Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective

Azar Zafari, Raul Zurita-Milla, and Emma Izquierdo-Verdiguier

***Abstract*— The classification of the ever-increasing collections of remotely sensed images is a key but challenging task. In this letter, we introduce the use of extremely randomized trees known as Extra-Trees (ET) to create a similarity kernel [ET kernel (ETK)] that is subsequently used in a support vector machine (SVM) to create a novel classifier. The performance of this classifier is benchmarked against that of a standard ET, an SVM with both conventional radial basis function (RBF) kernel, and a recently introduced random forest-based kernel (RFK). A time series of Worldview-2 images over smallholder farms is used to illustrate our study. Four sets of features were obtained from these images by extending their original spectral bands with vegetation indices and textures derived from gray-level co-occurrence matrices. This allows testing the performance of the classifiers in low- and high-dimensional problems. Our results for the high-dimensional experiments show that the SVM with tree-based kernels provide better overall accuracies than with the RBF kernel. For problems with lower dimensionality, SVM-ETK slightly outperforms SVM-RFK and SVM-RBF. Moreover, SVM-ETK outperforms ET in most of the experiments. Besides an improved overall accuracy, the main advantage of ETK is its relatively low computational cost compared to the parameterization of the RBF and RFK. Thus, the proposed SVM-ETK classifier is an efficient alternative to common classifiers, especially in problems involving high-dimensional data sets.**

***Index Terms*— Image classification, random forest, smallholder agriculture, support vector machine (SVM), very high spatial resolution satellite images.**

I. Introduction

With the advent of new sensors and open data policies, large data sets are becoming available. This includes large collections of remotely sensed (RS) images that often need to be classified to support their use in various domains and applications [1]. Yet, traditional classification methods cannot properly deal with the challenge of handling large and complex data sets [2]. Moreover, access to images with higher spatial, spectral resolutions facilitates the extraction of extra features from RS images. These features are required because elements in the scene may appear at various scales and orientations because of variable weather and lighting conditions [3]. Extra features often lead to high dimensionality, which is the most important challenge in recent RS image classification tasks [2].

Kernel methods can efficiently deal with nonlinear and high-dimensional problems. Support vector machine (SVM) is one of the most representative kernel-based classification methods, and radial basis function (RBF) is the most common kernel used with this classifier [4]. Using an SVM-RBF classifier requires optimization of two parameters (i.e., RBF bandwidth and SVM regularization parameters) through a computationally demanding cross-validation process [4]. This is a limitation of SVM-RBF [5]. Another limitation to accuracy and efficiency of SVM-RBF is experienced when the number of features increases for a certain amount of training data [6]. The reason for this is the curse of dimensionality, also called the Hughes phenomenon. Moreover, the RBF kernel is typically computed with all of the available features assuming that they are all informative. High-dimensional problems often require to select the most important features, and SVM-RBF cannot directly select the most important features. This is another known limitation of SVM-RBF [6], [7]. Another well-known classifier for high-dimensional problems is random forest (RF) [8]. RF grows trees based on recursive partitioning of nodes, and it generally uses the Gini index to select the best split in a node. RF is fast, not sensitive to the choice of parameters, produces good results with relatively low amounts of training samples, and is resistant to noise in training samples and to overfitting [9]. These characteristics alongside with its tree-based structure make it a suitable classifier to draw partitions in the data and to obtain an RF-based kernel (RFK) that quantifies similarities between samples [5], [10]. The pairwise similarities between the samples reflect whether they fall in the same end node or not [11]. By using default values for the RF parameters, the classification results of SVM-RFK are comparable to those obtained by SVM-RBF as shown for an AVIRIS data set often used in benchmarking studies (i.e., Salinas) and for a time series of Worldview-2 images over Sukumba, Mali [5]. Hence, RFK is an effective alternative to RBF, particularly when combined with RF-based feature selection methods [5]. Nevertheless, the structure of this kernel is highly dependent on the training labels since it is built based on classification results of RF. This means that the RFK can overfit to the training data especially when the number of features is low since the structure of trees would be correlated [12]. Moreover, RFK is negatively impacted by possible mislabeled samples in the training data. To overcome these downsides, the randomization level of the RF ensemble

Manuscript received June 12, 2019; revised September 6, 2019; accepted November 9, 2019. Date of publication November 26, 2019; date of current version September 25, 2020. This work was supported in part by the Bill and Melinda Gates Foundation through the STARS Grant Agreement under Grant 1094229-2014. *(Corresponding author: Azar Zafari.)*

A. Zafari and R. Zurita-Milla are with the Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500 AE Enschede, The Netherlands (e-mail: a.zafari@utwente.nl; r.zurita-milla@utwente.nl).

