Geo-information Science and Remote Sensing
Thesis Report GIRS-2020-03

Comparing Classification of Ghana's Complex Agroforestry Land Cover by a Random Forest and a Convolutional Neural Network with a Small Training Set

Thesis submitted in partial fulfilment of the requirements for the Degree of Master of Science

by

Anne-Juul Welsink

12-03-2020

WAGENINGEN
UNIVERSITY & RESEARCH

Comparing Classification of Ghana's Complex Agroforestry Land Cover by a Random Forest and a Convolutional Neural Network with a Small Training Set

by

Anne-Juul Welsink
Registration number: 96 03 24 938 040

Supervisors:

dr.ir. Jan Verbesselt
dr. Nandika Tsendbazar

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at Wageningen University and Research Centre,
The Netherlands.

12-03-2020

Wageningen, The Netherlands

Thesis code number: *GRS-80436*
Thesis Report: *GIRS-2020 -03*
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Abstract

Comparing Classification of Ghana's Complex Agroforestry Land Cover by a Random Forest and a Convolutional Neural Network with a Small Training Set

Anne-Juul Welsink, Wageningen University
MSc Geo-Information Science
03-2020

Accurate and up-to-date land cover maps are key to environmental research, monitoring of climate change and deforestation, resource management, and disaster prevention**.** Over the past years, Convolutional Neural Networks (CNNs) have replaced traditional algorithms such as the Random Forest (RF) as a dominant land cover classifier, thanks to their superior performance. However, little research has been done regarding the performance of either algorithm with different amounts of training data. This study asks how the performance of the RF and CNN compares for a 9 class land cover classification based on Sentinel-1 and Sentinel-2 imagery in West Africa. The performance of the RF and CNN was assessed with both 400 and 800 training points per class. In addition, this study asks which land cover classes were most affected by the choice of algorithm, and which of the two performed best on these classes. A study area in South-West Ghana was used; a region where several commercial crops and drivers of deforestation are grown in a country with one of the highest net deforestation rates worldwide. The results of this study show that with the relatively small training sets that were used, the CNN was more prone to overfitting than the RF due to its higher bias towards the training set. Neither algorithm structurally showed better performance in areas of major disagreement; the algorithm that classified an area as *cacao* was always favoured over the alternative. This implies that, even though the CNN may have a higher potential accuracy, a RF can outperform the CNN when little training data is available, when intra-class variability is high, and/or when the parameterization of the CNN is suboptimal. Data acquisition is a major challenge in deforestation monitoring and the remote sensing domain in general. The insight that this study provides into the performance of the RF and CNN with limited input data can guide the choice of a suitable algorithm in future research and puts the presumed superiority of the CNN into perspective.

Acknowledgements

This research would not have been possible without the support and constructive criticism of my supervisors at the WUR, Jan Verbesselt and Nandika Tsendbazar, and at CIAT, Louis Reymondin, Thibaud Vantalon, and Eric Rahn. They have guided me with patience through the use of new software, statistical formulas, and errors. I specifically would like to thank Thibaud for allowing me to make use of his efforts on pre-processing and training the CNN.

I owe an acknowledgement to CIAT's Christian Bunn for sharing his carefully collected ground truth data. At the WUR, I would like to thank Diego Marcos Gonzales for helping me understand the functioning of the CNN. At home, my parents, Fleur, Monne, and friends in the Netherlands and Vietnam have been a constant source of support, enthusiasm, and patience. I would not do this without you.

The cover art of this thesis was made by Gigi van Grevenbroek, whose work on socially and environmentally relevant themes can be found via gigivangrevenbroek.com.

Table of Contents

List of Tables

List of Figures

Glossary

CI	Confidence Interval (statistical term)
CNN	Convolutional Neural Network (algorithm)
IW	Interferometric Wide Swath (Sentinel-1 mode)
NDVI	Normalized Difference Vegetation Index
RF	Random Forest (algorithm)
SAR	Synthetic Aperture Radar (Sentinel-1)
Sieving	A post-processing method aimed at noise reduction
UAV	Unmanned Aerial Vehicle

1. Introduction

Anthropogenic activities have a strong impact on natural forests and other forms of land cover across the world (Tondoh et al., 2015). Accurate and up-to-date land cover maps are key to environmental research, monitoring of climate change and deforestation, resource management, and disaster prevention (Pelletier et al., 2019). Satellite imagery is used to create land cover classifications used to monitor such dynamics (DeFries et al., 2006). However, classification is often challenging due to the high intra-class variance of objects and spectral similarity, particularly in agroforestry landscapes with mixed land cover (Duguma et al., 2001; Yang et al., 2018). In those complex landscapes, the spectral signatures of different agroforestry crops are often difficult to distinguish, which complicates classification.

In order to optimize land cover classification, it is important to know which algorithm performs best in which situation. The RF is renowned for its high performance and often used as a baseline in classification (Zhong et al., 2019). The RF builds a number of binary decision trees, and uses a bootstrap sample of the data of each tree (Pelletier et al., 2019). The data are recursively split into two subsets and all possible splits are tested based on a feature value until all nodes are pure or when a user-defined criterion is met. The RF is appreciated for its easy-to-tune parameters and its robustness to the presence of mislabelled data (Pelletier et al., 2019). Over the past few years however, the CNN has emerged as a dominant algorithm in land cover classification, thanks to its superior performance (Scott et al., 2017). A CNN applies convolutions in both x and y dimensions (Pelletier et al., 2019). Each convolutional layer takes the outputs of the previous layer as inputs. Significant expertise is required to choose and optimize the hyperparameters, but the CNN has the potential to achieve a significantly higher accuracy than conventional classifiers (including the RF) (Helber et al., 2018; Huang et al., 2018; Liu et al., 2018). However, the CNN is prone to overfitting, especially with limited training data (Liu et al., 2018). There is a need to gain more insight into the performance of the RF and CNN with different training sample sizes, in order to better understand in which cases investment in the parameterization of a CNN is worthwhile.

Relatively little research has been done to compare how either algorithm performs with small or varying amounts of input data. In fact, Liu et al. (2018) claim that their study on object-based wetland mapping with Unmanned Aerial Vehicle (UAV) data is the first to comprehensively compare a CNN with conventional classifiers, taking the size of the training sample into account. Further research on the performance of different algorithms with very different amounts and qualities of training data could guide the choice for a particular algorithm, given an available set of data. This is particularly useful in the remote sensing domain, where acquisition of large and/or high-quality datasets is a major challenge (Helber et al., 2018; Scott et al., 2017).

This study builds on the novel research by Liu et al. (2018) by providing insight into the performance of a RF and CNN when trained with a relatively small amount of data in an agroforestry landscape. The comparison of these algorithms is based on a case study in South-West Ghana. The classification includes the most important tree crops and potential drivers of deforestation that are found in the area of interest (Anderman et al., 2014; Chiti et al., 2014). The following research questions are addressed:

1. How does the performance of RF and CNN algorithms compare for 9-class land cover classification in South-West Ghana?
2. Does the performance of the RF and CNN differ with reduced training data (400 points as compared to 800 training points per class)?
3. Which land cover classes are most affected by the choice of the two algorithms, and which algorithm performs best for these classes?

Sentinel-1 and Sentinel-2 data provide the basis for the classification. Sentinel-1's Synthetic Aperture Radar (SAR) provides observations in the C-band. This study used imagery collected in Interferometric Wide Swath (IW) mode with a spatial resolution of 5 by 20 meters and a 250 km swath. Sentinel-2's multi-spectral imagery is available in a spatial resolution of up to 10 meters, covering 13 bands.

This study compares the performance of a Random Forest (RF) and a Convolutional Neural Network (CNN) on land cover classification with 9 classes in Ghana based on Sentinel-1 and Sentinel-2 imagery. It investigates which algorithm generalizes best with 400 and 800 training points per class, respectively. Elaborate validation was performed on the classifications with 800 training points, in line with 'good practice' recommendations from Olofsson et al. (2014). Furthermore, this study investigates which land cover classes are most affected by the choice of the two algorithms and which algorithm performs best for these classes. The following sections provide a detailed description of the data and methods that were used. Next, an overview of the results is presented, followed by a discussion and limitations section in which the results are interpreted and placed in the context of existing research. A brief conclusion wraps it all up.

2. Area and Classes of Interest

The study area in South-West Ghana encompasses an area of around 67 800 km^2 (Figure 1). Ghana's tropical forests are located in the south and west, while the central and northern zones are savanna (Förster, 2009). The area of interest includes the denser forest and protected (and deforested) areas around Kumasi as well as some less dense savannahs in the northern part.

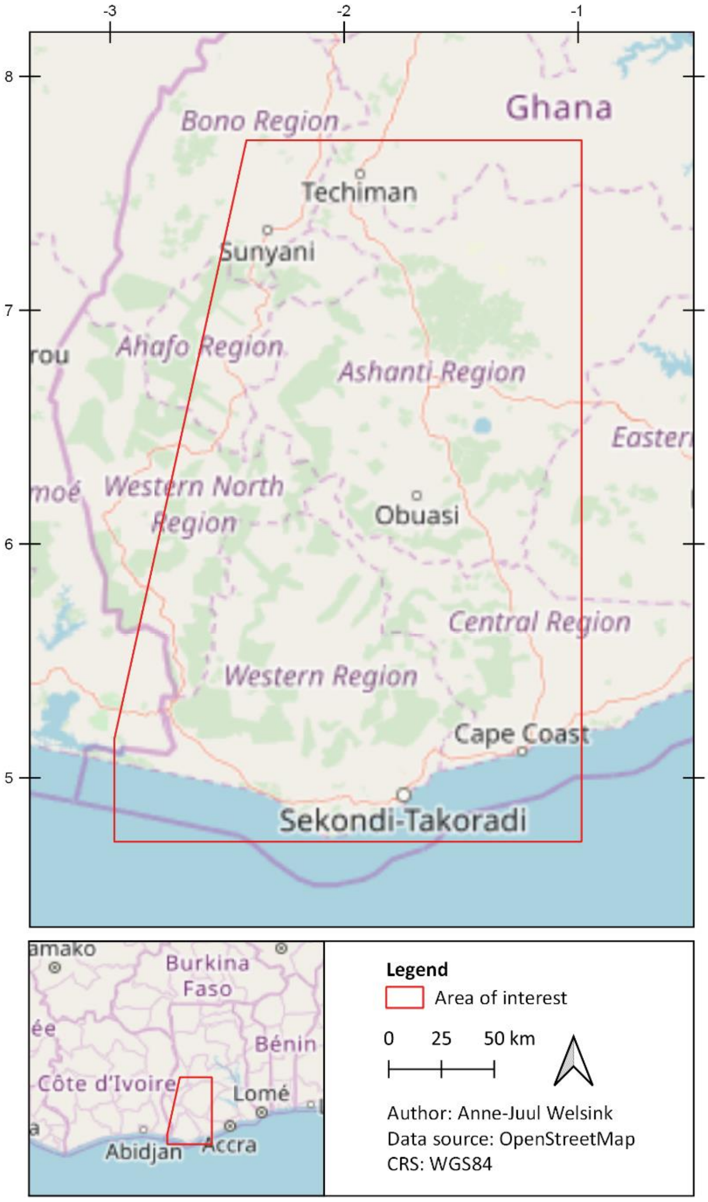

Figure 1. Study area in South-West Ghana.

