Using Isolation Forest and Alternative Data Products to Overcome Ground Truth Data Scarcity for Improved Deep Learning-based Agricultural Land Use Classification Models

Agustin Garcia Pereira
Insight Centre for Data Analytics
University of Galway, Ireland
agustin.garciapereira@insight-centre.org

Lukasz Porwol
Insight Centre for Data Analytics
University of Galway, Ireland
lukasz.porwol@insight-centre.org

Adegboyega Ojo
1) School of Public Policy and Administration, Carleton University, Canada
2) Department of Applied Informatics in Management, Gdansk University of Technology, Poland
AdegboyegaOjo@cunet.carleton.ca

Abstract

High-quality labelled datasets represent a cornerstone in the development of deep learning models for land use classification. The high cost of data collection, the inherent errors introduced during data mapping efforts, the lack of local knowledge, and the spatial variability of the data hinder the development of accurate and spatially-transferable deep learning models in the context of agriculture. In this paper, we investigate the use of Isolation Forest (IF), an anomaly detection algorithm, to reduce noise in a large-scale, low-resolution alternative ground truth dataset used to train land use deep learning models. We use a modest-size, high-resolution and high-fidelity manually collected ground-truth dataset to calibrate Isolation Forest parameters and evaluate our approach, highlighting the relatively low cost of the methodology. Our data-centric methodology demonstrates the efficacy of deep learning methods coupled with IF to create mid-resolution land-use models and map products for agriculture using an alternative ground-truth dataset. Moreover, we compare our deep learning approach with a traditional algorithm used in remote sensing and evaluate the spatial transferability of the created models. Finally, we reflect upon the lessons learnt and future work.

Keywords: Deep learning, agriculture, GIS, data, datasets, isolation forest, ground truth, data-centric AI.

1. Introduction

The 17 United Nations Sustainable Development Goals (SDGs), a set of worldwide objectives to be met by 2030, provide a pathway to eradicate poverty, improve socioeconomic inclusion, and provide greater protection for the environment. To this end, 232 indicators have been defined to measure the progress made towards these goals and the need for monitoring these indicators has led to an increased demand for quality data. Before the introduction of the Sustainable Development Goals, the United Nations highlighted the need for a data revolution to enhance data quality and quantity to monitor different indicators (United Nations, 2013). Almost ten years later, the lack of quality data is still a challenge for many developing countries to direct and monitor their efforts to address SDGs (Bali Swain & Yang-Wallentin, 2020), (Tassopoulou et al., 2019). Paradoxically, the amount of free, high-quality unlabelled EO (Earth Observation) data is ever-increasing thanks to the contributions of Landsat and Sentinel constellations (Gómez et al., 2016).

Specifically, SDGs 2 and 15 are defined as "End hunger, achieve food security and improved nutrition and promote sustainable agriculture" and "Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss" respectively. Each of these objectives presents a list of indicators to help assess the progress made in each case. For instance, indicators: 2.4.1 "Proportion of agricultural area under productive and sustainable agriculture", 15.1.1 "Forest area as a proportion of total land area"; and 15.3.1 "Proportion of land that is degraded over a total land area" rely heavily on land use and land cover data. To this end, the literature suggests that remote sensing has been an effective tool for monitoring the land surface properties resulting from human practices and can greatly contribute to measuring these indicators in a cost-effective way (Tassopoulou et al., 2019). However, ground truth data to train the models and verify the results is an essential part of the remote sensing process (Hoffer, 1971), (Holloway & Mengersen, 2018), and it is usually a bottleneck for many machine learning

URI: https://hdl.handle.net/10125/103244
978-0-9981331-6-4

initiatives in different domains (Sidike et al., 2019) and particularly in the remote sensing area. Moreover, the ground truth data scarcity issue is exacerbated when deep learning architectures are used to train models since they are more data-dependent (Bansal et al., 2021).

In Section 2, we introduce the traditionally used Random Forest algorithm and the new deep learning alternatives. We describe different approaches to addressing the ground truth data scarcity issue, and we describe the Isolation Forest algorithm. In Section 3, we present the datasets used in the study setting and the data pre-processing steps. In Section 4, we describe our methodology, the deep learning architectures used, and the random forest parameters selected, among other implementation details. Then, in Section 5, we describe the results of five different experiments. Finally, we discuss the results, conclusions, limitations and future work in Section 6.

2. Background

Random Forest (RF) has been traditionally used in Remote Sensing for different classification tasks. Schmidt et al. (Schmidt et al., 2016) used several machine learning techniques to create maps of cropping activity for the period 1987-2015 using Landsat imagery. In this study, Random Forest performed better when compared with SVM, multinomial logistic regression, and decision-tree classifiers. Tian et al. (Tian et al., 2016) used Random Forest to map wetland land cover surpassing SVM and Artificial Neural Networks by more than 10%. Chan et al. (Chan & Paelinckx, 2008) compared the performance of Random Forest and Adaboost to classify ecotopes using hyperspectral data showing that both algorithms perform similarly in terms of accuracy (Random Forest's results were more stable), outperforming neural network classifiers. Due to the good classification results and the capacity to handle high dimensionality data, RF is established as a popular algorithm in the remote sensing domain (Belgiu & Drăgu, 2016). Random Forest is an ensemble classification method, which means that uses not only one but many tree predictors that can accurately describe complex relationships among multiple variables (dos Reis et al., 2018). Once the trees in the forest output their decisions, a plurality vote is used to combine the final output using the same weight for each one (Chan & Paelinckx, 2008).

