remote sensing

MDPI

Article

KappaMask: AI-Based Cloudmask Processor for Sentinel-2

Marharyta Domnich [1,2,*], Indrek Sünter [1], Heido Trofimov [1], Olga Wold [1], Fariha Harun [1], Anton Kostiukhin [1], Mihkel Järveoja [1], Mihkel Veske [1], Tanel Tamm [1], Kaupo Voormansik [1,3], Aire Olesk [3], Valentina Boccia [4], Nicolas Longepe [4] and Enrico Giuseppe Cadau [4]

[1] KappaZeta Ltd., 51007 Tartu, Estonia; indrek.sunter@kappazeta.ee (I.S.); heido.trofimov@kappazeta.ee (H.T.); olga.wold@kappazeta.ee (O.W.); fariha.harun@kappazeta.ee (F.H.); anton.kostiukhin@kappazeta.ee (A.K.); mihkel.jarveoja@kappazeta.ee (M.J.); mihkel.veske@kappazeta.ee (M.V.); tanel.tamm@kappazeta.ee (T.T.); kaupo.voormansik@kappazeta.ee (K.V.)

[2] Institute of Computer Science, University of Tartu Estonia, 51009 Tartu, Estonia

[3] Tartu Observatory, University of Tartu, 61602 Tõravere, Estonia; aire.olesk@ut.ee

[4] European Space Agency, ESA-ESRIN, Largo Galileo Galilei, 1, 00044 Frascati, RM, Italy; valentina.boccia@esa.int (V.B.); Nicolas.Longepe@esa.int (N.L.); enrico.cadau@esa.int (E.G.C.)

* Correspondence: info@kappazeta.ee

check for updates

Citation: Domnich, M.; Sünter, I.; Trofimov, H.; Wold, O.; Harun, F.; Kostiukhin, A.; Järveoja, M.; Veske, M.; Tamm, T.; Voormansik, K.; et al. KappaMask: AI-Based Cloudmask Processor for Sentinel-2. *Remote Sens.* **2021**, *13*, 4100. https://doi.org/10.3390/rs13204100

Academic Editors: Claudio Piciarelli, Hyungtae Lee and Sungmin Eum

Received: 19 August 2021
Accepted: 8 October 2021
Published: 13 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: The Copernicus Sentinel-2 mission operated by the European Space Agency (ESA) provides comprehensive and continuous multi-spectral observations of all the Earth's land surface since mid-2015. Clouds and cloud shadows significantly decrease the usability of optical satellite data, especially in agricultural applications; therefore, an accurate and reliable cloud mask is mandatory for effective EO optical data exploitation. During the last few years, image segmentation techniques have developed rapidly with the exploitation of neural network capabilities. With this perspective, the KappaMask processor using U-Net architecture was developed with the ability to generate a classification mask over northern latitudes into the following classes: clear, cloud shadow, semi-transparent cloud (thin clouds), cloud and invalid. For training, a Sentinel-2 dataset covering the Northern European terrestrial area was labelled. KappaMask provides a 10 m classification mask for Sentinel-2 Level-2A (L2A) and Level-1C (L1C) products. The total dice coefficient on the test dataset, which was not seen by the model at any stage, was 80% for KappaMask L2A and 76% for KappaMask L1C for clear, cloud shadow, semi-transparent and cloud classes. A comparison with rule-based cloud mask methods was then performed on the same test dataset, where Sen2Cor reached 59% dice coefficient for clear, cloud shadow, semi-transparent and cloud classes, Fmask reached 61% for clear, cloud shadow and cloud classes and Maja reached 51% for clear and cloud classes. The closest machine learning open-source cloud classification mask, S2cloudless, had a 63% dice coefficient providing only cloud and clear classes, while KappaMask L2A, with a more complex classification schema, outperformed S2cloudless by 17%.

Keywords: convolutional neural network; cloud mask; Sentinel-2; KappaMask; active learning; image segmentation; remote sensing

1. Introduction

As part of Europe's Copernicus Earth Observation programme, Sentinel-2 is one of the core missions providing global and continuous multi-spectral observations of all the Earth's land masses. Thanks to the two operating satellites, images over a specific area can be acquired every five days at the equator, whilst at higher latitudes the temporal resolution of the data is even higher.

Clouds and cloud shadows are the main obstacles for frequent land monitoring, significantly reducing the usability of optical satellite data. For information retrieval with automatic processing chains, it is required to separate valid pixels from the ones contaminated by clouds and cloud shadows. Otherwise, extensive manual pre-processing is needed, or errors might propagate to higher level products. Hence, accurate cloud filtering algorithms could boost and improve optical satellite data application for Earth Observation (EO) services.

To tackle the cloud mask problem, various approaches have been developed. The most used are rule-based algorithms, such as Sen2Cor [1], Fmask [2] and MAJA [3]. These are further divided into single-scene and multi-temporal algorithms. MAJA [3] belongs to the multi-temporal-algorithm group that identifies clouded pixels based on temporal time-series of the satellite acquisitions. MAJA performs atmospheric correction and cloud detection for Sentinel-2 images using time series, which can help to avoid the over-classification of clouds by utilizing the correlation of the pixel neighbourhood with previous images. It is very unlikely that there are two different clouds of the same shape at the same location on successive dates. However, MAJA software is complex to set up and is computationally expensive to run.

Sen2Cor [1] is the algorithm currently used by the European Space Agency (ESA) for atmospheric correction and cloud masking on Sentinel-2 images, providing scene classification maps, cloud and snow probabilities at a ground resolution of 20 m. Sen2Cor's cloud detector belongs to the single-scene-algorithm group. Fmask [2] was developed by Zhe Zhu et al. and is another single-scene-algorithm allowing automated masking of clouds, cloud shadows, snow and water from Landsat 4-8 and Sentinel-2 images. Alternatively, S2cloudless [4] is a single-scene cloud detection algorithm which runs single pixel-based classification using machine learning models, such as decision trees, support vector machines (SVM) and neural networks using FastAI API [5]. The best performing method for S2cloudless is with neural networks using FastAI API; nevertheless, due to the slower inference time of neural networks S2cloudless utilizes the tree-based method with LightGBM [6]. However, S2cloudless focuses only on cloud detection, ignoring cloud shadows.

With the rapid development of deep learning, the advantages of using CNN-based networks for image segmentation have been highlighted by many researchers [7]. One data driven approach for solving the cloud segmentation task using a modified U-Net convolutional neural network was proposed by Marc Wieland et al. [8]. The authors trained the network on a global database of Landsat OLI images to segment five classes ("shadow", "cloud", "water", "land" and "snow/ice"). A similar type of network was used by Jeppesen et al. [9], who trained RS-Net on Landsat 8 BIOME [10] and SPARCS datasets [11]. RS-Net is a variation of U-Net which aims to improve the exploitation of spatial information. Landsat dataset was segmented in 2017 by Badrinarayanan et al. [12], who used their own adaptation of SegNet to produce segmentation masks. To reduce the computational load, all images of the Landsat Collection 1 scenes were split into non-overlapping 512 × 512 pixels image blocks with a 30 m spatial resolution. A publicly accessible GaoFen-1 dataset was released by Li et al. [13] with 108 full Landsat 7 and Landsat 8 scenes in different global regions of different land-cover types, including forest, barren, ice, snow, water and urban areas. ResNet with modifications for pixel-level segmentation was used on this dataset with products cropped into 512 × 512 pixel sub-tiles [14]. The developed model is called a multilevel feature fused segmentation network (MFFSNet). It segments input to cloud, cloud shadow and background, and does not differentiate between cloud types. A similar type of network was proposed by Li et al. in [13]. The network, named multi-scale convolutional feature fusion (MSCFF), is meant for remote sensing images of different sensors. It was validated on a dataset including Landsat 5/7/8, Gaofen-1/2/4, Sentinel-2, Ziyuan-3, CBARS-04 and Huanjing images. Regardless of the dataset size, the Northern European terrestrial area relevant for the current development is missing. In comparison to Fmask, the authors achieved an improvement of 3% in the overall accuracy on Landsat 7, and an almost 5% improvement on Landsat 8 dataset.

Training a deep learning model requires a balance between a good representation of the data domain and the ability to converge within the given space. The active learning approach that was introduced to remote sensing in 2009 [15] promotes the idea that a carefully selected subset of data will perform as good as a large dataset of random samples. With the rising interest in deep learning approaches, the active learning methodology has become a powerful tool for creating reference cloud mask datasets. Baetens et al. [16,17] used a random forest in active learning loop to create a reference cloud mask dataset. However, the use of a neural network in an active learning loop has yet to be explored.

In our work, we utilize the active learning methodology in training a CNN-based model that outputs L2A and L1C classification masks at 10 m resolution. Cloud mask output at 10 m is preferred, since it is the native spatial resolution of Sentinel-2 RGB and NIR spectral bands. The finer delineation helps increase the amount of usable data for a large variety of higher-level products generation. Unfortunately, until now, public 10 m Sentinel-2 cloud mask datasets were scarcely available. Recently, the reference dataset that covers Asian region was introduced in [18,19]. Therefore, one objective of this study is the creation of a novel active learning-based hand-annotated dataset "Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks" that covers Northern European terrestrial area at a 10 m resolution [20].

The other objectives of the current study are to further analyse the capabilities of CNN-based models for the creation of cloud and cloud shadow mask for Sentinel-2 in Northern European terrestrial area and to meet the need for a more accurate, categorised and finer resolution mask that has not been published before. The paper describes an active learning methodology with a deep learning network in the loop to create a dataset and train a CNN-based model. The novelty of the training procedure comprises of pre-training on the 20 m resolution reference dataset distributed globally and fine-tuning on the 10 m resolution dataset with a focus on the Northern European area. The approach leverages advantages of the 20 m resolution dataset and increases its robustness to different landscapes. The evaluation on the test dataset is shown, including comparison with rule-based methods—Sen2Cor, Fmask, MAJA—and AI-based methods—S2cloudless, DL-L8S2-UV (Deep Learning for Cloud Detection in Landsat-8 and Sentinel-2 Images [21]). The feature importance in the model for both L1C and L2A input is analysed. Finally, a brief timing comparison over KappaMask on GPU and CPU, Fmask, S2cloudless and DL-L8S2-UV methods is presented.

2. Materials and Methods

The open data policy of the Copernicus program makes Sentinel-2 one of the most powerful resources for EO applications. Data can be accessed through several distribution channels, such as CREODIAS [22] among the five DIASes of the DIAS Hub [23], the Copernicus Open Access Hub [24] and the Collaborative National Mirror Sites, e.g., [25,26]. Sentinel-2 data are provided at different processing levels, including Level-1C (L1C) and Level-2A (L2A). L1C is a cartographic UTM projection of the Top of Atmosphere (TOA) reflectances orthorectified using a digital elevation model (DEM) to correct geometric distortion. The L2A product provides Bottom of Atmosphere (BOA) reflectance images derived from the associated L1C products. Both L1C and L2A Sentinel-2 products are distributed in ~100 by ~100 km^2 ortho-image tiles resulting in up to 10,980 by 10,980 pixel images at 10, 20 and 60 m spatial resolution depending on the spectral band.

The pipeline for obtaining the training dataset can be split into several blocks (Figure 1). First, the "Sentinel-2 Cloud Mask Catalogue" dataset by Francis et al. [27] is used for initial model training. Next, we prepared our own Sentinel-2 (S-2) labelled dataset "Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks" [28] using Computer Vision Annotation Tool (CVAT) [29] and Segments.ai [30] labelling tools, which will be described in Section 2.2. On our labelled data (Figure 2), the model is fine-tuned (the process is described in more detail in Section 2.3). After this step, the prediction is performed on the new Sentinel-2 products, followed by a selection of the sub-tiles with the lowest prediction accuracy for further labelling.

