

## RESEARCH ARTICLE

# Pseudo-Labeling Approach for Land Cover Classification Through Remote Sensing Observations With Noisy Labels

ISLOMBEK MIRPULATOV<sup>1</sup>, SVETLANA ILLARIONOVA<sup>1</sup>, DMITRII SHADRIN<sup>1,2</sup>,  
AND EVGENY BURNAEV<sup>1,3</sup>

<sup>1</sup>Skolkovo Institute of Science and Technology (Skoltech), 121205 Moscow, Russia

<sup>2</sup>Institute of Information Technology and Data Science, Irkutsk National Research Technical University, 664074 Irkutsk, Russia

<sup>3</sup>Autonomous Non-Profit Organization Artificial Intelligence Research Institute (AIRI), 105064 Moscow, Russia

Corresponding author: Svetlana Illarionova (s.illarionova@skoltech.ru)


This work was supported by the Analytical Center through the Russian Federation (RF) Government (subsidy agreement 000000D730321P5Q0002), in 2 November 2021, under Grant 70-2021-00145.

**ABSTRACT** Satellite data allows us to solve a wide range of challenging tasks remotely, including monitoring changing environmental conditions, assessing resources, and evaluating hazards. Computer vision algorithms such as convolutional neural networks have proven to be powerful tools for handling huge visual datasets. Although the number of satellite imagery is constantly growing and artificial intelligence is advancing, the present sticking point in remote sensing studies is the quality and amount of annotated datasets. Typically, manual labels have particular uncertainties and mismatches. Also, a lot of annotated datasets available in low resolution. Available visual representation of the observed objects can be more detailed than annotation. This causes the need for markup adjustment, which can be referred to as a pseudo-labeling task. The main contribution of this research is that we propose a pipeline for pseudo-labeling to address the problem of inaccurate and low-resolution markup improvement for solving land-cover and land-use segmentation task based on the data from the Sentinel-2 satellite. Our methodology takes advantages both of classical machine learning (ML) and deep learning (DL) algorithms. We examine random sampling, uniform sampling, and K-Means sampling and compare it with the full dataset usage. U-Net, DeepLab, and FPN models are trained on the adjusted dataset. The achieved findings show that a simple yet effective approach of data preliminary sampling and further markup refinement leads to significantly higher results than just using raw inaccurate data in a deep neural network pipeline. Moreover, the considered sampling technique allows to use less data for ML model training. The experiments involve markup adjustment and up-scaling from 30m to 10m. We verify the proposed approach in precise test area with manual annotation and show the improvement in F1-score from 0.792 to 0.816.

**INDEX TERMS** Artificial intelligence, artificial neural networks, computer vision, data analysis, pseudo-labeling, remote sensing, sampling.

## I. INTRODUCTION

Data acquisition using various satellite constellations becomes more available for a number of environmental studies. Computer vision algorithms enable fast and precise data analysis of ecosystems, forest areas, nature events, and human

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang .

impact [1]. Advanced algorithms reduce computational time and show remarkable results.

However, the current main limitation for neural network-based approaches developing is heavily related to high-quality annotation [2]. Manual markup creation is a time-consuming process. Moreover, for particular environmental tasks, field-based measurements are required [3]. Typically, annotation for remote sensing semantic segmentation

tasks is rather complex and contains mistakes and inaccuracy [4], [5]. Moreover, there can be gaps in a dataset with proper labels. To resolve this issue, one can consider a task of pseudo-labeling or weakly supervised learning. Pseudo-labeling is a popular technique used in the field of machine learning to improve and enlarge training datasets. The idea behind pseudo-labeling is to use the predictions made by an already trained model on unlabeled data to generate labels for that data [6]. These pseudo-labels can then be used to train a new model, which can in turn be used to make more precise predictions on the previously unlabeled data. Another powerful technique to deal with labeled data limitations is weakly supervised learning [7]. In contrast to traditional supervised learning, where models are trained on large amounts of precisely labeled data, weakly supervised learning uses incomplete or inexact labels to train models [8].

The same challenges of annotated data quality and availability arise in land use and land cover semantic segmentation tasks [9]. This is a crucial task for environmental monitoring, as changes in land cover can signify a range of local and global processes [10], [11]. By gaining an understanding of the natural and anthropogenic effects on land cover, society can promptly take measures to mitigate their impact on the environment. Therefore, we prioritize this task and aim to address the challenges of data annotation and availability to improve the accuracy of our analysis. Specifically, the motivation of the study is the following. Deep neural networks have proven to be a powerful technique for image processing. However, they require well-annotated large datasets. There are existing datasets with middle or low spatial resolution and inaccuracies in labels. We set a hypothesis that such datasets can be automatically adjusted and used for higher spatial resolution precise land cover and land use mapping.

In this study, we investigate and propose a pseudo-labeling pipeline to enhance the quality of semantic segmentation using weak annotations and deep neural networks. Our experiments utilize satellite data captured by the Sentinel-2 satellite, where the spatial resolution is set to 10m per pixel. The initial annotations have a lower spatial resolution of 30m per pixel. Researchers commonly rely on either upscaling the annotations or downsampling the satellite images to match the resolution to each other. However, these methods often lead to a significant reduction in the accuracy of recognition. Instead, our approach focuses on selecting more relevant training samples and annotation refinement to improve the neural network model's accuracy. Furthermore, we incorporate both machine learning (ML) and deep learning (DL) techniques in our pipeline, leveraging the robustness of Random Forest in handling noisy labels [12] and the ability of DL to extract essential spatial features [13]. The primary objectives of this study is:

- To investigate sampling approaches to reduce dataset size preserving recognition quality of ML models. The selected samples should accurately represent the study area;

- To propose and verify an approach for markup enhancement for satellite data using pseudo-labeling technique. For this task, we take an advantage of ML algorithm robustness to noise in labels;
- To develop a pipeline for land cover type classification using weak markup with noisy or outdated labels utilizing a CNN model. The pipeline will incorporate the previous two issues: sampling and pseudo-labeling.
- To show the benefits and possible applications of the proposed approach in the specific domain such as land use and land cover tasks.

The rest of the paper is organized as follows. In Section II, we discuss the state-of-the-art in the remote sensing and land cover classification. In Section III-A, we describe the datasets used in the experiments. The proposed approach for markup enhancement is presented in Sections III-C and III-B. Numerical results, method limitations and future avenues are presented and discussed in Section IV.

## II. RELATED WORK

### A. PSEUDO-LABELING AND WEAK-MARKUP

Pseudo-labeling techniques in the remote sensing domain strive to enlarge or enhance datasets. It assumes assigning labels to new samples beyond the train and test datasets. Pseudo-labeling can be also used to deal with weak markup, specifically to correct some labels in the initial markup. Manual markup in remote sensing tasks can be not precise enough or cover limited regions. Therefore, such techniques have high importance due to vast study areas and varying environmental conditions that make the annotation process more complex. There are different approaches for the pseudo-labeling task. In [14], pseudo-labeling assumes a clustering algorithm utilizing and is used for dimensionality reduction. Patch-based pseudo-labeling procedure for hyperspectral satellite images is proposed in [15]. In [16], the authors leverage strong spectral correlation between labeled and unlabeled samples in the hyperspectral images to extend artificially training dataset. The key idea behind the proposed approach is to select pixels with low information entropy based on sparse representation. The discussed method allows to increase classification models results. An adaptive method for satellite image classification is presented in [17]. This approach relies on utilizing unlabeled samples with high confidence from a classifier to expand an initial training dataset, with the selection process repeated several times. This leads to an increase in the accuracy of the final model. CNN-based approach to handle unlabeled data in a remote sensing task was also described in [18].

One can consider a pseudo-labeling approach to deal with weak markup or noise in labels. The task of dealing with outdated maps in a remote sensing task has been previously discussed in [12]. It is highlighted that a Random forest classifier is able to cope with a particular amount of noise in labels. The authors proposed an iterative scheme involving adapted Random forest classifiers to process both changed and unchanged spatial regions. In [19], the authors investigate

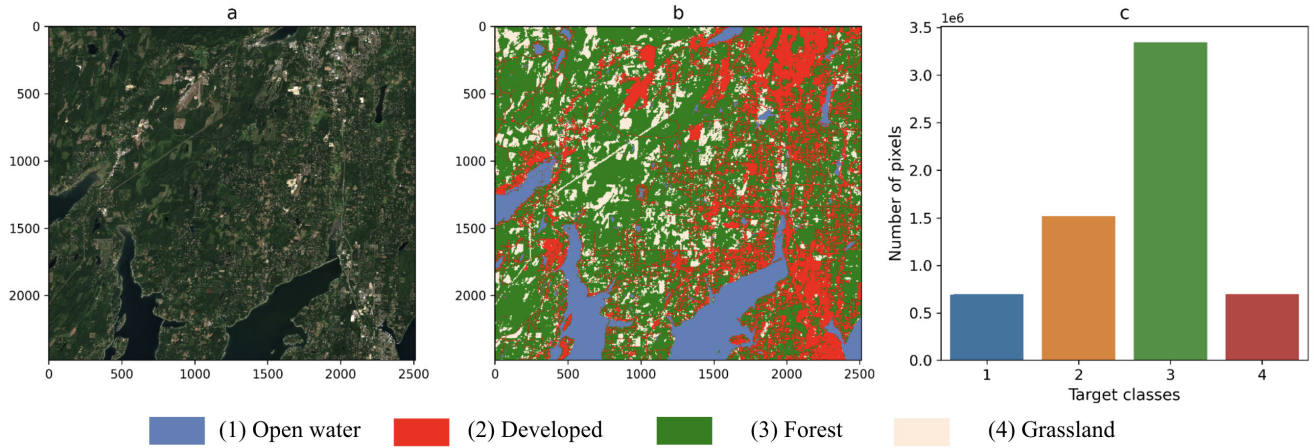


FIGURE 1. Sentinel-2 image (a), NLCD dataset with 30m per pixel (b), distribution of classes in image (c).