Recently, the use of artificial intelligence, the proliferation of volunteered geographic information culture, and the increasing availability of free EO data present new opportunities to address widescale problems. The advances in computing power and data availability come in parallel with significant developments in the field of artificial intelligence (AI) algorithms, in particular, deep learning. Deep learning models have improved the state-of-the-art in many domains, such as visual object recognition, speech recognition, object detection, and recently, the remote sensing domain (Lecun et al., 2015). In the GIS context, deep learning models have been used for different purposes such as geospatial modelling, remotely sensed imagery processing, navigation, governance and societal, and agriculture. Specifically, the efficacy of one-dimensional convolutional neural networks over RF has been studied when exploiting the temporal and spectral dimensions of remotely sensed imagery in other parts of the world (García Pereira et al., 2020; Pereira et al., 2019, 2021). Other studies have demonstrated the potential of Recurrent Neural Networks (RNNs) such as Long-Short Term Memory (LSTM) to classify multi-temporal Synthetic Aperture Radar (Ienco et al., 2017). The increasing volume and variety of collected geospatial big data creates new opportunities but also poses additional challenges (Li et al., 2016). One such challenge is the lack of high-quality, labelled and large-scale datasets to train deep learning models for land use classification (Sun et al., 2017), (Holloway & Mengersen, 2018). On the one hand, if large-scale datasets are made available, they usually provide a low-spatial resolution and are generally a product derived from computer models whose training and validation data samples are not accessible (Hao et al., 2020), (Sahajpal et al., 2014), (D et al., 2021). On the other hand, the high cost and sometimes prohibitive price of manually-mapped high-quality data collection initiatives make it difficult to repeat them periodically and/or over large extensions of land. Considering the current data-centric artificial intelligence drift, the use of these datasets to train highly performant models using mid-spatial EO resolution data is a challenge worth addressing.

To address the ground truth data scarcity issue, the literature describes several approaches. The use of weak supervision systems for creating training data using labelling functions is an emerging area that is being explored in other domains (Ratner et al., 2020) and in the context of remote sensing (Dao et al., 2019). However, the need for domain experts to define the labelling functions represents a limitation. Other authors have explored the use of Vectorized Code Projected Gradient Descent Unmixing (VPGDU) (Faran et al., 2019) to simulate the ground truth of mid-resolution data by applying unmixing techniques to high-resolution hyperspectral images. However, the selection of relevant endmembers sets is a key issue in achieving successful unmixing (Kizel & Shoshany, 2018). The use of UAVs has also been explored to create high-fidelity ground truth data (Hegarty-Craver et al., 2020), however, this approach is costly to be implemented on a

large scale and previous years' crop maps can not be derived using this technique. Nevertheless, the recent development of efficient unsupervised anomaly detection algorithms opens up new opportunities to address this problem. One of the most recent anomaly detection algorithms is Isolation Forest (IF) (Liu et al., 2008). Despite its simplicity, it excels at dealing with high-dimensional data while exhibiting low linear time complexity and a small memory requirement (Al Farizi et al., 2021). In this way, the quality of alternative large-scale ground truth datasets with low spatial resolution and high per cent error can be enhanced with the use of mid-resolution EO data and IF and used for training high-quality deep learning models.

3. Data Preparation

In this section, we present the study setting. We then introduce and detail the datasets used and finally, we depict the data pre-processing approach.

3.1. Context

This study is conducted in the District of Guaminí, SW of Buenos Aires Province, Argentina. The study setting encompasses an area of 4824.23 square kilometres delimited by the District political boundaries as shown in Figure 1.

Figure 1. Study Context

3.2. Data Sources

The data sources used in this study are presented in Table 1. Data source A is a raster product developed by the "Instituto Nacional de Tecnología Agropecuaria" (INTA) in the context of the Mapbiomas project within the JECAM-GEOGLAM network. The product development methodology consisted of the supervised classification of different indexes obtained from 30 meters resolution Landsat satellite imagery. Training and validation samples were obtained from field observation and complementary information during the agricultural season 2020/2021. From the samples gathered, the authors used 60% for training and 40% for validation. These samples are not part of the final product released. The Argentine agricultural surface was divided into 15 different zones, and independent classifiers were trained and evaluated in each of them. For zone XI, containing this study setting, authors reported a general accuracy of 0.82 and a Kappa score of 0.73 for 7 classes classification, named: Maize, Soybean, Sunflower, Sorghum Grain, Sorghum Grass, Fallow, and Nonagricultural. However, for the classes addressed in this study, Nonagricultural, Maize, Soybean, and Sunflower, overall accuracies reported were 0.91, 0.75, 0.76, and 0.82, respectively, having Maize and Soybean two of the lowest class accuracies in the data product.