Ghana is among the countries with the highest net deforestation rates worldwide (FAO, 2015). At the start of the 20th century, around one-third of Ghana was covered by natural tropical forest, of which an estimated 78% had disappeared by 1989 (Appiah et al., 2009; Hawthorne, 1989). More recently, an annual deforestation rate of around 3% has been recorded (Appiah et al., 2009). Deforestation is enticing to farmers, as newly cleared forest areas produce higher commercial crop yields than replanted areas (Rice & Greenberg, 2000). In addition, clearing a new forest area costs about half the effort of replanting (Rice & Greenberg, 2000). This renders Ghana's agroforestry landscape an important focus for deforestation monitoring.

Cacao *(Theobroma cacao)* is one of the most important cash crops in West Africa, and the area provides about 70% of global cacao demand (Ruf, Schroth, & Doffangui, 2015). The crop is the primary driver of deforestation in Ghana; an estimated 27% of deforestation has been attributed to cacao production (DeVries et al., 2015; Kroeger et al., 2017). However, cacao and natural forest are difficult to distinguish on satellite imagery due to their high spectral similarity (Duguma et al., 2001). Therefore, special attention is paid to the classification of this crop.

Besides cacao, oil palm and rubber plantations are among the most important drivers of deforestation (Asubonteng et al., 2018; Chiti et al., 2014). These crops, too, need to be distinguished from natural forest as well as cacao plantations in order to obtain reliable information on drivers of deforestation. Tree crops such as shea nuts or mangoes are included under the general class of *other tree crops*, because the distinction between these crops is not easily made on the basis of satellite imagery (Yaro et al., 2017). Furthermore, there is a class for *seasonal crops*, which have a different temporal signature than the other classes (which is reflected in the standard deviation). The classification is completed with classes for *low vegetation, urban*, and *water.*

3. Data

This section provides insight into the data that were used for this study. It starts with an overview of the data that were used for training and testing, followed by the validation set. Finally, the remote sensing imagery that was used as a basis for the classifications is introduced.

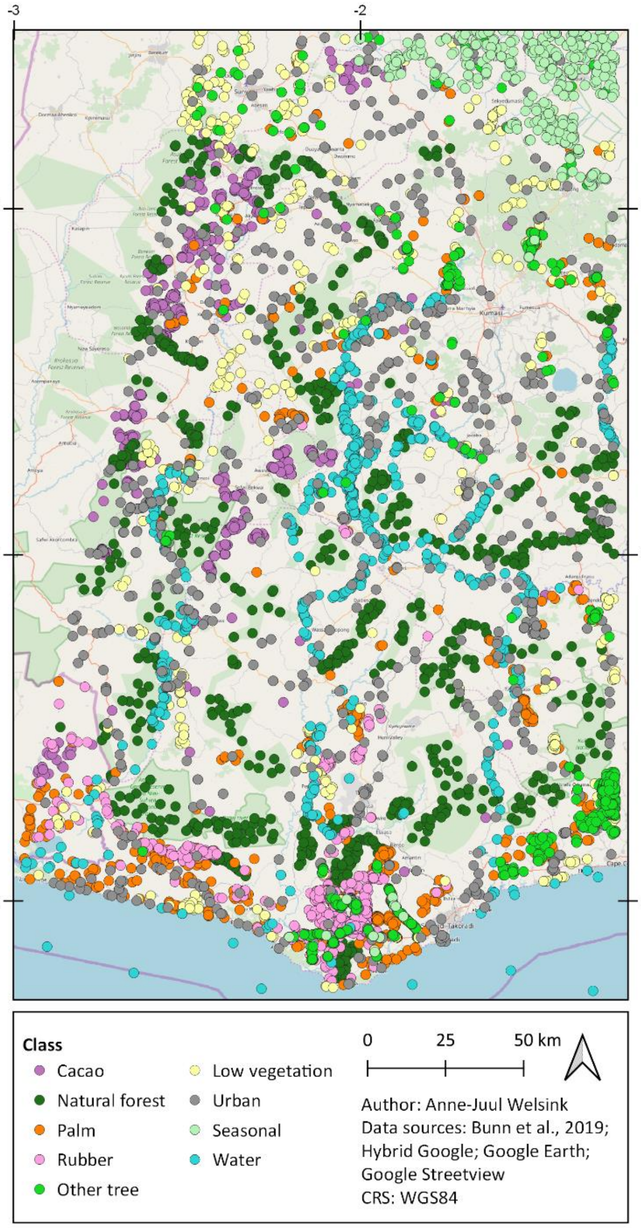

Figure 2. Points used to train the RF and CNN. Here, 800 points per class are depicted. A random stratified subset was used to train with 400 points per class.

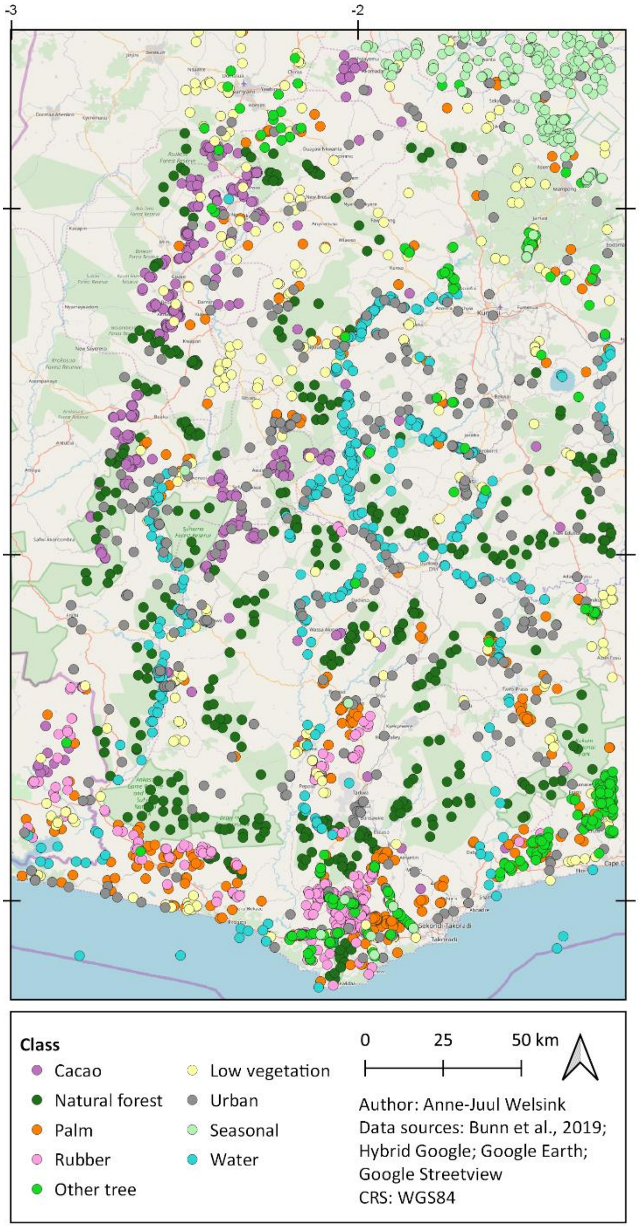

Figure 3. Validation points used in the testing phase (400 per class).

3.1 Training and testing

For each of the 9 classes of interest, 1200 points were collected, of which 800 were randomly selected for training and 400 were used for testing (Figure 2, Figure 3). For eight classes, points were collected in QGIS 3.10.0 based on a hybrid Google layer in combination with Google Earth Pro and Google Street View data from 2017. A circular support area with a diameter of 42m was used to establish the land cover (size based on personal communications with CIAT). For the 9th class of *cacao*, over 3000 ground truth points were available that were collected in the field in 2016 (Bunn et al., 2019). A verified subset of those points was used to train the RF and CNN. Verification took place based on imagery from 2017 on hybrid Google, Google Earth Pro and/or Google Street View. The benefit of using field data was considered greater than the drawback of using data from 2016 instead of 2017. The land cover was expected to remain stable over the time period of a year considering the fact that cacao is a perennial crop (Siebert, 2002). In addition, cacao plantations are generally profitable and therefore assumed to be relatively stable land covers (Duguma et al., 2001).

3.2 Validation

In the validation phase, the algorithms were tested on a new set of points, again collected in QGIS with the exception of the ground truth points for the class of *cacao* (Figure 3). The required number of validation points was calculated in R based on a desired confidence level of 2*2.5%, in accordance with Olofsson et al. (2014):

$$N = \frac{(\sum W_i S_i)^2}{[S(\hat{O})]^2 + (\frac{1}{N})\sum W_i S_i^2} \approx \left(\frac{\sum W_i S_i}{S(\hat{O})}\right)^2,$$

where N is the total number of points, S(Ô) is the standard error of the estimated overall accuracy (2.5%), W_i is the mapped proportion of the area of class i. S_i is the standard deviation of class i ($S_i = \sqrt{U_i(1-U_i)}$), where Ui is the precision of class i based on the test set. A proportional sampling design was used to calculate the number of validation points per class, with a minimum of 40 (Lillesand et al., 2015). Points were randomly sampled for each class. In case of unknown land cover or low image resolution, additional random sample points were added for the class in question, in order to ensure that the required number of validation points was maintained.

