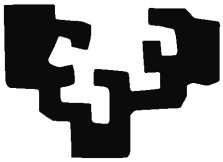


Contributions to Ensemble Classifiers with Image Analysis Applications

By

Borja Ayerdi Vilches

Submitted to the department of Computer Science and Artificial Intelligence in partial fulfillment of the requirements for the degree of Doctor of Philosophy

PhD Advisor:

Prof. Manuel Graña Romay
at The University of the Basque Country (UPV/EHU)

Euskal Herriko Unibertsitatea
Universidad del País Vasco
Donostia - San Sebastián
2015

Acknowledgments

Prof. Manuel Graña was my Computer Vision profesor in 2011. I was finishing my degree in Computer Science and his subject was really interesting and new for me. I sent him some emails about new ideas and the source code of a Tracking-Learning-Detection system (TLD). After some emails his response was: *"Yo lo que necesito es gente que trabaje, no gente que me maree."*

I can still remember this day. His answer shocked me. I was pissed and challenged by his response. This day I decided that I wanted to do a PhD and I wanted to do it with him.

I want to give him special thanks not only for mentoring me during this years also for his friendship and patience. Without him this Thesis would not be possible.

I am very grateful to Prof. Michał Woźniak for giving me the opportunity of doing an internship in the Machine Learning Group at Wrocław University of Technology. It was a great experience. *Dziękuję!*

I want to thank all my workmates and collegues at GIC who helped me in this task and all my friends that have always been there for me.

Also, and most importantly, I want to thank my family -specially my parents and brother- and Iraia for their unending support in my life. Thanks.

Borja Ayerdi Vilches

"Imagine the future, change the present."

Contributions to Ensemble Classifiers with Image Analysis Applications

by
Borja Ayerdi Vilches

Submitted to the Department of Computer Science and Artificial Intelligence on November 9th, 2015, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

The Thesis touches two main aspects, one is the proposal of new classifier architectures, and the other is their application to image analysis. From the point of view of proposing new classification architectures, this Thesis has two main contributions, one is an innovative ensemble of classifiers based on randomized archiectures, such as Extreme Learning Machines (ELM) and Random Forest (RF), called Hybrid Extreme Rotation Forest (HERF), and its enhance Anticipative HERF (AHERF) which carries a model selection based on predictions of individual classifier achitecture performance on the dataset at hand. Formal rationale for the AHERF model selection is provided. Moreover, we provided a formal proof of the convergence of ensembles of ELM regressors which improves its usability and reproducibility of results. On the application versant, we have dealt with two main types of images: hyperspectral images from remote sensing, and medical images from specific blood vessel related patologies, and the Alzheimer's Disease diagnosis. In all cases the main tool has been the ensembles of classifiers, and specific Active Learning strategies based on ensembles of classifiers. The specific blood vessel segmentation problems are the segmentation of Aortic Abdominal Aneurysm trombus in 3D Computerized Tomography Angiograms, and retinal image blood vessel segmentation. Results achieved both in terms of classification performance and time savings for the human interactive segmentation allow to recommend such approaches for the clinical practice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides a general introduction to this Thesis. In a very broad summary, the Thesis main axis are some contributions to the field of classification methods, and their application to image processing. Section 1.1 gives introductory comments, motivation and an overview of the Thesis contents. Section 1.2 lists the contributions of the Thesis; Section 1.3 enumerates the publications obtained as a result of the research associated with this Thesis studies. Lastly, Section 1.4 details the chapters structure.

1.1 Introductory comments and motivation

Overall, the Thesis deals with innovative classification algorithms and their application in two image processing domains, those of medical image and hyperspectral images. In this section we present introductory views of the topics covered in the Thesis, including the presentation of the applications tackled with the proposed computational methods. We start with introducing and motivating the ensemble classifiers, a hot topic in the Machine Learning literature, where we have realized the core of the computational contributions in the Thesis. Then we introduce Active Learning, which we have extensively used for the design of medical image segmentation algorithms. The last three subsections introduce the application domains, namely Alzheimer's Disease computer aided diagnosis, blood vessel segmentation in computerized tomography and retinal images, and hyperspectral image analysis.

1.1.1 Ensembles of classifiers

The idea of ensembles of classifiers was proposed by Hansen and Salamon [69] in the context of Artificial Neural Network learning. Their work showed that an ensemble of neural networks with a plurality consensus scheme can be expected

to improve over a single neural network performance. In general, ensembles of classifiers aim to obtain improved classification results by the combination of weak and diversified classifiers [137, 183]. Early propositions of ensembles of classifiers consisted of a homogeneous collection of individual classifiers. For instance, the Random Forest [17, 18] combine, by majority voting, the outputs of a collection of Decision Trees (DT), each built from bootstrapped training data on random variable selections. Other examples are the ensembles of Support Vector Machines (SVM) [99], and sequential selection approaches [168].