Table 1. Data Sources

ID	Data Source	URL
A	Argentina National Summer 2021 and Winter 2020 Crops. Resolution: 30 m.	http://www.geointa.inta.gob.ar/2021/09/07/mapa-nacional-de-cultivos-campana-2020-2021/
B	SENTINEL-2 is a European wide-swath, high-resolution, multi-spectral imaging mission. The satellite's multispectral imager provides a versatile set of 13 spectral bands spanning from the visible and near-infrared to the shortwave infrared and a 5-days revisit time.	https://scihub.copernicus.eu/
C	Ground truth dataset. Agricultural data was manually collected in the context of this study, while nonagricultural data was obtained using Open Street Maps data.	n/a

Data source C was created by the authors of this study in the context of a project named "Supporting Bee-Friendly Agriculture in Argentina". Agricultural fields were visited during February 2021 and crop information was mapped using GIS tools. The extent of the fields was manually digitized by humans using Sentinel 2 relevant satellite imagery as reference. Nonagricultural data was obtained from Open Street Maps. The polygons were filtered by *landuse = "grass*", and *highway = "unclassified" OR highway = "tertiary"* after a visual observation of the data confirmed these classes represented most of the nonagricultural, but rural, land use. A 20 meters buffer was applied to the filtered highway lines to include road verges (where natural vegetation usually grows) in the dataset. The classes mapped are depicted in Table 2.

In Table 3 we provide a mapping between Dataset A and Dataset C. We can observe that the class Sunflower-2nd is not available in Dataset A. This type of cropping practice where Sunflower is grown lately in the summer season after winter crops such as Barley and Wheat is not common in the area, but it is still practised by a small number of farmers.

Table 2. Dataset C Classes Description

Class	*Description*
Fallow	A field without summer crops in the period
Sunflower	Sunflower
Sunflower-2nd	Sunflowers sowed late in the season after a winter crop
Maize	Maize
Pasture	Pasture for cattle
Stubble	Straw and crown of plants left on the soil surface after harvest. No new crop growing.
Soybean	Soybean
Soybean-2nd	Soybean sowed late in the season after a winter crop
Nonagricultural	Rural roads, including road verges, and grasslands.

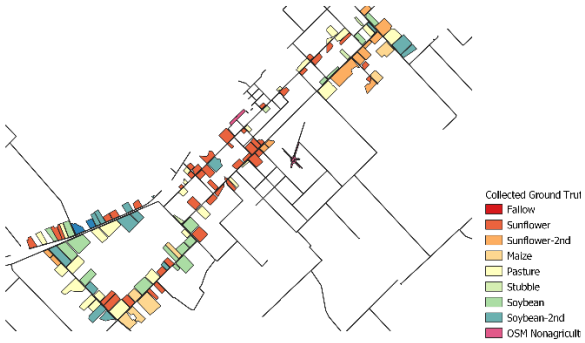

Figure 2. Dataset C

Table 3. Datasets Mapping

Dataset C	*Dataset A*	*Comment*
Sunflower	Sunflower	Direct mapping
Sunflower-2nd	-	Not classified in A
Maize	Maize	Direct mapping
Soybean	Soybean	Direct mapping
Soybean-2nd	Soybean	Direct mapping
Nonagricultural	Nonagricultural	Direct mapping

3.3. Pre-processing

Data pre-processing was performed using Python programming language and the library eo-learn[1]. This library provides a modular approach to defining Earth Observation (EO) data extraction workflows. During this work, we followed a pixel-based approach. Three main pre-processing tasks were performed. The first consisted in downloading satellite imagery data and computing extra information. For this task, the District of Guaminí's political boundaries was used to define the region of interest. A 500 meters buffer was added to ensure the availability of data on the borders. The resulting area was divided into 242 squared patches of 5 km sides. Six different bands (blue, green, red, NIR, SWIR1, and SWIR2) were downloaded, together with cloud mask and cloud probabilities data, and Normalized Difference Vegetation Index (NDVI) spectral index was calculated. The data acquired spanned from September 2020 to May 2021, a period of nine months covering the summer crops growing season. The second task consisted of the creation of two datasets based on two different reference data sources describing the Earth's surface. The first one is the Argentina National Summer 2021 Crops data. In this case, the data was resampled from a 30-meter spatial resolution to a 10-meter spatial resolution matching Sentinel 2 resolution. The second one is the manually collected ground truth data, both described in Table 1. For the two data sources, independent datasets were created by overlaying the geospatial data with the earth observation data gathered in the previous task, using the same coordinate reference systems. Only the directly mapped classes defined in Table 3 are used. Finally, for both datasets, training data were temporally sampled using linear interpolation with a five days interval, and previously removing cloudy data points, creating a 55 data point time series for each pixel in the datasets.

4. Approach

4.1. Methodology

Our methodology presented in Figure 3 focuses on experimentation aiming at training deep learning models using an alternative low-resolution, machine-derived reference data and Isolation Forest as an algorithm to previously reduce data noise. To this end, we evaluate the performance of different deep learning models trained with different IF contamination values. The contamination value represents the proportion of outliers in the dataset and acts as a parameter to control the threshold for the decision function to decide whether a scored data point should be considered an outlier or not. During the experiments we use scikit-learn library IF implementation, using 100 estimators, and a max number of 55 features, matching the number of data points in each pixel. We also compare the performance of the deep learning models with random forest models and we evaluate the spatial transferability of the resulting convolution-based deep learning models. During the training process, each 25-patch grid was split into 15 patches as a training set (60% of the entire grid), 5 patches for the validation set (20%), and 5 patches for the testing set (20%). This spatial splitting was important to reduce the chance that pixels from the same fields are present in training, testing, and/or validation

[1] https://eo-learn.readthedocs.io/

sets at the same time. Only the calculated NDVI index was used for training and as the IF feature input in all the experiments. Due to its ratio properties (using NIR and Red bands), NDVI can cancel out a large proportion of the noise caused by changing sun angles, topography, clouds or shadow, and atmospheric conditions (Huete et al., 1999). The distance used to study the spatial transferability of the models was calculated using two points: the centroid of each test patch, and the centroid of each grid used for model training.