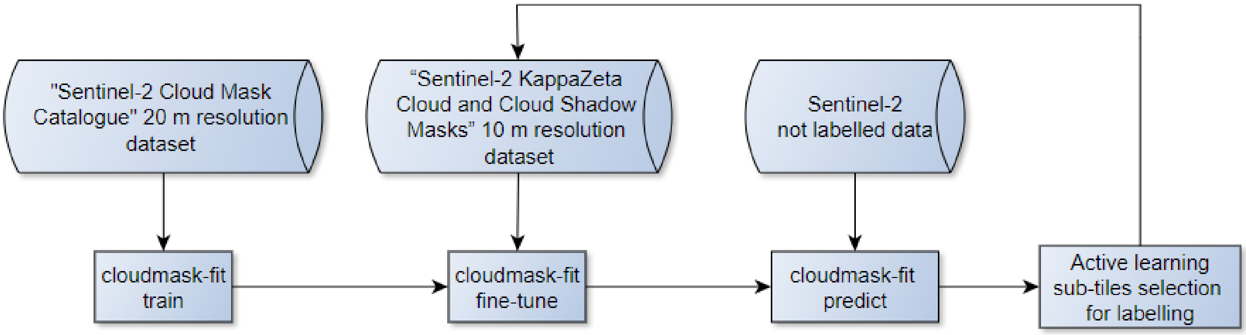

"Sentinel-2 Cloud Mask Catalogue" 20 m resolution dataset
"Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks" 10 m resolution dataset
Sentinel-2 not labelled data
cloudmask-fit train
cloudmask-fit fine-tune
cloudmask-fit predict
Active learning sub-tiles selection for labelling

Figure 1. General pipeline of the KappaMask model development.

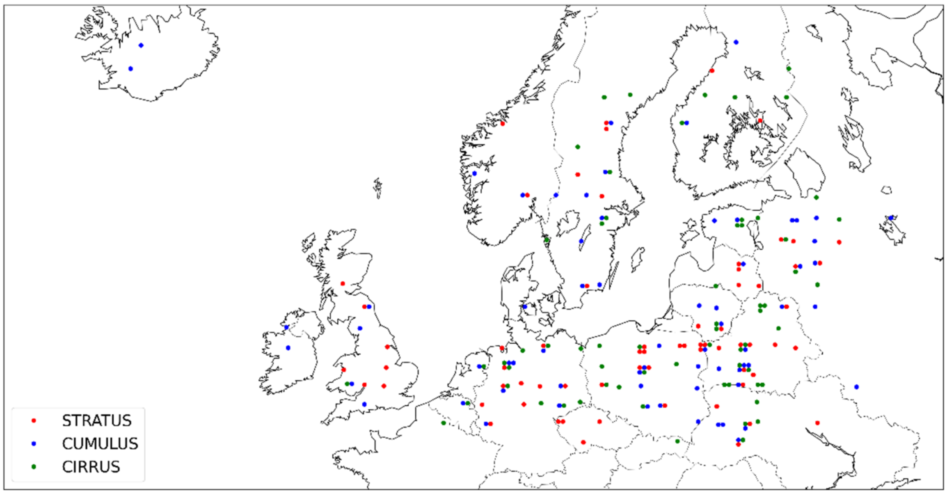


Figure 2. Sentinel-2 tiles used for labelling. Images with cumulus clouds are indicated as blue dots, images with stratus clouds are marked as red and images with cirrus clouds are marked as green. Each dot corresponds to one Sentinel-2 100 × 100 km data product. If the dots are exactly next to each other, it means the corresponding Sentinel-2 products are from the same location.

The final model is validated on the isolated test dataset (Figure 3) and the performance is compared to the existing masks, such as Sen2Cor, MAJA, Fmask and S2cloudless.

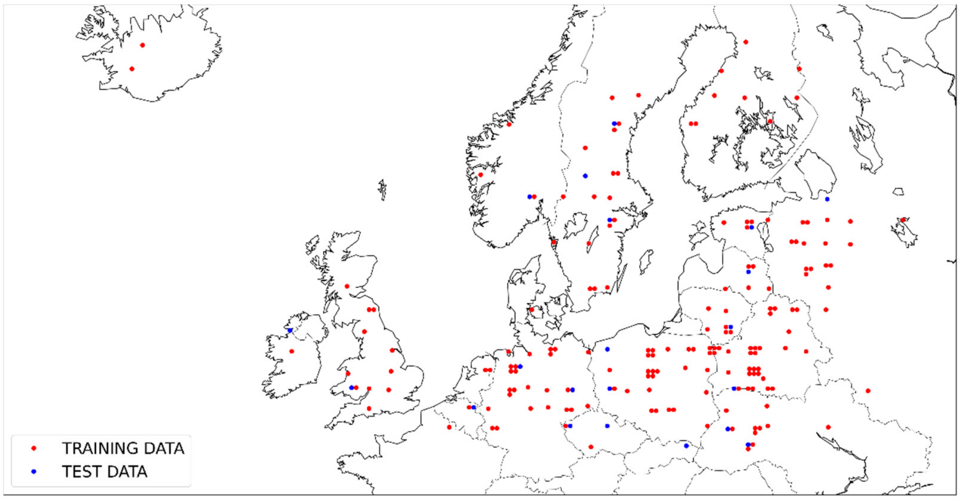


Figure 3. Distribution of the training and test data. Red dots indicate Sentinel-2 images reserved for training and validation dataset, and blue dots indicate images used for the test dataset only. If the dots are exactly next to each other, it means the corresponding Sentinel-2 products are from the same location.

2.1. Data Processing

KappaMask can generate the classification mask from both, L2A and L1C products input. Thus, the processing pipeline for each product level utilizes different input bands which, in the context of machine learning, are referred to as features.

The algorithm for Level-2A product is using features at spatial resolutions of 10 m, 20 m and 60 m. The following features are used for training:

- 10 m bands: "B02" (490 nm), "B03" (560 nm), "B04" (665 nm), "B08" (842 nm);
- 20 m resolution bands: "B05" (705 nm), "B06" (740 nm), "B07" (783 nm), "B8A" (865 nm), "B11" (1610 nm), "B12" (2190 nm);
- 60 m resolution bands: "B01" (443 nm), "B09" (940 nm);
- "AOT" (Aerosol Optical Thickness) and "WVP" (Water vapor map).

Level-1C product includes all listed above bands, but AOT and WVP are not available. Additionally, L1C includes 60 m Band 10 (1375 nm).

Each of the S-2 product is divided into 484 non-overlapping sub-tiles of 512 × 512 pixels before uploading the tiles to CVAT or Segments.AI for labelling. Since the channels of S-2 products have dimensions of either 10,980 × 10,980 pixels (10 m spatial resolution), 5490 × 5490 pixels (20 m spatial resolution) or 1830 × 1830 pixels (60 m spatial resolution), the 512 × 512 pixel sub-tiles on the right and bottom edges are only partially filled with data, hence zero padding is applied when needed.

The features are resampled to the same 10 m resolution with Sinc Infinite Impulse Response (IIR) filter that is windowed with a Blackman filter [7].

For training, both the open-source dataset "Sentinel-2 Cloud Mask Catalogue" by Francis et al. [27] and the "Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks" dataset [28] that we created ourselves, were processed. The "Sentinel-2 Cloud Mask Catalogue" contains cloud masks for 513 subscenes of 1022 × 1022 pixels at 20 m resolution, from the 2018 L1C Sentinel-2 archive. To use this dataset for pretraining the KappaMask model, the output mask is resampled to 10 m resolution and cropped to 512 × 512 pixel sub-tiles. The corresponding L2A products were processed for the same size and resolution and were used for training.

The "Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks" dataset consists of 150 Sentinel-2 products with 10 sub-tiles of 512 × 512 pixels labelled at 10 m resolution. The dataset contains roughly equal amounts of stratus, cumulus and cirrus clouds. Different cloud types were chosen in order to achieve maximal representation. For each month, ten products with stratus clouds, ten products with cumulus clouds and ten products with cirrus clouds were chosen from different locations. The labelled products are distributed uniformly over Northern Europe (Figure 2). Sentinel-2 products form a fixed grid which covers the globe; therefore, many dots on Figure 2 share similar latitude and/or longitude. The dots that are bundled together represent products from the same grid point. However, the products from the same location usually have different cloud types and they are from different months.

From each Sentinel-2 product represented with a red, blue or green dot, which correspond respectively to stratus, cumulus and cirrus cloud types, ten sub-tiles of 512 × 512 pixels were labelled. The selection of sub-tiles was done using active learning approach by selecting worst performing sub-tiles after model inference. All 484 sub-tiles were predicted by the model, and the prediction masks were laid half-transparently over the original images. Based on the visual evaluation of the resulting images, a human labeller chose the sub-tiles where the prediction mask was most erroneous. This way, only the highest impact sub-tiles were chosen for subsequent labelling. From each month, one product with stratus clouds, one product with cumulus clouds and one product with cirrus clouds were reserved for the test set, and the rest of the products were used for training. Additionally, six S-2 products covering Estonia were fully labelled, meaning that all 484 sub-tiles of 512 × 512 pixels were labelled with an additional 2821 sub-tiles. However, those fully labelled product sub-tiles are present only in the training dataset.

2.2. Output Classification Classes

The labelling was done using the Computer Vision Annotation Tool (CVAT) [29] and Segments.ai [30]. CVAT is a free, open-source tool that can be used to create segmentation mask annotations. Using CVAT, six S-2 images over Estonia were labelled, resulting in 6 × 484 sub-tiles of 512 × 512 pixels. The rest of the products was labelled using the Segments.ai platform that supports model-assisted labelling. Thanks to the possibility of integrating active learning process into Segments.ai by uploading new images with predictions, the labelling speed improved significantly. However, after the model predictions, all sub-tiles were still manually corrected. Therefore, a further 150 Sentinel-2 products were labelled using Segments.ai and resulted in ~1500 labelled sub-tiles of 512 × 512 pixel size. The labelling from both tools were peer-reviewed by labellers. CVAT and Segments.ai have significant differences in usage and resulting outputs. While Segments.ai has ability to draw with brush, CVAT uses polygon shapes. As a result, Segments.ai labels are smoother and less angular. The "Sentinel-2 Cloud Mask Catalogue" by Francis et al. [27] that was used for pre-training was labelled semi-automatically using the IRIS toolkit [31]. We believe that a combination of different tools is another advantage for training, which reduces a bias towards one specific way of labelling.

Labels in both CVAT and Segments.ai are split into the following categories:

- 0—MISSING: missing or invalid pixels;
- 1—CLEAR: pixels without clouds or cloud shadows;
- 2—CLOUD SHADOW: pixels with cloud shadows;
- 3—SEMI TRANSPARENT CLOUD: pixels with thin clouds through which the land is visible;
- 4—CLOUD: pixels with cloud;
- 5—UNDEFINED: pixels that the labeller is not sure which class they belong to.

Undefined class is used where the labeller is either not certain what is in the image or when the class borders are not clear. The UNDEFINED class is excluded from the model training and the predictor assigns the class with the highest confidence from the other five classes.

Table 1 illustrates the mapping of output classes from different classification masks for validation and performance comparison. The table includes the Cloud Masking Inter-Comparison Exercise (CMIX) standard notation [32], Sen2Cor, Fmask, MAJA and S2cloudless label comparison.

Table 1. Output correspondence for different cloud classification masks. KappaMask output scheme is used for model fitting with logic corresponded to CMIX notation. Sen2Cor, Fmask, MAJA and S2cloudless are mapped respectively to KappaMask output. FMSC to Francis, Mrziglod and Sidiropoulos' classification map [27] is used for pretraining.