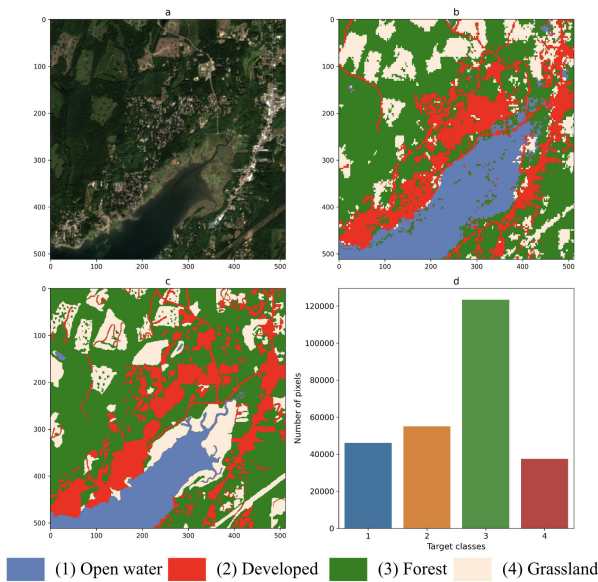


FIGURE 2. Sentinel-2 image (a), NLCD dataset with 30m per pixel (b), Manually Annotated 10m per pixel (c), distribution of classes in image (d).

the same problem of noisy labels. They constructed a spectral-spatial probability transform matrix (SSPTM) to assess the spectral similarity and spatial information. Some samples were randomly assigned as “clean” and propagated through the SSPTM. They repeated this process and finally defined likeliest labels for each sample. Another weak annotation challenge is described in [20]. To adjust forest species dataset, the authors utilized a custom loss function to take into account “clarity” of training samples. Then, an updated dataset was used to train a final neural network model.

**B. SAMPLING APPROACHES**

Sampling techniques in ML aim at selecting more relevant training data points. It allows one to facilitate fast and accurate ML algorithms training. In the remote sensing domain, this approach is widely used to determine more

TABLE 1. Description of sentinel-2 channels.

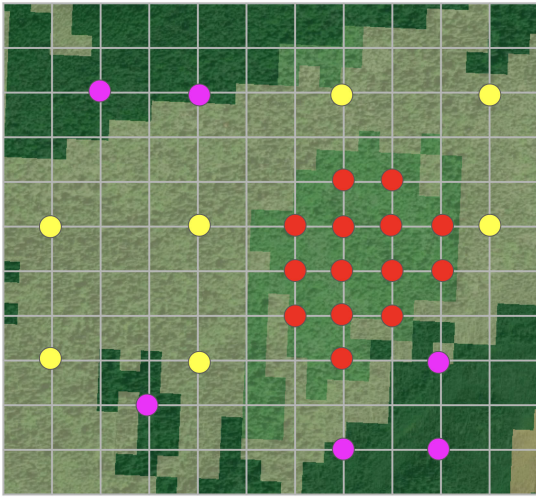
Sentinel-2 Bands	Central Wavelength ( $\mu\text{m}$ )	Resolution (m)
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.490	10
Band 3 - Green	0.560	10
Band 4 - Red	0.665	10
Band 5 - Vegetation Red Edge	0.705	20
Band 6 - Vegetation Red Edge	0.740	20
Band 7 - Vegetation Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Vegetation Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.610	20
Band 12 - SWIR	2.190	20

representative locations within large study areas [21]. In [22], the authors conduct a comprehensive analysis of different sample selection methods for a remote sensing task. They chose and considered four methods including simple random, proportional stratified random, disproportional stratified random, and deliberative sampling. It was found that stratified-statistical-based sampling methods are more valuable for further ML algorithm training. These approaches allow to reduce dataset size significantly. Moreover, it is important in case of imbalance classes, that can affect ML model performance [23].

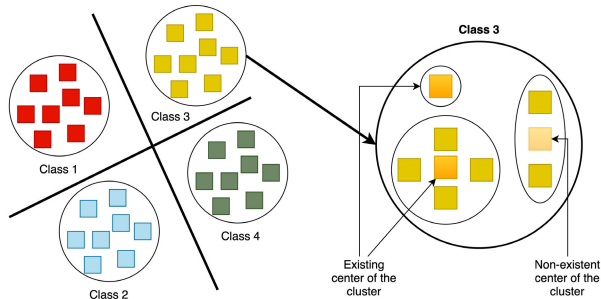
**C. LAND COVER AND LAND USE CLASSIFICATION**

Land cover type recognition plays a crucial role in the field of remote sensing. The specific tasks involved in this process can vary based on factors such as the scale of the target territory, the number of classes, and the spatial resolution. Low-spatial resolution satellite data is often used to create large-scale maps, such as those covering an entire country or the entire globe. In [24], a global database is presented for the mapping task based on MODIS data with spatial resolution of 500m per pixel. Medium-resolution satellite data support more detailed land cover analysis. A number of studies describe usage of the Landsat constellation data





**FIGURE 3.** Uniform sampling approach. Red points represent a smaller class, the selected samples for each class have higher density than representatives of a larger class (yellow points).



**FIGURE 4.** K-Means sampling approach. For each class, we identify cluster centers that correspond to the required quantity for the training dataset.

with spatial resolution of 30m per pixel [25]. For instance, Landsat-8 data was leveraged to create pan-European land cover and land use maps [26]. Sentinel constellation is also a widely-used source of medium-resolution data utilized for environmental studies. Wide spectral range, free-available access, and rapid revisit time make the data highly relevant for change detection in land use and land cover. This data has been used in a number of works aiming to classify land cover types [27], [28], [29].

On the other hand, high spatial resolution satellite imagery provides more detailed information on land cover characteristics. In [30], the authors conduct land cover classification using QuickBird satellite observations with the spatial resolution of 2.4m per pixel. The importance of precise land cover maps for carbon balance estimation is demonstrated on the Arctic tundra ecosystem. Another example of the high resolution maps is based on Gaofen satellite observations [31] and WorldView images [32]. The main disadvantages of such satellite data are its high cost, limited access, and typically a narrow wavelength range compared with middle resolution data.

The number of target classes is another important aspect of land cover classification. Some studies focus on a few general

**TABLE 2.** Results F1-score after pseudo-labeling.

Data	NLCD	Manually Annotated
<b>Full data</b>	<b>0.72</b>	<b>0.79</b>
Random data	0.71	0.75
Uniform	0.72	0.76
K-Means 30k data	0.70	0.74
K-Means 80k data	0.72	0.76
K-Means 200k data	0.72	0.77
K-Means 400k data	0.72	0.77
<b>K-Means 692k data</b>	<b>0.72</b>	<b>0.79</b>

classes, while others consider highly specialized classes such as forest species [33] or agriculture crop types [34].

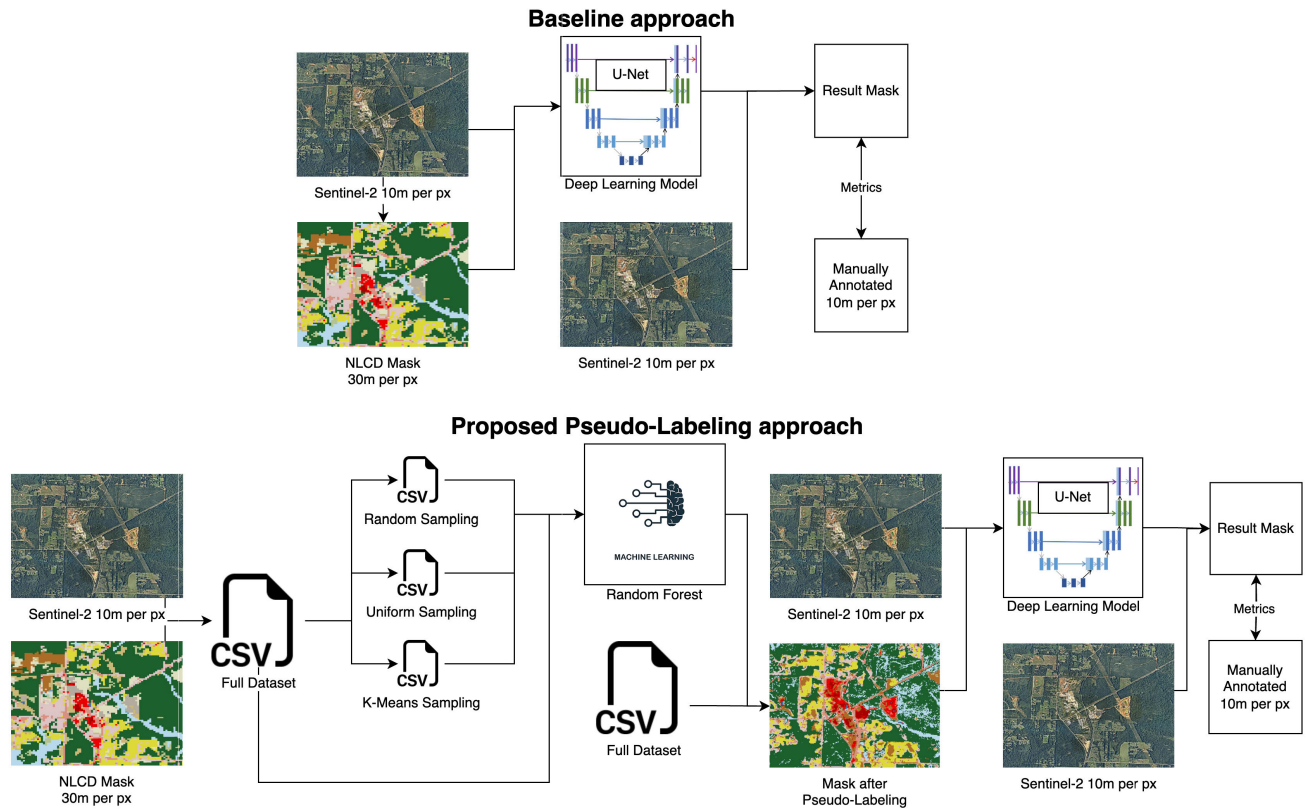
Depending on the tasks' characteristics, appropriate computer vision algorithms are selected. Classical ML algorithms have demonstrated robust results in complex environmental tasks, particularly of land cover and land use classification [26], [27], [29]. On the other hand, deep learning algorithms are capable of extracting and processing spatial information and applied in a number of studies. For instance, deep neural networks are considered in [28], [31], and [32]. In addition to semantic segmentation tasks, neural network approaches are used for land cover scene classification [35].