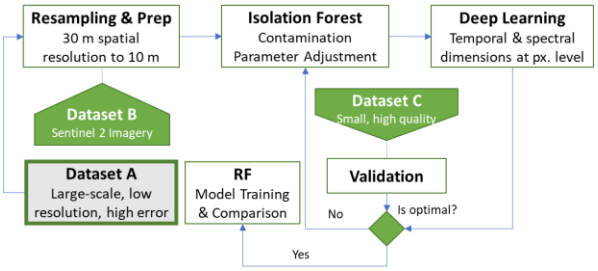

Figure 3. Methodology Schematic Diagram

4.2. Deep Learning Architectures and RF

To define the deep learning pipeline we utilized Ludwig[2], an open-source, declarative machine learning framework. Deep learning methods are characterized by Neural Networks built using more than two hidden layers. The composition of simple but non-linear modules allows DNNs to learn raw data representation at many levels. Starting from the raw input, each level transforms the representation into a more abstract level. In this way, many complex functions can be learned. Convolutional Neural Networks (CNNs) are deep neural networks where one or more convolutional layers are used. Convolution can be seen as applying and sliding a filter over different dimensions of the data representation. In our study, we focus on the use of one-dimensional convolutions involving the temporal and spectral dimensions of remotely sensed time series. To train a deep learning model using the temporal information at the pixel level, we defined the main architecture that works by first mapping the input time series sequence b x s (where b is the batch size, in this case, 128, and s is the length of the time series, in this study is 55) into a stack of one-dimensional convolutional layers with different filter sizes: (6 layers with filter size 7, 7, 3, 3, 3 and 3), followed by a final pool and by a flattening operation. This single flattened vector is then passed through a stack of fully connected layers and returned as a b x h tensor where h is the output size of the last fully connected layer. To compare this architecture with other common neural networks developed for sequential data, we later modify the previous architecture and replace the stack of CNNs with, first, a simple RNN unit and then, an LSTM unit. The preferred way to optimize neural networks in the literature are variants of Stochastic Gradient Descent (SGD), such as Adam and AdaDelta. In particular, Adam has been widely utilized in many optimization problems in the field of machine learning, especially in time series classification problems (Ismail Fawaz et al., 2019). Adam (Adaptive Moment Estimation) computes adaptive learning rates for each parameter and not only stores an exponentially decaying average of past square gradients, but also keeps an exponentially decaying average of past gradients, as in the momentum method. It is also computationally efficient and requires little. In our experiments, we used Adam as the optimization method, with standard parameters suggested by the authors $\beta1= 0.9$, $\beta2=0.999$, and $\alpha=0.001$. During our experiments with neural networks, we implemented an early stopping technique to mitigate the overfitting problem, a model that performs well in the data that has already been seen but does not generalize well with unseen data (low bias and high variance) (Zhang et al., 2019). Because stopping training too early may reduce variance but increment bias and stopping too late may reduce bias but increment variance (Yao et al., 2005), we utilized the validation set accuracy to stop the learning when validation accuracy decreases over two epochs. Studies have analyzed the impact of parameter selection on RF performance (Belgiu & Drăgu, 2016). RF requires the tuning of four parameters: 1) k, the number of trees; 2) m, the number of randomly selected features at each node; 3) max_depth, the maximal depth of each tree, and; and 4) min_samples, the minimal number of samples per node. Even though some studies have proven that RF parameter selection does not have a very significant impact on classification accuracy, some general recommendations can be followed to improve its performance. Rodriguez-Galiano et al. (Rodriguez-Galiano et al., 2012) demonstrated that the number of trees (k) is directly proportional to the classifiers' accuracy up to the number of 100 trees. Once this value is reached, the generalization error converges. Pelletier et al. (Pelletier et al., 2016) studied different values of k, ranging from 50 to 400, and also concluded that this value can be set to 100 without a major accuracy loss. Other studies have used k values of 500 for land use classification (Pelletier et al., 2018). The m parameter value suggested by the literature is the square root of p, where p is the number of features (Liaw & Wiener, 2002). However, small values of m have shown very good performance due to the reduction correlation

[2] https://ludwig-ai.github.io/ludwig-docs/0.5/

among individual trees (Pelletier et al., 2016). Finally, the values for max_depth and min_samples have been less explored in the literature. Pelletier et al. (Pelletier et al., 2016) used a max_depth of 25, and a min_samples of 10 or 25, and showed that the accuracy impacts of these parameters' selection are low. In our experiments, we used the RF implementation of the python library Sklearn with the following parameters: ***k***=*200*; ***m***=$\sqrt{p}$,*where* $p=n\_features$; ***max_depth*** = *None, nodes are expanded until all leaves are pure or until all leaves contain less than min_samples samples;* ***min_samples*** = *2.*

5. Results

In this section, we present the results of five different experiments conducted in the study region.

5.1. Experiment 1

To evaluate the performance of CNN models using the alternative ground truth Dataset A and different IF contamination thresholds, we defined two separate data grids using the spatial square patches grid defined in Section 3.3.

