SCALABLE DISTRIBUTED RANDOM FOREST CLASSIFICATION FOR PADDY RICE MAPPING ASIAN CONFERENCE ON REMOTE SENSING ACRS 2019

Dr. L. Nguyen[1], Dr. T. Hoang[1], Dr. P. Tran[1], Dr. Q. Pham[1]*

[1] Faculty of Medicine and Pharmacy, Ho Chi Minh City University of Medicine and Pharmacy, Ho Chi Minh City, Vietnam

KEY WORDS: Sentinel Missions, Machine Learning, Big Data, Earth Observation, High Performance Data Analytics

ABSTRACT: The present work[1] deals with the satellite based monitoring of agriculture, and specifically using Sentinel data, for the purposes of food security monitoring in South Korea. South Korea's food security concerns have to do with the overproduction of rice and the low self-sufficiency in the production of other major grains. For this reason the systematic and large scale monitoring of the paddy rice extent has been identified as key knowledge for the high-level decision-making in regard to food security. This work addresses the Big Data implications, derived from a large scale and high-resolution paddy rice mapping application. In this regard, a distributed Random Forest classifier has been implemented, using the cluster-computing framework Apache Spark in a High Performance Data Analytics environment. The input data to the classifier comprise of long time-series of Sentinel-1 and Sentinel-2 images, but also pertinent vegetation indices. The proposed paddy rice classification method achieves an accuracy of more than 85%, for a study site in Northwestern South Korea.

1. INTRODUCTION

Over the next decades, the estimated increase in the global population, combined with the climate change, is anticipated to have a significant impact on the food sector (Fritz et al., 2013). Earth Observation (EO) driven agriculture monitoring, for the purposes of food security, control of the implementation of sustainable agriculture policies and the improvement of the overall agricultural productivity, is a top priority for the European Union, but also global initiatives such as GEOGLAM and Asia-RiCE.

Paddy rice is a primary crop in South Korea, which is the area of interest for this study, and information about its spatial distribution and yield is of great importance for the environmental management, but also for food security related decision-making. South Korea, although of high food security index, has low food self-sufficiency that is decreasing in the long term. The country can be considered exposed to potential food security issues due to its high dependency on international supply for major crops and limited number of exporting countries for rice.

The monitoring of paddy rice extent and its overall production requires manual field visits for survey and is thus costly and time consuming, when compared to information gathered through EO means. Remote sensing is one of the most effective technologies to map the extent of crops. Rice area mapping at the parcel, regional and national scale has been extensively studied in the past, through several approaches, including mono-temporal and multi-temporal classification schemes that utilize both optical and microwave - Synthetic Aperture Radar (SAR) data (Qin et al., 2015; Nguyen et al., 2015). Tian et al. (2018) have introduced a novel multi-season paddy rice mapping method, using Sentinel-1 and Landsat-8 data under a k-means unsupervised

[1] This work was supported by EOPEN project, partially funded by the European Commission, under the contract number H2020-776019.

classification methodology. Torbick et al. (2017) have produced an updated land cover map, including the rice class, fusing Sentinel-1, Landsat-8 OLI and PALSAR-2 data using a Random Forest classifier. Finally, Pazhanivelan et al. (2015) have introduced a robust rule-based classification for mapping rice area with multi-temporal, X-band, HH polarized SAR imagery (COSMO Skymed and TerraSAR X), with site-specific parameters.

There have been multiple studies to exploit both optical and SAR data for the monitoring of agriculture, with a lot of research on the usage of Sentinel data, utilizing their unprecedented characteristics in temporal and spatial resolution (Inglada et al., 2016; Immitzer et al., 2016; Sitokonstantinou et al., 2018). Studies have used multiple classification techniques including neural networks, supervised and unsupervised machine learning techniques, but also custom rule-based systems. Nevertheless, all studies had to manually collect their training and validation samples, undermining the design of a fully transferable and site independent framework of application for the described systems.