Sen2Cor	CMIX	KappaMask	Fmask	S2Cloudless	FMSC
0 No data		0 Missing			
1 Saturated or defective		0 Missing			
2 Dark area pixels		1 Clear			
3 Cloud shadows	4 Cloud shadows	2 Cloud shadows	2 Cloud shadows		2 Cloud shadows
4 Vegetation	1 Clear	1 Clear	0 Clear	0 Clear	0 Clear
5 Not vegetated		1 Clear			
6 Water		1 Clear	1 Water		
7 Unclassified		5 Undefined			
8 Cloud medium probability		4 Cloud			
9 Cloud high probability	2 Cloud	4 Cloud	4 Cloud	1 Cloud	1 Cloud
10 Thin cirrus	3 Semi-transparent cloud	3 Semi-transparent cloud			
11 Snow		1 Clear	3 Snow		

2.3. Model Fitting

Each input Sentinel-2 L2A image has 13 features: "B01", "B02", "B03", "B04", "B05", "B06", "B07", "B08", "B8A", "B09", "B11", "B12", "AOT", "WVP". Meanwhile, the Sentinel-2 L1C model has 12 input features as an input: "B01", "B02", "B03", "B04", "B05", "B06", "B07", "B08", "B8A", "B09", "B10", "B11", "B12". The input features are normalised with min-max normalization and passed to the model as an input. The output of the network is a classification map that is identifying each pixel as clear, cloud shadow, semi-transparent cloud, cloud, or missing class.

The full dataset is divided into training, validation and test sets. The training and validation sets are split randomly, according to the ratio of 80/20%. The test dataset consists of separate Sentinel-2 products which were preselected with a diverse geographical and temporal distribution in mind. For each month from April to August, three Sentinel-2 products with different types of clouds were selected. Moreover, from each product, the 10 most complicated sub-tiles were selected by data labellers. As the most complicated sub-tiles we considered the tiles that covered all different classes and for which the model gave the most errors during prediction with manual observation. The geographical distribution for the test dataset is shown in blue, alongside the training data in red (which includes both training and validation set) in Figure 3.

In contrast to standard pixel classifications where a limited pixel neighbourhood is exploited, we aim to segment one sub-tile as a whole, giving both spatial context from the whole image and spectral information from all channels.

The input of the model has 512 × 512 × 14 dimensionality corresponding to the rows, columns and channels of the image, respectively. The output segmentation map consists of the confidence values for five classes and the values transformed with argmax for the final segmentation mask. U-Net architecture [33] has two sides: the encoding part which down-samples the image to generalize the features, and the decoding part which up-samples the image back to the original size. Between encoder and decoder, there are skip connections, which copy information between different levels of features. The architecture of the network is shown in Figure 4.

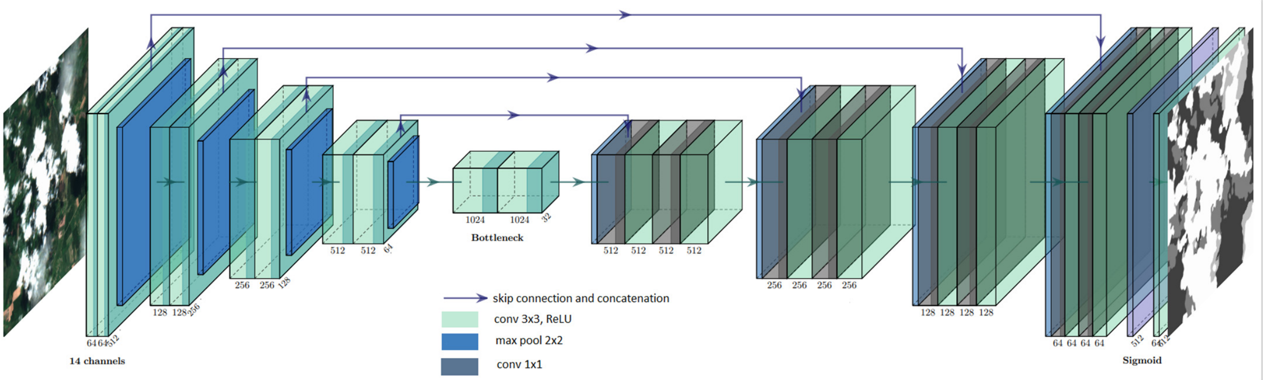


Figure 4. U-Net model architecture used for training.

The Adam optimizer [34] is used for fitting the model. During training, if the monitored validation metric does not improve for a specified number of epochs, the learning rate is reduced. To avoid overfitting, the training is stopped when the validation loss does not improve for a specified number of epochs.

In our initial experiments, we used categorical cross entropy loss, which is commonly used for CNN-based cloud mask methods [11–13,19,20]. However, we found that lower validation loss does not necessarily mean better segmentation results and the F1 score still increased when the model was already overfitted. Due to this, we are using dice coefficient loss [35] which shows better results for the current setup.

2.4. Inference

Once the network has been trained (the optimal values for model weights have been determined), inference is performed by forward passing an image through the network. The Sentinel-2 L1C or L2A product folder is an input to the inference tool [20]. The product is cropped into 512 × 512 pixel sub-tiles with overlapping (the overlapping size that we used in our experiments is 32 pixels) and the model generates the prediction for each individual crop. In the next step the whole product output classification mask is combined from individual crops. The process of sub-tiling and final mask generation is illustrated in Figure 5.

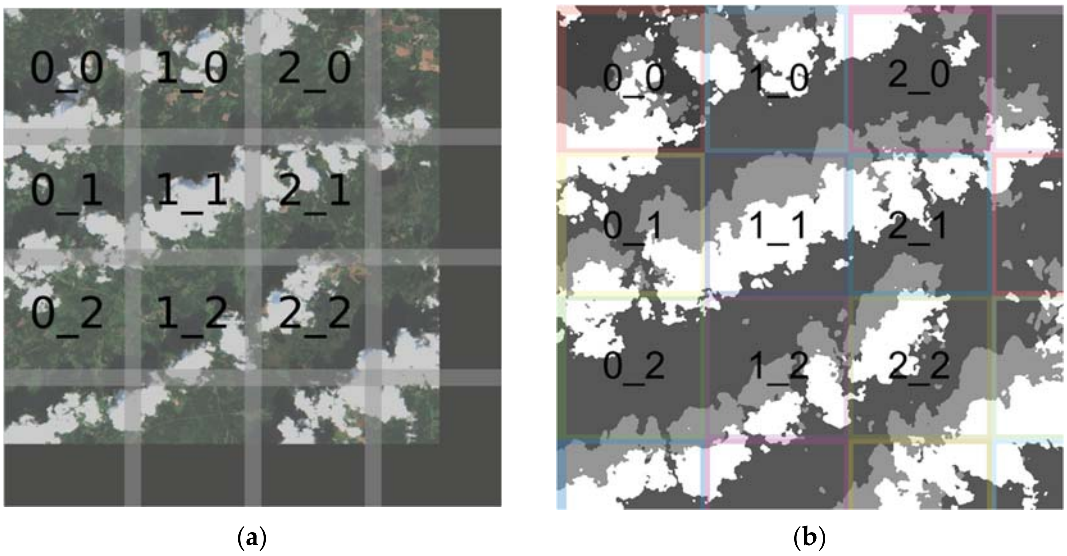


Figure 5. Illustrative images of how full Sentinel-2 image is cropped for prediction and how classification map is rebuilt into the full image with 32 pixels overlap. (**a**) Cropping original Sentinel-2 image for inference. (**b**) Prediction output mask combined into the final image.

3. Results

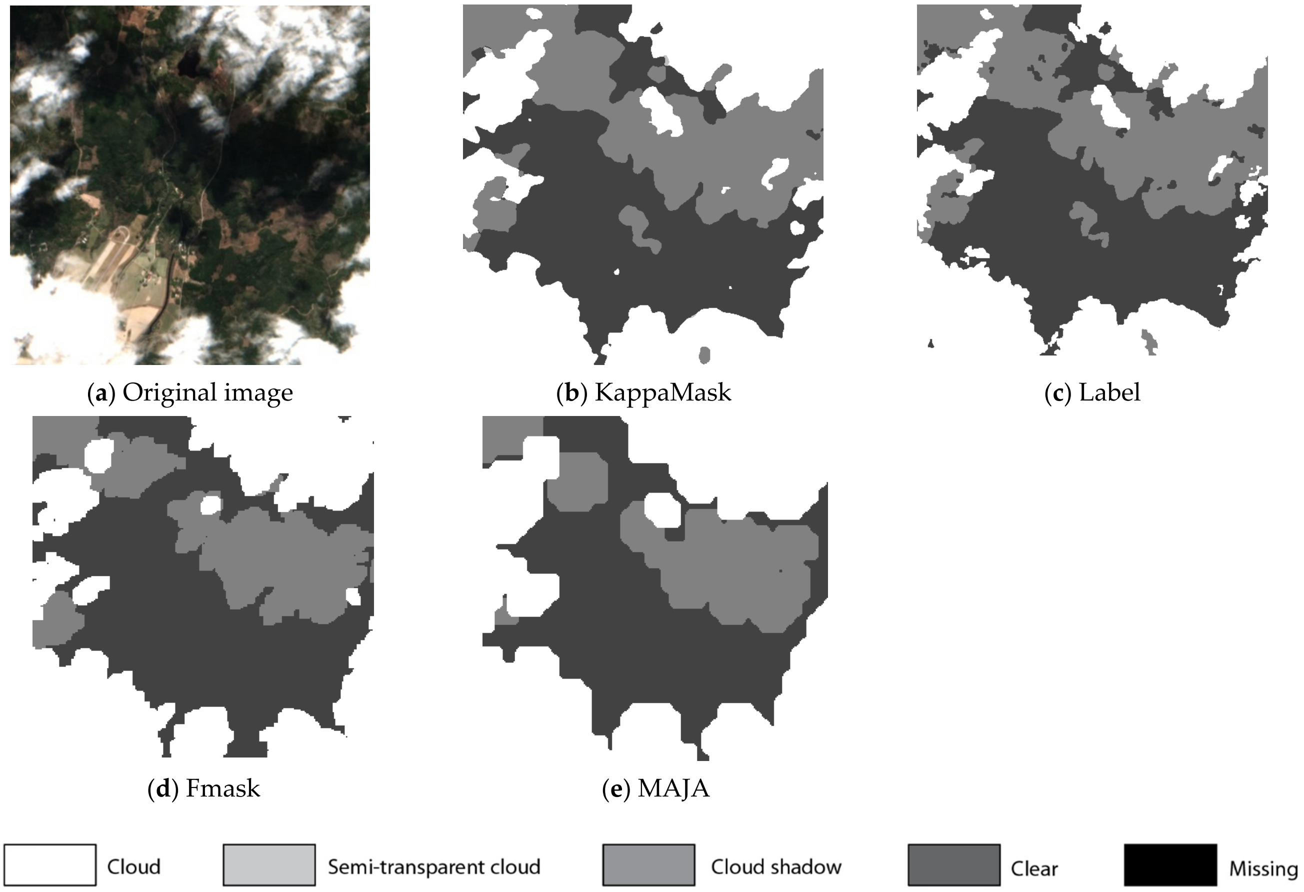
Once U-Net model was trained over the "Sentinel-2 Cloud Mask Catalogue" Francis [27] dataset, the network was finetuned on Northern European terrestrial dataset that we labelled using the active learning methodology. As mentioned in Section 2.1, the test dataset was comprehensively selected to showcase different kinds of errors over Northern Europe and consisted of 15 Sentinel-2 products. The test products were predicted by the pre-trained model, and the least performing sub-tiles were selected, based on a visual comparison between prediction mask and original image. However, these test images were not used in training or validation at any stage.

The results section is structured as follows. The first subsection provides KappaMask L1C and L2A comparison with rule-based methods—Sen2Cor, Fmask and MAJA. Afterwards, KappaMask performance is compared to machine learning methods—S2cloudless that uses tree algorithm with LightGBM and deep learning DL_L8S2_UV [21] network. The next sub-section shows the feature importance analysis of KappaMask L1C and L2A model, followed by the hyperparameter tuning of network depth and number of filters. The last sub-section presents the comparison of the time spent by KappaMask on CPU and GPU, Fmask, S2cloudless and DL_L8S2_UV networks.