3.3 Remote sensing imagery

The imagery used for classification was from Sentinel-1 and Sentinel-2 (level-1C) (*Sentinel-2 MSI*, 2015). Composites of the years 2017 and 2018 were used in order to obtain at least 10 cloud-free images on average, which was necessary for a reasonable estimate of the standard deviation. Images with over 60% cloud cover were rejected. Including bands for 2 years was suboptimal because the training and validation data were (mostly) from 2017, yet acceptable as the land cover was not expected to change significantly within a time period of one year. After all, the classification is mostly concerned with perennial crops (Fold, 2008).

A stack of 40 bands was used for training. The 20th percentile, 50th percentile, 80th percentile, mean, and standard deviation were included for each of the following wavelengths: blue, green, red, near-infrared, and two short-wave infrared bands. In addition, slope and elevation were included, because both influence the growth of the crops included in the current classification (Läderach et al., 2013). Slope, elevation, and the medians of Sentinel-1 VV- and VH-polarized ascending bands were collected in interferometric wide swath (IW) mode (descending is not available for Ghana).

Finally, the 20th, 50th and 80th percentiles, mean, and standard deviation of the Normalized Difference Vegetation Index (NDVI) were included. The NDVI is calculated using the visible and near-infrared bands. It was included separately because the relationship between red and infrared is key to vegetation classification (Jia et al., 2014; Krishnaswamy et al., 2004). Feeding this information directly to the algorithms was thought to save time for training and calibration.

3.4 Pre-processing

Pre-processing was performed in the Python API of Google Earth Engine using Python 3.7. The study area was limited to a single swath in order to remove strong differences in illumination and cloud coverage between different swaths. Every band was normalized independently. Only images with less than 60% cloud cover were maintained and the lower and upper 20 percentiles were excluded in order to further reduce cloud distortion. For Sentinel-1, Google Earth Engine's SAR GRD quality mask was used (*Sentinel-1 SAR GRD*, n.d.). For Sentinel-2, the QA60 band was used for cloud masking (*Sentinel-2 MSI*, 2015). A topographic correction was applied on the Sentinel-2 data based on the cosine correction for slope (Tan et al., 2013).

4. Methods

This section provides a detailed overview of the methods that were used for this research. It starts with the pre-processing stage, followed by the parameterizations of the RF and the CNN. It then continues with a description of the processing steps in the testing phase, and subsequently the validation phase. This includes calculation of confidence intervals for precision and recall, and validation in the land cover classes on which the RF and CNN disagreed relatively often.

4.1 RF parameterization

A pixel-based supervised random forest algorithm was used on the Google Earth Engine cloud computing platform (Breiman, 2001). The number of trees was varied with a stable validation set of 800 points in order to determine the number of trees. An optimal fit was achieved with 100 trees. The number of variables per split was set to the square root of the number of variables, and the minimum size of a terminal node was 1.

4.2 CNN parameterization

The CNN was constructed in Java using the DL4J open source software and included 1 convolution, 1 max pooling layer, and 3 dense neural network layers. A moving window of 7x7 pixels was used, of which the central pixel was classified. See Appendix I for the full parameterization of the CNN.

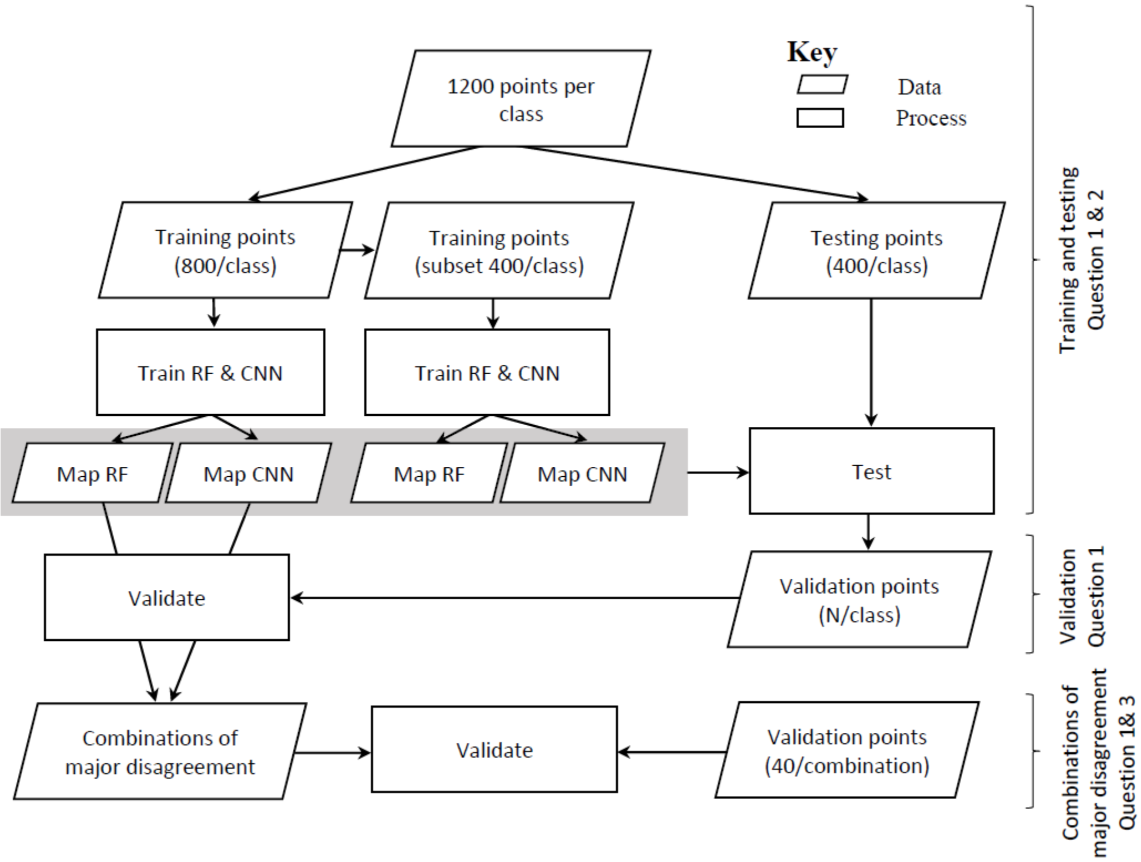
Key
Data
Process
1200 points per class
Training points (800/class)
Training points (subset 400/class)
Testing points (400/class)
Train RF & CNN
Train RF & CNN
Map RF
Map CNN
Map RF
Map CNN
Test
Validate
Validation points (N/class)
Combinations of major disagreement
Validate
Validation points (40/combination)
Training and testing Question 1 & 2
Validation Question 1
Combinations of major disagreement Question 1& 3

Figure 4. Succinct flowchart of the research methods.

4.3 Testing

The RF and CNN were both run with 400 and 800 training points per class (Figure 4 below). In both cases, the same test set of 400 points per class was used. Accuracy indicators of the resulting classified map were obtained with and without sieving, using a threshold of 10, and 8-connectedness. Whether sieving is desirable depends on the purpose of the project, as it reduces potential noise at the cost of resolution.

The results without sieving are more representative of the performance of the classifier and were therefore validated further. Raster values were sampled at the validation points in order to compute the confusion matrix. Accuracy metrics based on the test points were calculated in R (including overall accuracy, precision, recall, and F1-score).

4.4 Validation

Validation was performed following CEOS-WGCV level 3 requirements (*CEOS Working Group on Calibration and Validation*, 2019). The non-postprocessed maps resulting from training with 800 points per class were validated further (Figure 4). Random validation samples were taken for each class on the basis of a proportional sampling design. The total number of samples was calculated based on the proportional confusion matrix, where the initial test sample was used to estimate the accuracy. The estimated area proportions $\hat{p}_{ij}$ were calculated in accordance with the formula provided by Olofsson et al. (2014):

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i\cdot}},$$

where *W_i* is the proportion of area mapped as class i, $n_{i\cdot}$ is the total number of sample units of map class i, and n_{ij} refers to the number of samples of class j mapped as class i. Again, R was used to obtain accuracy metrics, including overall accuracy, precision, recall and F1-score.

4.4.1 Precision and recall

Confidence intervals were calculated for the precision and recall of the validation samples in accordance with (Olofsson et al., 2014). The test samples were not taken randomly, which implies that confidence intervals cannot be calculated for those results (Olofsson et al., 2014). For the precision of class i ($\hat{U}_i$), the estimated variance ($\hat{V}$) was calculated as follows:

$$\hat{V}(\hat{U}_i) = \hat{U}_i(1 - \hat{U}_i)/(n_{i\cdot} - 1),$$

where $n_{i\cdot}$ refers to the total number of sample units of map class i.

The estimated variance ($\hat{V}$) of the recall of reference class j $(\hat{P}_j)$ is:

$$\hat{V}(\hat{P}_j) = \frac{1}{\hat{N}_{\cdot j}} \left[\frac{N_{j\cdot}^2 (1-\hat{P}_j)^2 \hat{U}_j (1-\hat{U}_j)}{n_{j\cdot}-1} + \hat{P}_j^2 \sum_{i\neq j}^{q} N_{i\neq j}^2 N_{i\cdot}^2 \frac{n_{ij}}{n_{i\cdot}} \left(1 - \frac{n_{ij}}{n_{i\cdot}}\right) / (n_{i\cdot} - 1) \right],$$

where $\hat{N}_{\cdot j} = \sum_{i=1}^{q} \frac{N_{i\cdot}}{n_{i\cdot}} n_{ij}$ is the estimated marginal total number of pixels of reference class j, N$_j$ is the marginal total of map class j and $n_{j\cdot}$ is the total number of samples of map class j. The confidence intervals were calculated as $\pm 1.96\sqrt{\hat{V}(\hat{U}_i)}$, where $\hat{U}_i$ was replaced by $\hat{P}_j$ for the recall.

