14, 2020-04-01 - 2020-04-30
WW	**NO**	2020-09-15 - 2020-09-30,2020-09-01 - 2020-09-14, 2020-08-15 - 2020-08-31,2020-08-01 - 2020-08-14, 2020-07-15 - 2020-07-31,2020-07-01 - 2020-07-14, 2020-06-15 - 2020-06-30,2020-06-01 - 2020-06-14, 2020-05-15 - 2020-05-31,2020-04-01 - 2020-05-14, 2020-04-15 - 2020-04-30

Table 2: Query time intervals

Table 3 reports the accuracy performance metrics that were measured for the RF models trained with depth=20 and w=5 and evaluated using the Sentinel-2 images of Czech Republic collected using the operator MEDIAN and the operator BEST, respectively. The RF models were trained and evaluated with single-temporal Sentinel-2 images (i.e. single images acquired at the same time). The analysis is performed with models developed and evaluated at different time periods by considering Sentinel-2 images collected, monthly or bi-weekly, between May 2020 and September 2020. Notably, the RF model gains accuracy as it was trained and evaluated considering the Sentinel-2 images acquired in the late summer and early autumn of 2020. This trend is equally observed independently of the use of the operators MEDIAN or BEST to prepare the Sentinel-2 imagery dataset. However, we also note that the accuracy performance of the RF model is commonly higher when the Senetinel-2 images were prepared using the operator MEDIAN especially when the input images were collected over a query time horizon of 30-31 days (configuration M). In addition, we note that the configuration enabling the window overlap mode (configuration O) allows us to obtain a RF model that can be trained and evaluated every two weeks leveraging Sentinel-2 images retrieved with a query spanned on the time interval of 30-31 days.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

configuration		MEDIAN				BEST			
		F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-09-01 - 20-09-30	**M + NO**	**72.04**	**93.54**	**82.79**	**56.30**	70.17	93.31	81.74	54.05
20-08-01 - 20-08-31		**58.24**	**90.66**	**74.45**	**41.08**	48.81	87.48	68.15	32.30
20-07-01 - 20-07-31		**45.19**	86.35	**65.77**	**29.19**	43.37	**87.01**	65.19	27.68
20-06-01 - 20-06-30		**36.98**	84.94	**60.96**	**22.68**	31.07	**87.05**	59.06	18.39
20-05-01 - 20-05-31		**37.94**	84.84	**61.39**	**23.41**	33.22	**86.44**	59.83	19.92
20-09-01 - 20-09-30	**M + O**	**72.04**	**93.54**	**82.79**	**56.30**	70.17	93.31	81.74	54.05
20-08-15 - 20-09-14		**67.84**	**92.70**	**80.27**	**51.34**	56.82	90.94	73.88	39.68
20-08-01 - 20-08-31		**58.24**	**90.66**	**74.45**	**41.08**	48.83	87.48	68.15	32.30
20-07-15 - 20-08-14		**51.30**	**89.22**	**70.26**	**34.50**	46.82	86.32	66.57	30.57
20-07-01 - 20-07-31		**45.19**	86.35	**65.77**	**29.19**	43.37	**87.01**	65.19	27.68
20-06-15 - 20-07-14		**33.93**	86.92	**60.42**	**20.43**	25.74	**87.01**	56.38	14.77
20-06-01 - 20-06-30		**36.98**	84.94	**60.96**	**22.68**	31.07	**87.05**	59.06	18.39
20-05-15 - 20-06-14		38.40	84.86	61.63	23.76	**40.98**	**86.72**	**63.55**	**25.77**
20-05-01 - 20-05-31		**37.94**	84.84	**61.39**	**23.41**	33.22	**86.44**	59.83	19.92
20-09-15 - 20-09-30	**WW + NO**	71.26	93.24	82.25	55.36	**71.55**	**93.45**	**82.50**	**55.70**
20-09-01 - 20-09-14		**69.33**	93.04	**81.18**	**53.05**	68.54	**93.08**	80.81	52.14
20-08-15 - 20-08-31		**58.48**	**91.01**	**74.74**	**41.32**	56.68	90.89	73.79	39.55
20-08-01 - 20-08-14		**52.97**	**88.90**	**70.94**	**36.03**	48.34	87.82	68.08	31.87
20-07-15 - 20-07-31		45.13	**86.60**	65.87	29.14	**45.91**	86.42	**66.16**	**29.79**
20-07-01 - 20-07-14		**44.06**	86.45	65.25	**28.25**	43.76	**86.97**	**65.37**	28.01
20-06-15 - 20-06-30		15.01	87.96	51.49	8.11	**15.79**	**88.11**	**51.95**	**8.57**
20-06-01 - 20-06-14		**38.26**	85.40	**61.83**	**23.66**	35.50	**87.80**	61.65	21.58
20-05-15 - 20-05-31		**38.36**	85.35	**61.86**	**23.73**	36.13	**86.48**	61.30	22.05

Table 3: MEDIAN vs BEST: Random Forest. The best results are in bold.

Table 4 reports the accuracy performance metrics that were measured for the EMF models, trained and evaluated with the multi-temporal Sentinel-2 images acquired from April 2020 to the timestamp considered for the model development and evaluation in the Czech Republic case study. Similarly to the RF model analysis, the EMF model analysis was performed with EMF models developed and evaluated at different time periods. The EMF model outperforms

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

the RF model in the late summer and early autumn, while the RF model outperforms the EMF model in the late spring and early summer. Focusing the attention on the EMF performance, few configurations with the operator MEDIAN outperform the counterpart configurations with the operator BEST. In general, we obtain better performance with BEST. Focusing on the configurations obtained with the operator BEST, we note that the configurations providing a new image of a scene every 15 days (M+O and WW+NO) outperform the configurations providing a new image of a scene every 30-31 days (M+NO). However, considering the configurations M+O and WW+NO we note that the Sentinel-2 images extracted with a query spanned on a time horizon of 30-31 days (M) allows us to gain accuracy in the EMF models developed and evaluated in the late summer, while the Sentinel-2 images extracted with a query spanned on a time horizon of 15 days (WW) allows us to gain accuracy in the EMF models developed and evaluated in the late spring and early summer.

configuration		MEDIAN				BEST			
		F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-04-01 - 20-09-30	**M**	**75.12**	**95.18**	**85.15**	**60.15**	74.94	95.02	84.98	59.92
20-04-01 - 20-08-31	**+**	**58.96**	**92.92**	**75.94**	**41.81**	49.70	91.23	70.47	33.07
20-04-01 - 20-07-31		33.84	89.33	61.59	20.37	**37.16**	**89.55**	**63.35**	**22.82**
20-04-01 - 20-06-30	**NO**	23.06	88.72	55.89	13.03	**27.38**	**89.08**	**58.23**	**15.86**
20-04-01 - 20-05-31		23.67	**88.82**	56.25	13.43	**25.92**	88.39	**57.15**	**14.89**
20-04-01 - 20-09-30		75.13	95.19	85.16	60.17	**77.31**	**95.44**	**86.38**	**63.02**
20-04-15 - 20-09-14		71.74	94.65	83.19	55.93	**74.39**	**95.02**	**84.71**	**59.22**
20-04-01 - 20-08-31	**M**	58.67	**92.95**	75.81	41.51	**60.64**	92.70	**76.67**	**43.51**
20-04-15 - 20-08-14	**+**	44.78	**91.46**	68.12	28.85	**49.29**	91.06	**70.17**	**32.70**
20-04-01 - 20-07-31	**O**	34.83	**89.80**	62.32	21.09	**42.39**	89.47	**65.93**	**26.90**
20-04-15 - 20-07-14		28.25	88.61	58.43	16.45	**32.94**	**89.57**	**61.26**	**19.72**
20-04-01 - 20-06-30		26.27	88.61	57.44	15.12	**30.45**	**88.80**	**59.63**	**17.96**
20-04-15 - 20-06-14		25.52	88.23	56.88	14.63	**27.06**	**88.94**	**58.00**	**15.64**
20-04-01 - 20-05-31		21.62	**88.83**	55.23	12.12	**22.50**	88.73	**55.62**	**12.68**
20-04-01 - 20-09-30		**76.81**	**95.36**	**86.09**	**62.36**	75.29	95.11	85.20	60.37
20-04-01 - 20-09-14		**75.16**	**95.17**	**85.17**	**60.21**	61.51	93.12	77.31	44.41
20-04-01 - 20-08-31		**63.68**	**93.25**	**78.46**	**46.71**	61.50	93.10	77.30	44.40
20-04-01 - 20-08-14	**WW**	48.35	**91.72**	70.04	31.88	**49.70**	91.05	**70.37**	**33.07**
20-04-01 - 20-07-31	**+**	38.28	89.75	64.01	23.67	**41.37**	**90.48**	**65.92**	**26.08**

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

20-04-01 - 20-07-14	**NO**	30.13	89.13	59.63	17.74	**35.15**	**89.67**	**62.41**	**21.32**
20-04-15 - 20-06-30		24.79	88.83	56.81	14.15	**29.37**	**88.93**	**59.15**	**17.21**
20-04-01 - 20-06-14		25.00	88.71	56.85	14.28	**30.65**	**87.87**	**59.26**	**18.10**
20-04-01 - 20-05-31		22.15	88.24	55.20	12.45	**32.02**	**87.88**	**59.95**	**19.06**

Table 4: MEDIAN vs BEST: EMF

Finally, considering the configuration M+NO, Table 5 reports the accuracy performance metrics that were measured for the BigEarth-UNet model and evaluated using the Sentinel-2 images of Czech Republic collected using the operator MEDIAN and the operator BEST, respectively. The BigEarth-UNet models were trained and evaluated with single-temporal Sentinel-2 images (i.e. single images acquired at the same time) using the traditional fine-tuning strategy. The analysis was performed with models developed and evaluated in the configuration M+NO by considering the Sentinel-2 images collected monthly between May 2020 and September 2020. The BigEarth-UNet models were trained performing a step of image augmentation during the training stage. The deep neural models trained with the Sentinel-2 images prepared with the MEDIAN operator commonly outperform the counterpart models trained with the Sentinel-2 images prepared with the BEST operator.

configuration		**MEDIAN**				**BEST**			
		F(D)	**F(H)**	**macro F**	**IoU**	**F(D)**	**F(H)**	**macro F**	**IoU**
20-09-01 - 20-09-30	**M**	**70.04**	**93.32**	**81.68**	**53.90**	67.99	93.56	80.78	51.51
20-08-01 - 20-08-31	**+**	**54.14**	**91.16**	**72.65**	**37.12**	45.78	89.80	67.79	29.68
20-07-01 - 20-07-31		**38.61**	**89.04**	**63.83**	**23.93**	37.84	88.78	63.31	23.34
20-06-01 - 20-06-30	**NO**	**33.64**	**87.93**	**60.79**	**20.22**	39.18	87.44	63.31	24.36
20-05-01 - 20-05-31		**33.41**	86.45	**59.93**	**20.06**	30.02	**87.78**	58.90	17.66

Table 5: MEDIAN vs BEST: BigEarth-UNet with fine-tuning (configuration M+O)

Unsupervised Change Vector Analysis

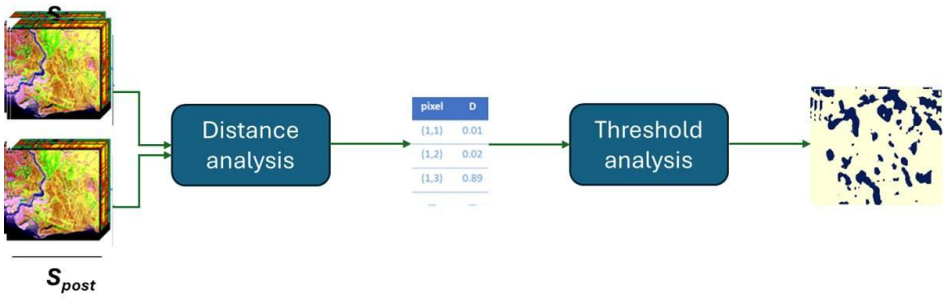


Figure 7: CVA methodThe CVA method

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

(Figure 7) works in an unsupervised manner, i.e. without a training stage supervised with ground truth labels that delineate annotated bark beetle outbreaks. For each scene, it takes **biS** -- a collection of bi-temporal Sentinel-2 images -- as input. Each Sentinel-2 image $(\mathbf{S}_{pre},\mathbf{S}_{post})$Î**biS** is a pair of tensors with shape W_S ´ H_S ´ C such that $\mathbf{S}_{pre}$ was acquired at the timestamp $\mathbf{t}_{pre}$ and $\mathbf{S}_{post}$ was acquired at the timestamp $\mathbf{t}_{post}$ with $\mathbf{t}_{post} > \mathbf{t}_{pre}$ and $\mathbf{t}_{post}$ denoting the timestamp at which the outbreak must be identified. For each pixel (i,j) with 1£ i £W_S and 1£ j £H_S the CVA method computes the distance between the Sentinel-2 spectral vectors $\mathbf{S}_{pre}$(i, j) and $\mathbf{S}_{post}$(i, j). In the remote sensing literature, spectral distances are commonly measured by resorting to either the Euclidean distance or the Spectral Angle (SAM) distance. The Euclidean distance is computed according to following formula:

$$Euclidean(\mathbf{S}_{pre}(i,j),\mathbf{S}_{post}(i,j)) = \sqrt{\sum_{k=1,\dots,C}(\mathbf{S}_{post}(i,j,k)-\mathbf{S}_{pre}(i,j,k))^2}$$

The SAM distance is computed according to the following formula:

$$SAM\left(S_{pre}(i,j),S_{post}(i,j)\right) = \arccos\left(S_{pre}(i,j)-S_{post}(i,j)\right) / \left(\left|\left|S_{pre}(i,j)\right|\right| \cdot \left|\left|S_{post}(i,j)\right|\right|\right)$$

The min-max scaling may be used to scale spectral bands between 0 and 1 before applying the distance analysis.

Let $\boldsymbol{D_S}$ the scene distance matrix with shape $W_S \times H_S$ such that:

$$\boldsymbol{D_S}(i,j)=distance(\mathbf{S}_{pre}(i,j), \mathbf{S}_{post}(i,j))$$

with *distance()* computed according to the Euclidean distance or the SAM distance formulas. Each scene distance matrix is binarized by performing a threshold analysis. The threshold analysis is done by using either the Otsu's algorithm or performing the clustering analysis.

The Otsu's algorithm allows us to automatically determine the upper threshold of SAM distances for separating pixels of the study scenes into background (``unchanged" pixels with low distance range) and foreground ("changed" pixels with high distance range). In particular, a pixel *(i, j)* with *distance*($\mathbf{S}_{pre}$(i,j), $\mathbf{S}_{post}$ (i,j)) higher than *otsuvalue* is assigned to the label "changed" (and, hence, "damaged"), otherwise it is assigned to the label "unchanged" (and, hence, "healthy").

The Otsu's threshold is determined by minimising the intra-class intensity variance defined as a weighted sum of variances of the two classes. We note that minimizing the intra-class variance is equivalent to maximizing the inter-class variance, since the total variance (the sum of the intra-class variance and the inter-class variance) is constant for different partitions. Let us assume that

the distances, computed pixelwise in the study scenes, are represented in an histogram with *L* equal-width bins (levels) denoted as [1,…, *L*]. Let n_i be the number of pixels at level *i*, so that that $\Sigma_{i=1,..,L} n_i$ corresponds to the total number of analysed pixels. Based upon these premises, the probability of each level *i* is computed as $p_i = n_i$ /nr.pixels. The Otsu's algorithm identifies the optimal threshold level *otsuvalue*, in order to divide the pixels of the processed scene into the background class "healthy", spanned over the distance levels [1,2, …, *otsuvalue*], and the

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

foreground class “damaged”, spanned over the distance levels] *outvalue*, … , *L*], respectively. Mathematically, the optimal *otsuvalue* is searched for minimizing the intra-class variance that is defined as a weighted sum of variances of the two classes:

$$otsuvalue = \underset{1 \le value \le L}{\operatorname{argmin}} \left(w_1(value)\sigma_1^2(value) + w_2(value)\sigma_2^2(value)\right)$$

where here σ^2_1(value) and σ^2_2(value) are the variance computed on the two classes separated by each candidate *value*. The weights w_1(value) and w_2(value) are the probabilities of the two classes, which are computed as follows:

$$w_1(value) = \sum_{i=1,\dots,value} p_i \quad \text{and} \quad w_2(value) = \sum_{i=value+1,\dots,L} p_i$$

In alternative, the clustering analysis can be used to binarize the distance matrices obtained for the processed scenes. A popular clustering algorithm based on the Gaussian Mixture Model (GMM) is adopted for the clustering analysis of distance information systems (Reynolds, 2009). Gaussian models are already used as pivotal components of various HS image analysis (Appice et al., 2020). They represent the probability density of the data with a weighted summation of a finite number of Gaussian densities, with different means and standard deviations (or covariance matrices in the case of multivariate GMM). In clustering, Gaussian models allow us to group data into a finite number of Gaussian clusters by modelling cluster conditional probability--maximum likelihood. In particular, the GMM algorithm estimates clustering parameters from training data using the iterative Expectation-Maximization (EM) algorithm. So, based upon the estimated parameters, for each pixel (i, j), the GMM algorithm determines the likelihood of (i, j) belonging to clusters C_1 and C_2, respectively. Pixel (i, j) is assigned to the cluster maximizing the likelihood. As clusters C_1 and C_2 are completely formed (each pixel of the scene has been assigned to one cluster), we label pixels according to the centre of the cluster they are assigned to. The cluster associated with the higher-valued centre is labelled as “damaged”, while the cluster associated with the lower-valued centre is labelled as "healthy”.

Th CVA method was implemented in Python by integrating the implementation of the Otsu’s algorithm available in scikit-image (https://scikit-image.org/docs/dev/auto_examples/segmentation/plot_thresholding.html) and the implantation of GMM available in scikit-learn (https://scikit-learn.org/1.5/modules/generated/sklearn.mixture.GaussianMixture.html) with covariance type set equal to “tied”.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

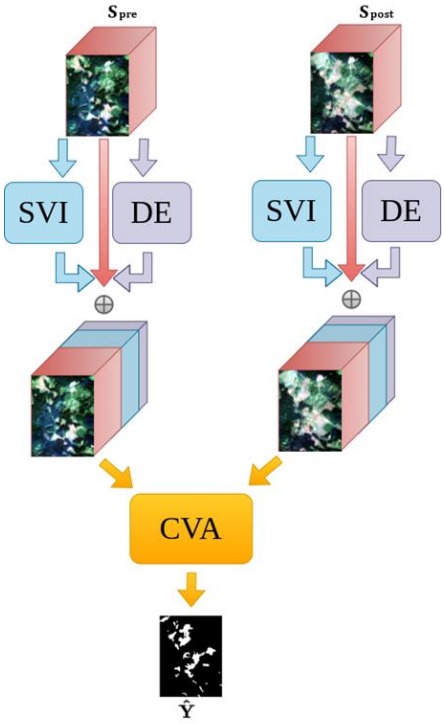


Figure 8: CVA -NN method

As a variant of the classical CVA, CVA-NN (see Figure 8) uses a deep neural network to obtaina Deep Embedding representation of spectral signatures used for the distance computation. In particular, the CVA-NN method uses a UNet pre-trained for the semantic segmentation of forest scenes labelled with forest type ground truth ("forest" vs "non-forest"). The UNet was implemented in Python 3 using the high-level neural network API Keras 2.11 integrated into TensorFlow. Specifically, it was implemented to process scenes tiles with size 32 × 32 × 12. The UNet architecture included an encoder repeating blocks with a 3×3 convolutional layer with a spatial dropout, a Batch Normalization layer and a 2×2 Max Pooling operation layer. At each downsampling step of the encoder, the number of feature channels was increased. In the decoder, the feature map was upsampled through a sequence of up-convolutions that halved the number of feature channels and concatenated the result with the features from the corresponding encoder layer. The classification of each pixel map was obtained by using the Sigmoid activation function, while the ReLu was used in all hidden convolution layers. The Tversky loss was used. An imagery tile augmentation strategy was considered, to improve the performance of the model. The tiler library (https://github.com/the-lay/tiler) was adopted for the imagery tile extraction, while the augmentation strategy was implemented by using the Albumentations library (https://albumentations.ai/). Specifically, the number of training imagery tiles was quintupled by creating new tiles with traditional computer vision transformation operators (i.e., Horizontal Flip, Vertical Flip, Random Rotate, Transpose and Grid Distortion). The hyperparameters of the final UNet model were optimised using the tree-structured Parzen estimator algorithm, implemented in the Hyperopt library to explore the mini-batch size in {4,8,16,32,64}, learning rate in [0.0001,0.01] and the augmentation in {True, False}. The optimization was performed using a random stratified split of 20% of the entire

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

training imagery tile set as a validation set. The hyperparameter configuration that achieved the highest F1 score on the class 1 ("forest") on the validation set, was automatically selected.

For each pixel (i,j) for which the CVA must be performed, the distance is computed as the Euclidean distance between the output values of the sigmoid layer obtained for $\mathbf{S}_{pre}(i,j)$ and $\mathbf{S}_{post}(i,j)$, respectively, or the SAM distance between the spectral bands, the output of the sigmoid layer and eventually the SVIs obtained for $\mathbf{S}_{pre}(i,j)$ and $\mathbf{S}_{post}(i,j)$, respectively.

Unsupervised method evaluation on D1.3 benchmark data

The performance of the CVA pipeline was evaluated in the Czech Republic case study already considered in the intermediary deliverable D1.3. As the CVA method is unsupervised, no split was done to partition scenes into a training set, used for model development, and a testing set, used for model evaluation. The evaluation was performed by considering Sentinel-2 images collected monthly from June 2020 to May 2020. Table 6 reports the accuracy metrics collected in all Sentinel-2 images of the Czech Republic case study prepared by using the BEST operator (as in D1.3). The CVA was conducted by performing the distance analysis using either the Euclidean distance or the SAM distance and the binarization analysis using either the Otsu's algorithm or the GMM. The min-max scaler was used in combination with the Euclidean distance. The results show that the best performance is achieved using the SAM distance and the Otsu's algorithm.

configuration		Euclidean				SAM			
		F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-09-01 - 20-09-30	**Otsu**	51.06	87.99	**69.52**	34.28	**52.19**	86.33	69.26	**35.31**
20-08-01 - 20-08-31		06.16	**89.84**	48.00	3.18	**29.43**	85.19	**57.31**	**17.25**
20-07-01 - 20-07-31		8.87	**90.28**	49.58	4.64	**21.35**	85.09	**53.22**	**11.95**
20-06-01 - 20-06-30		17.08	**86.51**	51.80	9.34	**19.07**	82.17	**50.62**	**10.54**
20-09-01 - 20-09-30	**GMM**	1.40	**91.73**	46.56	0.70	**52.61**	87.17	69.89	**35.69**
20-08-01 - 20-08-31		6.16	**89.85**	48.00	3.17	**27.38**	86.07	56.73	**15.86**
20-07-01 - 20-07-31		8.87	**90.28**	49.58	4.64	**19.80**	85.81	52.81	**10.99**
20-06-01 - 20-06-30		17.08	**86.51**	51.80	9.34	**18.50**	82.75	50.62	**10.19**

Table 6: VA analysis: Sentinel-2 images prepared with the BEST operator, SAM vs Euclidean, Otsu's algorithm vs GMM. The best results are in bold.