3.1. KappaMask L1C and L2A Comparison with Rule-Based Methods

Figure 6 presents individual sub-tile output for L2A model in comparison to label, Sen2Cor, Fmask and MAJA classification mask and Figure 7 showcases individual sub-tile prediction for L1C model. Overall, KappaMask is performing better on cloud and cloud shadow on both figures, while Sen2Cor is under-segmenting cloud and Fmask and MAJA are under-segmenting cloud shadows. As a multi-temporal algorithm, MAJA relies on the quality of the time-series, where the "backward mode" is used to obtain better quality results for the first product of a time-series. For this, at least one product in the time-series tiles should be considered as valid in the period of 45 days, which means that it has less than 90% of cloudy area. Otherwise, the result of the processing will be empty. These constrains may affect MAJA output results, since North European weather conditions for some locations might not have cloud-free images for the required length of the time series as it happened for the L1C example. The performance comparison of KappaMask, Sen2Cor and Fmask can be checked at [36]. Examples of full S-2 product classification maps can be seen in Appendix A (Figures A1–A3).

Table 2 shows the comparison of dice coefficients for KappaMask L2A, KappaMask L1C, Sen2Cor, Fmask and MAJA on the test dataset. The results were obtained from the same challenging test set for all methods. The greatest improvement in the dice coefficient of KappaMask L2A in comparison to other cloud mask methods is cloud shadow detection for which L2A is 20% more accurate than the closest competitor, Sen2Cor. The performance of cloud classification is also superior, with a dice coefficient of 86% against the closest result (62% for Sen2Cor). KappaMask L2A has the highest dice coefficient on the semi-transparent class, which is 29% higher than Sen2Cor. Finally, clear class KappaMask 82% is 7% higher than the closest result of Fmask. The total dice coefficient average over all classes for both KappaMask L2A and KappaMask L1C is bigger than Sen2Cor, Fmask, and MAJA with 80% and 76% against 59%, 61%, and 51%, respectively. Overall, the performance of KappaMask L2A is 4% better than performance of the KappaMask L1C which can be explained by the focus of the training. The choice of the architecture in Table 2 for the L2A model was influenced by the initial priority of applicability for agricultural use and therefore needed the availability of the atmospheric correction. Regarding this, we believe that the performance of KappaMask L1C can be further improved in the future. Precision, recall and accuracy in Tables 3–5, respectively, are presented to provide a deeper understanding of each model's performance.


(**a**) Original image

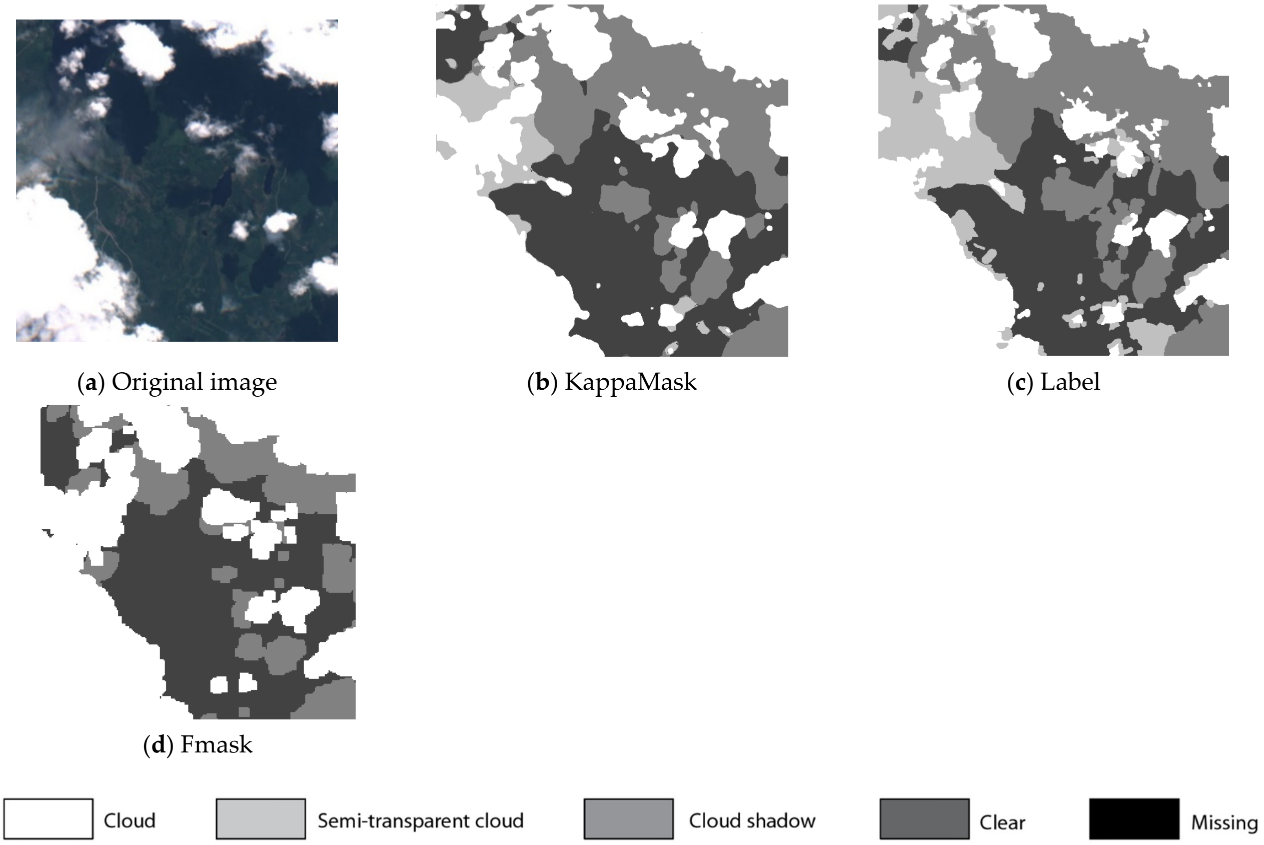
(**b**) KappaMask

(**c**) Label

(**d**) Fmask

(**e**) MAJA

Figure 6. Comparison of L2A prediction output for a 512 × 512 pixels sub-tile in the test dataset. (**a**) Original Sentinel-2 L2A True-Color Image; (**b**) KappaMask classification map; (**c**) Segmentation mask prepared by a human labeller; (**d**) Fmask classification map; (**e**) MAJA classification map.



(**a**) Original image (**b**) KappaMask (**c**) Label

(**d**) Fmask

Figure 7. Comparison of L1C prediction output for a 512 × 512 pixels sub-tile in the test dataset. (**a**) Original Sentinel-2 L1C True-Colour Image; (**b**) KappaMask classification map; (**c**) Segmentation mask prepared by a human labeller; (**d**) Fmask classification map.

Table 2. Dice coefficient evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, Sen2Cor, Fmask and MAJA cloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Dice Coefficient	**KappaMask L2A**	**KappaMask L1C**	**Sen2Cor**	**Fmask**	**MAJA**
Clear	**82%**	75%	72%	75%	56%
Cloud shadow	**72%**	69%	52%	49%	-
Semi-transparent	**78%**	75%	49%	-	-
Cloud	**86%**	84%	62%	60%	46%
All classes	**80%**	76%	59%	61%	51%

Table 3. Precision evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, Sen2Cor, Fmask and MAJA cloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Precision	**KappaMask L2A**	**KappaMask L1C**	**Sen2Cor**	**Fmask**	**MAJA**
Clear	75%	**79%**	60%	66%	64%
Cloud shadow	82%	79%	**87%**	51%	-
Semi-transparent	**83%**	71%	78%	-	-
Cloud	**85%**	83%	57%	44%	35%
All classes	**81%**	78%	71%	54%	50%

Table 4. Recall evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, Sen2Cor, Fmask and MAJA cloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Recall	KappaMask L2A	KappaMask L1C	Sen2Cor	Fmask	MAJA
Clear	**91%**	71%	90%	86%	50%
Cloud shadow	**64%**	61%	37%	48%	-
Semi-transparent	74%	**80%**	36%	-	-
Cloud	**87%**	85%	67%	60%	65%
All classes	**79%**	74%	58%	65%	58%

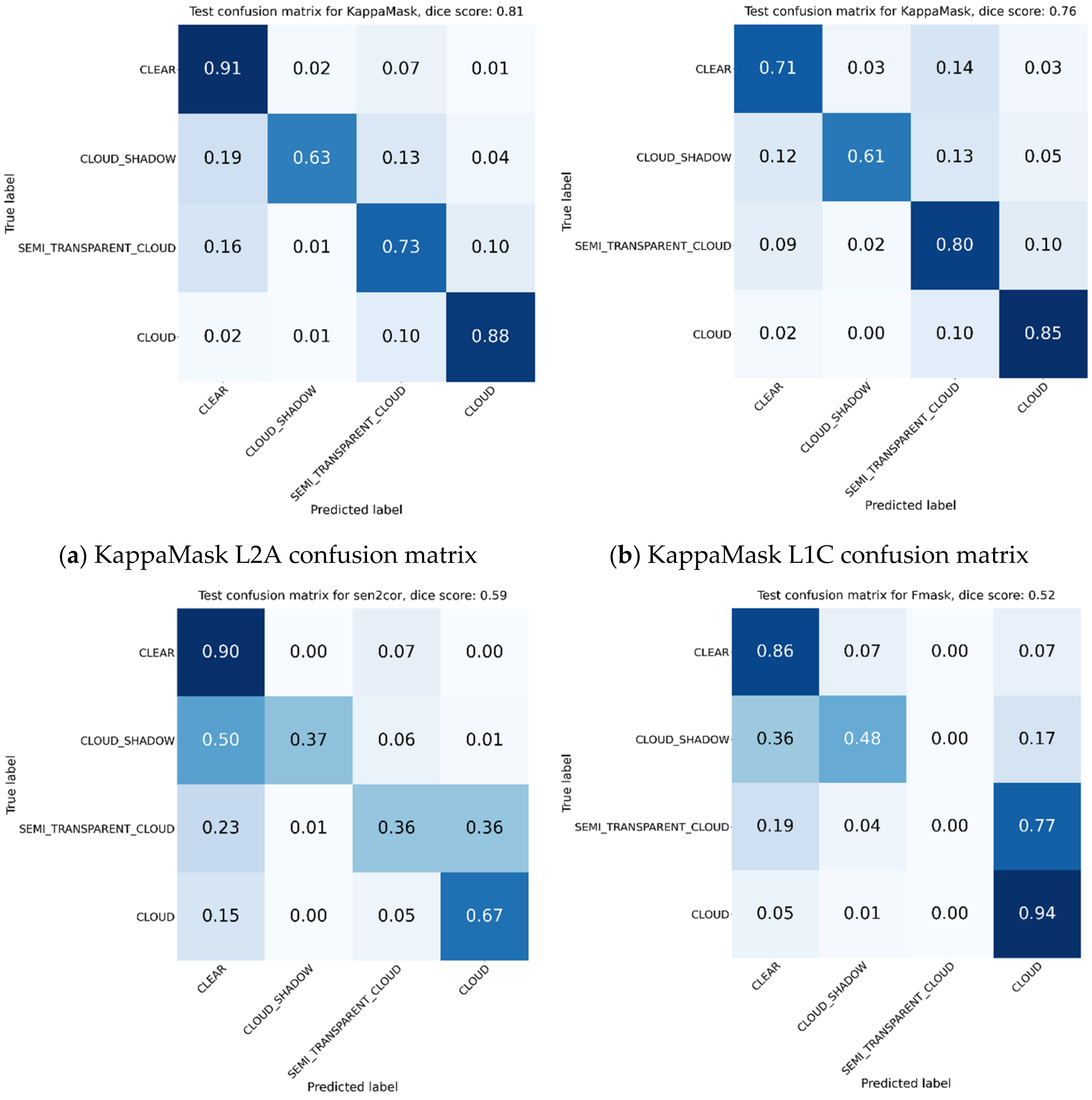
Table 5. Overall accuracy evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, Sen2Cor, Fmask and MAJA cloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Overall Accuracy	KappaMask L2A	KappaMask L1C	Sen2Cor	Fmask	MAJA
Clear	**89%**	86%	81%	84%	79%
Cloud shadow	**96%**	95%	95%	92%	-
Semi-transparent	**85%**	79%	72%	-	-
Cloud	**92%**	91%	78%	67%	63%
All classes	**91%**	88%	82%	81%	71%

Figure 8 shows the confusion matrices for KappaMask L2A, KappaMask L1C, Sen2Cor and Fmask. The highest recall for the clear class is 91% for KappaMask L2A (Figure 8a). Moreover, the cloud shadow has the highest recall of 63% for KappaMask L2A. The semi-transparent cloud class recall is the highest compared to Sen2Cor and Fmask (80% for KappaMask L1C, Figure 8b), being over two times more accurate than Sen2Cor semi-transparent class with 36% in Figure 8c. Fmask (Figure 8d) has the highest recall for the cloud class with 94%; however, 5% of the clear area in Fmask is falsely predicted as cloud, while for KappaMask it is only 2% for L1C and L2A. Fmask does not have a semi-transparent class; thus, if we respectively add confusion with the semi-transparent class for KappaMask, it will be 98% for L2A and 95% for L1C models, outperforming Fmask.