To achieve accurate results, both classical ML and DL algorithms require large and reliable datasets. There are several well-known datasets and web services available for land cover and land use classification tasks, including the following ones. Dynamic World is a near-real time global land cover dataset with 9 classes and the spatial resolution of 10m per pixel [28]. DeepGlobe Land Cover Classification dataset consists of RGB satellite images with very high spatial resolution of 50cm for 7 classes [36]. The Radiant MLHub platform provides datasets for various regions around the world with the spatial resolution of 10m [37]. The National Land Cover Database (NLCD) is a dataset collected in 2019 that includes segmentation mask for USA representing 8 classes and 20 sub-classes with the spatial resolution of 30m [38]. Satellite data providers also propose additional land cover classification products based on such satellites as Sentinel-2 or MODIS.

### III. DATA AND METHODS

#### A. SATELLITE DATA AND MARKUP

We use the National Land Cover Database (NLCD) 2019 [38] for segmentation masks. The dataset provides a comprehensive and consistent classification of land cover for the conterminous United States (CONUS). It includes 20 land cover classes, ranging from developed land and agriculture to wetlands and forests. Database was created using Landsat satellite imagery with a spatial resolution of 30m per pixel. Within our study we choose area that is located in USA with markup resolution of 30m per pixel and area  $62423 \text{ km}^2$ . We upsampled markup from 30 to 10m per pixel with nearest neighbor interpolation. Resulting mask size is  $2512 \times 2485$  pixels with classes distribution (Figure 1).



**FIGURE 5. Study workflow.** The initial approach uses original markup brought from 30m to 10m per pixel with nearest neighbor interpolation. The proposed approach applies sampling and pseudo-labeling techniques to adjust markup.

In our research we use four classes:

- Open water - areas of open water, generally with less than 25% cover of vegetation or soil.
- Developed - areas with a mixture of constructed materials and vegetation.
- Forest - areas dominated by trees generally greater than 5m tall, and greater than 20% of total vegetation cover.
- Grassland - areas dominated by grasses or cultivated vegetation, generally greater than 80% of total vegetation.

Instead of Landsat [39] data with the spatial resolution of 30m, we use Sentinel-2 data with higher spatial resolution. Sentinel-2 [40] is a multispectral satellite system that captures images in several bands of the electromagnetic spectrum. The satellite has 13 bands that cover a range of wavelengths, from visible light to near-infrared and shortwave infrared light with different spatial resolutions that interpolated to 10m per pixels (Table 1). These spectral bands can be combined to create a variety of composite images, such as false-color composites, which can enhance the contrast between different features on the Earth’s surface. In our experiments we use bands B01 - B08, B11, B12. The data is preprocessed with L2-Preprocessing [41]. L2-reprocessing is a set of steps taken to transform raw Sentinel-2 satellite data into a format that can be used for further analysis. This includes radiometric calibration, atmospheric correction, and geometric correction

to ensure accurate and consistent data for applications such as land use mapping and environmental monitoring. In this study, we use cloudless composite for the summer period of 2019. For additional visual assessment, we also considered summer composites for 2020 and 2021. The images for 2020 and 2021 were not used for models development.

We also expand the feature space by adding Vegetation Indices. NDVI (Normalized Difference Vegetation Index) and NDWI (Normalized Difference Water Index) are widely used vegetation indices that are derived from remote sensing data [42]. These indices are used to assess the presence and health of vegetation, and the presence of surface water, respectively.

By calculating NDVI and NDWI from satellite imagery, it is possible to obtain valuable information about the environment, which can be used for various applications such as land use mapping, environmental monitoring, and agricultural management. NDVI is especially useful for monitoring vegetation growth and health, which is important for crop management and monitoring of natural ecosystems. On the other hand, NDWI can help to detect and monitor changes in water bodies, such as lakes and rivers, which is useful for hydrological modeling, water resource management, and disaster response.

To validate the quality of land cover classification on more accurate annotated data, we choose sub-region

512 × 512 pixels covering the area of 26 km<sup>2</sup> and manually annotate it with CVAT-tool [43] based on 10m per pixel composite for 2019, as well as other sources of satellite data with higher spatial resolution. Figure 2 depicts the obtained markup and the distribution of the target classes.

## B. SAMPLING APPROACHES

Data sampling is the process of selecting a subset of information from a larger dataset for analysis and processing. The primary task of data sampling is to choose the most representative subset of data that will best reflect the structure and characteristics of the larger dataset.

In the context of our research, data sampling is necessary to ensure sample balance and to speed up the model training process by reducing the dataset size. We use a sample approach to create smaller yet representative dataset to train an ML algorithm. In contrast to a CNN-based approach, in a classical ML approaches, each sample is a pixel with the defined set of features including spectral satellite bands and vegetation indices.

For data sampling, we convert the original image and its corresponding mask into a CSV format. Each object in the file contains information about the pixel position on the image (x, y) and feature vectors B01 - B08, B11, B12, NDVI, NDWI with the target variable. The file contains a total of 6.2 million rows, which are normalized to a range of 0 to 1 for faster convergence.

The size of each class in the sampling process is chosen based on the size of the smallest class (Grassland), which comprises 692 thousand objects. As a result of the sampling process, we obtain a file with 2.7 million rows containing an equal number of training data for each class.

The main different between sampling in the general domain and in the remote sensing domain is a spatial distribution. For instance, forest in different regions varies by tree species, vegetation state, climate conditions. Therefore, to develop a robust algorithm, we should consider pixels from different areas, but we might exclude the repetitive entries of neighbor pixels from the same location. In this study, we consider a several sampling strategies to take into account spatial distribution of training samples.

### 1) RANDOM SAMPLING

Random sampling using the pandas sampling tool enables obtaining samples for each of the four classes from the training dataset. Selection of 692 thousand objects from each class provides a sufficiently large sample size to be used further for pseudo-labeling. After conducting this procedure, the overall sample size amounts to 2.7 million objects. Finally, we create a balanced dataset that is randomly distributed within the study area.

### 2) UNIFORM SAMPLING

The original image is divided into a grid, and within each class, 692 thousand samples are selected. Pixels from each

class are uniformly distributed within study area. The density of selected pixels for each class depends on the initial amount of observation for that class. For instance, for the smallest class, that is grassland, we take each pixel, while for a larger class, we select samples evenly distributed. Example of this sampling approach is shown in Figure 3.

### 3) SAMPLING USING CLUSTERING

The K-Means [44] algorithm is an ML algorithm used to cluster data into a predetermined number of clusters, k.

To select exemplary values, we divide the objects of each class into k clusters using this algorithm, where k in our case is 692,000 (the smallest class in the initial dataset). Then we select the centers of each cluster, thereby selecting the most diverse representatives within each class (Figure 4).

These exemplary objects can then be used to train a model representing the entire dataset.

In addition, k values of 30k, 80k, 200k, and 400k are also tested.

## C. PSEUDO-LABELING

Pseudo-labeling is an ML technique that is often used when only small amounts of labeled data are available, and the remaining data needs to be labeled. In our experiment, we use pseudo-labeling to improve the detailing in the existing labeling when processing satellite images for land use classification.

For the pseudo-labeling stage of the study workflow, we use datasets obtained during the sampling stage. The generated CSV files with samples corresponding to selected pixels are used to train an ML model. As features for each pixel, band values and vegetation indices from Sentinel-2 images are considered. It is known that the RF classifier is robust to noise in labels; therefore, it is utilized to reduce the effect of weak and low-resolution annotations. The ML model produces a new land cover map based on the selected pixels. It relies on the ability of the model to refine the initial labels from the markup with a 30m per pixel resolution based on a large statistical sample with more detailed satellite imagery. However, classical ML algorithms in such a setup cannot take into account the spatial patterns in data, and further improvements are required. Therefore, in the next step, a CNN model is trained on the pseudo-labeling maps produced by the ML model. The entire experimental workflow is shown in Figure 5.