Let us focus the attention on the CVA method performed with the SAM distance and Otsu's algorithm. Table 7 reports the accuracy metrics measured in all Sentinel-2 images of the Czech Republic case study prepared by using the BEST operator and the MEDIAN operator. This analysis shows that, although the operator BEST allows the CVA method to achieve the

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

better performance in September (later stage detection), the operator MEDIAN allows the CVA method to achieve the better performance in June, July and August (earlier stage detection).

SAM + Otsu's algorithm	BEST				MEDIAN			
	F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-09-01 - 20-09-30	**52.19**	**86.33**	**69.26**	**35.31**	51.65	86.00	68.82	34.81
20-08-01 - 20-08-31	29.43	85.19	57.31	17.25	**40.05**	**85.37**	**62.71**	**25.04**
20-07-01 - 20-07-31	21.35	**85.09**	53.22	11.95	**24.70**	82.39	**53.54**	**14.09**
20-06-01 - 20-06-30	19.07	**82.17**	50.62	10.54	**24.42**	81.06	**52.74**	**13.91**

Table 7: CVA analysis: Sentinel-2 images prepared with the BEST operator vs Sentinel-2 images prepared with the BEST operator. CVA performed with SAM and Otsu's algorithm. The best results are in bold.

By considering the Sentinel-2 imagery set prepared with the operator MEDIAN, we selected the following SVIs: 'NGDRI', 'NDWI_A', 'DRSBis', 'TCW', 'DRS' (see Table 8 for the mathematical formulation of the selected SVIs).

SVI 1	Algebraic formulation	Literature
DRS	$\sqrt{B4^2 + B12^2}$ $\sqrt{\mathrm{B4}^2 + \mathrm{B12}^2}$	Candotti et al., 2022; Huo et al., 2021; Zhang,et al., 2021
DRS2	$\sqrt{\mathrm{B4}^2 + \mathrm{B11}^2}$	Candotti et al., 2022
NGDRI	$\frac{\mathrm{B3} - \mathrm{B4}}{\mathrm{B3} + \mathrm{B4}}$	Huo et al., 2021; Barta et al., 2021; Zhang, et al 2021; Andresini et al., 2023b;
NDWI_A	$\frac{\mathrm{B8} - \mathrm{B11}}{\mathrm{B8} + \mathrm{B11}}$	Huo et al., 2021
TCW	0.1509 B2 + 0.1973 B3 + 0.3279 B4 + 0.3406 B8 - 0.7112 B11 - 0.4572 B12	Barta et al., 2021; Candotti et al., 2022

Table 8: SVI selection for CVA

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Table 9 reports the results of the CVA performed with SAM and Otsu's algorithm extending the Sentinel-2 spectrum (prepared with the MEDIAN operator) of each pixel with the selected SVIs' vector. As baseline, let us consider the results of the CVA performed with SAM and Otsu's algorithm on the Sentinel-2 spectrum. The use of the selected SVIs contribute to improve the performances of CVA analysis in all the tested periods.

SAM + Otsu's algorithm	Sentinel-2				Sentinel-2 + SVIs			
	F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-09-01 - 20-09-30	51.65	86.00	68.82	34.81	**56.94**	**88.85**	**72.90**	**39.81**
20-08-01 - 20-08-31	40.05	85.37	62.71	25.04	**45.42**	**87.42**	**66.42**	**29.38**
20-07-01 - 20-07-31	24.70	82.39	53.54	14.09	**27.59**	**85.31**	**56.45**	**16.00**
20-06-01 - 20-06-30	24.42	81.06	52.74	13.91	**25.24**	**83.35**	**54.29**	**14.44**

Table 9: CVA analysis: Sentinel-2 images prepared with the MEDIUM operator vs Sentinel-2 images prepared with the MEDIUM operator extended with the SVIs: 'NGDRI', 'NDWI_A', 'DRSBis', 'TCW', 'DRS'. CVA performed with SAM and Otsu's algorithm. The best results are in bold.

Finally, we compare the performance of CVA with the performance of its deep neural variant named CVA-NN. Table 10 reports the results of the evaluated configurations using the Sentinel-2 imagery datasets prepared with the operator MEDIAN. The UNets were pre-trained for forest type classification by using 800 scenes randomly sampled in 2020 across Czeck Republic in the same months considered for the bark beetle detection. The used forest type ground truth was created by the partner SRI within SWIFTT project (https://code.earthengine.google.com/5874048ea778501b60d67c353ab6a9d2; it refers to 2022). This analysis shows that CVA-NN outperforms CVA in the late detection task (experiments performed in August and September 2020), while CVA outperforms CVA-NN in the early detection task (experiments performed in June and July 2020).

Sentinel-2, MEDIAN operator	CVA				CVA-NN			
	F(D)	F(H)	macro F	IoU	F(D)	F(H)	macro F	IoU
20-09-01 - 20-09-30	51.65	86.00	68.82	34.81	**58.61**	**89.26**	**73.94**	**41.46**
20-08-01 - 20-08-31	40.05	85.37	62.71	25.04	**42.54**	**86.98**	**64.76**	**27.02**
20-07-01 - 20-07-31	**24.70**	82.39	**53.54**	**14.09**	22.54	83.58	53.06	12.70
20-06-01 - 20-06-30	**24.42**	**81.06**	**52.74**	**13.91**	18.72	80.26	49.49	10.33

Table 10: CVA vs CVA-NN analysis. The best results are in bold.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Finally, Table 11 reports the results of CVA-NN and CVA performed using SAM and Otsu's algorithm with the output of the UNet extracted Deep Embedding enriched with the S2 bands and/or the SVIs: 'NGDRI', 'NDWI_A', 'DRSBis', 'TCW', 'DRS' (configurations CVA-NN+S2, CVA-NN+SVIs and CVA-NN+S2+SVI).These results confirm the better performance of a CVA-NN configuration in the later detection than in the earlier detection. In addition, the consideration of SVIs, in addition to the Deep Embedding, allows us to achieve a further improvement in the earlier detection (in particular, June and July), but without outperforming the performance achieved with the traditional CVA+SVIs in the same months (see results reported in Tables 9 and 11).

Sentinel-2, MEDIAN operator	CVA method	F(D)	F(H)	macroF	IoU
20-09-01 - 20-09-30	**CVA-NN**	**58.61**	89.26	**73.94**	**41.46**
	CVA-NN+S2	52.86	86.86	69.86	35.93
	CVA-NN+SVIs	57.63	89.08	73.35	40.48
	CVA-NN+S2+SVIs	58.31	**89.50**	73.90	41.15
20-08-01 - 20-08-31	**CVA-NN**	42.54	86.98	64.76	27.02
	CVA-NN+S2	38.08	84.49	61.29	23.52
	CVA-NN+SVIs	**44.99**	**87.27**	**66.13**	**29.03**
	CVA-NN+S2+SVIs	42.36	87.91	65.13	26.87
20-07-01 - 20-07-31	**CVA-NN**	22.54	83.58	53.06	12.70
	CVA-NN+S2	24.70	82.39	53.54	14.09
	CVA-NN+SVIs	**25.71**	**84.82**	**55.26**	**14.75**
	CVA-NN+S2+SVIs	22.33	85.24	53.79	12.57
20-06-01 - 20-06-30	**CVA-NN**	18.72	80.26	49.49	10.33
	CVA-NN+S2	1.88	91.78	46.83	0.95
	CVA-NN+SVIs	13.11	87.79	50.45	7.01
	CVA-NN+S2+SVIs	**22.48**	**83.38**	**52.93**	**12.66**

Table 11: CVA-NN analysis using the SVIs selected with a wrapper approach. The best results are in bold.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

On-field data development and evaluation protocol

Selection of testers

In the project, we finally selected the AI method that was developed to train a pixelwise classification model using the RF algorithm in the configuration SVI+SS. The proof of concept of this method is also considered for the bark beetle outbreak mapping service in the SWIFTT platform. This method was selected as, in the preliminary investigation conducted by UNIBA in WP1, it achieved the best trade-off between accuracy and complexity (also in terms of data preparation). On the other hand, the development of a pixelwise classification model, instead of a sematic segmentation model, for bark beetle outbreak detection is coherent with the data collection protocol finally adopted in SWIFTT project to obtain the on-field data regarding bark beetle outbreaks.

This data collection protocol, as formulated in WP3, describes how the ground truth information regarding both the healthy areas and the damaged areas of monitored forests are expected to be collected on-the-ground by foresters involved in the project by using the data collection application (APP) developed by partner Timbtrack. Specifically, the polygons acquired with the APP, equipped with mandatory metadata (e.g., outbreak start date, acquisition date, stage at acquisition time, percentage of affected trees) are used to identify spatial and temporal coordinates of Sentinel-2 images to download, as well as to obtain the bark beetle outbreak masks of the downloaded Sentinel-2 images. These masks are used both to supervise the model development and evaluate the accuracy of the developed model. We note that semantic segmentation models could be developed only if the on-field labeling was done at rectangular forest scene granularity (i.e., a rectangular scene is fully annotated so that it can be fully considered for the model development). On the other hand, pixelwise classification models could be always developed also if the labeling was done at the forest pixel granularity (i.e., an irregularly shaped polygon is annotated, but no information is provided regarding the labeling of the polygon surrounding that must be neglected for the model development). In the in-field data collection (still ongoing) in SWIFTT, foresters are using the SWIFTT APP to annotate disjoint, healthy or damaged, polygons without providing any annotation regarding the polygon surrounding. Therefore, in-field data are really collected for pixelwise classification.

For the in-field testing of the RF-based model, we considered:

- **Ukraine testers**: They provided a map of bark beetle outbreaks in the north of Kyviska (Ukraine) created on August 11, 2018. This map was made available by SRI partner within the project SWIFTT. It was created without using the SWIFTT APP.
- **Latvia testers**: They provided a map of healthy spruce polygons observed in March 2025, some polygons of historical sanitary cuts performed between 2023 and 2024 and, finally, some polygons of damaged areas and healthy areas, that were acquired with the SWIFTT APP according to the guidelines formulated for the in-field data collection protocol jointly formulated by the SWIFTT partners in WP3. All these data were provided by the SWIFTT partner Rigas Mezi. They are in the Latvia forest in the Riga zone.

Test phase execution

The test phase execution was performed at the University of Bari Aldo Moro using labeled Sentinel-2 imagery data obtained from Ukraine tester and labeled Sentinel-2 imagery data

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

developed according to in-field data provided by Latvia testers and used the data preparation pipeline implemented by UNIBA.

The testing phases was conducted partitioning the imagery dataset in training set, considered for the model development phase, and testing set, considered for the model evaluation phase. The accuracy performance was evaluated by measuring Fscore (harmonic mean of Precision and Recall) on both “healthy” and “damaged” classes, macroF (average of F-scores as they were measured on both “healthy” and “damaged” classes) and IoU (intersection over Union). Whenever Sentinel-2 images were downloaded for the same area at multiple timestamps, the images of the same area were assigned to either the training set or the testing set for all considered timestamps. This is to avoid that any evaluation bias (possibly caused by considering differently timestamped views of the same area both for the development phase and the evaluation phase) may reduce the trustiness in the evaluation.

A detailed description of on-field data provided by testers, imagery Sentinel-2 dataset created for both model development and evaluation, as well as obtained results is reported in the following section.

Results

In this Section we illustrate results obtained in the on-filed evaluation and summarize lessons learned

Results and interpretation

UKRAINE 2018 bark beetle outbreak ground map

In this Section we illustrate back-testing results achieved on the ground-truth dataset provided by the partner SRI in the project SWIFTT. This dataset refers to ground truth data timestamped on August 11, 2018. The Sentinel-2 images were prepared from the partner SRI at the timestamps 17-9-3, 18-4-8, 18-4-13, 18-4-21, 18-05-01, 18-05-08, 18-05-11, 18-05-26, 18-06-02, 18-06-10 and 18-08-11. A single scene was acquired in the north of Kyviska (Ukraine) with size 1431×2493 at 10 mt^2. The ground truth map of the bark beetle outbreaks refers to 1.19% of the scene pixels identified as damaged on August 11, 2018. Figure 9 shows the RGB and ground truth map of the Sentinel-2 image acquired on August 11, 2018. Figure 10 shows the box plots of the Sentinel-2 bands grouped per classes and acquired on August 11, 2018.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

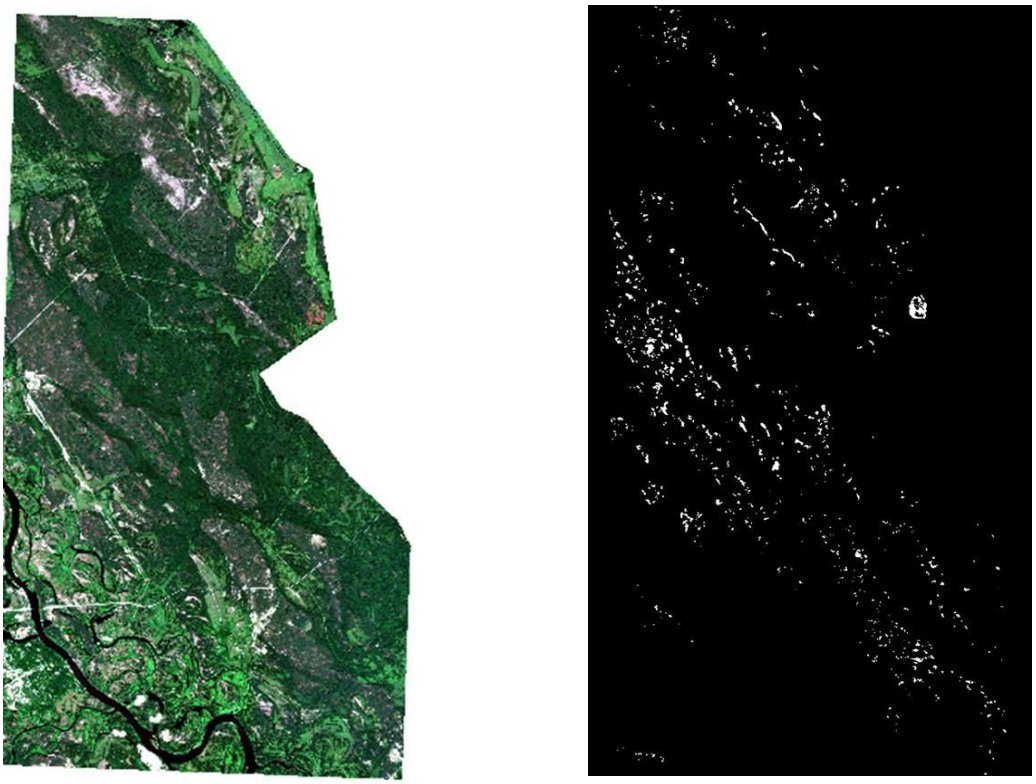


Figure 9: RGB of the Ukraine dataset acquired on August 11, 2018 and ground truth map of the bark beetle outbreak identified on August 11, 2018

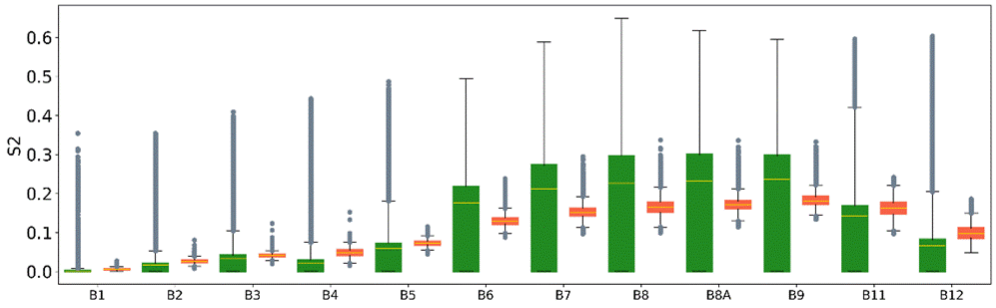


Figure 10: Box plots of Sentinel-2 bands (in logscale) of the Ukraine dataset acquired on August 11, 2018. Spectral bands are grouped per classes: orange boxes (class "damaged") and green boxes (class "healthy"). Box plots are computed for all pixels in the rectangular scene

We partitioned the scene in 897 tiles with size 64×64 and randomly selected 718 tiles for the model development and 179 for the model evaluation. We tested the performance of both RF model (trained with w=10 and depth=25) and the EMF model. The RF model was trained by using spectral-spatial features (Andresini et al. 2024b) and SVIs as identified in the studies of Andresini et al. (2024b), Barta et al. (2021), Candotti et al. (2022), Fernandez-Carillo et al. (2020), Huo et al. (2021) and Zhang, Cong et al. (2021). All models were developed and evaluated at the timestamps: 18-4-8. 18-4-13, 18-4-21, 18-05-01, 18-05-08, 18-05-11, 18-05-26, 18-06-02, 18-06-10 and 18-08-11. The RF models were developed and evaluated in the image acquired in a single timestamp. The EMF models were fine-tuned on the time-series of

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

images collected from 17-9-03 to the considered timestamp. Results are reported in Table 12. These results show that EMF systematically outperforms RF in all periods, although EMF requires more complex data preparation steps as multiple clean images should be obtained over time for model development and evaluation. This data preparation step is difficult to automate for frequent re-training and/or re-evaluation of the model. On the other hand, the RF achieved good performance with smaller data preparation constraints. The use of SVIs allows RF to gain accuracy compared to the baseline version (S2). In general, RF achieved similar accuracy performance with the different selections of SVIs used in literature study, although we note that good accuracy performance was achieved with the light RF – SVIs configuration identified by Andresini et al. 2024b by considering three SVIs only (NGDRI, NMDI and MCARI).

S2 imagery data	Feature configuration	F(D)	F(H)	macro F	IoU
18-08-11	**RF - S2 (Andresini et al. 2024b)**	69.61	99.53	84.57	53.39
	RF – SVIs (Andresini et al. 2024b)	70.70	99.53	85.11	54.67
	RF – SVIs (Barta et al., 2021)	70.73	99.53	85.13	54.71
	RF – SVIs (Candotti et al., 2022)	70.01	99.52	84.77	53.86
	RF – SVIs (Fernandez-Carillo et al., 2020)	70.02	99.54	84.78	53.87
	RF – SVIs (Huo et al., 2021)	70.95	99.53	85.24	54.98
	RF – SVIs (Zhang, Cong et al., 2021).	70.82	99.53	85.17	54.83
17-9-03 -- 18-08-11	**EMF**	**72.85**	**99.60**	**86.22**	**57.29**
18-06-10	**RF - S2 (Andresini et al. 2024b)**	61.21	99.46	80.34	44.11
	RF – SVIs (Andresini et al. 2024b)	62.59	99.46	81.02	45.55
	RF – SVIs (Barta et al., 2021)	62.65	99.47	81.06	45.62
	RF – SVIs (Candotti et al., 2022)	63.18	99.46	81.32	46.18
	RF – SVIs (Fernandez-Carillo et al., 2020)	61.67	99.45	80.56	44.59
	RF – SVIs (Huo et al., 2021)	62.50	99.46	80.98	45.45
	RF – SVIs (Zhang, Cong et al., 2021).	62.57	99.45	81.01	45.53
17-9-03 -- 18-06-10	**EMF**	**68.68**	**99.54**	**84.11**	**52.30**
18-06-02	**RF - S2 (Andresini et al. 2024b)**	55.96	99.40	77.68	38.85
	RF – SVIs (Andresini et al. 2024b)	56.93	99.40	78.17	39.79
	RF – SVIs (Barta et al., 2021)	57.61	99.42	78.51	40.46
	RF – SVIs (Candotti et al., 2022)	58.47	99.40	78.94	41.32
	RF – SVIs (Fernandez-Carillo et al., 2020)	56.64	99.40	78.02	39.51
	RF – SVIs (Huo et al., 2021)	57.38	99.40	78.39	40.23
	RF – SVIs (Zhang, Cong et al., 2021).	57.50	99.40	78.45	40.35
17-9-03 -- 18-06-02	**EMF**	**68.12**	**99.55**	**83.84**	**51.65**
18-05-26	**RF - S2 (Andresini et al. 2024b)**	53.91	99.38	76.65	36.91
	RF – SVIs (Andresini et al. 2024b)	54.65	99.38	77.01	37.60
	RF – SVIs (Barta et al., 2021)	55.06	99.39	77.22	37.99
	RF – SVIs (Candotti et al., 2022)	55.02	99.37	77.19	37.95
	RF – SVIs (Fernandez-Carillo et al., 2020)	54.02	99.38	76.70	37.00
	RF – SVIs (Huo et al., 2021)	55.36	99.39	77.37	38.27
	RF – SVIs (Zhang, Cong et al., 2021).	55.10	99.38	77.24	38.03
17-9-03 -- 18-05-26	**EMF**	**62.39**	**99.50**	**80.95**	**45.34**
18-05-11	**RF - S2 (Andresini et al. 2024b)**	51.56	99.37	75.47	34.74
	RF – SVIs (Andresini et al. 2024b)	52.31	99.37	75.84	35.42
	RF – SVIs (Barta et al., 2021)	52.47	99.37	75.92	35.57