Table 4. Grid 1 Contamination Results

Contamination	20%	30%	40%	50%
Accuracy	0.779	0.802	0.792	0.867
Loss	0.94	1.214	0.54	0.553
Avg F1 score weighted	0.778	0.805	0.795	0.864
Avg F1 score macro	0.764	0.793	0.791	0.852
Avg F1 score micro	0.779	0.802	0.792	0.867
Kappa score	0.622	0.665	0.651	0.771

Table 5. Grid 2 Contamination Results

Contamination	20%	30%	40%	50%
Accuracy	0.773	0.792	0.84	0.873
Loss	0.663	0.524	0.426	0.47
Avg F1 score weighted	0.775	0.791	0.84	0.87
Avg F1 score macro	0.766	0.78	0.83	0.859
Avg F1 score micro	0.773	0.792	0.84	0.873
Kappa score	0.610	0.641	0.723	0.775

Each grid encompassed 25 squared patches. Both grids can be observed in Figure 4 and Figure 5. For each grid, we trained a different model to classify the three more predominant crops in the study area, named: maize, soybean, and sunflower. We used contamination values of 20, 30, 40 and 50 per cent and we evaluated the resulting models using a subset (only 3 classes) of ground truth Dataset C, which location is detached from the two grids. Table 4 shows the results for Grid 1 and Table 5 the results for Grid 2.

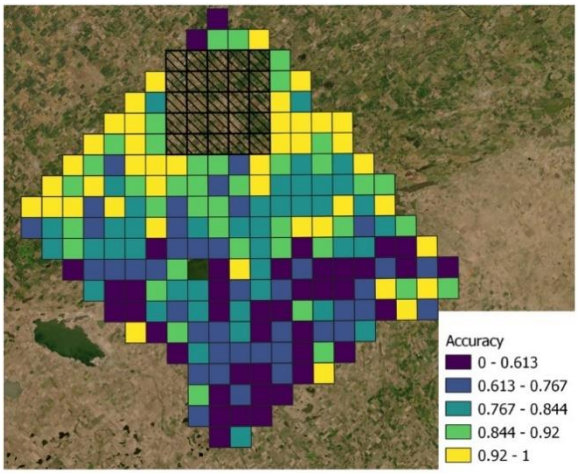

Figure 4. Grid 1 Spatial Transferability

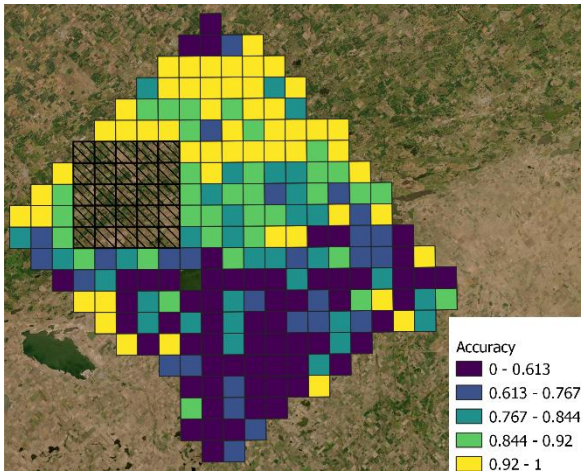

Figure 5. Grid 2 Spatial Transferability

5.2. Experiment 2

In Experiment 2 we evaluate the spatial transferability of the models created before.

Table 6. Grid 1 and Grid 2 Distance Correlation

	Grid 1	Grid 2
Accuracy	-0.44573	-0.36454
Loss	0.33925	0.31751
Avg F1 score weighted	-0.43624	-0.35079
Avg F1 score macro	-0.53468	-0.36397
Avg F1 score micro	-0.44573	-0.36454
Kappa score	-0.54482	-0.34515
Distance	1	1

From Experiment 1 we observe that the best model results were produced using contamination of 50%, for both grids. We use these models to evaluate their performance in different locations defined by the

remaining square patches. For the data available for each patch, a 50% contamination Isolation Forest was run before testing as well, removing the noise on the testing set. We then calculated the distance of each patch as explained in Section 4. The correlation between the distance and different model performance statistics is then calculated. Table 6 present the results for each grid and Figure 6 and Figure 7 present the regression results for the distance and the accuracy.

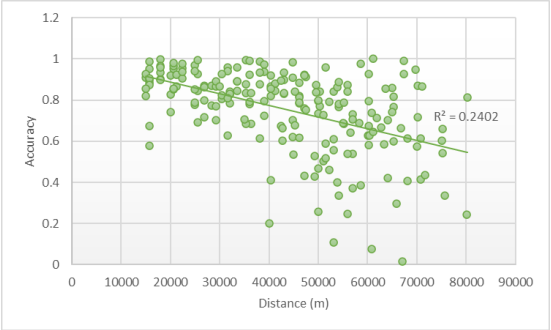

Figure 6. Grid 1 Distance and Accuracy Regression

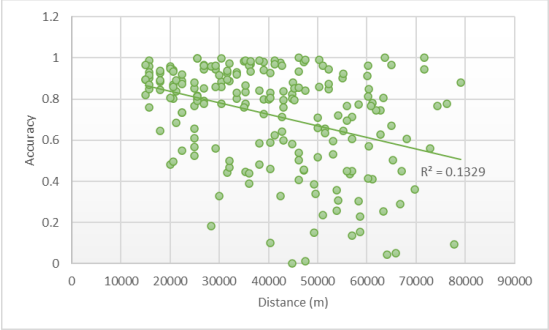

Figure 7. Grid 2 Distance and Accuracy Regression

5.3. Experiment 3

As can be observed from the previous two experiments, the models' performance decreases with the distance between the model training area and the area where they are used. Considering this, we conducted a third experiment to evaluate the results of a model trained with a combination of both grids. Because these grids are spatially separated, our intuition suggests that the features learnt by the model will be more comprehensive and the resulting model should perform better when used to predict Dataset C. To test this hypothesis, we joined Grid 1 and Grid 2 training, evaluation and testing set separately, creating a combined dataset with 50 total patches, 30 in the training set, 10 in the evaluation set, and 10 in the testing set. We used a contamination value of 50% as in the previous experiment, and we named this model "model 1". Table 7 presents the results obtained and compares them with the two single grid models trained before. Table 8 presents the class accuracy results.