False negatives are crucial in agricultural applications, since the area treated as clear would propagate to the higher-level products, such as NDVI. The result of false negatives for the clear class with KappaMask L1C is 23%, while false negatives for the clear class in Fmask and Sen2Cor are 60% and 88%, respectively. The pixels that are falsely predicted as clear cause the biggest disturbances when used for applications. We find that 16% out of 37% false negative errors in KappaMask L2A output appear with semi-transparent clouds. However, semi-transparent clouds are often very thin and their borders often not visible with human eye, making them difficult to delineate. The more accurate classification for KappaMask L1C for semi-transparent clouds is explained by the availability of cirrus B10 band in Level-1C product which is missing for Level-2A input.

At the end of the classification run on a Sentinel-2 product, the overlapping cloud mask sub-tiles are mosaicked together into a georeferenced .tiff file. Figures A1–A3 show examples of the final output for one of the test products.



(**a**) KappaMask L2A confusion matrix

(**b**) KappaMask L1C confusion matrix

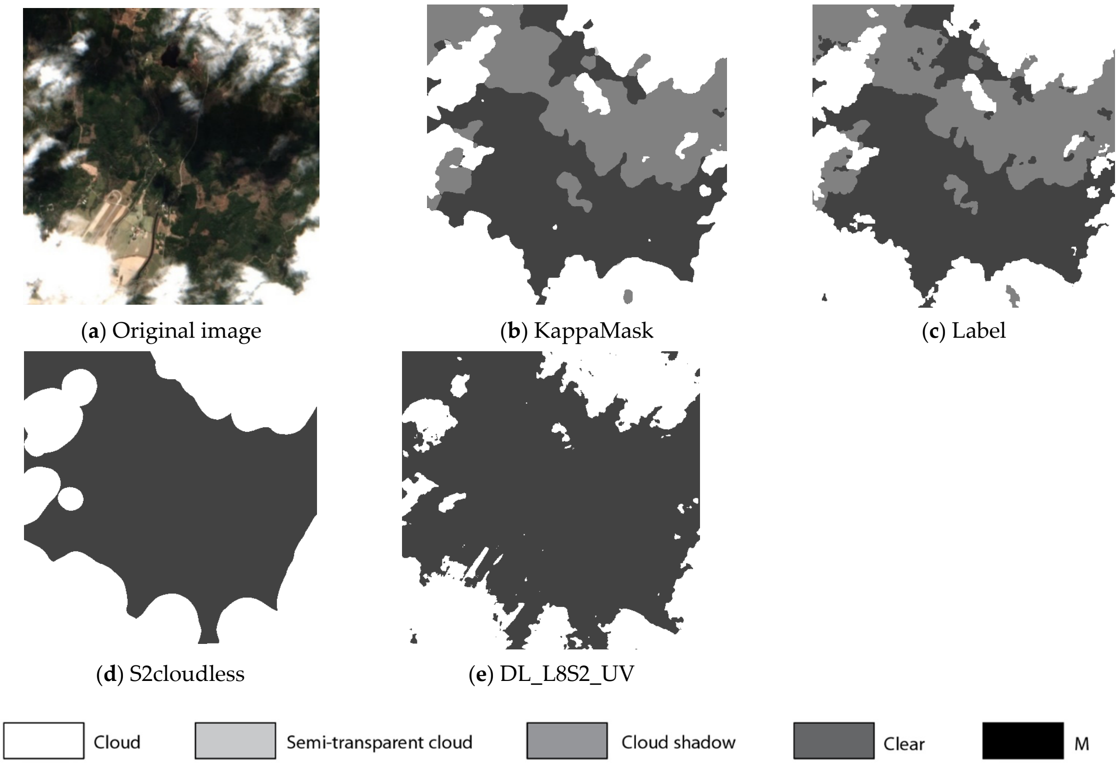
(**c**) Sen2Cor confusion matrix

(**d**) Fmask confusion matrix

Figure 8. Confusion matrices on test set for (**a**) KappaMask Level-2A; (**b**) KappaMask Level-1C; (**c**) Sen2Cor; (**d**) Fmask. Confusion matrix consists of clear, cloud shadow, semi-transparent cloud, cloud and missing class; however, the last one is removed from this comparison to make matrices easier to read.

3.2. KappaMask L1C and L2A Comparison with AI-Based Methods

A comparison between KappaMask L1C, KappaMask L2A, machine learning S2cloudless trained using tree algorithms with LightGBM and deep learning model DL_L8S2_UV is performed. In order to generate the S2cloudless mask at 10 m resolution, we contacted researchers to provide coefficients for 10 m run. DL_L8S2_UV outputs 10 m resolution mask by design. Figures 9 and 10 presents the comparison of one sub-tile prediction with S2cloudless and DL_L8S2_UV, respectively. A visual inspection shows that S2cloudless can miss smaller clouds or make their shapes inaccurate, while DL_L8S2_UV has more accurate shapes, but under-segments parts of clouds. Meanwhile, KappaMask segments cloud shape accurately, does not miss smaller clouds and tends to over-segment clouds.



(**a**) Original image

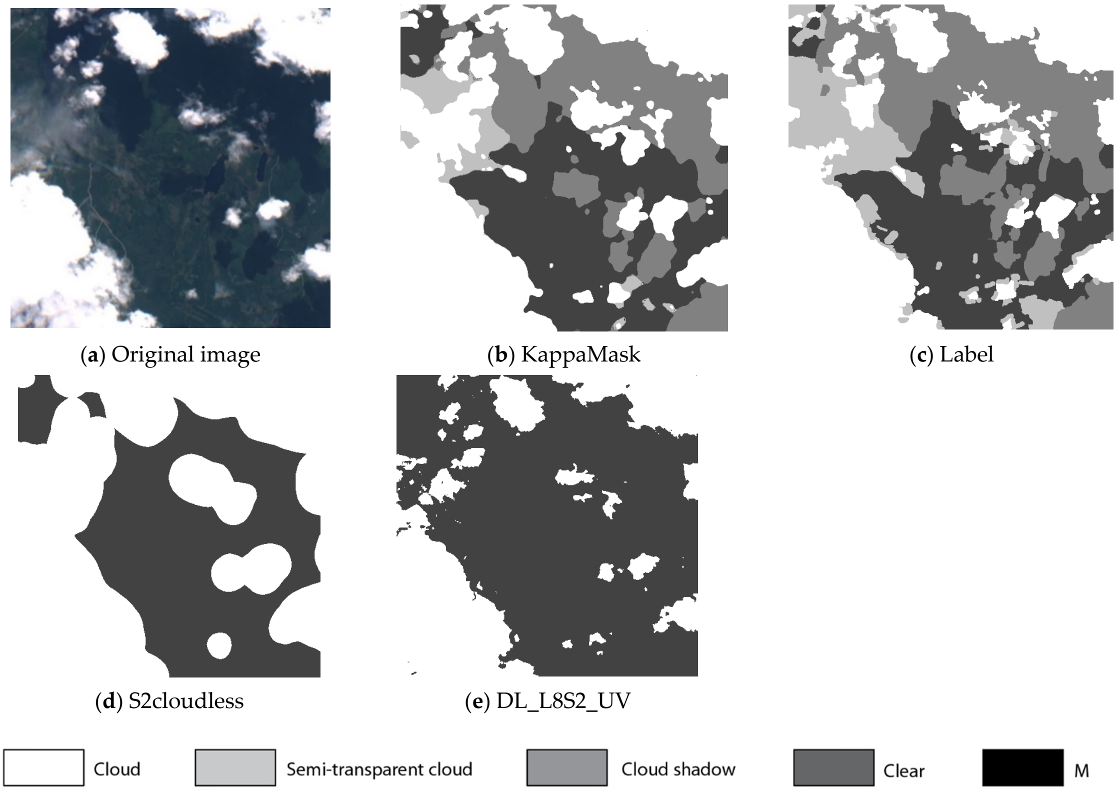
(**b**) KappaMask

(**c**) Label

(**d**) S2cloudless

(**e**) DL_L8S2_UV

Figure 9. Comparison of L2A prediction output for a 512 × 512 pixels sub-tile in the test dataset. (**a**) Original Sentinel-2 L2A True-Colour Image; (**b**) KappaMask classification map; (**c**) Segmentation mask prepared by a human labeller; (**d**) S2cloudless classification map; (**e**) DL_L8S2_UV classification map.



(**a**) Original image

(**b**) KappaMask

(**c**) Label

(**d**) S2cloudless

(**e**) DL_L8S2_UV

Figure 10. Comparison of L1C prediction output for a 512 × 512 pixels sub-tile in the test dataset. (**a**) Original Sentinel-2 L1C True-Colour Image; (**b**) KappaMask classification map; (**c**) Segmentation mask prepared by a human labeller; (**d**) S2cloudless classification map; (**e**) DL_L8S2_UV classification map.

Since both comparison methods output only clear versus clouds classification map, the metric comparison in Tables 6–9 is also presented for these two classes for KappaMask. Overall, KappaMask L2A outperforms both methods for two classes for all presented metrics with 21% difference for both classes average in dice coefficient with S2cloudless and 22% difference with DL_L8S2_UV. Both S2cloudless and DL_L8S2_UV have a recall close to KappaMask, but significantly lower precision and accuracy, confirming the observation from images examples that KappaMask detects more smaller clouds than the other methods.