Additionally, the ML algorithm provides us with a probability map for each pixel. The probabilities of those pixels that were misclassified are replaced with zero.

## D. EXPERIMENTAL SETUP

For the pseudo-labeling task, the RF [45] algorithm from the Scikit-Learn [46] library is chosen.

For land cover type segmentation, we use U-Net [47], FPN [48], and DeepLab [49] networks with a Resnet50 [50] backbone and input size of 128 × 128 pixels. These

TABLE 3. DeepLab model performance.

Data	Pseudo-Labeling	Probability Map	Precision	Recall	F1-score	IoU
NLCD	-	-	0.764±0.007	0.755±0.006	0.758±0.007	0.625±0.008
NLCD	-	+	0.760±0.004	0.755±0.005	0.757±0.005	0.622±0.006
NLCD	+	-	0.787±0.004	0.780±0.003	0.781±0.003	0.654±0.004
NLCD	+	+	<b>0.792±0.004</b>	<b>0.782±0.004</b>	<b>0.784±0.004</b>	<b>0.658±0.006</b>

TABLE 4. U-Net model performance.

Data	Pseudo-Labeling	Probability Map	Precision	Recall	F1-score	IoU
NLCD	-	-	0.799±0.005	0.785±0.004	0.792±0.005	0.675±0.007
NLCD	-	+	0.807±0.002	0.803±0.003	0.804±0.002	0.684±0.003
NLCD	+	-	0.821±0.003	0.812±0.001	0.815±0.002	0.701±0.002
NLCD	+	+	<b>0.823±0.004</b>	<b>0.813±0.001</b>	<b>0.816±0.001</b>	<b>0.702±0.002</b>

TABLE 5. Feature pyramid network model performance.

Data	Pseudo-Labeling	Probability Map	Precision	Recall	F1-score	IoU
NLCD	-	-	0.646±0.005	0.773±0.005	0.780±0.003	0.773±0.005
NLCD	-	+	0.658±0.003	0.784±0.002	0.788±0.001	0.785±0.002
NLCD	+	-	0.667±0.004	0.792±0.004	0.795±0.001	0.788±0.001
NLCD	+	+	<b>0.667±0.001</b>	<b>0.792±0.003</b>	<b>0.796±0.003</b>	<b>0.789±0.001</b>

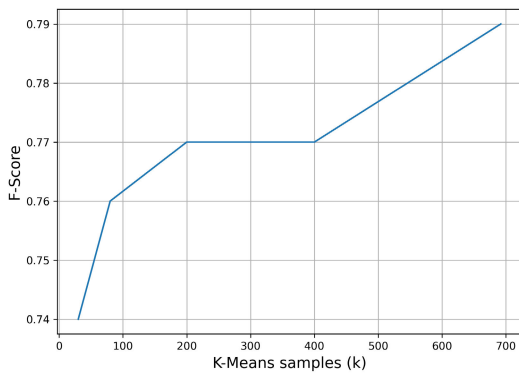


FIGURE 6. F1-score for different dataset size.

architectures are a common choice in the remote sensing domain and have shown high results in various environmental tasks [51], [52].

The training is performed using the Adam optimizer [50] with a learning rate of  $10^{-3}$  and the DiceLoss [53] loss function for 40 epochs.

In addition, we employ two types of augmentations, HorizontalFlip and VerticalFlip with a probability of 0.5, utilizing the Albumentations library [54].

The network implementation is performed on the Pytorch framework [55].

All experiments are conducted on the Zhores supercomputer [56]. To perform subsampling using clustering, the K-Means algorithm implemented in the CUMML [57] package is used.

### E. EVALUATION METRICS

We use F1-Score and IoU (Intersection over Union) metrics to evaluate the quality of models.

Precision and Recall are commonly used metrics to evaluate the quality of ML models. Precision is defined as the ratio of true positive results to the total number of positive results predicted by the model, while Recall is defined as the ratio of true positive results to the total number of actual positive results. The formulas are the following:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where  $TP$  is the number of true positive results,  $FP$  is the number of false positive results, and  $FN$  is the number of false negative results.

The F1-score (F-measure) is the harmonic mean between Precision and Recall, and is a measure of the overall effectiveness of the model. It is defined as the weighted harmonic mean between precision and recall:

$$F1 - score = \frac{2Precision Recall}{Precision + Recall} \tag{3}$$

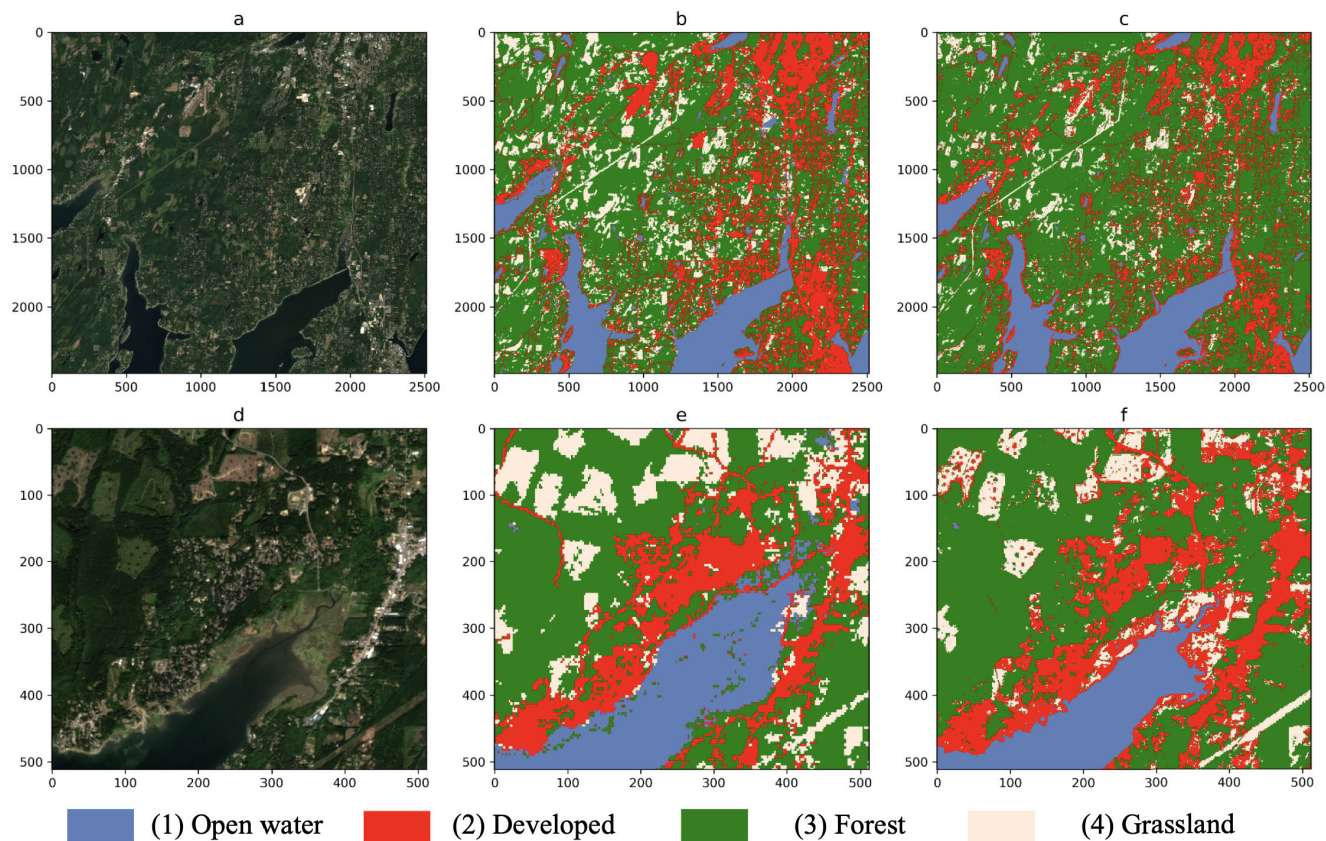
where  $Precision$  is precision, and  $Recall$  is recall.

The F1-score has the property of taking into account both metrics and is more robust to imbalanced classes than simple precision or recall. A high F1-score indicates that the model performs well and is capable of correct classifying.

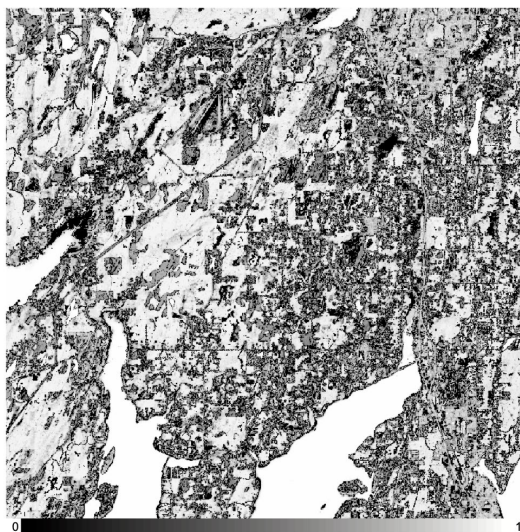
The IoU metric, also known as Jaccard Index, is one of the most common metrics for evaluating the quality of image segmentation. It is calculated as the ratio of the area of intersection between the ground truth and predicted masks to the area of their union. Formally, it is expressed as:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{4}$$





**FIGURE 7.** Sentinel-2 image (a) , NLCD dataset with 30m per pixel (b), new markup after Pseudo-Labeling with K-Means 692k (c), Sentinel-2 image tile with L2-Preprocessing (d), NLCD dataset with 30m per pixel (e), new markup after Pseudo-Labeling with K-Means 692k (f).