We have considered specially ensembles of Extreme Learning Machines (ELM) [80] and their ensemble constructs. The ELMs have a great appeal because they posses very quick learning, however they lack stability, i.e. repeated instances of training may provide wide different performance results. Ensembles have been proven to achieve such stability by a direct application of the law of large numbers. For instance, Voting ELM (V-ELM) [24] is a direct composition by majority voting of the outputs a collection of independently trained ELMs. This straightforward ensemble approach has been shown to enhance the stability of ELM results, by approaching the limit perfect classification performance when the number of training samples grows sufficiently. The Ensemble of Online Sequential ELMs (EOS-ELM) [102] aims to improve stability and reproducibility of results of the single OS-ELM classifier. The approach has also been applied to regression problems, combining the output of the individual OS-ELM regressors by averaging them. The addition of a forgetting mechanism in [196] improves the EOS-ELM performance in problem domains where data lifetime is limited, such as happens in financial market applications, and similar prediction tasks. The EOS-ELM discards outdated data samples in the training process by a filtering process that weights the prediction value of each datum. The Multiple ELM (MELM) approach [114] combine specific sampling and feature selection methods with an ensemble of ELMs to improve prediction in highly imbalanced datasets. The process of the data starts with the SMOTE algorithm in order to correct dataset imbalance. Then feature selection is performed on the basis of information gain. The MELM is trained on the dataset to play the role of knowledge repository for the last element of the architecture, which is a DT trained on the MELM outputs. The DT is used as a set of rules allowing interpretation of the system reasoning.

The emphasis on diversification follows from the expectation that classifiers with quite different fields of expertise may have a synergistic effect on the whole to improve classification performance. Approaches to obtain diversification include data preprocessing by randomized rotation matrices, as in Rotation Forests [146], or supervised projections [60]. Another way to obtain diversification is to build heterogeneous ensembles, where individual classifiers are trained with different approaches, so that intrinsic sensitivity of the models to data distribution may

be exploited to get classifier diversity. In this line of work, the Thesis includes the proposition of Hybrid Extreme Rotation Forest (HERF) [10], which is an heterogenous ensemble of ELMs and DTs over a randomized rotation of the bootstrapped training data as done by the Rotation Forest. The approach is able to profit from the best adapted method, DT or ELM, to the problem at hand, improving homogeneous ensembles in a series of benchmarking experiments. HERF has been shown to be successful in remote sensing and medical image segmentation [8]. However, the determination of the composition of the ensemble is very restricted, lacking flexibility. Therefore, we have proposed an improvement that carries model selection by trying to predict the individual model performances on the data, which we have called Anticipative HERF (AHERF).

1.1.2 Active Learning

Building a supervised classifier consists in learning a mapping of data features into a set of classes given a labeled training set. In many real life situations, obtaining the training data is costly, time consuming and error prone. This makes the construction of the training set a cumbersome task requiring extensive manual analysis of the image. This is typically done by visual inspection of the images and labeling of each sample pixel. Consequently, the training set is highly redundant and training phase of the model is significantly slowed down. Besides, noisy pixels may interfere the class statistics, which may lead to poor classification performances and/or overfitting. Therefore a desirable training set must be constructed in a smart way, meaning it must represent correctly the class boundaries by sampling discriminative pixels. Generalization is the ability of providing correct class labels to previously unseen data. Active learning tries to exploit the interaction with a user that provides the labels for selected training set samples, with the aim of obtaining the most accurate classification using the smallest possible training set. Samples are optimally selected for inclusion, ensuring that they will provide the greatest increase in accuracy [155]. To this end, the incremental data selection uses a classification uncertainty criterium that does not require actual knowledge of the data sample label, thus there is no circularity in the analysis. Besides economy of computation and data labeling, Active Learning assumes that the underlying data statistics are non stationary, so that the classifier built at one time instant for a subject data will not be optimal later on. In this Thesis, we have applied Active Learning for medical image analysis, specifically blood vessels segmentation in the CT images of Abdominal Aortic Aneurysm and retinal images.

