Article

Semantic Segmentation with High-Resolution Sentinel-1 SAR Data

Hakan Erten [1], Erkan Bostanci [1], Koray Acici [2], Mehmet Serdar Guzel [1], Tunc Asuroglu [3,*]
and Ayhan Aydin [4]

1   Department of Computer Engineering, Ankara University, 06830 Ankara, Turkey
2   Department of Artificial Intelligence and Data Engineering, Ankara University, 06830 Ankara, Turkey
3   Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland
4   Department of Computer Technologies, Ankara University, 06830 Ankara, Turkey
*   Correspondence: tunc.asuroglu@tuni.fi

Abstract: The world's high-resolution images are supplied by a radar system named Synthetic Aperture Radar (SAR). Semantic SAR image segmentation proposes a computer-based solution to make segmentation tasks easier. When conducting scientific research, accessing freely available datasets and images with low noise levels is rare. However, SAR images can be accessed for free. We propose a novel process for labeling Sentinel-1 SAR radar images, which the European Space Agency (ESA) provides free of charge. This process involves denoising the images and using an automatically created dataset with pioneering deep neural networks to augment the results of the semantic segmentation task. In order to exhibit the power of our denoising process, we match the results of our newly created dataset with speckled noise and noise-free versions. Thus, we attained a mean intersection over union (mIoU) of 70.60% and overall pixel accuracy (PA) of 92.23 with the HRNet model. These deep learning segmentation methods were also assessed with the McNemar test. Our experiments on the newly created Sentinel-1 dataset establish that combining our pipeline with deep neural networks results in recognizable improvements in challenging semantic segmentation accuracy and mIoU values.

Keywords: SAR; SAR segmentation; deep learning; semantic segmentation; Sentinel-1

1. Introduction

Land Cover Maps characterize essential information about the earth's surface and land usage. It helps to observe the subjects of some of the most critical topics of today's world, such as deforestation, urban and environmental monitoring and climate change. Namely, these maps give us tips about the future of our world. Generally, satellite optical imagery is used for land-cover mapping studies. Because several remote sensing platforms create representations with very high-resolution optical images, and detect all details about land, such as edge, shape and texture, and in very high-resolution images, resolution range is between submeter-scale and several meters [1–5].

Besides optical images, radar images are also used for mapping, and SAR is one of the most popular. SAR is an active sensor that transmits microwave signals to the earth and recovers signals that return or scatter from the earth's surface. SAR can be mounted on manned/unmanned aerial vehicles or satellites and is capable of high-resolution imaging, target detection in all weather conditions, including rainy or cloudy conditions, and provides images day and night. With these capabilities, SAR provides a wide range of uses due to its high-resolution images (some of which are exact and quickly processed) [6], target detection, images of the earth, evaluation of agricultural areas and forests and mass control of glaciers. With the increase in high-resolution SAR images [7], big SAR datasets have begun to form. Some frameworks help to label each pixel with different classifiers [8–10]. However, labeling SAR images still needs to be clarified and made easier. Thus, SAR images

have too much noise due to its imaging procedure [11], complex backscattering [12] and geometrically-significant distortion [13]. When collecting data, soil moisture, vertical and horizontal form of the scatterers, roughness on the surface and geometry can obscure the SAR image because of the backscattering microwave radiation. In short, seasonal, temporal and environmental effects on the earth change the dataset and can cause shortages of generalization capabilities. These difficulties when obtaining SAR images, combined with the increased resolution made creating a new dataset almost impossible. Additionally, finding a labeled free dataset for scientific research is almost impossible.

CORINE, a program started in 1985, aims to gather information for the European Union on priority environmental issues (e.g., air, water, soil, land cover, coastal erosion and biotopes). CORINE utilizes the land cover/use data produced by computer-aided visual interpretation methods over satellite images according to the Land Cover/Use Classification determined by the European Environment Agency (EEA) [14]. According to the EEA criteria and classification units (44 classes), changes in land cover/land use for monitoring the land via satellite images are detected with the help of remote sensing and geographic information systems. The main ideas of the CORINE projects are managing the environment, reducing the effects of climate change and helping to observe urban and environmental developments.

SAR image processing is a challenging task [15]. One of the significant steps in the processing of SAR images is segmentation. The main aim of segmentation with SAR image is to split the image into parts according to separate attributes without any intersection and to mark each pixel in the image according to one of the classes [16]. With most accessible datasets, methods based on the many different convolutional neural network (CNN) [17–19] and the recurrent neural network (RNN) [20] have been used for the SAR image segmentation task, and successful results have been achieved. With the increase in access to big datasets with high resolution, pixel-wise segmentation has become impossible, since labeling SAR images is difficult. Superpixel generation models use superpixel-wise segmentation methods [21,22] for big SAR datasets. The methods are divided into two categories according to their superpixel generation: clustering-based and graph-based [22]. Here, in order to use affective pixel-wise CNN segmentation models, we present a new pipeline.

As seen above, the need for labeled free datasets, together with the presence of SAR image noise, backscattering and geometric distortion are the main issues for pixel-wise segmentation. The previous research efforts have mainly concentrated on speckle noise [23,24] and on generating labels for these noisy datasets. While SAR images are available for free, reaching the noise-free labeled dataset is still difficult. To solve this problem, we propose a novel process for easily creating a new dataset. The process starts with downloading SAR images, uses CORINE and ends with creating labeled forms ready for SAR image segmentation tasks, as detailed in Section 3.

Once we created our dataset, we trained U-Net, PSPNet and HRNet deep learning models effectively for the semantic segmentation task. The overview of our approach is presented in Figure 1.