4.4.2 Classes of major disagreement

ArcMap 10.6.1 was used to analyse the prevalence of all the different combinations of classification results by the RF and CNN (trained with 800 points per class). The following raster calculation was performed to obtain this information: $10 * (map1 + 1) + (map2 + 1)$, where $map1$ was the classification result of the RF and $map2$ was the classification result of the CNN. This resulted in a two-digit number, of which the first represented the classification result of the RF, and the second represented the classification result of the CNN. The results were analysed in R. Special attention was paid to different classification results related to the class of *cacao*. This class is of special interest, because cacao production is among the most important drivers of deforestation in Ghana, while the spectral similarity between cacao plantations and natural forest is particularly high (DeVries et al., 2015; Duguma et al., 2001). The combinations of classification as *cacao* by one algorithm and a different class by the second were analysed. Combinations that occurred in more than 1% of the cases were validated more elaborately. For each of these combinations, 40 points were taken randomly in each of the 6 areas of disagreement (Lillesand et al., 2015). These points were again classified based on Hybrid Google, Google Earth, and/or Google Street View data from 2017 in order to assess which algorithm performed better in these areas.

5. Results

This section starts with the classification results, showing the performance of the RF and the CNN with 400 and 800 training points per class. In addition, the impact of sieving is shown. Next, the outcomes of the validation phase are presented, in which 800 independent new samples were taken for each class. Finally, the results of validation in the areas of major disagreement follow.

5.1 Testing

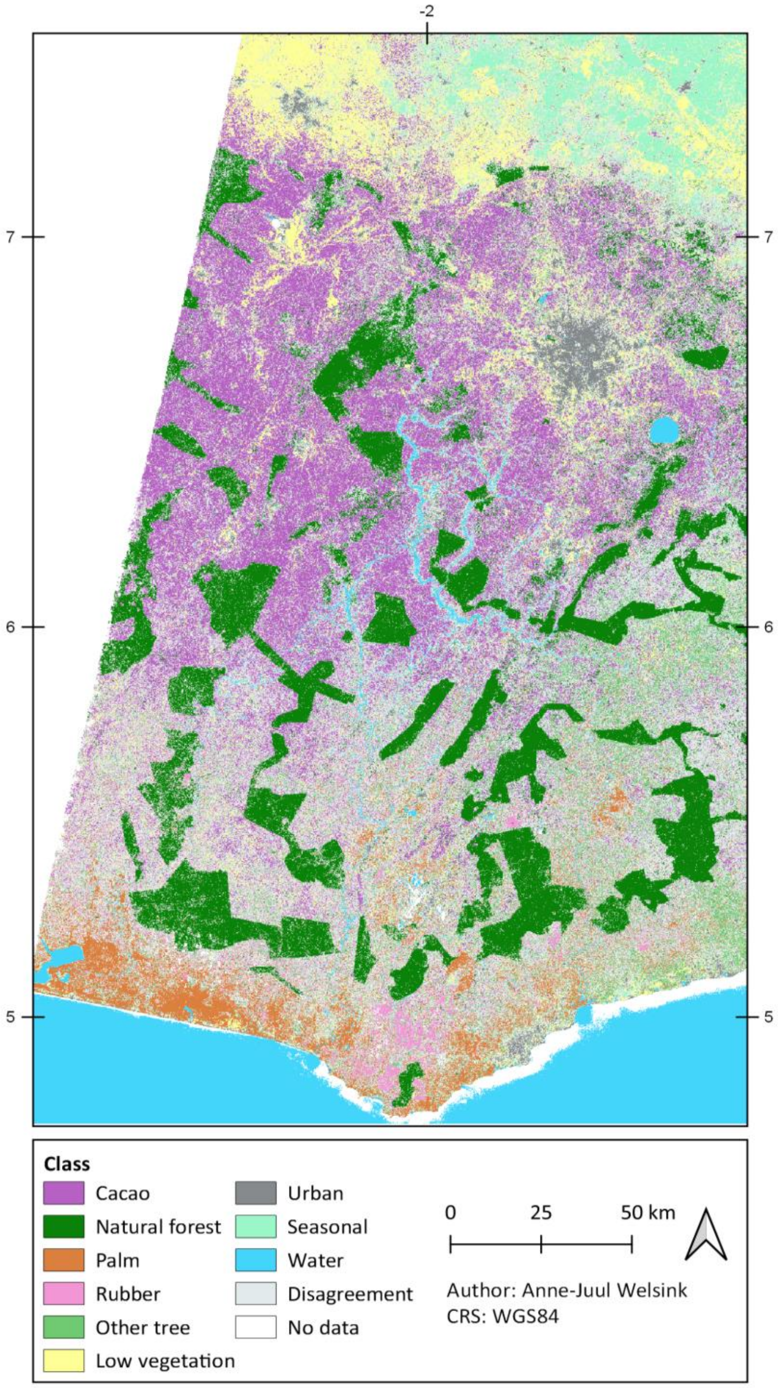

Figure 5. Areas in which the classification results of the RF and CNN correspond. Training was performed with 800 points per class. White areas inside the study area indicate disagreement between the RF and CNN, or a no-data value for either algorithm.

Table 1. Precision, recall, and F1-score (%) per class for the RF and CNN. These metrics are reported for training with 400 points per class and 800 points per class, with and without sieving. In addition, the overall accuracy of each training result is reported with the corresponding 95% confidence interval in brackets.

	Cacao	**Natural forest**	**Palm**	**Rubber**	**Other tree**	**Low vegetation**	**Urban**	**Seasonal**	**Water**
Random Forest									
400 points	Accuracy 81.2% (79.9, 82.5)								
Precision (%)	77.3	81.0	79.0	87.0	65.7	64.4	93.4	91.0	92.5
Recall (%)	78.3	86.3	65.8	86.8	64.8	69.9	91.8	91.3	96.0
F1 (%)	77.8	83.5	71.8	87.9	65.2	67.1	92.6	91.3	94.2
800 points	Accuracy: 82.5% (81.2, 83.7)								
Precision (%)	80.7	82.6	79.9	90.1	66.5	65.9	94.3	92.2	92.3
Recall (%)	81.5	87.8	68.5	88.5	68.0	72.2	91.0	89.0	96.0
F1 (%)	81.5	85.1	73.8	89.3	67.2	68.9	92.6	90.6	94.1
400 points, sieved	Accuracy: 82.8% (81.5.4, 84.0)								
Precision (%)	74.9	84.5	84.8	90.5	70.2	65.0	93.0	91.7	95.5
Recall (%)	83.5	91.0	65.8	88.5	68.8	71.2	89.8	91.3	95.5
F1 (%)	79.0	87.6	74.1	89.5	69.4	67.9	91.3	91.5	94.6
800 points, sieved	Accuracy: 83.1% (81.9, 84.3)								
Precision (%)	78.7	85.8	82.6	91.0	67.0	66.6	93.7	93.0	93.6
Recall (%)	83.3	89.3	67.8	88.0	71.0	74.4	89.5	89.5	95.5
F1 (%)	80.9	87.5	74.5	89.5	68.9	70.3	91.6	91.2	94.6
CNN									
400 points	Accuracy: 84.3% (83.0, 85.5)								
Precision (%)	85.1	88.7	87.6	82.4	69.2	69.7	96.1	93.1	93.0
Recall (%)	81.4	88.7	63.2	93.9	78.1	76.2	88.9	90.7	97.7
F1 (%)	83.2	88.7	73.4	87.8	73.4	72.8	92.3	91.9	95.3
800 points	Accuracy: 87.3% (86.1, 88.4)								
Precision (%)	81.6	96.1	83.1	90.1	78.0	76.4	94.0	92.5	94.0
Recall (%)	86.3	91.9	80.6	92.4	78.6	72.9	91.6	93.2	98.2
F1 (%)	83.9	94.0	81.8	91.2	78.3	74.6	92.8	92.9	96.1
400 points, sieved	Accuracy: 84.7% (83.4, 85.8)								
Precision (%)	84.1	89.2	88.6	83.9	69.4	70.3	96.3	93.3	93.3
Recall (%)	82.4	89.7	63.4	94.2	79.3	76.7	87.7	90.7	97.7
F1 (%)	83.3	89.5	73.9	99.7	74.0	73.4	91.8	92.0	95.4
800 points, sieved	Accuracy: 87.5% (86.3, 88.6)								
Precision (%)	80.9	96.3	84.5	90.5	78.1	76.7	94.2	92.8	93.9
Recall (%)	87.6	92.2	81.1	92.6	80.1	72.8	90.3	93.0	97.7
F1 (%)	84.1	94.2	82.8	91.6	79.1	74.7	92.2	92.9	95.8

5.1.1 Training with 400 points per class

The random forest, trained with 400 points per class, obtained an overall accuracy of 81.2% (95% CI: 79.9, 82.5) (Table 1; please refer to Appendix II for a graph of the precision, recall, and F1-scores). The CNN obtained a significantly higher accuracy of 84.3% (83.0, 85.5).

Precision is the probability of a reference pixel being correctly classified (Liu et al., 2018). The CNN had a higher precision than the RF in almost all classes, apart from *rubber* (RF: 87.0%, CNN: 82.4%) (Table 1). The difference in precision was bigger than 5% in the classes *cacao* (RF: 77.3%, CNN: 85.1%), *natural forest* (RF: 81.0%, CNN: 88.7%), and *palm* (RF: 79.0%, CNN: 87.6%).

Recall is the probability that the classification of a pixel represents the real land cover (Liu et al., 2018). The CNN also performed better than the RF in terms of recall, apart from the classification of *palm* (RF: 65.8%, CNN: 63.2%) and *urban* (RF: 91.8%, CNN: 88.9%). The difference between the two algorithms was bigger than 5% in the classes of *rubber* (RF: 86.8%, CNN: 93.9%), *other tree* (RF: 64.8%, CNN: 78.1%), and *low vegetation* (RF: 69.9%, CNN: 76.1%).