**FIGURE 8.** Probability map after pseudo-labeling.

where *Area of Overlap* is the intersection area between the ground truth and predicted masks, and *Area of Union* is the union area of the masks.

The IoU value ranges from 0 to 1, where 0 indicates no overlap between the masks, and 1 indicates a perfect

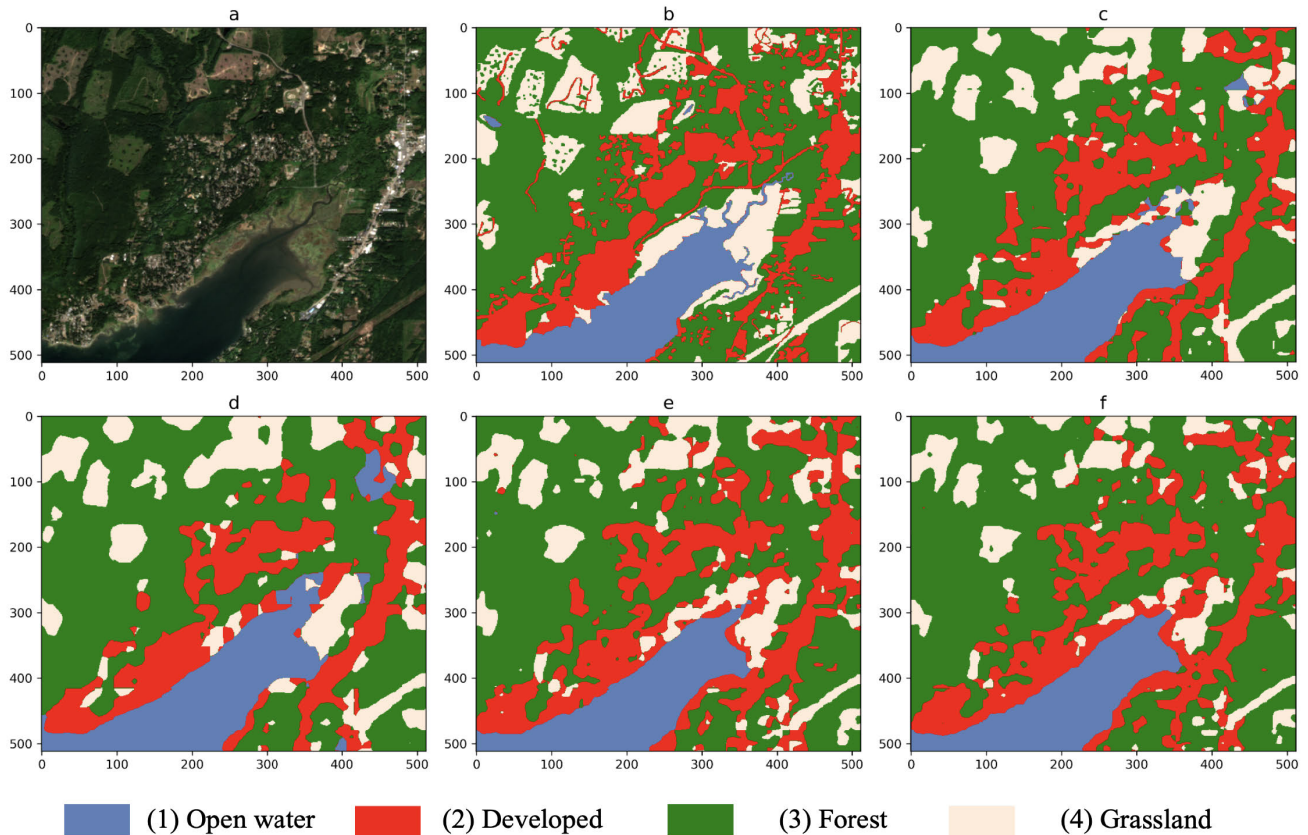
match. The higher the IoU value, the better the quality of segmentation.

#### IV. RESULTS AND DISCUSSION

##### A. NUMERICAL AND VISUAL RESULTS

To adjust the markup, we use different sampling techniques and RF classifier. We use manually annotated data to evaluate the achieved results for different sampling strategies (Table 2). We also show F1-score computed on the initial markup, called NCLD, adjusted from 30m per pixel to 10m per pixel. The first experiment involves usage of the entire dataset without training pixels selecting. The obtained F1-score equals to 0.79, while the total amount of training samples is 6.2 millions. To reduce the time of training process, we consider a sampling technique where random training points are selected and formed new training dataset [22]. Although the dataset size decreases to 2.7 million samples, the F1-score is also decreased from 0.79 to 0.75 for manually annotated dataset. The uniform sampling assumes selecting training pixels that are distributed evenly within each class [58]. It is supposed to cover various territories and reduce amount of training dataset selecting just a part of pixels for each area. The training dataset size is the same as for the random sampling, 2.7 million samples. However, it allows just slightly improve the classification results comparing with





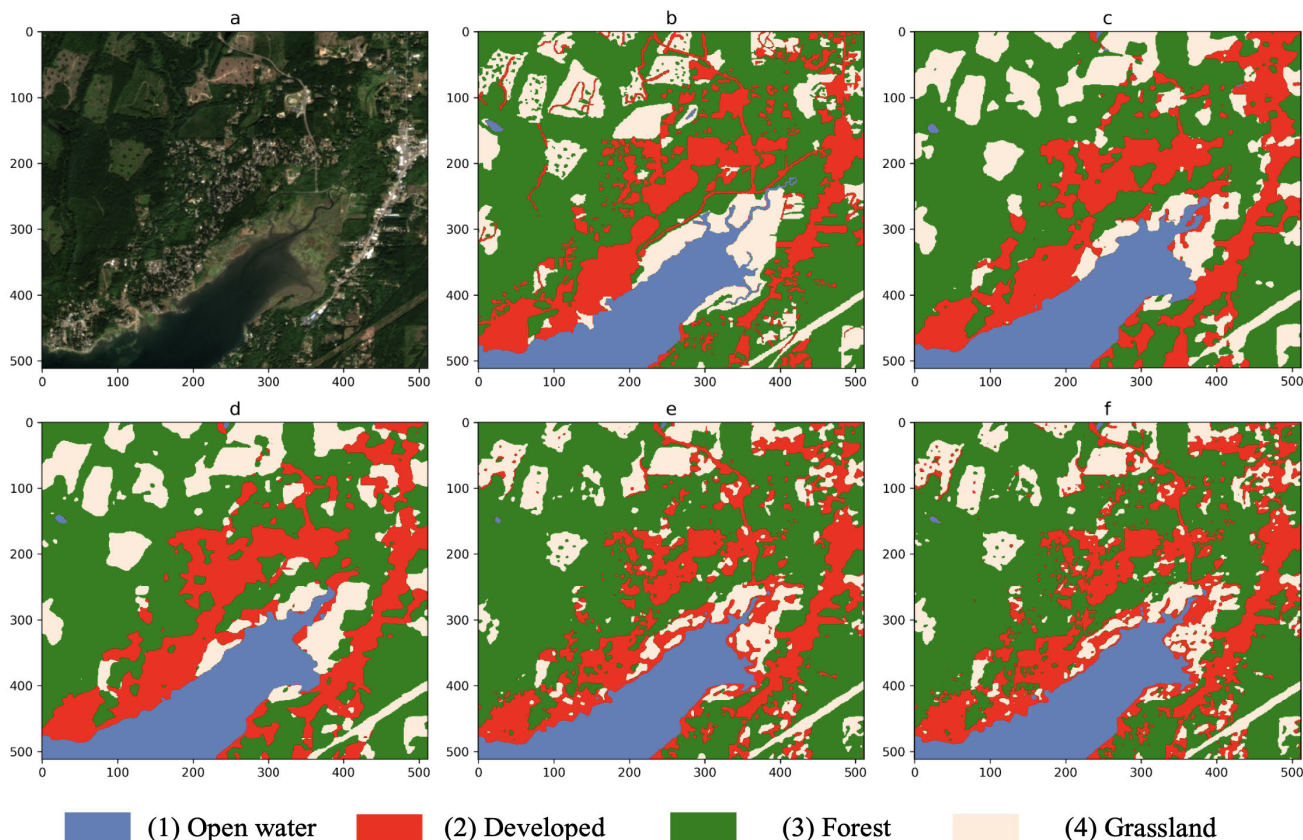
**FIGURE 9.** DeepLab model performance. Sentinel-2 image (a), Manually Annotated 10m per pixel (b), results after training on NLCD with 30m per pixel (c), results after training on NLCD with 30m per pixel with probability map (d), results after training on mask after Pseudo-Labeling (e), results after training on mask after Pseudo-Labeling with probability map (f).

the random sampling. F1-score equals to 0.76 for the uniform sampling. Other experiments involve the clustering strategy to create a reduced dataset. This approach assumes selecting more relevant training samples. Figure 6 shows results for different numbers of clusters that represent number of samples in each class. The ultimate dataset are balanced with the same number of samples for each class. This experiment depicts that we can significantly reduce the training dataset size from 6.2 to 2.7 millions training samples and preserve the classification quality equal to F1-score of 0.79. We create the final updated markup using dataset “K-Means 692k data”. The obtained map is presented in Figure 7. The proposed approach allow accelerate time of ML algorithm training. We also compute the probability map of the RF (Figure 8). It shows confidence of the model that each pixel has a particular class. This map can be used as an additional feature.