Identifying the qualities and shortcomings of deep learning methods is a significant task. To specify which methods are better and how methods differ from each other, we utilize McNemar's test to evaluate the success of the models: U-Net, PSPNet and HRNet. McNemar's test has been employed in various previous research, as discussed in our evaluations. Five machine-learning models were analyzed, with the classification results assessed according to the methods in [25]. In ref. [26], a similar evaluation was performed to compare the class-wise prediction against an extensive database.
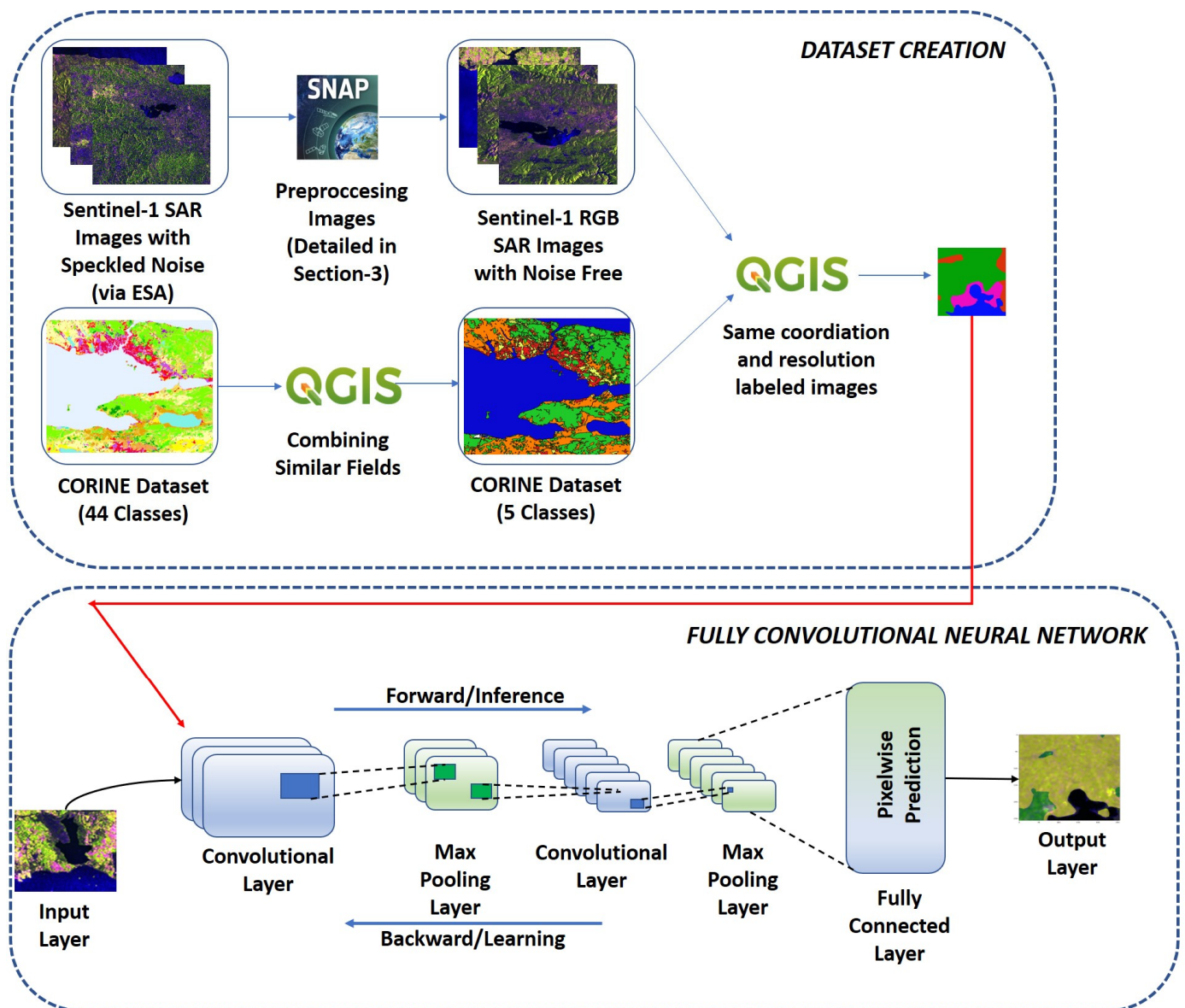
Figure 1. The overview of our proposal.

The essence of this work is explained as follows:

(1) We propose a dataset creation pipeline that can be performed easily;
(2) We train the latest deep learning segmentation models, instead of traditional ones, with the newly created dataset;
(3) We demonstrate the robustness of the proposed pipeline by training and then comparing results against noisy and noise-free versions of our dataset;
(4) We evaluate the performance of deep learning methods with McNemar's test.

The rest of this study is structured as follows. We give a brief review of related works in Section 2. In Section 3, the dataset creation pipeline is introduced. In Section 4, models utilized and experimental setups are detailed, step by step. Analyses and results are represented in Section 5. We then conclude this work and make recommendations for future work.

2. Related Works

Many different algorithms have been applied, and many datasets have been used for segmenting SAR images, by many earlier research efforts. At the same time, there are many

common problems encountered while segmenting. In this part, we first mention existing datasets, then provide an overview of common problems, and finally, briefly provide information about related work.

2.1. Most-Used SAR Datasets

SAR images are obtained from air and satellite platforms. The most-utilized SAR dataset for image processing is the MSTAR [27], created by the U.S. Air Force based on HH polarization X-band images. This dataset has images of armored vehicles of 15 different types. There are a few hundred images, all of which are convenient for image processing.

Some of the most recent SAR datasets are OpenSARShip 2.0 [28], JAXA's ALOS PALSAR [29] forest/non-forest, and the SARptical [30] dataset. OpenSARShip 2.0 includes 41 images of Sentinel-1 and nearly 11,000 various ships. JAXA's ALOS PALSAR forest/non-forest dataset is used for land-cover analysis. Additionally, the SARptical provides TerraSAR-X data.

The main differences between SAR datasets are that they can be obtained from different sources using separate bands, and that they have different resolutions. Resolution and noise of the datasets are some of the main factors affecting success.

2.2. Common Problems for SAR Image Segmentation

Main problems encountered in this area are as follows:

1.　SAR images have a low resolution for segmentation tasks;
2.　SAR image noise, backscattering and geometric distortion;
3.　No easy access to a free, labeled dataset, though some SAR images can be accessed through paid services;
4.　The numbers of training and test images in the datasets used in experiments are low due to the fact that the labeled dataset is not open source.

2.3. SAR Processing Models

Studies have shown that many different datasets have been used in many areas, such as road segmentation, land cover, wetland classification, etc.

For road segmentation, DeepLabV3+, FCN-8s and Deep Residual U-Net deep neural models were compared in [17]. X-band TerraSAR-X images were trained and the resulting IoU value for FCN-8s was 45.46%, the precision value was 71.69% and the recall value was 75.17%. Meanwhile, Deep Residual Unet IoU's value was 40.18%, while DeepLabV3+ IoU value was 40.18%.

