

Sentinel-2 Masking CNNs Trained on Physics-Supervised Labels

Efrain Padilla-Zepeda , Kevin Alonso , Raquel De Los Reyes , Deni Torres-Roman , Avi Putri Pertiwi ,
and Tobias Storch 

Abstract—This article presents a method to improve pixel-level classification of Sentinel-2 imagery by integrating spectral index-based masking with deep learning approaches using 1-D, 2-D, and 3-D convolutional neural networks (CNN1D, CNN2D, and CNN3D). Rather than relying on manually labeled data, the proposed method selects high-quality training samples from Python-based atmospheric correction software (PACO), using pixel selection strategies to remove ambiguous or inconsistent labels. Three selection strategies are explored: full inclusion, uniqueness-based filtering, and physics-based rules. Unlike traditional masking algorithms based only on spectral indices, the CNN models leverage spatial correlations among neighboring pixels across all spectral bands, plus auxiliary features like elevation and illumination, enabling the extraction of more informative representations and improved classification accuracy, particularly in complex scenes. The model is trained using a large global training dataset from PACO, while a separate validation dataset from the same source is used to monitor performance during learning and prevent overfitting. Final evaluation is performed using two independent manually labeled testing datasets (TD1 and TD2) that span diverse land cover types and atmospheric conditions. Compared to PACO's baseline classification, our CNN approaches achieve consistent improvements for normalized Matthews correlation coefficient, with maximum gains of +3.3 percentage points (pp) on TD1 (from 0.855 to 0.888) and +18.3pp on TD2 (from 0.665 to 0.848). The largest class-wise gains are observed for shadows and clear land-related classes, with up to +22.7pp improvement. These results confirm the effectiveness of the proposed training strategy and its potential for improving label quality in large-scale Earth observation pipelines.

Index Terms—Classification, deep learning, masking algorithm, multispectral, pixel-level, sentinel-2.

Received 18 February 2025; revised 17 April 2025 and 18 May 2025; accepted 15 June 2025. Date of publication 19 June 2025; date of current version 18 July 2025. The work of Efrain Padilla-Zepeda was supported in part by the Consejo Nacional de Humanidades, Ciencia y Tecnología (Conahcyt), Mexico, for the Ph.D. under Grant 789304, and in part by the Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav) for serving as the home institution for his Ph.D. studies. This work was supported by the Deutscher Akademischer Austauschdienst (DAAD) for supporting the research stay in Germany, funded by the Bundesministerium für Bildung und Forschung (BMBF) through DAAD fellowship number 57645446, and in part by the Deutsches Zentrum für Luft- und Raumfahrt (DLR) for providing a workplace and supporting this research. (Corresponding author: Efrain Padilla-Zepeda.)

Efrain Padilla-Zepeda is with the German Aerospace Center (DLR), Earth Observation Center, Remote Sensing Technology Institute, 82234 Wessling, Germany, and also with Center for Research and Advanced Studies of IPN (Cinvestav), Guadalajara 45017, Mexico (e-mail: Efrain.Padilla@dlr.de).

Kevin Alonso is with European Space Agency (ESA), 00044 Frascati, Italy (e-mail: kevin.alonso@esa.int).

Raquel De Los Reyes, Avi Putri Pertiwi, and Tobias Storch are with German Aerospace Center (DLR), Earth Observation Center, Remote Sensing Technology Institute, 82234 Wessling, Germany (e-mail: Raquel.delosReyes@dlr.de; Avi.Pertiwi@dlr.de; Tobias.Storch@dlr.de).

Deni Torres-Roman is with Center for Research and Advanced Studies of IPN (Cinvestav), Guadalajara 45017, Mexico (e-mail: deni.torres@cinvestav.mx).

Digital Object Identifier 10.1109/JSTARS.2025.3581058

I. INTRODUCTION

THE optical data products of the Earth's surface are essential for environmental monitoring and security, as they capture interactions between solar radiation and matter in the lithosphere and hydrosphere. This interaction produces a unique combination of light reflected at different wavelengths for each ground sample (or sensor pixel), known as the spectral signature. These spectral signatures support diverse applications such as vegetation monitoring, water management, climate studies, and wildfire detection, among many others [1].

In the context of the rapidly increasing implementation of machine learning techniques in remote sensing applications, the contribution of this work lies in providing a method for handling high data volumes for training deep learning classification models with a global-perspective implementation for current and future optical remote sensing missions. The goal of the proposed approach is to develop a robust methodology for automatically retrieving high-quality classification labels, reducing the need for manual labeling. In addition, the model was tested on independent datasets to assess its generalization across varying class definitions, and to compare the benefits of 1-D, 2-D, and 3-D convolutional feature extractors.

An important aspect of this work is that the testing datasets (TDs) were independently and manually labeled, while the models were trained exclusively using reference labels derived from Python-based atmospheric correction software (PACO). This demonstrates the generalization capability and robustness of the proposed approach. Although the PACO-generated masks are based on spectral indices such as the normalized difference water index and the normalized difference snow index, these indices are not directly used as inputs to the deep learning models. Instead, the model operates on the original reflectance bands, along with additional features such as elevation and illumination, and learns spatial and spectral patterns from raw data.

To analyze dataset-specific biases and identify common misclassification patterns, we analyze the resulting confusion matrices and selected classification maps. This analysis aims to draw conclusions and highlight the improvements needed to create better training datasets.

Multiple efforts from space agencies and the private sector have developed missions to monitor Earth, allowing public access to optical imagery and its associated products. Among the currently operational remote sensing missions [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], the Sentinel-2 mission [12] stands out as part of the European Union's Copernicus program. Sentinel-2

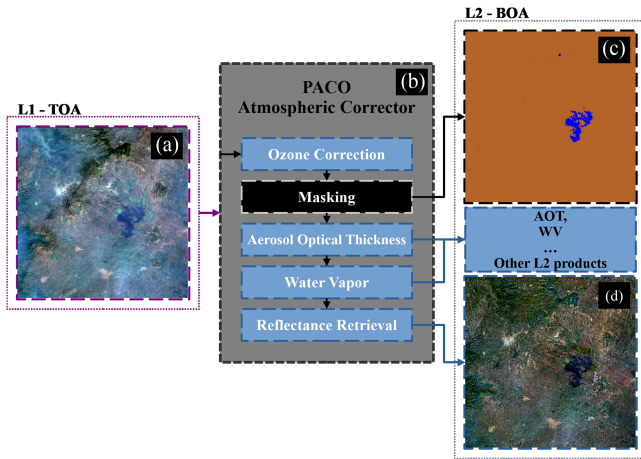


Fig. 1. TOA input image, atmospheric correction with PACO, and the corresponding products obtained during processing: BOA image, masks, among others [13]. Image: Sentinel-2, T32NPM, 23 December 2016.

consists of a constellation of two operational Sun-synchronous satellites equipped with optical imaging instruments. The mission's primary data products are multispectral images at Level-1 (L1) and Level-2 (L2) of processing [see Fig. 1(a) and (d)]. The multispectral sensor aboard Sentinel-2 captures bands in the visible (VIS), near-infrared (NIR), and Short-Wave Infrared (SWIR) spectra, with a 5-day revisit time, and each image covers a 290 km swath. The two satellites follow the same Sun-synchronous orbit with a 98.6° of inclination, but are 180° out of phase. They orbit the Earth at an altitude of 786 km with a period of 101 min.

Remote sensing optical data is captured at various spatial and spectral resolutions, depending on the optical sensor and its specific applications. For instance, multispectral sensors, such as the one onboard Sentinel-2 satellites, have a limited number of selected bands (typically between 5 and 15) with a broad full width at half maximum (FWHM), covering the most relevant wavelengths in the VNIR-SWIR to retrieve physical properties of the Earth's surface and atmosphere. The Sentinel-2 multispectral sensor has a total of 13 bands with a FWHM ranging from 13 to 184 nm, depending on the wavelength being captured.

Due to the growing range of optical remote sensing applications and the need for more effective use of distributed products for operative and future missions [14], [15], [16], atmospheric correction is essential to approximate surface reflectance as if the data were captured directly at ground level, allowing images to be compared across time despite variable atmospheric conditions. Fig. 1 presents a schematic of the atmospheric correction process, where the atmospheric correction software [see Fig. 1(b)] processes the top-of-atmosphere (TOA) reflectances [see Fig. 1(a)], which are typically distributed as L1 data. The resulting spectral image is called the bottom-of-atmosphere (BOA) reflectance [see Fig. 1(d)] and constitutes the L2 product. In addition, auxiliary products generated during BOA retrieval are also distributed, including the classification map, as shown in Fig. 1(c).

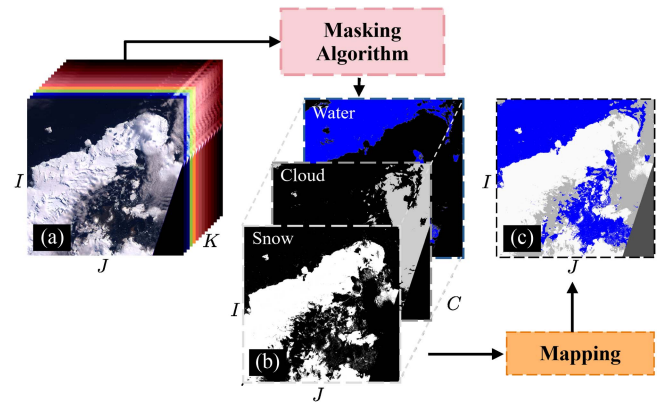


Fig. 2. Masking algorithm workflow, with a spectral image as an input and the resulting classification map of the same spatial dimensions at the output. Image: Antarctic. Sentinel-2, T21EVK, 3 February 2019.

During the atmospheric correction process, it is necessary to identify certain elements captured in the image that correspond to specific atmospheric or Earth's surface features, for instance: clear-sky, water, clouds, snow, among others [13]. For this task, masking algorithms take the spectral image as input [see Fig. 2(a)] and apply logical combinations of spectral tests based on thresholds of spectral indices, band ratios, and illumination conditions. These rules are used to identify known spectral signatures on a pixel-by-pixel basis, generating one binary mask per target class [see Fig. 2(b)]. Another common approach is pixel-level classification, also referred to as semantic segmentation or pixel-wise classification, where each pixel is assigned a unique class from a predefined set [see Fig. 2(c)].

The classification is typically delivered as part of the L2 products, either as a set of binary masks [referred to as the mask cube, see Fig. 2(b)] or as a complete labeled classification map [referred to as the preclassification, see Fig. 2(c)]. The preclassification is obtained by mapping each binary mask into a unique class using empirically defined rules. The classification products may also be delivered as a mixture of both for easier use. The final set of classification L2-BOA products is often defined based on the specific requirements of the mission. As a delivered product, this classification provides useful information for a wide range of remote sensing applications. As a result, most of the current remote sensing missions have a state-of-the-art implementation of masking algorithms, some of which will be described in Section I-A together with international efforts in the evaluation of these masking algorithms.

A. Literature Review

For the review of the current state of classification algorithms, we will consider as a main reference the latest cloud mask intercomparison exercise (CMIX) [17], which is an international collaborative effort initiated in the framework of the Committee on Earth Observation Satellites Working Group on Calibration and Validation (CEOS WGCV) and the participation of public and private entities. Although this exercise focuses on cloud masking, the classification algorithms evaluated in CMIX classify pixels into several classes, such as clear-sky, water,

TABLE I
PARTICIPANT MASKING ALGORITHMS IN THE LAST EDITION OF CMIX

Processor	Organization	Methodology	Application
ATCOR [18], [19]	DLR	Spectral tests	Multi-/hyper-spectral
CD-FCNN [20], [21]	University of Valencia	Machine learning	Multi-spectral
Fmask [22]–[24]	USGS	Spectral tests	Multi-spectral
FORCE [25]–[27]	Humboldt-Universität zu Berlin / Trier University	Spectral tests	Multi-spectral
Idepix [28]	Brockmann Consult	Spectral tests	Multi-spectral
S2cloudless [29]	Sinergise	Machine learning	Multi-spectral
InterSSIM [30]	Sinergise	Machine learning and spatio-temporal context	Multi-spectral
LaSRC [10], [31], [32]	NASA / University of Maryland	Spectral tests	Multi-spectral
MAJA [33], [34]	CNES / CESBIO	Multi-temporal data and spectral tests	Multi-spectral
Sen2Cor [35], [36]	ESA / Telespazio France	Spectral test and auxiliary data	Multi-spectral

Table extracted from [17].

snow, cloud shadows, thin clouds, and more. Table I provides a brief summary of the participating processors from the latest edition of CMIX.