High differences between precision and recall are undesirable, because they indicate that the classification in a class is either quite exact but not complete (error of omission), or complete but not exact (error of commission). The biggest difference between the precision and recall within each algorithm was found in the *palm* class. The difference was almost 15% for the RF (precision: 79.0% recall: 65.8%) and 25% for the CNN (precision 87.6%, recall: 63.2%). The second biggest difference was 6% for the RF in *low vegetation* (precision: 64.4%, recall: 69.9%). For the CNN, this was 9% in *other tree* (precision: 69.2%, recall: 78.1%).

The F-1 score is defined as $2 * \frac{Precision*Recall}{Precision+Recall}$. The CNN obtained a higher F1-score in all classes, apart from *urban* (RF: 92.6%, CNN: 92.3%). The RF obtained the lowest F1-score in the classes of *other tree* (65.2%) and *low vegetation* (67.1%). The CNN obtained the lowest scores on *palm, other tree, and low vegetation*, each with an F1 rounded to 73%. Both algorithms obtained the highest scores on *urban* (RF: 92.6%, CNN: 92.3%), and *water* (RF: 94.2%, CNN: 95.3%). In addition, the CNN obtained a high F1-score on *seasonal* (91.9%).

5.1.2 Training with 800 points per class

Classification with 800 points per class resulted in an agreement among the RF and CNN in 66% of the pixels (Figure 5). The RF achieved an accuracy of 82.5% (81.2, 83.7) (Table 1). The CNN achieved an accuracy of 87.3% (86.1, 88.4); significantly higher than that of the RF.

The precision of the CNN was higher than the precision of the RF in all classes, apart from *rubber* (both rounded to 90%) (Table 1). The difference in precision was higher than 5% in the case of *natural forest* (RF: 82.6%, CNN: 96.1%), *other tree* (RF: 66.5%, CNN: 78.0%), and *low vegetation* (RF: 65.9%, CNN: 76.4%). The recall of the CNN was higher than that of the RF in all classes, apart from *urban*, where both algorithms achieved a rounded recall of 92%. A difference above 5% was found for *cacao* (RF: 78.3%, CNN: 86.3%), *palm* (RF: 68.5%, CNN: 80.6%), and *other tree* (RF: 68.0%, CNN: 78.6%).

For the RF, the difference in precision and recall was highest for *palm* (precision: 79.9%, recall: 68.5%), followed by *low vegetation* (precision: 65.9%, recall: 72.2%). For the CNN, the highest difference was 4%, found in the classes of *cacao* (precision: 81.6%, recall: 86.3%), *natural forest* (precision: 96.1%, recall 91.9%), and *water* (precision: 94.0%, recall: 98.2%).

The CNN obtained a higher F1-score in all classes (apart from a tie for urban). The RF obtained the lowest scores for *other tree* (67.2%) and *low vegetation* (68.9%) and the highest scores for *urban* (92.6%) and *water* (94.1%). The CNN also obtained the lowest scores for *other tree* (78.3%) and *low vegetation* (75.6%). The highest scores were achieved on *natural forest* (94.0%) and *water* (96.1%).

5.1.3 Sieving

Sieving resulted in an improved overall accuracy for both the RF and the CNN (Table 1). Based on training with 400 points per class, the accuracy of the RF increased from 81.2% to 83.1%. When trained with 800 points, the accuracy increased from 82.5% to 83.1%. Sieving improved the accuracy of the CNN from 84.3% to 84.7% when trained with 400 points. With 800 points, the accuracy improved from 87.3% to 87.5%.

5.2 Validation

Table 2. Result of validation based on a proportional random sampling design and training with 800 points per class. For both algorithms, the overall accuracy is reported with a corresponding 95% confidence interval in brackets. Precision, recall, and F1-score are reported for each class. Precision and recall have corresponding 95% confidence intervals (CI).

	Cacao	**Natural forest**	**Palm**	**Rubber**	**Other tree**	**Low vegetation**	**Urban**	**Seasonal**	**Water**
Random Forest	Accuracy: 75.6% (71.6, 79.2)								
Precision (%)	93.5	74.6	70.7	61.0	22.6	89.0	82.2	64.3	100.0
95% CI	(93.3, 93.6)	(74.2, 75.1)	(70.3, 71.2)	(60.4, 61.5)	(22.2, 23.0)	(88.8, 89.3)	(81.9, 82.6)	(63.8, 64.8)	(100.0, 100.0)
Recall (%)	63.8	78.1	55.8	100.0	85.7	69.2	88.1	96.4	100.0
95% CI	(63.6, 63.8)	(78.0, 78.3)	(55.5, 56.1)	(100.0, 100.0)	(85.4, 86.0)	(69.1, 69.2)	(87.1, 89.1)	(95.9, 96.9)	(100.0, 100.0)
F1 (%)	75.8	76.3	62.4	75.8	35.8	77.8	85.1	77.1	100.0
CNN	Accuracy: 71.4% (67.3, 75.3)								
Precision (%)	84.6	78.9	76.2	41.7	16.4	86.3	86.1	71.4	97.8
95% CI	(84.3, 84.9)	(78.5, 79.3)	(75.8, 76.6)	(41.1, 42.2)	(16.1, 16.7)	(86.0, 86.5)	(85.8, 86.3)	(71.0, 71.9)	(97.8, 97.9)
Recall (%)	52.0	83.6	60.4	80.0	66.7	65.1	90.2	93.8	100.0
95% CI	(51.9, 52.0)	(83.4, 83.8)	(60.1, 60.6)	(78.1, 81.9)	(65.8, 67.5)	(65.0, 65.2)	(88.9, 91.6)	(93.4, 94.1)	(100.0, 100.0)
F1 (%)	64.4	81.2	67.4	54.8	26.3	74.2	88.1	81.1	98.9

In the validation phase, the accuracies of the RF and CNN reduced compared to those based on the test set. The overall accuracy of the RF was now 75.6% (95% CI: 71.6, 79.2) (Table 2) while the CNN achieved 71.4% (67.4, 75.3), 5% lower than that of the RF. The RF tended to confuse *cacao* for *natural forest*, while the CNN confused *cacao* for *rubber* relatively often (see Appendix III for the confusion matrices of the RF and the CNN in the validation phase). Both algorithms frequently confused *cacao* for *other tree*, and *low vegetation* for *seasonal*.

The precision of the RF was significantly higher than that of the CNN for *cacao, rubber, other tree, low vegetation, and water* (Table 2). On the other hand, the precision of the CNN is significantly higher for *natural forest*, *palm*, *urban*, and *seasonal*. In each class, the algorithm with the best precision also achieved the highest recall, apart from *seasonal*. In the case of *water*, both achieved a recall of 100.0%.

The differences between precision and recall were bigger in the validation phase than in the classification results as presented in Table 1. For both the RF and the CNN, the precision was at least 15% higher than the recall for the classes of *cacao*, *palm,* and *low vegetation*. Recall was more than 15% higher than precision for *rubber, other tree*, *and seasonal*.

The RF obtained a higher F1-score than the CNN for *cacao, rubber, other tree, low vegetation,* and *water.* The CNN obtained a higher score for *natural forest, palm, urban,* and *seasonal.* As expected, these are the same classes for which the CNN obtained a higher precision and recall than the RF. Both algorithms performed relatively well on the classification of *cacao, low vegetation, urban,* and *water.*

5.3 Areas of major disagreement

Table 3. Matrix that indicates what percentage of the total number of pixels was classified as class i by the RF and class j by the CNN. The combinations of [i,j] where either i or j was cacao and i ≠ j that occur in more than 1% of the pixels are highlighted in ***bold.*** *The combinations on which both algorithms agree are found on the diagonal and are highlighted in italics.*

		CNN								
		Cacao	**Natural forest**	**Palm**	**Rubber**	**Other tree**	**Low vegetation**	**Urban**	**Seasonal**	**Water**
RF	**Cacao**	*16.2*	0.8	**1.3**	0.8	**3.1**	**1.4**	0.1	0.0	0.0
	Natural forest	**1.8**	*12.5*	0.9	0.5	1.6	0.1	0.0	0.0	0.0
	Palm	0.8	0.7	*4.1*	0.3	1.1	0.2	0.0	0.0	0.1
	Rubber	0.3	0.3	0.6	*1.7*	0.8	0.2	0.0	0.0	0.0
	Other tree	**2.4**	0.4	2.1	0.5	*5.4*	1.8	0.1	0.1	0.1
	Low vegetation	**1.7**	0.0	0.6	0.2	2.1	*10.4*	0.6	1.8	0.2
	Urban	0.0	0.0	0.0	0.0	0.0	0.1	*2.0*	0.0	0.2
	Seasonal	0.0	0.0	0.0	0.0	0.1	0.6	0.1	*4.1*	0.0
	Water	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	*9.7*

The RF classified pixels as *cacao* while the CNN classified them as *palm* in 1.3% of the map. The combination of *cacao* and *other tree* occurred in 3.08%, and the combination of *cacao* and *low vegetation* 1.4%. Other combinations where the RF classified land cover as *cacao* and the CNN gave a different result [*cacao, i*] *occurred* less than 1% of the time.

The RF classified pixels as *natural forest* while the CNN classified them as *cacao* in 1.8% of the total pixel count. The combination of *other tree* and *cacao* occurred in 2.4%, and the combination of *low* *vegetation* and *cacao* occurred in 1.7%. Other combinations of [*i*, *cacao*] occurred in less than 1% of the map.

5.3.1 Validation in the areas of major of disagreement

Table 4. Percent of times the RF and CNN were correct during validation in each combination of major disagreement related to cacao, based on 40 extra validation points per combination of major disagreement. Both algorithms were correct in case of mixed canopy (e.g. cacao and palm trees planted in the same area). The proportion that both were correct should be added to the proportion that a single algorithm was correct in order to get the total proportion of validation points that was correctly classified by the algorithm.