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

	RF – SVIs (Candotti et al., 2022)	52.73	99.37	76.05	35.80
	RF – SVIs (Fernandez-Carillo et al., 2020)	52.01	99.37	75.69	35.15
	RF – SVIs (Huo et al., 2021)	52.70	99.37	76.03	35.78
	RF – SVIs (Zhang, Cong et al., 2021).	52.75	99.37	76.06	35.82
17-9-03 -- 18-05-11	**EMF**	**54.14**	**99.46**	**76.80**	**37.12**
18-05-08	**RF - S2 (Andresini et al. 2024b)**	50.74	99.35	75.05	33.99
	RF – SVIs (Andresini et al. 2024b)	51.74	99.36	75.55	34.90
	RF – SVIs (Barta et al., 2021)	51.57	99.36	75.46	34.75
	RF – SVIs (Candotti et al., 2022)	51.61	99.35	75.48	34.78
	RF – SVIs (Fernandez-Carillo et al., 2020)	51.25	99.36	75.31	34.45
	RF – SVIs (Huo et al., 2021)	51.72	99.36	75.54	34.88
	RF – SVIs (Zhang, Cong et al., 2021).	51.68	99.36	75.52	34.84
17-9-03 -- 18-05-08	**EMF**	**54.85**	**99.45**	**77.15**	**37.78**
18-05-01	**RF - S2 (Andresini et al. 2024b)**	50.07	99.35	74.71	33.39
	RF – SVIs (Andresini et al. 2024b)	51.34	99.35	75.34	34.54
	RF – SVIs (Barta et al., 2021)	51.27	99.35	75.31	34.47
	RF – SVIs (Candotti et al., 2022)	51.32	99.35	75.33	34.52
	RF – SVIs (Fernandez-Carillo et al., 2020)	50.80	99.35	75.07	34.05
	RF – SVIs (Huo et al., 2021)	51.28	99.35	75.32	34.48
	RF – SVIs (Zhang, Cong et al., 2021).	51.51	99.35	75.43	34.69
17-9-03 -- 18-05-01	**EMF**	**53.86**	**99.45**	**76.66**	**36.86**
18-04-21	**RF - S2 (Andresini et al. 2024b)**	41.76	99.31	70.53	26.39
	RF – SVIs (Andresini et al. 2024b)	44.66	99.32	71.99	28.75
	RF – SVIs (Barta et al., 2021)	43.54	99.32	71.43	27.83
	RF – SVIs (Candotti et al., 2022)	46.87	99.34	73.11	30.61
	RF – SVIs (Fernandez-Carillo et al., 2020)	45.30	99.33	72.32	29.28
	RF – SVIs (Huo et al., 2021)	45.43	99.33	72.38	29.39
	RF – SVIs (Zhang, Cong et al., 2021).	45.58	99.33	72.46	29.52
17-9-03 -- 18-04-21	**EMF**	**53.42**	**99.44**	**76.43**	**36.45**
18-04-13	**RF - S2 (Andresini et al. 2024b)**	46.74	99.31	73.02	30.50
	RF – SVIs (Andresini et al. 2024b)	47.64	99.32	73.48	31.26
	RF – SVIs (Barta et al., 2021)	47.53	99.32	73.42	31.17
	RF – SVIs (Candotti et al., 2022)	48.66	99.33	73.99	32.15
	RF – SVIs (Fernandez-Carillo et al., 2020)	47.38	99.32	73.35	31.04
	RF – SVIs (Huo et al., 2021)	48.12	99.32	73.72	31.69
	RF – SVIs (Zhang, Cong et al., 2021).	48.34	99.32	73.83	31.87
17-9-03 -- 18-04-13	**EMF**	**52.44**	**99.44**	**75.94**	**35.54**
18-04-08	**RF - S2 (Andresini et al. 2024b)**	45.08	99.31	72.19	29.10
	RF – SVIs (Andresini et al. 2024b)	45.77	99.31	72.54	29.67
	RF – SVIs (Barta et al., 2021)	45.95	99.30	72.63	29.83
	RF – SVIs (Candotti et al., 2022)	46.48	99.31	72.90	30.27
	RF – SVIs (Fernandez-Carillo et al., 2020)	45.37	99.31	72.34	29.34
	RF – SVIs (Huo et al., 2021)	47.26	99.31	73.28	30.94
	RF – SVIs (Zhang, Cong et al., 2021).	46.82	99.31	73.06	30.56
17-9-03 -- 18-04-08	**EMF**	**52.79**	**99.43**	**76.11**	**35.86**

Table 12: RF and EMF trained and evaluated in the Ukraine dataset. The best results are in bold.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

LATVIA historical sanitary cut ground data (V1)

In this Section, we illustrate preliminary back-testing results achieved on historical, ground-truth polygons provided by the Rigas Mezi partner. The polygons regard the sanitary cuts performed in January 2024 in the Latvia forest. Based on some communications with Rigas Mezi partner, the considered sanitary cuts were due to the bark beetle outbreaks occurred in 2023. These historical data were collected without additional information regarding the bark beetle outbreak history. So, we could not count on information on when each outbreak was in the early stage, when it passed to the red stage and, finally, when the tree dieback was observed due to the grey stage of the outbreak. In addition, each sanitary cut polygon covers both damaged forest patches that were cut due to the presence of tree dieback caused by the bark beetle outbreak and healthy forest patches that were cut to mitigate the risk of future outbreaks. Unfortunately, there was no ground truth referred to the correct segmentation of healthy areas and damaged areas staying in the same sanitary cut, which may allow us to obtain a good quality ground truth labels for the supervision of the model development. So, we considered the entire sanitary cut polygons as damaged areas. In addition, no healthy polygon ground truth was available in this data collection, so we assumed that unlabelled forest surrounding the sanitary cut polygons could be considered healthy.

Based on these premises we obtained a collection of 1000 scenes with size varying from 17×19 to 325×344 pixels for a total of 2775594 pixels. To obtain these scenes, we first created multi-polygons of sanitary cut polygons by iteratively merging neighbour polygons within the radius of 150 meters. Then, we determined the minimum bounding box of each multi-polygon. Finaly, a scene of the multi-polygon was obtained by enlarging the minimum bounding box of the multi-polygon by 50 meters on each side. By constructing the scene mosaic of the case study according to the process described above, we obtained that the percentage of territory labelled as “damaged” per scene varies from 10.17% to 56.43% of the scene surface. The total percentage of damaged territory of the entire scene collection is 26.16%. We randomly split the scenes in 800 scenes covering 2303030 pixels used for model development and 200 scenes (covering 472564) used for model evaluation. In addition, as we were interested in monitoring the performance of the model on forest patches we accessed to the map of the forest type in LATVIA developed by the partner SRI within the SWIFTT project (https://code.earthengine.google.com/5874048ea778501b60d67c353ab6a9d2; the map refers to 2022). Accordingly, we identified 19565558 pixels covering a forest area for the model development and 402354 pixels covering a forest area for the model evaluation steps.

The images were downloaded in the 3857 EPSG system by GEE by using the pipeline implemented to create the Czech Republic case study data set. We considered images downloaded monthly from June 2023 to September 2023 with both the MEDIAN and BEST operator. Due to the weather conditions in Latvia in 2023, the satellite images of the obtained scenes downloaded from October to December 2023 were very cloudy.

Supervised model development. We considered the RF (w=5, depth=20) and SVM (w=2) models trained with spectral-spatial features, UNet and UNet with attention, TITANIA, BigEarth-UNet and EMF for the supervised model development.

Results with Sentinel-2 images prepared with the MEDIAN operator

We started evaluating the impact of SVIs commonly used in literature for bark beetle outbreak detection in the model developed on September 2023 with RF. Results are reported in Table

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

13. RF - S2 refers to the configuration accounting for Sentinel-2 spectral bands only, RF - S2 (Andresini et al. 2024b) accounting for spectral spatial features of Sentinel-2 spectral bands, while RF – SVIs (Andresini et al. 2024b), RF – SVIs (Barta et al., 2021), RF – SVIs (Candotti et al., 2022), RF – SVIs (Fernandez-Carillo et al., 2020), RF – SVIs (Huo et al., 2021), RF – SVIs (Zhang, Cong et al., 2021) refer to the configurations accounting for spectral-spatial features of Sentinel-2 spectral bands and SVIs. The results achieved with these filed data confirm that the spectral-spatial features allow us to achieve a gain in accuracy in the model development. However, no group of SVIs identified in the literature allows us to achieve a gain in F(D) compared to the configuration RF-S2 that used spectral-spatial features, but neglected SVIs. The selection of SVIs identified by Andresini et al. (2024b) allows us to achieve a small improvement in terms of F(H) and macroF. However, due to the negligible differences observed using these SVIs, we conclude that the SVIs, commonly used in the literature study, do not provide any significant contribution to accurately predict future sanitary cuts caused by bark beetle outbreaks in this case study.

Feature configuration	F(D)	F(H)	macroF	IoU
RF - S2	56.18	71.73	63.95	39.06
RF - S2 (Andresini et al. 2024b)	**59.49**	72.59	66.04	**42.33**
RF – SVIs (Andresini et al., 2024b)	58.52	73.63	**66.08**	41.36
RF – SVIs (Barta et al., 2021)	58.46	73.65	66.05	41.30
RF – SVIs (Candotti et al., 2022)	58.64	73.21	65.93	41.48
RF – SVIs (Fernandez-Carillo et al., 2020)	58.55	72.99	65.77	41.39
RF – SVIs (Huo et al., 2021)	58.82	73.92	66.37	41.66
RF – SVIs (Zhang, Cong et al., 2021)	58.83	74.01	66.42	41.67

Table 13: RF configurations trained and evaluated in the LATVIA SANITARY CUT V1 imagery dataset of September 2023 prepared with the median operator. The best results are in bold.

Table 14 reports the results of RF, SVM, UNet with Attention, UNet, TITANIA, BigEarth-UNet, and EMF generated for the datasets created monthly between June and September 2023 using the operator MEDIAN. For the EMF, we considered the imagery timeseries monthly produced from April 2023 to the timestamp considered for the model development and evaluation. The results show that all methods train sematic segmentation models that achieve a smaller F1(H) than the models trained with the same methods in the Czeck Republic and Ukraine case studies. This depends on the fact that we are using sanitary cut information, which labels damaged areas that, although healthy, were cut to prevent infestation diffusion. This highlights the importance of accurate field data collection to better fuel the supervision of semantic segmentation models in the considered problem. This issue has a higher effect on machine learning methods (RF and SVM) than on deep learning methods. The better performance is achieved with the semantic segmentation model trained with BigEarth-UNet having the semantic segmentation model trained with TITANIA as runner-up. In addition, the accuracy performance of models decreases moving from September 2023 to June 2023.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Again, this decrease in accuracy is higher in machine learning models than in semantic segmentation models.

S2 imagery data	Method	F(D)	F(H)	macroF	IoU
September 2023	**RF**	59.49	72.59	66.04	42.33
	SVM	60.10	79.51	69.81	42.96
	UNet+A	66.42	83.60	75.01	49.72
	UNet	67.11	82.46	74.78	50.50
	TITANIA	67.61	82.16	74.88	51.07
	BigEarth-UNet	**76.88**	**88.87**	**82.88**	**62.45**
	EMF	64.82	85.93	75.37	47.95
August 2023	**RF**	57.99	71.76	64.88	40.84
	SVM	59.08	76.80	67.94	41.92
	UNet+A	63.41	74.65	69.03	46.42
	UNet	59.30	70.10	64.70	42.15
	TITANIA	65.32	80.04	72.68	48.50
	BigEarth-UNet	**71.16**	**87.16**	**79.16**	**55.23**
	EMF	62.68	85.27	73.98	45.65
July 2023	**RF**	56.10	65.64	60.87	38.98
	SVM	57.22	74.14	65.68	40.07
	UNet+A	61.51	75.10	68.31	44.41
	UNet	58.67	77.56	68.11	41.51
	TITANIA	63.92	79.90	71.91	46.97
	BigEarth-UNet	**71.75**	**87.76**	**79.76**	**55.95**
	EMF	61.61	85.08	73.34	44.51
June 2023	**RF**	55.79	63.98	59.88	38.69
	SVM	57.41	75.92	66.67	40.26
	UNet+A	50.99	64.17	57.58	34.22
	UNet	63.89	79.89	71.89	46.94
	TITANIA	64.70	81.77	73.24	47.82
	BigEarth-UNet	**71.40**	**87.06**	**79.23**	**55.52**
	EMF	60.63	84.61	72.62	43.50

Table 14: RF, SVM, UNet+A (Attention UNet), UNet, TITANIA, BigEarth-UNet and EMF trained and evaluated with the LATVIA SANITARY V1 CUT imagery datasets prepared with the operator MEDIAN between June 2023 and September 2023. The best results are in bold.

By selecting RF and SVM as machine learning methods, as well as TITANIA and BigEarth-UNet as deep learning methods, Table 15 reports the accuracy performance obtained evaluating these methods in the Sentinel-2 imagery datasets prepared with the MEDIAN operator and the BEST operator, respectively. The results confirm conclusions drawn from results reported in Table 14 regarding the better performance of deep learning models compared to machine learning models. In addition, these results show that preparing the Sentinel-2 dataset with the MEDIAN operator allows us to achieve better performances with machine learning methods. Differences are often negligible for deep neural models except for BigEarth-UNet that significantly outperforms TITANIA when Sentinel-2 imagery is prepared with the MEDIAN operator in September 2023.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

S2 imagery data	Method	MEDIAN				BEST			
		F(D)	F(H)	macroF	IoU	F(D)	F(H)	macroF	IoU
September 2023	**RF**	**59.49**	**72.59**	**66.04**	**42.33**	58.03	71.18	64.60	40.87
	SVM	**60.10**	**79.51**	**69.81**	**42.96**	58.71	77.39	68.05	41.55
	TITANIA	**67.61**	82.16	74.88	51.07	67.30	**82.63**	**74.96**	**50.71**
	BigEarth-UNet	**76.88**	**88.87**	**82.88**	**62.45**	73.54	88.16	80.85	58.15
August 2023	**RF**	**57.99**	**71.76**	**64.88**	**40.84**	56.33	70.37	63.35	39.21
	SVM	**59.08**	**76.80**	**67.94**	**41.92**	57.41	76.05	66.73	40.27
	TITANIA	**65.32**	**80.04**	**72.68**	**48.50**	63.94	79.32	71.63	46.99
	BigEarth-UNet	71.16	87.16	79.16	55.23	**72.46**	**87.56**	**80.01**	**56.82**
July 2023	**RF**	**56.10**	**65.64**	**60.87**	**38.98**	55.59	67.20	61.40	38.49
	SVM	**57.22**	**74.14**	**65.68**	**40.07**	56.26	73.9	65.08	39.14
	TITANIA	**63.92**	**79.90**	**71.91**	**46.97**	63.20	77.40	70.30	46.20
	BigEarth-UNet	**71.75**	**87.76**	**79.76**	**55.95**	70.29	86.99	78.64	54.19
June 2023	**RF**	**55.79**	63.98	59.88	**38.69**	55.60	**65.22**	**60.41**	38.51
	SVM	**57.41**	**75.92**	**66.67**	**40.26**	55.86	75.70	65.78	38.76
	TITANIA	**64.70**	**81.77**	**73.24**	**47.82**	64.29	81.12	72.70	47.37
	BigEarth-UNet	71.40	87.06	79.23	55.52	**71.99**	**87.51**	**79.75**	**56.23**

Table 15: RF, SVM, UNet+A (Attention UNet), UNet, TITANIA, BigEarth-UNet and EMF trained and evaluated with the LATVIA SANITARY CUT V1 imagery datasets prepared with the operator BEST between June 2023 and September 2023. The best results are in bold.

Unsupervised model development. We used the CVA method (Otsu+SAM) to map the bark beetle outbreaks that were involved in a sanitary cut in 2024. As the CVA did not require any supervision, we performed the evaluation of the performance of this method on the entire mosaic of scenes obtained in the LATVIA2023 sanitary cut dataset (without splitting these scenes in training and testing scene sets). Table 16 reports the results of traditional CVA (SAM + Otsu's algorithm) performed considering the entire bi-temporal imagery Sentinel-2 collection acquired in September 2022 and September 2023, respectively. Images were obtained using the data preparation pipelines with operators: BEST and MEDIAN. The evaluation was done by using the Forest Type map of 2022 (accessed at https://code.earthengine.google.com/5874048ea778501b60d67c353ab6a9d2) to neglect pixels that were classified as "non-forest". Table 16 also reports the results of CVA performed by enriching the Sentinel-2 spectrum with the selection of SVIs identified to perform the CVA in the Czech Republic case study (Table 9). As for the Czech Republic, better results are achieved by using the MEDIAN pipeline and the use of SVIs allows us to gain in accuracy. Again, the performance of CVA, due to the lack of supervision, is significantly lower than the performance of supervised methods. For this reason, we decide to select supervised methods.

S2 imagery operator	Method	F(D)	F(H)	macroF	IoU
BEST	**CVA**	11.65	78.87	45.26	6.19
	CVA+SVIs	20.37	77.91	49.14	11.34

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

MEDIAN	**CVA**	22.05	77.12	49.58	12.39
	CVA + SVIs	26.48	77.59	52.04	12.58

Table 16: CVA and CVA+ SVIs ('NGDRI', 'NDWI_A', 'DRSBis', 'TCW', 'DRS' from Table 9) in the LATVIA SANITARY CUT V1 imagery dataset prepared with the operator MEDIAN in September 2023. The best results are in bold.

Final considerations. Despite the amazing results achieved using the deep semantic segmentation models in the experimentation performed considering historical sanitary cut polygons and creating images of scenes surrounding the historical sanitary cut polygons, the analysis of the obtained results performed in collaboration with Rigas Mezi partner highlighted that the historical sanitary cut polygons used for this preliminary evaluation study were often surrounded by either young forest areas (reforested in recent years) or non-forest areas. This special condition created a bias in the model development that was caused by the limited amount of healthy coniferous areas considered in both the model development and evaluation stage. Accordingly, a revised version of the sanitary cut dataset was constructed in collaboration with the Rigas Mezi partner.

LATVIA historical sanitary cut ground data (V2)

In this Section, we illustrate a revised version of preliminary back-testing results achieved by UNIBA to train and evaluate models for bark beetle damage assessment using the historical, ground-truth sanitary cut polygons provided by the Rigas Mezi partner. For this experiment, the Rigas Mezi partner provided an annotated version of a subset of its historical bark beetle sanitary cut database. In particular, the Rigas Mezi partner annotated 30 multi-polygons with the sanitary cut date happened on May 17, 2023 (1), August 30, 2023 (18), October 25, 2023 (3), April 29, 2024 (2), May 7, 2024 (2), July 2, 2024 (2), August 15, 2024 (2). In addition, Rigas Mezi partner provided a checked spruce healthy layer covering an area of the Latvia forest hosting the considered sanitary cut polygons. This area was explored by foresters of Rigas Mezi partner in March 2025 and identified as covered by a “healthy” coniferous forest at the visit date. In this way, UNIBA partner was able to obtain sanitary cut polygons and healthy spruce polygons that were used to download Sentinel-2 data. Specifically, the Sentinel-2 images of healthy spruce polygons were monthly downloaded, using the MEDIAN operator, in May 2023, June 2023, July 2023, August 2023, September 2023, May 2024, June 2024, July 2024, August 2024 and September 2024. In addition, the Sentinel-2 images of sanitary cut polygons were monthly downloaded within the sanitary cut year from May to the sanitary cut date. As the field ground truth maps of this experiment covered isolated healthy and damaged polygons, the constructed Senitnel-2 dataset cannot be fairy used for semantic segmentation, while it can be considered for pixelwise classification. Hence, the RF algorithm used to train a pixelwise classification model was used for the bark beetle damage assessment analysis in this dataset.

For the evaluation scope we roughly spilt all available polygons, to use the 80% of all polygons to feed the training set and the remaining 20% of all polygons to feed the testing set. The split of both sanitary cut polygons and spruce healthy polygons was performed in a stratified manner. In this way, all multi-month Sentinel-2 acquisitions of the same polygon were assigned either to the training set or to the testing set.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

The obtained Sentinel-2 training set was composed of 499420 spruce pixels (i.e. 404215 healthy spruce pixels and 64260 damaged pixels). The obtained Sentinel-2 testing set was composed of 79861 pixels (i.e., 64260 healthy spruce pixels and 15601 damaged spruce pixels). We trained and evaluated both the RF model using the multi-month Sentinel-2 training set, ad the month-based RF models trained from Sentinel-2 training set data grouped per month. Each RF model was trained using spectral Sentinel-2 bands and spectral-spatial features of both spectral bands and spectral vegetation indexes described in Andresini et al. (2024b), Barta et al. (2021), Candotti et al. (2022), Fernandez-Carillo et al. (2020), Huo et al., (2021), and Zhang, Cong et al. (2021). The RF models were trained with:

- depth=20 and w=8 using the entire multi-month Sentinel-2 training set;
- depth=15 and w=8 using the monthly-grouped Sentinel-2 training set.

In both experiments we selected Randomization as None. Table 17 reports the accuracy results achieved on the multi-month Sentinel-2 dataset of this evaluation. The best performances are achieved by training the RF models on Sentinel-2 data grouped per month and enriched with SVIs identified by either Huo et al., (2021), or Zhang, Cong et al. (2021).

Feature configuration	F(D)	F(H)	macroF	IoU
RF	48.44	85.68	67.06	31.96
RF – SVIs (Andresini et al., 2024b)	48.70	83.79	66.24	32.18
RF – SVIs (Barta et al., 2021)	47.29	84.07	65.68	30.97
RF – SVIs (Candotti et al., 2022)	47.91	82.70	65.30	31.50
RF – SVIs (Fernandez-Carillo et al., 2020)	48.12	84.72	66.42	31.68
RF – SVIs (Huo et al., 2021)	49.91	85.21	67.56	33.25
RF – SVIs (Zhang, Cong et al., 2021).	49.88	85.05	67.47	33.22
Month-grouped RF	49.27	85.84	67.56	32.69
Month-grouped RF – SVIs (Andresini et al., 2024b)	48.84	85.05	66.94	32.31
Month-grouped RF – SVIs (Barta et al., 2021)	49.62	85.11	67.37	33.00
Month-grouped RF – SVIs (Candotti et al., 2022)	48.23	84.59	66.41	31.77
Month-grouped RF – SVIs (Fernandez-Carillo et al., 2020)	48.50	84.88	66.69	32.02
Month-grouped RF – SVIs (Huo et al., 2021)	**52.31**	**86.10**	**69.20**	**35.41**
Month-grouped RF – SVIs (Zhang, Cong et al., 2021).	**52.20**	**86.06**	**69.13**	**35.31**

Table 17: RF configurations trained and evaluated in the LATVIA SANITARY CUT (V2) imagery dataset prepared with the MEDIAN operator. The best results are in bold.

Figure 11 shows the map (colored in red) of the testing ground truth damaged polygons and the map (colored in blue) of the predicted damaged polygons obtained using the monthly-based RF trained with SVIs identified by Huo et al. (2021). We note that the different precisions of the same polygons delimited at different times may also depend on the quality of Sentinel-2 data downloaded in the month considered for both the model development and the evaluation. We also note that, in all months, all testing damaged polygons were roughly identified, although the predicted areas may not precisely cover some neighbor damaged areas within the predicted polygon surface. On the other hand, the developed model identifies several small false positive areas in the healthy spruce layer. We consider this behavior of the

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

developed model related to the use of sanitary cut polygons instead of precisely identified damaged polygons for the supervision of the model development phase. In fact, the use of sanitary cut information to produce the mask of the damaged class led us to consider healthy forest patches, that were cut to prevent future bark beetle swarming, as prototype examples of the “damaged” class.

May

June

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

July

August

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

September

Figure 11: Predicted map of testing images grouped per month (May, June, July, August and September) in the evaluation performed on the testing set of the LATVIA SANITARY CUT (V2) with the month-based RF model. The red polygons denote the ground truth map of the testing damaged polygons, and the blue polygons denote the predicted map of damaged areas shown in the spruce forest layer.

LATVIA damaged field ground data (V1)

Accounting for the conclusions drawn using both sanitary cut information and checked healthy spruce information provided by the Rigas Mezi partner, we re-trained the RF model by using the first batch of the damaged polygons in-field recorded using the SWIFTT APP by Rigas Mezi partner in February 2025. The batch of polygons acquired in February 2025 covered spruce areas in Latvia affected by a bark beetle damage that was in the stages 2 or 3 at the acquisition time. As no healthy spruce polygon was acquired by foresters of Rigas Mezi in this phase of the field data collection, we considered the healthy spruce layer previously provided by Rigas Mezi partner. Notably, the healthy spruce layer was constructed by Rigas Mezi partner for the area surrounding the historical sanitary cut polygons considered for the experimentation conducted with the “LATVIA historical sanitary cut ground data (V2)” data collection. The damaged polygons acquired using the SWIFTT APP in February 2025 refer to a different geographic area of LATVIA. However, the Rigas Mezi partner confirmed that the same family and type of spruce forest were availble in both areas (i.e., the area hosting the historical sanitary cuts happened in 2023 and 2024 and the area hosting the damaged patched recorded in February 2025 with the SWIFTT app). Based on these premises we created the version V1 of the dataset obtained from LATVIA damaged field ground data, by considering damaged polygons recorded in February 2025 with the SWIFTT app and healthy polygons identified in the checked healthy spruce layer observed in March 2025. We downloaded the Sentinel-2 images of both types of polygons in May 2024, June 2024, July 2024, August 2024 and September 2024 by covering an area of 35168 pixel at 10 mt^2 spatial resolution per month

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

(175840 pixels in total). We used the MEDIAN operator to download Sentinel-2 images. By splitting both damaged and healthy polygons in training polygons and testing polygons, we created a training Sentinel-2 spruce dataset containing 159145 pixels (129305 “healthy pixels” and 29840 damaged pixels) spanned on the period May-September 2024 and a testing Sentinel-2 spruce dataset containing 16695 pixels (12505 “healthy pixels” and 4190 damaged pixels) spanned on the period May-September 2024.

We trained and evaluated both the RF model using the multi-month Sentinel-2 training set. Each RF model was trained using spectral Sentinel-2 bands and spectral-spatial features of both spectral bands and spectral vegetation indexes described in Andresini et al. (2024b), Barta et al. (2021), Candotti et al. (2022), Fernandez-Carillo et al. (2020), Huo et al., (2021), and Zhang, Cong et al. (2021). The RF models were trained with depth=20 and w=11 and we selected Randomization as None. Table 18 reports the accuracy results achieved on the multi-month Sentinel-2 dataset of this evaluation. The best performances are achieved by training the RF models on Sentinel-2 data enriched with SVIs identified by Candotti et al., (2022). Notably, the accuracy performance increased significantly by using precise damage polygons in place of sanitary cut polygons.

Feature configuration	F(D)	F(H)	macroF	IoU
RF	72.62	91.90	82.26	57.01
RF – SVIs (Andresini et al., 2024b)	71.52	91.72	81.62	55.66
RF – SVIs (Barta et al., 2021)	70.29	91.48	80.88	54.19
RF – SVIs (Candotti et al., 2022)	**72.78**	**92.30**	**82.54**	**57.21**
RF – SVIs (Fernandez-Carillo et al., 2020)	71.87	91.98	81.93	56.10
RF – SVIs (Huo et al., 2021)	69.69	91.31	80.50	53.48
RF – SVIs (Zhang, Cong et al., 2021).	68.54	90.94	79.74	52.14

Table 18: RF configurations trained and evaluated in the LATVIA DAMAGED FIELD GROUND TRUTH (V1) imagery dataset prepared with the MEDIAN operator. The best results are in bold.

Finally, Figure 12 shows an example of a damaged polygon of the testing set as it was predicted using the model trained with the configuration **RF – SVIs** (Candotti et al., 2022) of Table 18. The results of this on-field evaluation were also presented by the partner UNIBA supported by the partner Rigas Mezi in the online webinar on “Leveraging AI Models for Insect Damage Detection in European Forest” held on July 11, 2025.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

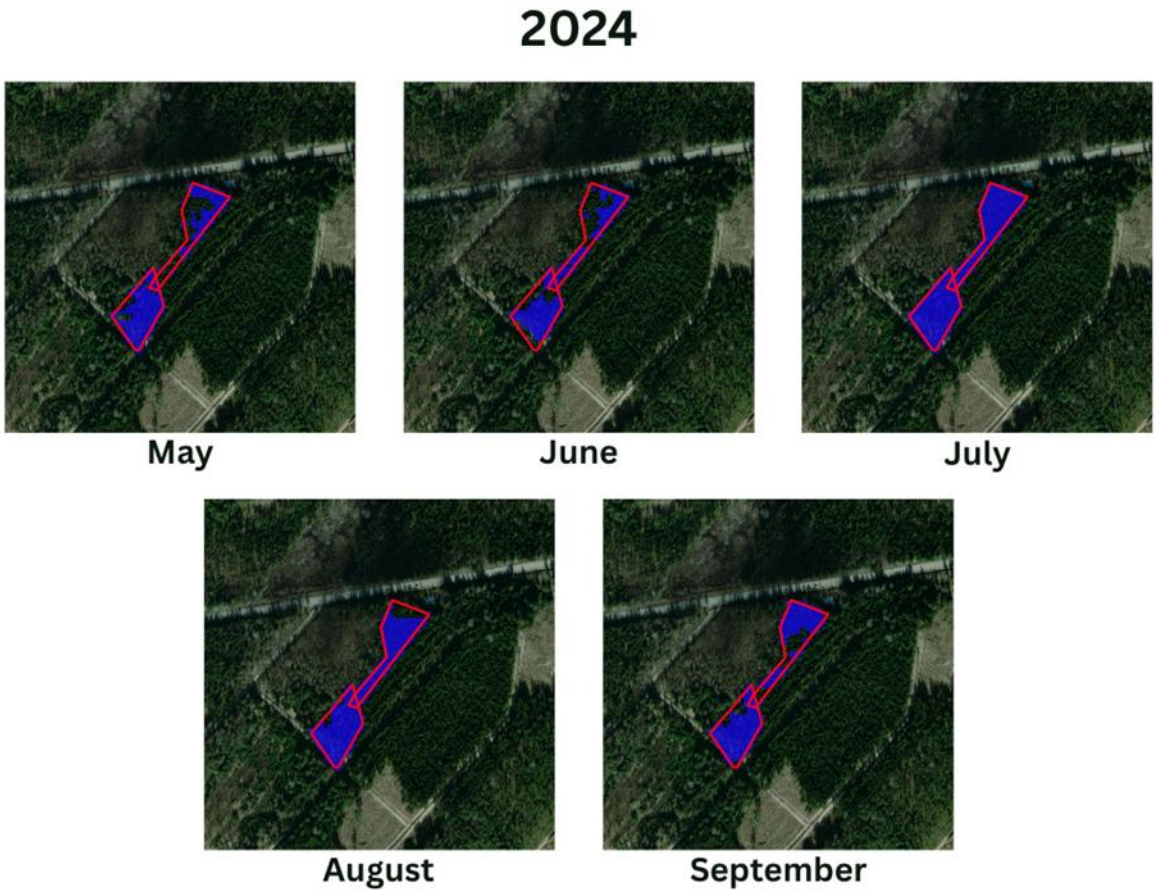


Figure 12: Map predicted for a testing damaged polygon of dataset LATVIA (V1) using the model developed with the configuration RF – SVIs (Candotti et al., 2022

LATVIA damaged field ground data (V2)

In this Section, we report the experiments performed using an extended version of the field ground data V1 provided by Rigas Mezi partner. Specifically, we extended field data used to create V1 with a new batch of both healthy and damaged polygons acquired by Rigas Mezi Foresters with the SWIFTT APP. Accordingly, we created the version V2 of LATVIA damaged field ground data. This dataset was obtained by including:

- The checked healthy spruce data layer provided by Rigas Mezi in February 2025 and already used in the evaluation done with the V2 of sanitary cut and V1 of damaged field ground truth data.
- The damaged polygons provided by Rigas Mezi partner and collected using the SWIFTT APP in February 2025. These polygons were already used in the version V1 of this field LATVIA dataset.
- The new batch of damaged polygons provided by Rigas Mezi partner and collected using the SWIFTT APP in April 2025. The polygons recorded with null metadata were neglected.
- The batch of polygons provided by Rigas Mezi partner and collected using SWIFTT application in April 2025, where the stage of the damage is early (stage 1) in 2025. We

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

assumed these polygons were in the healthy stage in the 2024 e we considered them as healthy to create the version V2 of the field LATVIA dataset.
- The batch of 4 healthy polygons checked by Rigas Mezi partner in June 2025.
- The batch of 9 damaged polygons checked by Rigas Mezi partner in June 2025. For these polygons, Rigas Mezi also provided the information regarding the start outbreak date. This information was unavailable with polygons acquired in the previous batch.

We downloaded Sentinel-2 images of all polygons listed above by considering the range of dates:

- May - September 2022, May - September 2023, and May - September 2024 for all polygons that are labeled as healthy (i.e., polygons from the checked healthy spruce layer, the stage 1 batch provided in April 2025, and the healthy batch checked in June 2025).
- May-September 2024 and May 2025 for all damaged polygons provided in February and April.
- May-September 2023 for polygons for which the start date of the outbreak was provided and this date was before 2023.

The final imagery dataset obtained as described above covers 530183 pixels in total. We used the MEDIAN operator to download Sentinel-2 images. After splitting both damaged and healthy polygons in training polygons and testing polygons, we created a labeled Sentinel-2 training set containing 470161 pixels (406265 “healthy pixels” and 63896 “damaged” pixels) and a labeled Sentinel-2 testing set containing 60022 pixels (46585 “healthy pixels” and 13437 “damaged” pixels). Notably, these pixels were acquired in different months (May, June, July, August, and September) for separate polygons.

Following the same methodology adopted to perform the model development and evaluation using the labeled imagery dataset produced by version V1 of LATVIA field ground data, we trained and evaluated the RF model using the multi-month Sentinel-2 imagery dataset, using spectral Sentinel-2 bands, SVIs and spectral-spatial features of both spectral bands and SVIs. We considered the SVI selections described in Andresini et al. (2024b), Barta et al. (2021), Candotti et al. (2022), Fernandez-Carillo et al. (2020), Huo et al., (2021), and Zhang, Cong et al. (2021). The RF models were all trained with depth=20 and w=3 and setting Randomization as None. Table 19 reports the accuracy results achieved on the multi-month Sentinel-2 dataset of this evaluation. The results show that also in this case the use of spectral data enriched with SVIs allow us to gain accuracy compared with configurations that consider the spectral bands of Sentinel-2 images only. In this case, the highest accuracy is achieved by training and evaluating the RF model on Sentinel-2 data enriched with the SVIs identified by Zhang, Cong et al. (2021).

Final considerations regard the fact the performance achieved in the evaluation done with V2 shows an improvement of both F1(D) and F1(H), probably thanks to the bigger amount and variety of field data available for the model development and evaluation.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Feature configuration	F(D)	F(H)	macroF	IoU
RF	77.73	94.33	86.03	63.58
RF – SVIs (Andresini et al., 2024b)	78.53	94.53	86.53	64.65
RF – SVIs (Barta et al., 2021)	79.60	94.77	87.18	66.11
RF – SVIs (Candotti et al., 2022)	78.05	94.49	86.27	64.00
RF – SVIs (Fernandez-Carillo et al., 2020)	78.29	94.45	86.37	64.32
RF – SVIs (Huo et al., 2021)	78.90	94.60	86.75	65.15
RF – SVIs (Zhang, Cong et al., 2021).	**80.08**	**94.85**	**87.47**	**66.78**

Table 19: RF configurations trained and evaluated in the LATVIA DAMAGED FIELD GROUND TRUTH (V2) imagery dataset prepared with the MEDIAN operator. The best results are in bold.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Lessons learned

The experimental evaluation, conducted by using the benchmark Sentinel-2 imagery dataset produced according to DEFID2 map of bark beetle outbreak in September 2020 in Czech Republic, confirmed that the pixelwise classification model trained with RF, SVIs and spectral-spatial features of Sentinel-2 bands and SVIs is still competitive compared to pixelwise classification models and sematic segmentation models trained with complex deep neural networks (comprising new methods developed and tested by UNIBA partner reusing deep foundation models). Hence, the model testing with labeled imagery datasets constructed according to the in-field data was mainly conducted with the RF configuration of a pixelwise classifier. Preliminary tests were done on a scene of a forest in the north of Kyviska (Ukraine) for which the damage was assessed in August 2018. Further tests were performed with historical sanitary cut polygons provided by Riga Mezi partner, and damage polygons collected by Riga Mezi partner using the APP developed by Timbtrack partner in SWIFTT. In the following we summarize key learnings and recommendations for future testing.

Key learnings from model testing

The quality of on-field data (both healthy and damaged polygons) collected with the data collection phase of the project SWIFTT will be paramount of developing accurate models for bark beetle damage assessment in Sentinel-2 images of different countries. At the same time, the set-up of parameters for RF model development must be carefully identified accounting for class imbalance and the real amount of labeled data available for model development. In short, key learnings from model testing are defined as follows:

- The use of sanitary cut polygons instead of precisely identified damaged polygons must be avoided. Sanitary cuts introduce data hallucinations regarding the class “healthy” for model development that may be the cause of a high number of false positives in the testing phase. This is related to the fact that using sanitary cuts labeled as damaged areas can lead to feed the model development with healthy patches handled as prototypical examples of the “damaged” area for the model supervision.
- The quantity as well as, the temporal and spatial variety of the on-field collected polygons used to construct the labeled imagery dataset considered to train the pixelwise classification model has a crucial effect on the model performance in terms of both F(D) and F(H) measured at the testing stage.
- The real-world distribution of healthy and damage classes is imbalanced (i.e., the number of pixels associated with damaged forest areas is significantly less than the number of pixels associated with healthy forest areas). A high number of polygons covering a large healthy area must be acquired to allow the construction of a labeled imagery dataset used to supervise model development in a realistic, imbalanced, scenario. A cost parameter must be set for training the RF accounting for the imbalance condition of the training dataset.
- An updated forest type layer must be used as a spatial filter to visualize damages detected in spruce areas only.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Recommendation for future testing

The main recommendations for the future testing regard the completion of the in-field data collection in the project, which is mandatory for the construction of the labeled imagery datasets to be considered for the final model development and evaluation. These recommendations are listed in the followings.

- No sanitary cut in place of damaged polygons.
- Both healthy and damaged polygons must be obtained, to allow the construction of an imbalanced labeled imagery dataset with the class “damage” appearing as a minority class with respect to the class “healthy”.
- Damaged and healthy polygons must precisely delineate areas that host only damaged or healthy trees, respectively.
- Damaged polygons must be acquired with all requested metadata (date of the beginning of the outbreak, date at which the polygon was recorded with the APP, outbreak stage at the recording time).
- Damaged polygons must be collected in stage 2 and 3.
- Both damaged and healthy polygons must cover areas larger than a few pixels at 10 square meters of resolution to allow us to leverage spatial correlation information and use spectral-spatial features for the model development
- Healthy polygons must regard spruce areas covered by the same type of forest observed in damaged areas.
- Healthy polygons must regard “old” spruce forest areas as bark beetles typically affect old spruce trees.
- Both healthy and damaged polygons must be collected at different timestamps between May and September, as well as in the different geographic areas (and multiple countries). However, it is crucial to also collect examples of healthy polygons and damaged polygons in the same area and at the same time.
- Both healthy and damaged polygons must refer to outbreaks occurred in the last two years. It is mandatory to obtain a high number of polygons in 2025 to explore the model development in the last year.

Final recommendations regard the need of retraining models over time and space to fit them to potential concept drifts occurring both in time and space.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Windthrow Model Development Update

Windthrow, defined as the uprooting or breaking of trees by wind events, represents a major form of forest disturbance across temperate and boreal ecosystems. It has significant ecological, economic, and operational consequences, particularly in the context of climate change, which is projected to increase the frequency and intensity of severe windstorms [Seidl et al., 2017]. Monitoring windthrow is thus crucial to improve response capabilities, assess damage, and calibrate predictive risk models.

Model Description

Literature Review Update – Windthrow Monitoring

This literature review summarizes recent advances in windthrow monitoring, with an emphasis on remote sensing techniques, integration with modeling frameworks, and operational applications.

Recent advances in remote sensing have enabled more reliable and scalable approaches to monitor windthrow events, particularly through the fusion of optical and radar satellite data. Numerous studies have investigated the use of Sentinel-1 and Sentinel-2 for forest disturbance mapping.

For example, Schiefer et al. (2020) employed dense time series from both sensors to detect storm damage across Central Europe, advocating for change detection approaches based on anomaly persistence rather than isolated values. Similarly, Oeser et al. (2021) explored Sentinel-1 coherence to map forest disturbances but noted its limitations in areas with high topographic complexity or persistent cloud cover.

However, purely optical approaches like those presented by Dalponte et al. (2021) often struggle to differentiate between logging and windthrow, especially in mono-temporal applications. Optical indices such as NDVI and NBR saturate in dense forests and are vulnerable to cloud contamination, making multi-sensor fusion essential.

Time-series models such as CCDC-SMA (Zhu & Woodcock, 2012; Kennedy et al., 2018) offer robust disturbance detection using long-term trends and breakpoints, but their computational cost and parameter tuning make them challenging to deploy across large European regions. To address these limitations, the SWIFTT model relies on a user-defined windstorm date and a delta-composite approach. This method narrows the detection window to six weeks pre- and post-event, significantly reducing false positives related to seasonal dynamics or unrelated disturbances (e.g., harvesting, pest outbreaks).

The use of SAR data, particularly Sentinel-1, offers valuable structural information insensitive to clouds or illumination. RVI and RFDI have been shown to effectively capture canopy disruption, as demonstrated in forestry applications by Albughdadi et al. (2021) and Kuntz et al. (2021). Moreover, combining radar with vegetation indices like NDMI or EVI improves classification accuracy, especially when disturbances reduce biomass but do not fully expose bare soil.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

However, single-event classification methods such as those developed by Dalponte et al. (2021) face limitations in distinguishing windthrow from other disturbances like logging or disease outbreaks. Unsupervised time series models (e.g., CCDC-SMA, Zhu & Woodcock, 2014) offer potential, but they are highly computationally intensive and can suffer from false positives. Therefore, we constrained our detection strategy to known storm dates provided by the client to improve reliability and reduce false alerts.

To conclude this review, the combination of radar and optical indices has proven effective. SAR-based indices like RVI and RFDI offer sensitivity to forest structure loss, while optical indices such as NDVI and NDMI respond to vegetation health and moisture content (Ullah et al., 2021; Singh et al., 2021). Studies from both Europe and global applications (e.g., Jakubowski et al., 2021; Paluba, 2023) have shown that data fusion improves classification performance under varying atmospheric and canopy conditions.

From Review to Model Design

From this review, key limitations emerged that guided our model development. First, reliance on time series alone is computationally demanding and lacks user control. Second, single-index approaches perform poorly when distinguishing windthrow from harvesting. Lastly, cloud contamination in optical data remains a challenge in temperate zones.

To address these, we designed a flexible, generalizable model based on a small set of radar and optical features, applied around known storm dates. The use of Sentinel-1 for structural changes, fused with Sentinel-2 vegetation indices and Dynamic World land cover, provides a compact yet powerful feature set. This enables high F1-score detection within a 3-month post-event window while maintaining broad European applicability through GEE.

Algorithmic Framework and Implementation

A fundamental distinction of the SWIFTT pipeline, compared to most academic approaches reviewed earlier, lies in its explicit objective to produce a commercially viable product. As such, the methodology has been intentionally optimized to minimize computational cost—both during training and inference—while maintaining acceptable levels of accuracy. In practice, this implies favoring server-side processing (via Google Earth Engine) and restricting the number of model features and algorithmic complexity to reduce the risk of overfitting. While this may initially appear as a constraint, the emphasis on a streamlined, interpretable pipeline ultimately enhances the model's generalization capability and scalability across large territories. This choice has also required particular care in preprocessing the satellite imagery and in selecting only the most informative vegetation and structural indices to capture relevant changes in forest condition due to storm events.

Principles and Computational Infrastructure

The entire pipeline is implemented within the cloud-native environment of Google Earth Engine (Gorelick et al., 2017), a platform that allows direct access to multi-temporal satellite archives, such as Sentinel-1 and Sentinel-2, without the need to download or host data locally. Through

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

its JavaScript and Python APIs, GEE provides server-side operations for image filtering, compositing, masking, and classification. This architecture is particularly well-suited to large-area environmental monitoring applications, where computational efficiency and data accessibility are critical. A Python wrapper orchestrates the logic and manages preprocessing, model inference, and exportation of results.

Overview of Processing Pipeline

The windthrow classification pipeline follows a streamlined sequence of operations, balancing computational efficiency and geospatial robustness. The workflow begins with the application of a forest mask derived from the Hansen Global Forest Change dataset (Hansen et al., 2023), ensuring that classification is restricted to areas with sufficient canopy cover (threshold set at 30% tree cover in the year 2000). Once the forested extent is delineated, Sentinel-1 and Sentinel-2 satellite imagery are pre-processed separately following advanced correction protocols as explained below.

Sentinel-1 Preprocessing and Index processing

Radar imagery from Sentinel-1 is particularly valuable for windthrow detection due to its sensitivity to forest structure, canopy roughness, and dielectric properties—features that are directly affected when trees are uprooted or broken. To maximize the reliability of SAR data, the preprocessing workflow adapts the advanced Sentinel-1 ARD (Analysis Ready Data) pipeline published by Adugna et al. (2021), implemented in the module gee_s1_ard. This routine includes five critical correction stages:

1. Radiometric calibration converts raw backscatter to sigma-naught (σ^0), in linear scale, enabling consistent inter-image comparison.
2. Border noise correction, based on ESA guidelines (Filipponi, 2019), removes bright edge artifacts that can bias signal interpretation in edge-of-swath regions.
3. Speckle filtering is applied using a 15×15 moving window and a temporal kernel of ten images. The multi-temporal boxcar filter (Lopes et al., 1993) is selected to balance noise reduction and spatial resolution preservation.
4. Terrain flattening is performed under the Volume model using the SRTM DEM, correcting for geometric distortions in mountainous areas due to layover and shadow effects (Small, 2011).
5. Geometric masking excludes highly sloped pixels or shadowed areas through an additional buffer.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

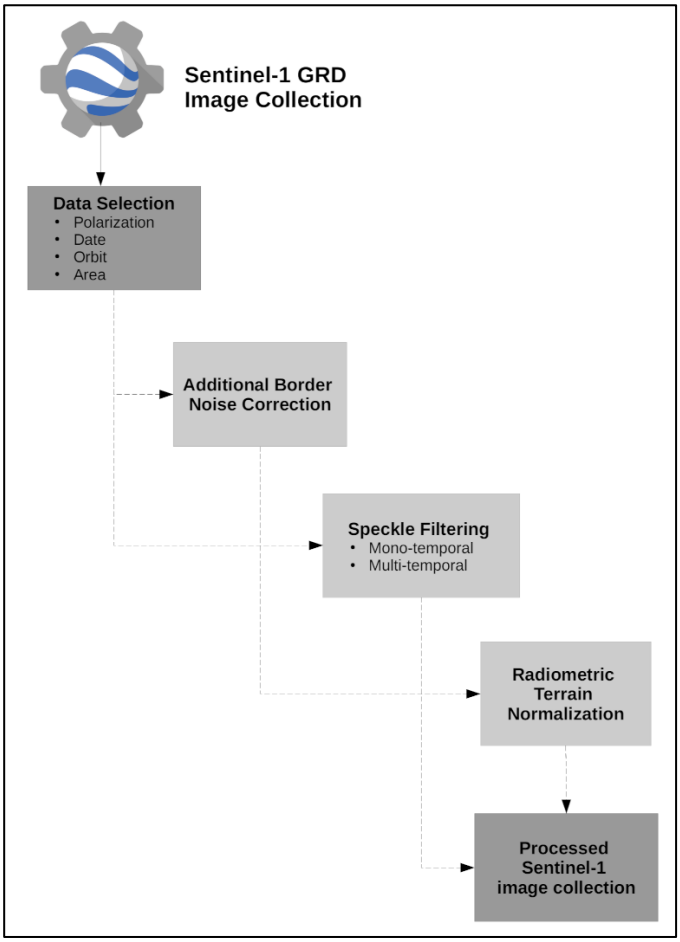


Figure 13: Corrected Sentinel-1 preprocessing pipeline adapted from Adugna et al. (2021) used in our algorithm

This robust sequence ensures that Sentinel-1 VV and VH bands are radiometrically and geometrically harmonized before further analysis.

In addition to VV and VH backscatter channels, two radar-derived indices are used to characterize structural changes the modified Radar Vegetation Index (mRVI) and Modified Radar Forest Degradation Index (mRFDI). Sentinel 1 does not provide full quad-pol SAR so mRVI and mRFDI adapt the RVI and RFDI for VV/VH dual-pol data while still provides a meaningful vegetation structure information.

- **Modified Radar Vegetation Index (mRVI)**:

$$mRVI = \frac{4 \times \sigma_{VH}}{\sigma_{VV} + \sigma_{VH}}$$

Where,

σ_VV = Backscatter in Vertical-Vertical polarization

σ_VH = Backscatter in Cross-Polarization (Vertical transmit, Horizontal receive)

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Reflects canopy volume and heterogeneity in scattering properties. Higher values generally correspond to dense, intact forests.

- **Modified Radar Forest Degradation Index (mRFDI)**:

$$mRFDI = \frac{\sigma_W - \sigma_{VH}}{\sigma_W + \sigma_{VH}}$$

Highlights degradation or deforestation. Values above 0.5 are typically indicative of severe canopy loss (Watanabe et al., 2018).

These indices are computed after converting amplitude data to decibels and are particularly suited for post-storm applications where canopy disaggregation is a key signal.

Index	Key Focus	Main Forestry Applications	Value interpretation
mRVI	Forest density & biomass	Forest mapping, biomass estimation	High values → Denser forests
mRFDI	Forest degradation	Deforestation, degradation monitoring	High values → More degradation

Table 20: Overview of the SAR modified vegetation indices and their interpretation

Sentinel-2 Preprocessing and Vegetation Index Computation

The optical counterpart is drawn from Sentinel-2's surface reflectance product (COPERNICUS/S2_SR_HARMONIZED), harmonized across platforms and corrected for atmospheric effects. To exclude cloud-contaminated pixels, we leverage the Cloud Score Plus dataset (GOOGLE/CLOUD_SCORE_PLUS/V1) and apply a conservative threshold of 60%. Importantly, masks are computed based on mosaics derived from 6-week pre- and post-event periods to identify and exclude consistently cloudy regions.

Subsequently, a suite of vegetation indices is computed per pixel, each designed to capture a specific biophysical parameter:

- **Normalized Difference Vegetation Index (NDVI)**:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

Indicates vegetation greenness and photosynthetic activity.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

- **Enhanced Vegetation Index (EVI)**:

$$EVI = G \times \frac{(NIR - Red)}{(NIR + C1 \times Red - C2 \times Blue + L)}$$

With:

G = gain factor (2.5 by default in our case)

C1, C2 = atmospheric correction coefficients

L = canopy background adjustments

The EVI emproves sensitivity in high-biomass areas and corrects for canopy background and atmospheric effects.

- **Normalized Burn Ratio (NBR)**:

$$NBR = \frac{(NIR - SWIR2)}{(NIR + SWIR2)}$$

The primary usage of NBR is to detect stress or fire-related damage., but it has been as a proven efficiency for different forest disturbances.

- **Normalized Difference Moisture Index (NDMI)**:

$$NDMI = \frac{(NIR - SWIR1)}{(NIR + SWIR1)}$$

Estimates canopy water content and drought stress.

Index	Key Focus	Main Forestry Applications
NDVI	Vegetation greenness	Forest health, biomass estimation, deforestation monitoring
EVI	Vegetation health (better for dense forests)	Monitoring seasonal growth, avoiding soil/atmospheric effects
NDMI	Vegetation moisture	Fire risk assessment, drought stress analysis
NBR	Burn severity	Fire impact assessment, post-fire recovery

Table 21: Optical vegetation index used by the model.

Only a selected subset (NDVI, EVI, NDMI, NBR) is retained in the final feature stack, chosen for their proven correlation with windthrow patterns (Fraser et al., 2005; Wang et al., 2010).

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Following preprocessing and computing spectral and structural indices a temporal change is assessed using a compositing approach based on median pixel values across a six-week window immediately following the storm event, and compared to the same period one year prior. This phenologically-aware differencing strategy reduces seasonal noise and isolates structural breaks attributable to wind damage. In parallel, pixel-level standard deviation is calculated over the post-storm period to capture local spectral heterogeneity, often a signature of partial damage or fragmentation.

In addition, the Dynamic World (Brown et al., 2022) dataset is integrated to enrich land cover context. The "trees" class, representing the probability of tree cover, is extracted for the pre- and post-storm periods, and its delta is included in the composite stack. This auxiliary feature helps disambiguate vegetation disturbance types and reinforce confidence in forested pixel classification.

To minimize classification artifacts, a spatial smoothing step is then applied using a circular focal mean and modal filter. This removes isolated outliers and enhances spatial coherence in the classified output. Finally, the cleaned composite image is passed through a pre-trained binary decision tree classifier—optimized for minimal complexity and maximum generalizability—and the resulting classification is exported as a GeoTIFF.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

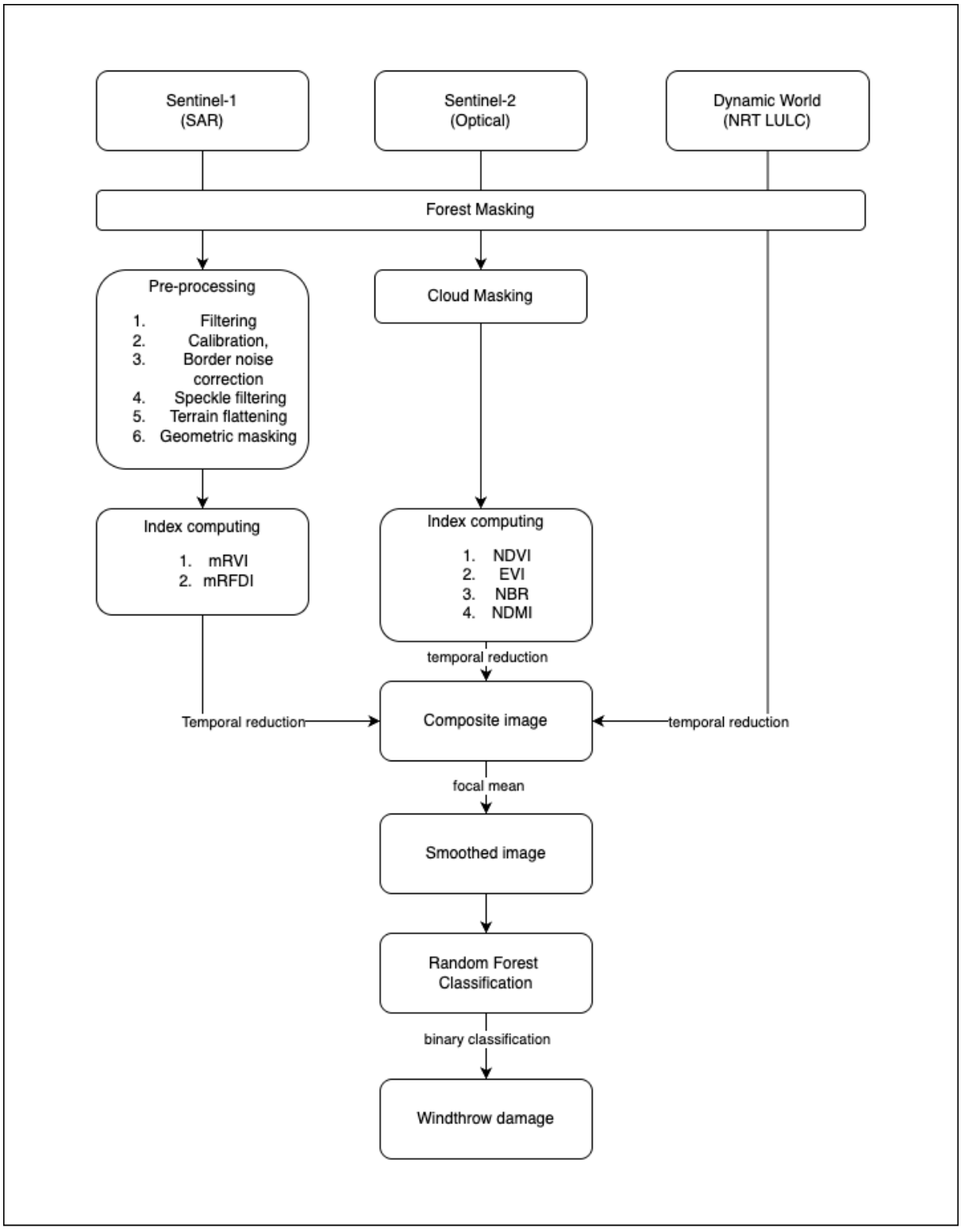


Figure 14: Diagram summarizing the pipeline of the windthrow model

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Methodology

Considering the windthrow, none has been yet recorded by our partner as of today (July 2025). To overcome of this issue, we have obtained private data from a partner in Ireland in the forestry sector following a major storm end of January 2025, Eowyn. The methodology proposed here is first train, test and tune the model on this data. Then in a second step, evaluate the model in another region, northern Italia which has suffering the storm Vaia in October 2018.

Model training, testing and tuning

Dataset description

The dataset has been built from field evaluation from our partner following the storm Eowyn on 24th January 2025. The initial dataset consists of 229 polygons classified as either *damaged* or *non-damaged* spread mostly in northern half of Ireland. It cumulates 5,657 ha in total, 1,324 ha damaged representing a potential of around 567,000 pixels of 10m resolution

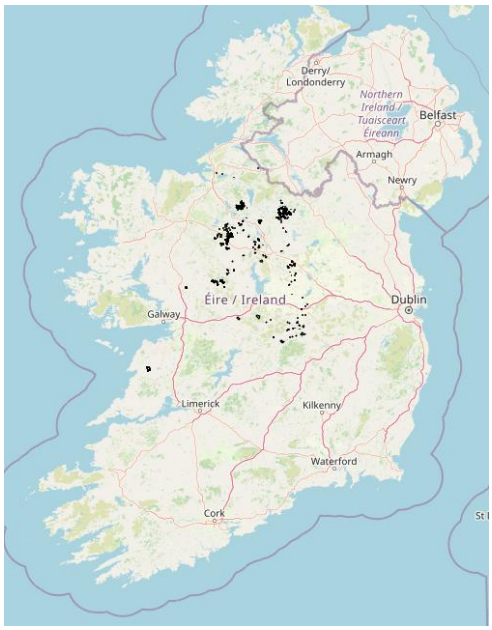


Figure 15: localisation of the field measurement following Eowyn

We can notice from the box plot some regions recorded only damaged area while other only healthy ones. While it appears unbalanced and unrealistic, it will be an interesting test for the model as in a real-world application the model should not learn from potential area damaged. Then, if the model performs well here we will be confident in his operational functioning.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

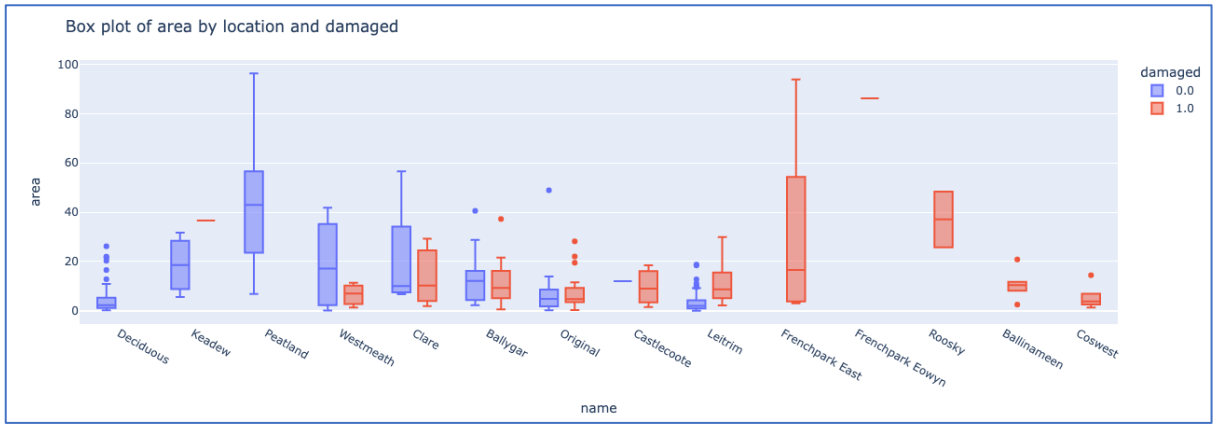


Figure 16: box plot representing the sum of area damaged and non-damaged by localisation.

To create a dataset readable for machine learning, we must combine the features from the composite image with the label (*damaged/non damaged*). To do so, we have used a stratified sample of 50,000 points from each label class, resulting in a 100,000 observation.

The final dataset is a dataframe of shape (100000, 10).

Before going any further and always in a care to make the model less complex as possible let's first have an overview of the correlation matrix:

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

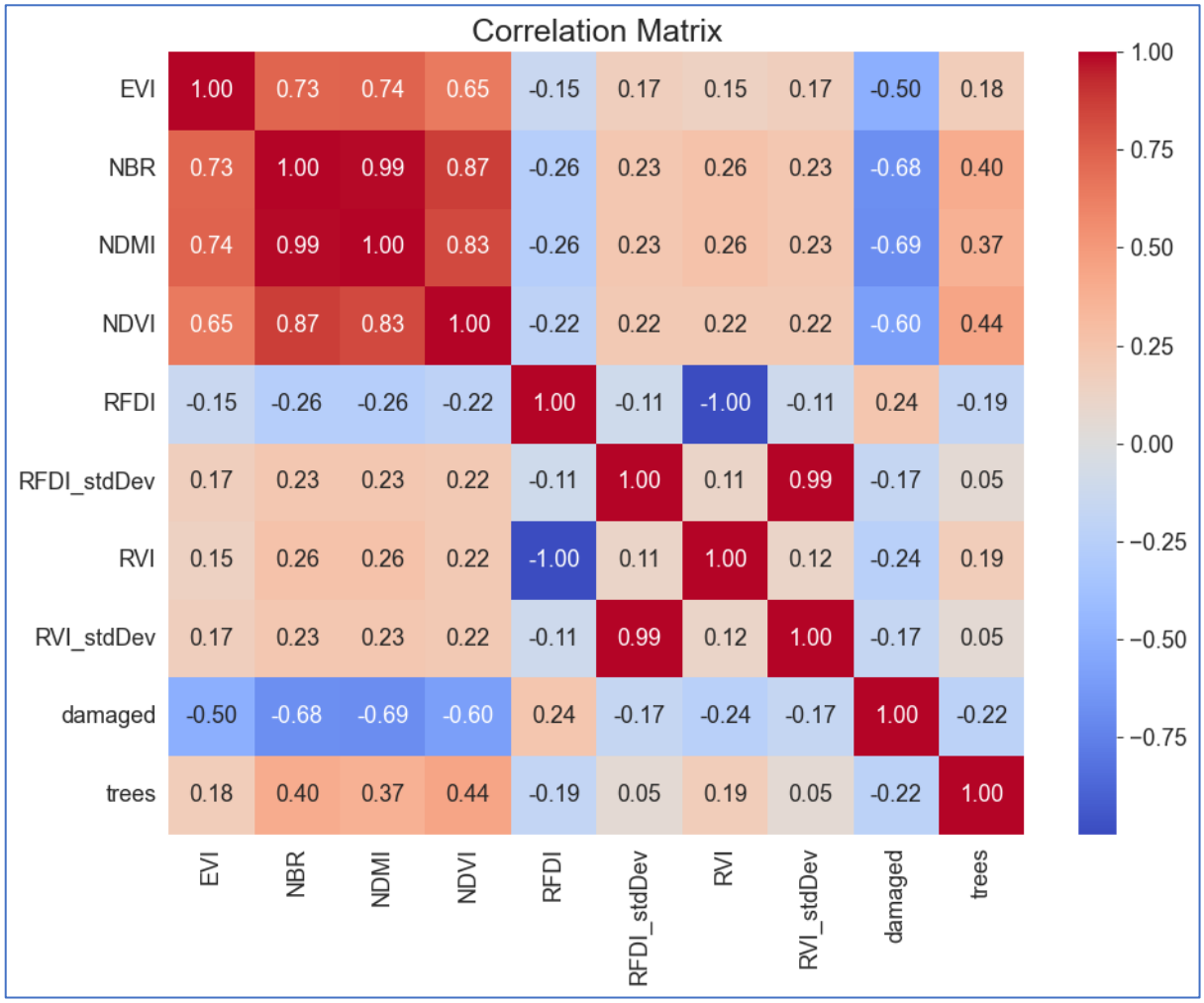


Figure 17: Correlation matrix (Pearson)

Even by using the Pearson correlation, getting mostly linear correlation between features, we can see the NBR is strongly correlated to other optical vegetation indices and will bring little added information to the label's variance. Similarly, the median of RVI and RFDI and the standard deviation are respectively 1 and 0.99. Therefore, we have made the choice to get rid of those features to keep only:

- Median
  - EVI
  - NDMI
  - NDVI
  - RFDI
  - Trees
- Standard Deviation
  - RFDI_stdDev

The final dataset is therefore compose of 100,000 observations, 6 features and one label.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Train, test and tuning the model

To get the most out of the output of the model in its possible application we have optimized the model following three constraints. Each time, we have tuned the model on unseen data. Therefore, out of the 100,000 observations, 60,000 was used for training based on a first custom score. 20,000 for the second constraint and the last 20,000 for the remaining one. Each subset was balanced, 50% damaged and 50% non-damaged.

First model: training on a custom score

To train the model we have conducted a grid search to optimize the following parameters of a RandomForestClassifier from the Sklearn library in python:

- `n_estimators`
- `max_depth`
- `min_samples_split`
- `min_samples_leaf`

We haven't found the need to use bootstrapping considering the large amount of balanced data and the criterion `entropy` was retained for is versality.

Finally, to mitigate the risk of overfitting, we have applied a 5-cross validation scores.

To select the best model out of it we have used a custom score:

$$custum\ score\ = \frac{abs(\ FP - FN)}{(TP\ + TN)}$$

with,

FP = False Positive

FN = False Negative

TP = True Positive

TN = True Negative

We have built this score to get a proxy of what could be the evaluation of a basis risk in insurance. The lower the score, the better the model. If we have high value for the denominator, it means we have high level of true predictions, while by evaluating the difference between false predictions we are encouraging the model to get as much as possible FP than FN, therefore even if some pixels are miss predicted, the total area damaged predicted will remain with a high level of accuracy.

We obtained then first encouraging results as displayed in the confusion matrix below (figure 6)

True 0	8824	1176
True 1	1035	8965
	Predicted 0	Predicted 1

Figure 18 Confusion matrix of the first trained model tested on 20,000 unseen data

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Leading to a F1-score of 0.89 and a ROC AUC of 0.96.

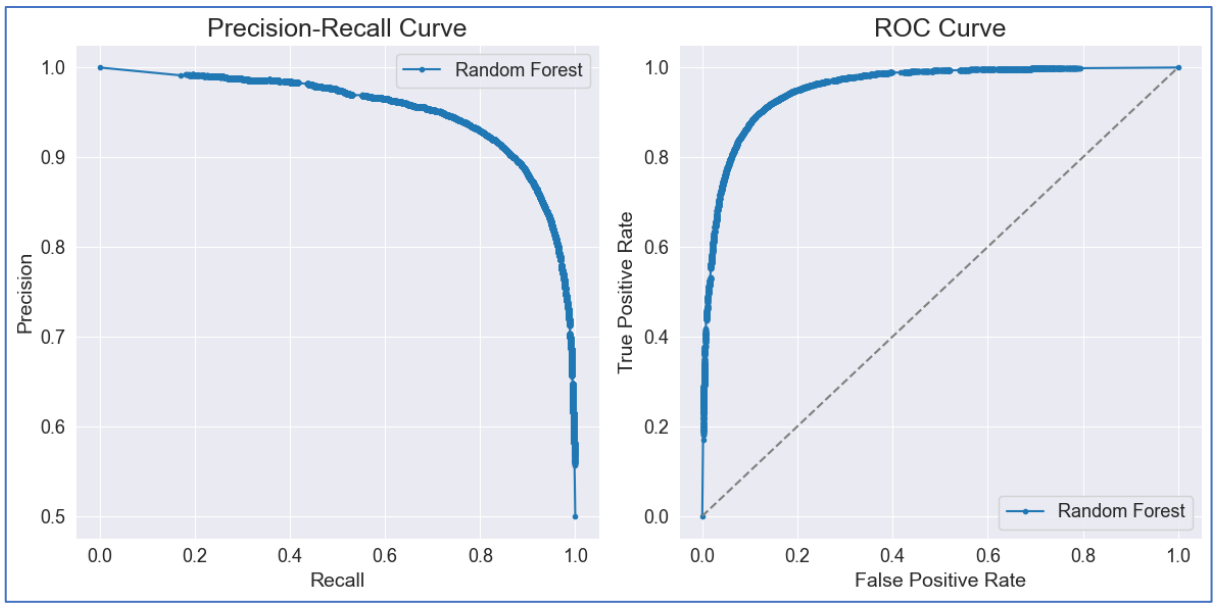


Figure 19: Precision_recall and ROC curves for the first model

Second model: tuning on optimized false positive/ false negative rate

Now before tuning the model on the F1-score, as it is the KPI to achieve for the SWIFTT project, we want to get adjust the best threshold. By default, in the SkLearn RandomForestClassifier, the threshold is set to 0.5. However, in certain condition moving this threshold can help to improve the model, even to get a balanced metrics such as the F1 Score. We have decided then to optimize this threshold to find the one minimizing the both the False Positive rate and False Negative Rate.