Table 7. Combined Grid Performance Model 1

	Grid_1	Grid_2	Combined
Accuracy	0.867	0.873	**0.890**
Loss	0.553	0.47	**0.367**
Avg F1 score weighted	0.864	0.87	**0.889**
Avg F1 score macro	0.852	0.859	**0.882**
Avg F1 score micro	0.867	0.873	**0.890**
Kappa score	0.771	0.775	**0.809**

Table 8. Combined Grid Class Accuracy

	Combined Grid Class Accuracy
Maize	0.891
Soybean	0.925
Sunflower	0.964

5.4. Experiment 4

Our previous experiments focused on the classification of maize, soybean and sunflower, the three main summer crops grown in the study region. However, the model is not able to classify agricultural and nonagricultural land.

Table 9. Agricultural Land Model Results

Accuracy	*Loss*	*Avg F1 weighted*	*Avg F1 macro*	*Avg F1 micro*	*Kappa SC*
0.96	**0.102**	**0.962**	**0.904**	**0.96**	**0.808**

In this experiment, we used Dataset A and the combined grid to train a model capable of classifying agricultural and nonagricultural land and we name it "model 2". This time, we combined all the agricultural classes of the dataset into a single class, and we used the nonagricultural class already available in the product. We used a 50% contamination percentage as per the findings of our previous experiments. To evaluate the model performance, we used Dataset C, combining agricultural classes into a single class, as in Dataset A. After this step, the resulting Dataset C is imbalanced, being the nonagricultural class underrepresented (10% of the dataset). This fact needs to be considered when evaluating the models' performance. Table 9 presents the results obtained.

5.5. Experiment 5

In this experiment, we first compare the performance of the CNN models developed before (Model 1 and Model 2) with other three methods: two

recurrent neural networks (simple RNN-based and LSTM-based), and the traditional RF. Following the approach applied in previous experiments, we used a contamination percentage of 50% and the datasets resulting from the combined grid to train three new models to classify maize, soybean and sunflower; and three new models to classify agricultural and nonagricultural land use. These results are presented in Table 10.

Table 10. Model 1 and Model 2 RF Comparison

	Model 1				**Model 2**			
	RF	***RNN***	***LSTM***	*CNN*	***RF***	***RNN***	***LSTM***	*CNN*
Accuracy	**0.834**	**0.786**	**0.725**	0.89	**0.933**	**0.936**	**0.945**	0.96
Loss	-	**0.438**	**0.592**	0.367	-	**0.192**	**0.11**	0.107
Avg F1 weight.	**0.83**	**0.791**	**0.727**	0.889	**0.94**	**0.938**	**0.949**	0.962
Avg F1 macro	**0.82**	**0.792**	**0.714**	0.882	**0.85**	**0.56**	**0.58**	0.904
Avg F1 micro	**0.834**	**0.786**	**0.725**	0.89	**0.933**	**0.936**	**0.945**	0.96
Kappa SC	**0.729**	**0.654**	**0.526**	0.809	**0.705**	**0.68**	**0.744**	0.808

Finally, to study the contribution of IF pre-processing to the algorithms, we used different contamination values to train RF models as we did in Experiment 1, for Grid 1 and Grid 2. These results and their respective comparisons are presented in Table 11 and Table 12.

Table 11. Grid 1 RF & CNN Comparison

Contam.	**20%**		**30%**		**40%**		**50%**	
	RF	*CNN*	*RF*	*CNN*	*RF*	*CNN*	*RF*	*CNN*
Accuracy	0.798	0.779	0.799	0.802	0.78	0.792	0.789	0.867
Kappa SC	0.654	0.622	0.659	0.665	0.623	0.651	0.63	0.771

Table 12. Grid 2 RF & CNN Comparison

Contam.	**20%**		**30%**		**40%**		**50%**	
	RF	*CNN*	*RF*	*CNN*	*RF*	*CNN*	*RF*	*CNN*
Accuracy	0.750	0.773	0.758	0.792	0.784	0.84	0.831	0.873
Kappa SC	0.575	0.61	0.591	0.641	0.630	0.723	0.70	0.775

6. Discussion and Conclusion

In this paper, we showed that the use of an alternative ground truth dataset together with Isolation Forest is an effective approach to train high-quality deep learning CNN-based models that significantly outperform the state-of-the-art random forest algorithm for land use classification. Our results show the improved CNN models' performance and proved that the approach can help alleviate the lack of high-quality ground truth labelled datasets in machine learning and remote sensing domain by using an alternative and freely available ground truth data source.