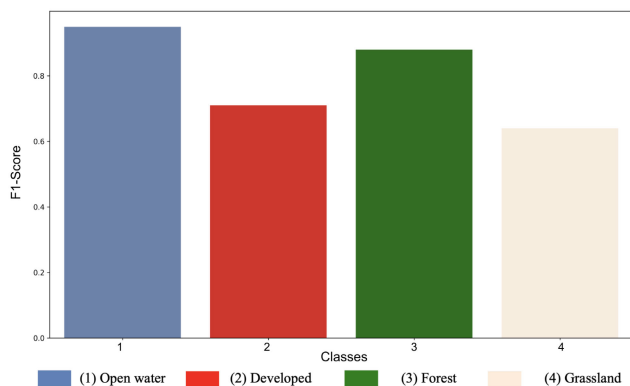
Next, we evaluate the performance of DeepLab and U-Net models on four types of data (Table 3 and 4). The first variant is obtained by training the model using markup with a spatial resolution of 30m per pixel adjusted to 10m per pixel by an interpolation. We achieve F1-score for DeepLab and U-Net equal to 0.758 and 0.792, respectively. The IoU equals to 0.625 and 0.675 for these two models. The next

experiment involves usage of additional input channel that accompanies multispectral data. This channel is a probability map representing the confidence of ML model in assigning particular class to each pixel (Figure 8). However, this probability map obtained after the pseudo-labeling does not lead to model performance increase. In the third variant, we train our model on data obtained through pseudo-labeling with K-Means sampling of 692k and we also test the addition of a probability map. The results are improved both for DeepLab and U-Net models. For DeepLab, IoU raises from 0.625 to 0.658, while for U-Net, it is increased from 0.675 to 0.702. Prediction results for DeepLab and U-Net are shown in Figures 9 and 10. The final results for each class are presented in Figure 11. The best segmentation quality is achieved for open water (F1-score is 0.95). F1-score for forested areas equals to 0.88. Settlements and grassland are identified with F1-score of 0.71 and 0.64, respectively.

In addition to these two models, an experiment is conducted with the FPN model. Although the overall performance of this model in the context of the given task falls behind the aforementioned models, our pseudo-labeling method yields improvements in terms of IOU, increasing from 0.646 to 0.667, and F-score, increasing from 0.773 to 0.789 (Table 5).



**FIGURE 10.** U-Net model performance. Sentinel-2 image (a) , Manually Annotated 10m per pixel (b), results after training on NLCD with 30m per pixel (c), results after training on NLCD with 30m per pixel with probability map (d), results after training on mask after Pseudo-Labeling (e), results after training on mask after Pseudo-Labeling with probability map (f).



**FIGURE 11.** F1-score results for the U-Net model on the test site with manual annotation.

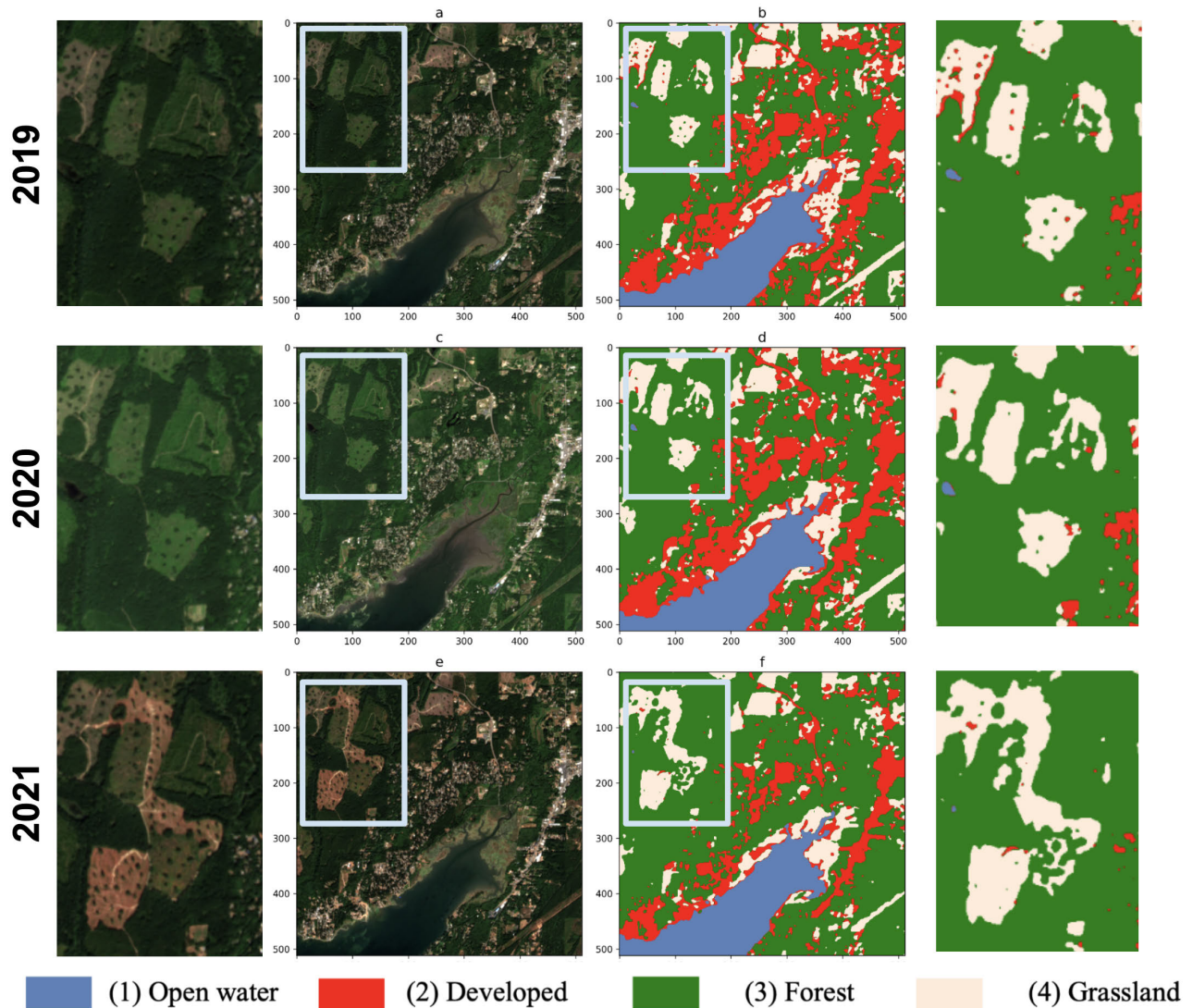
Furthermore, we examine the applicability of the best model trained on the pseudo-labeled dataset on Sentinel-2 composites obtained for the same region in 2020 and 2021 (Figures 12). Visual assessment of predictions for the years 2020 and 2021 supports the possibility of further model usage for land cover and land use change monitoring over time. For instance, it can be applied to estimate the absolute and percentage changes between particular land cover classes. The model successfully recognizes areas where logging was conducted in 2021 and areas with growing trees.

Land cover and land use markup often have noise and inaccuracy in labels. Figure 13 depicts a case with an inaccuracy in annotation in the initial dataset. The proposed approach allows us to create more detailed markup with precise labels for the target classes. For instance, areas with buildings are distinguished better from the forested areas. Although, the markup can be further improved, the conducted study shows the importance of processing of weak annotation and proposes a promising way how deep learning and classical ML can be combined to tackle the issue.

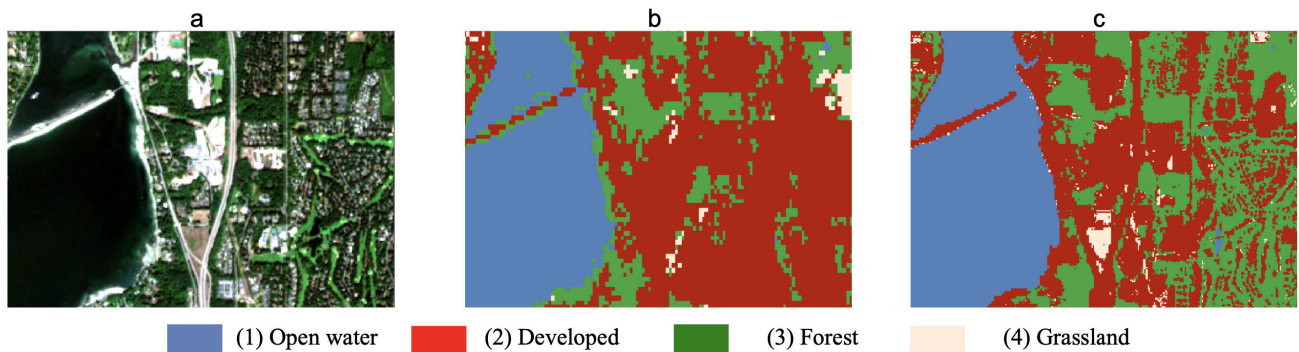
**B. COMPARISON WITH EXISTING APPROACHES**

Dealing with noisy labels is a crucial task in the remote sensing domain and different studies have been already conducted on this topic. They differ by the integrated methods and specificity of the solved problems. In [12], the authors utilized an adaptive RF classifier to process an outdated dataset. The approach has shown significant results for cases with changes in land cover types through the time. In our study, we combine both a classical ML approach to adjust the dataset and a DL approach to train a final model. It allows us to take advantage of RF robustness to noise and ability of CNN to process spatial data. Pseudo-labeling approaches for non annotated samples are also highly relevant for remote sensing. In [17], to expand the training dataset, a maximum





**FIGURE 12.** Sentinel-2 image for 2019 (a) , model prediction for 2019 (b), Sentinel-2 image for 2020 (c), model prediction for 2020 (d), Sentinel-2 image for 2021 (e), model prediction for 2021 (f).