1.1.3 Alzheimer's Disease

Alzheimer's Disease (AD) is one of the most important causes of disability in the elderly and with the increasing proportion of elderly in many country populations, the number of dementia patients will rise also. Due to the socioeconomic importance of the disease in occidental countries there is a strong international effort focus in AD. In the early stages of AD brain atrophy may be subtle and spatially distributed over many brain regions, including the entorhinal cortex, the hippocampus, lateral and inferior temporal structures, as well as the anterior and posterior cingulate. There are many computational algorithms design efforts trying to find image biomarkers that may be used for the non-invasive diagnosis of AD and other neurodegenerative diseases. The main emphasis of our works is to obtain the localization of the most discriminative regions, hence we preserve the anatomical localization of the features used for classification. We report in this Thesis an application of ensembles of classifiers over specific features extracted from anatomical MRI.

1.1.4 Blood vessel segmentation

Segmentation of blood vessels [12, 119, 107] is one of the essential medical computing tools for clinical assessment of vascular diseases, partitioning the image into non-overlapping vascular and background regions. Based on the partitioning results, surfaces of vasculatures can be extracted, modeled, manipulated, measured and visualized. These are very useful and play important roles for the endovascular treatments of vascular diseases. Vascular diseases are one of the major sources of morbidity and mortality worldwide. Therefore, developing reliable and robust image segmentation methods for angiographyis a priority in many research groups. Image segmentation in angiography poses strong challenges. For example, blood vessels can contain low or turbulent flow. This can lead to low signal-to-noise (SNR) ratio in the angiograms. The conventional segmentation methods based on image intensity alone may then fail when there is a significant signal drop in the vascular region. Furthermore, intensity spatial inhomogeneity often violates the assumption that intensity is a piecewise constant function in the spatial domain common in many segmentation processes. Finally, the intensity contrast between vessel and background regions, or inside vessel regions can vary from region to region. Therefore, the local intensity statistics in the vessel and background regions may not be reliable, or the intensity gradient magnitude may not be large enough on the vessel boundary for the conventional image segmentation methods. In this Thesis we apply an Active Learning approach to train ensemble classifiers on spatial feature to discriminate blood vessels.

1.1.4.1 Abdominal Aortic Aneurysm (AAA)

One of the applications worked upon in this Thesis is the segmentation of Abdominal Aortic Aneurysm (AAA) images obtained with 3D Computerized Tomography Angiography (CTA). The AAA is a local dilation of the Aorta that occurs between the renal and iliac arteries. The weakening of the aortic wall leads to its deformation and the generation of a thrombus. Generally, an AAA is diagnosed when the minimum anterioposterior diameter of the aorta reaches 3.0 cm [72]. The majority of aortic aneurysms are asymptomatic and without complications. Aneurysms that cause symptoms have a higher risk of rupture. Abdominal pain or back pain are the two main clinical features suggestive of either the recent expansion or leakage. The complications are often life threatening and can occur in a short space of time. Therefore, the challenge is to diagnose before the onset of symptoms. Asymptomatic aneurysms are often detected incidentally[95]. The statements about the natural course of the disease and the risk of rupture are mainly based on the results of the randomized observational studies that compared immediate repair with surveillance for small AAA[106]. The evidence report prepared for the Agency for Healthcare Research and Quality (AHRQ) analyzed these and further observational studies [180]. They came to the conclusion that the annual risk of rupture is 1% or lower for AAA less than 5.5 cm. The 1-year risk of rupture increases with aneurysm size and may exceed 10% in individuals with AAA > 6 cm. For AAA that attain a size of > 8 cm, the risk may exceed 25 % at 6 months. Female sex, higher mean arterial blood pressure, current smoking, and poor lung function increase the risk of aneurysm rupture in addition to the size of initial AAA diameter. CTA allows minimally invasive visualization of the Aorta's lumen, thrombus and calcifications. The segmentation of the AAA thrombus is a challenging task due to the low thrombus signal contrast, great shape variability, both intra and inter-subjects, and little availability of prior information.

1.1.4.2 Retinal Images

Ocular fundus image assessment is a diagnostic tool of vascular and non vascular pathology. Retinal vasculature inspection may reveal hypertension, diabetes, arteriosclerosis, cardiovascular disease and stroke. The major challenges for the retinal vessel segmentation are: (1) the presence of lesions which may be misdetected as blood vessels; (2) low contrast around thinner vessels, (3) multiple scales of vessel size.

Retinal image segmentation has been tackled from many points of view:

- as supervised classification problem, where classifiers are trained to discriminate vessel from non-vessel pixels,

- by unsupervised methods based on *a priori* assumptions on vessel properties in the image which are appropriately detected by specific filters, either matched filters or edge detection oriented filters.

- multi-scale approaches combining several scale filters into a single map for decision.

- many approaches end up with some kind of post-processing or cleaning of the result image in order to remove false positives. Often these processes are heuristics for removing isolated detections.