E. Izquierdo-Verdiguier is with the Institute for Surveying, Remote Sensing and Land Information (IVFL), University of Natural Resources and Life Sciences (BOKU), A-1190 Vienna, Austria (e-mail: emma.izquierdo@boku.ac.at).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2019.2953778

should be increased to have trees that are less correlated. This can be achieved by using extremely randomized trees, a method commonly known as Extra-Trees (ET) [13]. The ET also generates an ensemble of unpruned decision trees, but it splits nodes by choosing cut points fully at random and it uses all training sample rather than bootstrap subsets to grow the trees [13]. In extreme cases, ET builds totally randomized trees (ToRT). The structure of these trees is independent of the training labels [13]. The randomization level can be adjusted to the problem at hand by selecting suitable parameters [13]. Several articles have applied ET classifier for land cover classification and have shown that ET can outperform RF and SVM-RBF in terms of overall accuracy (OA) [14], [15]. Besides OA, the main strong point of ET is its computational efficiency [13]. Like RF, the tree-based structure of the ET can be used to create partitions in the data and to generate an ET kernel (ETK) that encodes similarities between samples based on these partitions [13]. Using ETK as an alternative to RBF and RFK, one can avoid the computational cost associated with parametrizing the RBF kernel and reduce the probability of getting an overfitted kernel. The main goal of this letter is to present and evaluate a novel classifier created by combining an ET-based kernel and SVM (SVM-ETK). We evaluate our approach by comparing it against ET, SVM-RFK, and SVM-RBF. Our evaluation is illustrated with a time series of very high spatial resolution data acquired over agricultural lands.

II. ET Kernel

ET grows an ensemble of unpruned decision trees using the classical top–down procedure through randomly recursively splitting the data into child nodes until reaching the terminal nodes defiend by a stopping criterion [13]. ET differs from other decision-tree-based ensembles, such as RF in two cases [14]. First, ET does not search extensively for an optimized cut point in the nodes; this causes the tree structures to be independent of the target variable values of the learning samples [14]. Second, it uses the same training sample for growing all trees rather than a bootstrap replica. The explicit randomization of cut point and feature combined with ensemble averaging reduces the variance among the trees. Using full training samples rather than bootstrapped samples reduces the bias [13]. Furthermore, the computational load of training for ET is less than that required to train RF since it does not search intensively for an optimal cut point [13]. ET has three parameters in common with RF. Like RF, a random subset of all the available features is evaluated when looking for the best split point. The number of features in the subset is controlled by the user and is typically called mtry. The second common parameter is N_t that is the number of the decision trees to be generated. It has been shown that for ET and RF, the prediction error is a monotonically decreasing function of a number of trees [12], [13]. The third common parameter is $n_{\min}$ that is the minimum sample size for splitting a node with the default value of one (or two) [12], [13]. The optimal value for $n_{\min}$ increases depending on the level of mislabeled samples in training data [13]. The higher values for this parameter result in smaller trees, smaller variance, and higher bias [12], [13]. In addition to mtry, N_t, and $n_{\min}$, the specific parameter to ET is the number of random cut points (N_{cp}) to consider for each selected feature in splitting a node. In the most extreme case, ET randomly picks a single feature (i.e., mtry is one) and a single cut point at each node [13]. This is typically called ToRT, and its structure is independent of the labels of training samples. However, the level of randomization can be optimized to the problem with $mtry$ and N_{cp} parameters [13], [14]. When ET uses more than one feature or/and random cut point in splitting the nodes, such as RF, it uses the Gini index or normalization of information gain to select the best cut point out of the randomly selected splits [13].