Table 6. Dice coefficient evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, S2cloudless and DL_L8S2_UV cloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Dice Coefficient	KappaMask L2A	KappaMask L1C	S2cloudless	DL_L8S2_UV
Clear	**82%**	75%	69%	56%
Cloud	**86%**	84%	57%	67%
All classes	**84%**	80%	63%	62%

Table 7. Dice coefficient evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, S2cloudless and DL_L8S2_UVcloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Precision	KappaMask L2A	KappaMask L1C	S2cloudless	DL_L8S2_UV
Clear	75%	**79%**	59%	41%
Cloud	**85%**	83%	41%	59%
All classes	**81%**	76%	50%	50%

Table 8. Dice coefficient evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, S2cloudless and DL_L8S2_UVcloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Recall	KappaMask L2A	KappaMask L1C	S2cloudless	DL_L8S2_UV
Clear	**91%**	71%	84%	90%
Cloud	87%	85%	**93%**	77%
All classes	**89%**	78%	**89%**	84%

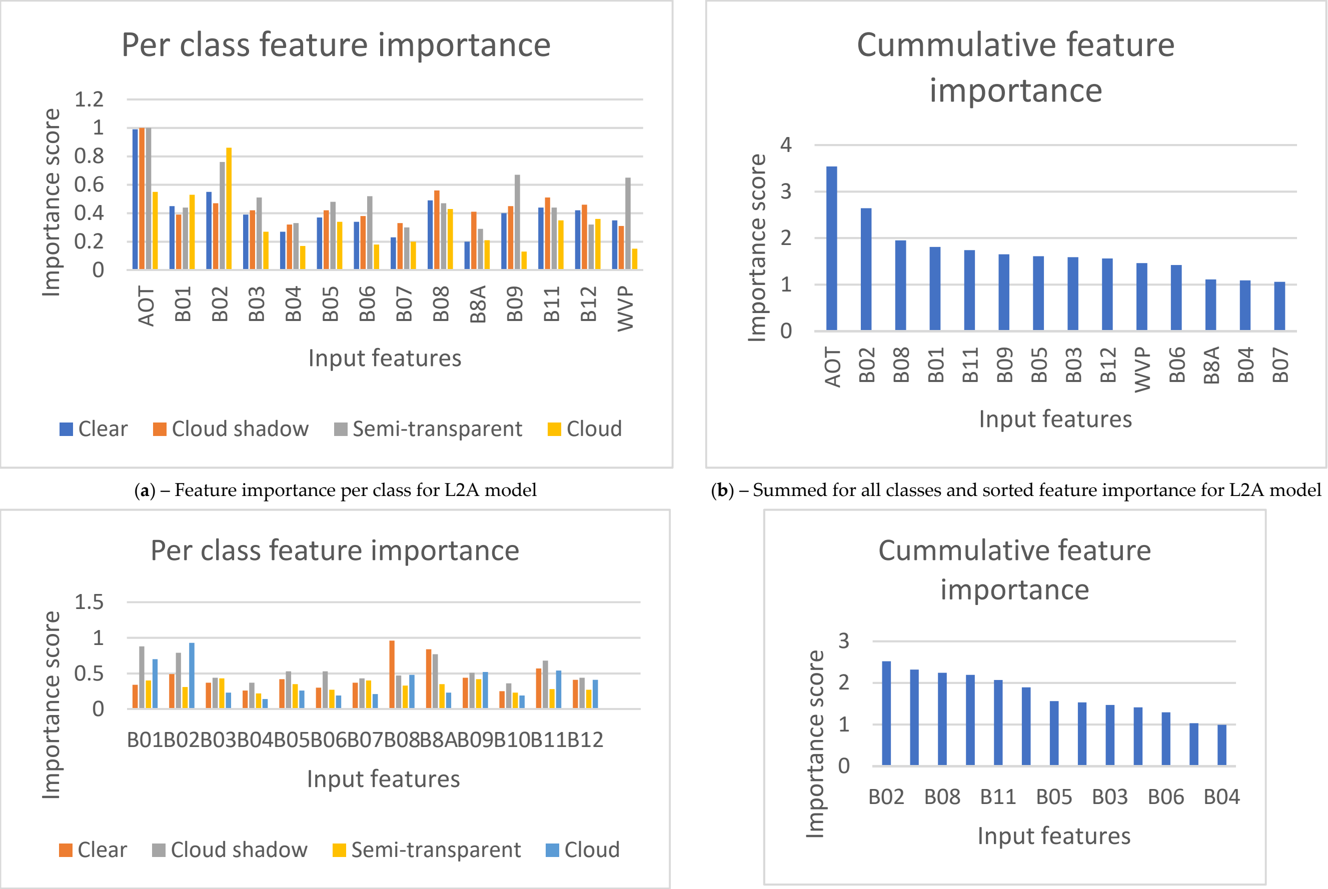
Table 9. Dice coefficient evaluation performed on the test dataset for KappaMask Level-2A, KappaMask Level-1C, S2cloudless and DL_L8S2_UVcloud classification maps. Evaluation is performed for clear, cloud shadow, semi-transparent and cloud classes.

Overall accuracy	KappaMask L2A	KappaMask L1C	S2cloudless	DL_L8S2_UV
Clear	**89%**	86%	80%	61%
Cloud	**92%**	91%	62%	79%
All classes	**91%**	89%	71%	73%

3.3. KappaMask Feature Importance Analysis

The KappaMask L2A model uses 14 features including 12 bands and AOT and WVP indexes as the input. Figure 11a illustrates the importance of each feature for each class during the inference. The importance score is obtained by filling the selected feature with

0 values and calculating the error as an inverse dice coefficient for each class, so 0 means the highest performance and 1 means the worst dice coefficient. This method is called permutation feature importance, suggested by Fisher et al. in 2019 [37]. The importance score per class is summed up in Figure 11b. Except for Aerosol Optical Thickness (AOT) and B02, the rest of the features contribute to model output almost equally. The network was retrained discarding the least important B07, the total dice coefficient metric decreased by 0.5%. Therefore, to reduce the network size, it is possible to discard some of the input features; however, there is a trade-off between accuracy and the input size. The feature importance analysis was also performed for the L1C model (Figure 11c,d). The highest importance features for clear class detection according to Figure 11c are B08 and B8A, B01 for cloud shadow and B02 for clouds. In conclusion, B02, B01, B08, B8A, B11, B09 are the most significant input features for the KappaMask L1C model.



(**a**) – Feature importance per class for L2A model

(**b**) – Summed for all classes and sorted feature importance for L2A model

(**c**) – Feature importance per class for L1C model

(**d**) – Summed for all classes and sorted for L1C model

Figure 11. Feature importance obtained by deleting input features during inference for L2A model: (**a**) feature importance per class; (**b**) sorted feature importance, summed up for all the classes, and for L1C model: (**c**) feature importance per class; (**d**) sorted feature importance, summed up for all the classes.

3.4. Parameter Tuning

A 5-level U-Net [33] proved to be the most accurate, yet light-weight network for our purpose. U-Net with 6 or 7 levels overfit more quickly without an improvement in accuracy. Additionally, the impact of the number of input filters (32, 64 or 128) was analysed on the performance of U-Net with different depths (5, 6 or 7 levels), listen in Table 10. For U-Net with 5 levels, the performance was similar with 32, 64 and 128 input filters, whereas 64 filters offered only a small improvement. All other combinations of U-Net depths and input filters performed worse on the validation dataset. Furthermore, the performance of Unet++ [38] architecture with a nested skip connection structure was analysed. However, due to the high memory consumption of the multi-spectral input, the model was trained with a smaller batch size than the original U-Net and with an enhanced capability for generalization [39]. The architecture of the best performing model setup is shown in Figure 4.

Table 10. Training experiments for different model architectures for the L2A model.

Architecture	U-Net Level of Depth	Number of Input Filters	Max Dice Coefficient on Validation Set
U-Net	5	32	83.9%
		64	84.0%
		128	84.1%
	6	32	80.7%
		64	80.8%
		128	82.9%
	7	32	75.1%
		64	83.1%
U-Net++	5	64	75.9%

3.5. KappaMask Time Usage Comparison with Other Methods

The processing time for KappaMask prediction was measured in comparison to Fmask and S2cloudless on the same hardware in Table 11. Sen2Cor was not measured, since it is already precalculated as part of L2A product. The results were obtained on a test computer with the following hardware and software characteristics: CPU – Intel Core i7-8700K, 64GB of RAM, GPU – NVIDIA GeForce GTX 1070 with 8GB of VRAM, Linux Ubuntu 18.04.5 LTS (Bionic Beaver). The models were trained using University of Tartu High Performance Computing Center [40]. The code for running the KappaMask predictor is available open source [20].

Table 11. Time comparison performed on one whole Sentinel-2 Level-1C product inference. KappaMask Level-1C with GPU and CPU, Fmask, S2cloudless and DL_L8S2_UV on generating 10 m resolution classification mask.

	KappaMask on GPU	KappaMask on CPU	Fmask	S2cloudless	DL_L8S2_UV
Running time	03:57	10:08	06:32	17:34	03:33

4. Discussion

The KappaMask L1C and L2A models were trained on a dataset that represents various global and temporal conditions, but were later fine-tuned and validated on a dataset that covers Northern European terrestrial April-October conditions as showed in Figure 3. The model has not been tested outside of this scope. Both models outperformed Sen2Cor, Fmask, MAJA and S2cloudless in dice coefficient on created test set. Initially, the main emphasis was on Level-2A training since atmospherically corrected products are directly used in agricultural applications. Therefore, we believe that KappaMask Level-1C model can be further improved to get comparable results with the Level-2A model.

We highlight the importance of false negatives for agricultural applications, since falsely predicted clear area propagate into NDVI calculation for parcel time series, affecting the performance of end user applications (e.g., mowing detection, crop classification, etc.). We managed to reduce the number of false negatives for Level-2A model significantly; while Sen2Cor produced 60% false negatives for the clear class on our test dataset, our model's total false negatives for the clear class were 23%. The highest per class improvement was for cloud shadow with 20% better dice coefficient than Sen2Cor.

Besides the accuracy, the processing time for cloud mask generation is equally important to make it practical for operational applications. We measured the time needed for 10 m mask generation on GPU and CPU for KappaMask, which was shown to be 3:57 and 10:08 min, respectively. We compared processing time with other methods, such as rule-based Fmask and machine-learning based S2cloudless; the results were 6:32 and 17:34 min, respectively. Another deep learning algorithm produces a more efficient processing time of 3:33; however, it has a simpler two class classification map which results in a lighter and more efficient network. Considering that KappaMask is a CNN-based algorithm, the biggest concern with the deep learning methods is the processing time. We proved that CNN-based methods can be fast enough for practical applications. All results were obtained and compared on 10 m resolution. Sen2Cor, Fmask and MAJA 20 m masks was oversampled to 10 m resolution and S2cloudless output already had a 10 m resolution. Cloud mask output at 10 m is preferred, since it is the native spatial resolution of Sentinel-2 RGB and NIR spectral bands. A finer delineation helps to increase the amount of usable data for a large variety of higher-level products generation.

In future work, we will address the limitations of our current L1C and L2A models, extending the models' scope into the global area and including winter conditions. Another avenue would be to further improve the performance of Level-1C model. A comparison with other 10 m reference datasets, if there will be any, is planned as well.

5. Conclusions

KappaMask provides an accurate 10 m classification map for Sentinel-2 Level-2A and Level-1C products. It was trained on an open-source dataset and fine-tuned on a Northern European terrestrial dataset which was labelled manually using the active learning methodology. The dataset was split into training, validation and test sets, where test set consisted of unique S-2 products that were not overlapping with model training products but had the same spatio-temporal distribution. The test set covered the Northern European terrestrial region and included sub-tiles of 512 × 512 pixels size that were selected with the active learning approach. A comparison with Sen2Cor, Fmask, MAJA, S2cloudless and DL_L8S2_UV networks was carried out on the test set. The total dice coefficient on the test set for KappaMask L2A is 21% higher than Sen2Cor and 19% higher than Fmask. MAJA comparison was performed for clear and cloud classes and the dice coefficient is 29% behind KappaMask L2A. The closest machine learning open-source cloud mask that uses machine learning is S2cloudless with a 21% lower dice coefficient. S2cloudless, however, only provides cloud and clear classes. Another deep learning model, DL_L8S2_UV, has a 62% dice coefficient against 84% KappaMask L2A if we compare the clear and cloud output classes. KappaMask classifies pixels into clear, cloud shadow, semi-transparent and cloud classes and supports both L1C and L2A S-2 products. The total dice coefficient of KappaMask L2A model is 80%, KappaMask L1C model is 76%, while Fmask and Sen2Cor are 61% and 59%, respectively.