In Experiment 1 we showed that a contamination value of 50% was optimal compared to lower values. The high contamination percentage can be explained by the fact that on top of the error carried by Dataset A data reference product, a new nature of the error was introduced during the pre-processing step described in Section 3, where we resampled Dataset A from 30 meters to 10 meters. In this step, a labelled pixel in Dataset A generated nine new pixels with the same label in the 10 meters spatial resolution data. Depending on the spatial context of this pixel, new errors can be introduced. In Experiment 2, we showed that the models' performance degrades with the increasing distance, and we provide evidence of the statistical significance of this relationship by using correlation and regression analysis. In addition, we show in Experiment 3 that the model performance can be improved by selecting training, validation, and testing data from detached locations. When compared with Dataset A performance report, our model outperforms the individual class accuracies reported in the data product description for the three crops, as shown in Table 8 and Section 3. In Experiment 4, we used the knowledge gained from previous experiments to train a CNN model to classify agricultural and nonagricultural land use showing outstanding results when compared with Dataset A data product description. Because Dataset C is imbalanced, the accuracy statistic is not a good performance indicator to evaluate the model. In this way, we can observe that the Kappa score, a performance metric that takes imbalanced classes into account, is higher than 0.80 for the CNN model, denoting a remarkable level of agreement between true values and predicted values. Moreover, the fact that the model is able to detect nonagricultural areas such as road verges and rural streets is of high importance to assessing ecosystems services for agroecosystems and agricultural landscapes (a desired property in the original Dataset A), highlighting the importance of having incremented the spatial resolution of the data using Sentinel 2 imagery.

Finally, in Experiment 5, we compared three deep learning models' performance with the traditional random forest models, and we provided evidence of the increased performance of the CNN models over the others. The lower performance of RNNs models might be explained by the fact that they excel at tasks that require a prediction at each time point, while land use classification aims at producing one label for all the time points. We show that the CNN approach proposed significantly outperformed the traditional RF for every performance metric; especially by almost 6% on Model 1 in terms of accuracy, and more than 10% in terms of the Kappa score coefficient for Model 2. Moreover, we studied how different IF contamination values help RF and the CNN method. Results presented in Table 11 and Table 12 indicate that the CNN method may benefit more from the use of IF in the data preprocessing step. When compared to other studies addressing the same

challenge, our approach is less dependent on domain expertise. However, local knowledge is needed to create a mapping between the datasets classes. Moreover, our approach does not depend on the definition of endmembers sets as in (Faran et al., 2019). When compared to the use of UAVs, our approach is cheaper to implement since we use a freely available alternative ground truth dataset, and a modest size but high-fidelity ground-truth dataset to calibrate IF parameters where part of the dataset was created using Open Street Maps volunteered data.

Additionally, our approach can be implemented using alternative ground truth data from Dataset A in different locations at the National level, alleviating the models' spatial transferability issue. Furthermore, it can be used after data creation using labelling functions to improve the training data quality. Another important property of our approach is that it relies only on temporal NDVI data and not on several independent spectral bands, minimizing the memory requirements during training.

In summary, the use of IF during the data preprocessing step can greatly improve the quality of the models developed using an alternative ground truth dataset during the training phase and a high-quality ground truth dataset to calibrate IF contamination value parameter. When comparing RF and CNN methods, our results indicate that the CNN approach may benefit more from the use of IF. Because this algorithm has a linear time complexity with a low constant and a low memory requirement, its application during data preprocessing is justified by the resulting model's quality improvement.

The limitations of this study include the fact that the experiments are carried out in a specific location, using two particular reference datasets and specific satellite imagery. Another limitation is that the approach presented here uses IF as the main anomaly detection algorithm and no comparison with other similar algorithms were investigated. Future work includes experimenting with deep learning models and different anomaly detection algorithms to implement a spatially progressive learning approach to overcome the spatial transferability issue of deep learning models without the need for large-scale datasets.

7. References

Al Farizi, W. S., Hidayah, I., & Rizal, M. N. (2021). Isolation Forest Based Anomaly Detection: A Systematic Literature Review. *2021 8th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2021*, 118–122. https://doi.org/10.1109/ICITACEE53184.2021.9617498

Bali Swain, R., & Yang-Wallentin, F. (2020). Achieving sustainable development goals: predicaments and strategies. *International Journal of Sustainable Development and World Ecology*, *27*(2), 96–106. https://doi.org/10.1080/13504509.2019.1692316

Bansal, M. A., Sharma, D. R., & Kathuria, D. M. (2021). A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *ACM Comput. Surv.* https://doi.org/10.1145/3502287

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011. https://doi.org/10.1016/j.rse.2008.02.011

D, D. A., Verón, S., Banchero, S., M, I. E., Valiente, S., Puig, O., Murray, F., Jp, M., Zelaya, K., Maidana, D., Varlamoff, N., Peiretti, J., Benedetti, P., Portillo, J., Melilli, M., Maidana, E., & Goytía, Y. (2021). *Mapa Nacional de Cultivos campaña 2020 / 2021*.