Tree-based models, such as RF and ET, can be used to generate kernels using a feature space defined by the terminal nodes of the trees; this is comprehensively proven and explained in [13]. The characteristics of ET, and consequently of the ETK, make it less dependent on the training labels than RF and its kernel version. This reduces the probability of getting an overfitted kernel. The dependence of the ETK on the training labels can be controlled with the randomization level, which can be adjusted to the problem at hand by selecting suitable ET parameters. Mathematically, ETK is a square matrix with the size of the training set, where the element (i, j) contains a number of times that the samples i and j fall in the same terminal node normalized by the number of trees in the ensemble. In other words, if two samples are fallen in the same terminal node of a tree, the similarity is equal to one; otherwise, it is zero. The similarity of each tree $[K_{t_n}(x_i, x_j)]$ is obtained by [13], [14]

$$K_{t_n}(x_i, x_j) = I\left[q(x_i) = q(x_j)\right] \tag{1}$$

where q is a terminal node and t_n is the nth tree of the ET. Then, the ETK matrix is calculated by the average of tree kernel matrices

$$\mathbf{ETK} = \frac{1}{N_t}\sum_{t_n=1}^{N_t}\mathbf{K}_{t_n} \tag{2}$$

where N_t is the number of trees used in the ET. ETK is the average of kernel matrices obtained with all trees and can be used in kernel-based methods, such as SVM (i.e., SVM-ETK).

III. Data and Experiments

A. Data and Study Area

The study area is located near Sukumba in Mali, West Africa. A time series of WorldView-2 images is used to illustrate this study. This data set includes seven multispectral images that cover the cropping season of 2014 [16]. Ground-truth labels for five common crops including cotton, maize, millet, peanut, and sorghum were collected for nine fields per crop (45 fields) through fieldwork. The Sukumba images are atmospherically corrected and coregistered, and trees and clouds are masked [16]. These images and the corresponding ground data are part of the STARS project. The Sukumba data set originally contains 56 bands (i.e., seven images with eight bands each). The number of features was extended by obtaining the normalized difference

TABLE I
EXPERIMENTS DESCRIPTION (N_f: NUMBER OF FEATURES)

Acronim	Features	N_f
B	Spectral features	56
BVI	Spectral &VIs features	525
BVITVI	BVI and GLCM textures of VIs	8498
ALL	BVI and textures of Spectral and VIs	10584

vegetation index (NDVI), the difference vegetation index (DVI), the ratio vegetation index (RVI), the soil-adjusted vegetation index (SAVI), the modified soil-adjusted vegetation index (MSAVI), the transformed chlorophyll absorption reflectance index (TCARI), and the enhanced vegetation index (EVI). The data set, the study area, a list, and a short explanation of VIs used in this study can be found in [5]. Furthermore, the pairwise band combinations by means of the difference, ratio, and normalization between bands 2 and 8 were generated increasing the number of the features until 525. Next, the number of features for Sukumba data set was extended by adding the gray-level co-occurrence matrix (GLCM) textures to the spectral features and VIs. These GLCM-based features capture spatial relationships across the pixels [17]. The GLCM textures derived from the Sukumba data set are also presented and explained comprehensively in [5] and [18]. Stacking all the spectral, VIs, and GLCM features of VIs, the total number of features reached 9450. This number further was increased by including extra features, namely, green leaf index (GLI) and local binary pattern (LBP) [18]. With this addition, the final number of features available for the case ALL reached 10 584. Table I shows the subsets and quantity of the features that are used in four tests to examine the proposed method in this letter. In such large data sets in last two experiments (i.e., BVITVI and ALL), there are many correlated features, and some of them might be not helpful for the classification task at hand (i.e., noise).

B. Experimental Setup

First, the polygons representing farms were split into four subpolygons. Two subpolygons were used to choose the training samples and the other two, the test samples. Then, the train and test sets were split into ten random subsets, with a balanced number of samples per class (130 and 100 samples per class for training and test, respectively). Final results were obtained by averaging the results obtained with ten subsets available for each spectral case (see Table I). To investigate the influence of ET parameters on ET kernel performance, the ranges $\{100, 300, 500\}$ and $\{1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ are used for N_t and N_{cp}, respectively. The mtry parameter is set to its default value of the square root of the number of features. The $n_{\min}$ parameter is also set to its default (i.e., $n_{\min} = 1$) because a moderate level of mislabeled samples is expected [13]. Moreover, ToRT results are obtained by setting mtry and N_{cp} to 1. To obtain RFK, N_t and mtry parameters in RF were set to their default values of 500 trees and the square root of the number of features because this stabilizes the error of the classification in the most applications [5]. The optimization of RF parameters for obtaining the RFK is skipped because of the marginal gain in OA of SVM-RFK compared to added computational cost [5]. Thus, the performances of the kernels derived from RF and ET are compared using models trained with default parameters for both methods. For the RBF kernel, the optimum bandwidth was found using the range $[0.1, 0.9]$ of the quantiles of the pairwise Euclidean distances ($D = \|x - x'\|^2$) between the training samples, and the optimal C value was found in the range of $[5, 500]$. For the RBF kernel, fivefold cross-validation was used to find the optimal bandwidth and C values. Fivefold cross-validation was also used to optimize C for the RFK and ETKs. In all the cases, the one-versus-one multiclass strategy implemented in LibSVM was used [19]. Classification results are compared in terms of their average OA ($\overline{\text{OA}}$) and Cohen's kappa index ($\bar{\kappa}$). Next, crop classifications maps are obtained through the classifiers pertinent to the set of train and test samples, which provides the highest test OA between the other ten sets of train and test samples. For visibility reasons, two classified fields per crop are shown for each classifier. In the end, OA and κ of the top-performing classifiers are obtained for all available labeled samples in the 45 fields.