Author Contributions: Conceptualization M.D., K.V., I.S.; Data annotation O.W, F.H.; Model development M.D.; Software I.S., H.T., A.K., M.V., T.T.; Writing—original draft M.D.; Writing—review and editing I.S., O.W., A.O., M.J., E.G.C., N.L., T.T.; Project administration M.D., K.V., M.J., V.B., Supervision—V.B., N.L., E.G.C., Resources—T.T., E.G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Space Agency, Contract No. 4000132124/20/I-DT.

Data Availability Statement: The data presented in this study is publicly available from https://zenodo.org/record/5095024#.YWVuuhpByUk, accessed on 7 October 2021.

Acknowledgments: We would like to thank European Space Agency for their directional support, review and suggestions. We also thank Gholamreza Anbarjafari at the University of Tartu for the feedback and directions. The project was funded by European Space Agency, Contract No. 4000132124/20/I-DT.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

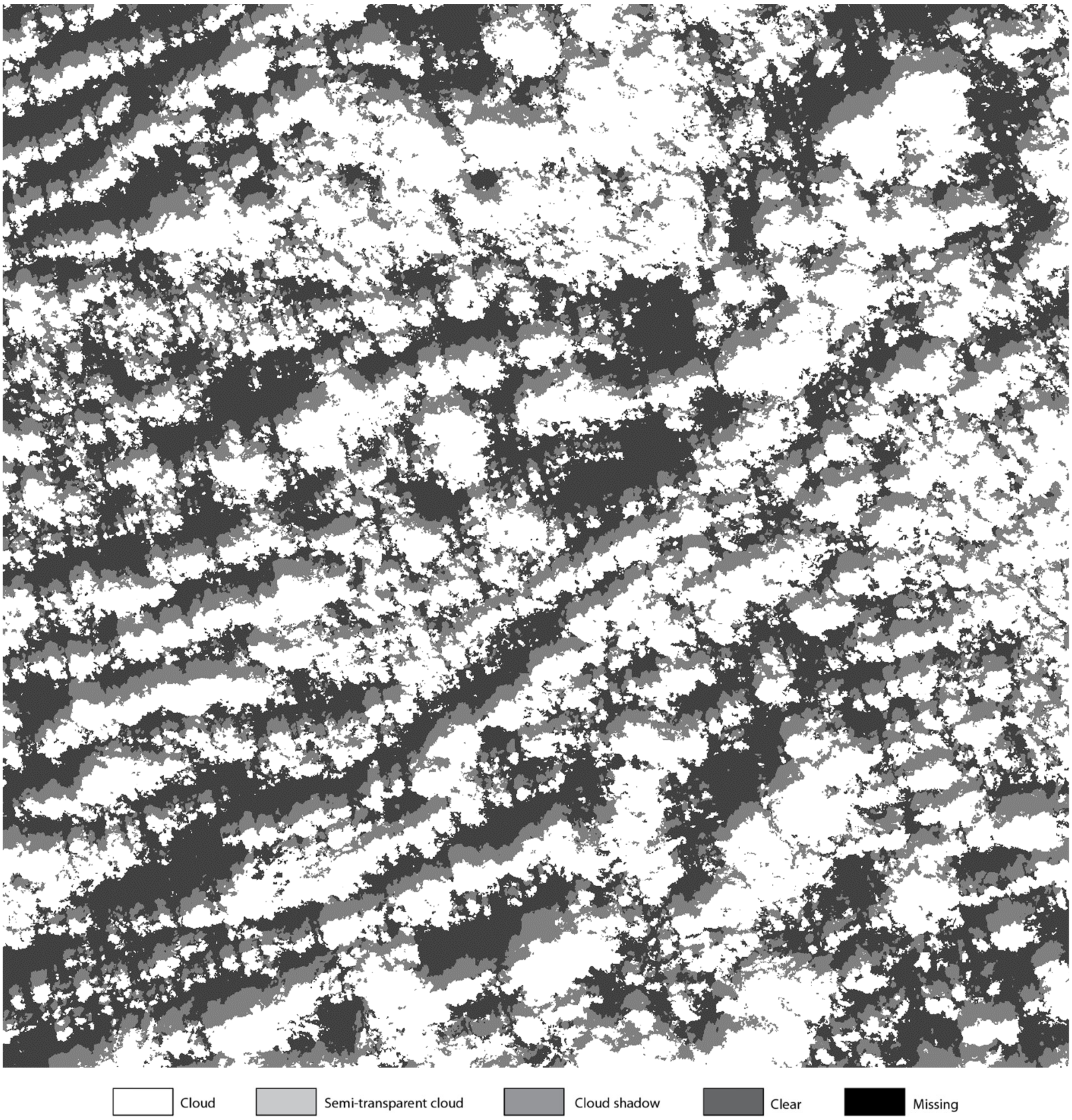
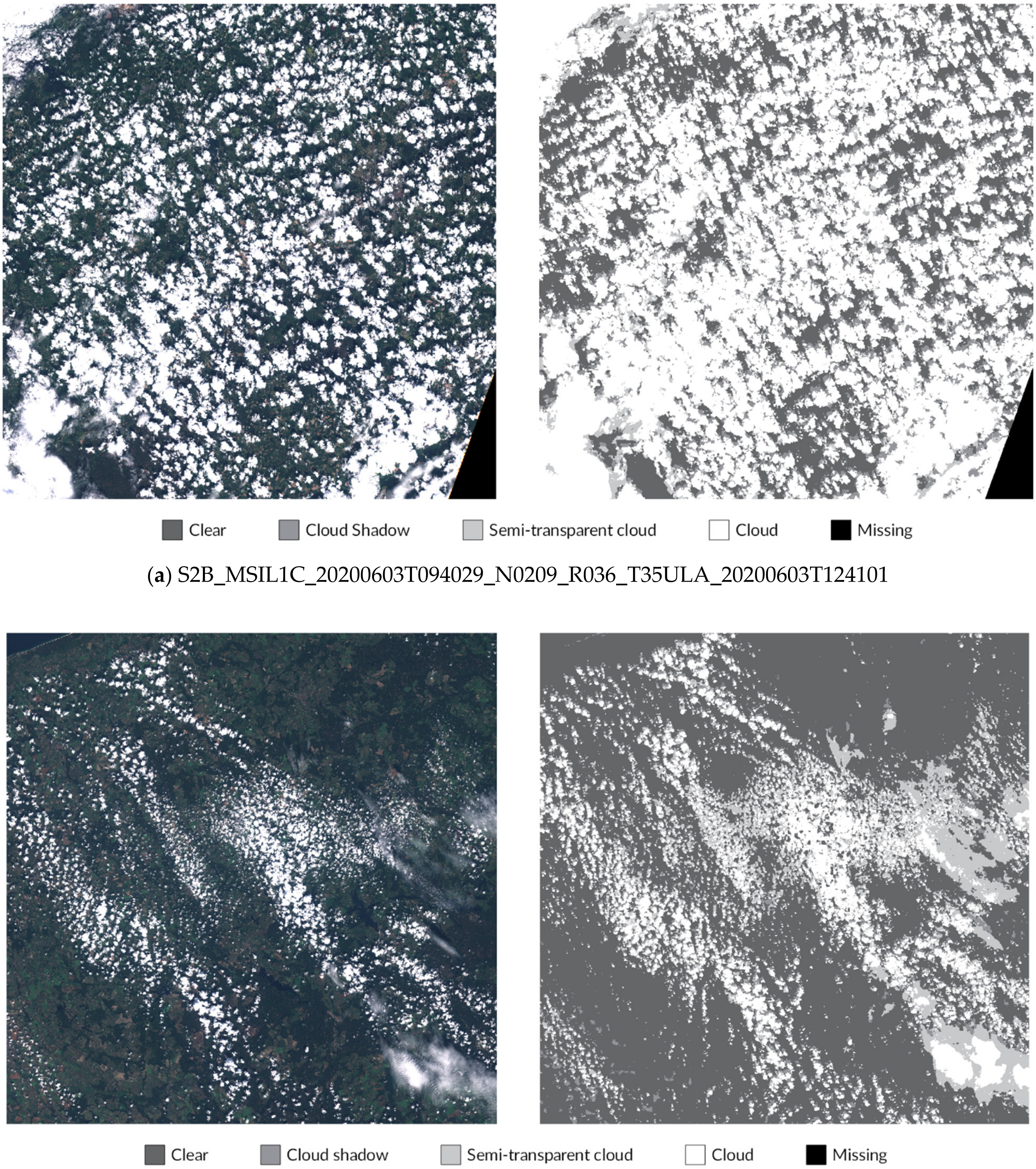


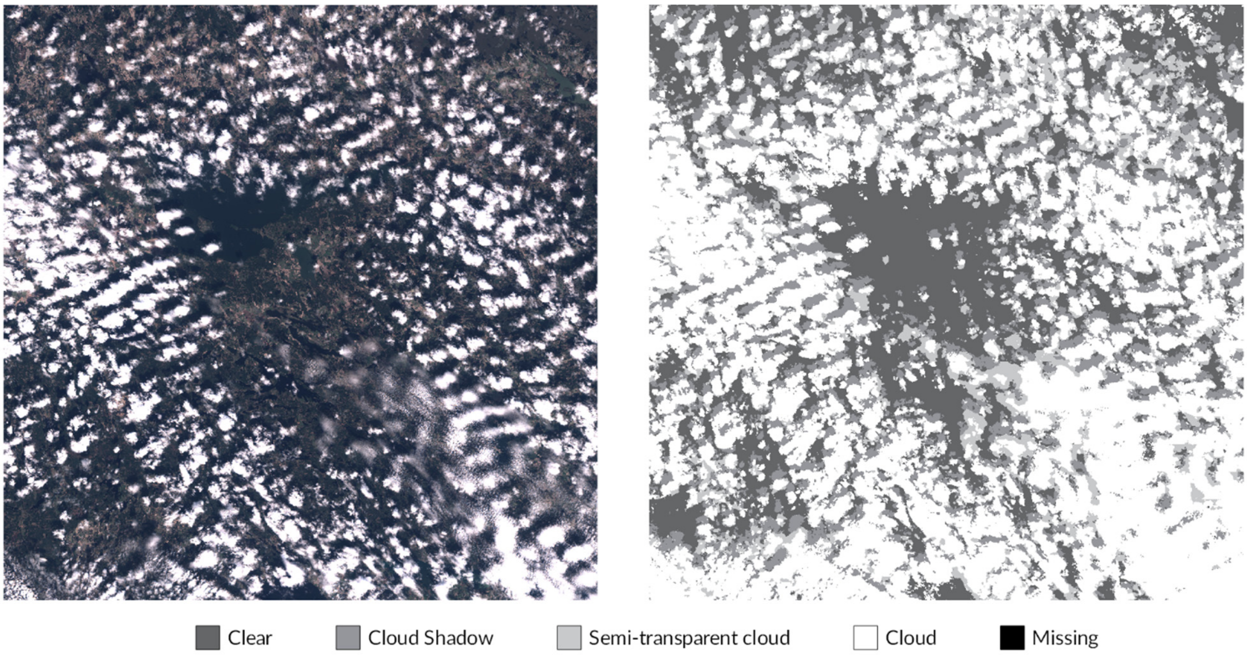
Figure A1. The entire Sentinel-2 L2A product classification map at 10 m resolution output (S-2 product S2A_MSIL2A_20200824T093041_N0214_R136_T35VND_20200824T121941).