The results are shown in the graph below:

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

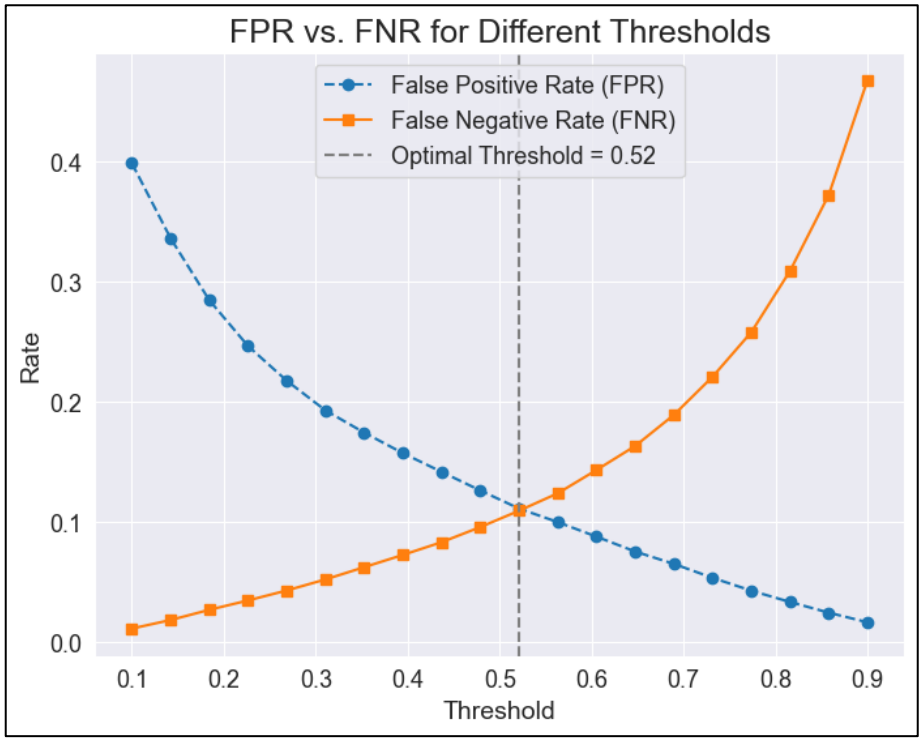


Figure 20: Optimal threshold to get a balanced FPR and FNR

We can see the optimal threshold is 0.52. We this applied we arrived to:

True 0	8966	1034
True 1	1196	8802
	Predicted 0	Predicted 1

Figure 21: Confusion matrix of the second model with a threshold at 0.52

With an F1 score of 0.89 and still a ROC AUC of 0.96

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

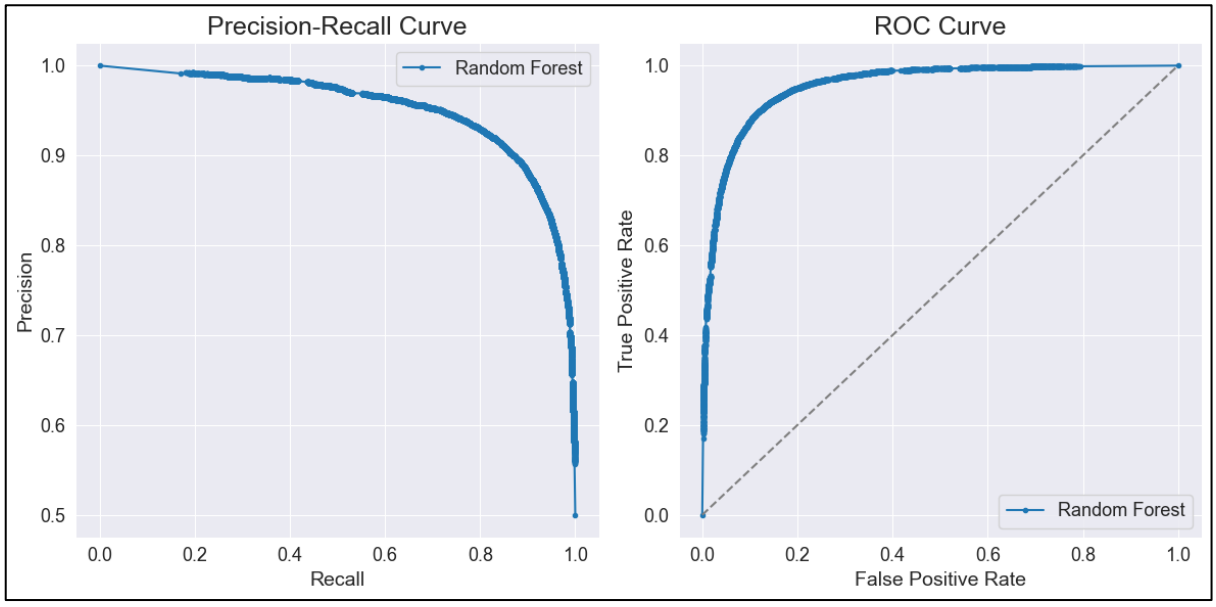


Figure 22: Precision_recall and ROC curves for the second model

This steps, while consistent in the approach doesn't have bring much adding value to the F1 score. However, we do now have an idea of the ideal threshold and for the last constraint we will optimized the F1-score around while moving the threshold around this point.

Final model: tuning on the F1 score around the optimal threshold value

We have then created a linear space between 0.48 to 0.56, meaning +/- 0.4 around the 0.52 optimal threshold with a 0.01 pace. While looking for that different value we will conduct one last time the GridSearchCV but using the F1-Score as the score to optimize.

In the final model we end up with:

True 0	8945	1055
True 1	1023	8977
	Predicted 0	Predicted 1

Figure 23: Confusion matrix of the final model

And a F-score of 0.90 with ROC AUC of 0.96

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

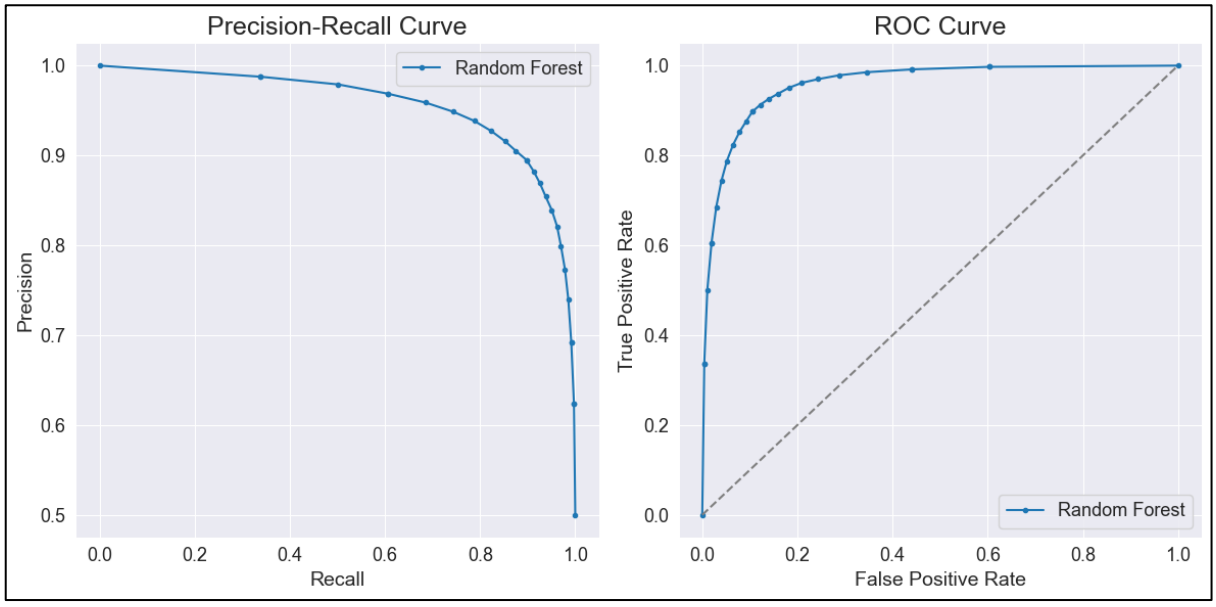


Figure 24: Precision_recall and ROC curves for the final model

This result is higher than expected. Indeed, the specification of SWIFTT ask to reach a F1 score of 0.75 with up to 3 months of imagery. Here we have a score of 0.90 with only 6 weeks of latency. We have then a better granularity and accuracy than the requirement.

We can also notice the number of FP (1055) and FN (1023) are extremely close to each other while getting 8945 TN and 8977 TP. Meaning the first custom score would be particularly low leading to a potential low basis risk, opening the door therefore to a potential insurance application.

Model Evaluation

As explained in the introduction of Methodology we have no data from forester in the consortium, but we manage nonetheless to find an external private partner providing us enough data to build a promising model with outreached result. Nonetheless we must consider the model has therefore been calibrated in one country, Ireland. Even if it has digested different types of forest and in different topographies, it would be much more comfortable to get an evaluation in a different country, with different vegetation, different topography and at a different time. This will be a strong test of the capacity of the model to generalized across Europe.

Methodology

To operate in this direction, we have used a subset of the FORWIND dataset (Forzieri et al.2019). The dataset also referenced as *"A spatially explicit database of wind disturbances in European forests over the period 2000–2018"* provides detailed geographical data on windstorm impacts across Europe. Spanning over 80,000 disturbance areas, it includes information from major wind events such as Gudrun, Kyrill, Klaus, Xynthia, and Vaia. The dataset, available in a harmonized vector format, is crucial for studying forest resilience, assessing the impact of wind disturbances, and supporting forest management strategies. It offers valuable insights into the temporal and spatial patterns of wind-related damages in European forests over nearly two decades. In our case, we will focus on Vaia, as it has

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

happened in October 2018, so sentinel imageries are operational. Other storms are too old to be of any used in our case. Plus, this is an interesting choice as it has had an impact in a different country (Italy), in a different period from Eowyn, respectively in October and February, in a very steep mountainous region, therefore challenging the limit of sentinel 1 application due to shadow casting. It will be a strong test of our sentinel one preprocessing pipeline.

Presentation of the Area Of Interest for the evaluation

One limitation of the FORWIND dataset is there are proposing only damaged area, not untouched and unfortunately, they are also mentioning it is not an extensive dataset, meaning it is not the damaged polygon is not in the dataset, there was no damaged. This is making our testing challenging as we need both value of the label. We have then, check the surroundings and inferred them as un-damaged. The figure below displays the polygons damaged from the initial dataset(white) and undamaged ones inferred from satellites imageries (black).

Figure 25: Evaluation label dataset, in white damaged area, in black undamaged surroundings.

We have then 2,014 ha of damaged out of 8,347 ha.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Results

Before applying the classifier, we have had first to apply all the preprocessing stages as explain in the Figure 14: Diagram summarizing the pipeline of the windthrow model. This is again crucial to make the model more robust, which is particularly true here in a mountainous region required therefore powerful algorithm for border noise reduction and terrain flattening before using SAR images.

When we have the composite image serving as inputs for the classifier we can run the change detection algorithm.

Again we have selected a stratified point sampling of 10,000 points of each label, resulting in 20,000 observations, as we have done in the model tuning phase. Similarly as before, let's have a look on both the confusion matrix and the F1-score to evaluate the model performance

True 0	8313	1687
True 1	1137	8863
	Predicted 0	Predicted 1

Figure 26: confusion matrix of the model evaluation on Vaia storm

This is leading to a F1 score of 0.86. This score indicates a very good performance of the model to predict damaged pixel in unseen and different area across Europe, and it is still much higher than the 0.75 required by the project specifications.

Let's have a look on the visualisation

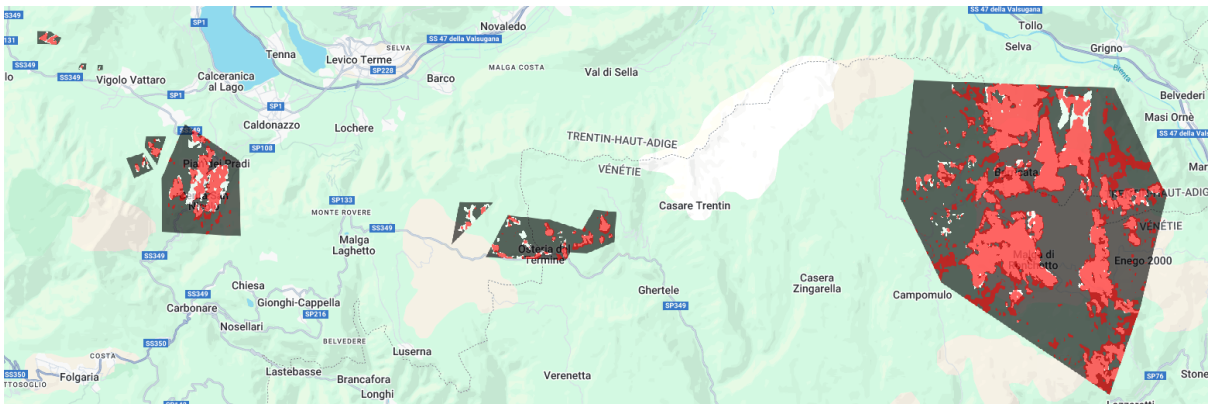


Figure 27: Visualisation of the predicted damaged (red) and the ground truth (white)

Overall, we can observe the prediction are quite close to the ground truth from FORWIND dataset. Most if the time we predicted a larger surrounding damaged on the border which can be due to the smoothing steps, and we are missing some true positive area. We can notice two areas where it seems to fail particularly. Let's deep dive into those.

Sub-area 1 evaluation:

Let's design the below area as AOI1:

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

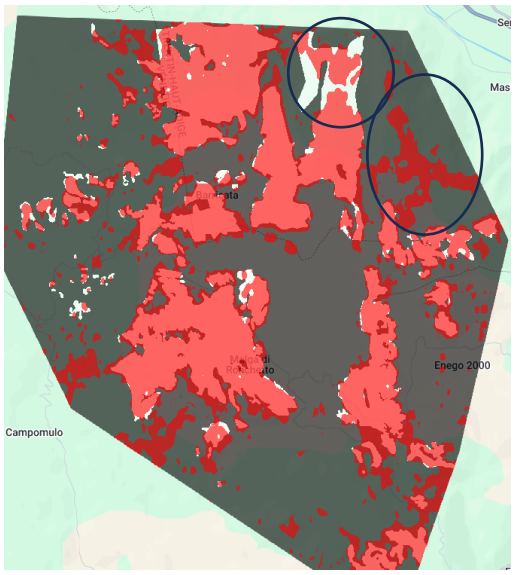


Figure 28: AOI1, zoom of figure 15

On this figure it appears quite clearly the predicted areas are often a bit larger, but let's have a focus on the two-area circled where the model seems to fail by either false positives or false negatives.

To get a high-quality image with a temporal activity we have used Google Earth Pro and here what we found thanks to images before and after the storm.

Figure 29: On the left image on 10/2017 before the storm. On the right in 10/2019 after the storm. In red outlined are predicted damaged.

We cannot help but noticed the model seems to have done surprisingly do great here and mention that the FORWIND dataset has been built by multiple sources (aerial and satelliute interpretation) therefore there is high chance the "ground truth" used for the evaluation of the model is not as good as it should.

Sub-area 2 Evaluation:

Same methodology:

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

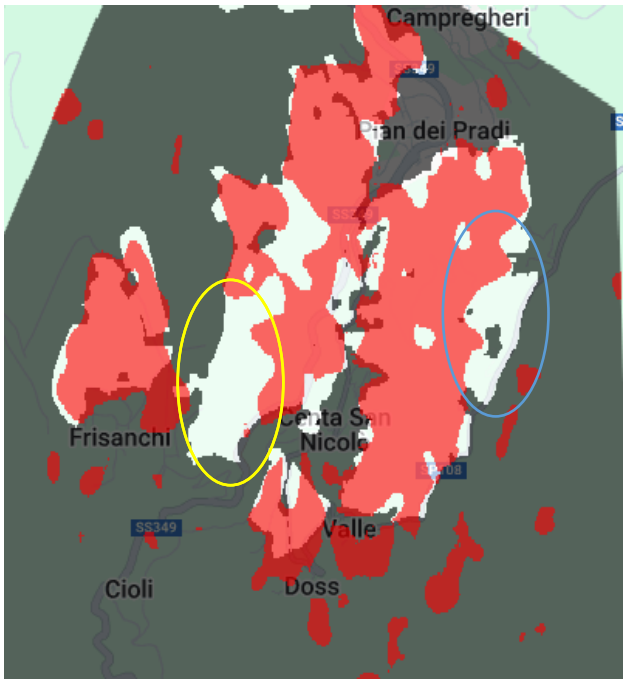


Figure 30: Zoom on the sub AOI2. In red predicted damaged. White ground truth.

Again, let's have a look on previous/after storm imagery and see if the errors is due to poor model prediction or again can be attribute to a poor data quality in the ground truth label.

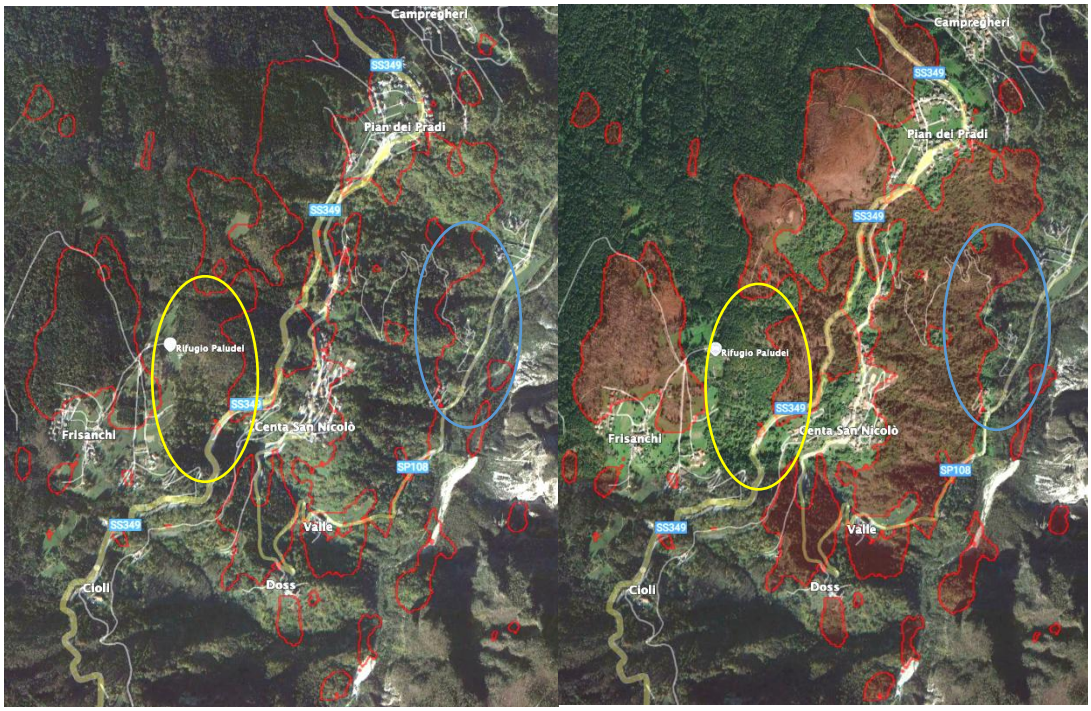


Figure 31: Satellite imagery in October 2017(left) before the storm and after in October 2019 on the left. In red outlined, the predicted damaged

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

In the yellow circle, it seems a priori the model miss some damaged area, but by looking at the imagery before the storm, we can notice the forest was already disturbed and it is not to be attributed to the storm. In the blue circle it is simply not damaged at all.

All this drives confidence in our model prediction. Even with a moderate good quality of the label we can end up to 0.86 F1 score, leaving us the possibility to achieve higher score.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Lessons learned

The development and testing of the SWIFTT model provided several key insights into both the capabilities and limitations of the approach. The model demonstrated a robust performance in terms of detection accuracy, achieving an F1 score of 0.90 on training data, and 0.86 on unseen validation data from the Vaia storm, which is substantially above the required threshold of 0.75 for the SWIFTT project. The following lessons emerged from this process:

1. **Multi-Sensor Data Fusion Improves Accuracy**

   The integration of radar and optical data, particularly Sentinel-1 SAR data with Sentinel-2 optical vegetation indices, significantly enhanced the model's ability to detect forest disturbances caused by windthrow. This fusion of data types helped reduce the impact of cloud cover, which is a significant challenge in optical-only approaches. Furthermore, SAR data provided valuable structural information that was crucial for detecting changes in forest canopy caused by wind damage. The model's performance demonstrated the effectiveness of combining indices like the Modified Radar Vegetation Index (mRVI) and Modified Radar Forest Degradation Index (mRFDI) with vegetation indices such as NDVI and NDMI (Albughdadi et al., 2021; Ullah et al., 2021).

2. **Temporal Differencing Reduces Seasonal Noise**

   The use of temporal change detection, particularly through a phenologically-aware differencing strategy, played a vital role in isolating disturbances attributable to windthrow from seasonal variations. This method, which involved comparing pre- and post-storm periods, successfully reduced the influence of seasonal noise, which is a common challenge in forest disturbance detection. This approach was particularly effective in regions with distinct seasonal patterns and variable atmospheric conditions, such as the European areas under study (Fraser et al., 2005; Wang et al., 2010).

3. **Model Generalization Across Different Regions**

   Although the model was initially calibrated using data from Ireland, its successful application to the Vaia storm dataset in Italy demonstrates the model's potential for generalization across different European regions. Despite the challenges posed by topographical variations and differing forest types, the model showed considerable robustness, indicating its potential for operational use across diverse geographies. However, further testing in a broader range of regions and forest types is necessary to ensure that the model's performance remains consistent under various environmental conditions.

4. **Challenge of Ground Truth Data Quality**

   One of the major challenges encountered during model evaluation was the quality of the ground truth data. The FORWIND dataset, used for model evaluation, provided only damaged areas and inferred non-damaged regions, which may not have always accurately reflected the true status of the forest cover. In some cases, the ground truth data may have been influenced by inconsistent interpretation methods, as the dataset combined various sources, including satellite and aerial imagery. This discrepancy

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

highlights the need for high-quality, consistent ground truth data in future testing (Forzieri et al., 2019).

5. **Balancing Model Complexity and Computational Efficiency**

   The decision to limit the number of features and use a simpler algorithm, such as the Random Forest classifier, was critical in ensuring that the model could operate efficiently on large datasets while maintaining an acceptable level of accuracy. The use of Google Earth Engine (GEE) for server-side processing further enhanced computational efficiency, allowing the model to scale across large European regions without excessive computational costs. This emphasis on simplicity and efficiency is particularly important in operational contexts where model deployment needs to be both fast and scalable (Gorelick et al., 2017).

6. **Threshold Optimization**

   Tuning the classification threshold to minimize both the false positive rate (FPR) and false negative rate (FNR) was a crucial step in optimizing the model's performance. By adjusting the threshold to 0.52, we achieved a more balanced trade-off between false positives and false negatives, which helped to improve the model's operational applicability. This strategy, which focused on reducing potential bias in the prediction of windthrow damage, aligns well with the needs of operational applications such as insurance.

Recommendations for Future Testing

1. **Expand Geographical Coverage for Evaluation**

   To further evaluate the model’s generalizability, it is recommended to apply the SWIFTT model to additional regions across Europe, particularly in areas with varied forest types, topographies, and climatic conditions. This will help assess the model's robustness in diverse environments and provide more comprehensive feedback on its accuracy and scalability.

2. **Improve Ground Truth Data**

   For future evaluations, it is essential to obtain high-quality, consistent ground truth data that includes both damaged and non-damaged areas. This will ensure more accurate model validation and help reduce the risk of errors introduced by the inconsistencies in the current ground truth datasets. Hopefully, this project will fill this important gap.

In conclusion, the SWIFTT model demonstrates strong potential for operational windthrow detection in Europe, with the ability to adapt to different storm events, topographies, and forest conditions. Further testing and model refinement will be crucial for ensuring its long-term success and scalability in operational environments, such as forest management and insurance applications.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Fire Risk Model Development Update

This section provides an update on the development and evaluation of the fire risk model, a component within the SWIFTT monitoring system.

Model Description

A. Literature review update (optional)

Recent studies increasingly suggest that pixel-based remote sensing approaches for wildfire prediction, while visually detailed, often struggle to generalize and perform reliably when used in isolation. For example, Porta et al. (2025) used deep learning on Sentinel-2 and MODIS imagery for wildfire forecasting across Canada and reported F1-scores around 0.60 at 0.1° resolution—highlighting the limited performance gains from high-resolution imagery alone. Similarly, Shen et al. (2023) evaluated multiple self-supervised vision models on the FireRisk benchmark and showed that even advanced architectures failed to achieve reliable multi-class ignition prediction accuracy using satellite data alone.

In contrast, coarse-resolution, feature-based models are showing greater promise in both predictive skill and operational utility. Moghim and Mehrabi (2024) demonstrated that Random Forest models using climate, topography, vegetation, and human presence achieved significantly higher accuracy for monthly fire susceptibility mapping in Germany than NDVI-based or pixel-level inputs. Similarly, Gelabert et al. (2025) showed that combining land cover, accessibility, and fuel moisture anomaly yielded strong AUCs (0.70–0.89) across multiple European regions.

A recent review by Wang et al. (2023) concluded that while physical fire models remain critical for spread simulation, machine learning models with meteorological and human activity features are superior for ignition probability estimation, especially when paired with coarse-grained inputs such as FWI, fuel anomalies, and population density. These findings reinforce the rationale for the feature-based, interpretable approach used in the present study.

B. Machine Learning and Deep Learning methods used

Three different machine learning models were evaluated and compared for the task of monthly wildfire risk prediction over Europe. For model training, default parameters of the specified libraries were utilized, unless otherwise specified in the subsequent sections:

1. Random Forest (RF)

A baseline ensemble of decision trees trained using Gini impurity. The RF model provides robustness to noise, simple interpretability via feature importance, and strong performance with tabular data.

2. XGBoost (XG)

A gradient-boosted decision tree model optimized for speed and accuracy. XGBoost is widely used in structured prediction tasks and offers regularization to prevent overfitting, as well as handling class imbalance effectively via internal weighting.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

3. Multi-layer Perceptron (MLP)

A shallow neural network classifier with fully connected layers, using ReLU activation and a softmax output layer. While less interpretable, the MLP provides a non-linear alternative to tree-based models and may capture different feature interactions.

4. Ensemble Model (Soft Voting)

An ensemble was built by averaging the predicted probabilities of the best RF, XG, and MLP models (see Figure 32). The soft voting strategy was chosen, to be able to output a combined predicted probability, if needed. This yielded a slight but consistent performance gain across most metrics and improved robustness.

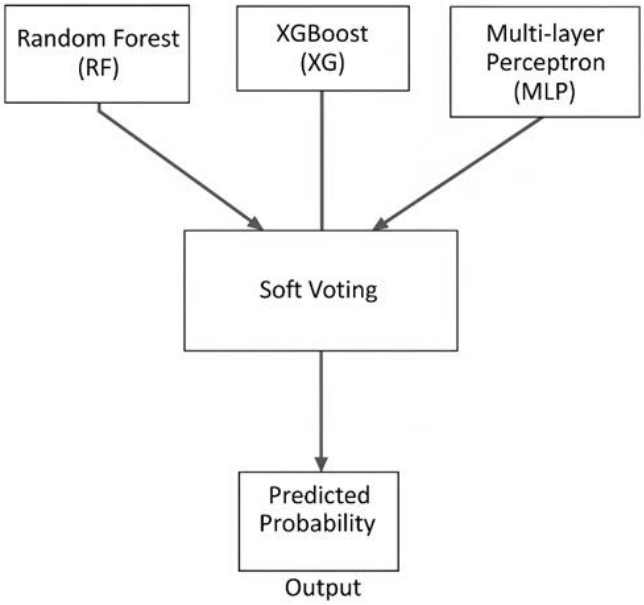


Figure 32: Diagram illustrating an ensemble soft voting model with RF, XGBoost and MLP as inputs

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Methodology

A. Selection of testers

The fire risk prediction task was framed as a binary classification problem: predicting whether at least one wildfire ignition would occur in a given 0.25° × 0.25° grid cell during a calendar month. Four machine learning models were evaluated as core "testers":

- **Random Forest (RF)** from *Scikit-learn 1.6.1*
- **XGBoost (XG)** from *XGBoost 3.0.0*
- **Multilayer Perceptron (MLP)** from *Scikit-learn 1.6.1*
- **Ensemble Model (soft voting)** combining RF, XGBoost, and MLP

These models were selected for their balance between interpretability, non-linearity, and practical performance in structured tabular data tasks. A standalone FWI-only Random Forest model served as a baseline for comparison.

To implement this baseline, a Random Forest model using only the Fire Weather Index (FWI) as input was trained. This minimalist configuration reflects a common operational assumption that fire occurrence correlates directly with climatological fire danger. The FWI values were aggregated to the 0.25° × 0.25° spatial grid used throughout the study, with no additional information (e.g., land cover, human pressure, or topography) included.

The target labels were derived from the EFFIS historical wildfire dataset, which contains satellite-based fire polygons across Europe. These were intersected with the spatial grid at a monthly resolution, and each cell was labeled as positive if it contained at least one fire event during a given month, and negative otherwise. This binary label definition was applied consistently across all model variants.

While the FWI-only model is easy to interpret and fast to compute, its performance was considerably weaker than that of the multi-feature models, achieving an F1-score of 0.225 and an ROC-AUC of 64.25%. This outcome underscores the limited standalone predictive power of FWI and highlights the importance of incorporating additional spatial and human-related features into fire risk models.

The models were trained and evaluated on a harmonized dataset (see Section B) covering continental Europe (excluding Ukraine, Belarus, and Moldova) with over 1.5 million monthly grid cell samples (2014–2024). Each model was trained using a fixed set of 15 engineered features, capturing spatial, meteorological, topographic, and anthropogenic influences (see Table 22).

Type	Features
Location	latitude, longitude
Weather	FWI, FWI_anomaly
Topography	elevation, slope, peak
Anthropogenic	Population density, Motorway

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Land Use	TCD, tcd_land_percentage, forest_and_shrub_percentage
Vegetation	FTY 1 (broadleaf), FTY 2 (coniferous), FTY 3 (mixed forest)

Table 22: A list of 15 features as input for the machine learning models

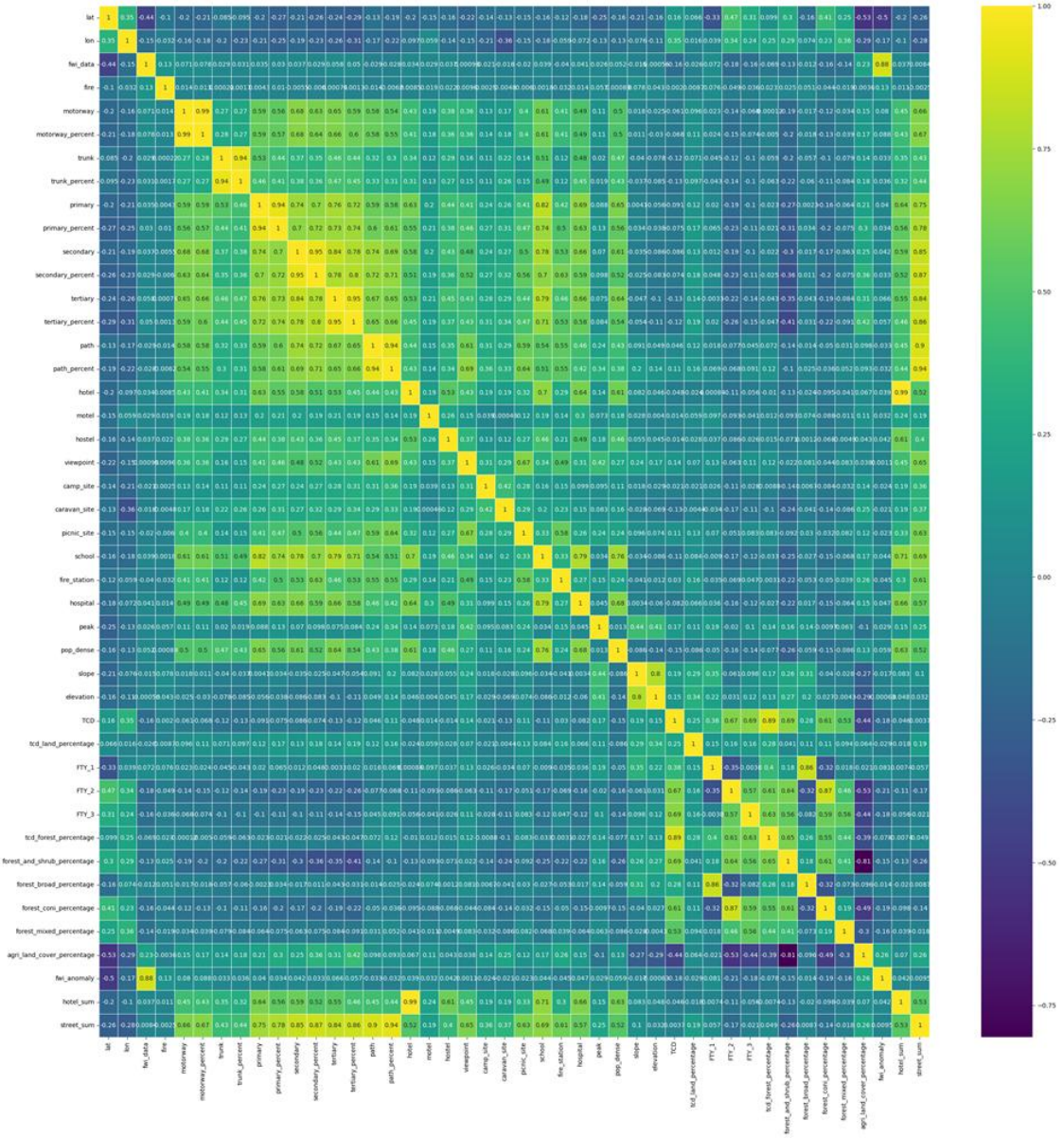


Figure 33: Full correlation matrix of all 43 features and the fire label data.

To deliver a model that is both accurate and genuinely useful to our primary customers, the foresters on the Swiftt platform, we undertook a crucial feature optimization process. We meticulously reduced our initial set of 43 diverse features (see Figure 33), sourced from detailed OpenStreetMap data and various, by Copernicus-services provided forest data maps, down to a focused set of just 15 (see Table 22). Notably, our analysis revealed that population density proved to be a particularly powerful and consolidating feature, effectively replacing the

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

predictive value of almost every other collected OpenStreetMap feature. This rigorous selection also prioritized features with low inter-correlation, ensuring each remaining variable provides unique and non-redundant information to the model. Critically, we emphasized selecting features that are intuitive and easily explainable, empowering foresters to understand the "why" behind fire risk predictions and build trust in the model's insights for their daily operational decisions.

B. Test phase execution

Temporal Data Splitting:

To ensure realistic generalization, data was split temporally:

- **Training**: 2014–2021
- **Validation**: 2022
- **Test**: 2023
- **Ground Check**: 2024

This allowed performance analysis on both unseen years and current events. No spatial leakage was introduced, and future data was never used during training.

Handling Class Imbalance:

Wildfires are rare events, leading to severe class imbalance. Of the roughly 1.5 million available monthly grid cells, only ~16,000 were labeled positive (at least one fire ignition in that cell and month). To manage this:

- Random sampling was performed with an 8:1 ratio of non-fire to fire samples to effectively address the severe class imbalance while retaining sufficient information from the majority class
- No cross-validation was used to avoid temporal leakage. Instead, five random seeds were applied to each train/test split and final metrics were averaged.
- Final performance metrics were averaged across these five runs

This strategy balances precision and recall while ensuring robustness against sample noise.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

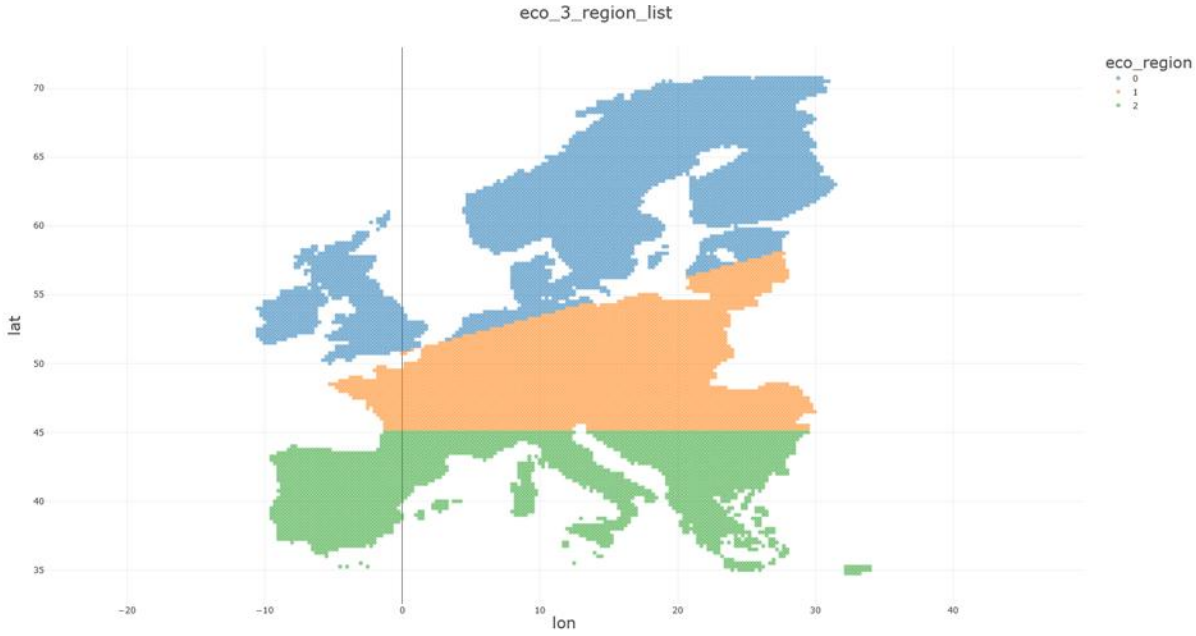


Figure 34: European grid structure split in specific Ecoregions (blue = Northern, orange = Middle, green = Mediterranean)

Spatial Evaluation Strategy:

In addition to continent-wide evaluation**,** three ecoregion-specific (Figure 34) models were trained to explore the hypothesis that regional specialization might improve performance. Each model was trained on a subset of the data corresponding to its respective ecological zone in mainland Europe.

Toolchain:

All experiments were conducted in Visual Studio Code, using:

- **Pandas 2.2.3**, **NumPy 2.2.5**, **GeoPandas 1.0.1** for data processing
- **Scikit-learn 1.6.1** for Random Forest and MLP models
- **XGBoost 3.0.0** for gradient boosting
- **Matplotlib 3.10.1** and **Seaborn 0.13.2** for visual analysis (used in results section)

Output Format:

The model's initial output for evaluation is a binary (0 or 1) value, where zero indicates a low chance of fire and one indicates a high chance of fire for a specific grid cell during a given month. However, while model evaluation focused on this binary classification, the real-world output of the Wildfire Forecast is intended to be a continuous probability score. This probability can then be mapped to up to five discrete fire danger levels (e.g., "Low, Moderate, High, Very High, Extreme"). For instance, probabilities from 0.0 to 0.2 might be 'Low', 0.2 to 0.4 'Moderate', 0.4 to 0.6 'High', 0.6 to 0.8 'Very High', and 0.8 to 1.0 'Extreme'. These danger levels and their associated probability intervals are currently illustrative examples; their precise

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

values will be defined during the operational calibration phase. This mapping enables clearer communication and operational use, in alignment with systems like EFFIS or local fire alert systems.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Results

A. Results and interpretation

The SWIFTT KPI for the fire model asks to reach an ROC-AUC score greater than 0.7. To assess the performance of the new *Forest Fire Forecast* model, several machine learning methods were evaluated: Random Forest (RF), XGBoost (XG), Multi-Layer Perceptron (MLP), and a soft-voting ensemble of these models. All were trained to classify monthly fire occurrence (binary) on 0.25°×0.25° grid cells across Europe using a feature set of 15 variables, including meteorological indicators (e.g., FWI), anthropogenic factors (e.g., population density, infrastructure), and geographic features (e.g., slope, elevation).

Model performance was measured using ROC-AUC, F1-score, precision, recall, and accuracy, averaged over five random seeds (no cross-validation). Table 23 summarizes the results for both the Europe-wide general model and the eco-region-specific models (In order: Northern / Middle / Mediterranean). For improved readability and focus on core trends, standard deviations across the five runs have been omitted from the table.

Model	ROC-AUC [%]	F1-score	Recall / Precision / Accuracy	Eco-Region ROC-AUC [%]
FWI-baseline RF	64.25	0.225	0.189 / 0.276 / 80.1%	67.32 / 52.39 / 53.39
Random Forest	92.77	0.579	0.460 / 0.784 / 89.8%	93.21 / 89.52 / 84.94
XGBoost	93.39	**0.668**	**0.791** / 0.578 / 88.0%	92.57 / 88.86 / **87.69**
MLP	92.79	0.532	0.395 / **0.812** / 89.4%	93.25 / 86.11 / 86.19
Ensemble (soft)	**93.79**	0.645	0.556 / 0.768 / **90.1%**	**93.53** / **89.63** / 87.21

Table 23: Performance comparison of classification models (Europe-wide and eco-region-specific ROC-AUC scores)

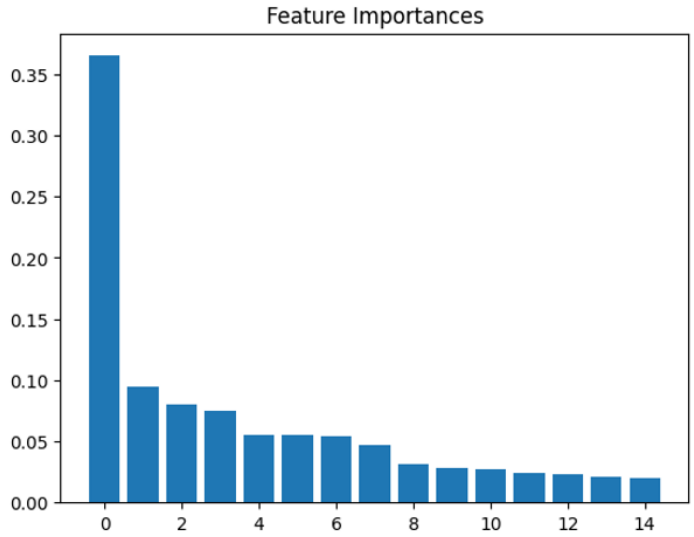
Among all evaluated models, XGBoost was selected as the final model due to its superior performance on the highly imbalanced fire occurrence task. It achieved the highest F1-score of 0.668, indicating the best balance between precision and recall, both crucial for minimizing missed ignition events and false alarms. While the ensemble model slightly outperformed XGBoost in overall ROC-AUC and accuracy, XGBoost's higher recall (0.791) and robust probabilistic output made it more suitable for practical deployment. The model was fine-tuned with a learning rate of 0.1 and a maximum tree depth of 5 to balance learning speed and generalization. To address class imbalance, the parameter scale_pos_weight was set to 8.0, roughly matching the 1:8 ratio of fire to no-fire samples in the training data. The binary logistic objective was used for classification, and model performance was monitored using the logloss evaluation metric.

The feature importance analysis for the best-performing XGBoost model (see Figure 35) highlights the most influential factors in predicting wildfire risk. Latitude emerged as the single most critical feature, contributing a significant 36.50% to the model's predictive power, underscoring the strong geographical patterns in fire occurrence across Europe. Longitude also played a notable role at 9.48%, further emphasizing the spatial distribution of risk. Meteorological conditions, primarily captured by the Fire Weather Index (FWI) and the "Peak"

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

value, were the next most impactful factors, with contributions of 8.03% and 7.51% respectively, reinforcing the established importance in fire prediction. Anthropogenic factors, represented by population density (5.55%), and topographic features like slope (5.49%) and elevation (5.39%), showed comparable and substantial influence. While all 15 features contributed to the model's accuracy, these top-tier variables demonstrate the model's reliance on a blend of geographical, weather-related, human, and terrain characteristics to accurately assess monthly wildfire risk.



Latitude	0.3650
Longitude	0.0948
FWI	0.0803
Peak	0.0751
Pop_dense	0.0555
slope	0.0549
elevation	0.0539
TCD	0.0465
tcd_land_percentage	0.0311
FTY_1	0.0279
FTY_2	0.0277
FTY_3	0.0238
FWI anomaly	0.0227
Motorway	0.0207
Forest_and_shrub_%	0.0202

Figure 35: Feature Importance of the best performing XGBoost model

While the specific ranking of features can vary slightly across different models, a consistent pattern emerged from our evaluations: latitude, longitude, FWI, and peak consistently ranked among the most important features across all evaluated models (Random Forest, XGBoost, and MLP). This strong and repeatable performance underscores the fundamental role of geographical location and fire weather conditions as primary drivers of wildfire risk, regardless of the underlying machine learning algorithm. Other features, such as population density, elevation, and land cover types, also maintained significant, though more variable, importance, complementing these core geographical and meteorological indicators.