IV. RESULTS AND DISCUSSION

Fig. 1 displays $\overline{\text{OA}}$s of ten test subsets versus different parameters configurations for ET and SVM-ETK classifiers. Fig. 1 shows that SVM-ETK always outperforms ET for all the cases with VIs (see Table I) and irrespective of the value of N_t and N_{cp}. For the experiment with B features, $\overline{\text{OA}}$ of SVM-ETK and ET overlaps in some ranges of N_{cp}. Yet, SVM-ETK outperforms ET in most ranges particularly for small values of N_{cp}. We can also observe in Fig. 1 that the peaks of $\overline{\text{OA}}$ for SVM-ETK correspond to higher levels of randomization (i.e., N_{cp} equal or less than 10). Nonetheless, the difference between $\overline{\text{OA}}$s obtained with the default (i.e., 1) and the best value of N_{cp} is less than 1% for the SVM-ETK classifier. For ET, lower levels of randomization lead to the best $\overline{\text{OA}}$s and optimizing the N_{cp} results in 1% improvement in OA for the experiments with BVITVI and ALL features. Fig. 1 also shows that the higher number of trees (i.e., 300 and 500) generates higher $\overline{\text{OA}}$s for both ET and SVM-ETK.

Table II compares $\overline{\text{OA}}$ and $\bar{\kappa}$ of the best and configurations of SVM-ETK with SVM-RBF, SVM-RFK, and ET classifiers. In Table II, the classifiers with the best parameters are shown with $*$ and with the default parameters are shown with d (i.e., ET* and ET$_d$). Focusing on the experiment with B features, the SVM-ETKs with an $\overline{\text{OA}}$ of 83.38% slightly outperform both SVM-RFK and SVM-RBF with $\overline{\text{OA}}$s of 80.68% and 82.08%, respectively. The default values of 500 and 1 for N_t and N_{cp} parameters of SVM-ETK give the best $\overline{\text{OA}}$ obtained in the tested ranges. Thus, the results for SVM-ETK$_d$ and SVM-ETK* are the same for this experiment.

Focusing on the experiment with BVI features, SVM-ETKs and SVM-RBF perform almost equally considering $\overline{\text{OA}}$ and SD of test subsets, and these two classifiers outperform ET and SVM-RFK. Increasing the number of features to 8498 in the experiment with BVITVI features results in a decrease of 6.62% in $\overline{\text{OA}}$ for SVM-RBF, while it slightly improves the results of ET*, SVM-ETKs, and SVM-RFK. Using the