Some efforts have also been made on the land cover segment. There were comparisons of roads, built-up areas, and water and vegetation via POLSAR image segmentation with MIoU values and FcN and Unet models in [18]. With the present study, we aim to increase the success rate by using small datasets. While the FCN mean IoU value was 0.50, the U-Net mean IoU value was 0.44. In ref. [31], Sentinel-1 SAR images were used with state-of-the-art segmentation models such as U-Net, DeepLabV3+, FC-DenseNet, BiSeNet, PSPNet, SegNet and FRRN-B; also, the results were compared based on their accuracy. The best overall accuracy through these experiments was achieved by the FC-DenseNet (Fully Convolutional DenseNets) model, with 92.78%.

In ref. [19], a novel architecture was proposed with polarimetric SAR (PolSAR) imagery, and it classified the wetland complexes. This work compared results from RF, FCN-16s, FCN-8s, FCN-32s and Segnet models. The F-1 score of the proposed FCN model was 0.84, and the F1-score of the RF model was 0.74. F1-scores achieved by FCN-32s, FCN-16s, FCN-8s and Segnet models were as follows: 0.69, 0.75, 0.81, 0.76, respectively. The proposed FCN model had an increased F1 score when compared with existing models.

In ref. [32], a novel matrix factorization active-contour model based on fused features was proposed for water/land segmentation in SAR images. Wavelet textual features, Gaussian (DoG) filter features, and Gabor filter features were compounded, and feature

matrixes were composed. Then, with the SAR image's edge and region data, the energy procedure was created.

Generally, unsupervised image segmentation models are preferred in SAR segmentation because they are easy and useful. In ref. [33], a novel model with region smoothing and label correction (RSLC) was explained in detail. To resemble the spatial information in the polynomial result of the image, smoothing results were used. This establishes that the process of segmentation may be accomplished rapidly and effectively. One of the other effective and thriving segmentation projects is a convolutional deep neural network (CDNN) [34], which has been proposed to classify crop areas within SAR satellite images, along with the cultivation status of the crop. Using a high pass linear spatial (HLS) filter, the segmented dataset is preprocessed in training. After that, a modified region growing (MRG) algorithm is used for classification in the testing phase, and with Euclidean distance (ED), the cultivation status of each crop can be specified. Additionally, a new method was proposed in [35], based on brief presentation and a hierarchical fuzzy C-means (FCM) advance. Firstly, the image is divided into pixel groups with their features, and the major pixels which consist of pixel groups are then used for constructing the thumbnail. After that, FCM is utilized for segmentation.

Our work differs from those mentioned above in that it explains the new dataset creation process from beginning to end. After that, popular semantic segmentation methods are used and evaluated with mIoU, MA, and PA metrics with our newly-created dataset. In this way, with this paper, a detailed study has been proposed, proceeding from the creation of a SAR dataset to explaining how it can be used with deep learning methods.

3. Dataset Creation

Although most used datasets are detailed above, here we propose a novel pipeline that easily generates labels for SAR images. With our proposal, datasets varying in size can be created in accordance with the needs of the models. First of all, as shown in Table 1, C-band Sentinel-1 images with IW sensor mode and a ground range detected (GRD) product type, released freely by ESA and representing specific parts of Turkey, are obtained. These images are then preprocessed with the Sentinel Application Platform (SNAP) application. One of the most critical factors is preprocessing, which affects the success rate when training with Sentinel images. For Sentinel-1 grid data, a standard workflow is presented in [36]. The following steps have been used for preprocessing: [36]

1. Thermal noise removal;
2. Application of orbit file;
3. Border noise removal;
4. Calibration;
5. Speckle filter;
6. Terrain correction.

After this preprocessing procedure, images are exported in the high resolution RGB format to be used in the segmentation process. There are three different methods for obtaining RGB images: dual pol ratio, dual pol multiple and dual pol difference. When dual pole ratio is preferred, it uses the red VV channel, green VH channel, and blue VV/VH channel. For our study, three high resolution images were acquired with dual pol ratio. After that, these were cut apart to a pixel resolution of 256×256, and 6722 images with 256×256 pixel were acquired. Example images are shown in Figure 2.

Table 1. Information about the newly-created dataset.

Istanbul, Turkey	
Size	12,288 × 12,288 px.
Geo Coordinates West, East	28.443, 29.546
Geo Coordinates South, North	39.96, 41.063
Reference Information	20210123 036260 0440DA 330D
Izmir, Turkey	
Size	12,228 × 12,228 px.
Geo Coordinates West, East	26.604, 27.703
Geo Coordinates South, North	38.654, 39.752
Reference Information	20210128 025357 03051D 2D81
Adana, Turkey	
Size	12,228 × 12,228 px.
Geo Coordinates West, East	34.906, 36.004
Geo Coordinates South, North	36.643, 37.742
Reference Information	20210120 025240 03015D DC9A
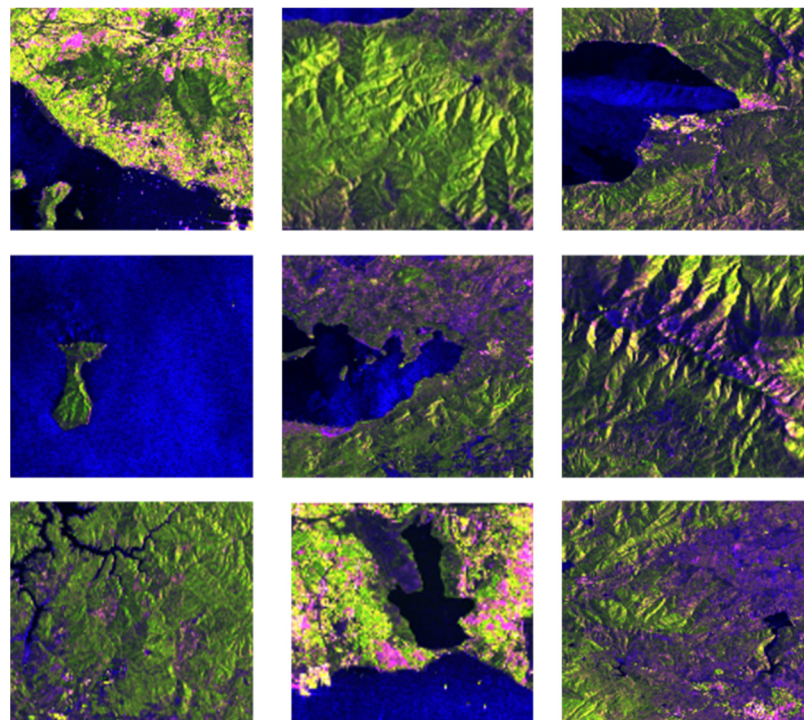
Figure 2. Examples of 256 × 256 RGB images.