It is important to note that some of the algorithms listed in this table have already been implemented in the processing chains of remote sensing missions, enabling global-scale deployment. In these types of exercises, the reference datasets were independently and manually labeled and were not used during the fitting process of any of the algorithms. This contrasts with most studies evaluating classifiers, which typically split a single dataset into training and test sets, relying on a single labeling source. This approach is not fully applicable to generic classification products in remote sensing missions with global coverage, due to the high variability of subclasses.

There are three major areas on which the classification techniques are based: spectral tests (or spectral indices thresholding), machine learning, and multitemporal analysis. Those algorithms based on spectral tests, which run tests on empirically defined thresholds tested over computed indices of different band combinations of the spectral image. Each of the indices is designed to highlight one of the physical properties of the materials that interact with the light in the Earth's surface or the atmosphere [19].

Machine learning approaches fit different types of artificial neural network models with high-quality data. The main challenge for training is to obtain high-quality classification maps on a global scale with sufficient representation of all types of Earth's surface materials under different atmospheric conditions. In addition, this process must be performed for every sensor that is intended to be supported. Alternatively, the support for multispectral and hyperspectral sensors creates the need to transform the data into an equivalent space to generalize the

model across a variety of sensors, usually resulting in a loss of spectral information.

For multitemporal analysis, repetitive observations with a short time difference between captures at similar viewing angles are needed to compare changes in reflectance with the background, obtaining promising results. However, these approaches are limited in their implementation due to the need for a location's temporal series.

The results reported in [17] show that machine learning-based masking algorithms do not perform significantly better than spectral test-based algorithms, as their predictions depend on the data quality used for training, and these datasets are not available for all sensors. Since the spectral tests of currently used masking algorithms typically do not consider all available spectral bands, deep learning-based classification algorithms could potentially identify hidden patterns within spectral signatures, but only if the training pixels are correctly labeled.

Recently published methods for pixel-level classification, which exploit the growing area of deep learning models [37], have been applied to remote sensing applications, usually based on convolutional neural networks (CNNs) [38], [39], generative adversarial networks (GANs) [40], [41], [42], or transformers [43], [44], [45] among other approaches. In addition, attention mechanisms [46], [47] enhance the detection of fine-grained features and, thus, raise classification accuracy on single-image datasets. All these works show promising results, and it is expected that emerging large models will make a huge impact on classification performance, as demonstrated in other study areas in recent years. However, the lack of high-quality harmonized labeled data limits researchers and developers to using datasets with the same labeling source for both training and testing sets. This approach demonstrates the model's capability of differentiating between the specified classes in the dataset, but it is limited by the sensitivity of the source labels and does not demonstrate generalization for other applications. Thus, this restriction limits the interpretability of the results for global-scale remote sensing missions.

Recent works in remote sensing have explored large-scale classification using self-supervised or weakly supervised approaches, often leveraging unlabeled or physically derived data. Notably, Tuia et al. [48] provided a perspective on physically informed learning as a key direction for advancing Earth observation. In related work [49], [50], [51], [52], [53], the Copernicus land cover product CORINE [54] has been used in conjunction with BigEarthNet [55] to train classification models while explicitly accounting for label noise. Our work follows a similar trend by combining physically based label generation with learning-based refinement, applied to globally distributed Sentinel-2 data.

Deep learning models based on CNNs for pixel-level classification are able to retrieve more spatial and spectral features compared to classic masking algorithms based on physics models. This is because they analyze high data volumes, considering all the spectral signatures and the information from neighboring pixels [56]. However, they are highly dependent on the data and reference labels for training, with the added challenge of interpreting the retrieved features. Although models

like CNN3D [57] can simultaneously extract spatial and spectral features, they also increase computational complexity. We leave the evaluation of approaches based on GANs, transformers, or attention mechanisms for future work, as the goal of this article is to assess the general concept introduced in [58] by improving the training dataset, providing a detailed methodology, and conducting a thorough evaluation of the results.

In this work, we compare three lightweight architectures: CNN1D [59], CNN2D [56], [60], and CNN3D [57] that represent purely spectral, spatial, and joint spectral–spatial feature extractors, respectively. These architectures have proved in previous work to achieve superior performance compared with classical machine-learning methods [38], [61]. Encoder–decoder networks such as U-Net [62] are designed for dense ground-truth masks. Because our supervision explicitly discards ambiguous pixels, those pixels do not contribute to the loss or to its backpropagation [63], making U-Net ill-suited for this setting. The chosen CNN1D, 2-D and 3-D variants, therefore, offer a clean spectrum of spectral only, spatial, and spectral–spatial modeling capabilities, mirroring the taxonomy in Paoletti et al.’s review [38]. The main motivation for implementing these deep learning approaches is that the current masking algorithm of PACO relies on simple spectral tests and does not extract complex spatial and spectral features, which CNNs are well-suited for.

For the evaluation of classification, the use of popular metrics, such as overall accuracy and the $F1$ -score, has been shown to lack robustness for imbalanced datasets [64], [65], exhibiting bias as the imbalance between classes increases. Therefore, it is necessary to use robust metrics to evaluate the performance of the analyzed processors.

To address this issue, the use of the Matthews correlation coefficient (MCC) is considered a very informative single-score metric [65]. MCC is also known as the R_k statistic for the multiclass case or the Phi coefficient for the binary case. For the analysis of performance for each class, it is helpful to include the confusion matrices together with Precision (P) and Recall (R), given their widespread usage and intuitive interpretation for class-wise analysis.

B. Contributions Structure

- 1) Presented in Section II is a description of the proposed methodology to generate improved classification products of Sentinel-2 scenes in the context of an atmospheric corrector, using PACO masking products as input, which is currently based on spectral indices [13]. PACO was developed at the German Aerospace Center (DLR) and is the new version of the well-known DLR IDL software ATCOR [19]. Also described are the pixel selection strategies for composing the training dataset from an extensive collection of scenes. In addition, the tested models are described, which extract spatial and spectral features from the spectral image to identify patterns in previously unused features and spatial context. This includes additional data, such as the digital elevation model (DEM) and the illumination of each pixel, considering the incident angle

of light over the scene. Finally, we outline the metrics we use to evaluate model performance.

- 2) Section III describes the independent test sets used for the evaluation of the classifier’s results.
- 3) The results are detailed in Section IV, and their contribution to classification in remote sensing is discussed in detail in Section V. Finally, Section VI concludes this article.

II. METHODOLOGY

This work addresses the pixel-level classification task using a supervised deep learning model trained with the current masks generated by PACO, which are derived from spectral indices. To improve classification performance, the model must extract more significant spatial and spectral features than the labeling source algorithm. This enables it to redefine the classes in the new classification space of the fitted model, differentiate between the classes, and thereby improve performance. In this section, the main characteristics of the implemented methodology are presented and discussed, covering the notation used, the problem statement together with an overview of the implemented workflow (shown in Fig. 3), and the followed guidelines, including data preprocessing, pixel selection, retrieval of higher quality pixels for training, and finally, the models’ architecture and fitting details, along with its evaluation.

For easy readability, in this article, the “reflectance” of a pixel refers to the TOA reflectance (also named at-sensor reflectance) represented by the Greek letter $\rho_{i,j,k}^{\text{TOA}}$, and computed for each of the k bands. It is defined in [13] and [19] as follows:

$$\rho_{i,j,k}^{\text{TOA}} = \frac{\pi \cdot d^2 \cdot L_{i,j,k}}{E_k \cdot \cos(\theta_s)} \quad (1)$$

where d represents the Earth–Sun distance in astronomical units, E_k is the extraterrestrial solar irradiance for band k , and θ_s denotes the solar zenith angle at the time of acquisition. $L_{i,j,k}$ is the measured radiance for the ground cell (i, j) at band k , defined by $L_{i,j,k} = c_0^k + c_1^k \cdot \text{DN}_{i,j,k}$, where DN is the scaled radiance (referred to as the digital number) and the radiometric offset c_0^k and gain c_1^k are obtained during radiometric calibration. Most of the classification or masking algorithms reviewed in Section I-B and our proposal have as input the TOA reflectance.

Using the notation from [66] and the representation of spectral cubes with tensors, where a is a scalar, $\mathbf{a}$ is a vector, $\mathbf{A}$ is a matrix, and $\mathcal{A}$ is a third- or higher order tensor, a spectral image of an observed scene $\mathcal{X}$ is represented as a third-order tensor

$$\mathcal{X} \in \mathbb{R}^{I \times J \times K} \quad (2)$$

where I (rows) and J (columns) are the spatial dimensions of the image, and K is the number of spectral bands. Then, the element $x_{i,j,k}$ is the reflectance value at position (i, j) in the band k , such that $\{\forall i, j, k \in \mathbb{Z} \mid (1 \leq i \leq I) \wedge (1 \leq j \leq J) \wedge (1 \leq k \leq K)\}$ and the vector containing the spectral signature of a pixel is represented by $x_{i,j,:}$. Each ordered pair (i, j) represents the coordinates of a specific location on the Earth’s surface in the observed scene.

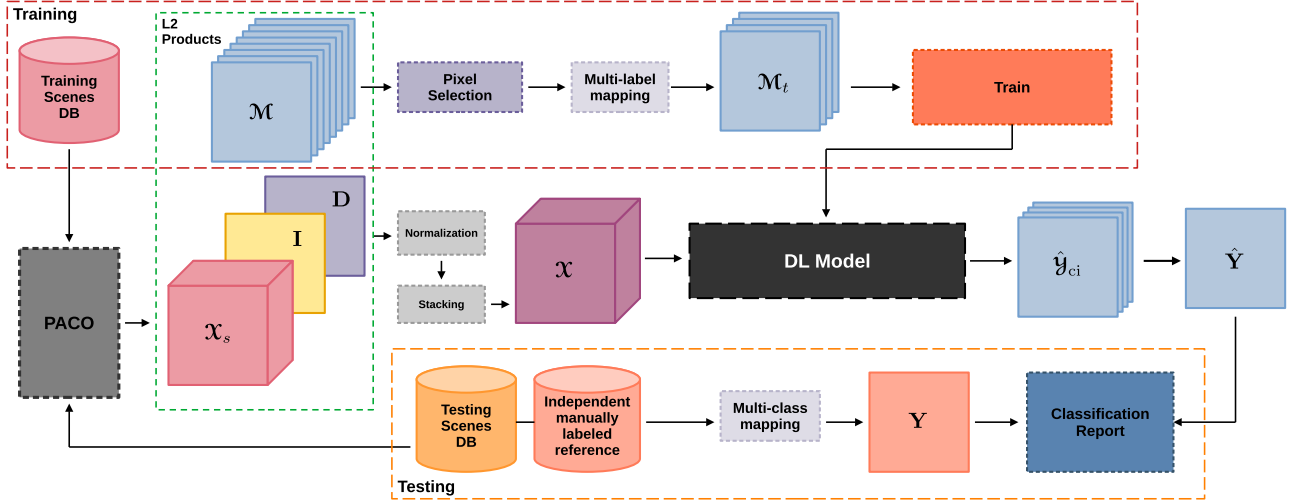


Fig. 3. Workflow of the implemented deep learning-based approach for classification. For the training stage (upper section), training scenes are preprocessed using PACO to generate TOA reflectances ($\mathcal{X}_s$), DEM ($\mathcal{D}$), and illumination ($\mathcal{I}$), which are normalized and stacked into $\mathcal{X}$. The masks cube ($\mathcal{M}$) is filtered by selecting pixels to create high-quality labels ($\mathcal{M}_t$), used to train the DL model. For evaluation (lower section), independent TDs with manually labeled references ($\mathcal{Y}$) are compared against predictions ($\hat{\mathcal{Y}}$), generating a classification report with performance metrics.

The classification algorithm is defined as a function

$$f_{\text{multiclass}} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}, \hat{\mathcal{Y}} \in \mathbb{Z}^{I \times J} \quad (3)$$

which maps from the spectral image $\mathcal{X}$ to the predicted classification map $\hat{\mathcal{Y}}$, matching the spatial resolution. This process is also known as pixel-based multiclass classification, where each element $\hat{y}_{i,j}$ represents a positive integer, so-called label, indicating the class to which it belongs.

Optionally, the masks could also be provided in a binary cube for multilabel classification (masks cube), such that

$$f_{\text{multilabel}} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}, \hat{\mathcal{Y}} \in \{0, 1\}^{I \times J \times C} \quad (4)$$

where C is the number of classes, allowing a pixel to belong to more than one class. Furthermore, $\hat{y}_{ci}$ could also represent the confidence index (ci) of each pixel to belong to a class, such that

$$f_{\text{confidence-index}} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}_{ci}, \hat{\mathcal{Y}}_{ci} \in \mathbb{R}^{I \times J \times C} \quad (5)$$

where each of the elements $\hat{y}_{i,j,c}^{ci}$ follows the axioms of probability, such that ($0 \leq \hat{y}_{i,j,c}^{ci} \leq 1$) and $\sum_{c=1}^C \hat{y}_{i,j,c}^{ci} = 1$. Although $\hat{y}_{i,j,:}^{ci}$ is often interpreted as a probability vector, neural networks are known to be mis-calibrated in practice [67]; we, therefore, avoid the use of the term and use the scores as a relative ci.