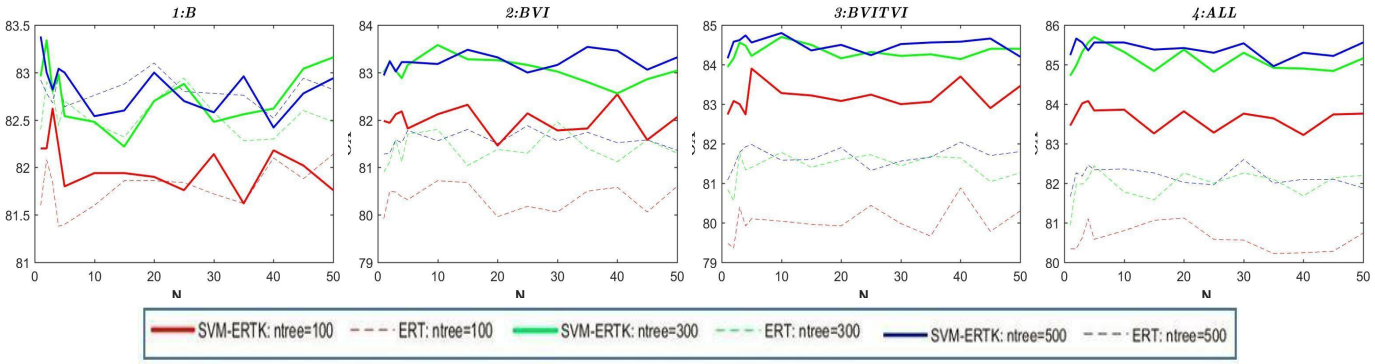

Fig. 1. $\overline{\text{OA}}$ for SVM-ETK and ET classifiers versus the number of random cut points for each candidate feature (N_{cp}) for the four experiments.

TABLE II

CLASSIFICATION RESULTS FOR DIFFERENT CASES AND CLASSIFIERS. N_t AND N_{cp} ARE, RESPECTIVELY, THE NUMBER OF TREES AND THE NUMBER OF RANDOM CUT POINTS PER CANDIDATE FEATURE. $*$ AND d ARE, RESPECTIVELY, THE BEST AND DEFAULT CONFIGURATIONS

Case	Classifier	N_t	N_{cp}	$\overline{OA}$	SD	$\bar{\kappa}$	SD$_\kappa$
B	ET*	500	20	83.10	0.83	0.79	0.01
	ET_d	500	1	82.92	0.85	0.79	0.01
	SVM-ETK*	500	1	83.38	1.26	0.79	0.01
	SVM-ETK$_d$	500	1	83.38	1.26	0.79	0.01
	SVM-RFK	500	–	80.68	1.13	0.76	0.01
	SVM-RBF	–	–	82.08	2.21	0.77	0.03
BVI	ET*	300	30	81.96	1.56	0.77	0.02
	ET_d	500	1	81.28	1.21	0.77	0.02
	SVM-ETK*	300	10	83.58	1.49	0.77	0.01
	SVM-ETK$_d$	500	1	82.94	1.30	0.77	0.01
	SVM-RFK	500	–	81.86	0.98	0.77	0.01
	SVM-RBF	–	–	83.44	1.46	0.79	0.02
BVITVI	ET*	500	40	82.04	0.99	0.78	0.01
	ET_d	500	1	81.08	1.52	0.76	0.02
	SVM-ETK*	500	10	84.80	1.02	0.77	0.02
	SVM-ETK$_d$	500	1	84.16	1.22	0.77	0.02
	SVM-RFK	500	–	84.36	1.01	0.80	0.01
	SVM-RBF	–	–	77.38	1.03	0.72	0.01
ALL	ET*	500	30	82.60	1.73	0.78	0.02
	ET_d	500	1	81.66	1.90	0.77	0.02
	SVM-ETK*	300	5	85.70	1.00	0.77	0.02
	SVM-ETK$_d$	500	1	85.24	1.72	0.77	0.02
	SVM-RFK	500	–	85.08	1.83	0.78	0.02
	SVM-RBF	–	–	78.72	1.04	0.73	0.01

tree-based kernels in an SVM for this experiment gives almost equal results considering their $\overline{\text{OA}}$ and SD.

In the fourth experiment with 10584 features, the SVM-ETKs and SVM-RFK perform almost equally and they considerably outperform ET and SVM-RBF. Our results show that SVM-RBF for the possibly noisy high-dimensional experiments (i.e., BVITVI and ALL) does not generate competitive results compared to the tree-based classifiers. The highest $\overline{\text{OA}}$ is 85.70% and obtained for SVM-ETK* with all features.

The McNemar test at 5% significance level shows that the difference between the classification results of SVM-ETK*, SVM-ETK$_d$, and SVM-RFK is not statistically significant for higher dimensional cases, including the experiment with BVI features. For the experiment with *B* features, the McNemar test shows that the results of both SVM-ETK* and SVM-ETK$_d$ are statistically significant compared to the results of SVM-RFK. This confirms that RFK and ETKs perform equally well for higher dimensional experiments, but ETKs outperform RFK for the lowest dimensional case.