	RF	**CNN**	**Both correct**	**Neither correct**
	Cacao	**Palm**		
%	40.0	37.5	22.5	0.0
	Cacao	**Other tree**		
%	66.7	7.1	4.8	21.4
	Cacao	**Low vegetation**		
%	51.1	23.4	23.4	2.1
	Natural forest	**Cacao**		
%	6.3	87.5	2.1	4.2
	Other tree	**Cacao**		
%	2.3	77.3	4.6	15.9
	Low vegetation	**Cacao**		
%	23.9	39.1	30.4	6.5

In all combinations of disagreement, the algorithm that classified the area as *cacao* was correct more often (Table 4). This is in line with the fact that the precision on *cacao* for both the RF and the CNN was significantly higher than the recall (Table 2). The algorithm that classified *cacao* over an alternative land cover was considered correct at least 15% more often than the other algorithm in each but one class. Only the classification as *cacao* by the RF (40.0% correct) and *palm* by the CNN (37.5% correct) was more evenly spread.

Both algorithms were correct relatively often in the combinations of *cacao* (RF) and *palm* (CNN) (22.5%), *cacao* (RF) and *low vegetation* (CNN) (23.4%) and *low vegetation* (RF) and *cacao* (CNN) (30.4%). In these cases, mixed land cover was present, which implies that in spite of disagreements in the classifications, to some extent both maps correctly represent the land cover in these areas. It occurred relatively frequently that neither algorithm classified correctly in the combinations of *cacao* (RF) and *other tree* (CNN) (21.4%) and *other tree* (RF) and *cacao* (CNN) (15.9%). This is explained by the fact that most new cacao plantations are planted in thinned forest with different types of trees, which are planted or maintained to provide shade to the subcanopy (Clough et al., 2009).

6. Discussion

This section interprets the findings that were presented in the previous section in light of existing research, and provides insight into the questions that were highlighted in the introduction. The first part of this section compares and explains the performance of the RF and CNN on land cover classification with 400 and 800 points in the testing phase, followed by an interpretation of the results in the validation phase. Next, the results of validation in the classes of major disagreement are discussed and interpreted. This section ends by highlighting a number of limitations of this study, and suggestions for further research.

6.1 The size of the training dataset

Previous research has shown that the RF has a relatively high accuracy on large scale studies compared to other traditional algorithms (Pelletier et al., 2019). Yet, CNNs can outperform RFs in classification efforts, as they can be more finely tuned than a RF through their many hyper-parameters (Helber et al., 2018; Huang et al., 2018; Pelletier et al., 2019). In the testing phase of this research, the CNN indeed achieved a higher overall accuracy than the RF, as well as a higher precision, recall, and F1-score in most classes (Table 1). This was the case for both training with 400 and 800 points. The results of both the RF and the CNN were better with more training points.

The CNN was more sensitive to the amount of training data than the RF. This can be explained by the fact that a CNN can achieve a very good fit, but is highly biased towards the training data; the algorithm is therefore prone to overfitting (Pelletier et al., 2019). The fact that the accuracy of the CNN saw a bigger improvement than the RF with 800 instead of 400 training points per class can be explained through the same line of reasoning. An increase in the number of training points had a bigger impact on the CNN than on the RF, as the RF had neared its potential more closely with fewer data. This is in line with the finding of Liu et al. (2018) in their study on wetland mapping, in which the RF also generalized better than a CNN. The high accuracies that are frequently reported for CNNs are generally achieved with a higher number of training points and through iterative sampling (Huang et al., 2018).

An additional difference between the RF and the CNN is the fact that the RF is a pixel-based method, while the CNN applies a moving-window approach (Zhang et al., 2019). As a result, speckle noise and intra-class variance have a higher influence on the classification accuracy of the RF than that of the CNN. Sieving reduces this influence, but also limits the level of spatial detail in the classification. Sieving resulted in an improvement of the accuracy of the RF and the CNN, although the improvement in the CNN was somewhat smaller than that of the RF. After all, the RF is more prone to noise as it classifies individual pixels without taking the neighbourhood into account, as opposed to the CNN. Whether the loss of spatial detail that results from sieving is desirable depends on the purpose of the study and the available training data. Besides post-processing, more elaborate pre-processing could help to reduce noise caused by clouds and atmospheric effects.

6.2 Validation

In the validation phase, the accuracy of the RF was higher than that of the CNN (Table 2). The RF achieved a better precision, recall, and F1-score (which is based on the former two) on *cacao, rubber, other tree, low vegetation*, and *water*. The CNN achieved better results for *natural forest, palm, urban,* and *seasonal*. Overall, the CNN performed more poorly than the RF in the validation phase (although not significantly). The relatively low accuracy of the CNN is likely due to its proneness to overfitting (Pelletier et al., 2019). The risk of overfitting is higher with little training data and in case of high intra-class variability (Yang et al., 2018). High intra-class variability implies that the spectral signature of a particular land cover is relatively unstable, which complicates the delineation of a class. The validation stage was based on a fresh, random sample. The RF generalized better than the CNN, resulting in better performance on these independent validation data. This is in line with research by Liu et al. (2018), which emphasizes that CNNs are prone to overfitting with limited training data. With even less training data, the RF may gradually start to further increase its advantage in relation to the CNN. The fact that CNNs outperform RFs in most existing literature is due to the fact that most previous research has used a larger amount of training data (Helber et al., 2018; Huang et al., 2018). According to Liu et al. (2018), none of these studies have used a varying sample size in the comparison of CNNs and conventional classifiers.

Besides general class accuracies, precision and recall are of interest. Precision is the complement of the omission error (100% - omission), while recall is the complement of the commission error (100% - commission). For both algorithms, the precision was more than 15% higher than the recall for *cacao*, *palm*, and *low vegetation* for both the RF and the CNN. These classes occupy a large proportion of the classified area, and therefore tend to have a relatively high commission error (Figure 1) (Olofsson et al., 2012). Conversely, *rubber, other tree*, and *seasonal* occupied relatively small areas in the classification and tended to be underrepresented (omitted) (Figure 1). Large differences between precision and recall are undesirable, because they may indicate a structural flaw in the (training of) the model. If it is undesirable to miss deforestation events, it is less problematic when the recall of drivers of deforestation is higher than the precision than the other way around. In that case, deforestation may be suspected in some areas where it is not present, rather than it being overlooked. In the current classification, it is particularly undesirable that the precision was (much) higher than the recall for the classes of *cacao* and *palm*. Iterative sampling could reduce these differences and further improve the accuracy of the models (Tuia et al., 2011).

Both algorithms performed relatively well on *cacao, low vegetation, urban,* and *water*. The RF performed better than the CNN on the classes of *cacao*, *rubber*, *other tree, low vegetation*, and *water*, while the CNN performed best on *natural forest, palm, urban, and seasonal*. Further research on the structure of the input data and the importance of different explanatory variables could lead to a better understanding of the large differences in some class accuracies. The RF may perform better in classes with high intra-class variance, while the CNN is perhaps favourable in case of high spectral similarity, due to its tendency to fit more closely to the training data (DeVries et al., 2015; Pelletier et al., 2019). The RF is will likely outperform the CNN if the variety in the spectral signature within a single class is high, since the CNN is suffers more from overfitting (Pelletier et al., 2019).

In the validation phase, the accuracy of both the RF and the CNN decreased compared to that of the testing phase (Table 2). This can be partly explained by the fact that the test data were an independent subset of the training data, while the validation set was collected separately as a subset of the reference data. Therefore, the test set was likely more biased towards the training set, which resulted in a higher accuracy than in the validation phase. A second explanation for the decrease in accuracy is the fact that the validation points were taken randomly, as opposed to the test set. The land cover at these random points was more often mixed than that in the training and testing phase, for which uniform land covers were selected relatively often.

6.3 Areas of major disagreement

In the areas of major disagreement related to *cacao*, the classes in which the choice of algorithm mattered most were characterized by a large difference in the accuracy, and a low percentage of cases in which both algorithms were correct (indicating mixed land cover) (Table 4). This was the case for the combinations of *cacao* (RF) and *other tree* (CNN), *natural forest* (RF) and *cacao* (CNN), and *other tree* (RF) and *cacao* (CNN). In all combinations of disagreement, the algorithm that classified the area as *cacao* was correct more often than the alternative (Table 4). This is explained by the fact that the precision on *cacao* for both the RF and the CNN was significantly higher than the recall (Table 2). Further validation is required in order to assess to what extent this is due to sampling bias.

Overall, validation in the classes of disagreement did not structurally favour one algorithm over the other. Previous research has shown that CNNs have the potential to outperform RFs when a large quantity of training data is available and when the hyper-parameters are optimized well (Helber et al., 2018; Pelletier et al., 2019). A small training dataset, erroneous data, or spectral limitations in satellite imagery seem to have had a relatively high impact on the classification accuracy of the CNN compared to the RF, due to its more flexible fit (Liu et al., 2018). The RF performed comparable to the CNN in spite of its inferior theoretical potential (Helber et al., 2018; Huang et al., 2018; Liu et al., 2018). Therefore, researchers should consider the size and quality of their datasets before investing in the complex parameterization of a CNN.

6.4 Limitations and further research

A number of limitations require attention. First of all, In the pre-processing stage, only basic cloud detection was performed. More elaborate cloud rejection could improve the quality of the image. In addition to that, inclusion of temporal metrics could improve land cover classification (Baamonde et al., 2019; Eberenz et al., 2016). This would foster recognition of crops based on their temporal variability, which is influenced by the seasonality of the Ghanaian climate (Lieberman, 1982). Future endeavours should account for the temporal variability of crops beyond the standard deviation.

Furthermore, the classification results are directly influenced by the quality of the input data. Cloud cover is highly prevalent in West-Africa, which limits the number of images that can be used for verification. In this study, the land cover in West-Ghana was assumed to be rather stable, as mostly perennial crops were included in the analysis (Fold, 2008; Siebert, 2002). At the same time, land cover changes, notably forest fires and cutting, are quick and common (Saatchi et al., 2001). Therefore, training data may quickly be outdated and classification errors result. Moreover, training points are most likely not always accurate because crops such as cacao are often planted under the shade of native trees, which makes it difficult to distinguish them from forest visually (Saatchi et al., 2001).

The training points in this project were collected based on a non-random sampling strategy. Although this speeds up the collection of training points, mixed and unknown land covers are likely to be underrepresented, as they could not be clearly marked as one of the land cover categories that were included. This problem could be mitigated through the use of a random sampling strategy, although this requires an extra time investment to classify unknown or mixed landcovers. Unfortunately, sub-optimal data is all we have access to at this point. The sampling error is likely to be non-negligible and should be included explicitly in further research.

That being said, it is important to emphasize that the findings of this research cannot be generalized. Future research should investigate how the algorithms perform with an incrementally increasing number of training points and rounds of iterative training data collection (Liu et al., 2018). Comparison of the classification results of a RF and CNN with different amounts and qualities of input data provides insight into the performance of these algorithms under various circumstances. In addition, comparisons between different study areas could provide insight into the respective performance of the algorithms with different types of land cover. This can help to inform the selection of the most appropriate algorithm, given a set of available training data.

7. Conclusion

This study has compared the performance of a RF and a CNN on land cover classification with 9 classes in West Africa using two different sets of training data (400 and 800 points per class). Remotely sensed imagery from Sentinel-1 (IW) and Sentinel-2 formed the basis for the classification. The results have shown that the CNN performed better than the RF in the testing phase for both training with 400 and 800 points. Sieving resulted in a slight improvement of the results of both the RF and the CNN. The results of the RF improved a little more, as its pixel-based approach is more prone to noise than the moving-window of the CNN.

The RF outperformed the CNN in the validation phase. The RF was found superior on the classes of *cacao*, *rubber*, *other tree, low vegetation*, and *water*, while the CNN performed best on *natural forest, palm, urban, and seasonal*. Both algorithms performed relatively well on the classification of the classes of *cacao, low vegetation, urban,* and *water.* Further research on the structure of the input data and the importance of different explanatory variables could lead to a better understanding of the large differences in some class accuracies.

The classes that were most dependent upon the choice of algorithm were *cacao, other tree*, and *natural forest*. The combinations on which the RF and CNN disagreed most frequently were validated more elaborately with 40 points per area of major disagreement (where either algorithm classified the area as *cacao*). The algorithm that classified as *cacao* was correct more often than the alternative in each combination. Apart from that, neither algorithm structurally performed better than the alternative.

Overall, the RF generalized relatively well in the validation phase, while the CNN was more prone to overfitting due to its higher bias towards the training set (Pelletier et al., 2019). Thus, even though the CNN may have a higher potential accuracy, use of a RF may be favourable when little training data is available and/or when intra-class variability is high. The fact that neither algorithm structurally performed best in the areas of disagreement reinforces the point that investment in the parameterization of a CNN may not be worthwhile when little or low-quality input data is used.

Data acquisition is a major challenge in deforestation monitoring and the remote sensing domain in general (Helber et al., 2018; Scott et al., 2017). Research into the performance of algorithms with small or varying amounts of input data can guide the choice of a suitable algorithm in remote sensing-based research. Liu et al. (2018) were the first to account for the amount of input data in their comparison of a CNN and more traditional algorithms. Their study focussed on object-based wetland mapping. This study has built on their work by comparing the performance of a CNN and RF in Ghana’s complex agroforestry landscape. It has provided insight into the performance of the RF and CNN with limited input data, which puts the presumed superiority of the CNN into perspective (Helber et al., 2018; Huang et al., 2018).

References

Anderman, T. L., Remans, R., Wood, S. A., DeRosa, K., & DeFries, R. S. (2014). Synergies and tradeoffs between cash crop production and food security: A case study in rural Ghana. *Food Security*, *6*(4), 541–554. https://doi.org/10.1007/s12571-014-0360-6

Appiah, M., Blay, D., Damnyag, L., Dwomoh, F. K., Pappinen, A., & Luukkanen, O. (2009). Dependence on forest resources and tropical deforestation in Ghana. *Environment, Development and Sustainability*, *11*(3), 471–487. https://doi.org/10.1007/s10668-007-9125-0

Asubonteng, K., Pfeffer, K., Ros-Tonen, M., Verbesselt, J., & Baud, I. (2018). Effects of Tree-crop Farming on Land-cover Transitions in a Mosaic Landscape in the Eastern Region of Ghana. *Environmental Management*, *62*(3), 529–547. https://doi.org/10.1007/s00267-018-1060-3

Baamonde, S., Cabana, M., Sillero, N., Penedo, M., Naveira, H., & Novo, J. (2019). Fully automatic multi-temporal land cover classification using Sentinel-2 image data. *Procedia Computer Science*, *159*, 650–657.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bunn, C., Läderach, P., Quaye, A., Muilerman, S., Noponen, M. R. A., & Lundy, M. (2019). Recommendation domains to scale out climate change adaptation in cocoa production in Ghana. *Climate Services*, *16*, 100123. https://doi.org/10.1016/j.cliser.2019.100123

CEOS Working Group on Calibration and Validation. (2019). Land Product Validation. https://lpvs.gsfc.nasa.gov/

Chiti, T., Grieco, E., Perugini, L., Rey, A., & Valentini, R. (2014). Effect of the replacement of tropical forests with tree plantations on soil organic carbon levels in the Jomoro district, Ghana. *Plant and Soil*, *375*(1–2), 47–59. https://doi.org/10.1007/s11104-013-1928-1

Clough, Y., Faust, H., & Tscharntke, T. (2009). Cacao boom and bust: Sustainability of agroforests and opportunities for biodiversity conservation. *Conservation Letters*, *2*(5), 197–205. https://doi.org/10.1111/j.1755-263X.2009.00072.x

DeFries, R., Achard, F., Brown, S., Herold, M., Murdiyarso, D., Schlamadinger, B., & Souza, C. de; (2006). *Reducing greenhouse gas emissions from deforestation in developing countries: Considerations for monitoring and measuring* (No. 46; Report of the Global Terrestrial Observing System (GTOS)). https://www.cifor.org/library/2526/

DeVries, B., Verbesselt, J., Kooistra, L., & Herold, M. (2015). Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat time series. *Remote Sensing of Environment*, *161*, 107–121. https://doi.org/10.1016/j.rse.2015.02.012

Duguma, B., Gockowski, J., & Bakala, J. (2001). Smallholder Cacao (Theobroma cacao Linn.) cultivation in agroforestry systems of West and Central Africa: Challenges and opportunities. *Agroforestry Systems*, *51*(3), 177–188. https://doi.org/10.1023/A:1010747224249

Eberenz, J., Verbesselt, J., Herold, M., Tsendbazar, N.-E., Sabatino, G., & Rivolta, G. (2016). Evaluating the Potential of PROBA-V Satellite Image Time Series for Improving LC Classification in Semi-Arid African Landscapes. *Remote Sensing; Basel*, *8*(12), 987. http://dx.doi.org.ezproxy.library.wur.nl/10.3390/rs8120987

FAO. (2015). *Global forest resources assessment 2015.* Food and Agriculture Organization of the United Nations.

Fold, N. (2008). Transnational Sourcing Practices in Ghana’s Perennial Crop Sectors. *Journal of Agrarian Change*, *8*(1), 94–122. https://doi.org/10.1111/j.1471-0366.2007.00164.x

Förster, J. (2009). *The Potential of Reducing Emissions from Deforestation and Degradation (REDD) in Western Ghana*. University of Bayreuth.

Hawthorne, W. D. (1989). *The flora and vegetation of Ghana’s forests* (Ghana Forest Inventory Project Seminar Proceedings, pp. 8–14). Forestry Department.

https://scholar.google.com/scholar_lookup?title=The%20flora%20and%20vegetation%20of%20Ghana%E2%80%99s%20forests&author=WD.%20Hawthorne&publication_year=1989

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2018). Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification—IEEE Conference Publication. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing Abbreviation*, *12*(7), 2217–2226.

Huang, B., Zhao, B., & Song, Y. (2018). Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, *214*, 73–86. https://doi.org/10.1016/j.rse.2018.04.050

Jia, K., Liang, S., Wei, X., Yao, Y., Su, Y., Jiang, B., & Wang, X. (2014). Land Cover Classification of Landsat Data with Phenological Features Extracted from Time Series MODIS NDVI Data. *Remote Sensing*, *6*(11), 11518–11532. https://doi.org/10.3390/rs61111518

Krishnaswamy, J., Kiran, M. C., & Ganeshaiah, K. N. (2004). Tree model based eco-climatic vegetation classification and fuzzy mapping in diverse tropical deciduous ecosystems using multi-season NDVI. *International Journal of Remote Sensing*, *25*(6), 1185–1205. https://doi.org/10.1080/0143116031000149989

Kroeger, A., Bakhtary, H., Haupt, F., & Streck, C. (2017). *Eliminating Deforestation from the Cocoa Supply Chain*. World Bank. https://doi.org/10.1596/26549

Läderach, P., Martinez-Valle, A., Schroth, G., & Castro, N. (2013). Predicting the future climatic suitability for cocoa farming of the world's leading producer countries, Ghana and Côte d'Ivoire. *Climatic Change*, *119*(3), 841–854. https://doi.org/10.1007/s10584-013-0774-8

Lieberman, D. (1982). Seasonality and Phenology in a Dry Tropical Forest in Ghana. *Journal of Ecology*, *70*(3), 791–806. JSTOR. https://doi.org/10.2307/2260105

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote Sensing and Image Interpretation, 7th Edition | Wiley*. John Wiley & Sons. https://www.wiley.com/en-us/Remote+Sensing+and+Image+Interpretation%2C+7th+Edition-p-9781118343289

Liu, Abd-Elrahman, A., & Wilhelm, V. L. (2018). Comparing Fully Convolutional Networks, Random Forest, Support Vector Machine, and Patch-based Deep Convolutional Neural Networks for Object-based Wetland Mapping using Images from Small Unmanned Aircraft System. *GIScience & Remote Sensing*, *55*(2), 243–264.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., Newell, J. D., Friedl, M. A., & Herold, M. (2012). A global land-cover validation data set, part I: Fundamental design principles. *International Journal of Remote Sensing*, *33*(18), 5768–5788. https://doi.org/10.1080/01431161.2012.674230

Pelletier, C., Webb, G., & Petitjean, F. (2019). Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series—ProQuest. *Remote Sensing*, *11*(5). https://search-proquest-com.ezproxy.library.wur.nl/docview/2303996454?OpenUrlRefId=info:xri/sid:wcdiscovery&accountid=27871

Rice, R. A., & Greenberg, R. (2000). Cacao Cultivation and the Conservation of Biological Diversity. *AMBIO: A Journal of the Human Environment*, *29*(3), 167–173. https://doi.org/10.1579/0044-7447-29.3.167

Saatchi, S., Agosti, D., Alger, K., Delabie, J., & Musinsky, J. (2001). Examining Fragmentation and Loss of Primary Forest in the Southern Bahian Atlantic Forest of Brazil with Radar Imagery. *Conservation Biology*, *15*(4), 867–875. https://doi.org/10.1046/j.1523-1739.2001.015004867.x

Scott, G. J., England, M. R., Starms, W. A., Marcum, R. A., & Davis, C. H. (2017). Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery.

IEEE Geoscience and Remote Sensing Letters, *14*(4), 549–553. https://doi.org/10.1109/LGRS.2017.2657778

Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected, log scaling. (n.d.). Google Developers. Retrieved 16 January 2020, from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S1_GRD?hl=nl

Sentinel-2 MSI: MultiSpectral Instrument, Level-1C. (2015). Google Developers. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2?hl=nl

Siebert, S. F. (2002). From shade- to sun-grown perennial crops in Sulawesi, Indonesia: Implications for biodiversity conservation and soil fertility. *Biodiversity & Conservation*, *11*(11), 1889–1902. https://doi.org/10.1023/A:1020804611740

Tan, B., Masek, J. G., Wolfe, R., Gao, F., Huang, C., Vermote, E. F., Sexton, J. O., & Ederer, G. (2013). Improved forest change detection with terrain illumination corrected Landsat images. *Remote Sensing of Environment*, *136*, 469–483. https://doi.org/10.1016/j.rse.2013.05.013

Tondoh, J. E., Kouamé, F. N., Martinez Guéi, A., Sey, B., Wowo Koné, A., & Gnessougou, N. (2015). Ecological changes induced by full-sun cocoa farming in Côte d'Ivoire. *Global Ecology and Conservation*, *3*, 575–595. https://doi.org/10.1016/j.gecco.2015.02.007

Tuia, D., Volpi, M., Copa, L., Kanevski, M., & Munoz-Mari, J. (2011). A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing*, *5*(3), 606–617. https://doi.org/10.1109/JSTSP.2011.2139193

Yang, C., Rottensteiner, F., & Heipke, C. (2018). Classification of land cover and land use based on convulutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science*, *4*(3). https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-3/251/2018/

Yaro, J. A., Teye, J. K., & Torvikey, G. D. (2017). Agricultural commercialisation models, agrarian dynamics and local development in Ghana. *The Journal of Peasant Studies*, *44*(3), 538–554. https://doi.org/10.1080/03066150.2016.1259222

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2019). Joint Deep Learning for land cover and land use classification. *Remote Sensing of Environment*, *221*, 173–187. https://doi.org/10.1016/j.rse.2018.11.014

Zhong, L., Hu, L., & Zhou, H. (2019). Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, *221*, 430–443. https://doi.org/10.1016/j.rse.2018.11.032

Appendix I. Parameterization of the CNN

Software: Deeplearning4j 1.0.0 beta 5

Input 7X7X39 (39 layers)

Convolution (1) with:

Convolution size: 4X4

Stride: 2X2

Padding: no padding

N mask: 128

Max pooling layer (1):

Size 2X2

Stride 2X2

Dense layers (3):

Unit: 2048

Drop-out survival rate: 0.6

General parameters:

Regularisation L2: 0.001*0.0020

Initial bias:1e-2

Parameter update: Adam(0.018*1e-2)

Bias updater: Adam(0.018*2*1e-2)

Loss Function: NEGATIVELOGLIKELIHOOD

Last layer activation: Activation function: SOFTMAX

Hidden Layer Activation Function: RELU

Gradient Normalization: Gradient Normalization, RenormalizeL2PerLayer

Optimization Algorithm: Optimization Algorithm STOCHASTIC_GRADIENT_DESCENT

Initial weight distribution: Weight Init Xavier

Appendix II. Graph of Precision, Recall, and F1-score

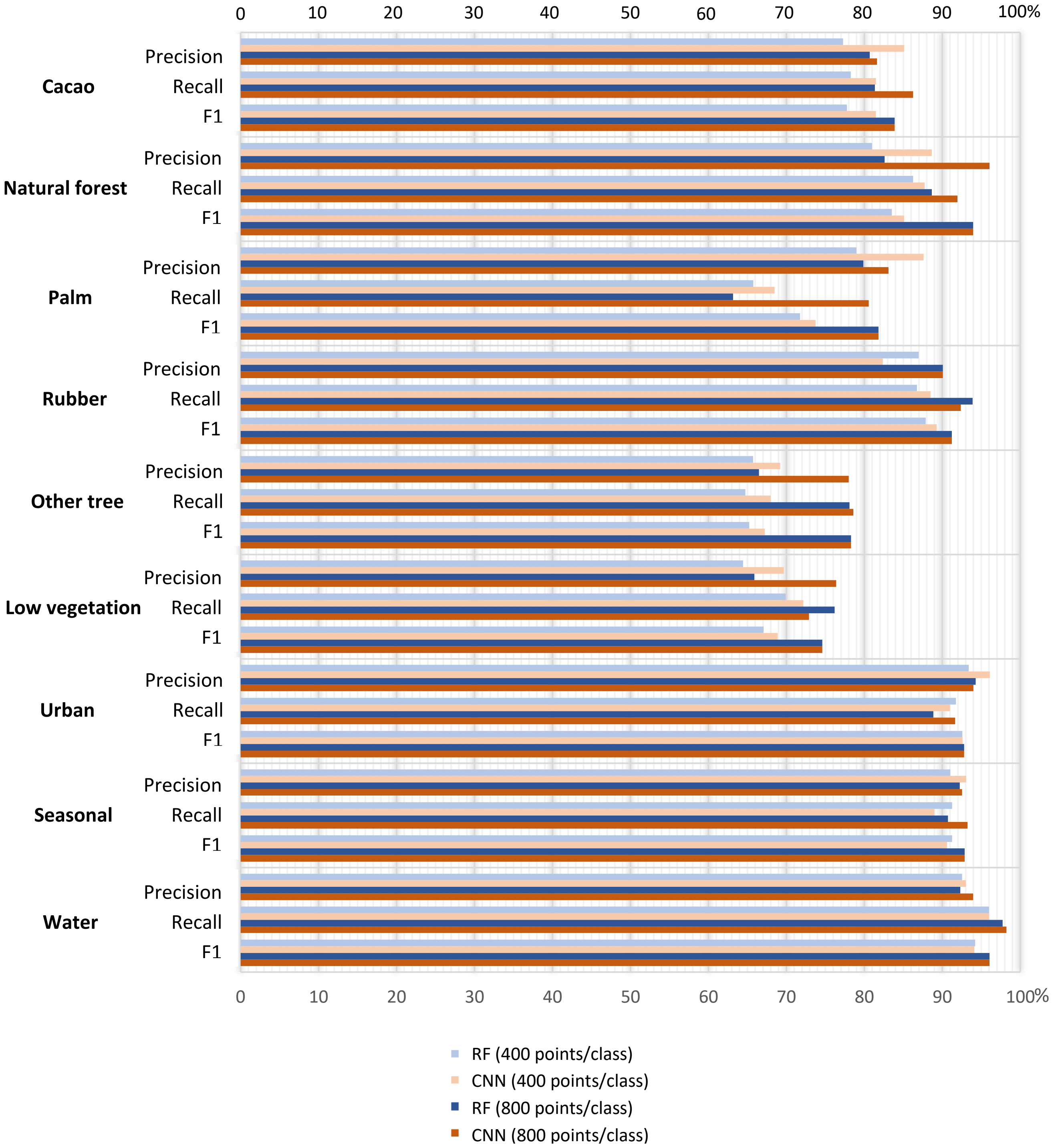

Figure 6. Precision, recall, and F1-score (%) per class for the random forest and CNN. These metrics are reported for training with 400 points per class and 800 points per class.

Appendix III. Confusion Matrices of the RF and CNN in the Validation Phase

Table 5. Confusion matrix of the results of the RF in the validation phase, where the classification results are indicated on the rows, and the ground truth is represented in the columns. The number of validation points per class was calculated based on the proportional confusion matrix (see p. 9).

		Reference								
		Cacao	**Natural forest**	**Palm**	**Rubber**	**Other tree**	**Low vegetation**	**Urban**	**Seasonal**	**Water**
Prediction	**Cacao**	*86*	2	2	0	0	2	0	0	0
	Natural forest	11	*50*	6	0	0	0	0	0	0
	Palm	4	6	*29*	0	1	1	0	0	0
	Rubber	9	1	3	*25*	1	2	0	0	0
	Other tree	20	5	10	0	*12*	6	0	0	0
	Low vegetation	5	0	1	0	0	*65*	1	1	0
	Urban	0	0	0	0	0	7	*37*	0	0
	Seasonal	0	0	0	0	0	11	4	*27*	0
	Water	0	0	0	0	0	0	0	0	*47*

Table 6. Confusion matrix of the results of the RF in the validation phase, where the classification results are indicated on the rows, and the ground truth is represented in the columns. The number of validation points per class was calculated based on the proportional confusion matrix (see p. 9).

		Reference								
		Cacao	**Natural forest**	**Palm**	**Rubber**	**Other tree**	**Low vegetation**	**Urban**	**Seasonal**	**Water**
Prediction	**Cacao**	*66*	2	4	0	2	4	0	0	0
	Natural forest	10	*56*	2	2	0	1	0	0	0
	Palm	6	3	*32*	1	0	0	0	0	0
	Rubber	14	4	4	*20*	3	3	0	0	0
	Other tree	23	1	11	1	*10*	11	3	1	0
	Low vegetation	7	1	0	1	0	*69*	1	1	0
	Urban	1	0	0	0	0	5	*37*	0	0
	Seasonal	0	0	0	0	0	12	0	*30*	0
	Water	0	0	0	0	0	1	0	0	*45*