In this regard, and under considerations of scalability, reproducibility and transferability, which are essential for a national scale application, it was decided to design a more dynamic rice monitoring system that is largely independent of hard-to-attain non-EO information. In this work we suggest a land cover map update mechanism based on change detection to produce appropriate training and validation datasets for any given year of inspection via utilizing older ground truth information. The updated land cover map is then used to train a distributed Random Forest classifier in a High Performance Data Analytics (HPDA) environment.

The recent and unprecedented availability to high resolution satellite imagery, such as the Sentinels, has introduced a paradigm shift in the field of remote sensing. Increasing number of satellites and sensors along with improved spatial and temporal resolution generate big EO data that demand increased computational power for their exploitation (Chebbi and Boulila, 2015). To accommodate for the storage and processing requirements, in this new era of EO science, decentralized and distributed environments and frameworks are utilized. High Performance Computing (HPC) combines technologies such system software, architecture and algorithms in order to increase the effectiveness and the speed of complex processing chains. Distributed processing is part of this umbrella technology. It makes use of parallel processing on multiple machines so that the data are distributed to the entire memory of the system.

Remote sensing research has been enhanced in recent years by high performance computing (Lee et al., 2011). Alongside with the infrastructure, several computing frameworks have been developed, which promise more effective and efficient image processing. Apache Hadoop is one of the most famous implementations of the MapReduce model, introduced by Dean and Ghemawat (2004). Apache Spark, an alternative framework, has advantages over Hadoop when it comes to speed, memory and high level operators, making it the best choice for big data tasks (Jai and Atul, 2016) and the chosen framework for this work.

2. AREA OF INTEREST AND DATA ACQUISITION

2.1 Area of Interest

The area of interest for this study comprises of the South Korean regions of Dangjin and Seosan. The two areas are recorded as the highest rice producing in the country (Park et. al, 2018). The annual mean temperature and precipitation in the region are 11.4 °C and 1158.7 mm, respectively (Kim et. al, 2017). South Korea has a rainy season during the summer due to the Asian Monsoon, therefore cloud free optical imagery is scarce. For this reason, weather independent SAR imagery from Sentinel-1 has also been exploited in the suggested methods of this study. The figure below illustrates the area of interest for which an accurate paddy rice map has been produced.

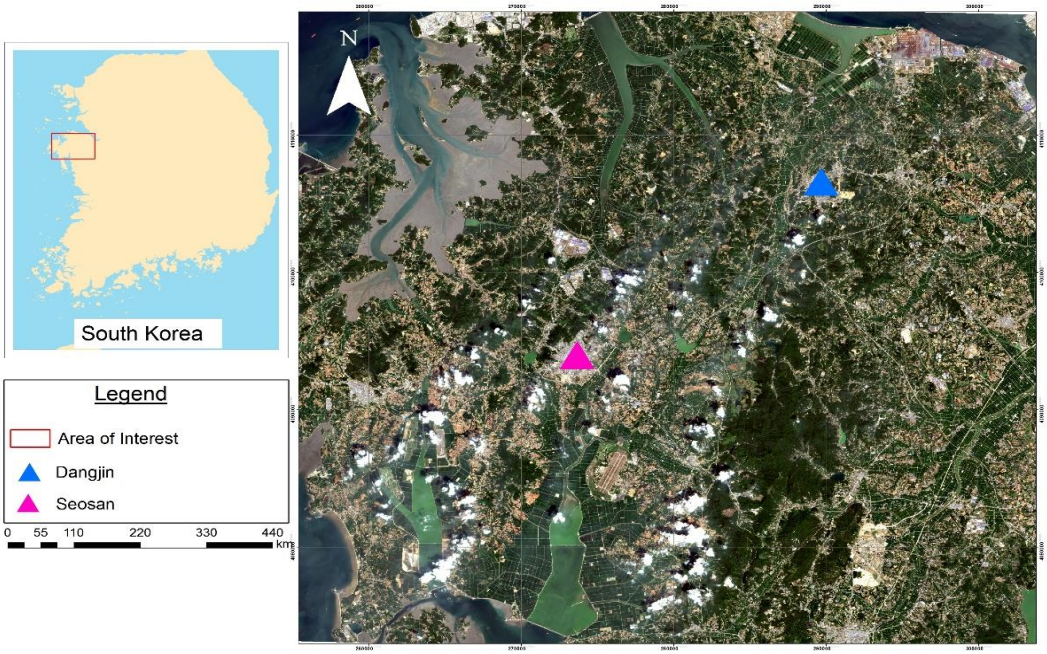

Figure 1. Area of interest located in the regions of Dangjin and Seosan in South Korea

2.2 Data acquisition

The training and validation data for the paddy rice classification algorithm stem from past land cover maps for the areas of Dangjin and Seosan in South Korea, as retrieved from the Korean Ministry of Environment database. Detailed land cover maps at the parcel level, distinguishing between rice and non-rice cultivations, are produced every few years. However, in order to systematically produce rice maps for any given year, updated land cover maps are required for the appropriate training and validation of the machine learning algorithms utilized. In this study the year of inspection is 2018, but the latest land cover maps for the area of interest were from 2015. For this reason, a framework has been developed for updating, through EO assisted change detection, the land cover map of year 2015 (reference map) for the South Korean regions of Dangjin and Seosan to reflect the reality of the year inspection (updated map).

Secondly, the input dataset for the paddy rice classification algorithm comprises of a time-series of Sentinel-1 and Sentinel-2 imagery to capture the entirety of the crop's phenology. For the creation of the pixel based feature space, monthly means of VV backscatter from Sentinel-1 images, along with Red-edge, Near Infrared (NIR) bands and Vegetation Indices (VIs) from Sentinel-2 images, were used as features. The VIs incorporated in the feature space include the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI) and Plant Senescence Reflectance Index (PSRI), which have been widely used in crop monitoring and crop mapping applications (Lebourgeois et al., 2017; Sitokonstantinou et al., 2018; Gao, 1996; Hatfield and Prueger, 2010). The resolution of the time-series of imagery is at 10 m and amounts to several tens of gigabytes of feature space, which is comprised of a total of 167 features. The figure below illustrates how the satellite image acquisitions capture the different phenological phases of rice in South Korea.

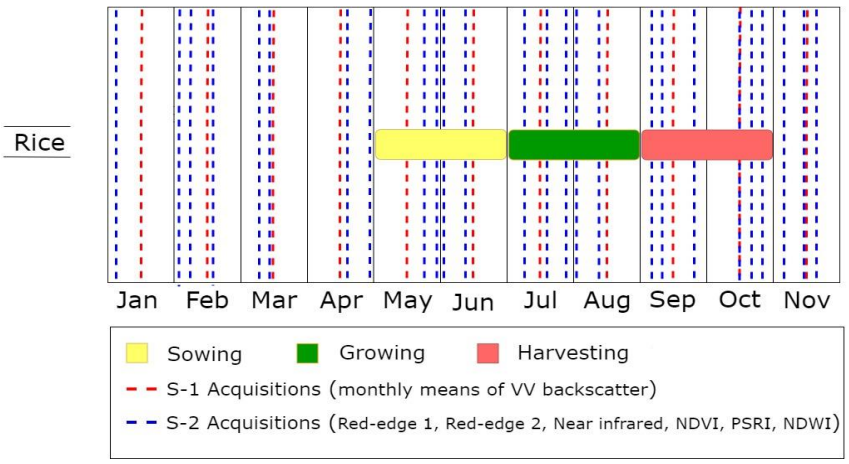

Figure 2. South Korea paddy rice phenology and the respective Sentinel-1 and Sentinel-2 acquisitions that capture it

The recent and unprecedented availability of open access high resolution satellite imagery, such as the Sentinels, has introduced a new era in the field of remote sensing. However, the automated and timely acquisition of Sentinel imagery from the plethora of available hubs becomes challenging. The different hubs offer Sentinel data with different specifications, such as their rolling archive policy, data availability, geographic coverage and latency of acquisition. Thus, an application has been developed to connect to multiple Sentinel hubs and automatically search for the pertinent Sentinel data. This broker of Sentinel data retrieves the requested products from the most efficient hub that is decided in terms of download speed and product availability. The overall architecture of this Sentinel hubs broker is shown in the figure below.

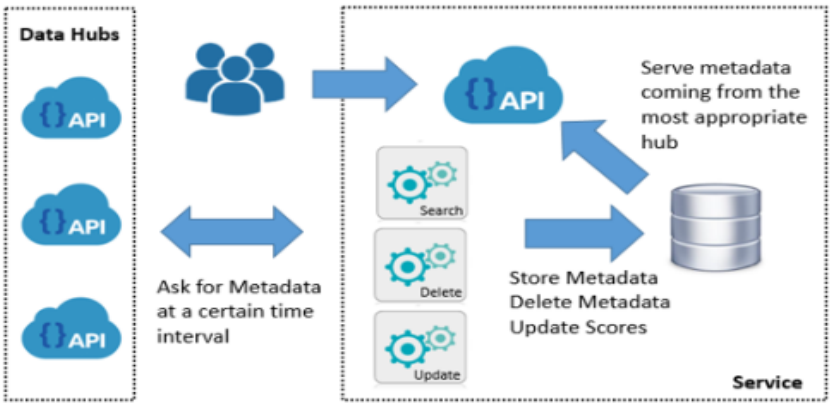

Figure 3. The federated Sentinel hubs API overall architecture

3. METHODS

3.1 Change Detection: Land cover map update

The output product of the below methodology does not attempt to accurately classify rice fields for the year of inspection but merely delete changes in the land cover map of 2015 (reference map), as described in the Area of Interest and Data Acquisition section. Therefore, using only the rice pixels from the reference map we eliminate outliers to produce the updated map for the year of inspection. This way the training dataset for the distributed Random Forest classifier is refined, by removing rice pixels of changed land cover.

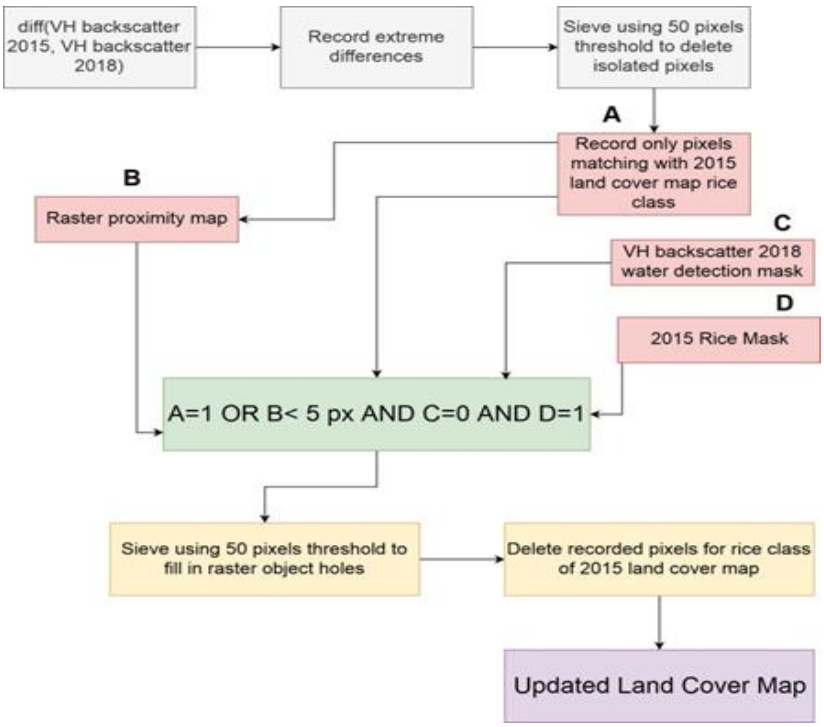

Figure 4. Framework of land cover map update method

Figure 4 illustrates the workflow for the land cover map update methodology. The processing chain starts by taking the difference of the VH polarized backscatter of Sentinel-1 images in mid-June of 2015 and 2018, respectively. VH polarization is especially sensitive in the water concentration of inundated rice fields. Imagery sensed in mid-June was selected, as it is the period that rice fields are flooded and thus more easily detectable. The preprocessing of Sentinel-1 imagery involved calibration, speckle filtering and terrain correction. The two images have been additionally co-registered in order to ensure one to one pixel match.