The $\overline{\text{OA}}$ and $\bar{\kappa}$ results for ToRT and ToRT kernel in an SVM (SVM-ToRTK) are shown in Table III. These results

TABLE III

CLASSIFICATION RESULTS OF TORT AND TORT KERNELS IN AN SVM (I.E., SVM-TORTK)

Case	Classifier	N_t	$\overline{OA}$	SD	$\bar{\kappa}$	SD$_\kappa$
1:B	ToRT	300	80.62	0.006	0.75	0.008
		500	81.32	0.01	0.76	0.01
	SVM-ToRTK	300	83.16	0.72	78.95	0.01
		500	83.70	0.77	79.29	0.01
2:BVI	ToRT	300	78.98	0.01	0.73	0.01
		500	79.36	0.009	0.74	0.01
	SVM-ToRTK	300	82.44	1.06	78.05	0.01
		500	82.44	1.03	78.05	0.01
3:BVITVI	ToRT	300	59.72	0.02	0.49	0.03
		500	60.96	0.02	0.51	0.03
	SVM-ToRTK	300	63.50	2.38	54.37	0.03
		500	65.50	2.35	56.88	0.03
4:ALL	ToRT	300	56.88	0.03	0.53	0.04
		500	58.14	0.02	0.54	0.03
	SVM-ToRTK	300	63.06	2.35	53.76	0.03
		500	64.38	2.49	54.58	0.03

were only obtained for 300 and 500 trees considering the results of the experiments shown in Fig. 1. Table III shows that using 300 and 500 trees generates similar results. These results show that SVM-ToRTK outperforms ToRTK. Comparing Tables II and III, SVM-ToRTK, SVM-ETKs, and SVM-RBF yield similar $\overline{\text{OA}}$ and $\bar{\kappa}$ for the experiment with *B* features. For the experiment with BVI features, SVM-ToRTK also gives competitive results compared to the other classifiers, but for the higher number of features, the performance of SVM-ToRTK decreases significantly. ToRT acts like an unsupervised classifier and results in a label independent kernel that cannot deal with high-dimensional noisy problems, but it can be used as an alternative to RBF, RFK, and ET when the number of features is low. This will also reduce the computational load for obtaining the kernel compared to ETK. Thus, increasing the level of randomization in ET to its most extreme case (i.e., ToRT) is only preferred for the smaller number of features since it slightly improves the results (improving $\overline{\text{OA}}$ and reducing its pertinent SD in comparison with ETK) and reduces the computational cost.

Finally, we present maps, $\overline{\text{OA}}$, and $\bar{\kappa}$ (see Table IV) corresponding to the whole available ground-truth labels in all 45 fields in the study area. The classification maps are obtained with *B* features and through the trained classifier corresponding to the set of train and test with the highest OA. Looking into Table IV, all classifiers perform good and at about same level when obtained based on *B* features and when applied to the whole study area, while SVM-ETK slightly

TABLE IV

$\overline{\text{OA}}$ AND $\bar{\kappa}$ OVER THE 45 FIELDS IN THE STUDY AREA

Classifier	ET	SVM-RBF	SVM-RFK	SVM-ETK
$\overline{OA}$ (in %)	78.63	79.59	79.57	**79.67**
$\bar{\kappa}$	0.72	0.73	0.73	**0.74**

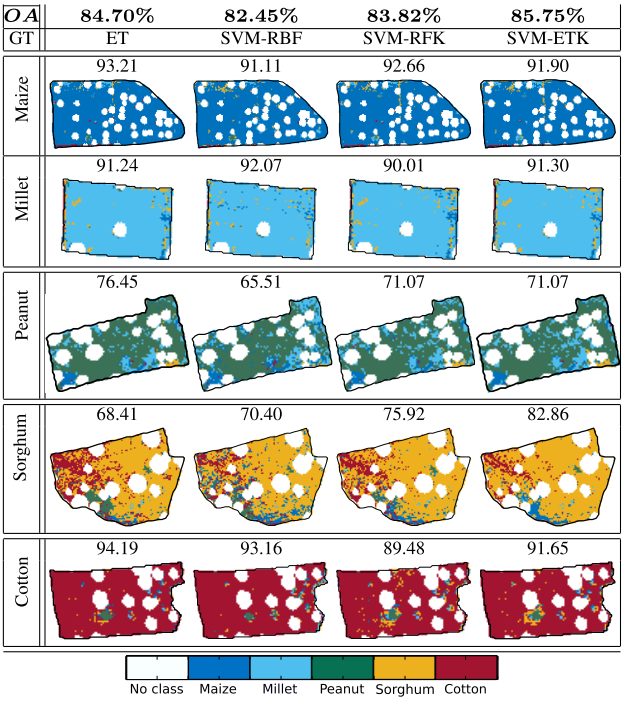

Fig. 2. Crop field per ground-truth class along with their OA obtained for the different classifiers using B and $\overline{\text{OA}}$*s* for five fields on top.