**FIGURE 13.** Sentinel-2 image (a) , NLCD dataset with 30m per pixel (b), new markup after Pseudo-Labeling with K-Means 692k (c).

likelihood classifier is implemented. Total accuracy raises from 0.76 to 0.89 compared with the conventional approach based on maximum likelihood classifier.

Previous works have proposed large global datasets with different properties. Although we do not provide new datasets

with precise land cover manual annotations, we propose an effective approach, how one can enhance existing datasets. It is vital, because even perfect annotation becomes less useful when they are utilized with new satellite constellations or observations for other periods. Model zoo with pre-trained



encoders are also known as a powerful tool for accurate environmental tasks and land cover classification. However, for fine-tuning demands for training data are more strict. Therefore, it is promising to combine such approaches with datasets enhancement as described in our work.

### C. LIMITATIONS OF THE PROPOSED APPROACH

There are two limitations of the proposed pipeline. Firstly, if a class is too small, the proposed pipeline may struggle with data imbalance and could potentially neglect rare classes. However, investigating this issue is beyond the scope of the current study. Secondly, computational cost is a limitation when estimating clustering centers. However, this computation is conducted only once, reducing the dataset size, and subsequent experiments for land cover classification are conducted faster due to smaller samples. Therefore, this limitation is not crucial for the study in general.

### D. FUTURE PERSPECTIVES

Among the promising avenues for future work, we can distinguish the following directions. In this study, we focus mainly on the markup enhancement, while this approach can be further combined with different advanced techniques to boost the ultimate results. For instance, multispectral [59] or object-based [60] augmentation approaches can be applied to the satellite images to extend the training dataset. Another area for refinement is to add more classes of land cover and land use. The proposed approach can be applied to various target classes according to the practical task definition. It might be highly relevant for some specific vegetation classes that are rarely available in open access datasets with proper spatial resolution [61]. For particular regions, this problem is more tangible. Besides land cover and land use classes, weak markup occurs also in other remote sensing tasks such as infrastructure object recognition [62]. The discrepancy between satellite images and object annotations introduces noise into the data. Searching for relevant pixels and conducting automatic markup adjustment can benefit such types of remote sensing tasks.

In this study, we consider markup enhancement to meet the satellite data resolution of 10m per pixel. High-resolution images are becoming more available and can be accompanied by previously created low-resolution markup. It is reasonable to study markup improvement for higher spatial resolutions, particularly for other datasets.

The primary objective of this study is to propose an effective approach for adjusting remote sensing markup. The study involves conducting experiments with U-Net, DeepLab, and FPN architectures to assess the improvements in land cover markup. Currently, advanced models such as transformers and diffusion-based models have demonstrated high levels of accuracy and performance in both the general domain and the remote sensing domain. Therefore, exploring the potential application of such advanced

architectures to enhance segmentation quality could be beneficial.

### V. CONCLUSION

In this study, we address the issue of weak markup in land cover classification tasks using satellite data and deep neural networks. Due to vast territories and diverse environmental conditions, obtaining high-quality precise data annotation can be challenging. While globe-maps may be available, more accurate labels are needed to produce detailed and accurate results for deep neural network models. To overcome this challenge, we propose an efficient pipeline that uses weak annotations with lower spatial resolution to create new markup. This approach combines pseudo-labeling and sampling techniques and includes classical ML algorithms with advanced deep neural network ones. We demonstrate our approach using Sentinel-2 imagery with a spatial resolution of 10m and an initial markup with a spatial resolution of 30m. Our approach includes two stages. The first stage involves selecting more relevant samples to reduce the size of the training dataset. The second stage focuses on training markup updates. The resulting dataset with updated semantic segmentation labels is then used to train a deep neural network. To validate our approach, we manually annotated a test image with a spatial resolution of 10m. Our results show a significant improvement in the recognition task on the test set. The F1-score increases to 0.816 compared to the F1-score of 0.792 when using the initial markup. This approach has the potential to benefit future environmental studies and change detection tasks of land cover and land use on a large scale.

### ACKNOWLEDGMENT

The authors would like to acknowledge the use of the Skoltech CDISE supercomputer Zhores [56] for obtaining the results presented in this article.

*(Islombek Mirpulatov and Svetlana Illarionova contributed equally to this work.)*

### REFERENCES

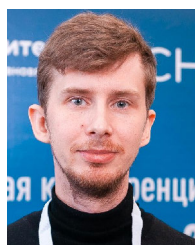
- [1] E. V. Burnaev, A. V. Bernstein, V. V. Vanovskiy, A. A. Zaytsev, A. M. Bulkin, V. Y. Ignatiev, D. G. Shadrin, S. V. Illarionova, I. V. Oseledets, A. Y. Mikhalev, A. A. Osipov, A. A. Artemov, M. G. Sharaev, and I. E. Trofimov, "Fundamental research and developments in the field of applied artificial intelligence," *Doklady Math.*, vol. 106, no. S1, pp. S14–S22, Dec. 2022.
- [2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.
- [3] S. Ecke, J. Dempewolf, J. Frey, A. Schwaller, E. Endres, H.-J. Klemmt, D. Tiede, and T. Seifert, "UAV-based forest health monitoring: A systematic review," *Remote Sens.*, vol. 14, no. 13, p. 3205, Jul. 2022.
- [4] C. Fasana, S. Pasini, F. Milani, and P. Fraternali, "Weakly supervised object detection for remote sensing images: A survey," *Remote Sens.*, vol. 14, no. 21, p. 5362, Oct. 2022.
- [5] S. Illarionova, D. Shadrin, P. Tregubova, V. Ignatiev, A. Efimov, I. Oseledets, and E. Burnaev, "A survey of computer vision techniques for forest characterization and carbon monitoring tasks," *Remote Sens.*, vol. 14, no. 22, p. 5861, Nov. 2022.
- [6] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Odonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 8, pp. 6912–6920.