From the difference product of the two VH backscatter images we record only the pixels of extreme differences, using the values of bottom 2% and top 98% of the cell data range as thresholds. Then raster polygons smaller than 50 pixels are replaced with the pixel value of the largest neighbor polygon, thus eliminating island pixels and filling raster polygon holes. The remaining outlier pixels are superimposed with the rice pixels of the reference map, keeping only matching records. Hence, we end up with a set of pixels identified as rice in the reference map that appear changed for the year of inspection, as illustrated in the following figure.

Figure 5. Unrefined outlier detection product

To further refine the product illustrated above, it was attempted to fill in missing pixels within the identified non-rice parcels. A raster proximity map was generated to indicate the distance from each pixel to the nearest pixels identified as an outlier. Then pixels that haven not yet been identified as outliers and have a distance of fewer than 5 pixels from outlier pixels are recorded (B in Figure 4). Product C in Figure 4 refers to an automated water detection map via thresholding, based on the VH backscatter, sensed in mid June of the inspection year. Since rice is inundated the algorithm also works for identifying the flooded agricultural land.

Figure 6. VH backscatter 2018 water mask via thresholding

Finally, product D in Figure 4 refers to rice pixels of the reference map. Combining products A-D under the Boolean expression in Figure 4, we record all outlier pixels of the unrefined detection product (Figure 5), but also pixels that are concurrently not included in the threshold based water mask and have a proximity value of fewer than 5 pixels (Figure 6). The final product is illustrated in Figure 8 in the Results section.

3.2 Distributed Random Forest

Food security monitoring at the national scale demands time series analysis of multiple satellite images, so as to capture the crop's phenology. It becomes apparent that the size of this dataset, which is several tens of gigabytes, exceeds the memory limits of conventional machines. To overcome this limitation, distributed architecture and parallel data processing had to be introduced. From the storing point of view, the Hadoop Distributed File System (HDFS) is used to store large amounts of data (Ruzgas et al., 2016) giving the potential of breaking down the dataset to separate blocks and distribute them to the multiple nodes of the cluster. Moving on to the processing of the dataset, Apache Spark reads the distributed pixel-based feature space from HDFS in the form of dataframes and transforms them into Resilient Distributed Datasets (RDDs), a read-only, partitioned collection of records, which allows for transformations and actions, such as filtering and sorting. Afterwards, data are prepared for the classification process by randomly splitting them into training and test subsets, with a split ratio of 30/70.

Considering the need of efficient and fast data analysis for this oversized dataset, an optimized and scalable machine learning algorithm was used to exploit the power of distributed computing. The distributed Random Forest implementation comes with MLLib, a library of Apache Spark. The Random Forest algorithm is an ensemble classifier based on the decision tree model. Training data are split into k different subsets, using the bootstrap method, and each subset creates a new decision tree. Finally, a forest is constructed from the different decision trees and predictions are made through a majority voting mechanism (Breiman, 2001). The algorithm was

parameterized to have 10 decision trees, a maximum tree depth of 30 and a maximum number of bins set to 32. The overall architecture of the system is depicted in Figure 7. The overall architecture of the distributed Random Forest classification

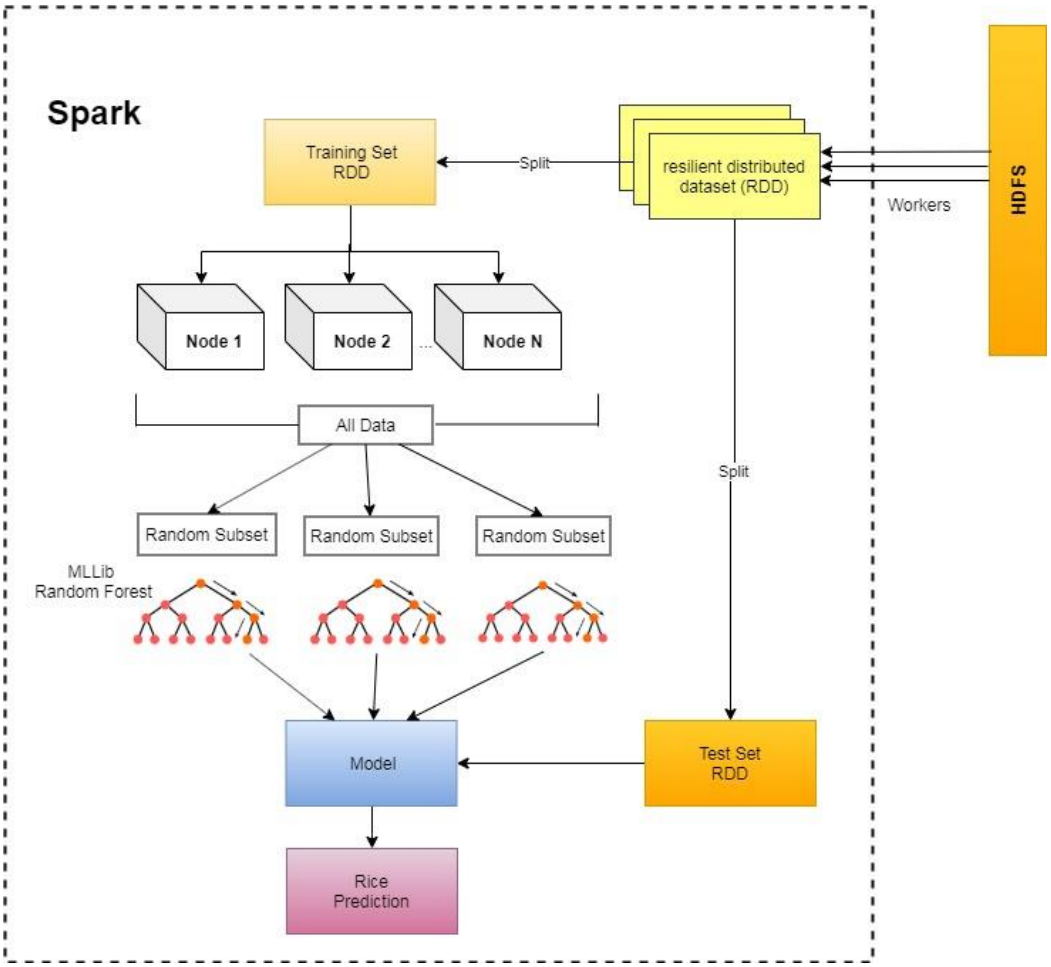

Figure 7. The overall architecture of the distributed Random Forest classification

4. RESULTS

4.1 Change Detection: Land cover map update

The figure below (Figure 8) illustrates the final refined outliers detection output at the top left corner. Rice fields based on the reference map and the unrefined outlier detection products are also shown for visual comparison.

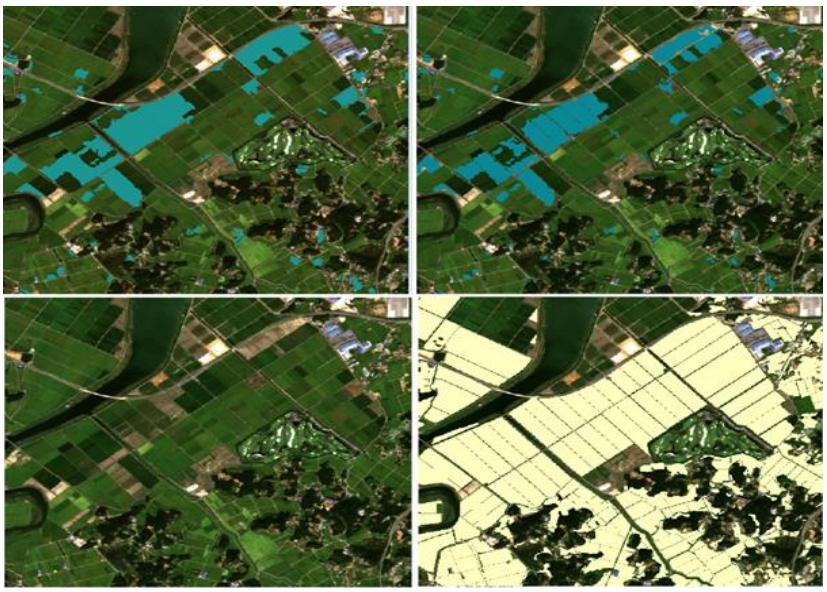

Figure 8.Top left - refined outliers detection, top right – unrefined outliers detection, bottom left – true colour composite September 2018, bottom right – rice pixels based on the reference map

The illustrated figures throughout this section represent an indicative snapshot of the area of interest. The method was tested in a total area of 7,185 ha of paddy rice. The detected outlier pixels amounted to 396 ha, therefore resulting in a 5.5% change. This percentage does not only refer to rice parcels that altered to other than rice cultivations in 2018, but also pixels on the parcel borders that were classified as rice due to less than optimal parcel digitization of the reference map. This further refines and removes noise from the training dataset.

4.2 Paddy rice classification

The classification performance was assessed based on the metrics of recall, precision and F1-score, computed as shown in Equations (1)–(3). Recall is the ratio of correctly classified pixels over the total number of pixels for a ground truth class. Alternatively, precision is the ratio of correctly classified pixels for a given class to the total number of pixels predicted to belong to that class. F1- score is the harmonic mean of precision and recall (Sokolova and Lapame, 2009). The metrics were computed based on the below confusion matrix. Rice is considered as the positive class and non-rice as the negative class.

Table 1. Confusion matrix of paddy rice classification

	Prediction Rice	**Prediction Non-Rice**
Truth Rice	4,151,744	616,558
Truth Non-Rice	697,320	19,547,543

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{F1}_{\text{score}} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

where TP is the number of correctly classified rice pixels or True Positive, FP is the number of non-rice pixels classified as rice or False Positive, TN the number of correctly classified non-rice pixels or True Negative and FN is the number of rice pixels classified as non-rice or False

Negative.

Table 2. Accuracy metrics for the Random Forest classification

Precision (%)	Recall (%)	F1 – score (%)
85.62	87.07	86.34
96.90	96.50	96.70

The results are very promising, with more than 85% precision and recall for the paddy rice class. It should be noted that the above accuracy metrics are computed against the updated map, as described in the Methods section, and not a validated dataset. The validity of the updated map is thus assumed for training and validation, even though it is not perfectly representative of the truth. The dataset does not account for the new rice cultivations in the year of inspection that were described as non-rice cultivations in the reference map, as the change detection method only removes outliers from the latter. For this reason the actual precision values are expected to be slightly higher than the ones indicated.

5. CONCLUSIONS

The distributed Random Forest implementation in the HPDA environment accommodates for the big EO data and allows for the large scale pixel-based paddy rice classification. The regional application presented in this work can be linearly scaled up for the entire country, providing an accurate indication of the total paddy rice extent. The classification product can then be used for the estimation of yield, providing essential information for food security and enabling high level decisions at the national scale.

The exploitation of exclusively freely available data, along with the employment of big data technologies, such HDFS and Apache Spark, but also big data infrastructure, such as the HPDA, constitutes the suggested rice monitoring scheme a scalable and transferable solution.

ACKNOWLEDGEMENTS

Authors acknowledge the Copernicus Open Access Hub (**https://scihub.copernicus.eu/**) and the Hellenic National Sentinel Data Mirror Site (**https://sentinels.space.noa.gr/**) for providing free access to Sentinel-1 and Sentinel-2 images. We are also grateful to the High Performance Computing Center Stuttgart (HLRS) for providing access to their HPDA resources.

6. REFERENCES

Breiman, L., 2001. Random Forests. Machine Learning, 45, pp. 5-32.

Chebbi, I.; Boulila W., 2015. Big data: Concepts, challenges and applications. In ICCCI, 2, pp. 638-647.

Dean, J.; Ghemawat, S., 2004. Mapreduce: Simplified data processing on large clusters. In USENIX Symposium on Operating Systems Design and Implementation.

Fritz, S.; See, L.; You, L.; Justice, C.; Becker-Reshef, I.; Bydekerke, L.; Cumani, R.; Defourny, P.; Erb, K.; Foley, J. et al., 2013. The Need for Improved Maps of Global Cropland. Eos Trans. Am. Geophys., 94, pp. 31–32.