(**a**) S2B_MSIL1C_20200603T094029_N0209_R036_T35ULA_20200603T124101

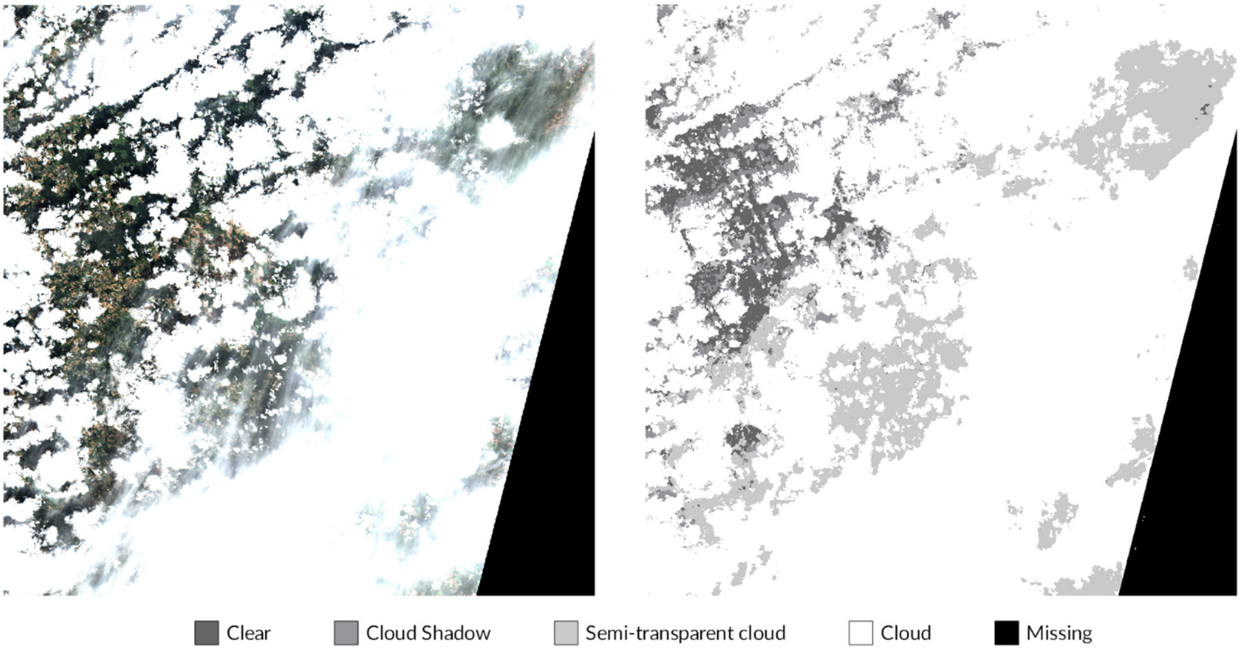
(**b**) S2A_MSIL1C_20200627T101031_N0209_R022_T33UWV_20200627T111749

Figure A2. *Cont.*



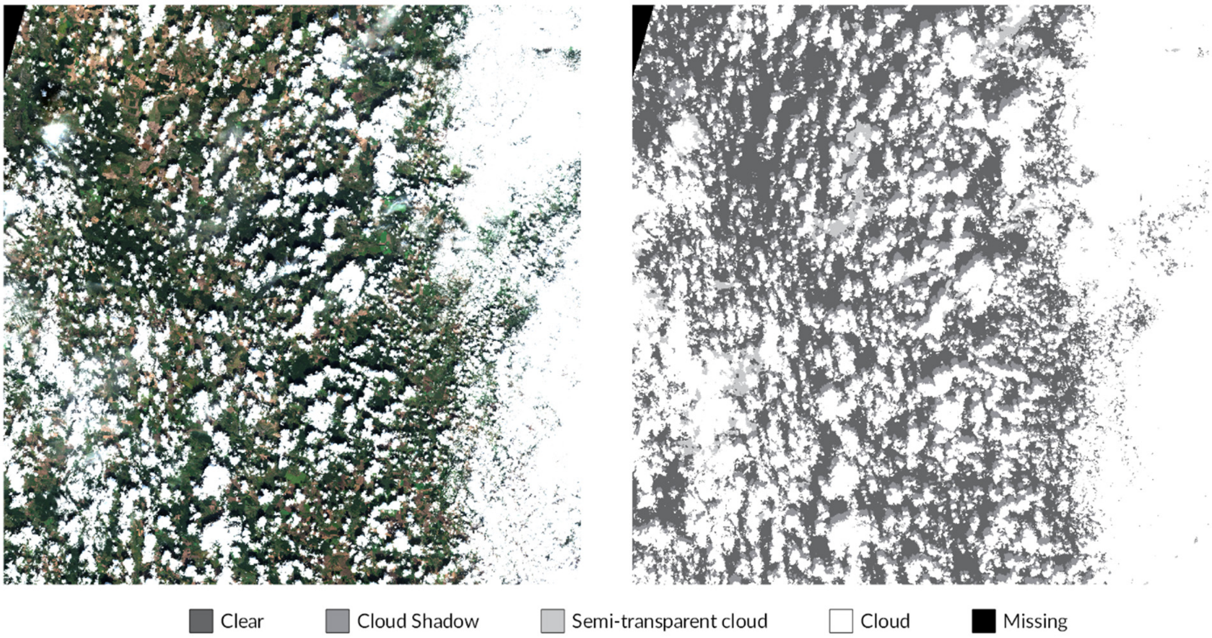
(**c**) S2B_MSIL1C_20200426T101549_N0209_R065_T33VWF_20200426T131809

Figure A2. The whole Sentinel-2 L1C product (**left**) and its classification map (**right**) at 10 m resolution.



(**a**) S2B_MSIL2A_20200824T101559_N214_R065_T32UQA_20200824T135410

Figure A3. *Cont.*



(**b**) S2B_MSIL2A_20200905T092029_N0214_R093_T35ULR_20200905T113748

Figure A3. The complete Sentinel-2 L2A product S2B_MSIL2A_20200905T092029_N0214_R093_T35ULR_20200905T113748 (**left**) and its classification map (**right**) at 10 m resolution.

References

1. Main-Knorn, M.; Pflug, B.; Louis, J.; Debaecker, V.; Müller-Wilm, U.; Gascon, F. Sen2Cor for Sentinel-2. In Proceedings of the Image and Signal Processing for Remote Sensing XXIII, Warsaw, Poland, 4 October 2017. [CrossRef]
2. Qiu, S.; Zhu, Z.; He, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sens. Environ.* **2019**, *231*, 111205. [CrossRef]
3. Hagolle, O.; Huc, M.; Pascual, D.V.; Dedieu, G. MAJA. Available online: https://github.com/CNES/MAJA (accessed on 10 October 2021).
4. Zupanc, A. Improving Cloud Detection with Machine Learning. 2017. Available online: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13 (accessed on 18 November 2020).
5. FastAI. Available online: https://github.com/fastai/fastai (accessed on 22 November 2020).
6. LightGBM. Available online: https://lightgbm.readthedocs.io/en/latest/ (accessed on 10 October 2021).
7. Drönner, J.; Korfhage, N.; Egli, S.; Mühling, M.; Thies, B.; Bendix, J.; Freisleben, B.; Seeger, B. Fast Cloud Segmentation Using Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1782. [CrossRef]
8. Wieland, M.; Li, Y.; Martinis, S. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sens. Environ.* **2019**, *230*, 111203. [CrossRef]
9. Jeppesen, J.H.; Jacobsen, R.; Inceoglu, F.; Toftegaard, T.S. A cloud detection algorithm for satellite imagery based on deep learning. *Remote Sens. Environ.* **2019**, *229*, 247–259. [CrossRef]
10. Foga, S.; Scaramuzza, P.L.; Guo, S.; Zhu, Z.; Dilley, R.D.; Beckmann, T.; Schmidt, G.L.; Dwyer, J.; Hughes, M.J.; Laue, B. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* **2017**, *194*, 379–390. [CrossRef]
11. L8 SPARCS Cloud Validation Masks. 2016. Available online: https://www.usgs.gov/core-science-systems/nli/landsat (accessed on 10 October 2021).
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
13. Li, Z.; Shen, H.; Cheng, Q.; Liu, Y.; You, S.; He, Z. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 197–212. [CrossRef]
14. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Hong, J.; Sun, J.; Zhang, Y.; Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
15. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.; Emery, W. Active Learning Methods for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [CrossRef]
16. Baetens, L.; Desjardins, C.; Hagolle, O. Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. *Remote Sens.* **2019**, *11*, 433. [CrossRef]

17. Baetens, L.; Hagolle, O. Sentinel-2 Reference Cloud Masks Generated by an Active Learning Method. Available online: https://zenodo.org/record/1460961#.YWMSJ9pByUk (accessed on 18 November 2020).
18. Li, J.; Wu, Z.; Hu, Z.; Jian, C.; Luo, S.; Mou, L.; Zhu, X.X.; Molinier, M. A Lightweight Deep Learning-Based Cloud Detection Method for Sentinel-2A Imagery Fusing Multiscale Spectral and Spatial Features. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1–19. [CrossRef]
19. Wu, Z.; Li, J.; Wang, Y.; Hu, Z.; Molinier, M. Self-Attentive Generative Adversarial Network for Cloud Detection in High Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1792–1796. [CrossRef]
20. KappaMask Predictor. Available online: https://github.com/kappazeta/cm_predict (accessed on 10 October 2021).
21. López-Puigdollers, D.; Mateo-García, G.; Gómez-Chova, L. Benchmarking Deep Learning Models for Cloud Detection in Landsat-8 and Sentinel-2 Images. *Remote Sens.* **2021**, *13*, 992. [CrossRef]
22. CREODIAS. Available online: https://creodias.eu/data-offer (accessed on 19 November 2020).
23. Data and Information Access Services (DIAS). Available online: https://www.copernicus.eu/en/access-data/dias (accessed on 18 January 2021).
24. Copernicus Open Access Hub. Available online: https://scihub.copernicus.eu/ (accessed on 10 October 2021).
25. PEPS: French Access to the Sentinel Products. Available online: https://peps.cnes.fr/rocket/#/home (accessed on 19 November 2020).
26. The Finnish Data Hub. Available online: https://nsdc.fmi.fi/services/service_finhub_overview (accessed on 19 November 2020).
27. Francis, A. Sentinel-2 Cloud Mask Catalogue. Available online: https://zenodo.org/record/4172871#.X6popcgzZaR (accessed on 7 March 2021).
28. Sentinel-2 KappaZeta Cloud and Cloud Shadow Masks. Available online: https://zenodo.org/record/5095024#.YQTuzI4zaUk (accessed on 10 October 2021).
29. Computer Vision Annotation Tool. Available online: https://cvat.org/ (accessed on 10 October 2021).
30. Segments.ai Dataset Tool. Available online: https://segments.ai/ (accessed on 10 October 2021).
31. Francis, A. 'IRIS Toolkit'. Available online: https://github.com/ESA-PhiLab/iris (accessed on 10 October 2021).
32. CEOS-WGCV ACIX II CMIX Atmospheric Correction Inter-Comparison Exercise Cloud Masking Inter-Comparison Exercise 2nd Workshop. Available online: https://earth.esa.int/eogateway/events/ceos-wgcv-acix-ii-cmix-atmospheric-correction-inter-comparison-exercise-cloud-masking-inter-comparison-exercise-2nd-workshop (accessed on 10 October 2021).
33. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
34. Kingma, D.P.; Ba, J. 'Adam: A Method for Stochastic Optimization'. 2017. Available online: https://arxiv.org/abs/1412.6980 (accessed on 10 October 2021).
35. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *BT—Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2017; pp. 240–248.
36. KappaMask Comparison with Rule-Based Methods. Available online: https://kappazeta.ee/cloudcomparison (accessed on 10 October 2021).
37. Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
38. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef] [PubMed]
39. Hoffer, E.; Hubara, I.; Soudry, D. Train Longer, Generalize Better: Closing the Generalization Gap in Large Batch Training of Neural Networks. May 2017. Available online: http://arxiv.org/abs/1705.08741 (accessed on 10 October 2021).
40. University of Tartu. "UT Rocket". share.neic.no. Available online: https://share.neic.no/marketplace-public-offering/c8107e145e0d41f7a016b72825072287/ (accessed on 10 October 2021).