- [7] S. Mukhamadiev, S. Nesteruk, S. Illarionova, and A. Somov, "Enabling multi-part plant segmentation with instance-level augmentation using weak annotations," *Information*, vol. 14, no. 7, p. 380, Jul. 2023. [Online]. Available: <https://www.mdpi.com/2078-2489/14/7/380>
- [8] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.
- [9] A. Vali, S. Comai, and M. Matteucci, "Deep learning for land use and land cover classification based on hyperspectral and multispectral Earth observation data: A review," *Remote Sens.*, vol. 12, no. 15, p. 2495, Aug. 2020.
- [10] A. H. Chughtai, H. Abbasi, and I. R. Karas, "A review on change detection method and accuracy assessment for land use land cover," *Remote Sens. Appl., Soc. Environ.*, vol. 22, Apr. 2021, Art. no. 100482.
- [11] S. Illarionova, D. Shadrin, V. Ignatiev, S. Shayakhmetov, A. Trekin, and I. Oseledets, "Augmentation-based methodology for enhancement of trees map detailization on a large scale," *Remote Sens.*, vol. 14, no. 9, p. 2281, May 2022.
- [12] A. E. Maas, F. Rottensteiner, and C. Heipke, "A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training," *Comput. Vis. Image Understand.*, vol. 188, Nov. 2019, Art. no. 102782.
- [13] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.
- [14] H. Wu and S. Prasad, "Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels," *Pattern Recognit.*, vol. 74, pp. 212–224, Feb. 2018.
- [15] A. Samat, E. Li, P. Du, S. Liu, and Z. Miao, "Improving deep forest via patch-based pooling, morphological profiling, and pseudo labeling for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9334–9349, 2021.
- [16] B. Cui, J. Cui, Y. Lu, N. Guo, and M. Gong, "A sparse representation-based sample pseudo-labeling method for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 4, p. 664, Feb. 2020.
- [17] M. Imani and H. Ghassemani, "Adaptive expansion of training samples for improving hyperspectral image classification performance," in *Proc. 21st Iranian Conf. Electr. Eng. (ICEE)*, May 2013, pp. 1–6.
- [18] W. Lin, J. Ma, X. Tang, X. Zhang, and L. Jiao, "RSMatch: Semi-supervised learning with adaptive category-related pseudo labeling for remote sensing scene classification," in *Proc. Int. Conf. Intell. Sci. (ICIS)*. Xi'an, China: Springer, Oct. 2022, pp. 220–227.
- [19] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, Feb. 2019.
- [20] S. Illarionova, A. Trekin, V. Ignatiev, and I. Oseledets, "Tree species mapping on Sentinel-2 satellite imagery with weakly supervised classification and object-wise sampling," *Forests*, vol. 12, no. 10, p. 1413, Oct. 2021.
- [21] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Boca Raton, FL, USA: CRC Press, 2019.
- [22] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sens.*, vol. 11, no. 2, p. 185, Jan. 2019.
- [23] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, May 2018.
- [24] R. Khaldi, D. Alcaraz-Segura, E. Guirado, Y. Benhammou, A. El Afia, F. Herrera, and S. Tabik, "TimeSpec4LULC: A global multispectral time series database for training LULC mapping models with machine learning," *Earth Syst. Sci. Data*, vol. 14, no. 3, pp. 1377–1411, Mar. 2022.
- [25] D. Phiri and J. Morgenroth, "Developments in Landsat land cover classification methods: A review," *Remote Sens.*, vol. 9, no. 9, p. 967, Sep. 2017.
- [26] D. Pflugmacher, A. Rabe, M. Peters, and P. Hostert, "Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey," *Remote Sens. Environ.*, vol. 221, pp. 583–595, Feb. 2019.
- [27] M. G. Sunde, D. D. Diamond, L. F. Elliott, P. Hanberry, and D. True, "Mapping high-resolution percentage canopy cover using a multi-sensor approach," *Remote Sens. Environ.*, vol. 242, Jun. 2020, Art. no. 111748.
- [28] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwehr, M. Weiss, C. Hanson, O. Guinan, R. Moore, and A. M. Tait, "Dynamic World, near real-time global 10 m land use land cover mapping," *Sci. Data*, vol. 9, no. 1, p. 251, Jun. 2022.
- [29] V. Syrris, P. Hasenohr, B. Delipetrev, A. Kotsev, P. Kempeneers, and P. Soille, "Evaluation of the potential of convolutional neural networks and random forests for multi-class segmentation of Sentinel-2 imagery," *Remote Sens.*, vol. 11, no. 8, p. 907, Apr. 2019.
- [30] C. C. Treat, M. E. Marushchak, C. Voigt, Y. Zhang, Z. Tan, Q. Zhuang, T. A. Virtanen, A. Räsänen, C. Biasi, G. Hugelius, D. Kaverin, P. A. Miller, M. Stendel, V. Romanovsky, F. Rivkin, P. J. Martikainen, and N. J. Shurpali, "Tundra landscape heterogeneity, not interannual variability, controls the decadal regional carbon balance in the Western Russian Arctic," *Global Change Biol.*, vol. 24, no. 11, pp. 5188–5204, Nov. 2018.
- [31] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.
- [32] J. Arndt and D. Lunga, "Large-scale classification of urban structural units from remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2634–2648, 2021.
- [33] S. Illarionova, A. Trekin, V. Ignatiev, and I. Oseledets, "Neural-based hierarchical approach for detailed dominant forest species classification by multispectral satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1810–1820, 2021.
- [34] V. Mazzia, A. Khaliq, and M. Chiaberge, "Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN)," *Appl. Sci.*, vol. 10, no. 1, p. 238, Dec. 2019.
- [35] I. Papoutsis, N. I. Bountos, A. Zavras, D. Michail, and C. Tryfonopoulos, "Benchmarking and scaling of deep learning models for land cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 250–268, Jan. 2023.
- [36] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.
- [37] MLHub. (2023). *Radiant MLHub*. [Online]. Available: <https://mlhub.earth/>
- [38] J. Wickham, S. V. Stehman, D. G. Sorenson, L. Gass, and J. A. Dewitz, "Thematic accuracy assessment of the NLCD 2016 land cover for the conterminous United States," *Remote Sens. Environ.*, vol. 257, May 2021, Art. no. 112357.
- [39] D. P. Roy et al., "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, Apr. 2014.
- [40] D. Phiri, M. Simwanda, S. Salekin, V. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sens.*, vol. 12, no. 14, p. 2291, Jul. 2020.
- [41] M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, "Sen2Cor for Sentinel-2," *Proc. SPIE*, vol. 10427, Oct. 2017, Art. no. 1042704.
- [42] Y. Zeng, D. Hao, A. Huete, B. Dechant, J. Berry, J. M. Chen, J. Joiner, C. Frankenberg, B. Bond-Lamberty, Y. Ryu, J. Xiao, G. R. Asrar, and M. Chen, "Optical vegetation indices for monitoring terrestrial ecosystems globally," *Nature Rev. Earth Environ.*, vol. 3, no. 7, pp. 477–493, May 2022.
- [43] B. Sekachev et al. (Aug. 2020). *opencv/cvat: v1.1.0*. [Online]. Available: <https://zenodo.org/record/4009388>
- [44] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, Jan. 1979.
- [45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
- [47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. Cham, Switzerland: Springer*, 2015, pp. 234–241.
- [48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

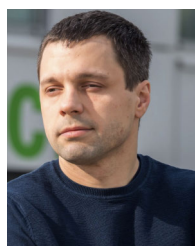
- [49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [51] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.
- [52] S. Illarionova, D. Shadrin, V. Ignatiev, S. Shayakhmetov, A. Trekin, and I. Oseledets, "Estimation of the canopy height model from multispectral satellite imagery with convolutional neural networks," *IEEE Access*, vol. 10, pp. 34116–34132, 2022.
- [53] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. 3rd Int. Workshop Deep Learn. Med. Image Anal. (DLMA)*. Québec City, QC, Canada: Springer, Sep. 2017, pp. 240–248.
- [54] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020. [Online]. Available: <https://www.mdpi.com/2078-2489/11/2/125>
- [55] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [56] I. Zacharov, R. Arslanov, M. Gunin, D. Stefonishin, A. Bykov, S. Pavlov, O. Panarin, A. Maliutin, S. Rykovanov, and M. Fedorov, "'Zhores'—Petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in Skolkovo Institute of Science and Technology," *Open Eng.*, vol. 9, no. 1, pp. 512–520, 2019.
- [57] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, Apr. 2020.
- [58] X. Zhang, L. Liu, C. Wu, X. Chen, Y. Gao, S. Xie, and B. Zhang, "Development of a global 30 m impervious surface map using multisource and multitemporal remote sensing datasets with the Google Earth Engine platform," *Earth Syst. Sci. Data*, vol. 12, no. 3, pp. 1625–1648, Jul. 2020.
- [59] M. Fawakherji, C. Potena, A. Pretto, D. D. Bloisi, and D. Nardi, "Multi-spectral image synthesis for crop/weed segmentation in precision farming," *Robot. Auto. Syst.*, vol. 146, Dec. 2021, Art. no. 103861.
- [60] S. Illarionova, S. Nesteruk, D. Shadrin, V. Ignatiev, M. Pukalchik, and I. Oseledets, "Object-based augmentation for building semantic segmentation: Ventura and Santa Rosa case study," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1659–1668.
- [61] L. He, W. Chen, S. G. Leblanc, J. Lovitt, A. Arsenault, I. Schmelzer, R. H. Fraser, R. Latifovic, L. Sun, C. Prévost, H. P. White, and D. Pouliot, "Integration of multi-scale remote sensing data for reindeer lichen fractional cover mapping in Eastern Canada," *Remote Sens. Environ.*, vol. 267, Dec. 2021, Art. no. 112731.
- [62] S. Illarionova, D. Shadrin, I. Shukhratov, K. Evteeva, G. Popandopulo, N. Sotiriadi, I. Oseledets, and E. Burnaev, "Benchmark for building segmentation on up-scaled Sentinel-2 imagery," *Remote Sens.*, vol. 15, no. 9, p. 2347, Apr. 2023.



**SVETLANA ILLARIONOVA** received the bachelor's and master's degrees in computer science from Lomonosov Moscow State University, Moscow, Russia, in 2017 and 2019, respectively, and the Ph.D. degree in computer science from the Skolkovo Institute of Science and Technology (Skoltech), Moscow, in 2023. Her research interests include computer vision, deep neural networks, and remote sensing.



**DMITRII SHADRIN** received the M.S. degree in applied physics and mathematics from the Moscow Institute of Physics and Technology (MIPT), in 2016, and the Ph.D. degree in data science from the Skolkovo Institute of Science and Technology (Skoltech), Russia, in 2020. He is currently a Research Scientist with Skoltech, where he is involved in the development of approaches for monitoring and modeling of the carbon footprint. His research interests include data processing, modeling of physical and bioprocesses in closed artificial growing systems, machine learning, and computer vision. He is responsible for the experimental research and several projects in the research center in artificial intelligence in the direction of optimization of management decisions to reduce carbon footprint.



**EVGENY BURNAEV** received the degree from the Moscow Institute of Physics and Technology, in 2006, the Candidate of Sciences degree from the Institute for Information Transmission Problems, in 2008, and the Ph.D. degree in physical and mathematical sciences from the Moscow Institute of Physics and Technology, in 2022.

He was the Head of the Data Analysis and Predictive Modeling Laboratory, Institute for Information Transmission Problems. He is currently the Director of the Skoltech Applied AI Center. The results are published in top computer science conferences, such as ICML, ICLR, NeurIPS, CVPR, ICCV, and ECCV, and journals. His current research interests include developing new algorithms in machine learning and artificial intelligence, such as deep networks for an approximation of physical models, generative modeling, and manifold learning, with applications to computer vision and 3D reconstruction, and neurovisualization.

Prof. Burnaev was honored with many awards for the research, including the Moscow Government Prize for Young Scientists, in 2017, the Ilya Segalovich Yandex Science Prize for the Best Research Director of Postgraduate Students in the field of computer sciences, in 2020, and the Best Paper Award for the research on modeling of point clouds and predicting properties of 3D shapes from the International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR), in 2020.



**ISLOMBEK MIRPULATOV** received the bachelor's and master's degrees in computer science from Lomonosov Moscow State University, Moscow, Russia, in 2019 and 2023, respectively. He is currently a Research Engineer with the Skolkovo Institute of Science and Technology (Skoltech). His research interests include ML, deep neural networks, and computer vision.