A more granular analysis of the best-performing XGBoost model revealed an important threshold dynamic (see Table 24). By plotting the False Positive Rate (FPR) and False Negative Rate (FNR) against varying thresholds (see Figure 36), we identified an intersection point at a threshold of 0.364. At this specific threshold, the model exhibits a compelling trade-off: while overall accuracy, F1-score, and precision are observed to be lower compared to the threshold that optimizes those metrics, the recall significantly increases. In the context of forest fire monitoring, this heightened recall is particularly valuable as it implies a reduction in missed fire events (false negatives). For early warning systems, preventing even a single major wildfire often outweighs the operational costs associated with a higher rate of false alarms (lower precision). Therefore, while not maximizing aggregate performance metrics, this threshold of 0.364 presents a potentially usable configuration for scenarios, where minimizing undetected fire ignitions and prioritizing comprehensive coverage for proactive measures is paramount.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

	Normal Threshold at 0.5	Optimized Threshold at 0.364
Accuracy	**88.01 %**	85.77 %
F1-Score	**0.669**	0.648
Recall	0.791	**0.858**
Precision	**0.579**	0.520

Table 24:: Results of the threshold optimization for the best performing XGBoost model

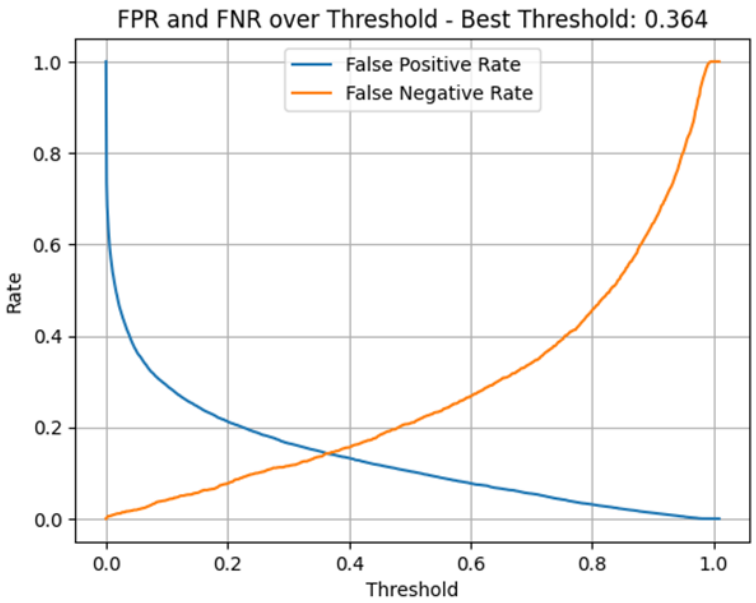


Figure 36: Threshold optimization for the best performing XGBoost model

B. Results limitations

Several limitations were identified during the model evaluation process:

- No true ground truth data: Fire occurrence is derived from remote-sensed EFFIS data. There is no official record of conditions leading to fire ignition, and many fire events remain undetected if too small or short-lived.
- Resolution mismatch: While the 0.25°×0.25° grid was appropriate for continental forecasting, many fire-related processes operate on a finer scale. Attempts to interpolate or upscale the weather features risk introducing noise and imbalance.
- Eco-region model weaknesses: Although the eco-region-specific models showed decent performance, especially in southern Europe, the model trained on northern regions achieved artificially high ROC-AUC scores due to low fire incidence—resulting

Funded by the European Union
Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

in systematic underprediction. As a result, the generalized model (with lat/lon as features) was preferred for its simplicity and robustness across Europe.

- Binary classification simplification: While the model was trained for binary classification, in practice the output will be converted into a 5-level fire danger scale, introducing another level of abstraction. The thresholding scheme for this translation remains to be finalized.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Lessons learned

A. Key learnings from model testing

1. Data Quality and Feature Selection Are Crucial

Deep learning models using only high-resolution imagery often fail to provide reliable wildfire ignition predictions due to noisy and indirect signals from vegetation indices or land use layers (Fusco et al., 2021). In contrast, machine learning models that incorporate weather variables, particularly the Fire Weather Index (FWI), along with anthropogenic features consistently achieve higher predictive performance (Di Giuseppe et al., 2022; Ruffault et al., 2020).

2. No True Ground Truth, only Fire Outcomes

Most wildfire datasets only log observed fires, without information on near-ignition events or suppressed fires. This absence of a complete ground truth introduces structural uncertainty into any classification task. It limits the model to approximating ignition likelihood from incomplete data, even with careful feature selection and preprocessing (Preisler et al., 2008).

3. FWI Is a Strong Predictor, but Its Resolution Is a Limitation

The FWI continues to serve as a robust and interpretable baseline across Europe. However, its resolution (typically 0.25° for ERA5 and 1.0° for forecast products) prevents its direct use in localized decision-making. Interpolating introduces noise and undermines performance rather than enhancing it (Bedia et al., 2018; Abatzoglou et al., 2018).

4. Simplicity and Generality Trump Regional Specialization

Region-specific models for fire risk prediction showed skewed performance, particularly in areas with low fire occurrence such as northern Europe. Although these models report high ROC-AUC values, their precision and recall are poor due to class imbalance. The general European model, enhanced with latitude and longitude as features, proved to be more robust and interpretable without needing explicit eco-region separation (Balch et al., 2017).

B. Recommendation for future testing

1. Interpolation Requires Careful Justification

Upscaling coarse input features such as FWI is tempting but can result in misleading gradients and overfitting. Interpolation should only be applied if there's evidence that higher resolution adds predictive value without distorting signal-to-noise ratio (Di Giuseppe et al., 2016).

2. Transition to a True Forecast Model

The Triple-F (Forest Fire Forecast) approach departs from pure FWI mimicry and creates a task-specific fire risk model by integrating key weather and anthropogenic features. This

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

approach enables easier retraining, region transferability, and higher accuracy in operational settings (Abatzoglou et al., 2018; Ruffault et al., 2020).

3. Use Forward-Year Data as a Proxy for Ground Truth

In the absence of high-quality ground-verified ignition reports, using future-year data (e.g. 2025) as a form of temporal validation helps estimate the model's robustness in deployment. While not perfect, this approach simulates real-world applicability better than random cross-validation (Preisler et al., 2008).

4. Fire Spread Prediction Could Be a Future Step

Although not necessary for this product's scope due to sparse fire occurrence per grid, fire spread modeling may be a valuable future direction. With tools like Google's FireSat expected to detect ignitions within minutes, predicting fire movement post-detection may become more critical than ignition risk itself (Filizzola et al., 2024; Abatzoglou et al., 2018).

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Conclusion

This intermediary report outlines substantial advancements in the development and validation of machine learning models for detecting major forest threats—bark beetle outbreaks, windthrow, and fire risk—using Sentinel satellite data. Key progress includes the integration of advanced deep learning architectures, the creation of more robust training datasets through collaboration with field partners, and the successful on-ground testing of model outputs across multiple European test sites.

Model performance has improved in terms of accuracy, with several models demonstrating good predictive capabilities and practical relevance for operational deployment.

These achievements confirm the scientific and technical viability of the SWIFTT platform as a scalable, cost-effective solution to support forest managers in risk monitoring and climate adaptation.

References

[Bark beetle outbreak model development and evaluation]

[Abdullah et al., 2019] Abdullah, H., Skidmore, A. K., and et al., Sentinel-2 accurately maps green-attack stage of European spruce bark beetle (Ips typographus, L.) compared with Landsat-8. Remote Sensing in Ecology and Conservation 5, 1, 87–106, 2019

[Andresini et al., 2022] Andresini, G., Appice, et al. Leveraging autoencoders in change vector analysis of optical satellite images. J. Intell. Inf. Syst. 58(3): 433-452, 2022

[Andresini et al., 2023a] Andresini, G., Appice, A., Ienco, D., Malerba, D., Seneca: Change detection in optical imagery using Siamese networks with active-transfer learning. Expert Systems with Applications 214, 119123, 2023

[Andresini et al., 2023b] G. Andresini, A. Appice, D. Malerba, SILVIA: An explainable framework to map bark beetle infestation in Sentinel-2 images, IEEE Journal of 643 Selected Topics in Applied Earth Observations and Remote Sensing 16 64, 2023

[Andresini et al., 2023c] G. Andresini, A. Appice, et al., SENECA: Change detection in optical imagery using Siamese networks with Active-Transfer Learning. Expert Syst. Appl. 214: 119123, 2023,

[Andresini et al. 2024a] Andresini G., Appice A., Ardimento P., Boffoli N., Carlucci M., Recchia V. Semantic Segmentation Model Reuse to Map Bark Beetle Outbreaks, DEARING@ECMLPKDD 2024 Workshops, Springer Post Proceedings, 2024

[Andresini et al. 2024b] Andresini G., Appice A, Ienco D., Malerba d., Recchia V., Potential of Spectral-Spatial Analysis to Map Forest Tree Dieback Due to Bark Beetle Hotspots in Sentinel-2 Images. IGARSS 2024, 5227-5230, 2024

[Appice et al., 2020] Appice, A., Guccione, P., Acciaro, E., Malerba, D., Detecting salient regions in a bi-temporal hyperspectral scene by iterating clustering and classification. Applied Intelligence 50(10), 3179–3200, 2020

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

[Barta et al., 2021] Barta, V., Lukes, P., Homolova, L., Early detection of bark beetle infestation in Norway spruce forests of central Europe using Sentinel-2, International Journal of Applied Earth Observation and Geoinformation 1—13, 102335, 2021

[Bruzzone & Prieto, 2000] Bruzzone, L., Prieto, D.F., Automatic analysis of the difference image for unsupervised change detection. IEEE Transactions on Geoscience and Remote sensing 38(2), 1171–1182, 2000

[Cai et al., 2023] Cai X., Bi Y., Nicholl P.N., Sterritt R., Revisiting the encoding of satellite image time series. In: 34th British Machine Vision Conference 2023, BMVC 2023, 402–404, BMVA Press, 2023

[Candotti et al., 2022] Candotti, A., De Giglio, M., Dubbini, M., Tomelleri, A., A Sentinel-2 based multi-temporal monitoring framework for wind and bark beetle detection and damage mapping, Remote Sensing 14 (23), 1—29, 2022

[Cheng et al., 2022] Cheng B., Misra I., Schwing A.G., Kirillov A., Girdhar R., Masked-attention mask transformer for universal image segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 1280–1289, 2022

[Choi et al., 2010] Choi, S., Cha, S., Tappert, C., A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8, 43–48, 2010

[Clasen et al., 2024] Clasen K.N., Hackel L.W., Burgert T., Sumbul G., Demir B., Markl V., reBEN: Refined BigEarthNet Dataset for Remote Sensing Image Analysis. CoRR abs/2407.03653 (2024)

[Deng et al., 2008] Deng, J.S., Wang, K., Deng, Y., Qi, G.J., Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data. International Journal of Remote Sensing 29(16), 4823–4838, 2008

[Devlin et al., 2019] Devlin J., Chang M.W., et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. ACL, 4171–4186, 2019

[Fernandez-Carillo et al., 2020] Fernandez-Carrillo, A., Patocka, Z., Dobrovolny, L., et al., Monitoring bark beetle forest damage in central Europe. A remote sensing approach validated with field data, Remote Sensing, 12(21), 1—19, 2020

[Forzieri et al., 2023] Forzieri G., Dutrieux L., et al., The database of European forest insect and diseasedisturbances: DEFID2. Global Change Biology 29(21), 6040–6065, 2023

[Gao et al., 2016] Gao, F., Dong, J., Li, B., Xu, Q., Automatic change detection in synthetic aperture radar images based on pcanet. IEEE Geoscience and Remote Sensing Letters 13(12),1792–1796, 2016

[Kalinicheva et al., 2019] Kalinicheva, E., Sublime, J., Trocan, M., Change detection in satellite images using reconstruction errors of joint autoencoders. In: I.V. Tetko, V. K˚urkov´a, P. Karpov, F. Theis (eds.) ICANN 2019: Image Processing, pp. 637–648. Springer International Publishing, 2019

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

[Kirillov et al., 2023] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, 3992–4003, IEEE, 2023

[Howard and Ruder, 2018] Howard J., Ruder S., Universal Language Model Fine-tuning for Text Classification. In 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, 2018

[Huo et al., 2021] Huo, L., Persson, H.J., Lindberg, E., Early detection of forest stress from European spruce bark beetle attack, and a new vegetation index: Normalized distance red & swir (ndrs), Remote Sensing of Environment, vol. 255, 1—18, 2021

[Ilsever & Unsalan, 2012]Ilsever, M., Unsalan, C.: Two-dimensional change detection methods: remote sensing applications. Springer Science & Business Media, 2012

[Li et al., 2024] Li, X., Wen, C. and et al., Vision-Language Models in Remote Sensing, Current progress and future trends. IEEE Geoscience and Remote Sensing Magazine 12, 2, 32–66, 2024

[Lopez-Fandino et al., 2019] Lopez-Fandino, J., B. Heras, D., Arguello, F., Dalla Mura, M., Gpu framework for change detection in multitemporal hyperspectral images. Int. J. Parallel Programm. 47, 272–292, 2019

[Louis et al., 2016] J. Louis, V. Debaecker, B. Pflug, M. Main-Knorn, J. Bieniarz, U. Mueller-Wilm, E. Cadau, F. Gascon, Sentinel-2 sen2cor: L2a processor for users. In Proceedings of the Living Planet Symposium 2016, pp. 1–8. Spacebooks Online, 2016

[Lu et al., 2010] Lu, D., Mause, P., Brondizio, E., Moran, E., Change detection techniques. International journal of remote sensing 25, 2365–2401, 2010

[Osco et al., 2023] Osco, L.P., Lopes de Lemos, E., and et al., The Potential of Visual ChatGPT for Remote Sensing. Remote Sensing 15, 13 , 3232,2023

[Pasquadibisceglie et al., 2025] Pasquadibisceglie V., Recchia V., Appice A., Malerba D., Fiameni G., GANDALF: A LLM-based approach to map bark beetle outbreaks in semantic stories of Sentinel-2 images, The 40th ACM/SIGAPP Symposium On Applied Computing - Track on Machine Learning and Its Applications, Catania, Italy, 2025

[Peng and Wang, 2020] Peng, P., Wang, J., How to fine-tune deep neural networks in few-shot learning? CoRR abs/2012.00204, 2020

[Raiaan et al., 2024] Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S., A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access, 12, 26839–26874, 2024

[Recchia et al., 2024] Recchia, V., Andresini G., Appice A., Fontana P., Malerba D., An attention-based CNN approach to detect forest tree dieback caused by insect outbreak in Sentinel-2 images, Proceedings of Discovery Science 2024, Springer, 2024

[Recchia et al., 2025] Recchia, V., Andresini G., Ardito L., Appice A., Reusing a BigEarthNet Deep Model to Map Bark Beetle Outbreaks in Sentinel-2 Forest Images, DEARING@ECMLPKDD 2025 Workshops, Selected for Post proceedings

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

[Reynolds, 2009] Reynolds, D., Gaussian Mixture Models. In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2009

[Sahoo et al., 1988] Sahoo, P., Soltani, S., Wong, A.C., A survey of thresholding techniques. Computer vision, Graphics and Image Processing 41(2), 233–260, 1988

[Sumbul et al., 2021] Sumbul, G., Charfuelan, M., Demir, B., Markl, V., Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 5901–5904, 2019

[Sundararajan et al., 2017] Sundararajan M., Taly A., and Yan O., Axiomatic Attribution for Deep Networks. In 34th International Conference on Machine Learning, ICML 2017, Vol. 70. PMLR, 3319–3328, 2017

[Sun et al., 2024] Wenfang Sun, Yingjun Du, and et al., Training-Free Semantic Segmentation via LLM-Supervision. CoRR abs/2404.00701, 2024

[Turc et al., 2019] Turc I., Chang M.W., and et al., Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation, CoRR abs/1908.08962,2019

[Turkulainen et al., 2023] Turkulainen, E., Honkavaara, E., Näsi, R., et al.: Comparison of deep neural networks in the classification of bark beetle-induced spruce damage using uas images. Remote Sensing 15(20), 2023

[Wang et al., 2021] Wang W., Xie E., Li X., Fan D.P., Song K., Liang D., Lu T., Luo P., Shao L., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021), 548–558, 2021

[Wang et al., 2022] Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X., SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. CoRR abs/2211.07044, 2022

[Zhang, Cong et al., 2022] Zhang, J., Cong, S., et al., Detecting pest-infested forest damage through multispectral satellite imagery and improved U-Net++, Sensors, vol. 22, no. 19, 1—21, 2022

[Zhao et al., 2020] Zhao H., Kong X., He J., Qiao Y., Dong C., Efficient image super-resolution using pixel attention. In: European Conference on Computer Vision. pp. 56–72 Springer, 2020

[Zwieback et al., 2024] Zwieback, S., Young-Robertson, J., et al., Low-severity spruce beetle infestation mapped from high-resolution satellite imagery with a convolutional network. ISPRS Journal of Photogrammetry and Remote Sensing 212, 412–42, 2024

[windthrow]

Albrecht, A., Hanewinkel, M., Bauhus, J., & Kohnle, U. (2019). How does silviculture affect storm damage in forests of South-Western Germany? *Forest Ecology and Management, 432*, 9–20.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Atalay, A. B., Aydinoglu, A. C., & Günay, E. (2021). The use of Sentinel-1/2 vegetation indexes with GEE time series data in detecting land cover changes in the Sinop nuclear power plant construction site. *ResearchGate Preprint*.

Cortini, F., Vescovi, F., Pirotti, F., & Tarantino, E. (2022). Sentinel-2 time series and Google Earth Engine for change detection and forest damage assessment. *Remote Sensing, 14*(6), 1472.

Filipponi, F. (2019). *Sentinel-1 ARD: ESA guidelines and implementation*. Remote Sensing Applications, 12(4), 210-223.

Forzieri, G., et al. (2020). *A spatially explicit database of wind disturbances in European forests over the period 2000–2018.* Earth System Science Data, 12(1), 257-276

Gardiner, B., Byrne, K. E., Hale, S. E., Kamimura, K., Mitchell, S. J., Peltola, H., & Ruel, J. C. (2000). A review of mechanistic modelling of wind damage risk to forests. *Forestry, 83*(3), 267–281.

Fraser, R. H., et al. (2005). *Evaluating forest disturbance using normalized difference vegetation index (NDVI) data*. Remote Sensing of Environment, 94(4), 412-421.

Gorelick, N., et al. (2017). *Google Earth Engine: Planetary-scale geospatial analysis for everyone*. Remote Sensing of Environment, 202, 18-27.

Jakubowski, M. K., Li, W., Guo, Q., & Kelly, M. (2023). Quantifying forest structure and

change using drone LiDAR: Applications in post-disturbance mapping. *Remote Sensing of Environment, 288*, 113520.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Kamoske, A. G., Fry, D. L., Kane, V. R., & North, M. P. (2019). Using LiDAR to characterize structural development in frequent-fire forests. *Forest Ecology and Management, 433*, 364–375.

Oeser, J., Asam, S., Kübert, A., & Dech, S. (2021). Windthrow detection using Sentinel-1 and Sentinel-2 time series. *ISPRS Archives, XLIII-B3-2021*, 865–872.

Pirotti, F., Tarantino, E., & Vescovi, F. (2021). Integration of Earth observation and forestry data for forest damage modeling. *Procedia Computer Science, 181*, 898–906.

Schindler, D., Völkel, J., & Hanewinkel, M. (2022). Predicting forest windthrow with machine learning: A comparison of models and variables. *Ecological Modelling, 471*, 110029. https://doi.org/10.1016/j.ecolmodel.2022.110029

Seidl, R., Schelhaas, M. J., Rammer, W., & Verkerk, P. J. (2014). Increasing forest disturbances in Europe and their impact on carbon storage. *Nature Climate Change, 4*(9), 806–810. https://doi.org/10.1038/nclimate2318

Senf, C., Pflugmacher, D., Zhiqiang, Y., Neumann, M., Sebald, J., Hostert, P., & Seidl, R. (2018). Canopy mortality has doubled in Europe's temperate forests over the last three decades. *Nature Communications, 9*, 4978.

Sinha, S., Heurich, M., & Krzystek, P. (2021). Evaluation of Sentinel-1 and Sentinel-2 data fusion for forest disturbance detection. *Remote Sensing of Environment, 263*, 112555.

Sobczak, R., Goczał, J., Dąbrowski, M. J., & Wojtuń, B. (2021). Modelling and prediction of wind damage in forest ecosystems of the Sudety Mountains, SW Poland. *Procedia Computer Science, 192*, 2419–2427.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Ullah, S., et al. (2021). *Fusion of optical and radar remote sensing data for forest disturbance monitoring: A review*. Remote Sensing, 13(4), 679-702.

Wang, X., et al. (2010). *Vegetation index-based approach for detecting windthrow and other forest disturbances*. Forest Ecology and Management, 259(12), 2159–2167.

Weng, Y., Peltola, H., Zeng, H., & Kellomäki, S. (2024). A hybrid modelling approach for wind damage risk assessment under future climate. *Forests, 15*(2), 221.

Zhu, J., Ni, W., Yang, Y., & Shang, Y. (2023). Detecting storm-induced forest disturbance using Sentinel-1 SAR time series. *Remote Sensing, 15*(4), 1123.

Zhu, Z., & Woodcock, C. E. (2012). *Continuous change detection and classification of land cover using all available Landsat data*. Remote Sensing of Environment, 122, 1-8.

[fire]

Gelabert, P., Libertà, G., Francos, A., & Camia, A. (2025). Modelling human-driven wildfire ignition in Europe using fuel and accessibility features. EGUsphere Preprints.

Moghim, S., & Mehrabi, A. (2024). Wildfire susceptibility mapping in Central Europe using topography, human activity and machine learning. Fire Ecology, 20(2), 33.

Porta, L., Clavet, B., & White, A. (2025). Sentinel-based wildfire forecasting using deep learning: performance limits and spatial transferability. Remote Sensing of Environment, in press.

Shen, L., Ma, Y., & Wu, J. (2023). FireRisk: A remote sensing benchmark for wildfire risk classification using masked autoencoders. arXiv preprint.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.

Wang, H., Zhu, Y., & Buch, M. (2023). Comparative assessment of physics-based and machine learning models for wildfire ignition prediction. Journal of Forestry Research, 34(9), 1087–1102.

Bedia, J., Herrera, S., Camia, A., Moreno, J. M., & Gutiérrez, J. M. (2014). Forest fire danger projections in the Mediterranean using ENSEMBLES regional climate change scenarios. *Climatic Change*, 122(1-2), 185–199.

Di Giuseppe, F., Rémy, S., Pappenberger, F., Wetterhall, F., & Camia, A. (2020). ERA5-based global reanalysis of fire weather indices. *Geophysical Research Letters*, 47(9).

Di Giuseppe, F., Pappenberger, F., Wetterhall, F., Krzeminski, B., Camia, A., Libertá, G., & San-Miguel, J. (2016). The potential predictability of fire danger provided by numerical weather prediction. *Journal of Applied Meteorology and Climatology*, 55(11), 2469–2491.

Di Giuseppe, F., Rutschmann, J., Krzeminski, B., & San-Miguel-Ayanz, J. (2022). Global fire danger modelling using ECMWF ensemble forecasts. *Scientific Data*, 9, Article 52.

Filizzola, C., Marchese, F., Pergola, N., & Tramutoli, V. (2024). Toward satellite-based near real-time wildfire detection: A review and perspective. *Scientific Reports*, 14, Article 12493.

Fusco, E.J., Finn, J.T., Balch, J.K., & Bradley, B.A. (2021). Spatial and temporal controls on fire likelihood across the western United States. *Environmental Research Letters*, 16(3), 034030.

Preisler, H.K., Brillinger, D.R., Burgan, R.E., & Benoit, J.W. (2008). Forecasting distributions of large federal-lands fires utilizing satellite and gridded weather information. *International Journal of Wildland Fire*, 17(5), 614–624.

Ruffault, J., Curt, T., Martin-StPaul, N., Moron, V., & Trigo, R.M. (2020). Increased likelihood of heat-induced large wildfires in the Mediterranean Basin. *Scientific Reports*, 10, Article 13790.

Ruffault, J., Martin-StPaul, N. K., Moron, V., & Trigo, R. M. (2018). Extreme wildfire events are linked to global-change-type droughts in the northern Mediterranean. *Natural Hazards and Earth System Sciences*, 18(3), 847–856.

Funded by the European Union

Funded by the European Union under Grant Agreement 101082732. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Union Agency for the Space Programme (EUSPA). Neither the European Union nor the granting authority can be held responsible for them.