A. Problem Statement

The goal is to obtain a classifier based on a deep learning (DL) model such that $f_{DL} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}_{ci}$, where $\mathcal{X} \in \mathbb{R}^{I \times J \times F}$ is a cube containing the spectral data, DEM, and illumination, normalized and stacked into one cube. Here, I and J correspond to the rows and columns of the image, respectively, and F to the number of features.

Given that the classification is performed pixel-wise, the input size of the model is a small patch $\mathcal{X}_{\text{patch}} \in \mathbb{R}^{W \times W \times F}$, where the window size W is odd to ensure that the pixel of interest is located at the center of the patch cube and surrounded by

neighboring pixels, providing the spatial context for feature extraction. This is particularly helpful for high-spatial-resolution sensors, where the spatial correlation is high.

The output is a cube $\hat{\mathcal{Y}}_{ci} \in \mathbb{R}^{I \times J \times C}$ containing the predicted ci, where C is the number of target classes. For the case of a multiclass classifier, $f_{DL} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}$, where $\hat{\mathcal{Y}} \in \mathbb{Z}^{I \times J}$ is a classification map, the highest ci from $\hat{y}_{i,j,:}^{ci}$ is selected as the most likely class, assigning a label represented as an integer in $\hat{\mathcal{Y}}$.

To fit the model f_{DL} , the starting point is the collection of an extensive database (DB) of scenes from a particular remote sensing monitoring mission to train the model. These scenes are pre-processed by the mission's atmospheric correction processor (PACO in this study), to obtain the TOA reflectance, DEM, and illumination products mentioned earlier. In addition, the masks cube $\mathcal{M} \in [0, 1]^{I \times J \times K}$ is obtained, where K is the number of classes from the source. The information from the multilabel product is then leveraged by filtering the pixels and keeping those with low classification uncertainty. Furthermore, the binary masks cube is mapped to the target set of classes, which contains C classes, usually $C \leq K$. The resulting cube is defined as $\mathcal{M}_t \in [0, 1]^{I \times J \times C}$, and it will be used by the optimizer to fit the model through supervised learning.

As a final step in the testing phase, we use a representative set of scenes that have been manually annotated by an independent expert to demonstrate the robustness and generalization ability of the trained model. The datasets belong to other validation exercises within the Sentinel-2 optical mission performance cluster [68] and DLR's internal validation of the Fmask, ATCOR, and Sen2Cor processors [69]. These reference labels (usually referred to as the ground truth) are composed of the tuple $(\mathcal{X}, \mathcal{Y})$. The classifier's evaluation is accomplished by generating a classification report, which includes the confusion matrix and classification metrics that are robust to the unbalanced availability of samples for each class.

TABLE II
TRAINING DATASET CLASS DISTRIBUTION AND PIXEL COUNT

Color	Target Class	Pixel Count
	Clear	69.92 %
	Atmosphere	11.41 %
	Shadows	2.50 %
	Water	12.81 %
	Snow	3.37 %
	Total	2,122,676,571

B. Training Dataset and Preprocessing

Deep learning approaches are designed to handle large data volumes for training, and this characteristic allows the global-scale application perspective for classification algorithms. Ideally, the training dataset should be as large as possible to ensure a representation of all possible atmospheric conditions over all Earth's surfaces. However, including the whole archive of an optical remote sensing mission is rather challenging and computationally intensive. Therefore, we assembled a compact DB of representative L1 scenes that can be extended over time, enabling follow-up continual learning.

The training dataset used in this work consists of 71 scenes selected to match the locations of the AERONET sites [70], comprising ≈ 2.1 billion pixels distributed across the globe in a wide variety of biomes and atmospheric conditions. In addition, some images from the same locations as the TDs were included, but these do not correspond to the same dates. Table II shows the pixel count by class for the training dataset.

The labels for training are generated by PACO's masking, and these consist of 21 nonmutually exclusive masks, each corresponding to a preclassification class. Therefore, a single pixel could belong to one or more masks, which in this work are also referred to as a multilabel masks cube. These masks are represented in a cube $\mathcal{M} \in \{0, 1\}^{5490 \times 5490 \times 21}$. A mapping is needed to match the set of classes from the TDs. Consequently, we have defined five common classes: clear-sky soil pixels (*clear*), semitransparent clouds, clouds and cirrus (*atmosphere*), shadows from cloud or terrain elevations (*shadows*), *water* and *snow*.

In addition, each mapping can incorporate the bias introduced by the subjective definition of each class (e.g., how transparent a cloud must be to be considered a semitransparent cloud). Therefore, each mapping must be manually defined according to the specific definition of each class. Ideally, the classes should be defined in terms of physical metrics to harmonize classification products across different studies and applications. However, given the available TDs, we have defined a mapping for these specific cases. The first and second columns in Table III show the mapping from PACO preclassification classes to the corresponding target classes addressed in this work, while the rest of the columns describe the mapping for the TDs, which will be explained in more detail in Section II-E. More details about the decisions made to define the mapping will be reviewed in Section III. Further details on PACO masking and preclassification can be found in [13].

TABLE III
MAPPING TO COMMON TARGET CLASSES FROM PACO MASKING FOR TRAINING AND FROM TD1 AND TD2 FOR TESTING

Target Class	Training	Testing	
	PACO	TD1	TD2
Background	- Background - Dead pixels	- No data - Saturated or defective	- Background
Clear	- Bright - Haze land - Land	- Vegetation - Non vegetated	- Clear land
Atmosphere	- Cirrus cloud - Cirrus cloud thick - Clouds - Medium cirrus land - Medium cirrus water - Saturated - Thick cirrus land - Thick cirrus water - Thin cirrus land - Thin cirrus water	- Cloud medium probability - Cloud high probability - Thin cirrus	- Semi transparent cloud - Cloud
Shadows	- Cloud Shadows - Litopo shadows - Spectral shadows	- Cast shadows - Cloud shadows - Unclassified	- Cloud shadow - Topographic shadow
Water	- Haze water - Water	- Water	- Clear water
Snow	- Snow	- Snow	- Clear snow/ice

The input data for the proposed classification algorithm is composed of three different sources pre-processed by PACO in the early stages of the atmospheric correction process and resampled to a resolution of 20 m: the spectral information from the Sentinel-2 mission, denoted as $\mathcal{X}_s \in \mathbb{R}^{5490 \times 5490 \times 13}$, the DEM from the Copernicus mission [71], $\mathcal{D} \in \mathbb{Z}^{5490 \times 5490}$ containing the altitude in meters above sea level; and the corresponding illumination map $\mathcal{I} \in \mathbb{R}^{5490 \times 5490}$, produced using the geometrical relation between the sunlight beam and the ground terrain elevation.

All the products are then normalized between 0 and 1 by dividing each input dataset by its theoretical maximum physical expected value: $\frac{\mathcal{X}_s}{1.5}$, $\frac{\mathcal{D}}{8850}$, and $\frac{\mathcal{I}}{100}$. The fixed-value normalization lets us add new scenes to the training DB without having to renormalize for updated maxima or minima. The products are then stacked in one final cube with dimensions $\mathcal{X} \in \mathbb{R}^{5490 \times 5490 \times 15}$, including elevation and illumination as extra features of the spectral signature.

C. Pixel Selection for Training

This work analyzes the model's behavior when trained using three pixel selection strategies ("Pixel Selection" in Fig. 3): first, using all the available pixels (*no filter*); second, selecting pixels classified into a unique class (*uniclass filter*); and third, selecting pixels based on the physics rules defined in Table IV (*physics filter*).

The proposed pixel selection strategies leverage the multilabel classification from spectral indices thresholding, allowing classification of a pixel into one or more preclassification classes. Furthermore, this characteristic can be used to identify pixels with high classification uncertainty when the active masks are

TABLE IV
PHYSICS RULES FOR PIXEL FILTERING

Group	Surface				Low Atmosphere (< 3-4 km)				High Atmosphere (> 3-4 km)				Hardware								
CLASS	Land	Water	Snow	Bright	Cloud shadows	Spectral shadows	Litopo shadows	Haze land	Haze water	Clouds	Thin cirrus land	Thin cirrus water	Medium cirrus land	Medium cirrus water	Thick cirrus land	Thick cirrus water	Cirrus cloud	Cirrus cloud thick	Background	Saturated	Dead pixels
Land	x																				
Water		x																			
Snow			x																		
Bright				x																	
Cloud shadows					x																
Spectral shadows						x															
Litopo shadows							x														
Haze land								x													
Haze water									x												
Clouds										x											
Thin cirrus land											x										
Thin cirrus water												x									
Medium cirrus land													x								
Medium cirrus water														x							
Thick cirrus land															x						
Thick cirrus water																x					
Cirrus cloud																	x				
Cirrus cloud thick																		x			
Background																			x		
Saturated																				x	
Dead pixels																					x

Pixels classified as coincident green cells are included in the training set.

Pixels classified as coincident green cells are included in the training set.

contradictory, e.g., when a pixel is classified as snow and thick clouds.

For baseline comparison, we created a training set using all the available pixels to evaluate how our pixel selection strategy performs relative to the complete dataset with $\approx 2,123$ million pixels. This filter is referred to as “*no filter*” in the results section. We also included a conservative approach that selects only the pixels classified into a unique class. This provides labels with low uncertainty at the expense of under-representing certain classes, as the classification remains conservative, leaving the neural network to handle borderline cases. This filter is referred to as the “*uniclass filter*” in the results section discarding ≈ 176 million pixels.

In addition to the proposed pixel selection criteria, we have defined physics-based rules (see Table IV), discarding the pixels that show contradictory classes in their multilabel classification. This approach does not include in the training set the pixels with nonphysical classification (gray cells in Table IV), allowing the model to infer the ci of a pixel belonging to one class from the average cases, based on the spatial context and the spectral features, while still preserving the information from multilabel classification cases when a pixel belongs to one or more classes. These rules were defined considering physical principles and the current implementation of the masking algorithm and class definitions of PACO masks cube, e.g., semitransparent clouds

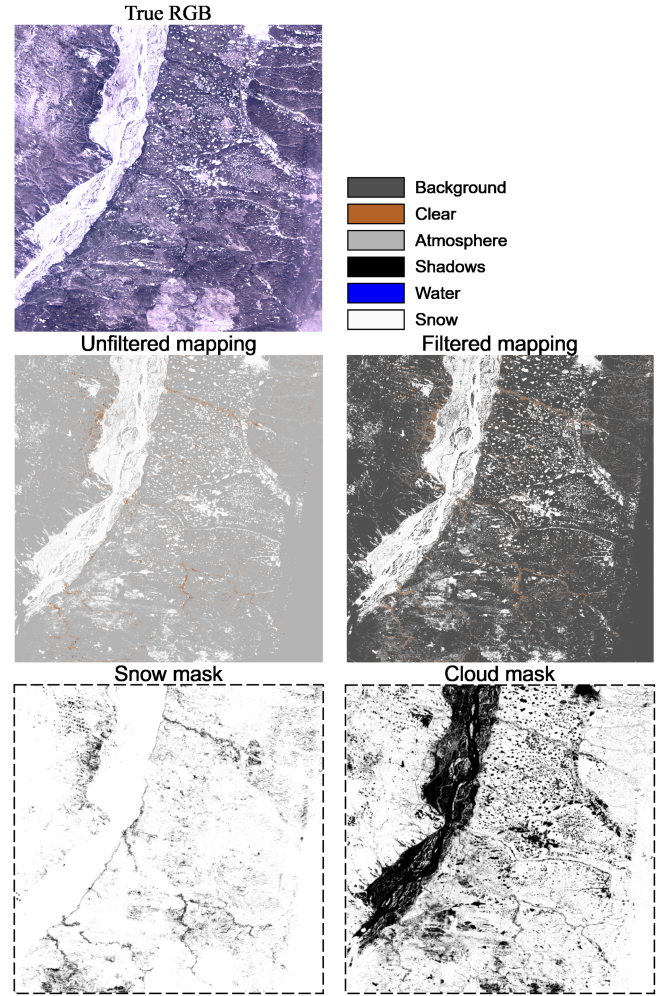


Fig. 4. Example of pixel selection for training. The binary masks at the bottom, from PACO’s masks cube, indicate contradictory classifications of snow and clouds. Pixels with both masks active are discarded from training by assigning them the background label. A comparison of unfiltered and filtered mapping is shown in the two middle images. The RGB image at the top serves as a reference. Image: Yakutsk, Russia. Sentinel-2, T52VEP, 17 March 2023, 11:45, -13°C .

or shadows over land or water. This filter is referred to as the “*physics filter*”; it discards ≈ 91 million pixels.