outperforms by improving $\overline{\text{OA}}$ and $\bar{\kappa}$. For visibility reasons, we only present classified fields. In particular, Fig. 2 shows one field for each of the classes considered. Looking into polygons individually, SVM-ETK significantly improves OA for the fields with the class Sorghum compared to ET, SVM-RBF, and SVM-RFK. In general, the SVM-ETK classifier slightly outperforms other classifiers in terms of $\overline{\text{OA}}$s for these polygons, and SVM-RBF gives the lowest $\overline{\text{OA}}$.

V. CONCLUSION

In this letter, we present and evaluate a novel classifier: SVM-ETK. The evaluation is done by comparing its performance against that of standard ET, SVM-RBF, and SVM-RFK. For this, we use a high spatial resolution time series over smallholder African farms expanded to create various dimensionality levels (from 56 to 10 584 features). In the experiments with low dimensionality, the average classification metrics show that the classifiers perform at about the same level although SVM-ETK slightly outperforms the results of the other classifiers. In the high-dimensional experiments, the tree-based kernels led to considerably higher overall accuracies compared to RBF while reducing the cost of the classification. Using ToRT (i.e., ET with the most extreme level of randomization) to create a kernel gives competitive results when the number of features is relatively low. This kernel reduces the computational costs, but being totally independent of the labels, it fails in our high-dimensional experiments. Overall, our results show that ET-based kernels are efficient and effective alternatives to the top-performing kernels used by the RS community. Further studies are required to evaluate the performance of the proposed methods on various benchmarking data sets.

ACKNOWLEDGMENT

The authors would like to thank all the STARS Partners and, in particular, the ICRISAT-Led Team for organizing and collecting the required field data in Mali and the STARS ITC team for preprocessing the WorldView-2 images. They would also like to thank M. Wright for his help and suggestions on the use of Extra-Trees in ranger package.

REFERENCES

[1] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, "A review of remote sensing image classification techniques: The role of spatio-contextual information," *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 389–411, Jan. 2014.
[2] M. Pal, "Kernel methods in remote sensing: A review," *ISH J. Hydraulic Eng.*, vol. 15, pp. 194–215, Jun. 2012.
[3] C. Cusano, P. Napoletano, and R. Schettini, "Remote sensing image classification exploiting multiple kernel learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2331–2335, Nov. 2015.
[4] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.
[5] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Evaluating the performance of a random forest kernel for land cover classification," *Remote Sens.*, vol. 11, no. 5, p. 575, 2019.
[6] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[7] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.
[8] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2007.
[9] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006.
[10] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," Feb. 2014, *arXiv:1402.4293*. [Online]. Available: https://arxiv.org/abs/1402.4293
[11] E. Scornet, "Random forests and kernel methods," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1485–1500, Mar. 2016.
[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[13] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.
[14] A. Samat, C. Persello, S. Liu, E. Li, Z. Miao, and J. Abuduwaili, "Classification of VHR multispectral images using extratrees and maximally stable extremal region-guided morphological profile," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3179–3195, Sep. 2018.
[15] B. Barrett, I. Nitze, S. Green, and F. Cawkwell, "Assessment of multi-temporal, multi-sensor radar and ancillary spatial data for grasslands monitoring in ireland using machine learning approaches," *Remote Sens. Environ.*, vol. 152, pp. 109–124, Sep. 2014.
[16] D. Stratoulias *et al.*, "A workflow for automated satellite image processing: From raw VHSR data to object-based spectral information for smallholder agriculture," *Remote Sens.*, vol. 9, no. 10, p. 1048, 2017.
[17] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[18] R. Aguilar, R. Zurita-Milla, E. Izquierdo-Verdiguier, and R. A. de By, "A cloud-based multi-temporal ensemble classifier to map smallholder farming systems," *Remote Sens.*, vol. 10, no. 5, p. 729, 2018.
[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm