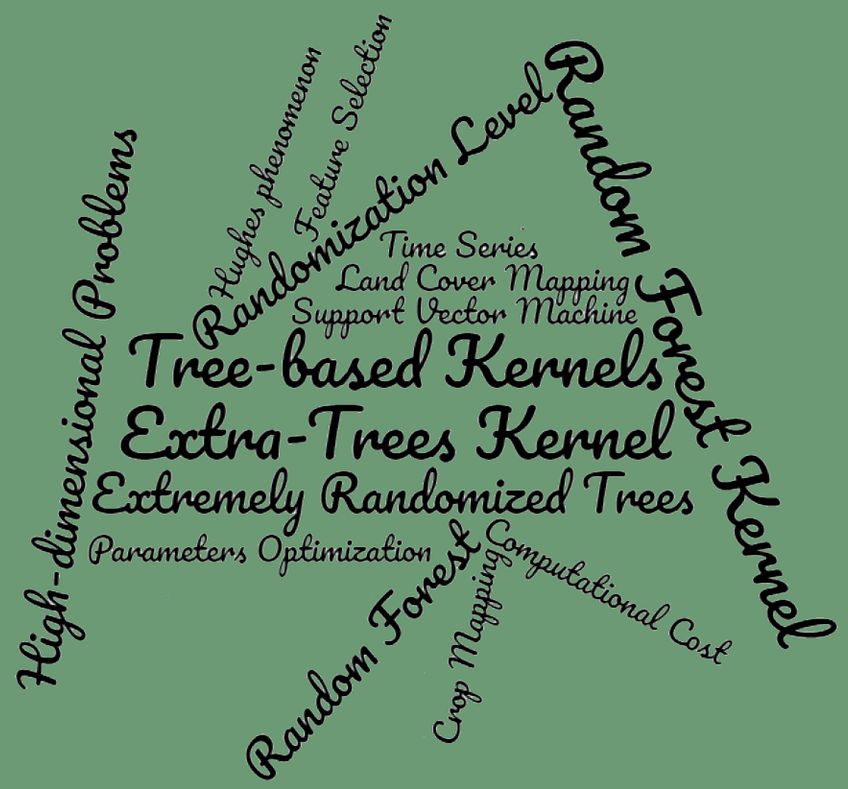


UNIVERSITY OF TWENTE. | ITC

ISBN: 978-90-365-5020-8
DOI: 10.3990/1.9789036550208
Diss.no: 383

Integrating Tree-Based Kernels and Support Vector Machine for Remote Sensing Image Classification Azar Zafari

Azar Zafari

Integrating Tree-Based Kernels and Support Vector Machine for Remote Sensing Image Classification

Integrating Tree-Based Kernels and Support Vector Machine for Remote Sensing Image Classification

Azar Zafari

INTEGRATING TREE-BASED KERNELS AND SUPPORT VECTOR MACHINE FOR REMOTE SENSING IMAGE CLASSIFICATION

D I S S E R T A T I O N

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. T. T. M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday, May 28, 2020 at 12.45

by

Azar Zafari
born on September 11, 1987
in Nahavand, Iran

This dissertation is approved by:

Prof. dr.ir. R. Zurita-Milla (promoter)

ITC dissertation number 383
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN: 978-90-365-5020-8
DOI: 10.3990/1.9789036550208
Printed by: ITC Printing Department

UNIVERSITY OF TWENTE.

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Graduation committee:

Chair/Secretary

Prof. dr. F. D. van der Meer	University of Twente

Supervisor

Prof. dr.ir. R. Zurita-Milla	University of Twente

Members

Prof. dr.ir. A. Stein	University of Twente
Dr.ir. T. A. Groen	University of Twente
Prof. dr. L. Gómez-Chova	University of Valencia
Prof. dr. B. Demir	Technical University of Berlin

Summary

There is an ever-increasing need for land cover information, since the population of the world is dependent on Earth as the source of food production and for various economic developments. Land cover maps are key inputs for policymakers in nurturing sustainable planning and management systems at the local, regional, and national levels.
Owing to advances in remote sensing (RS) technology, abundant sources of timely land cover data at various spectral and spatial resolutions have become available. Using big geo-data from recent Earth observation sensors providing very high spatial resolution (VHR) satellite images makes it possible to obtain land cover maps with higher levels of detail. However, the development of efficient classification methods for the new generations of VHR images has become one of the most challenging problems addressed by the RS community in recent years. The most important challenge associated with new generations of data is the Hughes phenomenon or curse of dimensionality that occurs when the number of features is much larger than the number of training samples. Hyperspectral images, time series of multispectral satellite images, and stacking additional features on top of the original spectral features are usually associated with the Hughes phenomenon. Tree-based ensemble learners such as the random forest (RF) and extra trees (ET) and kernel-based methods such as the support vector machine (SVM) are well-known classifiers in high-dimensional classification problems. The main objective of this dissertation is to investigate the integration of two of the most well-known and recurrently used classifiers by the geospatial community: tree and kernel-based methods.
The performance of the proposed methods is evaluated for crop classification over small-scale farms. The vast majority of low-income country farming is undertaken by smallholder farmers that often struggle to make ends meet. Currently, little is known in quantitative terms regarding the crop growth processes in smallholder farming. There are barely any systems in place that monitor such information, even though such knowledge is crucial for numerous stakeholders in the food production pyramid. Farmer communities (such as the agribusiness sector that supplies farm inputs and those marketing

farm outputs), the financial sector serving farmers, and the governmental agencies that work with farmers could utilize such information. Eventually, individual farmers could also use such information, of course, if given to them in the form of on-farm advice.
Unlike in high-income country farming (where plots are larger, only a single crop is grown, the farm inputs are well-documented, as are the weather conditions, and farm practices are more standardized), monitoring smallholder farming requires the addressing of a much higher variation in these parameters. Farm plots tend to have more irregular geometries and are often only vaguely delineated. In addition, plots are typically not formally registered in a farm cadastre. Moreover, smallholder plots include multiple crops and numerous crop varieties, there is little information about the soils, and unknown inputs are received and can be subject to variable field management. Therefore, research work in this thesis was focused on employing a number of specific VHR image sources to derive crop maps that can be used to improve the understanding of crop conditions in small-scale farms. Such image sources must be multispectral, of high spatial resolution, and the image series must be sufficiently temporally dense. This results in increasing the dimensionality of the dataset used for this study. Therefore, the research described in this dissertation concentrated on exploring the use of tree-based kernels in an SVM for land cover mapping of small-scale agriculture using VHR satellite images.

First, we studied the synergic use of RF and SVM as two well-known and recurrent classifiers for the production of land cover maps through using an RF-based kernel (RFK) in an SVM (SVM-RFK). The performance of this synergic classifier is evaluated by comparing it against using a customary radial basis function (RBF) kernel in an SVM (SVM-RBF) and standard RF classifiers. Two datasets were used to illustrate the analyses in this study—a time series of seven multispectral WorldView-2 images acquired over Sukumba (Mali) and a single hyperspectral AVIRIS image acquired over Salinas Valley (CA, USA). The features set for Sukumba was extended by obtaining vegetation indices (VIs) and grey-level co-occurrence matrices (GLCMs) and stacking them to spectral features. For Sukumba, SVM-RFK, RF, and SVM-RBF were trained and tested over 10 subsets once using only spectral features and once using the extended dataset. As benchmarking, the Salinas dataset with only spectral features was also trained and tested over 10 subsets. The results revealed that the newly proposed SVM-RFK performs at almost same level as that of the SVM-RBF and RF in terms of overall accuracy (OA) for the spectral features of both datasets. For the extended Sukumba dataset, the results showed that SVM-RFK yields slightly higher OA than RF and it considerably outperforms the SVM-RBF. Moreover, the SVM-RFK substantially reduced the time and computational cost associated with parametrizing the kernel compared to the SVM-RBF. In addition, RF

was also used to derive an RFK based on the most important features, which improved the OA of the previous SVM-RFK by 2%. In summary, the proposed SVM-RFK classier achieved substantial improvements when applied to high-dimensional data and when combined with RF-based feature selection methods; it is at least as good as the SVM-RBF and RF when applied to fewer features.

Second, we explored the connection between random forest and kernel methods by using various characteristics of RF to generate an improved design of RFK. The classic design of RFK is obtained based on the end-nodes of trees. Here, we investigated the possibility of developing the classic design of RFK by using tree depths, the number of branches among the leaves of trees, and the class probabilities assigned to samples with RF. Accordingly, we developed a multi-scale RFK which uses multiple depths of RF to create an RF-based kernel. All the obtained RFKs are evaluated by importing them into an SVM classifier (i.e., SVM-RFK) to classify the extended Sukumba dataset. The results showed that investigating the depth improves the OA of RFK, particularly for high-dimensional experiments. Other examined designs of RFKs also outperformed the RBF for the extended Sukumba datasets. Using the spectral features for Sukumba, all suggested designs of RFKs performed at almost the same level as that of the RBF kernel when they were used in an SVM.

Third, we introduced the use of ETs to create a kernel (ETK) that can be used in an SVM to overcome the limitations of RFK and RBF kernel. The use of these kernels in an SVM is also compared with the ET classifier. Four different sets of features were tested by dividing the extended Sukumba dataset. For datasets with fewer features, SVM-ETK slightly outperforms SVM-RBF and SVM-RFK. Moreover, SVM-ETK almost entirely outperforms ET. Apart from OA, the main advantage of ETK is the lower computational cost associated with parametrizing the kernel compared to the RBF and RFK. Our results showed that tree-based kernels (i.e., RFK and ETK) compete closely and yield higher OA than RBF in high-dimensional and noisy experiments. Thus, the proposed SVM-ETK classifier outperforms ET, SVM-RFK, and SVM-RBF in a majority of the cases.

Fourth, with regard to the context of open science, we include an R-function to implement the ideas of different designs of tree-based kernels evaluated in this thesis.

In a nutshell, the main conclusion of this PhD thesis is that the kernels obtained on the basis of supervised tree-based ensemble learning methods can be used as efficient alternatives to the conventional kernels in kernel-based classifications methods such as the SVM, in particular, in dealing with high-dimensional noisy problems such as mapping small-scale agriculture.

Samenvatting

Er is een steeds grotere behoefte aan landgebruiksinformatie aangezien de wereldbevolking afhankelijk van de aarde is als de bron van voedselproductie en voor diverse economische ontwikkelingen. Landgebruikskaarten zijn belangrijke input voor beleidsmakers bij het bevorderen van duurzame planning- en beheersystemen op lokale, regionale en nationale niveaus.

Als gevolg van de vooruitgang in technologie voor aardobservatie (AO) zijn overvloedige bronnen van tijdige landbedekkingsgegevens in verschillende spectrale en ruimtelijke resoluties beschikbaar gekomen. Met behulp van big geodata uit recente AO-sensoren die beelden met zeer hoge ruimtelijke resolutie (VHR) leveren, is het mogelijk om landgebruikskaarten te verkrijgen met meer details. De ontwikkeling van efficiënte classificatiemethoden voor zulke VHR-beelden is uitgegroeid tot een van de meest uitdagende problemen die de AO-gemeenschap bezighoudt. De belangrijkste uitdaging in verband met deze nieuwe AO gegevens is het Hughes-fenomeen of de dimensionaliteitsvloek. Dit doet zich voor wanneer het aantal dimensies of eigenschappen veel groter is dan het aantal trainingsobservaties. Hyperspectrale beelden, tijdseries van multispectrale satellietbeelden, en het stapelen van extra dimensies bovenop de oorspronkelijke spectrale kenmerken worden meestal geassocieerd met het Hughes fenomeen. Tree-gebaseerde ensembles zoals de random forest (RF) en extra trees (ET), en kernel methoden zoals support vector machines (SVM) zijn bekende classificatiemethoden voor hoogdimensionele classificatieproblemen. De belangrijkste doelstelling van dit proefschrift is de integratie van twee van de meest bekende en veelgebruikte classificaties door de geospatiale gemeenschap: tree- en kernel-gebaseerde methoden.

De prestaties van tree- en kernel-gebaseerde methoden worden beoordeeld voor gewasclassificatie op kleinschalige landbouw. In landen met lage inkomens wordt de overgrote meerderheid van boerderijen beheerd door kleinschalige boeren die worstelen om rond te komen. Er is nog weinig kwantitatief bekend over her verloop van de groei van gewassen in kleinschalige landbouw. Er zijn nauwelijks

systemen die dergelijke informatie monitoren, ook al is die kennis cruciaal voor tal van stakeholders in de voedselproductiepiramide. Uiteindelijk zouden de boeren zelf ook dergelijke informatie kunnen gebruiken als ze die krijgen in de vorm van gewasteeltadvies.

In tegenstelling tot landbouw in hoge-inkomenslanden waar de percelen groter zijn en er een gewas verbouwd wordt per perceel, waar de inputs en de weersomstandigheden goed gedocumenteerd, en waar landbouwpraktijken meer gestandaardiseerd zijn, vereist het monitoren van kleinschalige landbouw veel aandacht. In kleinschalige landbouw hebben percelen meestal meer onregelmatige geometrieën en zijn vaak slechts vaag afgebakend. Daarnaast zijn percelen doorgaans niet formeel geregistreerd in een agrarisch kadaster. Bovendien bevatten kleinschalige percelen meerdere gewassen en tal van gewasvariëteiten, is er weinig informatie over de bodem, en is onbekend welke en hoeveel inputs (zoals irrigatie en bemesting) worden gebruikt en kunnen de velden onderhevig zijn aan variabel beheer.
Daarom is dit proefschrift gericht op het gebruik van beelden van zeer hoge resolutie om gewaskaarten af te leiden die kunnen worden gebruikt om het inzicht in de gewasomstandigheden in kleinschalige landbouw te verbeteren. Dergelijke beeldbronnen moeten multispectraal en van zeer hoge ruimtelijk resolutie zijn en de tijdserie moet voldoende data bevatten. Dit resulteert in het vergroten van de dimensionaliteit van de data gebruikt in dit proefschrift. Daarom ligt de focus in het onderzoek hier beschreven op het verkennen van het gebruik van tree-gebaseerde kernels in SVM.

Als eerste bestudeerden we het synergetisch gebruik van RF en SVM als twee bekende en terugkerende classificatoren voor de productie van landgebruikskaarten door het gebruik van een op RF gebaseerde kernel (RFK) in een SVM (SVM-RFK). De prestaties van deze synergetische classificator worden geëvalueerd door te vergelijken met een gebruikelijke radiale basisfunctie (RBF) kernel in een SVM (SVM-RBF) en een standaard RF-classificator. Twee datasets zijn gebruikt om de analyses in deze studie te illustreren - een tijdreeks van zeven multispectrale WorldView-2-beelden verkregen over Sukumba (Mali) en een hyperspectraal AVIRIS-beeld verkregen over Salinas Valley (VS). De spectrale eigenschappen van de Sukumbabeelden zijn uitgebreid door het verkrijgen van vegetatie indices (VI's) en grey-level co-occurrence matrices (GLCM' s). Voor Sukumba werden de SVM-RFK, RF en SVM-RBF classificatoren getraind en getest over 10 subsets van data met originele en uitgebreide eigenschappen. Als benchmarking is de Salinas-dataset met alleen spectrale eigenschappen ook getraind en getest over 10 subsets. Uit de resultaten bleek dat de nieuw voorgestelde SVM-RFK op bijna hetzelfde niveau presteert als dat van de SVM-RBF en RF in termen van overall accuracy (OA). Voor de uitgebreide Sukumba-dataset toonden de resultaten aan dat SVM-

RFK een iets hogere OA oplevert dan RF en het presteert aanzienlijk beter dan de SVM-RBF classificator. Bovendien is de SVM-RFK sneller dan SVM-RBF door de benodigde optimalisering van de parametrisering van de RBF kernel. Daarnaast werd RF ook gebruikt om een RFK af te leiden op basis van de belangrijkste eigenschappen, wat de OA ten opzichte van de vorige SVM-RFK met 2% verbeterde. Samengevat behaalde de voorgestelde SVM-RFK classificator substantiële verbeteringen wanneer toegepast op hoog-dimensionale gegevens en in combinatie met de RF ingebouwde eigenschappenselectiefunctionaliteit is het minstens zo goed als de SVM-RBF en RF wanneer toegepast op problemen met lage dimensionaliteit.

Ten tweede hebben we het verband onderzocht tussen RF en kernelmethoden door verschillende kenmerken van RF te gebruiken om een verbeterd ontwerp van RFK's te maken. Het klassieke ontwerp van RFK is gebaseerd op de eindknopen van de bomen binnen de RF. Hier hebben we de mogelijkheden van ontwikkeling van verbeterde RFK's onderzocht door boomdiepten, het nummer van takken tussen de bladeren van de RF bomen en de RF toegewezen classificatiewaarschijnlijkheden te gebruiken als overeenkomstmetrieken. Daarmee hebben we verschillende "multi-scale" kernels ontwikkeld. Alle verkregen RFK's zijn gebruikt in een SVM-classificator (d.w.z. SVM-RFK) om de uitgebreide Sukumba-dataset te classificeren. De resultaten lieten zien dat RFKs die gebruikmaken van de boomdiepte een betere OA hebben, vooral voor hoog-dimensionale experimenten. De andere RFK's presteerden ook beter dan de standaard RBF kernel voor de uitgebreide Sukumba-datasets. Met alleen de spectrale eigenschappen voor de Sukumba dataset presteren alle RFK ontwerpen op bijna hetzelfde niveau als de RBF-kernel.

Ten derde hebben we het gebruik van extra trees (ET) geïntroduceerd om een kernel (ETK) te maken die kan worden gebruikt in een SVM om de nadelen van de RFK- en RBF-kernels te overwinnen. Het gebruik van deze ETK in een SVM wordt ook vergeleken met de ET classificator. Vier verschillende aantallen eigenschappen van de Sukumba dataset zijn getest om het effect van data dimensionaliteit te bestuderen. Voor de datasets met lagere aantallen eigenschappen presteert SVM-ETK iets beter dan SVM-RBF en SVM-RFK. Bovendien presteert SVM-ETK bijna altijd beter dan ET. Afgezien van een betere OA is het belangrijkste voordeel van ETK de lagere rekentijdkosten die gepaard gaan met de optimalisering van de parametrisering van de RBF kernel en het optimaliseren van de RFK. Onze resultaten tonen aan dat RFK en ETK nauw met elkaar concurreren en een hogere OA opleveren dan RBF in experimenten met hoge dimensies en ruis. De voorgestelde SVM-ETK presteert in de meeste gevallen beter dan ET, SVM-RFK en SVM-RBF.

Ten vierde hebben we in het kader van Open Science een R-functie

toegevoegd om de tree-gebaseerde kernels geëvalueerd in dit proefschrift te implementeren en testen.
In een notendop is de belangrijkste conclusie van dit proefschrift dat tree-gebaseerde kernels kunnen worden gebruikt als efficiënte alternatieven voor conventionele kernels (zoals RBF) in op kernel-gebaseerde classificatiemethoden zoals SVM. Dit geldt in het bijzonder bij het omgaan met hoog-dimensionele en ruizige problemen zoals het in kaart brengen van kleinschalige landbouw.

Acknowledgments

Here, it comes the end of long journey of the PhD. I still remember how enthusiastic and motivated I was when I began this journey. I am grateful for being selected as a PhD candidate at the University of Tewente, ITC faculty, GIP department. First and foremost, I express my heartfelt gratitude to Prof. Menno Jan Kraak, who believed in me and selected me as a PhD candidate in the GIP department. I deeply appreciate his support and encouragement during all these years. Further, I am sincerely grateful to Dr. Ali Abkar for supporting me in beginning and continuing this journey. I would also like to acknowledge the European Commission's Erasmus Mundus (SALAM2) and ITC foundation for awarding me a PhD fund and providing financial support during my PhD.

After undergoing several challenges during the five years of my research, I am happy that I remained dedicated and finally completed this thesis. Throughout the duration of my PhD, I received support from numerous people and without them this thesis would not be achieved. My research was conducted under the supervision of Prof. Raul Zurita-Milla; I am thankful to him for his scientific input and his effort to further push the boundaries of science in my research work.

I want to thank my colleagues in my research group for their advice and support over these years: Rosa, Irene, Hamed, Noorhakim, and Emma. I am also grateful to the staff of GIP and EOS departments for their help and feedback—special thanks goes to Rolf de By, Claudio Persello, Luis Calisto, and Andre Mano. I would also like to thank Prof. Saeid Homayouni at Centre Eau Terre Environnement, INRS-Quebec for his advice and help.

I am also incredibly grateful for the support of my amazing friends during these years. Parya, words cannot describe how much our friendship and your support during these years means to me, THANK YOU. Caroline, thank you for all your support and sweet friendship over these years, it means a lot to me. Shima, thank you for being there for me during the difficult times and for all your support and advice. Nina, thank you for your time, support, and encouragement over our long calls, it means a lot to me. Sara, Adish, and Lydia talking to you has been so inspiring for me—thank you for all your help

and advice. Vahid, Khatareh, Zahra, and Sajad, you all brought so many joyful moments to my life—thank you for all your help and sweet friendship. Manual, Ieva, Charis, Nga, thank you all for your help, advice, and for the pleasant times during our lunch and coffee breaks. I am also thankful to my friends back in Iran. Nasim and Negar, thank you and your parents for the extreme kindness you all showed to me and for your valuable friendship.

Last but not least, family members are the most beautiful gifts life gives. I am grateful to my parents and my brother, Ashkan, for their never-ending love and support during the entire time. I am also grateful to my in-laws; special thanks goes to Farhad for his support, advice, and sharing his personal experiences of his PhD journey with me.

Finally, my heartfelt and profound gratitude goes to the love of my life. Farzad, meeting you was the sweetest thing happened to me over these years and during all the time. Your optimism and enthusiasm enabled and encouraged me to complete my PhD. Thank you for believing in me and for always being so understanding and supportive.

In memory of Zahra Naghibi and all innocent victims of the flight PS752. Zahra was close to obtaining her PhD degree if her life was not taken away from her.

Contents

List of Figures

List of Tables

List of Nomenclatures

Abbreviations

AVIRIS	Airborne Visible Infrared Imaging Spectrometer
AVHRR	Advanced Very High Resolution Radiometer
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
DVI	Difference Vegetation Index
EVI	Enhanced Vegetation Index
ENVISAT	Environmental Satellite
GLI	Green Leaf Index
HSIC	Hilbert–Schmidt Independence Criterion
LSP	Land Surface Phenology
ML	Maximum Likelihood
MIF	Most Important Features
MSAVI2	Modified Soil-Adjusted Vegetation Index
NDVI	Normalized Vegetation Index
NN	Neural Networks
PRI	Photochemical Reflectance Index
OA	Overall Accuracy
OSAVI	Optimized Soil-adjusted Vegetation Index
RBF	Radial Basis Function
SVM-RBF	Radial Basis Function Support Vector Machine Classifier
RF	Random Forest
RF-BD	Best Depth Random Forest Classifier
RF-FG	Full-grown Random Forest Classifier
RFK	Random Forest Kernel
RFK-BD-SVM	Best Depth Random Forest Kernel Support Vector Machine Classifier
RFK-FG-SVM	Full-grown Random Forest Kernel Support Vector Machine Classifier
RS	Remote Sensing
RKHS	Reproducing Kernel Hilbert Space
RVI	Ratio-based Vegetation Indices
SAVI	Soil-adjusted Vegetation Index

SD	Standard Deviation
SVM	Support Vector Machine
TCARI	Transformed Chlorophyll Absorption Reflectance Index
CT	Classification Tree
VI	Vegetation Index
WBI	Water Band Index
WV2	WorldView-2
GLCM	Gray-level Co-Occurrence Matrix
ET	Extra-Trees
ETK	Extra-Trees Kernel
ToRT	Totally Randomized Trees
ToRTK	Totally Randomized Trees Kernel
VHR	Very High Spatial Resolution
B	Spectral Features
BVI	Spectral &VIs Features
BVITVI	BVI and GLCM Textures of VIs
ALL	BVI and GLCM Textures of Spectral and VIs

Symbols

RFK_{Nd}	Node-based (Classic) Random Forest Kernel
RFK_{Br}	Branch-based Random Forest Kernel
RFK_{Prob}	Probability-based Random Forest Kernel
κ	Cohen's kappa index
RFK_{Nd^*}	RFK_{Nd} at an optimized depth
RFK_{Prob^*}	RFK_{Prob} at an optimized depth
$RFK_{\overline{Nd}}$	Multi-scale RFK_{Nd}
$RFK_{\overline{Prob}}$	Multi-scale RFK_{Prob}

Introduction 1

1.1 Remote sensing image classification

1.1.1 Background

Understanding and quantifying land cover information is important for human beings, since the increasing population of the world is dependent on Earth as the source of food production and various economic developments [1, 2, 3]. Land cover is used to characterize and describe the Earth's surface in terms of soil, vegetation layers, and man-made structures [4, 5]. Land cover maps are of importance to policymakers in planning and management at the local, regional, and national levels [6]. Up-to-date and accurate land cover information makes a significant contribution to the development of sustainable economic and environmental plans [6]. In addition, land cover maps are key components for studying several governmental concerns such as flooding, soil erosion, run-off, climate change, and agricultural monitoring [7]. Therefore, it is essential to monitor ongoing changes and processes related to land cover patterns over time in order to ensure sustainable development [1, 2, 3].

The necessity of acquiring regular, precise, and accurate information regarding the Earth's surface over vast areas has resulted in the development of remote sensing (RS) over time [8, 9, 6]. The term RS was used for the first time in the 1950s and refers to obtaining information from objects without direct physical contact with them [6, 8]. Sputnik 1 was the first man-made satellite developed by Russia in 1957, and the first photo from space was obtained by United States Explorer 6 in 1959. Further, Landsat 1 is the pioneering United States RS satellite program that has acquired a continuous supply of synoptic, multispectral data. Landsat 1 is a key milestone for monitoring Earth and its natural resources in the history of the RS [10, 6]. The advances in RS have enabled the obtaining and monitoring of land cover information at different temporal and spatial resolutions; this opens opportunities for a wide range of operational applications in the environmental and agricultural domains [5]. Since the first satellite image, a series of sensors called Landsat thematic mapper, Advanced Very High Resolution Radiometer (AVHRR), Satellite Pour l'Observation de la Terre (SPOT1), Moderate Resolution Imaging Spectroradiometer (MODIS), Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Environmental Satellite (ENVISAT), and SPOT5 have been launched to map key landscape features and resources [6]. During this course of advancement, the spatial resolution of images has improved from 1.1 kilometers to 5 meters and multispectral information was made available through SPOT and ASTER images.

DigitalGlobe optical sensors—such as Ikonos, QuickBird, and Worldview-2—provided multispectral imagery with very high resolution (VHR) from two-to-four meters and a short revisit time. Currently, Worldview-3 provides multispectral imagery with a spatial

resolution of 1.2 meters. However, the narrow field of view corresponding to DigitalGlobe sensors limits their ability to capture the entire Earth in a timely fashion [11]. The latest generation of RS data is provided by Cubesat satellites—for example, the doves of Planet Labs that provide a unique combination of VHR multispectral imagery (i.e., a spatial resolution of five meters) and full-Earth coverage repeat rate (i.e., one day) [11]. Using big geo-data from recent Earth observation sensors enables the oversight of tasks related to environmental and agriculture monitoring in greater detail [12, 13]. In order to achieve these tasks, the development of effective data processing techniques for the latest generation of very high spatial resolution optical sensors has become one of the most challenging problems addressed by the RS community in recent years. Addressing these challenges enhances human beings' understanding of land cover information and results in the creation of sustainable management systems to mitigate issues related to land cover in various urban, environmental, and agricultural contexts.

1.1.2 Remote sensing for land cover mapping

Land cover maps have been created from a variety of RS data sources [14, 15, 16, 17, 18] and for a variety of applications, including socioeconomic, natural resources, agricultural, environmental, urban, and regional monitoring and planning [19, 20, 21, 22].
A few of the most important applications of RS are the obtaining of agricultural information and crop mapping and monitoring. Further, food security is one of the main concerns of governments and policymakers, particularly in developing countries [5]. The world population is expected to reach 9.3 billion in 2050 [23, 5]; to feed this population, the Food and Agriculture Organization estimates that the world's agricultural production will need to increase by approximately 70% by 2050 [24, 5] from the 2005 production levels. The demands of an increasing world population leave no doubt of the need to improve sustainable agricultural production in order to minimize both monetary and environmental costs [13, 5]. Geographic information systems, satellite imagery, and field data measurements are key in developing an information management system for agriculture monitoring. Crop maps are key initial components of agricultural monitoring, and satellite imagery has proven to be effective in revealing the type and variation of spatial and temporal characteristics of crop production. In large and single-type crop fields, multidate hyperspectral or time series of multispectral imagery are often used for crop mapping [25]. In small-scale agriculture, farm plots have irregular shapes and vaguely delineated boundaries. Smallholder farms commonly contain multiple crops and crop varieties and involve variable field management. Using lower spatial resolutions for crop mapping in smallholder farms can cause a single

pixel value to represent multiple crops. Therefore, monitoring smallholder farms requires image sources that are of very high spatial resolution; moreover, several studies have also revealed that the image series must be temporally sufficiently dense. Further, mapping methods applied to time series images have been proven to perform generally better than single-date mapping methods [26, 27, 28]. However, each type of satellite image has its own limitations, since there is an inevitable trade-off among spatial, spectral, and temporal resolutions. In addition, due to persistent cloud coverage during the growing season, the available image information is often sparse.
For rural regions where small-scale farming is predominant, it is necessary to expand our knowledge and develop RS image classification techniques so that they can address the complexity of the crop mapping in such areas.

1.1.3 Common remote-sensing image classification methods

RS image classification methods group image pixels into one of several land cover classes to reveal meaningful information [29]. Classifiers can be categorized as pixel-based and object-based. A pixel-based classifier assigns each pixel to one class based on spectral information [30]. An object-based classifier derives objects that consist of several pixels by considering the shape and texture variations among them [31]. One common characteristic of object-based classifications is that they are based on image segmentation [32, 33]. Image segmentation aims at building homogeneous blocks of pixels that are object candidates for further steps of processing [34].
Segments are generated using a criteria of homogeneity and have additional spectral and spatial information compared to pixels. Both pixel-based and object-based approaches, accompanied with machine learning methods, are widely applied for numerous land cover mapping applications [35]. Examples of applications using pixel-based approaches are forest mapping [36, 37], carbon emission monitoring [38, 39], climate dynamics [40, 41], biodiversity mapping [42, 43], damage assessment and disaster management [44, 45, 46], agricultural mapping [47, 28], and water and wetland monitoring [48, 49].
Focusing on crop mapping applications using pixel-based approaches, patterns of vegetation dynamics identified from time series images have been successfully used to classify crops in different study areas [50, 51, 28, 52]. A review of object-based classification approaches for various applications, including land cover mapping, urban mapping, forest cover types, shrub changes, texture analysis, structural damage, and change detection through the use of various satellite platforms is presented in [34]; several studies utilize object-based classification approaches for crop mapping [53, 54, 55, 56]. However, over-segmentation and under-segmentation errors that affect the accuracy of the classifications are known drawbacks

of the object-based approaches [57], and disregarding spatial and textural information is a limitation of pixel-based approaches [58]. When the resolution is coarse, pixel-based and sub-pixel approaches are recommended, while approaches based on extracting information regarding the neighborhood of the pixels are suitable when there is increased spatial resolution [55].

Within pixel and object-based approaches, different classification methods have been successfully used for land cover mapping. These classification methods are mainly categorized into two groups: supervised and unsupervised. In unsupervised classification, a clustering algorithm such as ISODATA or K-means divides the spectral data into groups based on statistical information derived from the image [59]. In supervised classification, sufficient additional reference data is used to train related classifiers, such as maximum likelihood, minimum distance, artificial neural networks, and decision trees [59]. According to the literature, supervised classifiers often outperform unsupervised classifiers [60]. The reason for this is that unsupervised classifiers require clear spectral separability between the classes of interest, which may not always be the case [60]. [61] examine the C4.5 decision tree, logistic regression, support vector machine (SVM), and neural network methods for crop classification in California; they conclude that the SVM outperforms other methods. Further, the use of vegetation indices (VIs) improves the accuracy of vegetation mapping for various classifiers, as VIs provide specific information to distinguish various types of vegetation. A few examples of VIs are normalized difference vegetation index, enhanced vegetation index, difference vegetation index, and ratio vegetation index.

The advent of recent RS technologies has led to the improvement of spatial, spectral, and temporal resolutions of satellite images; this offers new possibilities for very accurate mapping of the environment apart from the new challenges that an efficient supervised classifier must address. The most important challenge associated with new generations of data is the Hughes phenomenon or curse of dimensionality that occurs when the number of features is much larger than the number of training samples [62]. The Hughes phenomenon often occurs when combining abundant sources of data—including multi-source satellite images, hyperspectral images, time series of multispectral satellite images—and where spatial, spectral, and temporal features are stacked on top of the original spectral channels for modeling additional information sources [63].
Pixel reflectance is not only a function of the land cover captured in a particular pixel but also of the land cover in surrounding pixels. Therefore, the information regarding the neighborhood of the pixels must be extracted in order to improve our understanding of land cover. To this end, features can be defined as attributes that are calculated using functions of the original measurement variables, which

are useful for classification problems [64]. In the present research thesis, various types of features are extracted to include information of pixel neighborhoods in a pixel-based classification approach.

Feature extraction is the process of defining a set of features, or image characteristics, which will meaningfully represent the information that is important for analysis and classification [64]. Textural features (texture) are the most common features used to describe the neighborhood of a pixel. As scholars have noted, "texture is generally taken to mean whatever structure exists within a semantic region" [65], and structure represents properties and relationships of image components. Texture analysis includes texture recognition (feature extraction), segmentation, and classification in RS applications. There are several texture descriptors. The methods of texture extraction can be categorized into four groups[66, 67]: structural methods, statistical methods, model-based methods, and transform methods.
Statistical methods represent the spatial distribution of gray values in an image by deriving a set of statistical measures of the arrangement of intensities in a region. First-order statistics assess characteristics (e.g., average and variance) of individual pixel values, while higher-order statistics estimate properties of two or more pixel values relative to each other. The most important second-order statistical features for texture analyzing are gray-level co-occurrence matrices (GLCM). The GLCM functions characterize the texture of an image by computing how often pairs of pixels with specific values and in a specified spatial relationship (i.e., pixel relationships of varying direction and distance) occur in an image, creating a GLCM, and then extracting statistical measures (e.g., contrast, correlation, homogeneity, and energy) from this matrix [68, 69]. In this study, we focus on GLCM textures, which have been reported to enhance crop classification results and are successfully applied to tackle different RS image classification problems [68, 69, 70, 71, 72].
Stacking all spectral and spatial features further increases the dimensionality of the datasets.
The performance of kernel-based methods are widely well-reported among supervised classifiers in handling high-dimensional data [73, 74]. Kernel-based methods are successfully applied in the context of hyperspectral and multi-temporal image classification [59, 75]. The SVM is the most well-known kernel-based method that has been shown to outperform classical supervised classifiers for high-dimensional problems in several studies [76, 77]. Another group of supervised classifiers proven to perform well in handling high-dimensional data is the tree-based ensemble learning schemes—in particular, random forest (RF) [78, 79, 80] and extremely randomized trees [81]. The following sections provide a detailed background on tree-based ensemble classifiers and the SVM as the most well-known classification methods for their perform-

ance in dealing with high-dimensional land cover mapping problems [82, 73, 83, 84].

1.2 Tree-based ensemble classifiers

Ensemble methods generate multiple base learners and combine them to obtain better performance than that from any single constituent learning algorithms. Ensemble methods assign labels to new data samples by taking a weighted or unweighted vote of predictions [85]. Two common ensemble techniques are boosting and bagging [86]. Boosting sequentially builds different base learners trained on the basis of whole training samples [86]. In boosting, samples are weighted on the basis of the previous classifier's success [86]. After each training step, the weights of misclassified samples are increased to emphasize the most difficult cases [86]. Boosting uses the weighted average votes of base learners for a new prediction. On the other hand, bagging techniques in parallel generate multiple base learners and train them based on bootstrap samples of training data [87]. Bootstrap sampling is random sampling with replacements. Bagging uses voting to aggregate the output of base learners, thereby reducing the variance of the prediction [87]. Benchmarking results show that boosting approaches generally provide higher accuracies compared to bagging approaches [88]. However, the optimization of boosting approaches is more time-consuming and difficult because of the sequential process of training and higher number of training parameters. Moreover, boosting is more sensitive to overfitting, particularly if the training samples are noisy [86].

Classification trees (CTs) are the most popular base learners for generating ensembles introduced by Leo Breiman [89]. CTs are supervised tree-based (i.e., do not assume a particular data distribution) non-parametric classification (and regression) learners that are applied in several land cover classification problems [89, 90, 91, 92]. CTs utilize a hierarchical tree-based approach that divide the feature space of training data recursively into child nodes, until each of them contains very similar samples or until one stopping condition is met [89]. CTs divide each node by extensively searching for a best cut-point. Although CTs are simple to interpret and operate, a few major drawbacks are that they tend to overfit, are sensitive to noise and size of training data, and require pruning [93]. In order to improve the classification accuracies of CTs, [94] introduced RF, which is a group of CTs. RF is a well-known tree-based ensemble learner that works based on the bagging scheme. RF works on the concept of utilizing multiple unpruned CTs that are trained on the basis of bootstrap samples of training data and variables, with the remaining samples called out-of-bag samples that contribute to

evaluating classification accuracy. RF uses a maximum voting rule from the prediction of all CTs to assign class labels to new samples [93, 94]. RF is a non-parametric approach like its components, the CTs. Moreover, RF can be easily trained and implemented, as setting its parameters to their default values stabilize the error of the classification in most classification problems[95]. In addition, RF is not sensitive to overfitting and requires a small sample size with high-dimensional input compared to CTs and several other classifiers [95, 96, 93]. Several studies have shown that RF outperforms traditional machine learning classifiers and provides comparable classification accuracies, while requiring fewer user-defined parameters compared to SVM [97]. In addition, RF is fast and computationally much lighter compared to SVM in both the training and predicting phases [98]. Another tree-based bagging scheme is extremely randomized trees, known as Extra-Trees (ET), which has been reported to outperform the SVM and RF in several studies [99, 100]. ET also generates an ensemble of unpruned decision trees like RF, but the level of randomization in ET is higher and the computational load of ET is smaller compared to that of RF [99, 100]. In addition, ET employs all training samples rather than bootstrap subsets to grow the trees [99, 100].

1.2.1 Tree-based ensemble learners: Pros and cons

Several strong features of tree-based ensemble learners make them a good choice for RS image classification. First, the default parameter configurations turned out to be optimal in terms of accuracy, which highlights the fact that these methods are almost parameter-free but still able to learn non-linear data [101, 102]. Second, their computing times are also rather competitive on rather large and high-dimensional datasets, both for training and making predictions [101, 102]. Third, tree-based ensemble learners can be used to obtain feature importance measures based on total decrease in node impurity from removing each feature, averaged over all trees. Obtaining the feature importance measure can provide some insight regarding the problem at hand [103, 101, 102]. Fourth, the structure of the tree-based ensemble learners creates data partitions, and similarity among samples can be quantified on the basis of whether or not the samples end up in the same partition and the similarity values among samples can be used to define tree-based kernels. The connection of tree-based ensemble learners and kernel methods is emphasized in several studies [103, 104, 105, 106]. Last, tree-based ensemble learners can be used to detect outliers in data on the basis of the similarity values among the samples [107, 108, 109]. On the downside, tree-based ensemble learners are difficult to visualize and interpret in detail and they have been observed to overfit for certain noisy datasets [101].

1.3 Support vector machine

For linearly separable data samples, an SVM aims to find an optimal location for a hyperplane in an N-dimensional space (N being the number of features) that partitions training samples into a finite number of classes [29, 110]. Generally, all training samples are not used in defining the hyperplane; this is mainly done by a subset of points that is located closest to the hyperplane (called support vectors). The optimal location for a hyperplane is where it generates the greatest margin (i.e., the sum of the distances to the hyperplane from support vectors) between classes. The problem of maximizing the margin is solved using standard quadratic programming optimization techniques. SVM tolerates a few misclassified samples in the trade-off with identifying a hyperplane that maximizes the margin. This trade-off is controlled with a regularization parameter called the C parameter [29, 110]. If the classes are nonlinearly separable in the original high-dimensional space, the original data is mapped into a higher-dimensional feature space using a kernel function, thereby formulating a linear classification problem in that feature space [111, 112]. There are different types of nonlinear kernels, such as sigmoid, polynomial, and radial basis function (RBF) kernels. Among all types of kernel functions, the most well-known is the RBF kernel ($k(x_i, x_j) = \exp(-(x_i - x_j)^2/2\sigma^2)$, where σ is the bandwidth and controls the dependency of the hyperplane on the training samples that are far from and close to the hyperplane). SVM using the RBF kernel requires the fixing of two parameters, σ and C. These parameters are typically optimized by cross-validation of a grid space of (C, σ) [111, 112, 29].
The characterization of an SVM as a non-parametric kernel-based learning technique is an appealing classification technique in RS land cover classification [113, 29]. The successful use of SVM is reported for land cover classification of monotemporal [114], multitemporal [115], multisensor [116], and hyperspectral [117] datasets.

1.3.1 SVM using an RBF kernel: Pros and cons

An SVM using an RBF kernel that represents a Gaussian function is well-known because of its capability of handling nonlinear high-dimensional data [73]. However, the main challenge of this classifier is the selection of the hyperparameters, since hyperparameters strongly influence classification results. The hyperparameters are typically selected by defining appropriate ranges for each of them to find the best configuration through a computationally extensive cross-validation process. This approach is not efficient for large datasets; therefore, Bayesian hyperparameter optimization is employed in these cases [118]. Bayesian hyperparameter optimiza-

tion builds a probability model of the objective function and optimizes the probability model to select the most promising hyperparameters of the true objective function. Further, Bayesian hyperparameter optimization reduces the computational time by utilizing an iterative approach that maintains a record of previous iterations for searching the next part of the feature space. However, Bayesian hyperparameter optimization remains a complex non-convex optimization problem. Moreover, the performance of RBF in an SVM decreases significantly when the number of features is much higher than the number of training samples—particularly if there are correlated and non-informative (i.e., noise) features in the dataset. Several studies use various feature selection approaches to overcome this downside of using the RBF kernel in SVM (i.e., SVM-RBF). The main feature selection methods used with SVM can be divided into filters, wrappers, and embedded methods [119], but each group has its own drawbacks. Filters select the features that are independent of the classifier, wrappers tend to be computationally expensive, and embedded methods require building multiple models [119]. Recently, several studies have shown that the use of tree-based ensemble learners as feature selection methods for an SVM is efficient and competitive [119, 120].

1.4 Integrating tree-based ensemble learners and the SVM classifier

The SVM and ensemble classifiers are the most prominent supervised classifiers used by the RS community in high-dimensional classification problems. In order to combine the power of an SVM and ensemble classifiers and to overcome the downsides of each classifier, several studies present an integrated approach employing both classifiers. For example, using an RF-based feature selection method for dimensionality reduction of hyperspectral data [16, 121] leads to higher overall accuracy (OA) for SVM-RBF. In [122], a hybrid SVM-based approach that is inspired by RF and boosting classifiers is used for RS data classification. The idea in this hybrid approach is to subdivide the input dataset into smaller subsets and classify individual subsets using the SVM classifier. In an iterative approach, boosting is used in each subset to update a weight factor for every data item in the dataset. The weight factors are increased if misclassification has occurred and vice versa. Inspired by RF, the outcome for the complete dataset is obtained by implementing a majority voting mechanism to the individual subset classification outcomes [122].

Another approach that has been used to integrate SVM-RBF and RF is dynamic classifier selection in which a pool of base classifiers

with different parameters and initializations are generated and the base classifiers are selected on the fly, in accordance with each new sample to be classified [123, 124]. The outputs obtained by the selected classifiers are fused in accordance with a combination rule, such as that employed in the majority voting scheme [125, 124]. In [124], an ensemble of five base classifiers—including an SVM-RBF and an RF—is used that improves the OAs of the classifications compared to the base classifiers in mapping crops from a time series of VHR images. However, ensemble methods demand high computational capability [124].

Although the integration of an SVM and ensemble classifiers for RS image classification is addressed in several works with their own pros and cons, there is a knowledge gap in integrating these classifiers through the kernel connection. The potential of RF and other ensemble learners to be reformulated as kernel methods is emphasized in several studies [103, 104, 105, 106]. Therefore, prevalent ensemble learning methods like RF and most recent approaches like ET can be related to kernel-based methods, like an SVM, through the kernel connection. The strong features of tree-based ensemble learners, like feature importance, can also be exploited to enhance the design of tree-based kernels. In this thesis, SVM and prevalent ensemble classifiers (i.e., RF and ET) as supervised classification frameworks and the most prominent classifiers used by the RS community are applied for crop classification. We evaluate whether the combination of these classifiers through kernel connections can help overcome the limitations of each classifier while maintaining their strong points.

1.5 Research objectives and questions

The main research objective of this PhD thesis is to integrate two of the most prevalent classifiers used by the geospatial community: tree-based methods like RF and kernel methods like an SVM. In particular, this thesis concentrates on exploring the use of tree-based kernels in SVMs by addressing the following specific objectives and research questions:

1. **Evaluating the potential of using an RF-based kernel (RFK) to classify remotely sensed images**
   a) How do the classification results of SVM-RFK compare to those obtained by standard RF and SVM-RBF classifiers?

   b) How do RF's most important parameters affect the performance of SVM-RFK classifier?

c) How does RF's feature selection impact the classification results of SVM-RFK classifier?

2. **Investigating the pros and cons of alternative RFK formulations**

   a) How does the use of a branch-based distance compare to the standard similarity metric used to calculate the RFK?

   b) How does designing a multiscale RFK based on using multiple depths of RF compare to the standard similarity metric used to calculate the RFK?

   c) How does designing a multiscale RFK based on using multiple depths and class probabilities compare to the standard similarity metric used to calculate the RFK?

3. **Exploring the use of an alternative tree-based classifier, namely ET, to derive tree-based kernels**

   a) What is the influence of ET's most important parameters on the classification accuracy of the corresponding SVM-ETK classifier?

   b) How does the level of randomization influence the performance of the ET and SVM-ETK classifiers?

   c) How do the classification results of SVM-ETK compare with those obtained by the standard ET, SVM-RBF, and SVM-RFK classifiers?

4. **Present an R function that implements the various designs of tree-based kernels evaluated in this thesis to support the shift toward open science**

1.6 Thesis outline

This thesis has a total of six chapters, including the Introduction and Synthesis. Apart from the Introduction and Synthesis, three core chapters are based on papers that have been published in peer-review journals and are independently structured as Abstract, Introduction, Data and Study Area, Methods, Experiments, Discussion, and Conclusions. The five chapters, after the introduction, can be summarized in the following manner:

- **Chapter 2** presents a classification method based on integrating RF and SVM for the production of land cover maps through the use of an RFK in an SVM. The performance of the synergic classifier is evaluated for crop classification over agricultural lands by comparing it against using a radial basis function (RBF) kernel in an SVM (SVM-RBF) and standard RF classifiers. Two VHR datasets, including a time series of multispectral Worldview-2 images over Sukumba, West Africa, and a single hyperspectral AVIRIS image over Salinas, California are used for illustration in this chapter.
- **Chapter 3** explores the relationship between RF and kernel methods by investigating the various characteristics of RF in order to generate an improved RFK design. Accordingly, this chapter presents a multi-scale RFK that uses multiple depths of RF to create an improved design for RFK. Further, in this chapter, the performance of the newly designed RFKs is evaluated by comparing them with the performance of RBF and classic design of RFK in an SVM classifier to classify crops over the study area of Sukumba.
- **Chapter 4** presents the use of ET to create an ETK that is introduced in an SVM for land cover classification. In this chapter, the performance of ETK is benchmarked against that of RBF and RFK in an SVM and against the standard ET classifier. These methods are evaluated for crop classification in small-scale agriculture over the study area of Sukumba.
- **Chapter 5** presents an R-function implementing the different designs of tree-based kernels evaluated in this thesis, accompanied with a documentation of the function.
- **Chapter 6** summarizes the results obtained from Chapters 2–5, answers the research questions, presents research reflections, discusses the main contribution of this PhD thesis, and provides recommendations for future research.

Evaluating the performance of a random forest kernel for land cover classification

2

This chapter is based on the published papers:
A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Integrating support vector machines and random forests to classify crops in time series of worldview-2 images," in Image and Signal Processing for Remote Sensing XXIII, vol. 10427, p. 104270W, International Society for Optics and Photonics, 2017.
A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Evaluating the performance of a random forest kernel for land cover classification," Remote Sensing, vol. 11, no. 5, 2019.

Abstract

The production of land cover maps through satellite image classification is a frequent task in remote sensing. Random Forest (RF) and Support Vector Machine (SVM) are the two most well-known and recurrently used methods for this task. In this paper, we evaluate the pros and cons of using an RF-based kernel (RFK) in an SVM compared to using the conventional Radial Basis Function (RBF) kernel and standard RF classifier. A time series of seven multispectral WorldView-2 images acquired over Sukumba (Mali) and a single hyperspectral AVIRIS image acquired over Salinas Valley (CA, USA) are used to illustrate the analyses. For each study area, SVM-RFK, RF, and SVM-RBF were trained and tested under different conditions over ten subsets. The spectral features for Sukumba were extended by obtaining vegetation indices (VIs) and grey-level co-occurrence matrices (GLCMs), the Salinas dataset is used as benchmarking with its original number of features. In Sukumba, the overall accuracies (OAs) based on the spectral features only are of 81.34%, 81.08% and 82.08% for SVM-RFK, RF, and SVM-RBF. Adding VI and GLCM features results in OAs of $82.\%$, 80.82% and 77.96%. In Salinas, OAs are of 94.42%, 95.83% and 94.16%. These results show that SVM-RFK yields slightly higher OAs than RF in high dimensional and noisy experiments, and it provides competitive results in the rest of the experiments. They also show that SVM-RFK generates highly competitive results when compared to SVM-RBF while substantially reducing the time and computational cost associated with parametrizing the kernel. Moreover, SVM-RFK outperforms SVM-RBF in high dimensional and noisy problems. RF was also used to select the most important features for the extended dataset of Sukumba; the SVM-RFK derived from these features improved the OA of the previous SVM-RFK by 2%. Thus, the proposed SVM-RFK classifier is at least as good as RF and SVM-RBF and can achieve considerable improvements when applied to high dimensional data and when combined with RF-based feature selection methods.

Keywords: Image classification, Random forest, Support vector machine, Random forest kernel, Very high spatial resolution satellite images

2.1 Introduction

Remote sensing (RS) researchers have created land cover maps from a variety of data sources, including panchromatic [14], multispectral [15], hyperspectral [16], and synthetic aperture radar [17], as well as from the fusion of two or more of these data sources [18]. Using these different data sources, a variety of approaches have also been developed to produce land cover maps. According to the literature, approaches that rely on supervised classifiers often outperform approaches based on unsupervised classifiers [60]. This is because the classes of interest may not present the clear spectral separability required by unsupervised classifiers [60]. Maximum Likelihood (ML), Neural Networks (NN) and fuzzy classifiers are classical supervised classifiers. However, there are unsolved issues with these classifiers. ML assumes a Gaussian distribution, which may not always occur in complex remote sensed data [126, 127]. NN classifiers have a large number of parameters (weights) which require a high number of training samples to optimize particularly when the dimensionality of input increases [128]. Moreover, NN is a black-box approach that hides the underlying prediction process [128]. Fuzzy classifiers require dealing with the issue of how to best present the output to the end user [129]. Moreover, classical classifiers have difficulties with the complexity and size of the new datasets [130]. Several works have compared classification methods over satellite images, and report Random Forest (RF) and Support Vector Machine (SVM) as top classifiers, in particular, when dealing with high-dimensional data [131, 132]. Convolutional neural networks and other deep learning approaches require huge computational power and large amounts of ground truth data [133].

With recent developments in technology, high and very high spatial resolution data are becoming more and more available with enhanced spectral and temporal resolutions. Therefore, the abundance of information in such images brings new technological challenges to the domain of data analysis and pushes the scientific community to develop more efficient classifiers. The main challenges that an efficient supervised classifier should address are [95]: handling the Hughes phenomenon or curse of dimensionality that occurs when the number of features is much larger than the number of training samples [62], dealing with noise in labeled and unlabeled data, and reducing the computational load of the classification [98]. The Hughes phenomenon is a common problem for several remote sensing data such as hyperspectral images [134] and time series of multispectral satellite images where [60] spatial, spectral and temporal features are stacked on top of the original spectral channels for modeling additional information sources [63]. Over the last two decades, the Hughes phenomenon has been tackled in different ways by the remote sensing community [135, 136]. Among them, kernel-based

methods have drawn increasing attention because of their capability to handle nonlinear high-dimensional data in a simple way [73]. By using a nonlinear mapping function, kernel-based methods map the input data into a Reproducing Kernel Hilbert Space (RKHS) where the data is linearly separable. There is no need to work explicitly with the mapping function because one can compute the nonlinear relations between data via a kernel function. The function kernel reproduces the similarity of the data in pairs in RKHS. In other words, kernel-based methods require computing a pairwise matrix of similarities between the samples. Thus, a matrix is obtained using the kernel function in the classification procedure [137]. The kernel methods generally show good performance for high-dimensional problems.
SVM as a kernel-based non-parametric method [138] has been successfully applied for land cover classification of mono-temporal [114], multi-temporal [115], multi-sensor [116] and hyperspectral [117] datasets. However, the main challenge of the SVM classifier is the selection of the kernel parameters. This selection is usually implemented through computationally intensive cross-validation processes. The most commonly nonlinear kernel function used for SVM is Radial Basis Function (RBF), which represents a Gaussian function. In SVM-RBF classifier, selecting the best values for kernel parameters is a challenging task since classification results are strongly influenced by them. The selection of RBF kernel parameters typically requires to define appropriate ranges for each of them and to find the best combination through a cross-validation process. Moreover, the performance of SVM-RBF decreases significantly when the number of features is much higher than the number of training samples. To address this issue, here we introduce and evaluate the use of a Random Forest Kernel (RFK) in an SVM classifier. The RFK can easily be derived from the results of an RF classification [105]. RF is another well-known non-parametric classifier that can compete with the SVM in high-dimensional data classification. RF is an ensemble classifier that uses a set of weak learners (classification trees) to predict class labels [94]. A number of studies review the use of RF classifier for mono-temporal [139], multi-temporal [140], multi-sensor [141] and hyperspectral [142] data classification. Compared to other machine learning algorithms, RF is known for being fast and less sensitive to a high number of features, a few numbers of training samples, overfitting, noise in training samples, and choice of parameters. These characteristics make RF an appropriate method to classify high-dimensional data. Moreover, the tree-based structure of the RF can be used to create partitions in the data and to generate an RFK that encodes similarities between samples based on the partitions [104]. However, RF is difficult to visualize and interpret in detail, and it has been observed to overfit for some noisy datasets. Hence, the motivation of this work is to introduce the use of SVM-RFK as a way to combine the two most prominent classifiers used by the RS community and

evaluating whether this combination can overcome the limitations of each single classifier while maintaining their strong points. Finally, it is worth mentioning that our evaluation is illustrated with a time series of very high spatial resolution data and with a hyperspectral image. Both datasets were acquired over agricultural lands. Hence, our study cases aim at mapping crop types.

2.2 Methods

This section introduces the classifiers background. As SVM and RF are well-known classifiers, a summary of them is presented in this section. After that, we define the RFK and explain how it is generated from the RF classifier.

2.2.1 Random forest

The basics of RF have been comprehensively discussed in several sources during last decades [95], [94], and [143]. Briefly, RF classifiers are composed of a set of classification trees trained using bootstrapped samples from the training data [94]. In each bootstrapped sample, about two-thirds of the training data (in-bag samples) are used to grow an unpruned classification (or regression) tree, and the rest of the samples (the out-of-the-bag samples) are used to estimate the out of bag (OOB) error. Each tree is grown by recursive partitioning the data into nodes until each of them contains very similar samples, or until meeting one stopping condition [94]. Examples of the latter are reaching the maximum depth, or when the number of samples at the nodes is below a predefined threshold [94]. RF uses the Gini Index [87] to find the best feature and plot point to separate the training samples into homogeneous groups (classes). A key characteristic of RF is that only 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*. Hence, for large trees which is what RFs use, it is at least conceivable that all features might be used at some point when searching for split points whilst growing the tree. The final classification results are obtained by considering the majority votes calculated from all trees, and that is why RF is called a bagging approach [94]. A general design of RF is shown in Figure 2.1.
The operational use of RF classifiers requires setting two important parameters. First, the number of the decision trees to be generated N_t. Second, the number of the features to be randomly selected for defining the best split in each node *mtry*. Studies show the default value of 500 trees and the square root of the number of features in the most applications stabilize the error of the classification [95, 144]. Studies also show that classification results are most sensitive to the latter parameter. However, it is important to re-

mark that several studies consistently observe that the differences in Overall Accuracies (OAs) between the best configurations and other configurations for RF are small [130, 145, 146]. Moreover, RF is known for being fast, stable against overfitting and requiring small sample size with high dimensional input compared to many classifiers [95, 96]. Furthermore, RF is commonly used for feature selection by defining feature importance values based on total decrease in node impurity from splitting on the features, averaged over all trees (Mean decrease Gini index). These characteristics, besides the tree-based structure, make RF a good choice to be used as a partitioning algorithm that allows for the extraction of the similarity between samples. This similarity can then be used to create an RFK. In Section 2.2.3, we discuss how to obtain the similarity values between samples based on partitions created on data by trees in an RF.

2.2.2 Support Vector Machine

The base strategy of an SVM is to find a hyperplane in a high-dimensional space that separates the training data into classes so that the class members are maximally apart [135]. In other words, SVM finds the hyperplane that maximizes the margin, where the margin is the sum of the distances to the hyperplane from the closest point of each class [111]. The points on the margin are called support vectors. Figure 2.2a illustrates a two-class separable classification problem in a two-dimensional input space. Remote sensing data is often nonlinearly separable in the original high dimensional space [111]. In that case, the original data is mapped into a RKHS, where the data is linearly separable [112]. Figure 2.2b illustrates a two-class nonlinear separable classification problem in a two-dimensional input space.
Given training column vectors, $x_i \in \mathbb{R}^{N_f}$, where N_f is the number of dimensions. In addition, a binary class vector that denotes the labels, $y_i \in \{-1, 1\}$, where i represents the $i-th$ sample, the maximization of the margin can be formulated as a convex quadratic programming problem. One way to solve the optimization problem is using the Lagrange multipliers (dual problem) as follows:

$$\max_{\alpha} (\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j),$$
$$\text{subject to } 0 \leq \alpha \leq C \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0. \tag{2.1}$$

In Equation (2.1), α_i is a Lagrange multiplier, C is a penalty (regularization) parameter and $x_i x_j$ is the dot product between x_i and x_j. When the data is nonlinear separable in the original space (characteristic of remote sensing data), the data is mapped into RKHS through a mapping function $\Phi : x \rightarrow \varphi(x)$. The dot product in the RKHS space is

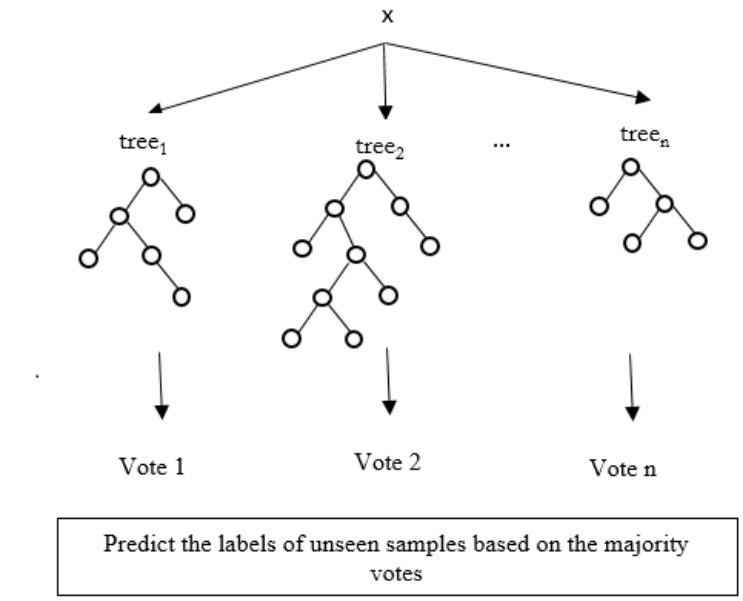


Figure 2.1 Example of general design of RF classifier with n number of trees.

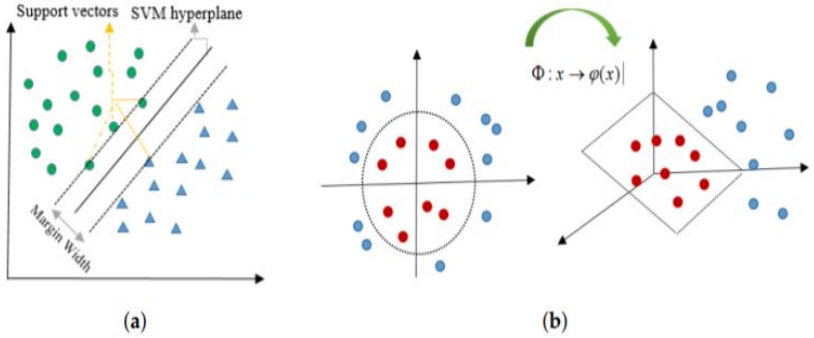


Figure 2.2 Example of a linear (**a**) and a nonlinear SVM (**b**) for a two-class classification problem. The nonlinear SVM maps the data into high dimensional space to separate linearly the classes of the data.

defined by a nonlinear kernel function $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. When the kernel function is calculated for all samples (N), the kernel function generates a square matrix ($\mathbf{K} \in \mathbb{R}^{N \times N}$) that containing pairwise similarities between the samples. Note that $\mathbf{K}$ is a positive definite and symmetric matrix.
Within all type of kernel functions, the most well-known is the Radial Basis Function (RBF) kernel ($k(x_i, x_j) = \exp(-(x_i - x_j)^2/2\sigma^2)$, where σ is the bandwidth). Thus, the SVM using the RBF kernel requires to fix two parameters, the σ and C. These parameters are tuned by cross-validation of a grid space of (C, σ). For a comprehensive review of kernel methods, see [147].

2.2.3 Random Forest Kernel

This section presents the RFK kernel. The main idea of the RFK is to calculate the similarities of pairwise data directly from the data by means of a discriminative model (i.e., learning the classification boundaries between classes) [148]. A discriminative approach divides the data into partitions through algorithms such as clustering or random forest [104]. In these cases, the fundamental idea is that the data that fall in the same partition are similar, and the data that fall in the different partitions are dissimilar (e.g., the Random Partition kernel [105]).
Let be ρ a random partition of the dataset, the Random Partition kernel is the average of occurrences that two samples (x_i and x_j) fall in the same partition, that is:

$$K(x_i, x_j) = \frac{1}{m} \sum_{g=1}^{m} I[\rho_g(x_i) = \rho_g(x_j)] \qquad i, j = 1, \ldots, N, \tag{2.2}$$

where I is the indicator function. I is equal to one when $\rho_g(x_i) = \rho_g(x_j)$, which means for this case that the samples x_i and x_j fall in the same partition; otherwise, it is zero [131]. In addition, g is the number of the partition in the data created by the eligible algorithms. Following the idea of the Random Partition kernel, the RFK is generated through creating random partitions by the RF classifier. As we have said before, RF is composed of trees. Each tree splits the data into homogeneous terminal nodes [149, 105]. Thus, the RFK uses the partitions obtained by the terminal nodes to calculate the similarity among data. In this instance, if two samples are landed 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 [105]:

$$K_{t_n}(x_i, x_j) = I\left[t(x_i) = t(x_j)\right], \tag{2.3}$$

where t is a terminal node and t_n is the $n - th$ tree of the RF. Then, the RFK matrix is calculated by the average of tree kernel matrices.

$$\mathbf{K}_{RFK} = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbf{K}_{t_n}, \tag{2.4}$$

N_t being the number of trees used in the RF.

Moreover, RF can also be used to identify the most important features (MIF) for high dimensional datasets, and an additional RFK can be derived from a subsequent RF model trained with those features only (RFK-MIF), which can be used in an SVM (SVM-RFK-MIF).

To assess the dependence of the applied kernels with an ideal kernel, we adopt the Hilbert–Schmidt Independence Criterion (HSIC) [150]. Given a kernel matrix for training dataset X (K_x) and the ideal kernel matrix for the class vector Y (K_y), the HSIC is obtained as follows [150]:

$$HSIC(K_X, K_Y) = \frac{1}{m^2} Tr(K_X H K_Y H), \tag{2.5}$$

where Tr is the trace operator, H is the centering matrix, and m is the number of samples. It has been proven that lower values of HSIC show the poorer alignment of the kernels with the target (ideal) kernel, and lower class separability subsequently.

2.3 Data and ground truth

Two high-dimensional data-sets including a time series of multispectral WorldView-2 (WV2) images and one hyperspectral AVIRIS image are used to evaluate the performance of the RFK. The first dataset was used to illustrate our work on a complex problem, namely that of classifying time series of VHR images to map crops. The second dataset was selected because it has been used as a benchmark dataset in several papers [151, 152].

2.3.1 WorldView-2

A time series of WV2 images acquired over Sukumba area in Mali, West Africa in 2014 is used to illustrate this study. The WV2 sensor provides data for eight spectral features at a spatial resolution of 2 m. This dataset includes seven multispectral images that span the cropping season [153]. The acquisition dates include May, June, July, October, and November. Ground truth labels for five common crops in the test area including cotton, maize, millet, peanut, and sorghum, were collected through fieldwork. These images and the corresponding ground data are part of the STARS project. This project, supported by the Bill and Melinda Gates foundation, aims to improve the livelihood of smallholder farmers. The Sukumba images are atmospherically corrected, co-registered and the trees and clouds are masked [153]. Figure 2.3a,b show the study area and the 45 fields contained within the database.

2.3.2 AVIRIS

A Hyperspectral image acquired by the AVIRIS sensor over Salinas Valley (CA, USA) on 9 October 1998 [132] is used to illustrate this study. The Salinas dataset is atmospherically corrected, and although the image contains 224 bands, they were reduced to 204 by removing water absorption bands (i.e., bands $[104-108]$, $[150-163]$, and 224). AVIRIS provides 3.7 meter spatial resolution. Ground truth labels are available for all fields and these labels contain 16 classes including vegetables, bare soils, and vineyard fields. Figure 2.3c,d show the area of interest and the RGB composite of the image.

2.4 Preprocessing and experimental set-Up

In this section, we describe the preprocessing and main steps of our work, which are also outlined in Figure 2.4.

2.4.1 Preprocessing

As shown in Figure 2.4, the accuracy of the classifiers was analyzed regarding the number of features. Table 2.1 shows the number of samples, features, and classes for each dataset. Additional features were generated (Table 2.2) for Sukumba dataset by obtaining Vegetation Indices (VIs) and Gray-Level Co-Occurrence Matrix (GLCM) features from spectral bands. These additional features were concatenated with the original spectral features to form an extended dataset for Sukumba.

The Sukumba dataset, which originally contains 56 bands, was extended by Normalized Difference Vegetation Index (NDVI), Difference Vegetation Index (DVI), Ratio Vegetation Index (RVI), Soil Adjusted Vegetation Index (SAVI), Modified Soil-Adjusted Vegetation Index (MSAVI), Water Band Index (WBI), Transformed Chlorophyll Absorption Reflectance Index (TCARI), and Enhanced vegetation index (EVI) increasing the number of the features until 105. Next, the number of features for Sukumba dataset was extended by adding the GLCM textures to the spectral features and VIs. Texture analysis using the Gray-Level Co-Occurrence Matrix is a statistical method of examining texture that considers the spatial relationship of pixels [160]. The GLCM textures derived for Sukumba dataset are presented and explained comprehensively in [124]. For each spectral feature, statistical textures including angular second moment, correlation, inverse difference moment, sum variance, entropy, difference entropy, information measures of correlation, dissimilarity, inertia, cluster shade, and cluster prominence are obtained [124]. Concatenating spectral, VI and GLCM features increase the number of features to

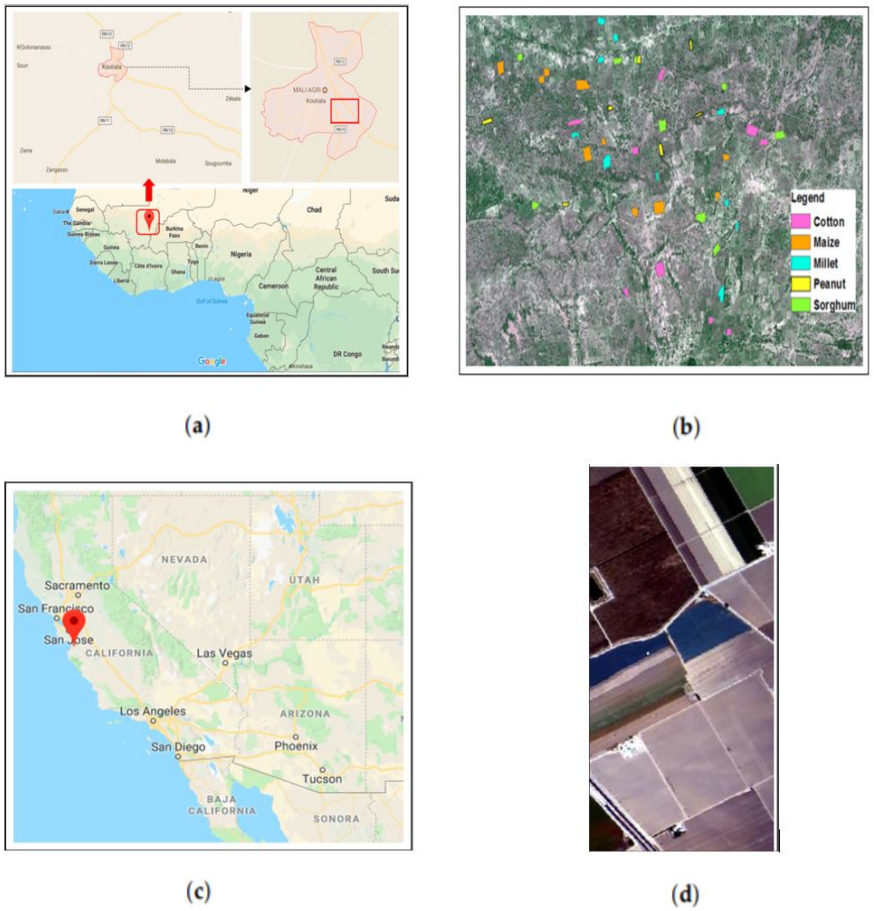


Figure 2.3 (**a**) study area of Sukumba site, southeast of Koutiala, Mali; (**b**) crop polygons for Mali and (**c**) study area of Salinas Valley, CA, USA and (**d**) RGB composite of Salinas.

Table 2.1 Dataset description (N_f: Number of features, N_{tr} total number training samples, N_{ts} total number test samples and N_{cl} number of classes).

Dataset	Features	N_f	N_{tr}	N_{ts}	N_{cl}
Sukumba	Spectral features	56	2043	1858	5
	Spectral &additional features	1057			
Salinas	Spectral features	204	24612	20782	16

Table 2.2 List of VIs used in this study together with a sort explanation of the them.

Formula	Description
$NDVI = \frac{NIR-Red}{NIR+Red}$	NDVI is a proxy for the amount of vegetation, and helps to distinguish the vegetation from the soil while minimizing the topographic effects, though does not eliminate the atmospheric effects [154].
$DVI = NIR - Red$	DVI also helps to distinguish between soil and vegetation, yet does not deal with the difference between the reflectance and radiance from atmosphere or shadows [155]
$RVI = \frac{NIR}{Red}$	RVI is the simplest ratio-based index showing high values for the vegetation and low values for soil, ice, water, etc. This index can reduce the atmospheric and topographic effects [155].
$SAVI = \frac{(NIR-Red)*(1+L)}{NIR+Red+L}$	SAVI is similar to the NDVI, yet it suppresses the soil effects by using an adjustment factor, L, which is a vegetation canopy background adjustment factor. L varies from 0 to 1 and often requires prior knowledge of vegetation densities to be set [156].
$MSAVI2 = \frac{2NIR+1-\sqrt{(2NIR+1)^2-8(NIR-RED)}}{2}$	MSAVI is a developed version of SAVI where the L-factor dynamically is adjusted using the image data and MSAVI2 is an iterated version of MSAVI [157].
$TCARI = 3[(R_{700} - R_{670}) - 0.2(R_{700} - R_{550})(\frac{R_{700}}{R_{670}})]$	TCARI indicates the relative abundance of chlorophyll using the reflectance at the wavelengths of 700 (i.e., R700), 670 and 550 and reduces the background (soil and non-photosynthetic components) effects compared to the initial versions of this index [158].
$EVI = \frac{2.5(NIR-Red)}{NIR+6Red-7.5Blue+1}$	EVI is developed to improve the NDVI by optimizing the vegetation signal with using blue reflectance to correct the soil background and atmospheric influences [159].

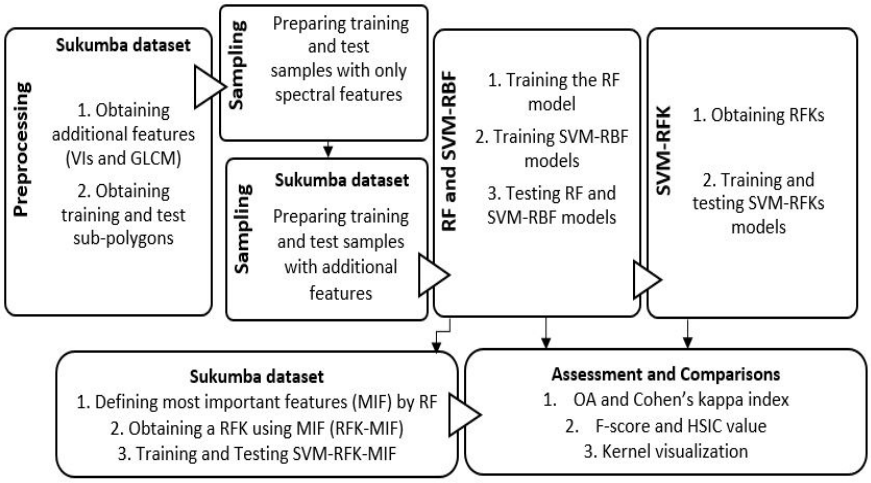


Figure 2.4 Overview of the steps followed to compare SVM-RFK with RF and SVM-RBF. Notation: The boxes with Sukumba dataset indicate steps that were only applied to this dataset, and the rest of the boxes indicate steps applied to both datasets

.

1057. Salinas dataset with 204 features used as a benchmarking dataset with its original number of features.

2.4.2 Experimental Set-Up

First, the polygons of the Sukumba dataset were split into four sub-polygons of approximately the same size to extract the training and test samples. Unlike a random selection of train and test samples, this step avoids selecting close samples in the training and test sets, which would inflate the performance of the classifiers. Two sub-polygons were selected to choose the training samples and the other two, the test samples. Both 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). A random sampling was used in the Salinas dataset (like in previous studies using this dataset). The samples were randomly split into train and test sets and 10 subsets are selected randomly from train and test sets separately, with the number of samples per class balanced (again, 130 and 100 samples per class for training and test).
In all the experiments, the optimization of the classifier parameters was required. The number of trees in RF was set to 500, according to the literature. The *mtry* parameter influence partially on the classification results of RF [130, 145]. Hence, we explored the influence of *mtry* on the SVM-RFK classification results. First, the RFK is obtained by training RF with the default value of this parameter. Next,

an RFK was obtained by optimizing *mtry* parameter for RF in a range of $[{N_f}^{(-1/2)} - 10, {N_f}^{(-1/2)} + 10]$ in steps of two. Then, the RFKs were obtained from the corresponding RF classifiers.
Taking the advantage of RF to select the most important features in high dimensional datasets, this method was used to select the top features in the extended dataset of Sukumba. The feature importance values provided by RF were used to select the 100 MIF, and an RFK was obtained using a subsequent RF model trained with the 100 features. Using RFKs in an SVM, a 5-fold cross-validation approach was used to find the optimal C value in the range $[5, 500]$. For the RBF kernel, we use the same range for the C parameter and 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. In all the cases, the one-versus-one multiclass strategy implemented in LibSVM [161] was used. An equal number of 11 candidates is considered when optimizing *mtry* for RF, as well as the bandwidth parameter of SVM-RBF. Classification results are compared in terms of their Overall Accuracy (OA), their Cohen's kappa index (κ), the F-scores of each class, and the timing of the methods. The computational times for each classifier were estimated using the ksvm function in the kernlab package of R [162]. The built-in and custom kernel of this package were respectively used to obtain RBF and RFKs classifications in an SVM. To obtain RF models and RFKs, randomForest package of R is used [108]. In addition, the generated RF-based and RBF kernels are compared through both visualization and HSIC measures. Finally, crop classifications maps are provided for the best classifiers.

2.5 Results and discussion

This section presents the classification results obtained with the proposed RF-based kernels and with the standard RF and SVM classifiers. All results were obtained by averaging the results of the 10 subsets used in each experiment. Results obtained with the default value of *mtry* are shown with RF_d and RFK_d, and those obtained with optimized *mtry* are shown by RF and RFK.

The OA and κ index averages of ten subsets are shown in Table 2.3 and Figure 2.5. In both cases, Sukumba and Salinas, results show high accuracies for all the classifiers for spectral features. The computational times for each classifier are depicted in Figure 2.6.

Table 2.3 and Figure 2.5 show that the three classifiers compete closely in the experiments using only spectral features. Comparing SVM-RFK and RF, SVM-RFK improves the results compared to RF in terms of OA and κ for all Sukumba and Salinas datasets. Focusing on only the spectral features, the RFK improvement is marginal. Op-

timizing the *mtry* parameter also helps the RF and SVM-RFK to outperform marginally compared to the models with the default values of the *mtry*. Although RF and RFK get better results by optimizing *mtry* parameter, the higher optimization cost required allows us to avoid it (Figure 2.6). This fact also make evident that optimizing the RF parameters is not crucial for obtaining an RFK.
Focusing on spectral features, the SVM-RBF yields slightly better results than SVM-RFK in terms OA and κ, reaching a difference of 1.41% and 0.74% in OA for Salinas dataset and Sukumba datasets, respectively. However, considering the Standard Deviation (SD) of these OAs, the performances of the classifiers are virtually identical (Table 2.3). Moreover, Figure 2.6 shows that the computational time for RFK is considerably lower than the RBF kernel for Salinas specifically without the *mtry* optimization. For spectral features of Sukumba, RFK and RBF computational times are at about the same level.
A notable fact is that SVM-RFK results improve considerably by extending the Sukumba dataset from 56 to 1057 dimensions, whereas RF and SVM-RBF classifiers get less accuracy with the extended dataset. For the extended Sukumba dataset, SVM-RFK outperforms SVM-RBF and RF with a difference of 4.34% and 1.48% in OA, respectively. Furthermore, RFK gets similar results for both *mtry* default and *mtry* optimized, whereas the computational time is three times higher using optimized parameter (Figure 2.6). Moreover, the time required to perform SVM-RFK$_d$ is also about seven times less than that of SVM-RBF (Figure 2.6). This fact could be seen as the first evidence of the potential of RFKs to deal with data coming from the latest generation of Earth observation sensors, which are able to acquire and deliver high dimensional data at global scales.
More evidence for the advantages of the RFKs is presented in Table 2.4 by exploiting the RF characteristics. This table shows that employing the RF to define the top 100 features (out of 1057 features) for Sukumba dataset, and obtaining the RFK based on a new RF model trained only with top 100 features improved the OA of the SVM-RFK by 2.66%.

Moreover, the HSIC measures presented in Table 2.5 reveal the alignment of the kernels with an ideal kernel for the training datasets. The lower separability of the classes results in poorer alignment between input and the ideal kernel matrices, and that leads in a lower value of HSIC [150]. Focusing on the spectral features, RFKs slightly outperform RBF for both Salinas and Sukumba datasets while both show almost equal alignment with an ideal kernel. The higher value of the HSIC measure for the RFKs compared to RBF is noticeable when the number of features is increased for the Sukumba dataset.
The analysis of the classifications results for each class is carried out by mean of the F-scores. Tables 2.6 and 2.7 show the results of $\overline{F}$ for each classifier, spectral case and dataset. In Sukumba (Table 2.6),

Table 2.3 Classification results of Sukumba with 56 features (Spectral features), and with 1057 features (Spectral features, VIs and GLCM textures), and Salinas with 204 features (Spectral features). Notation: $\overline{OA}$ (in %) is the overall accuracy averaged over 10 test samples, SD (in %) is the standard deviation for OA values, $\bar{\kappa}$ is the Cohen's kappa index averaged over 10 test samples, SD_κ is the standard deviation for κ values.

Tests	Methods	$\overline{OA}$	SD	$\bar{\kappa}$	SD_κ
		Sukumba			
Spectral features	RF	81.08	1.34	0.76	0.02
	RF_d	80.64	0.98	0.75	0.01
	SVM-RBF	82.08	2.21	0.77	0.03
	SVM-RFK	81.34	1.27	0.76	0.02
	SVM-RFK_d	80.68	1.12	0.75	0.01
Spectral features and additional features	RF	80.82	1.31	0.76	0.02
	RF_d	80.46	1.20	0.75	0.01
	SVM-RBF	77.96	1.26	0.72	0.02
	SVM-RFK	82.30	1.02	0.77	0.01
	SVM-RFK_d	82.14	0.84	0.77	0.01
		Salinas			
Spectral features	RF	94.16	0.5	0.93	0.004
	RF_d	94.10	0.48	0.93	0.005
	SVM-RBF	95.83	0.52	0.95	0.01
	SVM-RFK	94.42	0.56	0.94	0.005
	SVM-RFK_d	94.38	0.47	0.94	0.005

Table 2.4 Classification results for Sukumba with the top 100 features. Notation: $\overline{OA}$ (in %) is the overall accuracy averaged over 10 test samples, SD (in %) is the standard deviation for OA values, $\bar{\kappa}$ is the Cohen's kappa index averaged over 10 test samples, SD_κ is the standard deviation for κ values, and MIF is the most important features.

Methods	$\overline{OA}$	SD	$\bar{\kappa}$	SD_κ
RF-MIF	79.68	1.31	0.74	0.01
SVM-RFK-MIF	84.96	1.66	0.81	0.02

Table 2.5 HSIC measures for RF and RBF kernels. Notation: Sp is spectral features, Sp&Ad is spectral features and additional features.

Kernels	Sukumba: Sp	Sukumba: Sp&Ad	Salinas
RFK	0.016	0.021	0.041
RFK_d	0.018	0.021	0.042
RBF	0.010	0.004	0.029

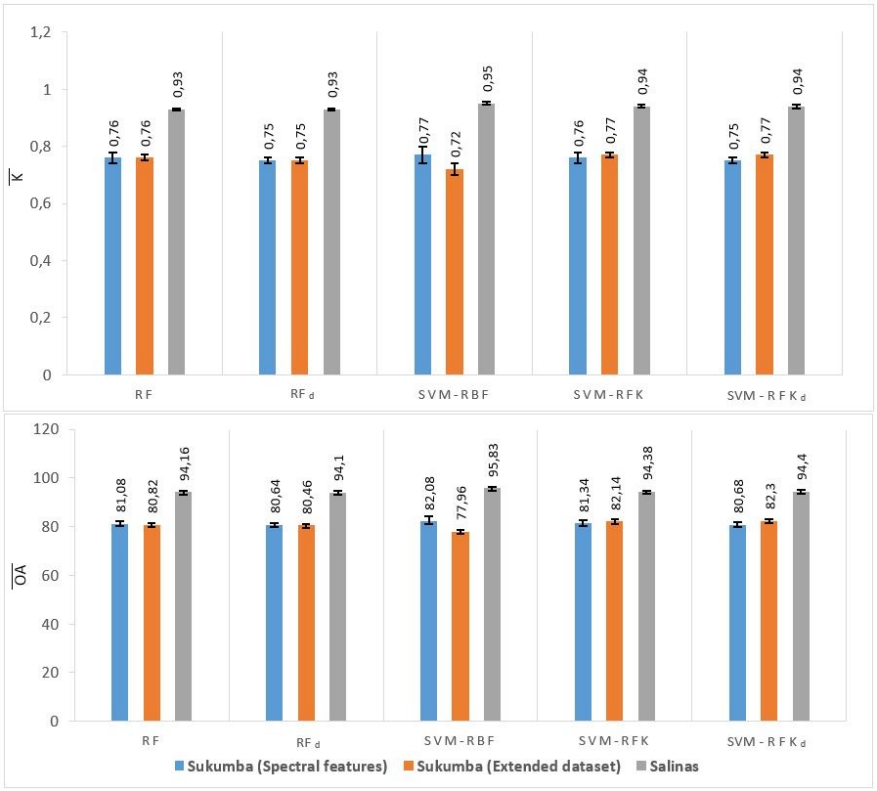


Figure 2.5 Comparison of $\overline{OA}$ and $\bar{\kappa}$ obtained for RF, SVM-RBF, and SVM-RFK classifiers. Notation: $\overline{OA}$ (in %) is the overall accuracy averaged over 10 test samples, $\bar{\kappa}$ is the Cohen's kappa index averaged over 10 test samples, and the standard deviations for OA and κ values are shown with error bars. RF and SVM-RFK denote classifiers created with an optimized *mtry* value, and RF$_d$ and SVM-RFK$_d$ denote classifiers created with the default *mtry* value.

the $\overline{F}$ has little variability, with standard deviations smaller or equal to 0.04. Furthermore, all classes have an $\overline{F}$ value larger than 0.75 (i.e., good balance between precision and recall). The classes Millet, Sorghum have the best $\overline{F}$ values, whereas the classes Maize and Peanut are harder to classify, irrespective of the chosen classifier. Focusing on the SVM-RBF and SVM-RFK classifiers, we see that the relative outperformance of SVM-RBF in terms of OA for spectral features (Table 2.3 and Figure 2.5) is mainly caused by the Maize and Millet classes, and this is while SVM-RFK and SVM-RBF show equal $\overline{F}$ values for classes Peanut and Sorghum, and SVM-RFK improves slightly the $\overline{F}$ value for the class Cotton compared to SVM-RBF. Moreover, SVM-RFK$_d$ competes closely with SVM-RFK and SVM-RBF while presenting slightly poorer $\overline{F}$ values.

Regarding Salinas, the $\overline{F}$ show results above 0.91 for all the classes

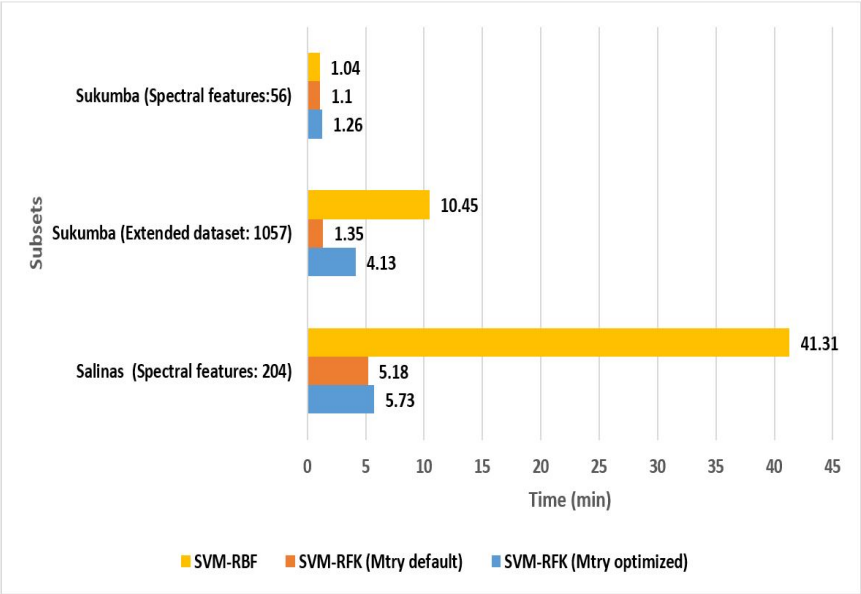


Figure 2.6 Classification time required by SVM classifiers.

except for Grapes untrained, and Vineyard untrained. For the latter two classes, the $\overline{F}$ are respectively around 0.69 and 0.71 for the RF-based classifiers. However, SVM-RFK improves the $\overline{F}$ values to 0.76 for both these classes. In this dataset, the SD values have also little variability (same as the ones found in Sukumba), with standard deviations smaller or equal to 0.05. For Salinas dataset, SVM-RFK$_d$ also competes closely with SVM-RFK and SVM-RBF while it presents slightly poorer $\overline{F}$ values.

Table 2.6 F-score average ($\overline{F}$) and standard deviation (SD) of the different classifiers using 56 features (Spectral features) and 1057 features (Spectral, VIs, and GLCM features) for the Sukumba dataset. Notation: RF and SVM-RFK denote classifiers created with an optimized *mtry* value, and RF$_d$ and SVM-RFK$_d$ denote classifiers created with the default *mtry* value.

Test	Classes	RF		RF$_d$		SVM-RBF		SVM-RFK		SVM-RFK$_d$	
		$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD
Spectral features	Maize	0.78	0.03	0.77	0.025	0.80	0.02	0.78	0.02	0.76	0.02
	Millet	0.86	0.02	0.85	0.02	0.87	0.03	0.85	0.02	0.84	0.02
	Peanut	0.78	0.02	0.78	0.02	0.79	0.04	0.79	0.02	0.77	0.01
	Sorghum	0.84	0.02	0.84	0.009	0.86	0.02	0.86	0.02	0.84	0.01
	Cotton	0.79	0.02	0.79	0.02	0.79	0.03	0.80	0.02	0.79	0.02
Spectral and additional features	Maize	0.77	0.04	0.76	0.03	0.75	0.03	0.77	0.03	0.76	0.02
	Millet	0.85	0.02	0.84	0.01	0.83	0.02	0.87	0.02	0.86	0.01
	Peanut	0.80	0.02	0.79	0.02	0.77	0.02	0.82	0.02	0.81	0.01
	Sorghum	0.82	0.02	0.82	0.02	0.81	0.03	0.84	0.02	0.84	0.02
	Cotton	0.80	0.02	0.80	0.02	0.73	0.02	0.82	0.02	0.83	0.01

A deeper analysis of the SVM-based classifiers can be achieved by visualizing their kernels. Figure 2.7 shows the pairwise similarity of training and test samples sorted by class. Here, we only visualize the RFK (with optimized *mtry*) because of the similarity of the results to RFK_d.
Focusing on the spectral features, this figure shows that the kernels obtained for Salinas are more "blocky" than those obtained for Sukumba. This makes it evident that a higher number of relevant features can improve the representation of the kernel. It also shows that the RFKs generated for Sukumba are less noisy than the RBF kernels. However, the similarity values of the RFKs are lower than those obtained for the RBF kernels. The visualization of the kernels confirms the higher $\overline{F}$ values found in the Salinas dataset. A detailed inspection of the RFKs obtained from this dataset shows low similarity values for classes 8 and 15, which correspond to Grapes untrained and Vineyard untrained. As stated before, these classes have the largest imbalance between precision and recall. Increasing the number of features to 1057 by extending the spectral features for Sukumba dataset represents a blockier kernel, by improving only the intraclass similarity values. However, the RBF kernel loses the class separability by increasing both intraclass and interclass similarity values by increasing the number of features for Sukumba dataset; this can be observed by RFK visualizations in Figure 2.7 and f-score values in Table 2.6. Focusing on the RFK, there are samples that their similarity values to other samples in their class are low for the RFK (Gaps inside the blocks), these samples could be outliers since RFK is based on the classes and the features while the RBF kernel is based on the Euclidean distances between the samples. Thus, removing outliers using RF can improve the representation of the RFK. Figure 2.8 shows the kernel visualization of RFK based on the 100 most important features selected by RF. As it can be observed in this figure, the similarity between the samples in the same classes is increased in particular for the classes one and five compared to the kernel using all 1057 features.
Finally, we present the classification maps obtained using the trained classifiers with spectral features. For Sukumba dataset, we also obtain the classification maps using SVM-RFK based on the top 100 features. For visibility reasons, we only present classified fields for Sukumba and classification maps for Salinas. In particular, Figure 2.10 shows two fields for each of the classes considered in Sukumba. These fields were classified using the best training subset of the ten subsets, and the percentage of pixels correctly classified are included on the top of each field. In general, the SVM classifiers perform better than the RF classifiers. Focusing on the various kernels, the RFKs outperform the results of RBF for the majority of the polygons.
Moreover, we observe a great improvement in the OA for all polygons by using the SVM-RFK-MIF. This means that RF can be used intuitively

to define an RFK based on only the top 100 features, and this kernel can improve the results significantly compared to RF, SVM-RBF, and SVM-RFK.
Classification maps for Salinas and their corresponding OAs are depicted in Figure 2.9. In this dataset, all classifiers have difficulties with fields where Brocoli_2 (class 2) and Soil_Vineyard (class 9) are grown. Moreover, it is worth mentioning that the performance of three classifiers is at about the same level. However, the SVM-RFK classifier has a marginally higher OA than the RF classifier, and SVM-RBF slightly outperforms SVM-RFK. This can be explained by the relatively high number of training samples used to train the classifiers compared with the dimensionality of the Salinas image. However, the computational time of classification for SVM-RBF is higher compared to RF and SVM-RFK (Figure 2.6).

Table 2.7 F-score average ($\overline{F}$) and standard deviation (SD) of the different classifiers using 204 features (Spectral features). Notation: RF and SVM-RFK are respectively RF and SVM-RFK with optimized *mtry*, and RF_d and SVM-RFK_d are respectively RF and SVM-RFK with default *mtry*.

Test	Classes	RF		RF_d		SVM-RBF		SVM-RFK		SVM-RFK_d	
		$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD
Spectral features	1:Brocoli_1	1.00	0.008	1.00	0.007	1.00	0.005	1.00	0.005	1.00	0.007
	2:Brocoli_2	0.99	0.009	0.99	0.009	1.00	0.005	1.00	0.006	0.99	0.007
	3:Fallow	0.97	0.018	0.97	0.017	0.98	0.012	0.97	0.014	0.97	0.015
	4:Fallow_rough	0.99	0.008	0.99	0.008	0.99	0.007	0.99	0.007	0.99	0.007
	5:Fallow_smooth	0.98	0.010	0.98	0.009	0.99	0.012	0.98	0.010	0.98	0.011
	6:Stubble	1.00	0.003	1.00	0.003	1.00	0.002	1.00	0.004	1.00	0.005
	7:Celery	0.99	0.006	0.99	0.005	1.00	0.004	0.99	0.007	0.99	0.007
	8:Grapes_untr.	0.69	0.032	0.69	0.039	0.76	0.026	0.70	0.042	0.69	0.041
	9:Soil_Vineyard	0.99	0.009	0.98	0.009	0.99	0.006	0.99	0.007	0.99	0.007
	10:Corn	0.91	0.011	0.91	0.014	0.94	0.019	0.91	0.009	0.91	0.009
	11:Lettuce_4wk	0.96	0.011	0.96	0.008	0.98	0.010	0.97	0.011	0.97	0.011
	12:Lettuce_5wk	0.98	0.010	0.98	0.011	0.98	0.008	0.98	0.011	0.98	0.010
	13:Lettuce_6wk	0.97	0.012	0.97	0.011	0.99	0.010	0.98	0.012	0.98	0.012
	14:Lettuce_7wk	0.95	0.018	0.95	0.018	0.98	0.014	0.96	0.016	0.96	0.017
	15:Vineyard_untr.	0.71	0.036	0.72	0.045	0.76	0.033	0.71	0.051	0.71	0.044
	16:Vineyard_vertical	0.98	0.013	0.98	0.014	0.99	0.006	0.98	0.013	0.98	0.012

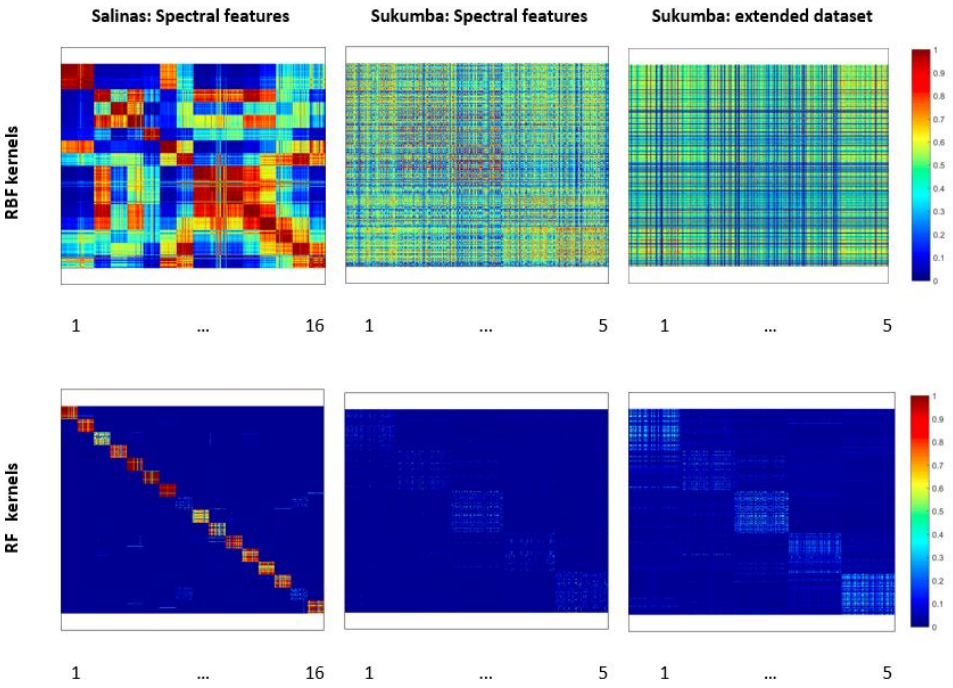


Figure 2.7 RBF Kernels (**top**) and RFKs (**bottom**) for the datasets from left to right: Salinas (Spectral features), Sukumba (Spectral features), and Sukumba (Spectral features and additional features). Class labels are shown on the bottom of the kernels. The class labels go from 1 to 5 for Sukumba, and from 1 to 16 for Salinas.

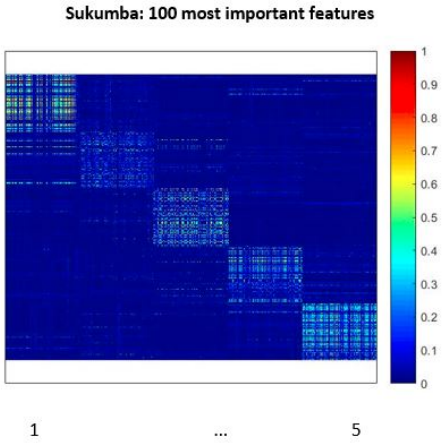


Figure 2.8 RF Kernel for top 100 features selected by RF (out of 1057). Class labels are shown on the bottom of the kernel. The clafss labels go from 1 to 5 for Sukumba.

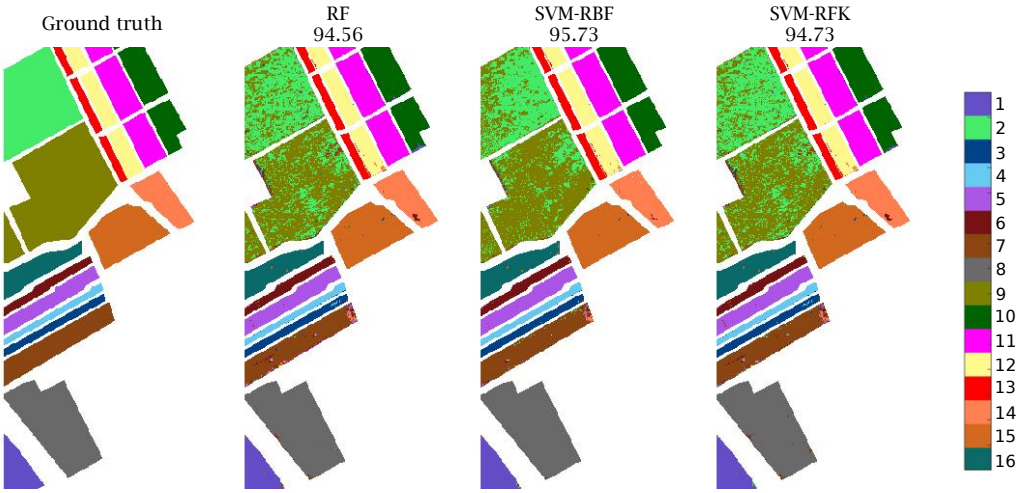


Figure 2.9 Ground truth and three classification maps (and the OA (%) calculated using all the pixels in the dataset on the top) for the RF, SVM-RBF, and SVM-RFK classifiers using the AVIRIS spectral features.

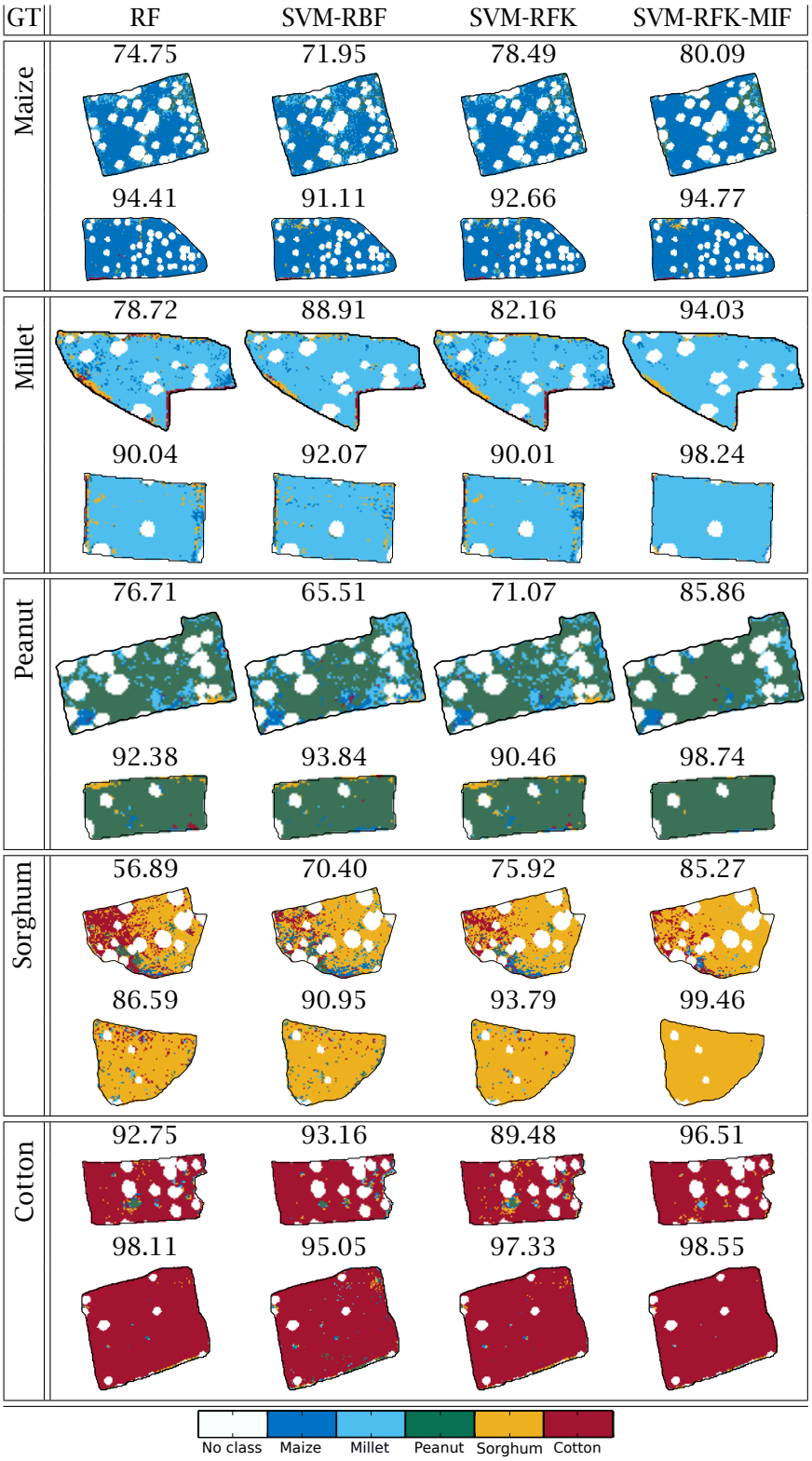


Figure 2.10 Two crop classified fields per ground truth class along with the overall accuracy for the different classifiers using spectral features, and the top 100 features for SVM-RFK-MIF. The trees within the crops were excluded from the classification (masked, unclassified).

2.6 Conclusions

In this chapter, we evaluate the added value of using an RF-based kernel (i.e., RFK) in an SVM classifier (i.e., RFK) by comparing its performance against that of standard RF and SVM-RBF classifiers. This comparison is done using two datasets: a time series of WV2 images acquired over Sukumba (Mali), and a hyperspectral AVIRIS image over Salinas (CA, USA). The obtained OAs and their SD values indicate that three classifiers perform at about the same level in most of the experiments. Our findings show that there are alternatives to the expensive tuning process of SVM-RBF classifiers. The proposed RFK led to competitive results for the datasets with a lower number of features while reducing the cost of the classification. Our findings prove that optimizing the *mtry* for RF leads to minor changes in the SVM-RFK. Thus, with a small trade-off in OA for the datasets with a low number of features, the cost of the classification can be reduced through skipping the *mtry* optimization. More importantly, our results show that RFKs created using high dimensional and noisy features considerably improve the classification accuracies obtained by the standard SVM-RBF while reducing the cost of classification. For the higher number of features, SVM-RFK results are also slightly better than the ones obtained by the standard RF classifier. Moreover, by exploiting the RF characteristics through defining the most important features, the results of the classification for SVM-RFK considerably improve, with OA around 7% better than those obtained with an SVM-RBF classifier. In short, our results indicate that RFK can outperform standard RF and SVM-RBF classifiers in problems with high data dimensionality. Further work is required to evaluate this kernel in additional classification problems and against other land cover classification approaches (e.g., based on deep learning). Other characteristics of RF (outlier detection) can be exploited to estimate the RFK more accurately. Furthermore, the proposed RFK is based on a rough estimation of the similarity between samples according to their terminal node. Future work is required to design and test more advanced and alternative estimations of similarity using RF classification results.

A multi-scale random forest kernel for land cover classification

3

This chapter is based on:
A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier. "A Multi-Scale Random Forest Kernel for Land Cover Classification. Remote Sensing." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Accepted for publication.

Abstract

Random forest (RF) is a popular ensemble learning method that is widely used for the analysis of remote sensing images. RF also has connections with kernel-based method. Its tree-based structure can generate a Random Forest Kernel (RFK) that provides an alternative to common kernels such as Radial Basis Function (RBF) in kernel-based methods such as Support Vector Machine (SVM). Using RFK in an SVM has been shown to outperform both RF and SVM-RBF (i.e., using an RBF kernel in an SVM) in classification tasks with a high number of features. Here, we explore new designs of RFKs for remote sensing image classification. Different RF structural parameters and characteristics are used to generate various RFKs. In particular, we explore the use of RF's depth, the number of branches between terminal nodes of trees, and the predicted class probabilities for designing and evaluating new RFK. Two depth-based kernel are proposed: an RFK at the optimal depth, and a multi-scale one created by combining RFKs at multiple depths. We evaluate the proposed kernels within an SVM by classifying a time series of Worldview-2 images, and by designing experiments having a various number of features. Benchmarking the new RFKs against RBF shows that the newly proposed kernels outperform RBF kernel for the experiments with a higher number of features. For the experiments with a lower number of features, RFKs and RBF kernel perform at about the same level. Benchmarking against standard RF also shows the general outperformance of the proposed RFKs in SVM. In all experiments, the best results are obtained with a depth-optimized RFK.
Keywords: Image classification, random forest kernel designs, support vector machine

3.1 Introduction

Remotely sensed images are one of the most important sources of data for land cover mapping. However, producing high-quality land cover maps using remotely sensed data is still challenging because the necessary use of time series of images leads to high-dimensional problems and because land cover classes typically are non-linearly separable [132]. The curse of dimensionality or Hughes phenomenon occurs when the number of features is much larger than the number of training samples [163]. The Hughes phenomenon is a common problem for several remote sensing data such as hyperspectral images and time series of multispectral satellite images [63]. Moreover, Hughes phenomenon occurs when spatial, textural or other types of extracted features are stacked on top of the original spectral features for modeling additional information sources [63]. Several works have reviewed land cover classification methods, and findings show that kernel-based methods outperform traditional classification meth-

ods particularly in dealing with Hughes phenomenon [163, 73, 164]. Kernel-based methods map the non-linear data into a Reproducing Kernel Hilbert Space (RKHS) where the data is linearly separable. Instead of explicitly using a mapping function, a kernel function is used to reproduce the pairwise similarities matrix by computing the inner products among the samples in RKHS [147]. The most well-known kernel-based classifier and kernel function are Support Vector Machine (SVM) and Radial Basis Function (RBF), respectively. Using SVM-RBF (i.e., using an RBF kernel in an SVM), one needs to optimize two parameters (i.e., RBF bandwidth and SVM regularization parameters) through a high computational cross-validation process [163], this is a known limitation of using the RBF kernel in an SVM.
Another well-known classifier able to handle high-dimensional and non-linear problems is Random Forest (RF) [140, 141, 142]. RF is fast and comparatively robust to a high number of features, a few numbers of training samples, overfitting, noise in training samples, and the choice of parameters [94, 132]. RF can be used for feature selection and outlier detection [107, 165]. The operational use of RF classifiers requires setting two parameters: the number of the decision trees to be generated (N_t) and the number of the features to be randomly selected for defining the best split in each node ($mtry$). Using the default value of 500 trees and the square root of the number of features stabilize the error of the classification in most applications [95]. However, RF is difficult to visualize and it can get overfitted [95]. Despite this, integrated approaches of RF and SVM-RBF can be used to exploit strong points of both classifiers and avoid their limitations. For instance, using RF to find the most important features and importing these features into an SVM-RBF classifier is shown to improve the Overall Accuracy (OA) of the classification compared to single use of RF or SVM-RBF [165].
In addition, the tree-based structure of RF allows the extraction of kernels that can be integrated with kernel-based methods such as SVM [94, 104]. The tree-based structure of RF draws partitions in the data that can be used to generate a Random Forest Kernel (RFK), which encodes similarities between data samples based on the partitions [104]. The classic RFK uses the average of a Kronecker delta function of pairwise leaf node samples as similarity values [94, 104]. Using an SVM classifier to compare kernels, we found that the classic RFK performs competitively in terms of Overall Accuracy (OA) while reducing the computational time with respect to RBF, as shown in classification tasks involving time series of multispectral satellite images and an airborne hyperspectral image [166]. Moreover, integration of RFK and SVM is shown to yield slightly higher OAs than traditional RF in high dimensional and noisy experiments, and it provides competitive results in low dimensional experiments [166]. However, the problem associated with classic RFK is that similarity values are binary metrics, and this rough binary estimation may not be always compatible with real-world data classes with similar spec-

tral signatures (i.e, crops) [167]. In other words, it is more realistic to think that pairs of samples can be similar to a certain degree instead of assuming that they are either similar or dissimilar [167]. A large number of trees is necessary to get accurate estimations for the classic RFK. When a relatively low number of trees are used, a more elaborate estimate of similarity values is required [167]. To overcome this problem, the similarity values can be obtained based on the number of tree branches between the terminal nodes containing the samples. This design was evaluated over two-class classification problems, and it was found to improve the data similarity estimation, especially when RF is made of a small number of trees [167]. Accordingly, the first goal of this paper is to compare the performance of branch-based RFK for a multi-classification problem against that of classic RFK and RBF kernel when used within an SVM. We also compare their performance against that provided by a standard RF classifier. Another problem with classic RFK is that the samples from the same class can land in different nodes in fully grown trees, and their similarity value gets assigned zero consequently. The second goal of this paper is to address this issue by exploring the influence of RF depth on the RFK performance. Following on this, the third goal is to design a multi-scale RFK based on multiple depths, inspired by [168]. The idea in [168] is to encode similarity values among samples using the probability that they are grouped together at different scales through Gaussian mixture models clustering. Different scales are defined by varying the number of clusters and initialization. Here, a multi-scale RFK is designed based on multiple depths of RF.
In a nutshell, the contribution of this paper lies on the investigation of alternative designs of RF-based kernels for remote sensing image classification. Four alternative designs of RFK are evaluated: a branch-based RFK based on the distance among terminal nodes, a depth-optimized RFK, a multi-scale RFK based on obtaining classic RFK at multiple depths, and a probabilistic multi-scale RFK based on obtaining RF-based class probabilities at multiple depths. We benchmarked the performances of these kernels against those provided by the RBF kernel in an SVM and the standard RF model. Our work is illustrated with a time series of very high spatial resolution data acquired over agricultural lands.

3.2 Background

The main idea of the so-called kernel trick is to allow mapping nonlinear separable data in the original space into RKHS without the explicit knowing of the mapping function $\Phi : x \rightarrow \varphi(x)$ [147]. The dot product of two training samples vectors (x_i and x_j) in the RKHS space is defined by a kernel function $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. When the kernel function is calculated for all samples (N), the kernel function generates a square matrix ($\mathbf{K} \in \mathbb{R}^{N \times N}$) of pairwise similarities

between the samples.
Kernel-based methods belong to the generative or discriminative categories [169]. Generative models aim at learning probability density functions and discriminative learning methods are based on learning class boundaries [170]. Generative approaches assume a data model which is often improper for the remote sensing data [169]. Discriminative learning methods obtain the class boundaries directly from the data [169]. Discriminative approaches can partition the data through several algorithms such as clustering and RF [168]. The key idea of discriminative kernels is that samples located in the same partition are similar and those ending up in different partitions are dissimilar [104]. In the present article, RF is used to create random partitions. The reason is that RF is known for being fast, stable against overfitting and requiring a small sample size with high dimensional input compared to several classifiers [95, 94]. Moreover, RF is robust to the choice of parameters [95, 94]. The strong points of RF along with its tree-based scheme can be used to partition data into homogeneous groups, and these partitions can be used to create an RF-based kernel that can subsequently be used in a kernel-based method such as SVM [166]. In the following, we present the background on different possible RFKs designs selected for the experimental tests of this study.

3.3 Methods

3.3.1 Classic Node Based RFK

The classic RFK uses the terminal nodes as partitions created on data by trees to calculate the pairwise similarity values among the samples. If two samples fall into the same terminal node of a tree, the similarity is equal to one; otherwise, it is zero. The classic RFK suggested by [94] is extensively described in our previous work [166]. For each tree, one pairwise similarity matrix is generated and RFK is the average of the matrices obtained for all trees. Here, we indicate this node-based RFK by RFK_{Nd}.

3.3.2 Branch Based RFK

To get accurate similarity values in RFK_{Nd}, a large number of trees is required [167]. In the experiments of using few trees, more elaborate estimate of data proximity is needed. A novel approach to estimate data proximity in RF is proposed in [167]. This approach is based on measuring distance between two terminal nodes containing samples i (s_i) and j (s_j). In this study, we indicate this kernel by RFK_{Br}, and it can be assessed with following equation:

$$\mathbf{RFK_{Br}}(s_i, s_j) = \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{1}{e^{w.g_{ijt_n}}} \tag{3.1}$$

Where N_t is number of trees in RF and n runs over the number of trees, the parameter w controls the influence of the distance between two terminal nodes occupied by the samples s_i and s_j, and g_{ijt_n} is the number of branches between two terminal nodes containing s_i and s_j in the $n-th$ tree of the RF (i.e.,t_n). For example, $g_{131} = 3$ between the terminal nodes 1 and 3 of $tree_1$ in Figure 3.1. If s_i and s_j land on the same terminal node, then $g_{ijt_n} = 0$ and RFK_{Br} will be increased by one as in the original way (i.e., RFK_{Nd}) to assess the similarity of two samples.

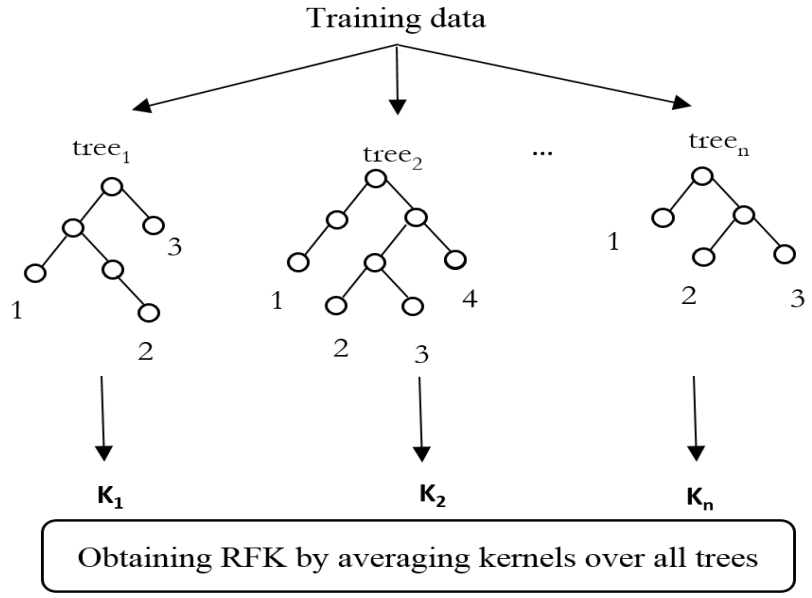


Figure 3.1 The general design of RFK for a RF classifier with n number of trees

3.3.3 Multi-Scale Probabilistic RFK

The pairwise similarity matrix is computed based on a kernel function ($k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$) which can be also designed through probabilistic approaches [168]. A probabilistic kernel function can be designed by considering a probability density function as the mapping function [170]. If we show the probabilistic mapping function as $\phi(s_i) = \pi_i$, the probabilistic kernel can be defined as:

$$K(s_i, s_j) =< \pi_i(s_i), \pi_j(s_j) >_H \tag{3.2}$$

Recently, [168] introduced a probabilistic cluster kernel by computing the composition of dot products between the posterior probabilities obtained via Gaussian mixture models. In this approach, the posterior probability of two samples belonging to the same cluster is considered as the similarity between such samples. Thus, the probabilistic cluster kernel is obtained through a generative and unsupervised approach. Here, we introduce a Probabilistic RFK that is obtained through a discriminative and supervised approach. In the present work, we indicate this kernel by RFK_{Prob}.
RF assigns a probability of membership to each one of the class labels of interest for all samples, namely the predicted probability vector. This vector contains the proportion of votes of the trees for each class. For each sample, the class label with the highest probability is selected as the class label for that sample by RF. Accordingly, the predicted probability vectors for samples i and j can be defined respectively as $\pi_i(s_i) = (p_{i1}, p_{i2}, p_{i3}, ..., p_{iC})$ and $\pi_j(s_j) = (p_{j1}, p_{j2}, p_{j3}, ..., p_{jC})$ where p_{iC} is the probability that sample i belongs to class C. The similarity value for samples i and j using the RFK_{Prob} can be defined as the inner product of the vectors $\pi_i(s_i)$ and $\pi_j(s_j)$ [170] and [105]:

$$\begin{aligned} \mathbf{RFK_{Prob}}(s_i, s_j) &= < \pi_i(s_i), \pi_j(s_j) >_H \\ &= p_{i1}.p_{j1} + p_{i2}.p_{j2} + ... + p_{iC}p_{jC} \end{aligned} \tag{3.3}$$

As a matrix notation, all the predicted probability vectors are placed in matrix $\boldsymbol{P}$ as following:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1C} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2C} \\ & & \cdots\cdots & & \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nC} \end{bmatrix} \tag{3.4}$$

$\boldsymbol{P}$ is the matrix of predicted probabilities by RF, one column per class and one row per observation. Once $\boldsymbol{P}$ is calculated, the RFK_{Prob} is defined as $RFK_{Prob} = \boldsymbol{P}.\boldsymbol{P}^T$. The kernel also can be obtained at different depths. Therefore, the average of probabilistic RF kernels obtained at different depths is defined as below:

$$\mathbf{RFK}_{\overline{\mathbf{Prob}}} = \frac{1}{N_d} \sum_{n=1}^{N_d} (\mathbf{P}.\mathbf{P^T})_{d_n} \tag{3.5}$$

Where N_d is the number of the depths considered to obtain RFK_{Prob}, n runs over the number of depths, and d_n is the $n-th$ depth in RF. Moreover, the RFK_{Prob} obtained at each depth is separately imported in an SVM and the RFK_{Prob} that generates the best OA is considered as a depth-optimized probabilistic kernel. Here, we indicate this depth-optimized probabilistic kernel with RFK_{Prob^*}.

3.3.4 Multi-Scale Node Based RFK

Similar to RFK_{Prob}, RFK_{Nd} can be obtained at different depths as well and average of these kernels can be obtained as a multi-scale RFK based on the terminal nodes. The average of classic RF kernels obtained at different depths is defined as below:

$$\mathbf{RFK}_{\overline{\mathbf{Nd}}} = \frac{1}{N_d} \sum_{n=1}^{N_d} (\mathbf{RFK}_{\mathbf{Nd}})_{d_n} \tag{3.6}$$

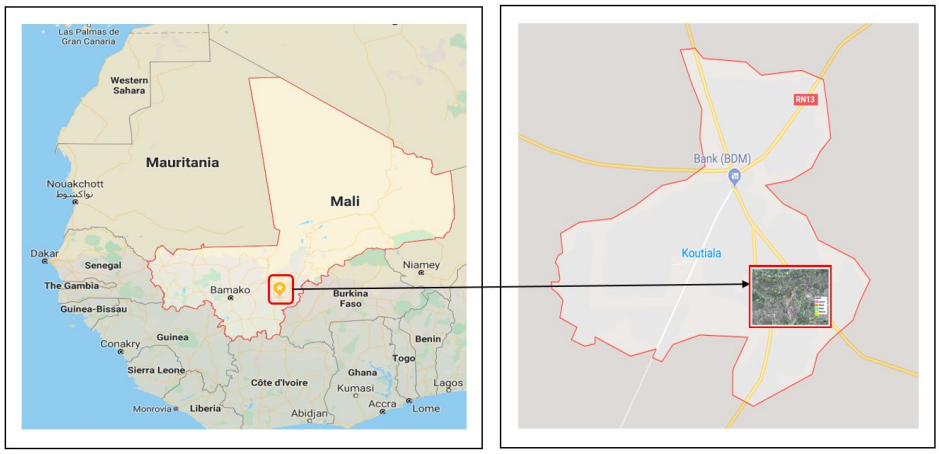
Where N_d is the number of the depths considered to obtain RFK_{Nd}, n runs over the number of depths, and d_n is the $n-th$ depth in RF. Here, we indicate this kernel by $RFK_{\overline{Nd}}$ and the depth-optimized classic RFK with RFK_{Nd^*}.

3.4 Experimental set-up

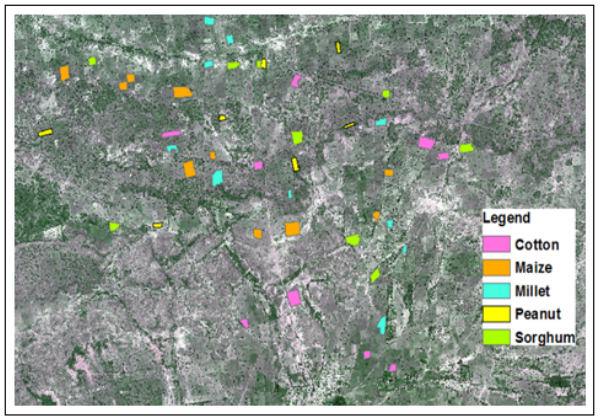
In this section, we first describe the data and study area used to illustrate this study. Next, we explain the experimental set up followed to evaluate the newly proposed RFKs (Figures 3.3 and 3.4).

3.4.1 Data and study area

A time series of WorldView-2 (WV2) images acquired over Sukumba area in Mali, West Africa in 2014 is used to illustrate this study. WV2 sensor provides data for eight spectral features at a spatial resolution of 2 m. There are seven multispectral images in this dataset that defines the cropping season in 2014 [153]. Ground truth labels for 5 common crops including cotton, maize, millet, peanut, and sorghum, were collected through fieldwork. The images were atmospherically corrected using the 6S radiative transfer model [171], and co-registered using the centroid of the trees located in the study area [153]. Tree masks were automatically created by applying a series of Gaussian filters [153], and cloudy pixels were removed by eliminating the pixels with the highest reflectance value in the blue band which fall within the percentage of cloud coverage reported in the metadata of the image delivery [153].
This dataset and the ground data are part of the STARS project which aims to improve the livelihood of smallholder farmers. The Sukumba dataset contains a total of 45 labeled polygons. Each polygon corresponds to a single farm management unit, and the average size of these units is about 1.35 ha. This means that the average farm contains about 3500 pixels. Figure 3.2a,b show the study area and the 45 fields contained within the database.
This dataset originally contains 7 multi-temporal images with 8 bands each image (56 bands). The acquisition dates include May,



(a)



(b)

Figure 3.2 (**a**) study area of Sukumba site, southeast of Koutiala, Mali; (**b**) crop polygons for Mali.

June, July, October, and November [153]. The Vegetation Indices (VIs) including Normalized Difference Vegetation Index (NDVI), Difference Vegetation Index (DVI), Ratio Vegetation Index (RVI), Soil Adjusted Vegetation Index (SAVI), Modified Soil-Adjusted Vegetation Index (MSAVI), Transformed Chlorophyll Absorption Reflectance Index (TCARI), and Enhanced vegetation index (EVI) were obtained and added to spectral features to increase class separability of the crops. The definitions of these VIs are given in [166]. Besides the above-mentioned classic VIs, the combinations of bands 2 to 8 through difference, ratio, and normalization of these bands were obtained and added to spectral and classic VIs, and this increases the number of feature to 525. Next, the number of features was extended by obtaining the Gray-Level Co-Occurrence Matrix (GLCM) textures to the

spectral features and VIs. Texture analysis based on the GLCM is a statistical method to define the spatial relationship of pixels [160]. The GLCM textures derived for Sukumba dataset are comprehensively described in [166] and [124]. For each spectral and VI feature, 17 GLCM textures were computed using a window of 3 by 3 pixels and by averaging the values obtained along four directions (0, 45, 90 and 135). For all spectral and VI features separately, statistical textures including angular second moment, correlation, inverse difference moment, sum variance, entropy, difference entropy, information measures of correlation, dissimilarity, inertia, cluster shade, and cluster prominence were obtained [124].
Stacking all the spectral, VIs and GLCM features, the total number of features reached 8498. This number further was increased by including extra features, namely Green Leaf Index (GLI) [172] and Local Binary Pattern (LBP) [173]. With this addition, the final number of features available for the experiment ALL equals 10584. Table 3.1 shows the sequence of adding the features which are used in four tests to examine the proposed methods in this study.

Table 3.1 Experiments description (N_f: Number of features used in each case.)

Case	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

3.4.2 Comparing RFK_{Br} and RFK_{Nd}

In the pre-processing and sampling step of Figure 3.3, we divided the polygons representing the labeled farms in the study area into four almost equal sized sub-polygons. Two of these sub-polygons were used to select the training samples and the other two to select test samples, so that samples from the same neighborhood do not end up in both training and test sets, and this prevents inflating the performance of the classifiers. Next, each of the train and test sets were randomly sampled to get ten random subsets, with a balanced number of samples per class (130 and 100 samples per class for training and test, respectively). We used these ten randomly selected subsets with a different number of features (for each of four experiments shown in Table 3.1) to obtain representative results (i.e. all presented results correspond to the average of the results obtained for the 10 subsets). In the high-dimensional 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.,

they might be considered noise).

After obtaining training and test samples for four experiments, as it is shown in Figure 3.3, RF models with 500 fully grown trees are obtained in the first step. RF models are trained using 5, 10, 20, 50, 100 and 500 trees, and for each model the *mtry* parameter is optimized in a range of $[{N_f}^{(-1/2)} - 10, {N_f}^{(-1/2)} + 10]$ in steps of two, where N_f is the number of features.
Based on these RF models, we obtained RFK_{Nd} and RFK_{Br}. First, we investigated the performance of RFK_{Nd} and RFK_{Br} in an SVM against the different number of trees. Next, we benchmarked their performance against RBF kernel in an SVM and a standard RF. The SVM using the RBF kernel requires to fix two parameters, the σ (i.e., bandwidth of the) and C (i.e., a penalty or regularization parameter) [135]. A 5-fold cross-validation approach is used to find the optimum bandwidth in 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 in the range of $[5, 500]$. Using the RFKs in an SVM requires to fix the C parameter as well. A 5-fold cross-validation approach is also used to optimize the C parameters for all the models. Besides the C parameter, RFK_{Br} requires optimizing w parameter which is optimized in a range of $w = [0.1, 2]$ in steps of 0.1 [167].
Obtaining RFK_{Br} requires to compute the pairwise distances among the end-nodes in terms of number of the branches between the end-nodes. To find number of number of branches between the end-nodes containing samples s_i and s_j in the $k-th$ tree (g_{ijk}), we obtained the paths of the end-nodes containing these samples to the root node, and we found the first ancestor of these end-nodes by comparing their corresponding paths to the root node. Then, the number of branches is counted between each end-node and the first ancestor. Finally, two number of branches are summed up to obtain the number of branches between the end-nodes containing s_i and s_j. For each tree, a g-matrix$_{TRTR}$ and a g-matrix$_{TSTR}$ is obtained. The g-matrix$_{TRTR}$ contains the pairwise distance among training samples and the g-matrix$_{TSTR}$ contains the pairwise distance among test and training samples. Next, the g-matrices$_{TRTR}$ together with the ranges of w and C parameters are used in an SVM to optimize w and C parameters in the cross-validation process. Next, optimal values of w and C parameters are used to train the SVM-RFK_{Br} model. After training the model, the g-matrices$_{TSTR}$ are used to evaluate the model.

3.4.3 Multi-scale RFKs

In Figure 3.4, we start by following the same pre-processing and sampling used in the previous workflow. Here, however, we use 10 different depths to grow 500 trees in the RF ensemble. Then, we obtain a RFK_{Nd} and RFK_{Prob} at each depth. The depth of RF is

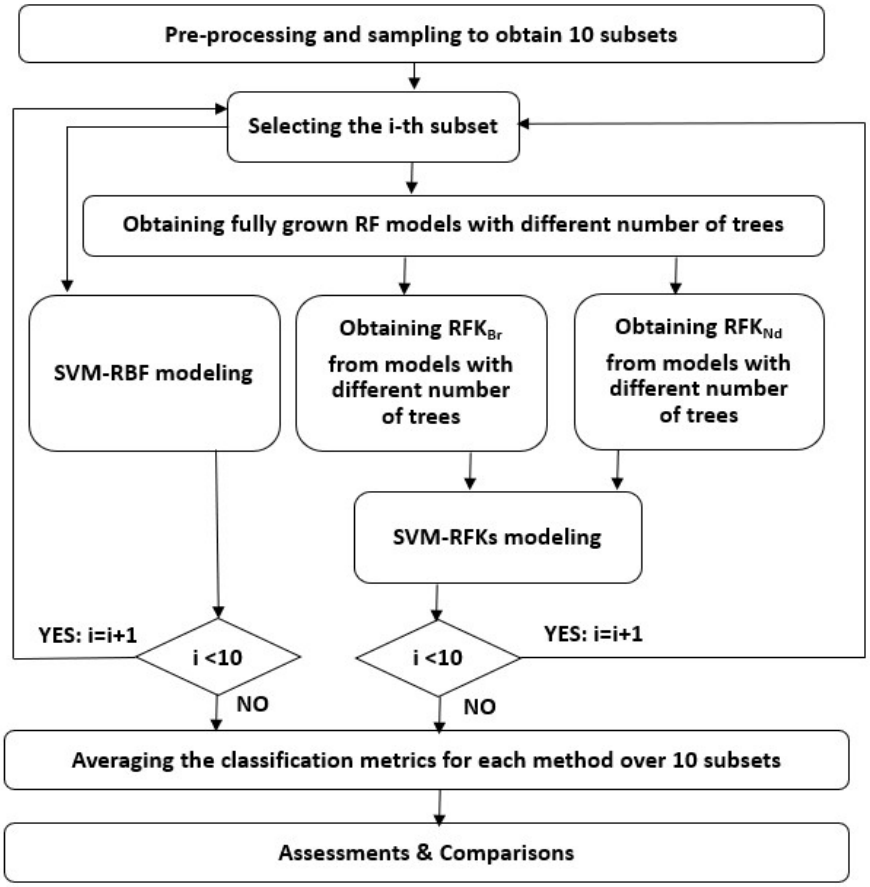


Figure 3.3 Overview of the steps followed to compare RFK_{Br} (i.e., RFK obtained based on the distance of nodes) with RFK_{Nd} (i.e., classic design of RFK) through importing them an SVM.

controlled by setting a maximum number of nodes as the threshold for splitting the nodes in the training phase of the model. Different depths are defined by changing the number of terminal nodes in the trees.
The range considered for the number of terminal nodes is $[3, N_n - 3]$, where N_n is the number of terminal nodes of trees in the fully grown RF and 10 different depths are selected with almost equal intervals. First, we compare the performance of RFK_{Nd} over the 10 different depths for all the 10 subsets. After that, we obtain two multi-scale kernels by averaging RFK_{Prob} and RFK_{Nd} over the 10 depths. The multi-scale RFKs are noted as $RFK_{\overline{Prob}}$ and $RFK_{\overline{Nd}}$. The results for the depths with the best Overall Accuracy (OA) are shown with RFK_{Nd^*} and RFK_{Prob^*}.

Moreover, we evaluate the performance of the multi-scale RFK_{Nd} by varying the number of depths used to generate this kernel. This means that rather than using 10 depths, we use 2 to 9 depths (with almost equal intervals) to obtain the multi-scale RFK_{Nd} and we evaluate the performance of these kernels in an SVM. We compare these kernels in terms of averaged OA, κ index, and computational time. The computational times for each classifier were estimated using the ksvm function in the kernlab package of R [108]. The custom kernel of this package were used to obtain RBF and RFKs classifications in an SVM.

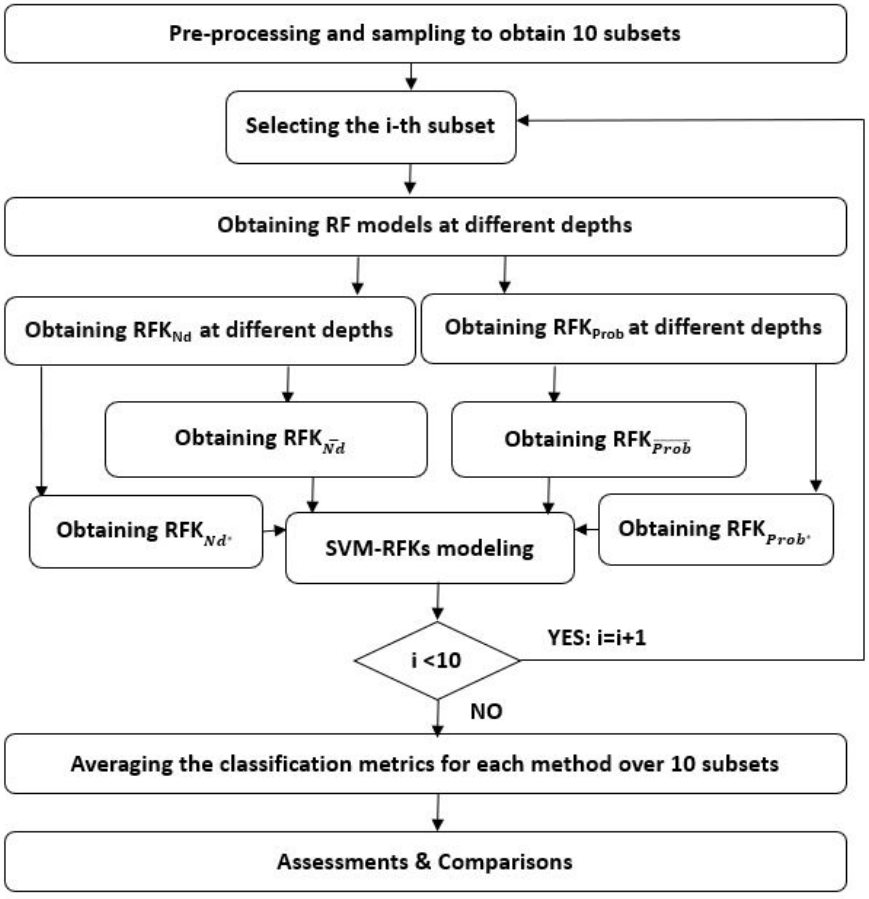


Figure 3.4 Overview of the steps followed to compare depth-based RFKs. Notation: $RFK_{\overline{Nd}}$ and $RFK_{\overline{Prob}}$ denote multi-scale RFKs obtained respectively with RFK_{Nd} and RFK_{Prob} at different depths. RFK_{Nd^*} and RFK_{Prob^*} denote the kernels at the depth with the best Overall Accuracy (OA).

3.5 Results and discussion

In this section, we discuss and compare the classification results for the experiments described in sections 3.2 and 3.3. The results are compared in terms of averaged OA, κ index, and computational time for 10 test subsets.

3.5.1 Comparing RFK_{Br} and RFK_{Nd}

The performance of the SVM-RFK_{Nd} and SVM-RFK_{Br} classifiers are compared in Table 3.2 which displays the $\overline{OA}$s versus the different number of trees. Table 3.2 shows that these two classifiers perform at about same level for different number of trees and different tests shown in Table 3.1. However, the computational load of RFK_{Br} is higher compared to RFK_{Nd}. The reason is that w parameter should be optimized for RFK_{Br}, and finding the number of branches between two nodes requires more computational load than checking if two samples land in the same terminal node.
The performance of the SVM-RFK_{Nd} and SVM-RFK_{Br} classifiers was also compared against that of the SVM-RBF classifier (Table 3.3). Focusing on experiments with B and BVI features, all three kernels perform at about the same level. The experiments with $BVITVI$ and All features show that RFKs considerably outperform RBF kernel for high-dimensional and possibly noisy problems, increasing the difference in $\overline{OA}$ up to around 7%. The performance of RFK_{Br} and RFK_{Nd} are about at the same level for all experiments while RFK_{Br} shows marginal improvements in experiments with B, BVI, and All features. However, considering the trade-off between the computational load and $\overline{OA}$, the use of RFK_{Nd} is generally preferred. It is worth mentioning that using RFKs in SVM improves the $\overline{OA}$s of the classifications in comparison to standard RF. The higher the dimensionality of the experiments, the higher the improvement in $\overline{OA}$s, reaching to 2.5% for RFK_{Nd}.

3.5.2 Multi-scale RFKs

The performance of RFK_{Nd} versus different depths of RF is compared in Table 3.5 and Figure 3.5. Table 3.5 shows that for almost all subsets and experiments (except for s_6 in experiment with BVI features) optimizing the depth results in an improved OA of the classification for RFK_{Nd}. In Figure 3.5, we illustrate the performance of RFK_{Nd} against different depths for the subset 9 (i.e., sub_9). The reason of selecting sub_9 is that optimizing depth yielded to the largest improvement in OA of SVM-RFK_{Nd} in the experiment with ALL feature. Figure 3.5 shows that for the low dimensional tests (i.e., B and BVI), the optimized depth is closer to the depth of fully

Table 3.2 Classification results obtained in terms overall accuracies ($\overline{OA}$) over 10 test subsets for SVM-RFK$_{Nd}$ and SVM-RFK$_{Br}$ classifiers versus the number of trees for four candidate feature subsets (N_f) defined in Table 3.1.

Tests	Methods	Number of Trees											
		5		10		20		50		100		500	
		$\overline{OA}$	SD	$\overline{OA}$	SD	$\overline{OA}$	SD	$\overline{OA}$	SD	$\overline{OA}$	SD	$\overline{OA}$	SD
1:B	SVM-RFK$_{Br}$	68.96	1.78	73.38	1.25	76.72	1.81	79.10	2.03	80.16	1.33	81.36	1.29
	SVM-RFK$_{Nd}$	70.00	1.90	73.40	1.28	76.68	1.75	79.10	2.00	80.12	1.35	81.34	1.27
2:BVI	SVM-RFK$_{Br}$	67.64	1.25	73.66	2.15	77.20	2.08	80.08	1.27	80.98	1.38	81.66	1.23
	SVM-RFK$_{Nd}$	68.40	1.61	73.50	2.36	77.12	1.97	80.08	1.27	81.00	1.37	82.14	1.05
3:BVITVI	SVM-RFK$_{Br}$	66.44	1.83	73.12	1.74	78.48	2.28	82.38	1.42	83.16	1.55	84.60	1.24
	SVM-RFK$_{Nd}$	66.80	1.94	72.94	1.85	78.32	2.47	81.78	1.41	82.80	1.49	84.66	1.17
4:ALL	SVM-RFK$_{Br}$	66.82	3.38	75.22	1.50	79.10	1.72	83.00	2.11	83.66	2.48	84.80	1.59
	SVM-RFK$_{Nd}$	67.46	3.48	75.36	1.35	79.08	1.60	83.00	2.11	83.60	2.52	85.16	1.32

Table 3.3 Classification results obtained for the experiments in Table 3.1. RF models trained with 500 fully grown trees are used to obtain RFK_{Br} and RFK_{Nd}. $\overline{OA}$ (in %) is the averaged overall accuracy, SD (in %) is its standard deviation, $\bar{\kappa}$ is the averaged Cohen's kappa index, and SD$_\kappa$ is its standard deviation.

Tests	Methods	$\overline{OA}$	SD	$\bar{\kappa}$	SD$_\kappa$
1:B	SVM-RFK$_{Br}$	81.36	1.29	0.76	0.02
	SVM-RFK$_{Nd}$	81.34	1.27	0.76	0.02
	SVM-RBF	82.08	2.16	0.77	0.03
	RF	81.08	1.34	0.76	0.02
2:BVI	SVM-RFK$_{Br}$	81.66	1.23	0.77	0.02
	SVM-RFK$_{Nd}$	82.14	1.05	0.78	0.01
	SVM-RBF	83.44	1.46	0.79	0.02
	RF	80.40	1.34	0.76	0.02
3:BVITVI	SVM-RFK$_{Br}$	84.60	1.24	0.81	0.02
	SVM-RFK$_{Nd}$	84.66	1.17	0.81	0.01
	SVM-RBF	77.38	1.03	0.72	0.01
	RF	82.12	1.71	0.78	0.02
4:ALL	SVM-RFK$_{Br}$	84.80	1.59	0.81	0.02
	SVM-RFK$_{Nd}$	85.16	1.32	0.81	0.02
	SVM-RBF	78.72	1.04	0.73	0.01
	RF	82.68	1.32	0.78	0.02

grown trees while for the high dimensional tests (i.e., $BVITVI$ and ALL), shallower depths are found to be optimal.
The performance of the two multi-scale RFKs over 10 depths based on the terminal nodes and class probabilities are compared in Table 3.6. This table also presents the results for the best depth (RFK_{Nd^*} and RFK_{Prob^*}). In the following paragraph, we compare

the multi-scale and best depth results with the ones obtained with the RBF and RFK_{Nd} kernels (depicted in Table 3.3).
Focusing on the experiment with B features, multi-scale RFKs (i.e., $RFK_{\overline{Prob}}$ and $RFK_{\overline{Nd}}$) and RFK_{Nd} perform at about same level considering their $\overline{OA}$s and SD values. However, RFK_{Nd^*} and RFK_{Prob^*} with 1.5% gain in OA slightly outperform RFK_{Nd}. For the experiment with B features, RFK_{Nd^*} and RFK_{Prob^*} slightly outperform RBF kernel considering $\bar{\kappa}$ and SD of $\overline{OA}$. Focusing on the experiment with BVI features, RFK_{Nd^*} with $\overline{OA}$ of 83.62% outperforms all other RFKs. RFK_{Nd^*} also outperforms RBF kernel considering $\bar{\kappa}$ and SD of $\overline{OA}$. The lowest $\overline{OA}$ and $\bar{\kappa}$ obtained for the experiment with BVI features are obtained by RFK_{Prob^*} and $RFK_{\overline{Prob}}$.
Focusing on the experiment with $BVITVI$ features, the highest $\overline{OA}$s of 88.62% and 86.04% are respectively obtained for RFK_{Nd^*} and $RFK_{\overline{Nd}}$. All RF-based kernels outperform RBF in this experiment while RFK_{Prob^*} and $RFK_{\overline{Prob}}$ obtain the worst results among the RF-based kernels. Focusing on the experiment with All features, the highest $\overline{OA}$s of 89.48% and 86.18% are respectively obtained for RFK_{Nd^*} and $RFK_{\overline{Nd}}$. Again for this experiment, RFK_{Prob^*} and $RFK_{\overline{Prob}}$ obtain the worst results among the RF-based kernels while all RF-based kernels outperform RBF results. Using the depth-based RFKs in an SVM also improves the $\overline{OA}$s of the classifications compared to standard RF. This improvement is more noticeable in case of higher dimensional experiments by about 3.5 to 6.8% for $RFK_{\overline{Nd}}$ and RFK_{Nd*}. Overall, RFK_{Nd^*} outperforms other kernels including RFK_{Nd} in all experiments. In other words, we found that SVM-RFK_{Nd^*} is the best classifier in terms of OA and Kappa.

The outperformance of RFK_{Nd^*} is small compared to RFK_{Nd} in experiments with B and BVI features, but it makes a considerable improvement up to around 4% in $\overline{OA}$ when the number of features is highly increased in experiments with $BVITVI$ and All features. This evidences that optimizing the depth of RF can improve the results for the classic RFK.We have also benchmarked our approach against the proposed method in [165]. To do so, we have obtained 100 most important features using RF for the subset with ALL features. These features are imported into an SVM-RBF. The obtained $\bar{OA}$ and SD for this approach are 89.02 % and 1.83 %. These results are slightly lower than best results of our work which is obtained for SVM-RFK_{Nd^*} with $\bar{OA}$ and SD of 89.48 % and 1.32 %.
For the higher number of features, obtaining a multi-scale RFK by averaging RFK_{Nd} over multiple depths also led to an improvement of around 1% in $\overline{OA}$ compared to classic RFK. Obtaining a multi-scale RFK based on the class probabilities gives competitive results only in case of the experiment with B features. The reason is the higher dependency of obtained similarity values among samples for this kernel on the class labels. For high dimensional and possibly noisy datasets, the probability that a sample is correctly classified

Table 3.4 Computational time

Experiments	Methods	Time (Minutes)
1:B	SVM-RFK$_{Nd}$	7.30
	SVM-RFK$_{\overline{Nd}}$	8.87
	SVM-RFK$_{Nd^*}$	71.42
	SVM-RFK$_{\overline{Prob}}$	8.72
	SVM-RFK$_{Prob^*}$	71.27
	SVM-RBF	7.77
	g-matrices	1322.70
2:BVI	SVM-RFK$_{Nd}$	8.61
	SVM-RFK$_{\overline{Nd}}$	21.23
	SVM-RFK$_{Nd^*}$	83.78
	SVM-RFK$_{\overline{Prob}}$	21.39
	SVM-RFK$_{Prob^*}$	83.94
	SVM-RBF	75.55
3:BVITVI	SVM-RFK$_{Nd}$	33.31
	SVM-RFK$_{\overline{Nd}}$	228.95
	SVM-RFK$_{Nd^*}$	291.50
	SVM-RFK$_{\overline{Prob}}$	218.75
	SVM-RFK$_{Prob^*}$	281.30
	SVM-RBF	288.00

decreases and this affects directly the similarity values in the kernel. This explains the relatively poor performance of this kernel in our experiments.

Table 3.5 Improvement of the classification results of SVM-RFK$_{Nd^*}$ compared to SVM-RFK$_{Nd}$ in terms of OA. The results are shown for 10 pairs of training and test subsets in the experiments with different dimensionality (Table 3.1). Notation: sub_i denotes subset i.

Tests	sub_1	sub_2	sub_3	sub_4	sub_5	sub_6	sub_7	sub_8	sub_9	sub_{10}
1:B	2,00	2,00	3,40	3,20	1,40	0,40	0,80	1,40	0,20	3,20
2:BVI	0,80	0,20	0,40	1,60	1,20	0,00	0,40	1,40	0,20	2,00
3:BVITVI	3,00	2,20	2,60	3,80	2,40	3,60	3,20	3,80	3,40	2,00
4:ALL	3,20	2,40	3,20	2,00	4,20	1,60	3,20	2,80	4,80	1,40

The analysis of the classifications results for each class is carried out by mean of the average of F-scores ($\overline{F}$) obtained over 10 subsets. Table 3.7 shows the results of $\overline{F}$ for the top performing classifiers. Table 3.7 depicts that the highest class separabilities also are achieved for RFK_{Nd^*} and $RFK_{\overline{Nd}}$. Moreover, the RF-based kernels frequently obtain higher scores than the RBF.
The computational times for each classifier and for the subsets are shown in Table 3.4. The time required for obtaining g-matrices is considerably higher than the one required for the other methods.

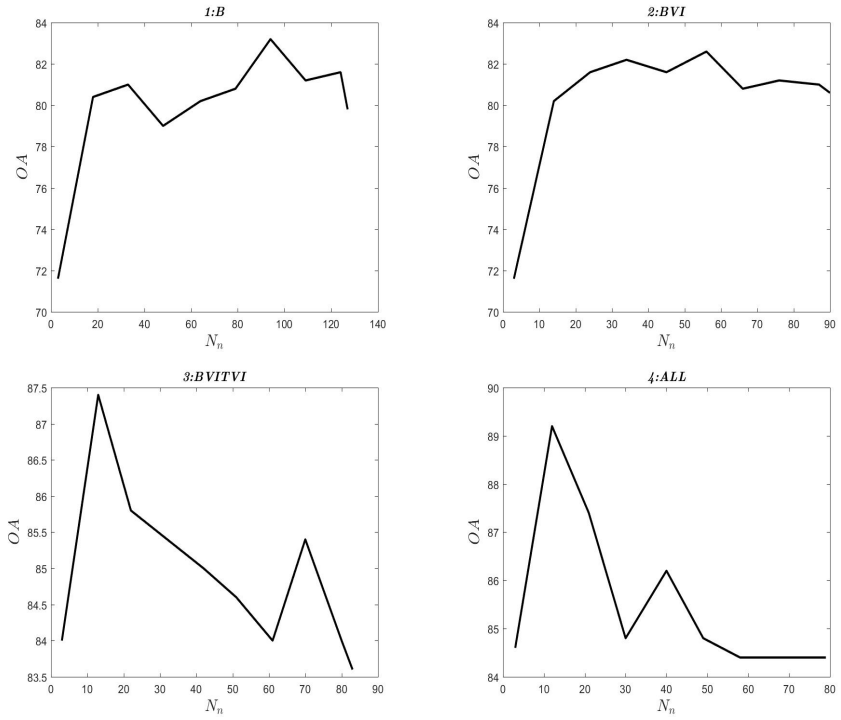


Figure 3.5 The OA obtained for SVM-RFK$_{Nd}$ classifier at 10 different depths of RF for four tests. Different depths are defined by changing the number of terminal nodes (N_n) in the trees. The panels in this figure show the classification results corresponding to the sub_9 which yields the greatest improvement in OA of RFK_{Nd^*} compared to RFK_{Nd}.

Therefore, we skipped obtaining the time required for importing g-matrices in an SVM and we did not calculate the time required to get these matrices for the experiment with other subsets of features. As it is shown, the time required for obtaining SVM-RBF cannot beat $SVM - RFK_{Nd}$ when the number of features grows.

At the end, the performance of the $RFK_{\overline{Nd}}$ against the number of the depths used to generate this kernel is evaluated in terms of $\overline{\boldsymbol{OA}}$ in Figure 3.6. Figure 3.6 shows how the $RFK_{\overline{Nd}}$ performs in an SVM against using different number of depths from 2 to 10 depths. Figure 3.6 shows that optimizing the number of depths marginally improves the $\overline{\boldsymbol{OA}}$ compared to the use of 10 depths for the experiments with B, BVI and ALL features. Thus, considering the trade-off between the gained $\overline{\boldsymbol{OA}}$ and the added computational load, the use of a default number of depths which generally stabilizes the $\overline{\boldsymbol{OA}}$ for $RFK_{\overline{Nd}}$ is preferred.

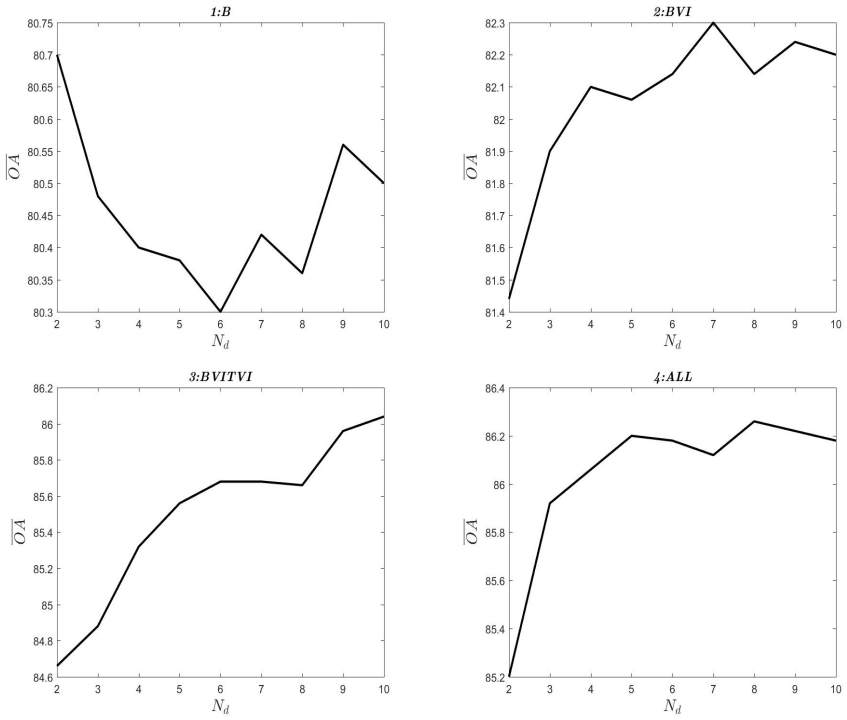


Figure 3.6 The performance of multi-scale SVM-RFK$_{Nd}$ in terms of $\overline{OA}$ (i.e., averaged OA over 10 subsets) against varying the number of the depths used to generate this kernel. N_d shows the number of depths.

3.6 Conclusion

In this chapter, we investigate the connection between RF and kernel methods by exploring different RF characteristics. To overcome the limitations of classic RFK, we designed novel RFKs by using the distance among terminal nodes, obtaining classic RFK and RF-based class probabilities at multiple depths. We evaluated these novel kernels by comparing their performances in an SVM against classic RBF and classic RFK for a crop classification problem over small-scale farms. We also compared the performances of the RFKs in an SVM against that provided by a standard RF classifier. A time series of WV2 images was used to illustrate the study. In general, using the proposed kernels in an SVM outperformed standard RF. In all experiments, the RFKs obtained based on the number of branches performed at about the same level of classic RFK while the computational cost for classic RFK is considerably lower. For low dimensional experiments, RBF kernel and the classic RFK at an optimized depth slightly outperform other kernels while all kernels perform at about the same level considering the OAs and their SDs. It is worth men-

Table 3.6 The influence of using 10 depths on the classification results obtained for the cases in Table 3.1. $\overline{OA}$ (in %) is the averaged overall accuracy, SD (in %) is the standard deviation, $\bar{\kappa}$ is the averaged Cohen's kappa index , SD$_\kappa$ is the standard deviation for κ values.

Tests	Methods	$\overline{OA}$	SD	$\bar{\kappa}$	SD$_\kappa$
1:B	SVM-RFK$_{\overline{Nd}}$	80.50	0.73	0.76	0.01
	SVM-RFK$_{Nd^*}$	82.84	1.18	0.79	0.01
	SVM-RFK$_{\overline{Prob}}$	81.12	1.48	0.76	0.02
	SVM-RFK$_{Prob^*}$	82.84	1.27	0.79	0.02
2:BVI	SVM-RFK$_{\overline{Nd}}$	82.20	0.95	0.78	0.01
	SVM-RFK$_{Nd^*}$	83.62	0.57	0.80	0.01
	SVM-RFK$_{\overline{Prob}}$	80.88	1.28	0.76	0.02
	SVM-RFK$_{Prob^*}$	81.13	1.94	0.76	0.02
3:BVITVI	SVM-RFK$_{\overline{Nd}}$	86.04	0.70	0.83	0.01
	SVM-RFK$_{Nd^*}$	88.62	0.33	0.86	0.00
	SVM-RFK$_{\overline{Prob}}$	82.42	1.32	0.78	0.02
	SVM-RFK$_{Prob^*}$	83.91	0.65	0.80	0.01
4:ALL	SVM-RFK$_{\overline{Nd}}$	86.18	1.71	0.83	0.02
	SVM-RFK$_{Nd^*}$	89.48	1.32	0.87	0.02
	SVM-RFK$_{\overline{Prob}}$	82.74	1.78	0.78	0.02
	SVM-RFK$_{Prob^*}$	84.84	1.34	0.81	0.02

Table 3.7 F-score average ($\overline{F}$) and the corresponding standard deviation (SD) for the different classifiers.

Test	Classes	SVM-RFK$_{Nd^*}$		SVM-RFK$_{\overline{Nd}}$		SVM-RFK$_{Nd}$		SVM-RBF		SVM-RFK$_{Br}$	
		$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD	$\overline{F}$	SD
1:B	Cotton	0.81	0.02	0.80	0.02	0.8	0.02	0.79	0.03	0.8	0.02
	Maize	0.78	0.02	0.77	0.02	0.78	0.02	0.8	0.02	0.78	0.02
	Millet	0.86	0.02	0.83	0.02	0.85	0.02	0.87	0.03	0.85	0.02
	Peanut	0.79	0.02	0.77	0.02	0.79	0.02	0.79	0.04	0.79	0.02
	Sorghum	0.86	0.02	0.85	0.02	0.86	0.02	0.86	0.02	0.86	0.02
2:BVI	Cotton	0.84	0.03	0.84	0.03	0.83	0.03	0.81	0.02	0.83	0.03
	Maize	0.79	0.02	0.78	0.02	0.78	0.02	0.82	0.02	0.78	0.02
	Millet	0.86	0.02	0.86	0.03	0.85	0.03	0.89	0.01	0.85	0.03
	Peanut	0.8	0.02	0.79	0.01	0.79	0.02	0.8	0.03	0.79	0.02
	Sorghum	0.85	0.02	0.85	0.02	0.84	0.03	0.86	0.02	0.84	0.03
3:BVITVI	Cotton	0.88	0.02	0.87	0.01	0.85	0.03	0.76	0.02	0.85	0.03
	Maize	0.86	0.02	0.84	0.03	0.81	0.03	0.73	0.02	0.81	0.03
	Millet	0.88	0.02	0.87	0.02	0.86	0.02	0.8	0.02	0.86	0.02
	Peanut	0.86	0.02	0.85	0.02	0.83	0.02	0.76	0.02	0.83	0.02
	Sorghum	0.88	0.02	0.87	0.02	0.86	0.01	0.82	0.02	0.86	0.01
4:ALL	Cotton	0.89	0.02	0.85	0.03	0.86	0.03	0.76	0.03	0.86	0.03
	Maize	0.88	0.02	0.81	0.03	0.83	0.02	0.77	0.02	0.83	0.02
	Millet	0.89	0.02	0.82	0.04	0.85	0.03	0.81	0.02	0.85	0.03
	Peanut	0.87	0.02	0.81	0.03	0.85	0.01	0.77	0.02	0.85	0.01
	Sorghum	0.87	0.02	0.84	0.03	0.85	0.02	0.83	0.01	0.85	0.02

tioning that all RF-based kernels obtained for high dimensional and possibly noisy features considerably improve the classification res-

ults obtained by the standard SVM-RBF classifier and this improvement is 4 to 11 % in terms of OA. Overall, SVM-RFK_{Nd^*} which corresponds to a classic RFK at an optimized depth, leads to the best results. In particular and compared to classic RFK, it results in an improvement of 4 to 5.46 % in terms of OA for higher dimensional experiments. Although the proposed kernels show high overall accuracies in a complex classification problem, future work is required to evaluate their performance with other datasets, land cover types, and kernel-based methods. Future work is also required to improve the rough binary estimation of similarity values through a better notion of probability.

4 Land cover classification using extremely randomized trees: a kernel perspective

This chapter is based on the published paper:
A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land cover classification using extremely randomized trees: A kernel perspective," IEEE Geoscience and Remote Sensing Letters, pp. 1–5, 2019.

Abstract

The classification of the ever-increasing collections of remotely sensed images is a key but challenging task. In this chapter, we introduce the use of Extremely Randomized Trees known as Extra-Trees (ET) to create a similarity kernel (ETK) which 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 small-holder 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 grey-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 datasets.
Keywords: Image classification, random forest, support vector machine, smallholder agriculture, very high spatial resolution satellite images

4.1 Introduction

With the advent of new sensors and open data policies, large datasets are becoming available. This includes large collections of remotely sensed (RS) images, which often need to be classified to support their use in various domains and applications [22]. Yet, traditional classification methods cannot properly deal with the challenge of handling large and complex datasets [174]. 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 [175]. Extra features often lead to high-dimensionality which is the most important challenge in recent RS image classification tasks [174].

Kernel methods can efficiently deal with non-linear 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 [29]. 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 [29]. This is a limitation of SVM-RBF [166]. Another limitation to accuracy and efficiency of SVM-RBF is experienced when the number of features increases for a certain amount of training data [62]. The reason for this is the curse of dimensionality, also called 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 [62, 176]. Another well-known classifier for high dimensional problems is Random Forest (RF) [177]. 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 are resistant to noise in training samples and to overfitting [98]. These characteristics alongside with its tree-based structure make it a suitable classifier to draw partitions in the data and to obtain an RF Kernel (RFK) that quantifies similarities between samples [104, 166]. The pairwise similarities between the samples reflect whether they fall in the same end-node or not [178]. 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 dataset often used in benchmarking studies (i.e. Salinas) and for a time series of Worldview-2 images over Sukumba, Mali [166]. Hence, RFK is an effective alternative to RBF, particularly when combined with RF-based feature selection methods [166]. 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 [94]. Moreover, RFK is negatively impacted by possible mislabeled samples in the training data. To overcome these downsides, the randomization level of the RF ensemble 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) [106]. 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 [106]. In extreme cases, ET builds totally randomized trees (ToRT). The structure of these trees is independent of the training labels [106]. The randomization level can be adjusted to the problem at hand by selecting suitable parameters [106]. Several papers 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) [99, 100]. Besides OA, the main strong point of ET is its computational efficiency [106]. 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 [106]. Using ETK as an alternative to RBF and RFK, one can avoid computational cost associated with parametrizing the RBF kernel and reduce the probability of getting an overfitted kernel.
The main goal of this chapter 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.

4.2 Extra-trees 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 defined by a stopping criterion [106]. ET differs from other decision-tree based ensembles such as RF in two cases [99]. 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 [99]. 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 [106]. Furthermore, the computational load of training for ET is less than that of required to train RF since it does not search intensively for an optimal cut-point [106]. 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*. Second common parameter is N_t which 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 deceasing function of number of trees [94, 106]. Third common parameter is n_{min} which is minimum sample size for splitting a node with the default value of one (or two) [94, 106]. The optimal value for n_{min} increases depending on the level of mislabeled samples in training data [106]. The higher values for this parameter result in smaller trees, smaller variance, and higher bias [94, 106]. 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 [106]. This is typically called

totally randomized trees, 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 [99, 106]. When ET uses more than one feature or/and random cut-point in splitting the nodes, like RF, it uses the Gini index or a normalization of information gain to select the best cut-point out of the randomly selected splits [106].
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 [106]. 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 dependency 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 number of times that 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 [99, 106]:

$$K_{t_n}(x_i, x_j) = I\left[q(x_i) = q(x_j)\right], \tag{4.1}$$

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

$$\mathbf{ETK} = \frac{1}{N_t}\sum_{t_n=1}^{N_t}\mathbf{K}_{t_n}, \tag{4.2}$$

N_t being 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).

4.3 Data and experiments

4.3.1 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 dataset includes seven multispectral images that cover the cropping season of 2014 [153]. Ground truth labels for 5 common crops including cotton, maize, millet, peanut, and sorghum were collected for 9 fields per crop (45 fields) through fieldwork. The Sukumba images are atmospherically corrected, co-registered, and trees and

clouds are masked [153]. These images and the corresponding ground data are part of the STARS project.
The Sukumba dataset originally contains 56 bands (i.e. 7 images with 8 bands each). The number of features was extended by obtaining Normalized Difference Vegetation Index (NDVI), Ratio Vegetation Index (RVI), Soil Adjusted Vegetation Index (SAVI), Modified Soil-Adjusted Vegetation Index (MSAVI), Transformed Chlorophyll Absorption Reflectance Index (TCARI), and Enhanced Vegetation Index (EVI). The dataset, the study area, a list and short explanation of VIs used in this study can be found in [166]. Furthermore, the pairwise band combinations by means of the difference, ratio and normalization between bands 2 to 8 were generated increasing the number of the features until 525. Next, the number of features for Sukumba dataset 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 [160]. The GLCM textures derived for Sukumba dataset are also presented and explained comprehensively in [124, 166]. For each spectral and VI feature, statistical textures including angular second moment, correlation, inverse difference moment, sum variance, entropy, difference entropy, information measures of correlation, dissimilarity, inertia, cluster shade, and cluster prominence are obtained [124]. Concatenating spectral, VIs and GLCM features obtained for both spectral and VIs features increases the number of features to 10536. Table 4.1 shows the subsets and quantity of the features which are used in four tests to examine the proposed method in this study. In such large datasets 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).

Table 4.1 Experiments description (N_f: Number of features.)

Acronim	Features	N_f
1:B	Spectral features	56
2:BVI	Spectral and VIs features	525
3:BVITVI	BVI and GLCM textures of VIs	8498
4:ALL	BVITVI, textures of spectral features, and additional features	10584

4.3.2 Experimental set-up

First, the polygons representing farms were split into four sub-polygons. Two sub-polygons were used to choose the training samples and the other two, the test samples. Then, the train and test sets were split into 10 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 10 subsets available for each spectral case (Table 4.1). 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 [106]. 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 [166]. The optimization of RF parameters for obtaining the RFK is skipped because of marginal gain in OA of SVM-RFK compared to added computational cost [166]. 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, a 5-fold cross-validation was used to find the optimal bandwidth and C values. A 5-fold 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 [161]. Classification results are compared in terms of their average Overall Accuracy ($\overline{OA}$) 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 other 10 sets of train and test samples. For visibility reasons, 2 classified fields per crop are shown for each classifier. At the end, OA and κ of the top performing classifiers are obtained for all available labeled samples in the 45 fields.

4.4 Results and discussion

Fig. 4.1 displays the $\overline{OA}$s of 10 test subsets versus different parameters configurations for ET and SVM-ETK classifiers. Fig. 4.1 shows that SVM-ETK always outperforms ET for all the cases with VIs (Table 4.1) and irrespective of the value of N_t and N_{cp}. For the experiment with B features, the $\overline{OA}$ 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. 4.1 that the peaks of $\overline{OA}$ for SVM-ETK correspond to higher levels of randomization (i.e., N_{cp} equal or less than 10). Nonetheless, the difference between the $\overline{OA}$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 $\overline{OA}$s, and optimizing the N_{cp} results in 1 % improvement in OA for the experiments with $BVITVI$ and ALL features. Fig. 4.1 also shows that the higher number of trees (i.e., 300 and 500) generate higher $\overline{OA}$s for both ET and SVM-ETK.

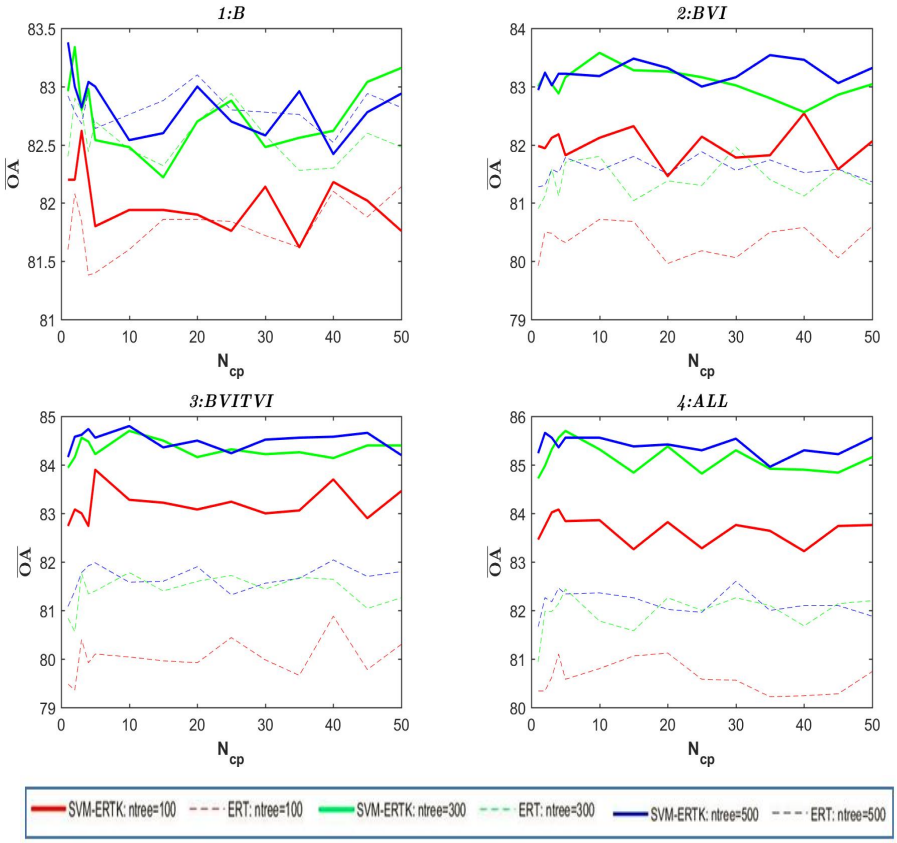


Figure 4.1 The $\overline{OA}$ for SVM-ETK and ET classifiers versus the number of random cut-points for each candidate feature (N_{cp}) for the four experiments.

Table 4.2 compares $\overline{OA}$ and $\bar{\kappa}$ of the best and default configurations of SVM-ETK with SVM-RBF, SVM-RFK, and ET classifiers. In Table 4.2, 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 $\overline{OA}$ of 83.38 % slightly outperforms both SVM-RFK and SVM-RBF with $\overline{OA}$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 gives the best $\overline{OA}$ 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 the $\overline{OA}$ and SD of test subsets, and these two classifiers outperform ET and SVM-RFK.
Increasing the number of features to 8450 in the experiment with $BVITVI$ features results in a decrease of 6.62 % in $\overline{OA}$ for SVM-RBF while it slightly improves the results of ET*, SVM-ETKs, and SVM-RFK.

Using the tree-based kernels in an SVM for this experiment gives almost equal results considering their $\overline{OA}$ and SD.
In the fourth experiment with 10536 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 tree-based classifiers. The highest $\overline{OA}$ 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 are not statistically significant for higher-dimensional cases including the experiment with BVI features. For the experiment with B features, the McNemar test shows that 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 outperforms RFK for the lowest dimensional case.

Table 4.2 Classification results for different cases and classifiers. N_t and N_{cp} are respectively number of trees and number of random cut-points per candidate feature. $*$ and d are respectively best and default configurations.

Case	Classifier	N_t	N_{cp}	$\overline{OA}$	SD	$\bar{\kappa}$	SD$_\kappa$
	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
B	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
	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
BVI	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
	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
BVITVI	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
	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
ALL	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

The $\overline{OA}$ and $\bar{\kappa}$ results for ToRT and totally randomized trees kernel in an SVM (SVM-ToRTK) are shown in Table 4.3. These results were only obtained for 300 and 500 trees considering the results of the experiments shown in Fig. 4.1. Table 4.3 shows that using 300 and 500 trees generates similar results. These results show that SVM-ToRTK outperforms ToRTK. Comparing Tables 4.2 and 4.3, SVM-ToRTK, SVM-ETKs, and SVM-RBF yield similar $\overline{OA}$ 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. Totally randomized trees act 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 $\overline{OA}$ and reducing its pertinent SD in comparison with ETK) and reduces the computational cost.

Table 4.3 Classification results of totally randomized trees (ToRT) and totally randomized trees 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

Finally, we present maps, $\overline{OA}$, and $\bar{\kappa}$ (Table 4.4) 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 4.4, all classifiers perform good

Table 4.4 $\overline{OA}$ 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**

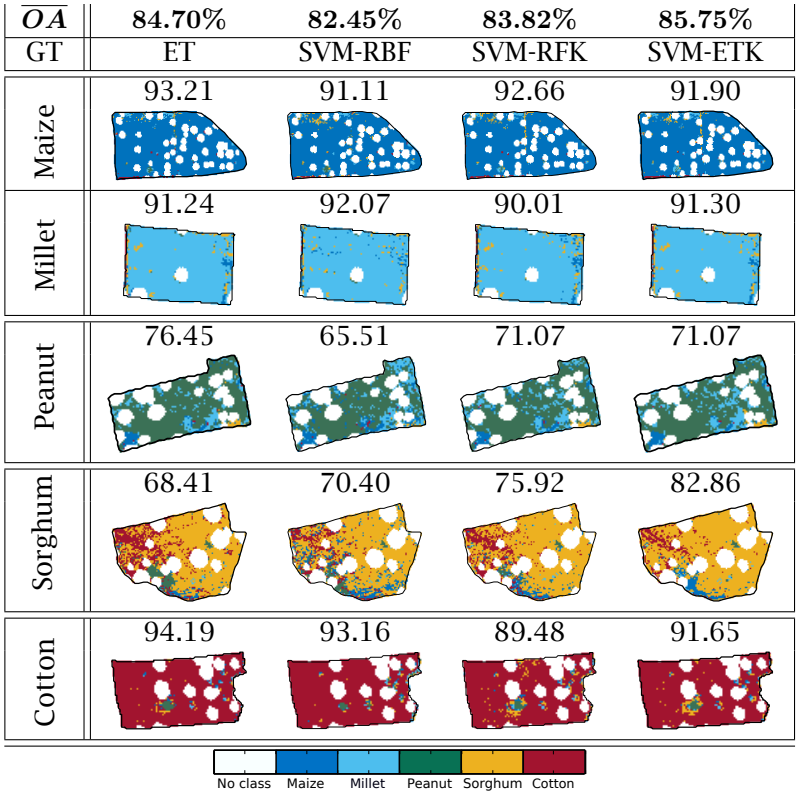


Figure 4.2 A crop field per ground truth class along with their OA obtained for the different classifiers using B, and the $\overline{OA}s$ for 5 fields on top.

and at about same level when obtained based on B features and when applied to whole study area while SVM-ETK slightly outperforms by improving $\overline{OA}$ and $\bar{\kappa}$. For visibility reasons, we only present classified fields. In particular, Fig. 4.2 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 $\overline{OA}$s for these polygons, and SVM-RBF gives the lowest $\overline{OA}$.

4.5 Conclusion

In this chapter, 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 10536 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 totally randomized trees (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 datasets.

5 TreeBasedKernels: an R-function for obtaining kernels based on tree-based ensemble learners

This chapter presents an R-function to obtain the tree-based kernels introduced and evaluated in the previous chapters. The R-function is freely distributed in the DANS repository. The instructions and the source code are available at DOI: 10.17026/dans-247-y9x3

This chapter presents an R-function that implements the various designs of tree-based kernels introduced and evaluated in this thesis. This function provides support for the shift toward open science, facilitates the reproducibility of the results presented in this thesis, and enables researchers who are interested in further testing or expanding tree-based kernels.
This chapter is organized into two main sections. The first one briefly summarizes the mathematical formulation of the kernels. The second section documents the function and its parameters.

5.1 A brief review of tree-based kernels

The structure of tree-based ensemble learners such as Random Forest (RF) and Extra-Trees (ET) can be used to create partitions in the data and to generate tree-based kernels that encode similarities among samples based on these partitions. The basic idea is that samples falling in the same partition are similar and those falling in different partitions are dissimilar. The similarity values among samples can be obtained based on various partitions and with different configurations of RF or ET. Thus, different structural parameters of RF and ET are used to generate various designs of tree-based kernels in the following manner:

1. ***RFK$_{Nd}$***: This is the classic design of an RF kernel. It uses the terminal nodes as partitions of a trained RF to calculate the similarity among data. In this design, if two samples land in the same terminal node of a tree, the similarity is equal to one; otherwise, it is zero. For each tree, one pairwise similarity matrix (K_{t_n}) is generated and RFK_{Nd} is the average of the matrices obtained for all trees [166, 105].

$$\mathbf{RFK_{Nd}} = \frac{1}{N_t}\sum_{n=1}^{N_t}\mathbf{K}_{t_n}, \tag{5.1}$$

   where t_n is the $n-th$ tree of RF and N_t is the number of trees.

2. ***RFK$_{Br}$* and g-matrices:** This RF kernel is based on measuring distance (i.e., the number of branches) between the terminal nodes containing samples i (s_i) and j (s_j). Here, we indicate this kernel by RFK_{Br} and define it as [167]:

$$\mathbf{RFK_{Br}}(s_i, s_j) = \frac{1}{N_t}\sum_{n=1}^{N_t}\frac{1}{e^{w.g_{ijt_n}}}, \tag{5.2}$$

   where n runs over the number of trees in RF, the parameter w controls the influence of the distance between two terminal

nodes occupied by the samples s_i and s_j, and g_{ijt_n} is the number of branches between two terminal nodes containing s_i and s_j in the t_n ($n-th$ tree of RF). If s_i and s_j land on the same terminal node, then g = 0 and RFK_{Br} will be increased by one as in the classical approach (i.e., RFK_{Nd}) to assess the similarity of two samples.
In the TreeBasedKernels function, for each tree, a g-matrix$_{TRTR}$ and a g-matrix$_{TSTR}$ is obtained. The g-matrix$_{TRTR}$ contains the pairwise distance among training samples and the g-matrix$_{TSTR}$ contains the pairwise distance among test and training samples. For obtaining RFK_{Br}, the g-matrices$_{TRTR}$—together with a range of w—must be imported in a kernel-based model to optimize w and the parameters of the model. After training the kernel-based model with the optimal value of the parameters, the g-matrices$_{TSTR}$ obtained with the TreeBasedKernels function can be used to evaluate the kernel-based model.

3. ***MultRFKprb***: A trained RF assigns a probability of membership to each of the classes of interest in each sample—namely, the predicted probability vector. Using a matrix notation, all the predicted probability vectors can be placed in matrix P in the following manner:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \dots & p_{1C} \\ p_{21} & p_{22} & p_{23} & \dots & p_{2C} \\ & & \dots\dots\dots & & \\ p_{n1} & p_{n2} & p_{n3} & \dots & p_{nC} \end{bmatrix} \tag{5.3}$$

P is the matrix of predicted probabilities by RF—one column per class and one row per sample, where C is the total number of classes. Once P is calculated, the RFK_{Prob} is defined as $RFK_{Prob} = P.P^T$. The kernel is obtained at different depths. Therefore, the averaged probabilistic RF kernels obtained at different depths are defined in the following manner:

$$\mathbf{RFK}_{\overline{\mathbf{Prob}}} = \frac{1}{N_d} \sum_{n=1}^{N_d} (\mathbf{P}.\mathbf{P^T})_{d_n,} \tag{5.4}$$

where N_d is the number of the depths considered to obtain RFK_{Prob}, n runs over the number of depths, and d_n is the $n-th$ depth in RF. Different depths can be defined by changing the number of terminal nodes in the trees.

4. ***MultRFKNd***: RFK_{Nd} can be obtained at different depths as well. The average of these kernels can be considered a multi-scale RFK_{Nd}.
5. ***ETK***: Like RF, the tree-based structure of ET can be used to generate a kernel that encodes similarities between samples

based on the terminal nodes [106, 179]. The level of randomization increases in the structure of ET compared to that of RF. Thus, ET kernels are less dependent on the labels of training samples compared to RFK.

5.2 R-function-TreeBasedKernels

5.2.1 Description

This function returns similarity kernels (or proximities) obtained from tree-based ensemble classifiers, namely, RF and ET. For implementing RF and ETs, randomForest (implementing Breiman's random forest algorithm) and ranger packages are used, respectively. These kernels can be used as alternatives to common kernels like the radial basis function (RBF) in kernel-based classifiers like the support vector machine (SVM). To use this function, it is required that the following packages are installed and loaded: caret, e1071, randomForest, ranger, edarf, PEIP, and doSNOW.

Usage

TreeBasedKernels (kernelType, xtrain, xtest, ltr, lts, MtryOp=FALSE, Nt=500, numsplits=1, Vis=TRUE)

Arguments

kernelType	the kernel design used in training and predicting. kernelType provides the above mentioned tree-based kernel designs by setting the kernelType parameter to one of the following strings: **RFKNd:** The classic design of RFK that encodes similarities among samples based on the terminal nodes of a fully grown RF classifier are returned by the function. **gmatrices:** The number of branches between the terminal nodes containing the samples are returned by the function. The g-matrix$_{TRTR}$ and g-matrix$_{TSTR}$ are stored in the working directory. To obtain RFK_{Br}, the g-matrices with a range of w must be imported in a kernel-based model to optimize the w parameter of RFK_{Br} and the parameters of the model. **MultRFKprb:** The average of probabilistic RFKs obtained at 10 different depths are returned by the function.

	MultRFKNd: The average of RFK_{Nd} obtained at 10 different depths are returned by the function. RFK_{Nd} obtained at different depths is stored in the working directory. An optimized depth for RFK_{Nd} can be defined by evaluating RFKs obtained at different depths in kernel-based classifiers like SVM. **ETK:** A kernel that encodes similarities between samples based on the terminal nodes of ET is returned.
xtrain	a data frame or matrix containing training data. In xtrain, rows contain the samples and columns contain the variables (or features).
xtest	a data frame or matrix containing test data. In xtest, rows contain the samples and columns contain the variables (or predictors).
ltr	A response vector, dataframe, or matrix containing the labels of training data.
lts	A response vector, dataframe, or matrix containing the labels of test data.
MtryOp	If FALSE (default), mtry is not optimized. If TRUE, $mtry$ optimization is implemented in case of RF-based kernels (i.e., the ETK is always calculated with the default value of $mtry$). Notation: $mtry$ is the number of variables randomly sampled as candidates at each split. The default value of $mtry$ is sqrt(p), where p is the number of variables in xtrain.
Nt	Number of trees to grow in the tree-based ensemble learners. The default is set to 500 trees, which stabilizes the error in most of the applications.
numsplits	If kernelType is set to "ETK"; numsplits is used as a parameter in ET. Numsplits is the number of random cut-points to consider for each selected feature in splitting a node as defined in the documentation of the ranger package.
Vis	If TRUE, a visualization of the selected tree-based kernel by the user is returned.

Value

A $(ntrain + ntest)$ by $(ntrain + ntest)$ matrix containing the pairwise similarity values among the samples, where ntrain is the number of training samples and ntest is the number of test samples (except if gmatrices is set as kernelType).
Subsetting the output matrix as $[1 : ntrain, 1 : ntrain]$ gives the train kernel containing the similarity values among training sample. Subsetting the output matrix as $[(ntrain + 1) : (ntrain + ntest), 1 : ntrain]$ yields the test kernel containing the similarity values of test samples to train samples.

Availability

This function is freely distributed in the DANS repository. The instructions, the source code, and 10 pairs of training and test subsets obtained in [179] are available at DOI: 10.17026/dans-247-y9x3.
Contact: zafari.azar@gmail.com

Examples

TrKr=TreeBasedKernels (kernelType='RFKNd', xtrain=xtr2, xtest=xts2, ltr=ytr2, lts=yts2, MtryOp=FALSE, Nt=100, numsplits=1, Vis=TRUE)

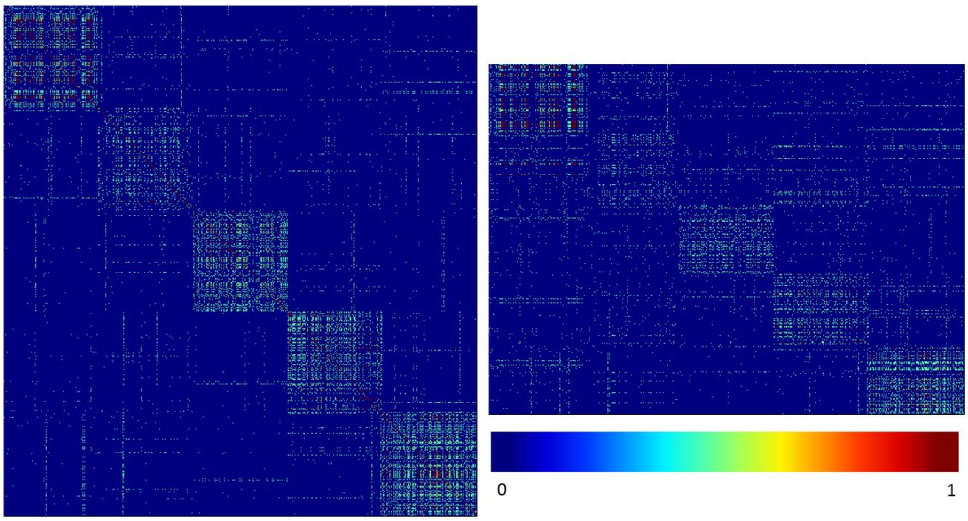


Figure 5.1 The visualization of RFK_{Nd} as train-train kernel in the left and test-train kernel in the right.

TrKr=TreeBasedKernels (kernelType='MultRFKprb', xtrain=xtr2, xtest=xts2, ltr=ytr2, lts=yts2, MtryOp=FALSE, Nt=100, numsplits=1, Vis=TRUE)

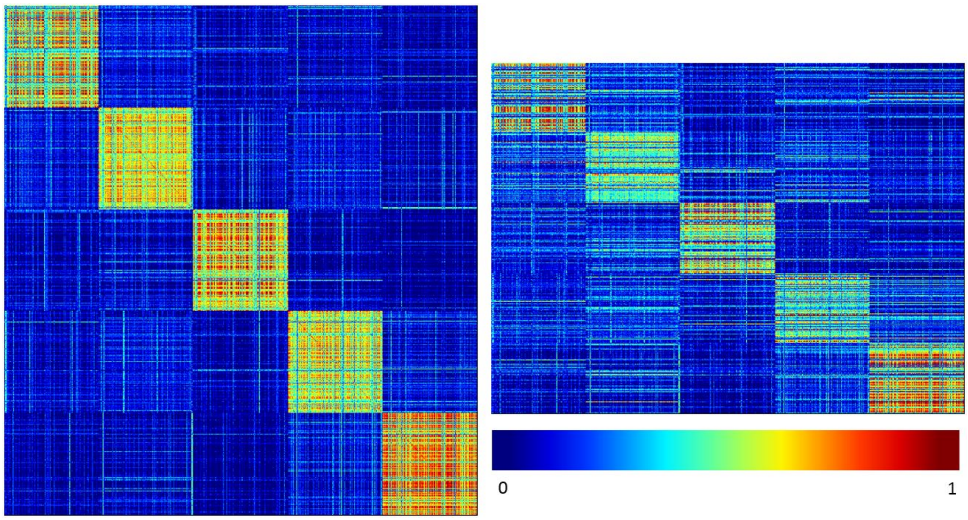


Figure 5.2 The visualization of $RFK_{\overline{Prob}}$ as train-train kernel in the left and test-train kernel in the right.

TrKr=TreeBasedKernels (kernelType='MultRFKprb', xtrain=xtr2, xtest=xts2, ltr=ytr2, lts=yts2, MtryOp=FALSE, Nt=100, numsplits=1, Vis=TRUE)

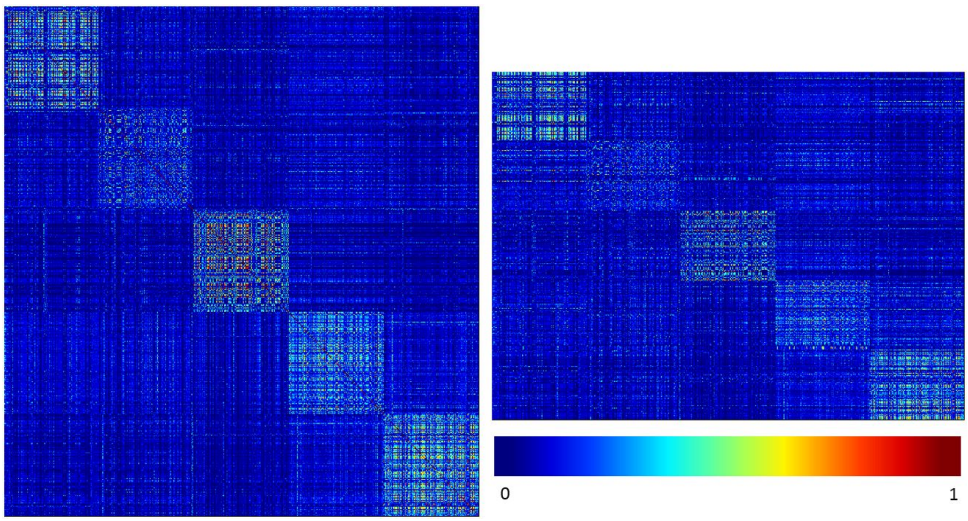


Figure 5.3 The visualization of $RFK_{\overline{Nd}}$ as train-train kernel in the left and test-train kernel in the right.

TrKr=TreeBasedKernels (kernelType='ETK', xtrain=xtr2, xtest=xts2, ltr=ytr2, lts=yts2, MtryOp=FALSE, Nt=100, numsplits=1, Vis=TRUE)

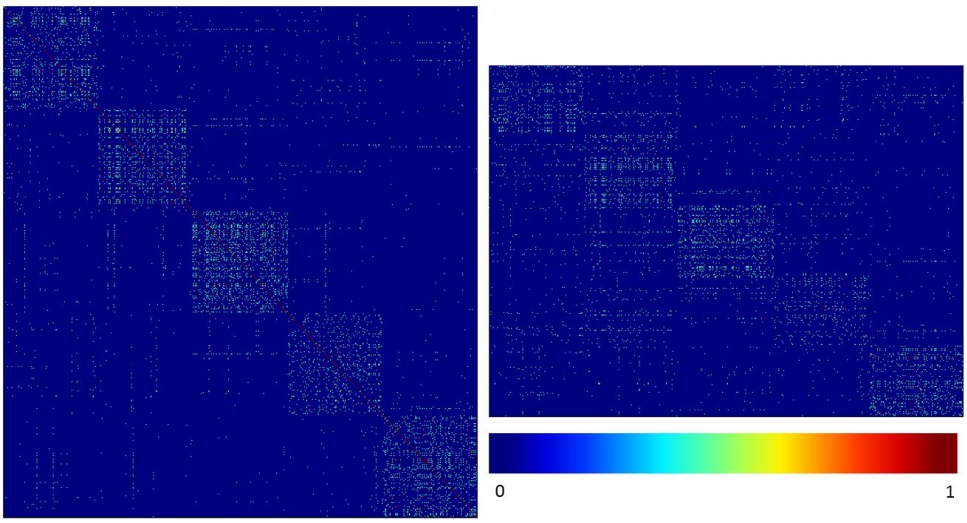


Figure 5.4 The visualization of ETK as train-train kernel in the left and test-train kernel in the right.

Synthesis

6

6.1 Research findings and conclusions

The aim of this dissertation was to integrate two of the most well-known and recurrently used classifiers by the geospatial community: tree-based methods like RF and kernel methods like SVM. In particular, this research concentrated on exploring the use of tree-based kernels in SVMs. Developing classification methods to generate high accuracy land cover maps is highly significant. Obtaining land cover maps is the first step in environmental and agricultural monitoring that leads to sustainable development and agricultural systems. In this chapter, we summarize the research findings and conclusions with respect to each research objective, as described in section 1.5. This chapter also presents a reflection on the link among Chapters 2 to 4, the main contribution of the dissertation, and the direction and recommendations for future research.

- **Evaluating the potential of using an RF-based kernel (RFK) to classify remotely sensed images**
  In this dissertation, we explored the possibility of crop classification over smallholder farms through the synergic use of prevalent classification methods. The SVM and RF are two well-known classifiers used for image classification, but each one has its own strong and weak aspects. We employed a synergic approach for these two methods in order to integrate the advantages of both and minimize their disadvantages. This was achieved by obtaining an RF-based kernel and importing it into an SVM rather than using a conventional RBF kernel. Using RFK, one can avoid the the high computational load of parameterizing the RBF kernel. The performance of the synergic method was benchmarked against conventional methods. We tested the proposed synergic method once for a time series of WV2 images over the Sukumba study area in Africa and once for an extended feature set of this dataset by obtaining vegetation indices and GLCM textures and stacking them on to spectral features.
  We discovered that our proposed synergic method yields slightly higher OAs than RF and it considerably outperforms SVM-RBF for the extended Sukumba dataset. We also exploited RF's characteristic to derive an RFK based on the most important features of the extended Sukumba dataset and this experiment achieves further improvements in the OA of the kernel when used in an SVM compared to RFK obtained with all features. Thus, RFK can inherit the strong points of RF. Using only the spectral features, our proposed synergic method performs at almost the same level as that of the classic SVM and RF in terms of OA. However, the computational cost required for the synergic method is much less than that for SVM-RBF for both spectral cases of Sukumba. We showed this through an experiment using the KSVM function in the kernlab package of R.

We used both the built-in and custom kernels of this package to import the RBF and RFK kernels, respectively, in an SVM. In addition, RF models and RFKs are obtained through the randomForest package of R. Using these packages in R, we found that importing RFK in an SVM reduces the computational time associated with parametrizing the kernel compared to RBF in an SVM. Using the default parameter of $mtry$, SVM-RFK is eight times faster for the higher-dimensional Sukumba dataset.
As benchmarking, we tested the proposed synergic method for the Salinas dataset with its original number of features. We found that the synergic method performs at almost the same level as that of RF and classic SVM for the Salinas dataset. However, using RFK in an SVM performs eight times faster compared to RBF and marginally improves the OA compared to the RF. Further, we investigated the effect of $mtry$ as the most influential parameter of the RF on the performance of RFK in an SVM for both the Sukumba and Salinas datasets. $mtry$ is the number of random subsets of all available features that is used in splitting nodes. We found that optimizing $mtry$ marginally improves the performance of RFK in an SVM compared to using the default value of this parameter. This indicates that RFK can perform at least as well compared to the RBF kernel without optimizing any parameter for RF and subsequently for RFK. Thus, considering the trade-off between the added computational load and gain in OA, optimizing the $mtry$ parameter for obtaining RFK can be skipped. Further, the newly tested RFK was revealed to be of importance, particularly for future RS classification tasks that are aimed at high-dimensional problems. Thus, the following research questions have been answered:

a) How do the classification results of SVM-RFK compare to those obtained by the standard RF and SVM-RBF classifiers?
   The results reveal that the selected supervised methods perform well for the selected classification problem, compete closely, and achieve high OAs for spectral features of the Sukumba and Salinas datasets. For the Sukumba dataset, the OAs are 81.08%, 82.08%, 81.34% for RF, SVM-RBF, and SVM-RFK, respectively. For the Salinas dataset, the OAs are 95.83%, 94.16%, and 94.42,% respectively. The proposed RFK led to competitive results for the datasets with a lower number of features, while reducing the cost of the classification compared to using RBF in an SVM. In the Sukumba dataset, adding VI and GLCM features to spectral features resulted in OAs of 80.82%, 77.96%, and 82.30% for RF, SVM-RBF, and SVM-RFK, respectively. The results of the extended Sukumba dataset indicate that RFKs created using high-dimensional and noisy features

considerably improve the results of the classification for SVM-RFK, with OA approximately 4% better than that obtained with an SVM-RBF classifier while reducing the cost of classification. For a higher number of features, the SVM-RFK results in approximately 1.5% improvement in OA, which is also slightly better than that obtained by the standard RF classifier.

b) How does RF's most important parameters affect the performance of SVM-RFK classifier?
The *mtry* parameter partially influences the classification results of RF, while the default values of other parameters generally stabilize the classification error [130, 145]. Hence, we explored the influence of *mtry* on the results of SVM-RFK classification. Our findings prove that optimizing the *mtry* for RF leads to minor changes in the classification results of SVM-RFK. Thus, with a small trade-off in OA, the cost of the classification can be reduced for SVM-RFK by skipping the *mtry* optimization.

c) How does RF's feature selection impact the classification results of SVM-RFK classifier?
RF can identify the most important features using feature importance scores. For the extended Sukumba dataset, we derived an additional RFK based on the most important features from a subsequent RF model trained only with the most important features. SVM-RFK-MIF improved the results of the classification by approximately 2.5 % and 7% compared to those obtained with an SVM-RFK classifier and SVM-RBF, respectively.

► **Investigating the pros and cons of alternative RFK formulations**
In order to meet this objective, we explored various alternative formulations of RFK by conducting a deeper examination into the structure and different characteristics of RF. In particular, we explored the distance between the end-nodes of trees, the role of trees' depths, and class probabilities assigned to samples by RF in improving the RFK notion. We evaluated the performance of the alternative formulation by importing them into an SVM. The evaluation was done using four levels of dimensionality derived for the Sukumba dataset. These four levels were obtained by extending the original spectral bands with vegetation indices and grey-level co-occurrence textures (i.e., B, BVI, BVITVI, and ALL).
The classic design of RFK originates from rough binary estimations, which is not always compatible with real-world data classes with similar spectral signatures (i.e, crops) [16]. We in-

vestigated the possibility of using the distance between end-nodes by computing the number of branches among the end-nodes to improve the performance of RFK for crop mapping in small-scale farms. We compared the performance of this kernel with the classic design of RFK by importing them in an SVM. We obtained the results for both designs, for different numbers of trees, and for four levels of dimensionality. We found that the performance of both kernels are approximately at the same level, while RFK obtained based on the distance between the end-nodes marginally improves the OA in certain cases. Obtaining RFK on the basis of branches requires significantly higher computational load, which makes the use of the classic design preferable, considering the similar results obtained for crop mapping in small-scale farms.

Another issue of the classic RFK is that the samples with the same class labels may end up in different end-nodes in a fully grown RF and be assigned a similarity value of zero. We addressed this issue by exploring the influence of tree depth on RFK performance. Accordingly, we obtained two multi-scale RFK based on multiple depths. The first one is the average of the classic RFKs obtained over multiple depths. The second one is the average of probabilistic RFKs over multiple depths. The idea of a probabilistic RFK is to assign the probability that two samples belong to the same class as a similarity value between the samples; we used the probabilistic notion of an RFK as an alternative to the binary notion of classic RFK.

We found that multi-scale RFKs and the classic RFK perform at almost the same level for experiments with B features, considering the OAs and standard deviation (SD) values obtained for these methods. The findings of this research indicate that for experiments with BVI, BVITVI, and ALL features, averaging the classic RFKs over multiple depths improves the OA and κ index of the classification compared to classic RFK.

Averaging the probabilistic RFK over multiple depths yielded competitive results only in the case of the experiment with B features. The relatively poorer performance of probabilistic RFK can be explained with higher dependency of this kernel on class labels. The direct dependency of the probabilistic RFK on the class labels can result in an overfitted training kernel that cannot predict for the new samples. In case of the presence of mislabeled samples, kernel values in probabilistic RFK are more strongly affected compared to the classic RFK. Despite the poorer performance of multi-scale probabilistic RFK compared to the multi-scale classic RFK and classic RFK, all RF-based kernels outperformed the RBF results for experiments with BVITVI and ALL features.

Ultimately, we also explored the performance of classic and probabilistic RFKs individually at each single depth. We dis-

covered that finding an optimal depth considerably improves the performance of both classic and probabilistic RFKs compared to the average of the kernels obtained at multiple depths. Moreover, the classic RFK obtained in an optimal depth outperformed the one obtained with fully grown RF in terms of OA and κ index in all cases. The findings of this research show that the classic RFK obtained at an optimal depth outperforms all other designs of RFKs and the RBF kernel. In our experiments, the highest OA is 89.48 %, which is obtained for the experiment with ALL features using the classic RFK at an optimum depth—this is an improvement of 11% with regard to SVM-RBF. We concluded that among all the explored characteristics of RF, the depth of trees plays an important role and the has highest contribution in improving the design of an RFK. In this regard, the following research questions have been answered:

a) How does the use of a branch-based distance compare to the standard similarity metric used to calculate RFK?
   In all experiments, RFKs obtained based on the distance between the nodes performed at almost the same level as that of classic RFK. It is worth mentioning that RFKs obtained based on the distance between the end-nodes marginally improve the OA of the classifications in certain cases, although at a high computational cost. Therefore, the use of the classic design of RFK is preferable for our application of crop mapping in small-scale farms.

b) How does designing a multi-scale RFK based on using multiple depths of RF compare to the standard similarity metric used to calculate RFK?
   Obtaining a multi-scale RFK led to an improvement of approximately 1% in OA compared to classic RFK for higher-dimensional experiments and competitive OAs in case of lower-dimensional experiments. However, finding RFKs at an optimum depth results in an improvement of up to 4% with regard to standard RFK for higher-dimensional experiments. Further, finding an optimum depth for trees prevents over-partitioning of the nodes; this avoids samples with same-class labels ending up in different end-nodes and being assigned a similarity value of zero, as tends to happen in a fully grown RF. This implies that there is a reduced likelihood of obtaining an overfitted training kernel.

c) How does designing a multi-scale RFK based on using multiple depths and class probabilities compare to the standard similarity metric used to calculate RFK?
   Including the class probabilities in RFK design did not

improve the performance of the kernel compared to the classic RFK in any of the experiments. Yet, for high-dimensional noisy problems, the probabilistic RFK outperformed the RBF kernel. We explained the poorer performance of probabilistic RFKs by the higher dependency of this kernel on the class labels of training samples. In our notion of probabilistic RFK, the similarity values are the probabilities that two samples fall in the same class. This can increase the likelihood of obtaining an overfitted training kernel to the class labels of the training samples. A model trained with an overfitted training kernel cannot generalize for the test kernel and this causes a poorer performance. Moreover, the possible presence of mislabeled samples misleads the similarity values to a greater extent in the probabilistic RFK as compared to the classic RFK.

- **Exploring the use of an alternative tree-based classifier, namely ET, to derive tree-based kernels**
  In order to achieve this objective, we investigated the performance of the synergic use of ET and SVMs through a kernel connection for crop classification of small-scale farms. To do so, we defined a kernel based on the tree-based structure of ET and imported the ET-based kernel (ETK) into an SVM. We evaluated the use of ETK in an SVM by comparing its performance against that of standard ET, use of the RBF kernel in an SVM, and use of RFK in an SVM. We tested the performance of the classifiers in low- and high-dimensional problems by obtaining four levels of dimensionality from these images by extending their original spectral bands with vegetation indices and grey-level co-occurrence textures (i.e., B, BVI, BVITVI, and ALL). This enabled testing the performance of the classifiers in low- and high-dimensional problems.
  The computational cost and dependency of the trees structure on the class labels for ET and ETK are smaller than those for RF and RFK. Using ETK as an alternative to RFK and RBF, we were able to reduce the probability of obtaining an overfitted kernel and avoid the computational cost associated with parametrizing the RBF kernel.
  First, we compared the added value of using an ETK in an SVM compared to the classic ET under different configurations of ET. We explored the influence of various numbers of random cut-points per candidate feature and different number of trees on the performance of ET and SVM-ETK; we set $mtry$ to its default value. We found that SVM-ETK always outperforms ET for all the cases with VIs (BVI, BVITVI, ALL), irrespective of the value for the number of trees and number of random cut-points per candidate feature. For the experiment with B features, the OA of SVM-ETK outperforms ET in most ranges, particularly for high

levels of randomization.
Second, we compared the performances of the ETK, RFK, and RBF kernels within an SVM. The performance of ETK was obtained once with default values of parameters and then with best performance over the tested ranges of parameters in the first step. For the experiment with the B features, we found that both SVM-ETKs classifiers slightly outperform SVM-RFK and SVM-RBF. For experiments with BVI features, we found that ETKs and RBF kernels perform at approximately the same level and outperform SVM-RFK. More interestingly, we discovered a gradual improvement in the OAs of tree-based kernels by increasing the number of features from BVI to BVITVI and then to ALL, while we found a sudden decrease of 6.62 % for SVM-RBF. Our results for the experiments with BVITVI and ALL features reveal that tree-based kernels perform at almost the same level and considerably outperform the RBF kernel. The highest OA is 85.70 % and was obtained for SVM-ETK trained with all features.
Third, we used totally randomized trees (ToRT) (i.e., ET with the most extreme level of randomization) to obtain a kernel (ToRTK) and evaluated the performance of ToRTK through the four experiments. These analyses revealed that SVM-ToRTK outperforms ToRT in all experiments. We found that for the experiment with B features, SVM-ToRTK yields similar and marginally higher OAs compared to SVM-ETKs and improves the results compared to classic ET, SVM-RFK, and SVM-RBF. Moreover, for the experiment with BVI features, SVM-ToRTK yields competitive results compared to the other classifiers. It is important to note that for higher-dimensional experiments, the performance of ToRTK decreases substantially. Therefore, using the most extreme level of randomization is only preferable in experiments with low dimensionality, since it reduces the cost of classification compared to ETK, RFK, and RBF kernels and yields marginally higher OAs.

Ultimately, we applied the trained classifiers to all the available ground truth labels (with B features) in the study area and obtained OA and κ index for each method. The results for the entire study area indicate that all classifiers perform well and at approximately the same level, while SVM-ETK slightly outperforms in terms of the OA and κ index. Moreover, we found that SVM-ETK significantly improves the OA of the fields with class Sorghum.

In a nutshell, this study concludes that ETKs are efficient and effective alternatives to most well-known kernels used by the RS community. In this regard, the following research questions have been answered:

a) What is the influence of ET's most important parameters

on the classification accuracy of the corresponding SVM-ETK classifier?
We explored the influence of number of trees and number of random cut-points per candidate feature on the performance of ETs and SVM-ETK. Irrespective of these parameters, SVM-ETK outperformed ETs in a majority of the cases. A higher number of trees (up to 500 trees were tested) led to higher OAs for both methods. The number of random cut-points also influenced the performance of the classifiers in terms of OA. This parameter controls the level of randomization and optimizing it to the problem at hand helps to improve the OAs of the classification. However, we found that the default value of this parameter yields competitive results for both ET and SVM-ETK classifiers compared to the optimized cases.

b) How does the level of randomization influence the performance of the ET and SVM-ETK classifiers?
We found that the level of randomization must be adjusted to the problem at hand. Increasing the level of randomization from RF to ET results in an improved OA for SVM-ETK compared to SVM-RFK in most of the cases and competitive results in other cases. The outperformance of ETK was found at both optimized and default levels of randomization in ET. Further increasing the level of randomization to its most extreme case resulted in competitive and in certain cases, higher classification OAs for experiments with lower dimensionality. Yet, for high-dimensional experiments, kernels obtained with the most extreme level of randomization performed significantly poorly.

c) How do the classification results of SVM-ETK compare with those obtained by the standard ET, SVM-RBF, and SVM-RFK classifiers? The results of this study indicate that SVM-ETK outperforms other tested methods in terms of OA for land cover mapping over small-scale agricultural lands. Using ETK in an SVM also reduces the computational cost of the classification compared to use of RFK and RBF kernels in an SVM.

6.2 Reflections

The research work in this dissertation revolves around investigating the integration of tree- and kernel-based classifiers by obtaining and importing tree-based kernels into an SVM and testing its use for land cover classification in an agricultural context. This was accomplished in the following three steps:

- Obtaining an RFK based on the classification results of RF and importing it into an SVM. The complexity of land cover classification using the recent generation of very high resolution satellite images with enhanced spectral and temporal resolutions has directed the need to develop more efficient classifiers that can address the curse of dimensionality associated with these types of data. To this end, SVM [29] and RF [95], as well-known kernel-based and ensemble classifiers, have been applied in several high-dimensional classification problems. Inspired by these works and concerned with the connection between RF and kernel-based methods introduced in [94, 104], Chapter 2 developed an integrated approach of RF and SVM classifiers through kernel connection. Our approach is based on obtaining an RFK and importing it into an SVM. Using the Sukumba and Salinas datasets, Chapter 2 explored the performance of RFKs obtained from fully grown RF and examined the influence of $mtry$. Chapter 3 explored the influence of the depth of trees as the most important parameters of RF on the performance of RFK in an SVM. Our approach enabled avoidance of the cost associated with parametrizing the RBF kernel and led to competitive or better results in terms of OA. The HSIC values indicated that RFK is closer to an ideal kernel compared to RBF. However, RFK does not yield the best results in terms of OA in all cases. This is evidence that RFK might be overfitted to training class labels and cannot generalize well for test data. Another reason is that the samples from the same class can land in different end-nodes of a fully grown RF and be assigned a similarity value of zero. This motivated us to use depth of trees in Chapter 3, which prevents oversplitting the nodes, and ET in Chapter 4, which decreases the dependence of the kernels on the class labels of training samples by increasing the level of randomization of trees. Chapter 2 also combined RFK with RF-based feature selection; this resulted in further improvements in the OAs of the classification. The promising results in this step proved that RFK is a good alternative to the RBF kernel and this led us to study the different characteristics of RF in order to define improved and advanced notions of RFK in the next steps.
  In the classic notion of RFK, for each pair of samples, binary similarity values are obtained based on whether or not two samples fall into the same terminal nodes of a tree. This rough binary estimation of similarities is not always compatible with real world problems [167]. This problem was addressed in Chapter 2 by investigating the use of probability and the distance between the nodes in the notion of RFKs. The same Sukumba dataset was used in Chapters 2 to 4, while the dimensionality of experiments in Chapter 2 is different. The first experiment with B features is the same in all chapters, and ex-

periments with BVI, $BVITVI$, and ALL features are the same in Chapters 3 and 4.

- Providing solutions for handling the binary design of classic RFK and the problem of RFKs obtained with fully grown RF
  Chapter 3 addressed the overcoming of the shortcomings of the classic notion of RFK in Chapter 2 through the full exploration of RF's characteristics. Improving the binary estimations of RFK is determined by obtaining an RFK based on the number of tree branches between the end-nodes. This kernel marginally improves the OAs of the classifications in several cases. We investigated the overcoming of the problem of oversplitted nodes in trees by proposing two designs of multi-scale RFK based on multiple depths of RF. The first design of multi-scale RFK is the average of the classic RFKs obtained at multiple depths. The second design of multi-scale RFK is the average of probabilities that samples fall into the same class at multiple depths. The idea of the probabilistic multi-scale RFK is overcoming both drawbacks of binary estimations and oversplitting the nodes that were mentioned earlier for the classic notion of RFK. The first design of multi-scale RFK improves the OAs of the classifications almost in all cases compared to the classic design. The probabilistic multi-scale RFK yields competitive results compared to classic RFK and multi-scale RFK only in case of experiments with the lowest dimensionality. As we described earlier in the notion of a probabilistic RFK, the similarity values are the probabilities that two samples fall in the same class. This can also result in a training kernel that is overfitted to the class labels of the training samples and cannot generalize well for the test kernel. We verified this inference by obtaining HSIC values among training samples (i.e., $HSIC_{TrTr}$) and among test samples (i.e., $HSIC_{TsTs}$). Table 6.1 shows that by increasing the number of features, the HSIC values of multi-scale probabilistic RFK ($RFK_{\overline{Prob}}$) increase for training samples and the training kernel comes closer to an ideal kernel, while this does not happen for the test kernel.

  However, $RFK_{\overline{Prob}}$ outperforms the RBF kernel for high-dimensional problems. Another aspect that was covered in Chapter 3 is identifying an optimal depth that considerably improves the performance of the kernel compared to averaging over multiple depths. Using this approach, RFK avoids oversplitting of the nodes that the classic RFK in Chapter 2 was unable to overcome. As expected, the classic RFK obtained in an optimal depth improves OA and κ the index of the classifications compared to the classic RFK obtained with fully grown RF for all the cases. Among all proposed design of RFK in Chapters 2 and 3, the highest level of improvement in OAs

Table 6.1 HSIC values obtained for training samples and test samples

Tests	Methods	$HSIC_{TrTr}$	$HSIC_{TsTs}$
B	RFK_{Nd}	0.016	0.008
	$RFK_{\overline{Prob}}$	0.074	0.083
BVI	RFK_{Nd}	0.025	0.014
	$RFK_{\overline{Prob}}$	0.069	0.014
BVITVI	RFK_{Nd}	0.028	0.015
	$RFK_{\overline{Prob}}$	0.067	0.015
All	RFK_{Nd}	0.028	0.014
	$RFK_{\overline{Prob}}$	0.066	0.014

of the classifications was achieved with a classic RFK obtained at an optimal depth.

It is worth noting from the outcome of Chapters 2 and 3 that higher dependency of kernel values can increase the likelihood of obtaining a training kernel that is overfitted to the class labels of the training samples. This was observed for multi-scale probabilistic RFK compared to classic RFK and for classic RFK compared to RBF in certain cases. Moreover, the possible presence of mislabeled samples disturbs the kernel values with higher dependency on class labels to a greater extent. Chapter 3 addressed this issue by optimizing the depth. However, optimizing the depth led to higher computational load for RFK. The results of these analyses indicated that there are two aspects that can be further improved: The first aspect is the use of a tree-based kernel with a reduced level of dependency on the class labels. The second aspect concerns reducing the computational load of the kernel. These two aspects were achieved in Chapter 4 by increasing the level of randomization in trees.

- Assessing the influence of reducing the dependency of tree-based kernel values on the class labels of training samples.
  Chapter 4 improved the notion of tree-based kernels by using ET, which enables increasing the level of randomization compared to RF. Futher, ETs result in less correlated trees and this reduces the likelihood of obtaining an overfitted forest. Obtaining ETK improved the OAs of the classifications compared to classic RFK and RBF kernels (Chapter 2) in the problems with lower dimensionality. For high-dimensional problems, classic RFK (Chapter 2) and ETK performed at almost the same level; moreover, they significantly improved the classification metrics compared to the RBF kernel. In comparison with ETK, the classic RFK at an optimized depth in Chapter 3 performed equally

well for lower-dimensional experiments and outperformed for higher-dimensional experiments. However, considerable higher computational load is required to obtain RFK at an optimized depth compared to ETK. ETK also reduces the computational cost of the classification compared to classic RFK and RBF for all the cases. It is worth noting that among all the tree-based kernels introduced and tested in Chapters 2 to 4, the highest OA is obtained for the classic RFK at an optimized depth in case of higher-dimensional experiments at an expense of higher computational load.

These three steps contributed to improving the accuracy of crop classification and reducing the computational cost of the classification in comparison with well-known classifiers used in the RS community. The proposed methods are of value, in particular, for high-dimensional problems that are possibly noisy, which is the case in the present research work. Land cover mapping over small-scale farms using VHR satellite images is associated with multiple challenges. The high accuracy crop maps produced with the proposed methods are of value in handling the problem of food security, which is the main concern of governments and policymakers in the countries with vast areas under small-scale farming. Agriculture in low-income countries is performed by smallholder farmers to a large extent—this is often over 50% of a country's population—who are at the bottom of the economic pyramid and often struggle to make ends meet. Thus, high-quality maps are valuable in improving sustainable agricultural production by minimizing the economic and environmental costs in these areas. One of the important applications of crop maps is in phenological studies. Phenology is the study of timing of plant' growth stages and how these are influenced by seasonal and interannual variations in climate and habitat factors [180]. Phenological observations provide basic information for numerous purposes in practical agriculture for farmers [181]. The data can be used to define the duration of the growing season in a region. "The growing season is the time of the year in which plants germ, grow, flower, fructify, and ripen" [181]. Phenological information helps decision-making for farmers to practice timely appropriate operations including planting, fertilizing, irrigating, crop protection, to predict phenophases, and to select favorable and unfavorable areas for agricultural production [181]. Phenological information can be acquired both through ground-based in situ observations and images from satellite sensors [182]. Satellite-based phenology is often called land surface phenology (LSP), as the satellite sensor signals are an integration of the reflectance from the plants and land surface [183]. LSP has been widely used to assess large-scale phenological patterns [184]. However, LSP observations often do not provide a precise match for ground-observed phenology. Therefore, it is necessary to produce validated phenology maps by linking the timing

of key phenophases obtained from VHR satellite images and ground truth data. The highly accurate crop maps obtained as an output of this study are of value as inputs to produce higher accuracy phenology maps.

6.3 Recommendations

This dissertation investigated the integration of tree-based kernels with SVM classifiers for land cover extraction in small-scale agriculture. In line with the research performed as part of this dissertation, there are several future avenues and interesting experiments that are recommended in the remainder of this section.

6.3.1 Outlier-free RFKs

In order to define an improved RFK design, other advantages of RF can be exploited. RF is a proximity-based outlier detection method. The idea in proximity-based methods is to model outliers as points that are isolated from the remaining data on the basis of similarity among samples [185, 107]. The similarity values obtained with RF can be used to identify outliers [107]. A new RF model can be trained with outlier-free samples and RFK can be obtained using the outlier-free samples. In order to identify the outliers, RF assigns an outlier index to all samples based on proximity. By defining a threshold, a sample with different outlier indices can be found and eliminated. For a class label, each sample is given a value for its "outlyingness," which is computed in the following manner [107]:

$$OutlierIndex = [\frac{n}{\sum proximity^2} - median]/MAD \tag{6.1}$$

where MAD is the median absolute deviation within each class. For each class label, Equation computes the sum of the squares of the similarity values with all the other observations in the same class; thereafter, it takes the inverse. It does the same for all other observations in that class. One can think of these values as unstandardized. Next, by subtracting the median and dividing by the mean absolute deviation, standardized values for outlier indices are obtained [185]. Samples with an outlier index of larger than 10 can be considered as outliers. However, it can be instructive to identify an optimal threshold value in an iterative approach. A primary result of combining RF's outlier detection method and RFK is presented in Table 6.2. The default threshold of 10 is considered to isolate the outliers in these results.

Table 6.2 Classification results obtained through the synergic use of RFK and RF's outlier detection method for different subsets of features introduced in Chapter 2. RFK_{Nd} shows classic RFK obtained based on the end nodes. Moreover, the depth that results in the best OA for RFK_{Nd} is shown with RFK_{Nd^*}.

Tests	Methods	$\overline{OA}$	SD	$\bar{\kappa}$	SD_{κ}
B	SVM-RFK_{Nd}	83.92	1 .59	0.80	0.02
	SVM-RFK_{Nd^*}	84.76	1.39	0.81	0.02
BVI	SVM-RFK_{Nd}	84.54	1.38	0.81	0.02
	SVM-RFK_{Nd^*}	85.54	1.32	0.82	0.02
BVITVI	SVM-RFK$_{Nd}$	86.44	1.8	0.83	0.02
	SVM-RFK_{Nd^*}	89.6	1.58	0.87	0.02
All	SVM-RFK_{Nd}	88.08	1.58	0.85	0.02
	SVM-RFK_{Nd^*}	90.84	1.31	0.89	0.02

Comparing Table 6.2 with the results obtained in Chapter 2 reveals that the OAs of RFKs (both obtained with a fully grown RF and at an optimal depth) considerably improve when applied to oultlier-free samples for all the experiments. Outlier-free RFK also outperforms RFK at an optimized depth and ETK for all the experiments. These results show that further research can be conducted to create outlier-free RFKs by finding an optimal threshold for eliminating outliers.

6.3.2 Improving the performance of the newly tested tree-based kernels

The newly tested tree-based kernels presented in this research can be improved with respect to several aspects. In order to deal with the curse of dimensionality, which is the case in several RS problems, feature selection methods are inevitable. In Chapter 2, the feature selection method of RF is used to rank features. We used an arbitrary number of 100 to select top-ranked features by RF. RFKs from new RF models trained with 100 most important features are obtained for the high-dimensional experiment. However, 100 features may not be an optimal number and further research could be conducted to identify an optimal number of features to be selected in high-dimensional datasets. This can be achieved with a guided regularized RF [186].
In Chapter 3, HSIC values revealed that the multi-scale probabilistic RFK resulted in a training kernel that is overfitted to the class labels of the training samples and cannot be generalized for the test kernel. Further research could be conducted to improve the design of the probabilistic tree-based kernel by reducing the dependence level of kernel values on the class labels and by making sense of RF's probabilities [187].

In Chapter 4, the influence of optimizing the depth of forest on the performance of ETK in an SVM is not investigated. Moreover, the forest characteristics including outlier detection, feature selection, and use of probabilities can be investigated and integrated with the notion of ETK. Using the synergies of these characteristics can also be a direction for future research.

In the present research, we used labeled samples for learning purposes; however, unlabeled data can also unveil valuable patterns for classification tasks. Semi-supervised approaches aim at learning from both labeled and unlabeled samples and have been shown to outperform supervised approaches in certain cases, particularly in problems with labeled data scarcity [188, 189]. The newly tested tree-based kernels can be further exploited in the framework of semi-supervised approaches, like the Laplacian Support Vector Machine, to exploit both labeled and unlabeled data [190]. Moreover, there is a rather limited selection of a single kernel that can fit complex data structures [191]. Several studies have showed that selecting inappropriate kernels leads to suboptimal or poor performances [192]. The performance of kernel-based classification methods depends on the choice of the kernel function and its parameters. In order to address this problem, the multiple kernel learning (MKL) approach, which combines a set of base kernels into a composite kernel, is employed in several studies [192, 191]. The basis of kernels can be defined by using different kernel functions or different values for the hyperparameters of a single kernel function [192, 191]. MKL algorithms can effectively be applied in the context of feature fusion by obtaining basis kernels for different subsets of features, such as spectral, textural, and multisource features [192, 191]. The performance of tree-based kernels in an MKL framework can also be further explored in future research.

6.3.3 Applications of the newly tested tree-based kernels

There are several applications that can benefit from the new classifiers. We efficiently applied the tree-based kernels for a complex crop classification problem using a time series of a single data source. Considering their efficient performance in high-dimensional noisy problems, tree-based kernels are recommended in the classification tasks of multisource data and for various land cover types. Combining feature selection and outlier detection properties of RF with tree-based kernels can also play an important role in further improving the performance of the kernel-based methods for multisource datasets. This is likely to result in high-quality land cover maps that are input to several agricultural, environmental, and urban management systems.
One good example is applying the high-quality crop maps obtained

in the present research work as an input in producing phenological maps that reveal a plant's growth stages [180]. As an example, phenological information helps farmers to practice timely appropriate planting, fertilizing, irrigating, and crop protection [182].
In a two-sided connection, such phenology maps can also be further exploited to improve the quality of crop maps. However, phenological information obtained trough satellite images often does not precisely match ground-based in situ observations [182]. Therefore, the performance of newly tested and newly developed tree-based kernels can also be examined in a kernel-based regression model, like support vector regression, to relate the timing of key phenophases obtained from satellite images and ground truth data.

Bibliography

[1] J. A. Foley, N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. O'Connell, D. K. Ray, P. C. West, *et al.*, "Solutions for a cultivated planet," *Nature*, vol. 478, no. 7369, p. 337, 2011.

[2] S. S. Paul, *Analysis of land use and land cover change in Kiskatinaw River watershed: A remote sensing, GIS modeling approach*. PhD thesis, 01 2013.

[3] D. P. C. Peters, J. R. Gosz, W. T. Pockman, E. E. Small, R. R. Parmenter, S. L. Collins, and E. Muldavin, "Integrating patch and boundary dynamics to understand and predict biotic transitions at multiple scales," *Landscape Ecology*, vol. 21, pp. 19–33, Jan 2006.

[4] S. Anwar, *Spatial point process modelling of land use and land cover (LULC) change*. PhD thesis, University of Twente, 11 2014. ITC Dissertation; 257.

[5] A. Bégué, D. Arvor, C. Lelong, E. Vintrou, and M. Simoes, "Agricultural systems studies using remote sensing," *Remote sensing Handbook*, vol. 2, pp. 113–130, 2015.

[6] A. Ali, "Hyper-temporal remote sensing for land cover mapping and monitoring," 2014.

[7] E. H. Helmer, S. Brown, and W. Cohen, "Mapping montane tropical forest successional stage and land use with multidate landsat imagery," *International Journal of Remote Sensing*, vol. 21, no. 11, pp. 2163–2183, 2000.

[8] T. Lillesand, R. W. Kiefer, and J. Chipman, *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

[9] C. Josef, "Land cover mapping of large areas from satellites: Status and research priorities," *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1093–1114, 2000.

[10] S. E. Franklin, *Remote sensing for sustainable forest management*. CRC press, 2001.

[11] C. Boshuizen, J. Mason, P. Klupar, and S. Spanhake, "Results from the planet labs flock constellation," 2014.

[12] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, pp. 778–782, May 2017.

[13] "Challenges and opportunities in mapping land use intensity globally," *Current Opinion in Environmental Sustainability*, vol. 5, no. 5, pp. 484–493, 2013.

[14] P. N. Rao, M. S. Sai, K. Sreenivas, M. K. Rao, B. Rao, R. Dwivedi, and L. Venkataratnam, "Textural analysis of irs-1d panchromatic data for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 17, pp. 3327–3345, 2002.

[15] H. Carrão, P. Gonçalves, and M. Caetano, "Contribution of multispectral and multitemporal information from modis images to land cover classification," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 986–997, 2008.

[16] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by svm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, 2010.

[17] M. C. Dobson, F. T. Ulaby, and L. E. Pierce, "Land-cover classification and estimation of terrain attributes using synthetic aperture radar," *Remote Sensing of Environment*, vol. 51, no. 1, pp. 199–214, 1995.

[18] R. Zurita-Milla, J. G. P. W. Clevers, J. A. E. V. Gijsel, and M. E. Schaepman, "Using meris fused images for land-cover mapping and vegetation status assessment in heterogeneous landscapes," *International Journal of Remote Sensing*, vol. 32, no. 4, pp. 973–991, 2011.

[19] C. Homer, J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, J. Wickham, *et al.*, "Completion of the 2001 national land cover database for the counterminous united states," *Photogrammetric Engineering and Remote Sensing*, vol. 73, no. 4, p. 337, 2007.

[20] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[21] J. R. Jensen, *Remote sensing of the environment: An earth resource perspective 2/e*. Pearson Education India, 2009.

[22] 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," *European Journal of Remote Sensing*, vol. 47, no. 1, pp. 389–411, 2014.

[23] U. Nations, *World Population Prospects*. 2012.

[24] P. Conforti, *Looking Ahead in World Food and Agriculture: Perspectives to 2050*. Food and Agriculture Organization of the United Nations, 2011.

[25] I. Mariotto, P. S. Thenkabail, A. Huete, E. T. Slonecker, and A. Platonov, "Hyperspectral versus multispectral crop-productivity modeling and type discrimination for the hyspiri mission," *Remote Sensing of Environment*, vol. 139, pp. 291–305, 2013.

[26] C. Gómez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55–72, 2016.

[27] J. A. Long, R. L. Lawrence, M. C. Greenwood, L. Marshall, and P. R. Miller, "Object-oriented crop classification using multitemporal etm+ slc-off imagery and random forest," *GIScience & Remote Sensing*, vol. 50, no. 4, pp. 418–436, 2013.

[28] M. Belgiu and O. Csillik, "Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis," *Remote Sensing of Environment*, vol. 204, pp. 509–523, 2018.

[29] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[30] G. Yan, J. Mas, B. Maathuis, Z. Xiangmin, and P. van Dijk, "Comparison of pixel-based and object-oriented image classification approaches: A case study in a coal fire area, wuda, inner mongolia, china," *International Journal of Remote Sensing*, vol. 27, no. 18, pp. 4039–4055, 2006.

[31] A. Carleer, O. Debeir, and E. Wolff, "Assessment of very high spatial resolution satellite image segmentations," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 11, pp. 1285–1294, 2005.

[32] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3, pp. 239–258, 2004.

[33] G. J. Hay, T. Blaschke, D. J. Marceau, and A. Bouchard, "A comparison of three image-object methods for the multiscale analysis of landscape structure," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 57, no. 5, pp. 327–345, 2003.

[34] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[35] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sensing of Environment*, vol. 177, pp. 89–100, 2016.

[36] G. P. Asner, E. N. Broadbent, P. J. C. Oliveira, M. Keller, D. E. Knapp, and J. N. M. Silva, "Condition and fate of logged forests in the brazilian amazon," *Proceedings of the National Academy of Sciences*, vol. 103, no. 34, pp. 12947–12950, 2006.

[37] J. R. Townshend, J. G. Masek, C. Huang, E. F. Vermote, F. Gao, S. Channan, J. O. Sexton, M. Feng, R. Narasimhan, D. Kim, K. Song, D. Song, X.-P. Song, P. Noojipady, B. Tan, M. C. Hansen, M. Li, and R. E. Wolfe, "Global characterization and monitoring of forest cover using landsat data: Opportunities and challenges," *International Journal of Digital Earth*, vol. 5, no. 5, pp. 373–397, 2012.

[38] R. Birdsey, G. Angeles-Perez, W. A. Kurz, A. Lister, M. Olguin, Y. Pan, C. Wayson, B. Wilson, and K. Johnson, "Approaches to monitoring changes in carbon stocks for redd+," *Carbon Management*, vol. 4, no. 5, pp. 519–537, 2013.

[39] R. S. DeFries, R. A. Houghton, M. C. Hansen, C. B. Field, D. Skole, and J. Townshend, "Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 1990s," *Proceedings of the National Academy of Sciences*, vol. 99, no. 22, pp. 14256–14261, 2002.

[40] K. M. Keegan, M. R. Albert, J. R. McConnell, and I. Baker, "Climate change and forest fires synergistically drive widespread melt events of the greenland ice sheet," *Proceedings of the National Academy of Sciences*, vol. 111, no. 22, pp. 7964–7967, 2014.

[41] Y. Knyazikhin, M. A. Schull, P. Stenberg, M. Mõttus, M. Rautiainen, Y. Yang, A. Marshak, P. L. Carmona, R. K. Kaufmann, P. Lewis, *et al.*, "Hyperspectral remote sensing of foliar nitrogen content," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 185–192, 2013.

[42] C. D. Mendenhall, C. H. Sekercioglu, F. O. Brenes, P. R. Ehrlich, and G. C. Daily, "Predictive model for sustaining biodiversity in tropical countryside," *Proceedings of the National Academy of Sciences*, vol. 108, no. 39, pp. 16313–16316, 2011.

[43] G. P. Asner, S. R. Levick, T. Kennedy-Bowdoin, D. E. Knapp, R. Emerson, J. Jacobson, M. S. Colgan, and R. E. Martin, "Large-scale impacts of herbivores on the structural diversity of african savannas," *Proceedings of the National Academy of Sciences*, vol. 106, no. 12, pp. 4947–4952, 2009.

[44] M. Haq, M. Akhtar, S. Muhammad, S. Paras, and J. Rahmatullah, "Techniques of remote sensing and gis for flood monitoring and damage assessment: A case study of sindh province, pakistan," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 15, no. 2, pp. 135–141, 2012.

[45] F. Yamazaki, "Applications of remote sensing and gis for damage assessment," *Structural Safety and Reliability*, pp. 1–12, 2001.

[46] K. Kaku, "Satellite remote sensing for disaster management support: A holistic and staged approach based on case studies in sentinel asia," *International Journal of Disaster Risk Reduction*, vol. 33, pp. 417–432, 2019.

[47] C. Alcantara, T. Kuemmerle, A. V. Prishchepov, and V. C. Radeloff, "Mapping abandoned agriculture with multi-temporal modis satellite data," *Remote Sensing of Environment*, vol. 124, pp. 334–347, 2012.

[48] M. C. Anderson, R. G. Allen, A. Morse, and W. P. Kustas, "Use of landsat thermal imagery in monitoring evapotranspiration and managing water resources," *Remote Sensing of Environment*, vol. 122, pp. 50–65, 2012. Landsat Legacy Special Issue.

[49] B. Hong, K. E. Limburg, M. H. Hall, G. Mountrakis, P. M. Groffman, K. Hyde, L. Luo, V. R. Kelly, and S. J. Myers, "An integrated monitoring/modeling framework for assessing human–nature interactions in urbanizing watersheds: Wappinger and onondaga creek watersheds, new york, usa," *Environmental Modelling Software*, vol. 32, pp. 1–15, 2012.

[50] D. Arvor, M. Jonathan, M. S. P. Meirelles, V. Dubreuil, and L. Durieux, "Classification of modis evi time series for crop mapping in the state of mato grosso, brazil," *International Journal of Remote Sensing*, vol. 32, no. 22, pp. 7847–7871, 2011.

[51] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. de Queiroz, "A time-weighted dynamic time warping method for land-use and land-cover mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, pp. 3729–3739, Aug 2016.

[52] C. Senf, P. J. Leitão, D. Pflugmacher, S. van der Linden, and P. Hostert, "Mapping land cover in complex mediterranean

landscapes using landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery," *Remote Sensing of Environment*, vol. 156, pp. 527–536, 2015.

[53] Q. Yu, P. Gong, N. Clinton, G. Biging, M. Kelly, and D. Schirokauer, "Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 7, pp. 799–811, 2006.

[54] K. Johansen, N. C. Coops, S. E. Gergel, and Y. Stange, "Application of high spatial resolution satellite imagery for riparian and forest ecosystem classification," *Remote Sensing of Environment*, vol. 110, no. 1, pp. 29–44, 2007.

[55] B. Carolyn and T. Blaschke, "A multi-scale segmentation/object relationship modelling methodology for landscape analysis," *Ecological Modelling*, vol. 168, no. 3, pp. 233–249, 2003.

[56] I. L. Castillejo-González, F. López-Granados, A. García-Ferrer, J. M. Peña-Barragán, M. Jurado-Expósito, M. S. de la Orden, and M. González-Audicana, "Object- and pixel-based analysis for mapping crops and their agro-environmental associated measures using quickbird imagery," *Computers and Electronics in Agriculture*, vol. 68, no. 2, pp. 207–215, 2009.

[57] D. Liu and F. Xia, "Assessing object-based classification: advantages and limitations," *Remote Sensing Letters*, vol. 1, no. 4, pp. 187–194, 2010.

[58] S. Bhaskaran, S. Paramananda, and M. Ramnarayan, "Per-pixel and object-oriented classification methods for mapping urban features using ikonos satellite data," *Applied Geography*, vol. 30, no. 4, pp. 650–665, 2010.

[59] G. Camps-Valls, "Machine learning in remote sensing data processing," in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, 2009.

[60] M. Song, D. L. Civco, and J. D. Hurd, "A competitive pixel-object approach for land cover classification," *International Journal of Remote Sensing*, vol. 26, no. 22, pp. 4981–4997, 2005.

[61] J. M. Peña-Barragán, M. K. Ngugi, R. E. Plant, and J. Six, "Object-based crop identification using multiple vegetation indices, textural features and crop phenology," *Remote Sensing of Environment*, vol. 115, no. 6, pp. 1301–1316, 2011.

[62] G. P. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[63] E. Izquierdo-Verdiguier, L. Gómez-Chova, L. Bruzzone, and G. Camps-Valls, "Semisupervised kernel feature extraction for remote sensing image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 9, pp. 5567–5578, 2014.

[64] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Berlin, Heidelberg: Springer-Verlag, 2nd ed., 1994.

[65] K. I. Laws, *Textured Image Segmentation*. PhD Thesis-Microfilm, University of Southern California, 1980.

[66] M. Tuceryan and A. K. Jain, "*Handbook of Pattern Recognition; Computer Vision*," ch. Texture Analysis, pp. 235–276, River Edge, NJ, USA: World Scientific Publishing Co., Inc., 1993.

[67] M. T. S. Taji and D. V. Gore, "Overview of texture image segmentation techniques," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 12, 2013.

[68] M. De Martinao, F. Causa, and S. B. Serpico, "Classification of optical high resolution images in urban environment using spectral and textural information," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, vol. 1, pp. 467–469 vol.1, July 2003.

[69] X. Huang, X. Liu, and L. Zhang, "A multichannel gray level co-occurrence matrix for multi/hyperspectral image texture representation," *Remote Sensing*, vol. 6, no. 9, pp. 8424–8445, 2014.

[70] P. Gong, D. J. Marceau, and P. J. Howarth, "A comparison of spatial feature extraction algorithms for land-use classification with spot hrv data," *Remote Sensing of Environment*, vol. 40, no. 2, pp. 137 – 151, 1992.

[71] C. Conrad, S. Fritsch, J. Zeidler, G. Rücker, and S. Dech, "Per-field irrigated crop classification in arid central asia using spot and aster data," *Remote Sensing*, vol. 2, no. 4, pp. 1035–1056, 2010.

[72] M. Aguilar, A. Vallario, F. Aguilar, A. Lorca, and C. Parente, "Object-based greenhouse horticultural crop identification from multi-temporal satellite imagery: A case study in almeria, spain," *Remote Sensing*, vol. 7, no. 6, pp. 7378–7401, 2015.

[73] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.

[74] L. Bruzzone, R. Cossu, and G. Vernazza, "Detection of land-cover transitions by combining multidate classifiers," *Pattern Recognition Letters*, vol. 25, no. 13, pp. 1491–1500, 2004.

[75] B. Zheng, S. W. Myint, P. S. Thenkabail, and R. M. Aggarwal, "A support vector machine to identify irrigated crop types using time-series landsat ndvi data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, pp. 103–112, 2015.

[76] R. Pande-Chhetri, A. Abd-Elrahman, T. Liu, J. Morton, and V. L. Wilhelm, "Object-based classification of wetland vegetation using very high-resolution unmanned air system imagery," *European Journal of Remote Sensing*, vol. 50, no. 1, pp. 564–576, 2017.

[77] Y. Qian, W. Zhou, J. Yan, W. Li, and L. Han, "Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery," *Remote Sensing*, vol. 7, no. 1, pp. 153–168, 2015.

[78] R. Saini and S. Ghosh, "Ensemble classifiers in remote sensing: A review," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1148–1152, IEEE, 2017.

[79] Wen Yang, Tongyuan Zou, Dengxin Dai, and Yongmin Shuai, "Supervised land-cover classification of terrasar-x imagery over urban areas using extremely randomized clustering forests," in *2009 Joint Urban Remote Sensing Event*, pp. 1–6, May 2009.

[80] G. J. Briem, J. A. Benediktsson, and J. R. Sveinsson, "Multiple classifiers applied to multisource remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, pp. 2291–2299, Oct 2002.

[81] T. Zou, W. Yang, D. Dai, and H. Sun, "Polarimetric sar image classification using multifeatures combination and extremely randomized clustering forests," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 4:1–4:12, Jan. 2010.

[82] F. Del Frate, F. Pacifici, G. Schiavon, and C. Solimini, "Use of neural networks for automatic classification from high-resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 800–809, April 2007.

[83] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *The Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, Jan. 2014.

[84] G. M. Foody and A. Mathur, "Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification," *Remote Sensing of Environment*, vol. 93, no. 1, pp. 107–117, 2004.

[85] G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers," *The Journal of Machine Learning Research*, vol. 9, pp. 2015–2033, June 2008.

[86] Jonathan Cheung-Wai Chan, Chengquan Huang, and R. DeFries, "Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 693–695, March 2001.

[87] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[88] R. Lawrence, A. Bunn, S. Powell, and M. Zambon, "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis," *Remote Sensing of Environment*, vol. 90, no. 3, pp. 331–336, 2004.

[89] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. wadsworth int," *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[90] M. Hansen, R. Dubayah, and R. DeFries, "Classification trees: an alternative to traditional land cover classifiers," *International Journal of Remote Sensing*, vol. 17, no. 5, pp. 1075–1081, 1996.

[91] A. Versluis and J. Rogan, "Mapping land-cover change in a haitian watershed using a combined spectral mixture analysis and classification tree procedure," *Geocarto International*, vol. 25, no. 2, pp. 85–103, 2010.

[92] S. Griffin, J. Rogan, and D. M. Runfola, "Application of spectral and environmental variables to map the kissimmee prairie ecosystem using classification trees," *GIScience & Remote Sensing*, vol. 48, no. 3, pp. 299–323, 2011.

[93] B. Ghimire, J. Rogan, V. R. Galiano, P. Panday, and N. Neeti, "An evaluation of bagging, boosting, and random forests for land-cover classification in cape cod, massachusetts, usa," *GIScience & Remote Sensing*, vol. 49, no. 5, pp. 623–643, 2012.

[94] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[95] M. Belgiu and L. Dr gu , "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[96] J. C.-W. Chan, P. Beckers, T. Spanhove, and J. V. Borre, "An evaluation of ensemble classifiers for mapping natura 2000 heathland in belgium using spaceborne angular hyperspectral (chris/proba) imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 18, pp. 13–22, 2012.

[97] S. E. Sesnie, B. Finegan, P. E. Gessler, S. Thessler, Z. R. Bendana, and A. M. S. Smith, "The multispectral separability of costa rican rainforest types with support vector machines and random forest decision trees," *International Journal of Remote Sensing*, vol. 31, no. 11, pp. 2885–2909, 2010.

[98] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[99] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 3179–3195, Sep. 2018.

[100] 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 Sensing of Environment*, vol. 152, pp. 109–124, 2014.

[101] P. Geurts and G. Louppe, "Learning to rank with extremely randomized trees," in *JMLR: Workshop and Conference Proceedings*, vol. 14, pp. 49–61, 2011.

[102] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[103] L. Breiman, "Some infinity theory for predictor ensembles," tech. rep., Technical Report 579, Statistics Dept. UCB, 2000.

[104] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," *arXiv preprint arXiv:1402.4293*, 2014.

[105] E. Scornet, "Random forests and kernel methods," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1485–1500, 2016.

[106] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, Apr 2006.

[107] C. C. Aggarwal, "Proximity-based outlier detection," in *Outlier Analysis*, pp. 111–147, Springer, 2017.

[108] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[109] G. Prashanth, V. Prashanth, P. Jayashree, and N. Srinivasan, "Using random forests for network-based anomaly detection at active routers," in *2008 International Conference on Signal Processing, Communications and Networking*, pp. 93–96, Jan 2008.

[110] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007–1011, 2005.

[111] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[112] V. N. Vapnik and S. Kotz, *Estimation of dependences based on empirical data*, vol. 40. Springer-Verlag NY, 1982.

[113] F. Roli and G. Fumera, "Support vector machines for remote sensing image classification," in *Image and Signal Processing for Remote Sensing VI*, vol. 4170, pp. 160–166, International Society for Optics and Photonics, 2001.

[114] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[115] I. Nitze, U. Schulthess, and H. Asche, "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil*, vol. 79, p. 3540, 2012.

[116] K. Chureesampant and J. Susaki, "Land cover classification using multi-temporal sar data and optical data fusion with adaptive training sample selection," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6177–6180, 2012.

[117] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, vol. 1, pp. 288–290, IEEE, 2003.

[118] W. M. Czarnecki, S. Podlewska, and A. J. Bojarski, "Robust optimization of svm hyperparameters in the classification of bioactive compounds," *Journal of cheminformatics*, vol. 7, no. 1, p. 38, 2015.

[119] Houtao Deng and G. Runger, "Feature selection via regularized trees," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, June 2012.

[120] P. Hao, Y. Zhan, L. Wang, Z. Niu, and M. Shakir, "Feature selection of time series modis data for early crop classification using random forest: A case study in kansas, usa," *Remote Sensing*, vol. 7, no. 5, pp. 5347–5369, 2015.

[121] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, pp. 2880–2889, July 2010.

[122] T. Rao and T. Rajinikanth, "A hybrid random forest based support vector machine classification supplemented by boosting," *Global Journal of Computer Science and Technology*, 2014.

[123] P. Du, J. Xia, J. Chanussot, and X. He, "Hyperspectral remote sensing image classification based on the integration of support vector machine and random forest," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 174–177, IEEE, 2012.

[124] 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 Sensing*, vol. 10, no. 5, p. 729, 2018.

[125] R. M. Cruz, R. Sabourin, and G. D. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Information Fusion*, vol. 41, pp. 195–216, 2018.

[126] A. Gil, Q. Yu, A. Lobo, P. Lourenço, L. Silva, and H. Calado, "Assessing the effectiveness of high resolution satellite imagery for vegetation mapping in small islands protected areas," *Journal of Coastal Research*, vol. 64, pp. 1663–1667, 2011.

[127] Y. Xie, Z. Sha, and M. Yu, "Remote sensing imagery in vegetation mapping: a review," *Journal of Plant Ecology*, vol. 1, no. 1, pp. 9–23, 2008.

[128] M. Pal and P. M. Mather, "A comparison of decision tree and backpropagation neural network classifiers for land use classification," in *In IEEE International Geoscience and Remote Sensing Symposium*, vol. 1, pp. 503–505 vol.1, 2002.

[129] F. Wang, "Fuzzy supervised classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 2, pp. 194–201, 1990.

[130] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, "Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas," *Remote Sensing of Environment*, vol. 187, pp. 156–168, 2016.

[131] K. Q. Ye, "Indicator function and its application in two-level factorial designs," *The Annals of Statistics*, vol. 31, no. 3, pp. 984–994, 2003.

[132] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery," *Sensors*, vol. 18, no. 1, 2018.

[133] P. Liu, K.-K. R. Choo, L. Wang, and F. Huang, "Svm or deep learning? a comparative study on remote sensing image classification," *Soft Computing*, vol. 21, pp. 7053–7065, Dec 2017.

[134] C.-I. Chang, *Hyperspectral data exploitation: theory and applications*. John Wiley & Sons, 2007.

[135] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[136] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: A review," *Sensors (Basel, Switzerland)*, vol. 12, no. 4, pp. 4764–4792, 2012.

[137] D. Tuia and G. Camps-Valls, "Cluster kernels for semisupervised classification of vhr urban images," in *2009 Joint Urban Remote Sensing Event*, pp. 1–5, May 2009.

[138] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[139] C. Deng and C. Wu, "The use of single-date modis imagery for estimating large-scale urban impervious surface fraction with spectral mixture analysis and machine learning techniques," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, pp. 100–110, 2013.

[140] M. Karlson, M. Ostwald, H. Reese, J. Sanou, B. Tankoano, and E. Mattsson, "Mapping tree canopy cover and aboveground biomass in sudano-sahelian woodlands using landsat 8 and random forest," *Remote Sensing*, vol. 7, no. 8, p. 10017, 2015.

[141] S. Tian, X. Zhang, J. Tian, and Q. Sun, "Random forest classification of wetland landcovers from multi-sensor data in the arid region of xinjiang, china," *Remote Sensing*, vol. 8, no. 11, 2016.

[142] J. Ham, C. Yangchi, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492–501, 2005.

[143] R. Colditz, "An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms," *Remote Sensing*, vol. 7, no. 8, pp. 9655–9681, 2015.

[144] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," in *Data Science & Engineering (ICDSE), 2012 International Conference*, pp. 64–68, IEEE, 2012.

[145] A. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[146] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[147] E. Izquierdo-Verdiguier, L. Gómez-Chova, and G. Camps-Valls, *Kernels for Remote Sensing Image Classification*, pp. 1–23. Wiley Encyclopedia of Electrical and Electronics Engineering, John Wiley Sons, 2015.

[148] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller, "A new discriminative kernel from probabilistic models," *Neural Comput.*, vol. 14, pp. 2397–2414, oct 2002.

[149] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[150] C. Persello and L. Bruzzone, "Kernel-based domain-invariant feature selection in hyperspectral images for transfer learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 2615–2626, May 2016.

[151] Y. Zhou, J. Peng, and C. L. P. Chen, "Extreme learning machine with composite kernels for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 2351–2360, June 2015.

[152] Q. Gao, S. Lim, and X. Jia, "Hyperspectral image classification using convolutional neural networks and multiple feature learning," *Remote Sensing*, vol. 10, no. 2, 2018.

[153] D. Stratoulias, V. Tolpekin, R. A. de By, R. Zurita-Milla, V. Retsios, W. Bijker, M. A. Hasan, and E. Vermote, "A workflow

for automated satellite image processing: from raw vhsr data to object-based spectral information for smallholder agriculture," *Remote Sensing*, vol. 9, no. 10, p. 1048, 2017.

[154] J. Rouse, R. Haas, J. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with erts," *NASA Special Publication*, vol. 351, p. 309, 1974.

[155] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, 1979.

[156] A. R. Huete, "A soil-adjusted vegetation index (savi)," *Remote Sensing of Environment*, vol. 25, no. 3, pp. 295–309, 1988.

[157] J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian, "A modified soil adjusted vegetation index," *Remote Sensing of Environment*, vol. 48, no. 2, pp. 119–126, 1994.

[158] D. Haboudane, J. R. Miller, N. Tremblay, P. J. Zarco-Tejada, and L. Dextraze, "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture," *Remote Sensing of Environment*, vol. 81, no. 2–3, pp. 416–426, 2002.

[159] A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira, "Overview of the radiometric and biophysical performance of the modis vegetation indices," *Remote Sensing of Environment*, vol. 83, no. 1–2, pp. 195–213, 2002.

[160] R. M. Haralick, K. Shanmugam, *et al.*, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6, pp. 610–621, 1973.

[161] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[162] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[163] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Integrating support vector machines and random forests to classify crops in time series of worldview-2 images," in *Image and Signal Processing for Remote Sensing XXIII*, vol. 10427, p. 104270W, International Society for Optics and Photonics, 2017.

[164] L. Gómez-Chova, J. Muñoz-Marí, V. Laparra, J. Malo-López, and G. Camps-Valls, *A Review of Kernel Methods in Remote Sensing Data Analysis*, pp. 171–206. Springer Berlin Heidelberg, 2011.

[165] F. Löw, U. Michel, S. Dech, and C. Conrad, "Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines," *ISPRS Journal of Photogrammetry and Remote sensing*, vol. 85, pp. 102–119, 2013.

[166] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Evaluating the performance of a random forest kernel for land cover classification," *Remote Sensing*, vol. 11, no. 5, 2019.

[167] C. Englund and A. Verikas, "A novel approach to estimate proximity in a random forest: An exploratory study," *Expert Systems with Applications*, vol. 39, no. 17, pp. 13046–13050, 2012.

[168] E. Izquierdo-Verdiguier, R. Jenssen, L. Gómez-Chova, and G. Camps-Valls, "Spectral clustering with the probabilistic cluster kernel," *Neurocomputing*, vol. 149, Part C, pp. 1299–1304, 2015.

[169] L. van der Maaten, "Learning discriminative fisher kernels," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, (USA), pp. 217–224, Omnipress, 2011.

[170] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, July 2004.

[171] E. F. Vermote, D. Tanré, J. L. Deuze, M. Herman, and J.-J. Morcette, "Second simulation of the satellite signal in the solar spectrum, 6s: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 3, pp. 675–686, 1997.

[172] M. Louhaichi, M. M. Borman, and D. E. Johnson, "Spatially located platform and aerial photography for documentation of grazing impacts on wheat," *Geocarto International*, vol. 16, no. 1, pp. 65–70, 2001.

[173] C. Song, F. Yang, and P. Li, "Rotation invariant texture measured by local binary pattern for remote sensing image classification," in *2010 Second International Workshop on Education Technology and Computer Science*, vol. 3, pp. 3–6, IEEE, 2010.

[174] M. Pal, "Kernel methods in remote sensing: a review," *ISH Journal of Hydraulic Engineering*, vol. 15, no. sup1, pp. 194–215, 2009.

[175] C. Cusano, P. Napoletano, and R. Schettini, "Remote sensing image classification exploiting multiple kernel learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2331–2335, Nov 2015.

[176] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, pp. 674–677, Oct 2007.

[177] M. Pal, "Random forest classifier for remote sensing classification," *International Journal Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[178] E. Scornet, "Random forests and kernel methods," *IEEE T Inform Theory*, vol. 62, pp. 1485–1500, March 2016.

[179] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land cover classification using extremely randomized trees: A kernel perspective," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019.

[180] H. Xu, T. E. Twine, and X. Yang, "Evaluating remotely sensed phenological metrics in a dynamic ecosystem model," *Remote Sensing*, vol. 6, no. 6, pp. 4660–4686, 2014.

[181] F.-M. Chmielewski, *Phenology and Agriculture*, pp. 505–522. Dordrecht: Springer Netherlands, 2003.

[182] H. Lieth, *Purposes of a Phenology Book*, pp. 3–19. Berlin, Heidelberg: Springer Berlin Heidelberg, 1974.

[183] W. Sun, S. Liang, G. Xu, H. Fang, and R. Dickinson, "Mapping plant functional types from modis data using multisource evidential reasoning," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1010–1024, 2008.

[184] J. Robin, R. Dubayah, E. Sparrow, and E. Levine, "Monitoring start of season in alaska with globe, avhrr, and modis data," *Journal of Geophysical Research: Biogeosciences*, vol. 113, no. G1, 2008.

[185] R. A. Berk, *Statistical learning from a regression perspective*, vol. 14. Springer, 2008.

[186] E. Izquierdo-Verdiguier, R. Zurita-Milla, and R. A. de By, "On the use of guided regularized random forests to identify crops in smallholder farm fields," in *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp. 1–3, June 2017.

[187] M. A. Olson and A. J. Wyner, "Making sense of random forest probabilities: a kernel perspective," 2018.

[188] X. J. Zhu, "Semi-supervised learning literature survey," tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 2005.

[189] J. Muñoz-Marí, L. Gómez-Chova, G. Camps-Valls, and J. Calpe-Maravilla, "Image classification with semi-supervised one-class support vector machine," in *Image and Signal Processing for Remote Sensing XIV*, vol. 7109, p. 71090B, International Society for Optics and Photonics, 2008.

[190] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari, and J. Calpe, "Semisupervised image classification with laplacian support vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, pp. 336–340, July 2008.

[191] S. Niazmardi, B. Demir, L. Bruzzone, A. Safari, and S. Homayouni, "Multiple kernel learning for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1425–1443, 2017.

[192] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple kernel learning for classification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2852–2865, 2012.