For instance, in Fig. 4, an example of the performed pixel selection is shown, where pixels classified as snow and clouds (a gray cell in Table IV) are discarded from the training datasets by assigning them the background label. In contrast, the pixels of snow/ice in the river are included, as they pass the filter (green cells in Table IV).

After filtering the pixels, two subsets were selected: one for training and one for validation. Both disjoint datasets were derived from the same global PACO-labeled collection and were separated using a stratified random sampling strategy to preserve class balance. The validation set was held out and used exclusively to monitor model performance during training. Its purpose was to prevent overfitting by evaluating the model at the end of each epoch and halting further updates if no improvement was observed in the validation performance.

TABLE V
CNN1D, CNN2D, AND CNN3D MODEL DESCRIPTIONS [38]

Model	Main Layer	Norm	Activation function	Down sampling
CNN1D	Linear input (1×15)	-	-	-
	Conv1D(50×15)	-	ReLU	-
	Dense(100)	BN	ReLU	-
	Dense(n_{classes})	-	Softmax	-
CNN2D	Linear input ($11 \times 11 \times 15$)	-	-	-
	Conv2D($50 \times 5 \times 5$)	-	ReLU	-
	Conv2D($100 \times 5 \times 5$)	-	ReLU	POOL(2×2)
	Dense(100)	-	ReLU	-
	Dense(n_{classes})	-	Softmax	-
CNN3D	Linear input ($11 \times 11 \times 15$)	-	-	-
	Conv3D($32 \times 5 \times 5 \times 10$)	BN	ReLU	-
	Conv3D($64 \times 5 \times 5 \times 5$)	BN	ReLU	POOL($2 \times 2 \times 1$)
	Dense(300)	BN	ReLU	-
	Dense(n_{classes})	-	Softmax	-

D. Models Description and Fitting

As mentioned in Section I, we have selected the models CNN1D [59], CNN2D [56], [60], and CNN3D [57] due to their relatively low computational complexity and different feature extraction domains [38]. In this specific application, where we discard pixels as explained in Section II-C, input data are supplied per-pixel. In the case of CNN2D and CNN3D, the pixel is fed along with neighboring pixels, resulting in spatial patches using a window size of 11×11 , incorporating spectral information. Therefore, the input size is a third-order tensor $\mathcal{X}_{\text{patch}} \in \mathbb{R}^{11 \times 11 \times 15}$, with the pixel of interest located at the center at coordinates (6, 6). The architectures of the models used in this work are shown in Table V. Originally presented in [38] for hyperspectral imagery, the models have been adapted in this case for multispectral imagery.

Each model performs feature extraction in different domains. CNN1D analyzes local relationships among neighboring wavelengths by performing 1-D convolutions with kernels of size ($1 \times F$). Therefore, the feature extraction only depends on the spectral signature for each pixel. CNN2D first extracts spatial features for each band independently with kernels of size (5×5). These band-specific feature maps are then passed to the detection stage, where spectral relationships across bands are jointly modeled. Similarly, CNN3D extracts spatial and spectral features jointly with 3-D convolutions using kernels of size ($5 \times 5 \times 10$), extracting complex relationships between the spatial and spectral information, which are then passed to a detection stage.

In subsequent layers of the three mentioned architectures, the spectral, spatial, or spatial-spectral features are progressively reduced and passed through a fully connected layer, which assigns confidence indices at the output of the CNNs using a softmax activation function [38].

In order to include the information from multilabel masks cube during training, we have used the softmax activation function $f(s)_c$ for each class $c \in \{1, 2, \dots, C\}$, defined in (6), in conjunction with the categorical cross-entropy loss $\text{CE}(\mathbf{t}, \mathbf{s})$, defined in (7)

$$f(s)_c = \frac{e^{s_c}}{\sum_i^C e^{s_i}} \quad (6)$$

$$\text{CE}(\mathbf{t}, \mathbf{s}) = - \sum_c^C t_c \cdot \log(f(s)_c) \quad (7)$$

where $\mathbf{s}$ is the vector containing the scores s_c inferred at the output of the model, and $\mathbf{t}$ is the reference (also called truth) multilabel vector. This loss function compares the ci at the output of the softmax activation functions, $[f(s)_1, f(s)_2, \dots, f(s)_C]$, with the multilabel classification used as a reference for training. A ci vector is obtained from the one-hot encoded classes by dividing each binary value t_c^b (casted to float) by the number of active masks, such that

$$\mathbf{t} = \frac{[t_1^b, t_2^b, \dots, t_C^b]}{\sum_{i=1}^C t_i^b} \quad (8)$$

in this way, the multilabel reference matches the ci format of the softmax function by assigning equal values to each class. While softmax remains a practical choice in this study, it is known to produce polarized confidence scores (one class is rated almost certain while the others receive little weight) [67]. In future work, we plan to derive the reference confidence from physics-based models and test alternative activation functions that may distribute confidence more evenly in ambiguous pixels. Lastly, to address the naturally unbalanced representation of each class in the training set, a weighting scheme is applied for each class as follows [72]:

$$w_c = \frac{|\mathbb{P}|}{C \cdot |\mathbb{P}_c|} \quad (9)$$

where $|\mathbb{P}|$ is the cardinality of the set of pixels and $\mathbb{P}_c$ the set of pixels of class c for training. To obtain $|\mathbb{P}_c|$, a mapping is performed from the multilabel mask cube to a single-label classification map using a fixed priority hierarchy in the following order: *atmosphere*, *shadows*, *snow*, *water*, and *clear*. The obtained pixel count provides an approximate class distribution for computing loss weights. Hence, the loss function leverages the multilabel classification from PACO, as it compares the ci of a pixel belonging to a class with the confidence indices of other classes, which enables a more informed classification criterion for mixed pixels, such as cloud edges, or for pixels with similar spectral signatures, such as snow and clouds.

An important aspect to consider during the fitting process is that domain generalization is achieved by avoiding overfitting. This is crucial since the TD labels originate from an independent source and correspond entirely to unseen scenes with manual labeling. The results presented in this work are from models trained with the Adam optimization algorithm [73], using 1% of the available data (≈ 20 million pixels) and a batch size of 100 000 samples. This is feasible due to the high spatial correlation between pixels and the uniformly distributed selection of pixels. These fitting parameters were determined through an exhaustive grid search, fitting, and evaluating different configurations. The validation set consists of 1% of the available data for training and is fixed for consistent comparison across training runs. The loss curves obtained by training for the three pixel selection strategies, using the parameters previously mentioned, are shown in Fig. 5(a)–(c), for CNN1D, CNN2D, and CNN3D, respectively.

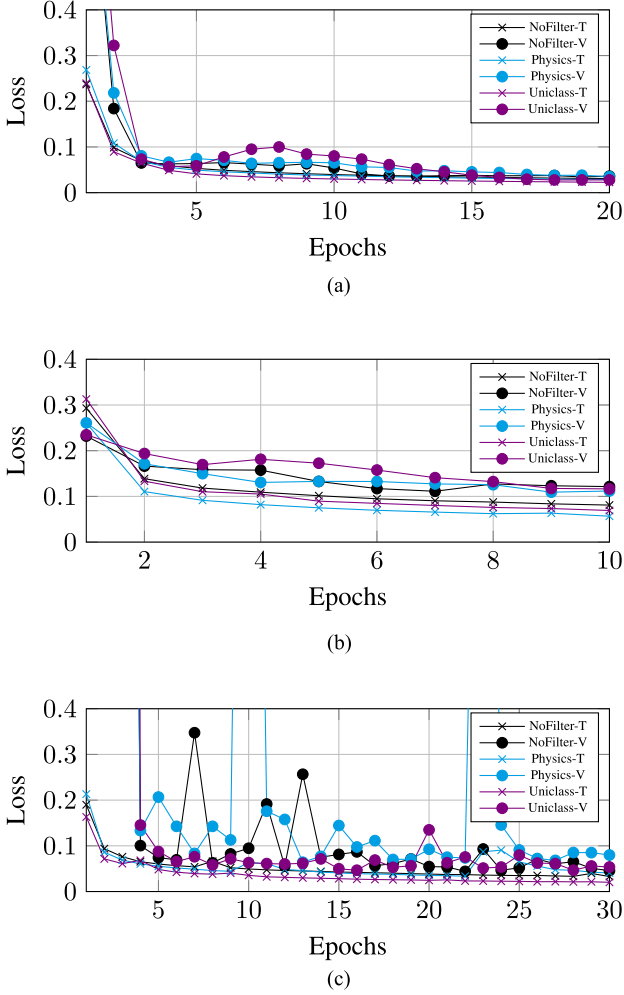


Fig. 5. Training (T) and validation (V) loss curves for the three CNN models under the three pixel-selection strategies. (a) CNN1D. (b) CNN2D. (c) CNN3D.

TABLE VI
CNNs SIZE AND TIME COMPLEXITY

Model	Parameters	Epochs	Training time (min)	Prediction time (min)
CNN1D	76,805	20	161.9	0.6
CNN2D	154,505	10	162.0	3.0
CNN3D	305,885	30	482.9	3.0

Table VI reports the parameters counts and wall-clock times measured using a GPU NVIDIA A100 (80 GB) with a 12-core CPU and 256 GB of RAM. For CNN2D and CNN3D, the spatial patches require extensive use of memory (with size $11 \times 11 \times 15$ per pixel), so a custom batch generator streams the data during both training and inference. All Sentinel-2 scenes were resampled to 20 m spatial resolution; times, therefore, reflect the operational cost at that resolution.

The reported times in Table VI show that CNN2D and CNN3D take practically the same wall-clock time per epoch, indicating that the custom batch generator is the training bottleneck rather than GPU computation. CNN1D, which ingests single pixels instead of spatial patches, trains roughly twice as fast per epoch

and, at inference, finishes a scene in about one-fifth the time of the 2-D and 3-D models, underscoring the speed advantage of a single-pixel input.

E. Prediction and Testing

Once the model is fitted, the testing phase must be conducted using different scenes and independent sources of information. This step is necessary to align with the global implementation perspective of this work and to demonstrate the generalization ability and limitations of the proposed approach. To evaluate this, multiclass classification testing sets are compared with the model's predictions using performance metrics, analyzing the bias of each while considering the imbalanced availability of samples of each class.

Given a confusion matrix C of K classes

$$C \in \mathbb{Z}^{K \times K} \quad (10)$$

which provides a general overview of correctly classified pixels (in diagonal) and misclassifications with other classes, while allowing the computation of any needed metric for comparison. In addition, most of the results are presented in terms of MCC for the general evaluation and comparison between the baseline and the prediction obtained. It is used for the multiclass case (as a single-score metric for each TD) and also for class-wise comparisons. However, also the inclusion of precision and recall metrics allows for an intuitive interpretation of overestimation or underestimation of a class.

For the multiclass case [74], [75], with K classes, MCC is defined as

$$\text{MCC} = \frac{c \cdot s - \sum_k p_k \cdot t_k}{\sqrt{\left(s^2 - \sum_k p_k^2\right) \left(s^2 - \sum_k t_k^2\right)}}, [-1, 1] \quad (11)$$

where $c = \sum_k C_{k,k}$ is the total number of correctly classified samples, $s = \sum_i \sum_j C_{i,j}$ is the total number of samples, $t_k = \sum_i C_{i,k}$ is the number of times class k truly occurred, and $p_k = \sum_i C_{k,i}$ is the number of times class k was predicted.

To individually evaluate each class, the multiclass confusion matrix is transformed into a binary confusion matrix, where the positive class is the class of interest k , and the negative class includes all other classes. The used metric is $\text{MCC}_k, [-1, 1]$, for the binary classification case [76], [77], defined for class k as follows:

$$\text{MCC}_k = \frac{T_p \cdot T_n - F_p \cdot F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \quad (12)$$

where T_p , T_n , F_p , and F_n are the true positives, true negatives, false positives, and false negatives, respectively.

For both cases, the results of MCC are presented in their normalized version [normalized Matthews correlation coefficient (nMCC)], linearly projected to the interval $[0, 1]$ as most of the popular metrics, defined as

$$\text{nMCC} = \frac{\text{MCC} + 1}{2}, [0, 1]. \quad (13)$$

TABLE VII
TDS CLASS DISTRIBUTION AND PIXEL COUNT

Color	Target Class	Pixel Count	
		TD1	TD2
	Clear	34.93 %	83.86 %
	Atmosphere	31.56 %	5.75 %
	Shadows	14.59 %	7.01 %
	Water	15.03 %	2.53 %
	Snow	3.88 %	0.84 %
Total		1,831,038	1,040,116

In addition, Precision (P_k) and Recall (R_k) metrics are also included, also known as user's and producer's accuracy, respectively, defined as

$$P_k = \frac{T_p}{T_p + F_p} \quad (14)$$

$$R_k = \frac{T_p}{T_p + F_n} \quad (15)$$

Other metrics provide a different interpretation of the results, and depending on the application, users may prioritize one over the others. For example, when the priority is to guarantee truly clear-sky pixels, users tend to prioritize the overestimation of clouds. In that scenario cloud-class recall is the preferred metric, because low recall would mean clouds were mislabeled as clear-sky. The classification report (including precision, recall, and the confusion matrices) is included in the Appendix to let the end-users judge such biases and tradeoffs. Nevertheless, MCC is used as the primary summary metric in this study because it shows less bias under class-imbalance of test datasets [64].