Gao, B.C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ., 58, pp. 257–266.

Hatfield, J.L.; Prueger, J.H., 2010. Value of using different vegetative indices to quantify agricultural crop characteristics at different growth stages under varying management practices. Remote Sens., 2, pp. 562–578.

Immitzer, M.; Vuolo, F.; Atzberger, C., 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. Remote Sens., 8, pp. 166.

Inglada, J.; Vincent, A.; Arias, M.; Marais-Sicre, C., 2016. Improved early crop type identification by joint use of high temporal resolution sar and optical image time series. Remote Sens., 8, pp. 362.

Jai, V.; Atul P., 2016. Comparison of mapreduce and spark programming frameworks for big data analytics on HDFS. IJCSC, 7, pp. 80-84.

Kim, M.; Ko, J.; Jeong, S.; Yeom, J.M.; Kim, 2017. H.O. Monitoring canopy growth and grain yield of paddy rice in South Korea by using the GRAMI model and high spatial resolution imagery. Gisci. Remote Sens., 54, pp. 535-536.

Lebourgeois, V.; Dupuy, S.; Vintrou, É.; Ameline, M.; Butler, S.; Bégué, A., 2017. A combined random forest and OBIA classification scheme for mapping smallholder agriculture at different nomenclature levels using multisource data (simulated Sentinel-2 time series, VHRS and DEM). Remote Sens., 9, pp. 259.

Lee, C. A.; Gasster, S.D.; Plaza, A.; Chang, C. I.; Huang, B., 2011. Recent developments in high performance computing for remote sensing- A review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 4.3, pp. 508-527.

Nguyen, D.; Clauss, K.; Cao, S.; Naeimi, V.; Kuenzer, C.; Wagner, W., 2015. Mapping rice seasonality in the mekong delta with multi-year envisat asar wsm data, Remote Sensing 7, pp. 15868–15893.

Park, S.; Im, J.; Yoo, C.; Han, H.; Rhee, J., 2018. Classification and Mapping of Paddy Rice by Combining Landsat and SAR Time Series Data. Remote Sens., 10, pp. 22.

Pazhanivelan, S.; Pandian, K.; Christy, P.; Mary, N.; Elangovan, S.; Jeyaraman, S. et al., 2015. Rice crop monitoring and yield estimation through COSMO Skymed and TerraSAR-X: A SAR-based experience in India, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XL-7/W3. 10.5194/isprsarchives-XL-7-W3-85-2015.

Qin, Y.; Xiao, X.; Dong, J.; Zhou, Y.; Zhu, Z.; Zhang, G.; Du, G.; Jin, C.; Kou, W.; Wang, J.; Li, X., 2015. Mapping paddy rice planting area in cold temperate climate region through analysis of time series landsat 8 (oli), landsat 7 (etm+) and modisimagery, ISPRS Journal of Photogrammetry and Remote Sensing, 105, pp. 220–233.

Ruzgas T.; Jakubėlienė K.; Buivytė A., 2016. Big Data Mining and Knowledge Discovery, Journal of Communications Technology, Electronics and Computer Science, 9, pp. 7.

Sitokonstantinou V.; Papoutsis I.; Kontoes C.; Arnal A.; Andrés A.; Zurbano J., 2018. Scalable Parcel-Based Crop Identification Scheme Using Sentinel-2 Data Time-Series for the Monitoring of the Common Agricultural Policy, Remote Sensing, 10, pp. 911.

Sokolova, M.; Lapame G., 2009. A systematic analysis of performance measures for classification tasks, Information Processing and Management, 45, pp. 427-437.

Tian, H.; Wu, M.; Wang, L; Niu, Z., 2018. Mapping Early, Middle and Late Rice Extent Using Sentinel-1A and Landsat-8 Data in the Poyang Lake Plain, China, Sensors, 18, pp. 185.

Torbick, N.; Chowdhury, D.; Salas, W.; Qi, J., 2017. Monitoring Rice Agriculture across Myanmar Using Time Series Sentinel-1 Assisted by Landsat-8 and PALSAR-2, Remote Sensing, 9, pp. 119.