In order to use these images for supervised pixel semantic segmentation, the next step is acquiring labels for images. The CORINE dataset [14], freely available in Copernicus, was downloaded. In the original, this dataset consists of labels for 44 classes. In this study, with the help of QGIS program, we decreased the count of labels to five, as: forests, urban, peatland bogs and marshes, agricultural areas, and water.

After the number of classes in the CORINE Land Cover dataset was reduced to five, three images with the same resolution and coordinates were obtained from the SNAP via the QGIS. As a result, labeled RGB images with the same resolution and coordinates were made ready for semantic segmentation.

4. Experiments and Models

In this section, we first give general information about Sentinel-1, which is followed by a description of the training process, the evaluation metrics, and the models used in this study. We will finish the section by explaining McNemar's test and how it measures success among deep learning models.

4.1. Sentinel-1

SAR is a method of creating images using radio waves. Due to the movement of the radar, a large artificial gap will be created when a gap in the synthesis of data is formed by taking measurement values at certain intervals and collecting these values at the same time [37]. The artificial aperture creates the equivalent of a large physical antenna, and it provides high resolution images. There are many different types of satellites with discrete bands or modes.

Sentinel is the name of the program called Copernicus run by the European Space Agency (ESA). Sentinel-1 consists of 2 satellites (Sentinel-1A and Sentinel 1-B), which both share the same orbit. Sentinel-1A and Sentinel-1B were launched, respectively, in 2014 and 2016. Sentinel-1 carries a C-band, dual-polarized synthetic aperture radar that enables data collection day and night in all weather conditions. Sentinel-1 can collect data in 4 different modes, which are: wave (WV), strip-map (SM), extra-wide swath (EW) and interferometric wide swath (IW). The main mode is the IW mode, which acquires a total representation of the earth's surface and satisfies the mass of service demands [38]. This is the reason why IW has been used in this paper.

4.2. Training

Models used in this study were implemented using MMSegmentation [39] and trained on a single GTX 1080TI with 11 gigabytes of VRAM. MMSegmentation is a framework based on the PyTorch library, and its toolbox is open-source and appropriate for segmentation. This framework represents various semantic segmentation methods. One can construct customized datasets and framework by uniting diverse modules. The training speed is outstanding. For training, 80% of our noise-free newly-created dataset images were used, and the rest of the images were used for testing. Data augmentation was performed and input data was normalized on the training set. Flipping training images horizontally and vertically is used for data augmentation. A drop in our performance was noticed when pre-trained models, which MMSegmentation incorporates, were used in our models. The ADAM optimizer was used, with a learning rate 0.003 and an exponential learning rate decay of 0.90 applied after each iteration. Each model was trained for 80 k iterations and, during this training phase, the best model's checkpoint was preserved and then used to evaluate the test set.

We used mIoU, *MA*, and *PA* for the evaluation, as shown in Equations (2) and (3) [27]. IoU is a robust metric for the assessment of semantic segmentation; it shows performance of the overlapping ratio between intersection and all surface data (union). IoU is calculated as shown in Equation (1). FP and TP stand for false positive and true positive, and FN means false negative.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$PA = \frac{\sum_i n_{i,i}}{\sum_i \sum_j n_{i,j}}, MA = \frac{1}{C} \sum_i \frac{n_{i,i}}{\sum_i j} \tag{2}$$

$$\text{mIoU} = \frac{1}{C} \sum_i \frac{n_{i,i}}{\sum_j n_{i,j} + \sum_j n_{j,i} + n_{i,i}} \tag{3}$$

where $n_{i,j}$: number of pixels, *i*: target class, *j*: predicted class, *C*: number of classes, IoU: intersection over union, *PA*: pixel accuracy, *MA*: mean accuracy, and mIoU: mean intersection over union.

4.3. Models

We selected the most vital segmentation models to test our dataset, as: U-Net, PSPNet and HRNet.

U-net was used successfully in the biomedical field at first and become one of the most-applied models for the SAR image segmentation task. The reason why it is preferred for SAR images is that it can obtain more successful results with less training data. U-Net's name comes from its architecture, which resembles the letter U, as presented in the figure. This model has two parts that are symmetric. The first has convolutional layers which are transposed and 2D, and, on the other hand, the other one consists of known convolutional layers.

A pyramid scene separation network (PSPNet) [40] is designed to better acquire the broad contextual representation of the scene. Patterns are derived from the input images using an extended mesh strategy and with the help of a feature extractor. Feature maps are then fed into a pyramid pooling module to separate models at divergent scales. Then, they are summarized on four different ranges, each corresponding to a pyramid level, and modified with a 1×1 layer to decrease their size. By this method, each pyramid level analyzes the lower regions of the image with different positions. The outputs of the pyramid levels are eventually upgraded and combined into initial feature maps to include local and general context information. Then, these outputs are processed by a last layer to output pixel-based predictions.

The high resolution net (HRNet) is proposed at first for estimation of human-readable data. This model processes high resolution images into the process. Then, the model connects resolution convolutions high-to-low in parallel and generates very high-resolution images by repetitively managing fusions across parallel convolutions [41]. HRNet is used for a wide range of visualization tasks, such as semantic segmentation and object detection. Experimental results shown in [41] prove that HRNet is one of the recent and popular models.

4.4. McNemar's Test Results

This test determines whether a learning model is better than another at a specific task. The McNemar test performs analysis by taking the number of couples that are positive but then negative and the number of couples that are negative at first but later become positive. The test utilizes a z-score for measuring the differences, as shown in Equation (4).

$$Z = \frac{(|A_{1,0} - A_{0,1}| - 1)}{\sqrt{A_{1,0} + A_{0,1}}} \tag{4}$$

Suppose that there are two algorithms X and Y. $A_{1,0}$ represents X algorithm succeeding and Y algorithm failing and $A_{0,1}$ shows X algorithm failing and Y algorithm succeeding. After the calculation, when the z-score = 0 we can say that the two algorithms have shown similar performance. If the value of the z-score increases past 0 in the positive direction, this points out that their performance differs significantly; confidence levels corresponding to the z-score are presented in Table 2 [42]. One-tailed prediction shows the effect of a change in one direction and not the other. A two-tailed prediction does not have a direction and does not have to be specified prior to testing. A two-tailed prediction has the possibility of both a positive and a negative effect.

Table 2. Confidence levels corresponding to z scores for one-tailed and two-tailed predictions.

Z-Score	One-Tailed Prediction	Two-Tailed Prediction
1.645	95%	90%
1.960	97.5%	95%
2.326	99%	98%
2.576	99.5%	99%

For our study, first we calculated the PA value of each image used for validation for each segmentation model. After that, we compared the PA values for each image with each other (using HRNet-PSPNet, HRNet-UNet and PSPNet-UNet). The algorithm which had a higher PA value was set to 1, and the smaller one was set to 0. Then, the z-scores were calculated and models were compared with each other.

5. Experimental Evaluation and Discussion

Here we will show the results and performance analysis of the models. After that, we will evaluate the performance of methods by using McNemar's test.

5.1. Results

First, we obtained SAR images of certain regions of Turkey, which ESA provides free of charge. Before labeling the SAR images, we performed six preprocesses to eliminate noise and make geometric corrections. In order to label the data, the 44-class CORINE dataset was reconstructed as five classes by combining similar features using the QGIS program. When combining the CORINE dataset and the SAR images, the exact coordinates and the same pixels were superimposed via the QGIS program. As a result, we created a new dataset with the acquisition of SAR images, and after the preprocessing steps, due to the labelling of the SAR images, we combined CORINE and noise-free SAR images. As detailed in Section 4, we evaluated our dataset using three semantic segmentation deep learning models, namely, U-Net, PSPNet and HRNet. Figure 3 shows ground-truth annotations on the left and images resulting from HRNet segmentation on the right column. By this, results obtained are compared with the original annotated images.
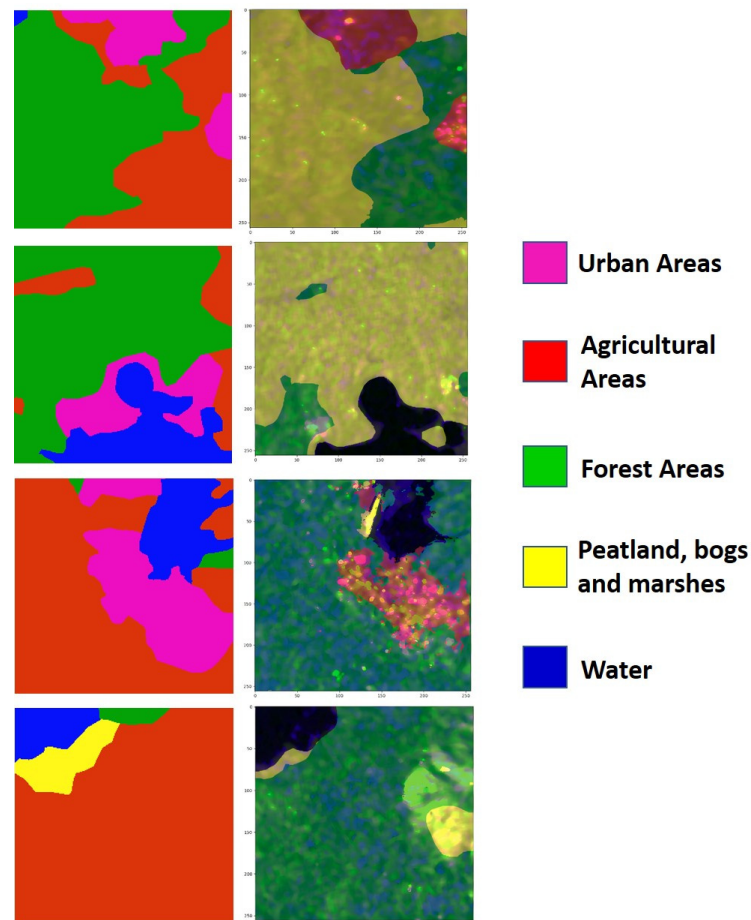

Figure 3. Illustration of the HRNet model performance: segmentation results, ground-truth CORINE annotations (**left column**) and segmented examples (**right column**).

5.2. Segmentation Performance

In this section, we evaluate the results of the UNet, PSPNet, and HRNet models given in detail above. For our models, the training times took from 1 to 2 days. Using multi-GPU systems could offer significant improvements, and more images for the dataset could then be made available. All models shown on the newly-created dataset achieved an overall pixel accuracy above 88%. We reached respective overall PAs of 92.23%, 90.74%, and 88.59% and mIoUs of 70.60%, 65.99%, and 62.87% with the HRNet, PSPNet and UNet models. These results can be seen in Table 3. The overall segmentation results for all studied models are given in Table 4. There we present MA and mIoU values for five classes and compare the results with the new dataset. As seen in the table, preprocessing steps increased success by approximately 3% in every class and proved that preprocessing is essential for SAR segmentation. The purpose of choosing the three biggest cities of Turkey was that the urban areas were more prevalent in terms of the surface area in the dataset. However, urban areas' scores were the lowest ranked in this study, as they represented the smallest part of the dataset. The highest success obtained in these areas was water, as the water areas were readily apparent.

Table 3. Summary of the segmentation performance of deep learning models.

	U-Net	PSPNet	HRNet
mIoU (%)	62.87	65.99	**70.6**
MA Value (%)	73.41	75.33	**80.47**
Overall PA (%)	88.59	90.74	**92.23**

Table 4. McNemar test results for deep learning models.

	HRNet	PSPNet	U-Net
HRNet	X	← 3536	← 18,685
PSPNet	X	X	← 15,423
U-Net	X	X	X

In Table 5, a summary of all segmentation evaluation results is shown. Here, PA values of all validation images are presented in order to compare each model with the others. When the images with the lowest PA values against the validation images are examined, it is seen that the urban areas are more often in the images. The PA values of two of the trained deep neural network models are higher than 90%, and the best-performing one among these models was the HRNet model, which achieved an overall PA of 92.23%.

We obtained different results in our study with images of the same three cities in our dataset taken at different times. It can be concluded that the results of images taken from identical regions at different times may differ within identical models because the backscattering of SAR diversifies the dataset.

Table 5. PA values of each validation image and results of comparison of each model with each of the others.

Number of Image	HRNet PA Value	PSPNet PA Value	U-Net PA Value	HRNet-PSPNet	
				A(1,0)	A(0,1)
1	1.000	1.000	1.000	0	0
2	0.921	0.770	0.766	1	0
3	0.892	0.847	0.768	1	0
4	0.800	0.829	0.741	0	1
5	0.967	0.967	0.964	0	1
6	0.986	0.997	1.000	0	1
7	1.000	1.000	1.000	0	0
8	0.898	0.903	0.868	0	1
9	1.000	1.000	1.000	0	0
10	0.809	0.764	0.846	1	0
11	1.000	1.000	1.000	0	0
12	0.884	0.684	0.730	1	0
13	1.000	1.000	1.000	0	0
14	1.000	1.000	0.996	0	0
15	0.975	0.979	0.979	0	1
1329	0.843	0.875	0.843	0	1
1330	0.872	0.817	0.723	1	0
1331	0.943	0.777	0.777	1	0
1332	0.764	0.742	0.876	1	0
1333	0.662	0.641	0.629	1	0
1334	0.855	0.860	0.773	0	1
1335	1.000	1.000	1.000	0	0
1336	0.971	0.971	0.971	0	0
1337	0.974	1.000	1.000	0	1
1338	0.885	0.901	0.900	0	1
1339	0.974	0.972	0.982	1	0
1340	0.767	0.789	0.767	0	1
1341	0.872	0.872	0.777	1	0
1342	1.000	1.000	1.000	0	0
1343	1.000	1.000	1.000	0	0
1344	0.834	0.836	0.733	0	1
1345	0.948	0.981	0.980	0	1

5.3. McNemar's Test Results

As can be seen from the experiments, the most successful results were obtained with HRNet model. Nonetheless, we also want to evaluate the models with McNemar's test to decide which model performs better. In Section 4, how we apply the McNemar test is explained in detail. Some of the PA values of the 1345 validation images and the comparison values of the models are shown in Table 5.

After the comparison made for each validation image, we calculated a z-score for each model against each of the others. Z-score values are represented in Table 6. The arrows

(←,↑) show which segmentation model presented better. In Table 6, it is clearly shown that HRNet has given a better segmentation performance than PSPNet or U-Net. PSPNet shows better performance than U-Net. The result of McNemar's test establishes the PA value that is presented in Table 5, distributed over almost all validation images. The results show that the confidence levels corresponding to the z-scores are more than 99%.

Table 6. Summary of the segmentation MA and mIoU values of various deep neural networks.

Classes	UNet				PSPNet				HRNet			
	Images with Noise		Images with Noise Free		Images with Noise		Images with Noise Free		Images with Noise		Images with Noise Free	
	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)
Urban Areas	12.29	25.21	18.06	34.84	14.86	27.91	15.20	32.06	21.13	39.77	27.13	44.27
Agricultural Areas	57.57	65.12	58.64	67.74	62.41	68.94	64.32	71.23	71.54	83.24	74.47	85.2
Forest Areas	69.09	80.34	71.35	83.46	73.87	84.04	77.43	87.92	72.33	82.74	74.06	84.71
Peatland, Bogs and Marshes	70.67	83.78	72.17	85.42	74.45	86.63	78.36	89.53	74.49	86.21	79.59	89.67
Water	92.67	94.15	94.15	95.58	92.11	94.21	94.66	96.90	93.51	95.56	97.78	98.49
Summary (Average)	60.46	69.72	62.87	73.41	63.54	72.34	65.99	75.53	66.6	77.50	70.60	80.47

5.4. Discussion and Future Work

The biggest problems in SAR segmentation are the need for labeled data, the uncertainty of how to label data, and the inherent problems of SAR images. The noise of the SAR image, geometric corrections, and backscattering errors must be minimized before beginning the labeling process. This paper explains in detail how to segment SAR images by obtaining all kinds of data free of charge, and in order to prove the success of new dataset creation, three CNN models have been applied.

We compared our results with similar previous studies [43–45]. For land cover segmentation/classification with SAR images, refs [43–45] obtained results between 79–88% using machine-learning algorithms. Refs. [46,47] attained a nearly 93% accuracy. However, in ref. [46], optical images with 5 m spatial resolution were used, and [19] utilizes fully polarimetric images with better resolution than Sentinel-1 SAR resolution.

Another land cover mapping study [47] used SAR C-band Sentinel-1 (S-1) time-series data for classification with machine-learning models. The classes used in the paper were water, shrubs and scrubs, trees, bare soil, built-up areas and cropland, categories which are similar to ours. Overall accuracy obtained with the random forest classifier in [47] was 84%.

The prior study most similar to ours is [31]. In ref. [31], a new Finland dataset was created, the same classes as ours were used, and the accuracy was 92.78% with FC-DenseNet. Comparing our work with [31], we obtained nearly the same accuracy, with 92.23%. Our method provides extra information about how to label datasets in a pipeline.

In our proposed method, 6722 images were processed by deep learning models. However, use of a more extensive dataset for deep learning is crucial to achieving more success. Additionally, using datasets with more urban areas can increase accuracy. In order to increase the success of urban areas, [48] combined the use of Sentinel-2B and Sentinel-1A imagery. Fusing Sentinel-2 Multispectral Instrument (MSI) and Sentinel-1 SAR images, they reached an overall accuracy of 92.12%, with a support vector machine with composite kernels. Furthermore, including Sentinel-1A SAR data with Sentinel-2B MSI data improves the model's capability to differentiate between urban and water areas. Applying our pipeline over a combined dataset as proposed in [48] may be pursued as future work.

Reducing speckle noise is one of the most crucial steps in order to obtain more successful results. In ref. [49], a modified U-net with temporal attenuation encoder (U-TAE) was used as a semantic segmentation method. First, the temporal median filter was used to

reduce the noise in the images. Then, Sentinel-2, Sentinel-1, spectral indices, and ALOS elevation data were evaluated with various band combinations. As a result, with modified U-TAE, a mIoU of 57.25%; with random forest, a mIoU of 39.69%; with U-Net, a mIoU of 55.73%; and with SegFormer, an mIoU of 53.5% is reached.

We focused primarily on auto-labeled noise-free dataset creation, then, by using these enhanced noise-free datasets, attempted to improve accuracy in this field with state-of-the-art deep learning models. We applied three state-of-the-art CNN models to try our new dataset. Additionally, after applying the models, we evaluated the performance of the models and proved which model provided the best accuracy. Confirmation of the model with the highest success with the McNemar test increases the reliability of our proposal.

For future work, it would be interesting to develop deep learning models specially for SAR segmentation and it would also be interesting to achieve more success for urban areas segmentation; another type of CNN model or RNN model could also be examined. Additionally, our pipeline could be evaluated by using another dataset for ground-truth instead of CORINE.

6. Conclusions

Generally, for SAR image segmentation, there are two main troubles. First, there are no freely available and labeled datasets, and images are not convenient for use due to noise. Secondly, a great success was lacking, although a certain degree of success was achieved with deep learning methods. Most of the researchers pay no attention how to obtain a new dataset and use the datasets that are detailed in Section 2. Additionally, unsupervised image segmentation methods are preferred due to being simple and practical.

By this study, a pipeline was suggested with a dataset which had been created by us. We experimented with this dataset using state-of-the-art segmentation models for testing. We obtained a 92.23% overall PA, with HRNET as the best model according to performance. When results achieved were examined, we realized that models trained on our newly created dataset had not achieved success only in the segmentation of urban areas. When we examined the proportions of labeled classes in our dataset, we found that the percentage of urban areas in the dataset is very few compared to that of the others. The effectiveness of our proposed denoising pipeline for creating a dataset, one which takes care of noises caused by clouds and weather conditions, is apparently seen with examination of the results. As a result, the compared deep learning models achieved better accuracy results on a noise-free dataset setup. Additionally, we evaluated the segmentation performance of models with McNemar's test. This test proved that HRNet is the best model for our dataset.

References

1. Zhang, J.S.; He, C.Y.; Pan, Y.Z.; Li, J. The high spatial resolution RS image classification based on SVM method with the "ulti-source dat". *J. Remote Sens. Beijing* **2006**, *10*, 49.
2. Zhao, W.; Du, S.; Wang, Q.; Emery, W.J. Contextually guided very-high-resolution imagery classification with semantic segments. *ISPRS J. Photogramm. Remote. Sens.* **2017**, *132*, 48–60. [CrossRef]
3. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very high resolution multiangle urban classification analysi. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 1155–1170. [CrossRef]
4. Chan, J.C.W.; Bellens, R.; Canters, F.; Gautama, S. An assessment of geometric activity features for per-pixel classification of urban man-made objects using very high resolution Satellite imager. *Photogramm. Eng. Remote Sens.* **2009**, *75*, 397–411. [CrossRef]
5. Pacifici, F.; Chini, M.; Emery, W.J. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote. Sens. Environ.* **2009**, *113*, 1276–1292. [CrossRef]
6. Ding, Y.; Li, Y.; Yu, W. SAR Image Classification Based on CRFs With Integration of Local Label Context and Pairwise Label Compatibility. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2013**, *7*, 300–306. [CrossRef]
7. Mason, D.; Davenport, I.; Neal, J.; Schumann, G.; Bates, P.D. Nearreal-time flood detection in urban and rural areas using high resolution synthetic aperture radar images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3041–3052. [CrossRef]
8. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
9. González, A.; Pérez, R.; Romero-Zaliz, R. An Incremental Approach to Address Big Data Classification Problems Using Cognitive Models. *Cogn. Comput.* **2019**, *11*, 347–366. [CrossRef]
10. Padillo, F.; Luna, J.M.; Ventura, S. A Grammar-Guided Genetic Programing Algorithm for Associative Classification in Big Data. *Cogn. Comput.* **2019**, *11*, 331–346. [CrossRef]
11. Gou, S.; Zhuang, X.; Zhu, H.; Yu, T. Parallel Sparse Spectral Clustering for SAR Image Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2013**, *6*, 1949–1963. [CrossRef]
12. Lu, H.; Zhang, H.; Fan, H.; Liu, D.; Wang, J.; Wan, X.; Zhao, L.; Deng, Y.; Zhao, F.; Wang, R. Forest height retrieval using P-band airborne multi-baseline SAR data: A novel phase compensation metho. *ISPRS J. Photogramm. Remote. Sens.* **2021**, *175*, 99–118. [CrossRef]
13. Sun, Y.; Montazeri, S.; Wang, Y.; Zhu, X.X. Automatic registration of a single SAR image and GIS building footprints in a large-scale urban area. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *170*, 1–14. [CrossRef] [PubMed]
14. Corine Land Cover. Available online: https://land.copernicus.eu/pan-european/corine-land-cover (accessed on 30 September 2020).
15. Feng, J.; Cao, Z.; Pi, Y. Multiphase SAR Image Segmentaion with G0-statiscal-model-based active contours. *IEEE Trans. Geosci. Remote Sens* **2013**, *51*, 4190–4199. [CrossRef]
16. Jiao, L.; Gong, M.; Wang, S.; Hou, B.; Zheng, Z.; Wu, Q. Natural and Remote Sensing Image Segmentation Using Memetic Computing. *IEEE Comput. Intell. Mag.* **2010**, *5*, 78–91. [CrossRef]
17. Henry, C.; Azimi, S.M.; Merkle, N. Road Segmentation in SAR Satellite Images With Deep Fully Convolutional Neural Networks. *IEEE Geosci. Remote. Sens. Lett.* **2018**, *15*, 1867–1871. [CrossRef]
18. Wu, W.; Li, H.; Li, X.; Guo, H.; Zhang, L. PolSAR Image Semantic Segmentation Based on Deep Transfer Learning—Realizing Smooth Classification With Small Training Sets. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 977–981. [CrossRef]
19. Mohammadimanesh, F.; Salehi, B.; Mahdianpari, M.; Gill, E.; Molinier, M. A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *Int. Soc. Photogramm. Remote Sens.* **2019**, *151*, 223–236. [CrossRef]
20. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649. [CrossRef]
21. Ma, F.; Gao, F.; Sun, J.; Zhou, H.; Hussain, A. Attention Graph Convolution Network for Image Segmentation in Big SAR Imagery Data. *Remote. Sens.* **2019**, *11*, 2586. [CrossRef]
22. Ma, F.; Zhang, F.; Yin, Q.; Xiang, D.; Zhou, Y. Fast SAR Image Segmentation with Deep Task-Specific Superpixel Sampling and Soft Graph Convolution. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–16. [CrossRef]
23. Kwak, Y.; Song, W.-J.; Kim, S.-E. Speckle-Noise-Invariant Convolutional Neural Network for SAR Target Recognition. *IEEE Geosci. Remote. Sens. Lett.* **2018**, *16*, 549–553. [CrossRef]
24. Singh, P.; Shree, R. A new SAR image despeckling using directional smoothing filter and method noise thresholding. *Eng. Sci. Technol. Int. J.* **2018**, *21*, 589–610. [CrossRef]
25. Bostanci, B.; Bostanci, E. An Evaluation of Classification Algorithms Using Mc Nemar's Test. In Proceedings of the Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), India, 14–16 December 2012; Bansal, J.C., Singh, P.K., Deep, K., Pant, M., Nagar, A.K., Eds.; Springer: New Delhi, India, 2013; pp. 15–26, ISBN 978-81-322-1038-2.
26. Abdi, A.M. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience Remote. Sens.* **2019**, *57*, 37–48. [CrossRef]
27. Wang, X.; Cavigelli, L.; Eggimann, M.; Magno, M.; Benini, L. HR-SAR-Net: A Deep Neural Network for Urban Scene Segmentation from High-Resolution SAR Data. In Proceedings of the IEEE Sensors Applications Symposium, SAS, Kuala Lumpur, Malaysia, 9–11 March 2020.

28. Huang, L.; Liu, B.; Li, B.; Guo, W.; Yu, W.; Zhang, Z.; Yu, W. OpenSAR-Ship: A Dataset Dedicated to Sentinel-1 Ship Interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 195–208. [CrossRef]

29. Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; Lucas, R. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sens. Environ.* **2014**, *155*, 13–31. [CrossRef]

30. Wang, Y.; Zhu, X.X. The SARptical Dataset for Joint Analysis of SAR and Optical Image in Dense Urban Area. In Proceedings of the IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 6840–6843.

31. Scepanovic, S.; Antropov, O.; Laurila, P.; Rauste, Y.; Ignatenko, V.; Praks, J. Wide-Area Land Cover Mapping With Sentinel-1 Imagery Using Deep Learning Semantic Segmentation Models. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 10357–10374. [CrossRef]

32. Meng, Q.; Wen, X.; Yuan, L.; Xu, H. Factorization-Based Active Contour for Water-Land SAR Image Segmentation via the Fusion of Features. *IEEE Access* **2019**, *7*, 40347–40358. [CrossRef]

33. Shang, R.; Lin, J.; Jiao, L.; Li, Y. SAR Image Segmentation Using Region Smoothing and Label Correction. *Remote. Sens.* **2020**, *12*, 803. [CrossRef]

34. Natteshan, N.; Kumar, N.S. Effective SAR image segmentation and classification of crop areas using MRG and CDNN techniques. *Eur. J. Remote. Sens.* **2020**, *53*, 126–140. [CrossRef]

35. Shang, R.; Chen, C.; Wang, G.; Jiao, L.; Okoth, M.A.; Stolkin, R. A thumbnail-based hierarchical fuzzy clustering algorithm for SAR image segmentation. *Signal Process.* **2020**, *171*, 107518. [CrossRef]

36. Filipponi, F. Sentinel-1 GRD Preprocessing Workflow. *Proceedings* **2019**, *18*, 11. [CrossRef]

37. Soumekh, M. *Synthetic Aperture Radar Signal Processing with MATLAB Algorithms*; Wiley: New York, NY, USA, 1999.

38. European Space Agency (ESA) Copernicus Open Access Hub. Available online: https://scihub.copernicus.eu/dhus/#/home (accessed on 10 March 2021).

39. Contributors, M. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 12 October 2020).

40. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.

41. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.

42. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]

43. Laurin, G.V.; Liesenberg, V.; Chen, Q.; Guerriero, L.; Del Frate, F.; Bartolini, A.; Coomes, D.; Wilebore, B.; Lindsell, J.; Valentini, R. Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa. *Int. J. Appl. Earth Observ. Geoinf.* **2013**, *21*, 7–16.

44. Antropov, O.; Rauste, Y.; Lönnqvist, A.; Häme, T. PolSAR Mosaic Normalization for Improved Land-Cover Mapping. *IEEE Geosci. Remote. Sens. Lett.* **2012**, *9*, 1074–1078. [CrossRef]

45. Lonnqvist, A.; Rauste, Y.; Molinier, M.; Hame, T. Polarimetric SAR Data in Land Cover Mapping in Boreal Zone. *IEEE Trans. Geosci. Remote. Sens.* **2010**, *48*, 3652–3662. [CrossRef]

46. Mahdianpari, M.; Salehi, B.; Rezaee, M.; Mohammadimanesh, F.; Zhang, Y. Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1119. [CrossRef]

47. Dahhani, S.; Raji, M.; Hakdaoui, M.; Lhissou, R. Land Cover Mapping Using Sentinel-1 Time-Series Data and Machine-Learning Classifiers in Agricultural Sub-Saharan Landscape. *Remote. Sens.* **2022**, *15*, 65. [CrossRef]

48. Hu, B.; Xu, Y.; Huang, X.; Cheng, Q.; Ding, Q.; Bai, L.; Li, Y. Improving Urban Land Cover Classification with Combined Use of Sentinel-2 and Sentinel-1 Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 533. [CrossRef]

49. Tzepkenlis, A.; Marthoglou, K.; Grammalidis, N. Efficient Deep Semantic Segmentation for Land Cover Classification Using Sentinel Imagery. *Remote. Sens.* **2023**, *15*, 2027. [CrossRef]