Dao, D., Rausch, J., & Zhang, C. (2019). GeoLabels: Towards Efficient Ecosystem Monitoring using Data Programming on Geospatial Information. *NIPS Workshop, NeurIPS*. https://daviddao.org/papers/geolabels.pdf

dos Reis, A. A., Carvalho, M. C., de Mello, J. M., Gomide, L. R., Ferraz Filho, A. C., & Acerbi, F. W. (2018). Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods. *New Zealand Journal of Forestry Science*, *48*(1), 1–17. https://doi.org/10.1186/s40490-017-0108-0

Faran, I., Netanyahu, N. S., Omid David, E., Shoshany, M., Kizel, F., Chang, J. G., & Rud, R. (2019). *Ground Truth Simulation for Deep Learning Classification of Mid-Resolution Venus Images Via Unmixing of High-Resolution Hyperspectral Fenix Data*. *July*, 807–810. https://doi.org/10.1109/igarss.2019.8900186

García Pereira, A., Ojo, A., Curry, E., & Porwol, L. (2020). Data Acquisition and Processing for GeoAI Models to Support Sustainable Agricultural Practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences 2020 (HICSS 2020)*, 922–931. https://doi.org/http://hdl.handle.net/10125/63854

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Hao, P., Di, L., Zhang, C., & Guo, L. (2020). Transfer Learning for Crop classification with Cropland Data Layer data (CDL) as training samples. *Science of the Total Environment*, *733*, 138869. https://doi.org/10.1016/j.scitotenv.2020.138869

Hegarty-Craver, M., Polly, J., O'Neil, M., Ujeneza, N., Rineer, J., Beach, R. H., Lapidus, D., & Temple, D. S. (2020). Remote crop mapping at scale: Using satellite imagery and UAV-acquired data as ground truth. *Remote Sensing*, *12*(12), 1–15. https://doi.org/10.3390/rs12121984

Hoffer, R. M. (1971). The Importance of "Ground Truth" Data in Remote Sensing. *United Nations Panel Meeting on the Establishment and Implementation of Research Programs in Remote Sensing*, 1–12. https://ntrs.nasa.gov/search.jsp?R=19730007768%0APaper presented at the %22United Nations Panel Meeting on the Establishment and Implementation of Research Programs in Remote Sensing. Held at the National Institute for Space Research. November 29 - Decemb

Holloway, J., & Mengersen, K. (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, *10*(9). https://doi.org/10.3390/rs10091365

Huete, A. R., Didan, K., & Van Leeuwen, W. (1999). Modis Vegetation Index. *Vegetation Index and Phenology Lab*, *3*(1), 129.

Ienco, Di., Gaetano, R., Dupaquier, C., & Maurel, P. (2017). Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, *14*(10), 1685–1689. https://doi.org/10.1109/LGRS.2017.2728698

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Kizel, F., & Shoshany, M. (2018). Spatially adaptive hyperspectral unmixing through endmembers analytical localization based on sums of anisotropic 2D Gaussians. *ISPRS Journal of Photogrammetry and Remote Sensing*, *141*, 185–207. https://doi.org/10.1016/j.isprsjprs.2018.03.021

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Pelletier, C., Valero, S., Inglada, J., Champion, N., & Dedieu, G. (2016). Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, *187*, 156–168. https://doi.org/10.1016/j.rse.2016.10.010

Pelletier, C., Webb, G. I., & Petitjean, F. (2018). *Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series*. *March*. https://doi.org/10.3390/rs11050523

Pereira, A. G., Porwol, L., Ojo, A., & Curry, E. (2019). Towards a Temporal Deep Learning Model to Support Sustainable Agricultural Practices. *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, 1–12.

Pereira, A. G., Porwol, L., Ojo, A., & Curry, E. (2021). Exploiting the Temporal Dimension of Remotely Sensed Imagery with Deep Learning Models. *54th Hawaii International Conference on System Sciences*, 5317–5326. http://www.edwardcurry.org/publications/0520.pdf

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *VLDB Journal*, *29*(2–3), 709–730. https://doi.org/10.1007/s00778-019-00552-1

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(1), 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Sahajpal, R., Zhang, X., Izaurralde, R. C., Gelfand, I., & Hurtt, G. C. (2014). Identifying representative crop rotation patterns and grassland loss in the US Western Corn Belt. *Computers and Electronics in Agriculture*, *108*, 173–182. https://doi.org/10.1016/j.compag.2014.08.005

Schmidt, M., Pringle, M., Devadas, R., Denham, R., & Tindall, D. (2016). A framework for large-area mapping of past and present cropping activity using seasonal landsat images and time series metrics. *Remote Sensing*, *8*(4). https://doi.org/10.3390/rs8040312

Sidike, P., Sagan, V., Maimaitijiang, M., Maimaitiyiming, M., Shakoor, N., Burken, J., Mockler, T., & Fritschi, F. B. (2019). dPEN: deep Progressively Expanded Network for mapping heterogeneous agricultural landscape using WorldView-3 satellite imagery. *Remote Sensing of Environment*, *221*(December 2018), 756–772. https://doi.org/10.1016/j.rse.2018.11.031

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 843–852. https://doi.org/10.1109/ICCV.2017.97

Tassopoulou, M., Verde, N., Mallinis, G., Georgiadis, C., Kaimaris, D., & Patias, P. (2019). *Demonstrating the potential of remote sensing to support sustainable development goals implementation: case studies over Greece*. *June 2019*, 42. https://doi.org/10.1117/12.2533634

Tian, S., Zhang, X., Tian, J., & Sun, Q. (2016). Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China. *Remote Sensing*, *8*(11), 1–14. https://doi.org/10.3390/rs8110954

United Nations. (2013). *A new global partnership: Eradicate poverty and transform economies through sustainable development, the report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, June 2013*. 81. http://www.post2015hlp.org/wp-content/uploads/2013/05/UN-Report.pdf

Yao, Y., Rosasco, L., & Caponnetto, A. (2005). *On Early Stopping In Gradient Descent Boosting*. 1–22.

Zhang, C., Recht, B., Bengio, S., Hardt, M., & Vinyals, O. (2019). Understanding deep learning requires rethinking generalization. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.