III. TDS DESCRIPTION

This section describes the TDs used in this work, including their labeling criteria, observed biases, and sample availability. The model is evaluated using two independently and manually labeled datasets, which were originally developed for external classification evaluation exercises unrelated to PACO or the proposed approach. For simplicity, these are referred to as testing dataset # (TD#). The class distribution and pixel counts for each evaluated TD are provided in Table VII.

Given that the TDs were originally developed for other validation exercises, including Sentinel-2 scene classification map (SCL) [78] and the one presented by Zekoll et al. [69], a mapping is needed for the common target classes defined for this work. The mapping described in Table III (columns three and four) ensures consistency between the test datasets and the target classes defined for this work. By standardizing the class definitions, the performance metrics can be uniformly evaluated, enabling a fair comparison of the model's predictions across different datasets. This mapping was empirically defined by analyzing the class definitions in the testing sets and comparing the results from PACO preclassification with the available labels for testing. This step is necessary due to subjectivity in class definitions and potential overlap between them. For instance, we observed that the PACO's preclassification class *cirrus cloud* is mapped almost

equally to the classes *cloud high probability* and *thin cirrus* from Sentinel-2 L2 A SCL. Thus, mapping a single class to multiple target labels is not possible without additional information. In this example, all cloud-related classes were merged into a unified class labeled *atmosphere*.

A. TD1: Sentinel-2 Fixed- and Random-Sites 2020–2023

This dataset was generated for the Copernicus Sentinel-2 SCL validation activities [68]. A labeling tool was used to generate visually labeled reference pixels in randomized subsets as the validation data. The annotations were based on the visualization of the selected Sentinel-2 L2 granule in true- and false-color composites, as well as spectral profiles. The two SWIR bands and the narrow NIR band were selected for the false color composite, along with the L1 C cirrus band, to help differentiate land surface, water, and clouds.

The scenes were labeled into the following SCL classes [78]: *vegetation*, *nonvegetated*, *water*, *snow*, *topographic*, and *cast shadows*, *cloud medium probability*, *cloud high probability*, *thin cirrus*, and *cloud shadow*. The number of pixels per class was manually controlled to be balanced (stratified); however, this depends on the land cover composition of the scenes, as some scenes might not naturally include all the SCL classes.

It comprises 53 manually labeled Sentinel-2 scenes, selected using two different strategies: fixed and randomly selected sites. The fixed site scenes set comprises 40 scenes from the time range of 2020 to 2023, corresponding to the validation exercise of Sentinel-2 [68]. This validation includes Germany, Spain, the United States, Russia, and Indonesia. These 5 fixed sites were selected because they are distributed across different continents, covering various climate zones, seasons, and environments. The scene selection is rotated monthly for each site in order to observe seasonality. The chosen monthly products are those with cloud cover closest to 20% and with no-data pixels of less than 33%.

The randomized sites were selected in proportion to the number of product requests per continent, comprising 13 annual randomized site scenes from all continents captured in 2023. To select the scenes, the Sentinel-2 archive was consulted in the Google Earth Engine cloud platform. The time range of the products was selected, then filtered by continent, cloud percentage between 15–25%, snow less than 15%, water less than 20%, and no-data pixels less than 10%. Then, random numbers were assigned to each image, and the first image based on these random numbers was selected. Final selection includes three scenes in Asia, four scenes in Europe, two scenes in North America, one scene in South America, two scenes in Africa, and one scene in Australia.

B. TD2: Sentinel-2 High Uncertainty Pixels

This dataset was created for a previous internal validation study by DLR using Sentinel-2 images, presented by Zekoll et al. [69], to test the Fmask, ATCOR, and Sen2Cor masking algorithms. The tests performed on 20 scenes from 2017 to 2019 included a manually labeled reference by a human expert, corresponding to pixels with high uncertainty. The selected

TABLE VIII
OVERVIEW OF THE OBTAINED CLASSIFICATION RESULTS ACROSS ALL TDs

Model		TD1	TD2
PACO		0.855	0.665
CNN1D	<i>No filter</i>	0.867	0.723
	<i>Physics filter</i>	0.877	0.848
	<i>Uniclass filter</i>	0.879	0.764
CNN2D	<i>No filter</i>	0.884	0.730
	<i>Physics filter</i>	0.883	0.755
	<i>Uniclass filter</i>	0.888	0.732
CNN3D	<i>No filter</i>	0.872	0.734
	<i>Physics filter</i>	0.879	0.741
	<i>Uniclass filter</i>	0.874	0.833

The results are expressed in *nMCC*; darker blue indicates higher scores per testing dataset.

pixels for labeling were chosen from those where the three mentioned classifiers disagreed in classification. Therefore, poor performance and low metrics are expected for this dataset, as it does not include average classification cases. Further details and results for this dataset can be found in the aforementioned paper.

IV. RESULTS

This section presents the evaluation of the predictions obtained from the trained models compared with the base pre-classification maps from PACO, hereafter PACO “*hcw*” (see [13]). Table VIII provides an overview of the results obtained for each TD (in terms of *nMCC*, darker blue = higher), produced by the three tested models, CNN1D, 2-D and 3-D, for the three pixelselection strategies: *no filter*, *physics filter*, and *uniclass filter* (see Section II-C). The metrics are then compared in terms of absolute improvement, expressed as percentage points (pp).

Table VIII summarizes the obtained *nMCC* by predicting both test datasets with every trained CNN variant.

The results show that all models outperform PACO’s *hcw* baseline. Models trained with *physics* and *uniclass* pixel selection strategies achieve larger margins of improvement: 1.9–3.3pp on TD1, and from 6.7–18.3pp on TD2.

- 1) On TD1, the best result is 0.888 *nMCC* obtained by CNN2D-*uniclass filter*, a gain of +3.3pp over PACO “*hcw*.”
- 2) On TD2, the largest improvement comes from CNN1D-*uniclass filter* (0.848 *nMCC*, +18.3pp), followed closely by CNN3D-*uniclass filter* (0.833 *nMCC*, +16.8pp)

These gains are observed across the diverse surfaces and atmospheric conditions represented in TD1/TD2. We attribute them to the richer feature extraction of the CNNs and to the integration of interclass relationships within a single model.

TD2 includes many complex or ambiguous pixels, hence baseline scores are lower than for TD1. This gap is narrowed by CNN1D-*physics filter* and CNN3D-*uniclass filter*. These results suggest that 2-D spatial filtering best suits the average cases of TD1 set, whereas spectral focused 1-D kernels cope better the ambiguous pixels in TD2. CNN3D offers a strong but more expensive compromise (≈ 305 k versus 77 k parameters).

TABLE IX
OBTAINED CLASSIFICATION RESULTS FOR TD1 PRESENTED FOR EACH CLASS SEPARATELY

		TD1				
Model		<i>Clear</i>	<i>Atmos</i>	<i>Shadow</i>	<i>Water</i>	<i>Snow</i>
PACO		0.861	0.855	0.769	0.905	0.906
CNN1D	<i>No filter</i>	0.874	0.873	0.809	0.911	0.822
	<i>Physics filter</i>	0.876	0.878	0.834	0.915	0.889
	<i>Uniclass filter</i>	0.872	0.885	0.834	0.916	0.905
CNN2D	<i>No filter</i>	0.899	0.885	0.844	0.910	0.821
	<i>Physics filter</i>	0.898	0.878	0.850	0.915	0.832
	<i>Uniclass filter</i>	0.902	0.881	0.859	0.918	0.861
CNN3D	<i>No filter</i>	0.878	0.873	0.832	0.911	0.825
	<i>Physics filter</i>	0.884	0.877	0.842	0.912	0.878
	<i>Uniclass filter</i>	0.872	0.866	0.845	0.922	0.867

The results are expressed in *nMCC*; darker blue indicates higher scores per class.

TABLE X
OBTAINED CLASSIFICATION RESULTS FOR TD2 PRESENTED FOR EACH CLASS SEPARATELY

		TD2				
Model		<i>Clear</i>	<i>Atmos</i>	<i>Shadow</i>	<i>Water</i>	<i>Snow</i>
PACO		0.630	0.746	0.712	0.895	0.558
CNN1D	<i>No filter</i>	0.702	0.710	0.914	0.905	0.543
	<i>Physics filter</i>	0.857	0.807	0.900	0.896	0.655
	<i>Uniclass filter</i>	0.757	0.794	0.891	0.883	0.577
CNN2D	<i>No filter</i>	0.712	0.749	0.880	0.891	0.556
	<i>Physics filter</i>	0.748	0.758	0.881	0.882	0.561
	<i>Uniclass filter</i>	0.716	0.756	0.879	0.901	0.564
CNN3D	<i>No filter</i>	0.715	0.756	0.898	0.894	0.551
	<i>Physics filter</i>	0.730	0.761	0.885	0.871	0.565
	<i>Uniclass filter</i>	0.843	0.772	0.893	0.877	0.726

The results are expressed in *nMCC*; darker blue indicates higher scores per class.

Pixel selection strategies improve over the *no filter* baseline. On TD2, the *physics filter* lifts CNN1D by +12.5pp, while the *uniclass filter* adds +9.9pp to CNN3D. On TD1 the gains are smaller (≤ 1.2 pp), and CNN2D-*physics filter* shows a marginal drop of -0.1 pp.

The larger gains on TD2 reflect how each filter targets the dominant source of label noise in that set. The *physics* and *uniclass filters* remove pixels whose spectral indices fall near class-decision thresholds, eliminating many mixed or ambiguous spectra; this particularly benefits the spectral feature extraction of CNN1D and CNN3D, which rely on the spectral signature to resolve uncertainty, but also benefits CNN2D in a more limited way. In TD1, the labels are more balanced and represent average classification cases, hence, both filters bring only marginal additional benefit (or a slight loss when informative edge pixels are discarded in the context of spatial feature extraction with CNN2D-*physics filter*).

A. Per-Class Results Analysis

In order to achieve a deeper understanding of the behavior of the proposed pixel selection strategies, the results for each class are presented individually in Tables IX and X, corresponding to

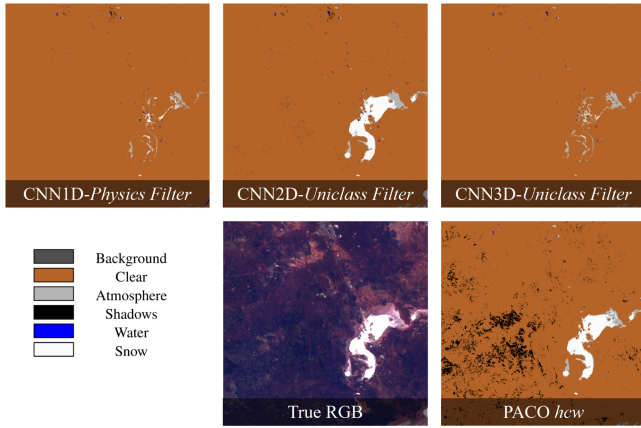


Fig. 6. Qualitative comparison between PACO *hcw* and the prediction from tested models with true RGB image for reference. Image: Lake Lefroy, Australia. Sentinel-2, TD2 T51JUF, 19 August 2018.

TD1 and TD2, respectively. For ease of reading, only significant improvements or losses are discussed in the text.

For both TDs, the biggest improvement is observed for the *shadow* class, where nMCC is considerably improved by 4–9pp for TD1 and 16.7–20.2pp for TD2, consistently across both datasets.

The next best improvement is observed for the *clear* class, improved by 1.1–4.1pp for TD1 and 7.2–22.7pp for TD2, highlighting CNN1D-*physics filter* and CNN3D-*uniclass filter* for TD2, which achieve the best results by a considerable margin. This improvement is correlated with the improvement of the *snow* class in TD2 for the aforementioned CNNs, which is not improved remarkably for any other CNN, and even loses performance in some cases.

The *atmosphere* class is generally improved, except for CNN1D-*no filter* for TD2, with improvements of ≤ 3 pp for TD1 and ≤ 6.1 pp for TD2. On the contrary, the *water* class shows improvements of ≤ 1.3 pp for TD1, but most of the trained models show losses of ≤ 2.4 pp for TD2, except for CNN1D-*no filter* and CNN2D-*uniclass filter*, which show an improvement of ≤ 1 pp.

B. Qualitative Analysis

Because the TDs contain only sparse reference (or ground-truth) labels, this section complements the quantitative scores with a qualitative comparison of classification maps for representative scenes.

The observed improvement of the *snow* class for TD2 is shown in Fig. 6, where a dried salty lake is misclassified as snow and cloud by PACO *hcw*. In addition, most of the *clear* pixels of TD2 are labeled in this scene, causing the observed low nMCC for the *snow* class. The reason this problem frequently occurs is likely due to high-reflectance bright soil being often misclassified as snow or clouds. This issue is almost solved by CNN1D-*physics filter* and CNN3D-*uniclass filter*, where the majority of pixels are correctly classified in the *clear* class, while CNN2D-*uniclass filter* continues misclassifying these pixels. On the other hand,

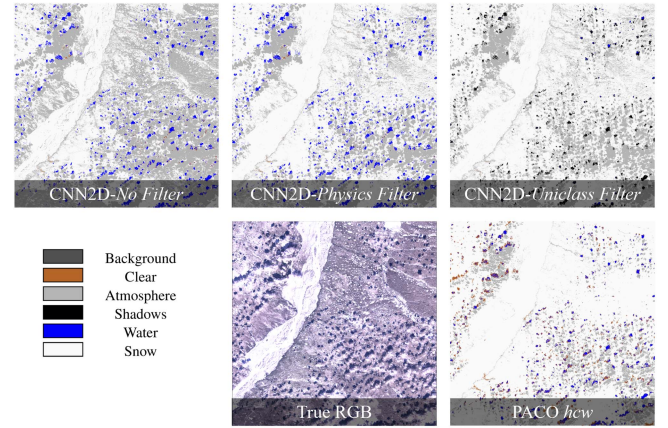


Fig. 7. Qualitative comparison between PACO *hcw* and the CNN2D prediction with true RGB image for reference. Image: Yakutsk, Russia. Sentinel-2, TD1 T52VEP, 16 April 2021.



Fig. 8. Qualitative comparison between PACO *hcw* and the CNN3D-*uniclass filter* prediction with true RGB image for reference. Because the remaining models perform similarly, their classification maps are omitted to avoid redundancy. Image: Indonesia. Sentinel-2, TD1 T48MZT, 22 November 2021.

all models correct the misclassification of shadows caused by low-reflectance soil (black soil).

Fig. 7 shows how the three pixel-selection strategies affect CNN2D. While Table VIII suggests only a 0.5pp spread between the CNN2D variants, the qualitative map reveals larger practical differences. Only the *uniclass filter* variant succeeds in labeling the narrow cloud-shadow pixels that lie on top of the highly reflective snow. By contrast, the *physics* and *no filter* models, like the PACO *hcw* baseline, misclassify these shadow pixels as water.

A persistent compromise of all three CNN variants is their systematic confusion of *snow* with *cloud*. In Table IX, the *snow* class is the only one that consistently loses accuracy. Fig. 7 makes this clear: regions that should be labeled *snow* (and in a few cases even *clear*) are frequently reclassified as *cloud* by the CNNs, whereas the PACO *hcw* baseline keeps them as *snow*. The error is not surprising: fresh *snow* and optically thick *clouds* have similar high-reflectance spectra, making them hard to separate.

Conversely, Fig. 8 highlights the opposite side of the tradeoff. Here, the PACO *hcw* baseline mistakes a bank of optically thick *clouds* for *snow*, whereas all three CNN models correctly assign the *cloud* label.

A case that we would like to highlight is shown in Fig. 9, where a significant amount of *water* pixels are misclassified in the *clear* class. This error is due to a well-known parallax effect that causes a striping signature in Sentinel-2 imagery [79], [80] caused by different viewing angles of onboard sensors. This error is not

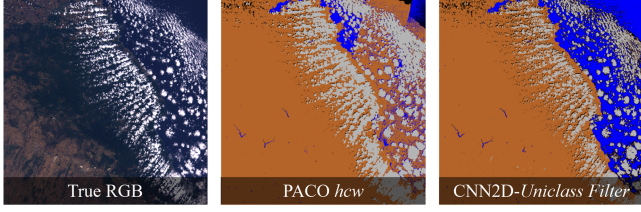


Fig. 9. Qualitative comparison between PACO *hcw* and the CNN2D-*uniclass filter* prediction with true RGB image for reference. Because the remaining models perform similarly, their classification maps are omitted to avoid redundancy. Image: Bulgaria/Turkey. Sentinel-2, TD1 T35TNG, 5 August 2023.

reflected in the quantitative analysis because almost no pixels of TD1 are labeled in this location, but the misclassification is clearly visible. Our approaches solved this problem, which is attributed to the improvement in the generalization of spectral signature recognition.

The gains in performance for all trained models are nevertheless consistent, and most evident for the models trained with the *physics* and *uniclass* pixel-selection strategies. Outside the presented scenes, most classification maps look visually similar to the PACO *hcw* baseline, with only modest refinements. More information about the best performing classifiers can be found in the Appendix, which includes the confusion matrices of CNN2D-*uniclass filter* for TD1 and CNN1D-*physics filter* for TD2, compared to the baseline PACO *hcw*. Per-class Precision and Recall calculations are also included to support the interpretation of class-level over- or underestimation.

V. DISCUSSION

This work proved that by selecting high-quality pixels from a given multispectral DB with reference labels obtained by thresholding spectral indices from the atmospheric corrector PACO, classification performance can be improved on independent test scenes by +3.3pp on TD1 and +18.3pp on TD2 in terms of nMCC (see Table VIII) compared with the operational baseline *hcw* classification map from PACO. Notably, the greatest improvement was observed for the pixel selection using the *uniclass filter* for CNN2D and using the *physics filter* for CNN1D, which both discard high uncertainty pixels for composing the training set. This improvement is particularly evident in cases where pixels are mixed at object borders or where the spectrum is similar, such as the *shadow* and *clear* classes, with improvements of 4–20.2 pp and 1.1–22.7 pp, respectively, thanks to the advanced feature extraction capabilities of the deep learning models tested in this work. In addition, the rest of classes performed acceptably, with moderate improvements of ≤ 6.1 pp, but losing in the snow class for TD1 up to 7.4pp.

Although qualitative analysis is subjective, some regions demonstrated clearly visible improvements that were not included in the TDs, potentially concealing the true classification performance of the proposed models.

A. Effect of Pixel-Selection Strategy

The presented experiments with different training dataset selection strategies demonstrated significant changes in the

behavior of the implemented classifier, and even if it is not reflected in the nMCC metric, the qualitative analysis demonstrated a clear improvement. For example, CNN2D results for TD1 appear similar in Table VIII, but a noticeable improvement can be observed in Fig. 7 for CNN2D-*uniclass filter*. The improvements are more remarkable for TD2, which contains challenging classification cases, where the *physics filter* supports the spectral feature extraction of CNN1D, while the *uniclass filter* benefits the joint spatial-spectral feature extraction of CNN3D.

Despite the significant feature extraction capabilities of CNNs, their performance heavily depends on the quality of the training set and its reference labels. The variety of TDs evaluated in this work highlighted that the training dataset is limited by the lack of representation for all possible subclasses under diverse atmospheric conditions, illumination angles, and noisy labels. Consequently, fully representing all classes for previously unseen pixels remains a significant challenge.

B. Dataset Coverage and Computational Complexity

To further improve this approach, additional sources of information should be incorporated to reduce label noise and enhance the representation of each class. From a global application perspective, this remains challenging due to the computational and spatial complexity of training deep learning models. This highlights the need for advanced data selection strategies and novel training approaches to fully exploit the potential of multispectral imagery, and by extension, hyperspectral imagery. In addition, incorporating supplementary features can improve data characterization, as deep learning models are capable of learning patterns from diverse sources. For instance, viewing and solar angles could help distinguish between cloud or topographic shadows and clear-sky land under low illumination conditions. This differentiation can be further supported by spatial information. For example, if a shadowed region is adjacent to a detected cloud and the solar-viewing geometry is consistent, it is likely that the shadow belongs to that cloud.

We observed that CNN1D, with only 77 k parameters, achieves outstanding performance compared with CNN2D (155 k parameters) and CNN3D (305 k parameters), highlighting the importance of pure-spectral feature extraction. In our prototype, the wall-clock time (per epoch) of CNN2D and CNN3D is nearly identical due to the custom data generator, which dominates total runtime by creating batches of spatial patches, hiding the theoretical differences on computational complexity between 2-D and 3-D convolutional layers [81]. We leave an in-depth computational cost analysis for future work because of the current implementation's limitations.

Since we aimed to avoid introducing any information from the TD into the training process and to study methods for obtaining a generalized model applicable to any case, we observed that increasing the number of training pixels or the number of training epochs, despite the validation set improving slightly with each epoch, does not result in better performance for the TDs. This behavior is likely because enlarging the training set also increases the likelihood of outliers or nonrepresentative samples. In addition, by introducing highly correlated pixels

corresponding to average cases, the training process may overfit the original model, replicating its behavior, and losing the generalization that comes from a sparse input. However, these are only hypotheses and should be further studied.

To strengthen domain generalization, the sample size must grow, but instead of including more pixels within the same limited number of scenes, this should be done by selecting an extensive set of additional scenes spanning diverse locations, biomes, seasons, and atmospheric conditions. Capturing a wider spectroscopic variability should yield a more representative training set and, therefore, improved generalization.

C. Testing Protocols

The per-class classification analysis reveals the biases of the classifier. Since each dataset provides slightly different conclusions, it is evident that the independently and manually labeled TDs have varying class definitions or are not uniformly distributed across all subclasses. Achieving such uniformity is nearly impossible with manual labeling due to the significant workforce required.

In addition, more robust testing methods for automatically retrieving highly sensitive reference data should be developed to achieve completeness and avoid biases, as well as by incorporating other missions specifically designed for the atmospheric phenomenon study that retrieve physical parameters for comparison [82]. Furthermore, a harmonized definition based on physical parameters should be established for each class, considering a compilation of the requirements from a wide range of applications that make use of the distributed L2 products. This would enable comparisons across classifiers and allow the interpretation of the generalization for global-scale applications.

D. Future Directions: Explainable Artificial Intelligence (XAI) and Physics-Informed Neural Networks (PINNs)

Future work will consider the use of XAI [83], [84], [85], [86], [87], to identify the most important features for classification in each class, since the spectral characteristics are basic features for pixel classification. In addition, the potential incorporation of well-studied physics models from other remote sensing areas will be explored through the use of PINNs [88], [89]. This integration aims to build robust models, not relying only on a statistical perspective, addressing the issue of sparse subclass representation while still leveraging the complex spatial and spectral feature extraction capabilities of deep learning models.

E. Scalability to Other Sensors

Regarding the scalability of this work, while this study focuses on Sentinel-2 imagery, the proposed methodology could be applied to other multispectral or hyperspectral sensors (by training a model for each sensor), as long as a preclassification input is available for training. While the current implementation is built using PACO, which supports Sentinel-2 [12], Landsat-8 [9], [10], EnMAP [2], [3], and DESIS [7], [8], the physical rules used for pixel selection are sensor-independent and derived from fundamental reflectance-based principles. In cases where a

different masking algorithm is used to generate training masks, the same logic can be adapted to new class definitions. Further evaluation on additional optical remote sensing missions with global-scale implementation is left for future work.

VI. CONCLUSION

This work employed the concept of 1-D, 2-D, and 3-D CNNs trained with the masking products (masks cube) of the atmospheric corrector PACO, achieving a higher classification performance compared with the masking algorithm based on spectral tests from the atmospheric processor PACO. This improvement is enabled by the advanced spectral, spatial, and joint spatial-spectral feature extraction capabilities of the CNNs and the training pixel selection strategies based on physics-informed rules and the ci of a pixel belonging to a class. The results suggest that this approach is particularly effective for challenging classification cases, such as when spectral signatures are similar or when pixels are mixed at object borders.

The learned techniques from this work can be applied in the near future to further applications using hyperspectral imagery from in-operation missions like EnMAP [2], [3] and DESIS [7], [8]. Given that the results presented in this work for multispectral imagery from Sentinel-2 are limited by spectral information, we expect that increasing the number of spectral bands with hyperspectral imagery will allow this deep learning approach to fully exploit all newly available features.

Addressing the case of recently launched missions, although this methodology requires data for training the model, the imagery obtained during the commissioning phase should be sufficient to set up an operational model. The reduced time frame makes it very challenging to manually label enough data to obtain a generalized model; therefore, there is a need to retrieve high-quality labels automatically, such as the one presented in this work.

However, some constraints were observed, including the significant impact of the quality of training datasets, which must include representative Earth's surfaces under varying atmospheric conditions. Furthermore, a harmonized definition of classes is necessary for testing, allowing TDs to provide a more generalized interpretation of the results, especially if they are based on physical parameters characteristic of the mask itself (e.g., snow grain size, cloud altitude, albedo, among others). The techniques learned will help define guidelines for the further development of robust and scalable methodologies for deep learning approaches in optical remote sensing missions with a global-scale application perspective.

A. Proposed Future Work

- 1) The automatic compilation of extensive training datasets with high-quality labels.
- 2) The development of new testing methods based on physical parameters and the harmonization of class definitions accordingly.
- 3) The integration of advanced training techniques, such as the use of XAI and PINNs.

TABLE XI
CONFUSION MATRICES FOR TD1

		TD1 - Reference labels							
Model	Class	Clear	Atmos	Shadows	Water	Snow	P	R	
CNN2D	Clear	89.1	10.0	10.9	2.4	0.0	0.86	0.89	
	Atmos	7.6	85.0	11.1	6.8	10.0	0.83	0.85	
	Shadows	1.8	0.2	71.0	11.8	0.0	0.81	0.71	
	Water	0.8	0.0	3.7	79.0	0.0	0.93	0.79	
	Snow	1.0	4.7	3.4	0.1	89.9	0.60	0.90	
PACO	Clear	91.4	16.2	30.0	7.6	0.2	0.75	0.91	
	Atmos	6.0	78.3	16.5	6.6	0.4	0.82	0.78	
	Shadows	2.0	0.1	43.9	7.2	0.0	0.78	0.44	
	Water	0.5	0.0	8.8	78.6	0.0	0.89	0.79	
	Snow	0.1	5.3	0.8	0.0	99.4	0.68	0.99	
Total		637,991	577,872	267,190	274,884	71,124			

Top: CNN2D-uniclass filter; bottom: PACO hcw baseline. Each cell is column percentage; darker blue = higher.

TABLE XII
CONFUSION MATRICES FOR TD2

		TD2 - Reference labels							
Model	Class	Clear	Atmos	Shadows	Water	Snow	P	R	
CNN1D	Clear	92.2	20.3	10.3	5.6	25.7	0.97	0.92	
	Atmos	4.6	77.9	0.6	0.0	20.8	0.53	0.78	
	Shadows	1.3	1.1	85.4	22.6	0.1	0.78	0.85	
	Water	0.1	0.3	0.9	71.0	1.7	0.89	0.71	
	Snow	1.8	0.5	2.8	0.8	51.7	0.20	0.52	
PACO	Clear	57.4	15.1	25.9	25.9	24.7	0.93	0.57	
	Atmos	8.5	75.8	3.6	0.2	2.5	0.37	0.76	
	Shadows	10.8	0.3	67.3	4.2	0.0	0.34	0.67	
	Water	0.1	0.3	0.5	68.9	1.2	0.91	0.69	
	Snow	23.3	8.6	2.7	0.7	71.6	0.03	0.72	
Total		872,169	59,803	72,961	26,300	8,781			

Top: CNN1D-physics filter; bottom: PACO hcw baseline. Formatting as in Table XI.

- 4) The implementation of the proposed methodology for hyperspectral missions, where the high number of spectral bands reduces the effectiveness of traditional threshold-based methods, and spatial-spectral learning can offer a more scalable alternative.

APPENDIX

To better understand the improvements, errors, and class-wise relationships, we compare the confusion matrices of the best-performing CNN models for each test set against the PACO hcw baseline. For TD1, the top performer is CNN2D-uniclass filter, whereas for TD2 it is CNN1D-physics filter. Tables XI and XII show, for each model, the percentage of pixels predicted in every class (rows) versus the reference labels (columns). The last two columns report class-wise precision (P) and recall (R), allowing over- and under-estimation to be read directly.

According to the confusion matrix in Table XI, the CNN2D model reduces a considerable number of errors that the PACO hcw baseline makes between shadow pixels classified as clear: shadow recall rises from 0.44 to 0.71, and precision for the clear class increases from 0.75 to 0.86. The cost of this gain is an increase in the number of snow pixels misclassified as atmosphere, with recall dropping from 0.99 to 0.90 and precision from 0.68 to 0.60.

In Table XII, the CNN1D model largely removes misclassification in the clear class, improving recall from 0.57 to 0.92. On the other hand, shadow precision goes up from 0.34 to 0.78, with recall rising from 0.67 to 0.85. The main remaining weakness is that some snow pixels are labeled as atmosphere, so snow recall

drops from 0.72 to 0.52, although their precision increases from 0.03 to 0.20.

ACKNOWLEDGMENT

The authors would like to thank the computational and data resources provided through the joint high-performance data analytics (HPDA) project “terabyte” of the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ). They also would like to thank the supervisors-D. Torres, R. de los Reyes, and K. Alonso-for their invaluable feedback and support throughout this research.

REFERENCES

- [1] W. Emery and A. Camps, *Introduction to Satellite Remote Sensing: Atmosphere, Ocean, Land and Cryosphere Applications*. Amsterdam, The Netherlands: Elsevier, 2017.
- [2] L. Guanter et al., “The EnMAP spaceborne imaging spectroscopy mission for Earth observation,” *Remote Sens.*, vol. 7, no. 7, pp. 8830–8857, 2015. [Online]. Available: <https://www.mdpi.com/2072-4292/7/7/8830>
- [3] T. Storch et al., “The EnMAP imaging spectroscopy mission towards operations,” *Remote Sens. Environ.*, vol. 294, 2023, Art. no. 113632.
- [4] M. Urai et al., “Initial onboard calibration results of the HISUI hyperspectral sensor,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1608–1610.
- [5] D. S. Connelly et al., “The EMIT mission information yield for mineral dust radiative forcing,” *Remote Sens. Environ.*, vol. 258, 2021, Art. no. 112380.
- [6] S. Cogliati et al., “The PRISMA imaging spectroscopy mission: Overview and first performance analysis,” *Remote Sens. Environ.*, vol. 262, 2021, Art. no. 112499.
- [7] A. Eckardt et al., “DESI (DLR Earth sensing imaging spectrometer for the ISS-MUSES platform),” in *Proc. Int. Geosci. Remote Sens. Symp.*, Nov. 2015, pp. 1457–1459.
- [8] K. Alonso et al., “Data products, quality and validation of the DLR earth sensing imaging spectrometer (DESI),” *Sensors*, vol. 19, no. 20, Oct. 2019, Art. no. 4471, doi: [10.3390/s19204471](https://doi.org/10.3390/s19204471).
- [9] J. A. Barsi, K. Lee, G. Kvaran, B. L. Markham, and J. A. Pedelty, “The spectral response of the Landsat-8 operational land imager,” *Remote Sens.*, 2014, vol. 6, no. 10, pp. 10232–10251, 2014. [Online]. Available: <https://www.mdpi.com/2072-4292/6/10/10232>
- [10] E. Vermote, C. Justice, M. Claverie, and B. Franch, “Preliminary analysis of the performance of the landsat 8/OLI land surface reflectance product,” *Remote Sens. Environ.*, vol. 185, pp. 46–56, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425716301572>
- [11] J. G. Masek et al., “Landsat 9: Empowering open science and applications through continuity,” *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 111968.
- [12] M. Drusch et al., “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.
- [13] R. de los Reyes et al., “PACO: Python-based atmospheric correction,” *Sensors*, vol. 20, no. 5, Mar. 2020, Art. no. 1428, doi: [10.3390/s20051428](https://doi.org/10.3390/s20051428).
- [14] J. Nieke et al., “The copernicus hyperspectral imaging mission for the environment (CHIME): An overview of its mission, system and planning status,” in *Sensors, Systems, and Next-Generation Satellites XXVII*. Bellingham, WA, USA: SPIE, 2023, pp. 21–40.
- [15] M. Celesti et al., “The copernicus hyperspectral imaging mission for the environment (CHIME): Status and planning,” in *Proc. Int. Geosci. Remote Sens. Symp.*, 2022, pp. 5011–5014.
- [16] D. R. Thompson et al., “Ongoing progress toward NASA’s surface biology and geology mission,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 5007–5010.
- [17] S. Skakun et al., “Cloud mask intercomparison exercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2,” *Remote Sens. Environ.*, vol. 274, Jun. 2022, Art. no. 112990.
- [18] R. Richter and D. Schl pfer, “Atmospheric/Topographic Correction for Satellite Imagery (ATCOR-2/3 User Guide, Version 9.3.0),” *ReSe Appl. LLC*, Wil, Switzerland, Manual, pp. 1–75, Oct. 2019.
- [19] R. Richter and D. Schl pfer, “Atmospheric and Topographic Correction (ATCOR Theoretical Background Document),” *DLR IB*, Stuttgart and Augsburg, Germany, Document 564, 2019, Art. no. 3.

- [20] G. Mateo-García, V. Laparra, D. López-Puigdollers, and L. Gómez-Chova, "Transferring deep learning models for cloud detection between Landsat-8 and Proba-V," *ISPRS J. Photogrammetry Remote Sens.*, vol. 160, pp. 1–17, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271619302801>
- [21] D. López-Puigdollers, G. Mateo-García, and L. Gómez-Chova, "Benchmarking deep learning models for cloud detection in Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 5, Mar. 2021, Art. no. 992, doi: [10.3390/rs13050992](https://doi.org/10.3390/rs13050992).
- [22] S. Foga et al., "Cloud detection algorithm comparison and validation for operational Landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425717301293>
- [23] S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery," *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111205. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425719302172>
- [24] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images," *Remote Sens. Environ.*, vol. 159, pp. 269–277, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425714005069>
- [25] D. Frantz, "Force—Landsat Sentinel-2 analysis ready data and beyond," *Remote Sens.*, vol. 11, no. 9, May 2019, Art. no. 1124, doi: [10.3390/rs11091124](https://doi.org/10.3390/rs11091124).
- [26] D. Frantz, E. Haß, A. Uhl, J. Stoffels, and J. Hill, "Improvement of the fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects," *Remote Sens. Environ.*, vol. 215, pp. 471–481, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425718302037>
- [27] D. Frantz, A. Röder, M. Stellmes, and J. Hill, "An operational radiometric landsat preprocessing framework for large-area time series applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3928–3943, Jul. 2016. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=s2.0-84960539907&doi=10.1109%2Ftgrs.2016.2530856&partnerID=40&md5=8dbf97c027e46854bcfbdc2619e68695>
- [28] J. Wevers, D. Müller, J. Scholze, G. Kirches, R. Quast, and C. Brockmann, "IdePix for Sentinel-2 MSI algorithm theoretical basis document," *Zenodo*, Tech. Rep., Dec. 17, 2021, pp. 1–25, doi: [10.5281/zenodo.5788067](https://doi.org/10.5281/zenodo.5788067).
- [29] A. Zupanc, "Improving cloud detection with machine learning," *Planet Stories (Medium)*, Dec. 19, 2017. [Online]. Available: <https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13>
- [30] J. Puc and L. Žust, "On cloud detection with multi-temporal data," *Planet Stories (Medium)*, Oct. 14, 2019. [Online]. Available: <https://medium.com/sentinel-hub/on-cloud-detection-with-multi-temporal-data-f64f9b8d59e5>
- [31] S. Skakun, E.F. Vermote, J.-C. Roger, C.O. Justice, and J. G. Masek, "Validation of the LaSRC cloud detection algorithm for Landsat 8 images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2439–2446, Jul. 2019.
- [32] S. Skakun, E.F. Vermote, A.E.S. Artigas, W.H. Rountree, and J.-C. Roger, "An experimental sky-image-derived cloud validation dataset for Sentinel-2 and Landsat 8 satellites over NASA GSFC," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 95, 2021, Art. no. 102253. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243420308965>
- [33] O. Hagolle, M. Huc, D. V. Pascual, and G. Dedieu, "A multi-temporal method for cloud detection, applied to FORMOSAT-2, VEN μ S, LANDSAT and SENTINEL-2 images," *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1747–1755, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425710000908>
- [34] O. Hagolle, M. Huc, C. Desjardins, S. Auer, and R. Richter, "MAJA algorithm theoretical basis document," *Zenodo*, Tech. Rep., Dec. 7, 2017, pp. 1–40, doi: [10.5281/zenodo.1209633](https://doi.org/10.5281/zenodo.1209633).
- [35] J. Louis et al., "SENTINEL-2 SEN2COR: L2A processor for users," in *Proc. Living Planet Symp.*, 2016, pp. 1–8.
- [36] J. Louis, O. Devignot, and L. Pessiot, "Level-2A algorithm theoretical basis document," *ESA*, Tech. Rep. S2-PDGS-MPC-L2A-ATBD-V2.10.0, 2021, pp. 1–78.
- [37] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.
- [38] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," in *ISPRS J. Photogrammetry Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.
- [39] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417420310836>
- [40] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
- [41] S. Jozdani, D. Chen, D. Pouliot, and B. A. Johnson, "A review and meta-analysis of generative adversarial networks and their applications in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102734. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243422000605>
- [42] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of hyperspectral images based on multiclass spatial-spectral generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5329–5343, Aug. 2019.
- [43] A. A. Aleissae et al., "Transformers in remote sensing: A survey," *Remote Sens.*, vol. 15, no. 7, Apr. 2023, Art. no. 1860, doi: [10.3390/rs15071860](https://doi.org/10.3390/rs15071860).
- [44] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, pp. 5514715.
- [45] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, Feb. 2021, Art. no. 516, doi: [10.3390/rs13030516](https://doi.org/10.3390/rs13030516).
- [46] S. Ghaffarian, J. Valente, M. van der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sens.*, vol. 13, no. 15, Aug. 2021, Art. no. 2965, doi: [10.3390/rs13152965](https://doi.org/10.3390/rs13152965).
- [47] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.
- [48] D. Tuia et al., "Artificial intelligence to advance Earth observation: A review of models, recent trends, and pathways forward," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–25, 2024, doi: [10.1109/MGRS.2024.3425961](https://doi.org/10.1109/MGRS.2024.3425961).
- [49] G. Sumbul and B. Demir, "Label noise robust image representation learning based on supervised variational autoencoders in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 646–649.
- [50] T. Burgert, M. Ravanbakhsh, and B. Demir, "On the effects of different types of label noise in multi-label remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5413713.
- [51] G. Sumbul and B. Demir, "Generative reasoning integrated label noise robust deep image representation learning," *IEEE Trans. Image Process.*, vol. 32, pp. 4529–4542, 2023.
- [52] A. K. Aksoy, M. Ravanbakhsh, and B. Demir, "Multi-label noise robust collaborative learning for remote sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6438–6451, May 2024.
- [53] A. K. Aksoy, M. Ravanbakhsh, T. Kreuziger, and B. Demir, "A consensual collaborative learning method for remote sensing image classification under noisy multi-labels," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 3842–3846.
- [54] Eur. Environ. Agency, "Corine land cover—CLC 2018 technical guidelines," European Environment Agency, Copenhagen, Denmark, Tech. Rep. EEA Report No 07/2018, 2018. [Online]. Available: <https://land.copernicus.eu/en/technical-library/clc-2018-technical-guidelines/download/file>
- [55] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.
- [56] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [57] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, Jan. 2017, Art. no. 67, doi: [10.3390/rs9010067](https://doi.org/10.3390/rs9010067).
- [58] E. Padilla-Zepeda, K. Alonso, R. De Los Reyes, D. Torres-Roman, and A. P. Pertiwi, "A deep learning approach for imagery masking of spectral sensors," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2024, pp. 8541–8546.
- [59] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, 2015.

- [60] P. Jia, M. Zhang, W. Yu, F. Shen, and Y. Shen, "Convolutional neural network based classification for hyperspectral data," in *Proc. Int. Geosci. Remote Sens. Symp.*, Nov. 2016, pp. 5075–5078.
- [61] E. Padilla-Zepeda, D. Torres-Roman, and A. Mendez-Vazquez, "A semantic segmentation framework for hyperspectral imagery based on Tucker decomposition and 3DCNN tested with simulated noisy scenarios," *Remote Sens.*, vol. 15, no. 5, Mar. 2023, Art. no. 1399, doi: [10.3390/rs15051399](https://doi.org/10.3390/rs15051399).
- [62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28
- [63] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1818–1827.
- [64] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019.
- [65] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, 2017, Art. no. 35.
- [66] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.
- [67] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.
- [68] ESA S2 MSI ESL Team, "Sentinel-2 annual performance report-year 2023," ESA, Paris, France, Tech. Rep. OMPC-CS.APR.003, 2024.
- [69] V. Zekoll et al., "Comparison of masking algorithms for sentinel-2 imagery," *Remote Sens.*, vol. 13, no. 1, Jan. 2021, Art. no. 137, doi: [10.3390/rs13010137](https://doi.org/10.3390/rs13010137).
- [70] B. Holben et al., "AERONET—A federated instrument network and data archive for aerosol characterization," *Remote Sens. Environ.*, vol. 66, no. 1, pp. 1–16, 1998. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425798000315>
- [71] L. Cenci, M. Galli, C. Santella, V. Boccia, and C. Albinet, "Analyzing the impact of the different instances of the copernicus DEM dataset on the orthorectification of VHR optical data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 6001–6004.
- [72] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices," *Remote Sens.*, vol. 16, no. 3, Feb. 2024, Art. no. 533, doi: [10.3390/rs16030533](https://doi.org/10.3390/rs16030533).
- [73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [74] J. Gorodkin, "Comparing two k-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, no. 5, pp. 367–374, 2004. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1476927104000799>
- [75] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," Aug. 13, 2020, *arXiv:2008.05756*. [Online]. Available: <http://arxiv.org/abs/2008.05756>
- [76] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0005279575901099>
- [77] A. Gonzalez-Ramirez, J. Lopez, D. Torres-Roman, and I. Yañez-Vargas, "Analysis of multi-class classification performance metrics for remote sensing imagery imbalanced datasets," *ECORFAN J. Quantitative Statist. Anal.*, vol. 8, pp. 11–17, 2021. [Online]. Available: www.ecorfan.org/bolivia
- [78] ESA, "Sentinel-2 Level-2A Algorithm Theoretical Basis Document (ATBD)," ESA, Paris, France, Tech. Rep. S2-PDGS-MPC-ATBD-V2.10.0, Nov. 17, 2021.
- [79] C. B. Lee, D. Traganos, and P. Reinartz, "A simple cloud-native spectral transformation method to disentangle optically shallow and deep waters in Sentinel-2 images," *Remote Sens.*, vol. 14, no. 3, Feb. 2022, Art. no. 590, doi: [10.3390/rs14030590](https://doi.org/10.3390/rs14030590).
- [80] MPC, "Data quality report," ESA, Paris, France, Tech. Rep. S2-PDGS-MPC-DQR, Mar. 2018.
- [81] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
- [82] E. Padilla-Zepeda et al., "Cross-mission methodology for masking validation: EnMAP cloud mask with Sentinel-5p," in *Proc. 3rd Workshop Int. Cooperation Spaceborne Imag. Spectrosc.*, Nov. 2024, pp. 1–1. [Online]. Available: <https://elib.dlr.de/209086/>
- [83] A. Höhl et al., "Opening the black box: A systematic review on explainable artificial intelligence in remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 4, pp. 261–304, Dec. 2024.
- [84] S. O. Atik and M. E. Atik, "Optimal band selection using explainable artificial intelligence for machine learning-based hyperspectral image classification," *J. Appl. Remote Sens.*, vol. 18, no. 4, 2024, Art. no. 42604.
- [85] D. E. Turan, E. Aptoula, A. Ertürk, and G. Taskin, "Interpreting hyperspectral remote sensing image classification methods via explainable artificial intelligence," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5950–5953.
- [86] A. H. Oveis, E. Giusti, G. Meucci, S. Ghio, and M. Martorella, "Explainability in hyperspectral image classification: A study of XAI through the shap algorithm," in *Proc. 13th Workshop Hyperspectral Imag. Signal Processing: Evol. Remote Sens.*, 2023, pp. 1–5.
- [87] G. De Lucia, M. Lapegna, and D. Romano, "Towards explainable AI for hyperspectral image classification in edge computing environments," *Comput. Elect. Eng.*, vol. 103, 2022, Art. no. 108381. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0045790622005985>
- [88] M. H. Dehghan-Shoar, G. Kereszturi, R. R. Pullanagari, A. A. Orsi, I. J. Yule, and J. Hanly, "A physically informed multi-scale deep neural network for estimating foliar nitrogen concentration in vegetation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 130, 2024, Art. no. 103917. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1569843224002711>
- [89] B. Li, Q. Hu, M. Gao, T. Liu, C. Zhang, and C. Liu, "Physical informed neural network improving the WRF-CHEM results of air pollution using satellite-based remote sensing data," *Atmospheric Environ.*, vol. 311, 2023, Art. no. 120031. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1352231023004570>



Efrain Padilla-Zepeda received the bachelor's degree in mechatronics engineering from the Technical and Industrial Teaching Center (CETI), Guadalajara, Mexico, in 2018 and the M.Sc. degree in electrical engineering from the Center for Research and Advanced Studies of the National Polytechnic Institute (Cinvestav), Mexico City, Mexico, in 2020. He is currently working toward the Ph.D. degree in electrical engineering with Cinvestav, in collaboration with the German Aerospace Center (DLR), Wessling, Germany.

He was an Assistant Professor for Engineering undergraduate students with the Autonomous University of Guadalajara (UAG), Guadalajara, and the University of Digital Arts (UAD), Guadalajara, where he taught courses on numerical methods, complexity analysis of algorithms, linear algebra, and discrete mathematics. His research interests include optical remote sensing imagery classification, deep learning, hyperspectral image processing, atmospheric correction, feature engineering, noise modeling, dimensionality reduction, and tensor algebra.



Kevin Alonso received the Technical Telecommunication Engineering Bachelor's degree in communication systems and the Telecommunication Engineering Master's degree in telematic systems from the Bilbao School of Engineering, Bilbao, Spain, in 2005 and 2007, respectively, the master's degree in space science and technology, and the Ph.D. (Dr.-Ing.) degree from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2017, respectively. His Ph.D. dissertation was titled "Heterogeneous Data Mining of Earth Observation Archives: Integration and Fusion of Images, Maps, and In Situ Data."

From July 2008 to March 2013, he was as a Staff Researcher with Vicomtech-4K, San Sebastian, Spain, where he worked in the field of digital television and multimedia services, participating in the European Research Project RUSHES. During 2009 and 2010, he combined his research work on semantic web and geographic information systems (GIS) technologies with the study of the master's degree in space science and technology. From 2013 to 2022, he was with the Department of Photogrammetry and Image Analysis, German Aerospace Center (DLR), Wessling, Germany. In September 2022, he joined Starion Group as a Mission Engineer to support the development of the Copernicus Expansion Missions CHIME and LSTM, ESA-ESRIN, Frascati, Italy. In January 2025, he joined ESA as Earth Observation Engineer developing Copernicus Satellite operational processors within the ESA's Earth Observation Framework.

Raquel De Los Reyes received the Ph.D. degree in physics from the Complutense University of Madrid, Madrid, Spain, in 2008.

She is currently an expert in the atmospheric correction of optical remote sensing data. She is also a Research Scientist with the Department of Photogrammetry and Image Analysis, Remote Sensing Technology Institute (MF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. She has more than 22 years of experience in atmospheric correction, data quality, and calibration, 15 of those in the field of ground-based astronomy and 7 years in the field of Earth observation. During this time, she has been a Developer and Manager of software packages for the calibration of the atmospheric influence on scientific data and its quality. She is currently a Software Manager of the python-based PACO software developed with DLR for atmospheric correction of multispectral and hyperspectral optical remote sensing data. She is also the Subsystem Manager of L2A processor of the DESIS and the EnMAP hyperspectral missions and Project Manager of the CHIME L2 processor.



Deni Torres-Roman received the Graduate degree in telecommunications engineer from the University of Oriente, Santiago de Cuba, Cuba, in 1972, and the Ph.D. degree in telecommunications from the University of Dresden, Dresden, Germany, in 1986.

He was a Professor with the University of Oriente. Since 1996, he has been a Researcher 3-C with Cinvestav, Guadalajara Unit, Zapopan, Mexico, where his areas of interest have included the development of low- and high-speed switches, software development for telecommunications, and the mathematical

modeling of the Internet. His current main areas of interest are video and image processing for intelligent transportation systems and the processing of multi- and hyperspectral remote sensing images.

Prof. Deni was the recipient of the Telecommunication Research Prize in 1993 from AHCIET Association and the 1995 Best Paper Award from AHCIET Review, Spain.

Avi Putri Pertiwi received the B.Sc. degree in ocean engineering from the Bandung Institute of Technology, Bandung, Indonesia, in 2015, and the M.Sc. degree in environmental engineering from the Technical University of Munich (TUM), Munich, Germany, in 2019.

She is currently a Researcher with the German Aerospace Center (DLR), Remote Sensing Technology Institute, Berlin, Germany. Her research interests include remote sensing, coastal ecosystem monitoring, and machine learning applications in geospatial analysis.

Tobias Storch received the Diploma (Dipl.-Inform.) and Doctorate (Dr. rer. nat.) degrees in computer science from the University of Dortmund, Dortmund, Germany, in 2003 and 2007, respectively.

He is currently the Head of Department Imaging Spectroscopy, Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany, specializing in hyperspectral remote sensing. From 2009 to 2023, he was the Ground Segment Project Manager for the EnMAP mission with DLR. In 2024, he was a DLR Senior Scientist.