1.1.5 Hyperspectral images: classification

Interest in hyperspectral image analysis has grown in recent years due to both growing number of deployed sensors and the richness of their applications as well as the challenges for basic research posed by related computational problems [13, 14, 22, 105, 140]. Hyperspectral images provide high resolution sampling of the light spectrum at each pixel, promising direct recognition of the scene materials from the pixel spectra by either machine learning or statistical methods. Such works have been pursued since the early days of multispectral data availability [104, 144]. The literature makes a broad distinction between pixel and sub-pixel resolution analysis [14], the later referred as hyperspectral unmixing [16].

Approaches dealing with spectral classification are abundant in the literature: Maximum Likelihood [66, 82, 179], Linear Discriminant Analysis [44], Support Vector Machines (SVM) [98, 163, 96, 64, 131], Artificial Neural Networks [70, 176], Multinomial Logistic Regression [109]. Most of them use single classifiers, but recently some ensemble approaches have been proposed, such as Adaboost [27], Random Forests [184, 26, 147], multiple kernel ensembles trained by active learning [195], or the Hybrid Extreme Rotation Forest (HERF) [**?**, 11]. A relevant issue in remote sensing data is the variability of observations in time, which impedes the reuse of the trained classifiers on new image captures from the same scene, encouraging Domain Transfer approaches which we have not pursued in this Thesis. Local correlation change measures can be computed on multiple time images [85], on the other hand domain transfer tries to reuse as much as possible the past trained classifier systems, for instance by way of active learning [138].

1.1.6 Hyperspectral images: unmixing

Hypespectral image pixels are high dimensional vectors produced by high resolution sampling of the light spectrum in the visual and infrared regions [103]. Research on hyperspectral image analysis is driven by the expectation that it will be

possible to identify the materials in the imaged scene through the interpretation or classification of the pixel spectra. Among the challenges posed by hyperspectral images [15], one of the most salient is unmixing, aka sub-pixel resolution material identification, which provides fractional abundances of the composing materials at each pixel. Linear methods assuming a linear mixing of constituent spectra, i.e. the endmembers, are the earliest approaches [97]. The linear mixture model is usually justified by the existence of well defined regions inside the area covered by a pixel (which can be hundred of square meters in some sensors) of materials with well differentiated spectra. In general, the problem of linear unmixing is decomposed into two problems:

- Finding the library of endmembers, i.e. the spectra of the constituent materials, that will serve as the basis for the decomposition of the image pixel spectra. This library may be extracted from some public database of material reflectance measurements, (i.e. USGS spectral library[1]), or may be induced from the image by some endmember extraction algorithm. A critical problem is the *a priori* determination of the number of endmembers, which is an instance of the intrinsic dimensionality problem [71]. Endmember induction relies on the assumption that the data lies in a convex region. When this assumption does not hold, methods to decompose non-convex data into convex partitions [192] allow to extend conventional endmember induction to this case.

- Solving the inverse linear problem to obtain the abundances of the endmembers in the pixel. Solving this problem may be tricky if we enforce the constraints of the mixing process, that is, that all abundances are positive and add up to 1. Least squares solutions may break these constraints, so it is required to resort to constrained optimization methods.

Some approaches, such as the Independent Component Analysis (ICA) [100] or the Non-Negative Matrix Factorization (NNMF) [128] solve both problems simultaneously. Other approaches are sequential, performing first the endmember identification and afterwards the unmixing analysis of the image. Non-linear unmixing approximations, discussed in Chapter 3, need to be provided with the identification of material spectra (which can no longer be called endmembers because of the non-linear transformations involved), because it is not possible to induce them through the inversion of the non-linear mixing transformation. In this thesis work, we do not try to find the constituent material spectra. Instead we seek to obtain the abundances of the target classes by training a regression model implemented by Extreme Learning Machines (ELM). In fact, the approach is a generalization of

[1] http://speclab.cr.usgs.gov/spectral-lib.html

the unmixing process encompassing non-linear unmixing, and it is not restricted by the limitations of other methods, such as the assumption of statistical independence underlying ICA, or the non-negative vector basis required by NNMF. The identification of endmember vectors is embedded in the weights from the input to the hidden units of the ELM, where the hidden units activation functions introduce the non-linear aspect of the unmixing process.

1.2 Contributions

This Thesis mainly has two types of contributions. Computational contributions and application-oriented or practical contributions.

From a computational point of view, this Thesis has the following contributions:

- A novel Active Learning scheme using Random Forests and uncertainty calculation that allows fast accurate interactive image segmentation.
- Hybrid Extreme Rotation Forest.
- Adaptative Hybrid Extreme Rotation Forest.
- Spectral-spatial semi-supervised learning methods.
- Nonlinear unmixing and reconstruction by ELM regression ensembles.

The application-oriented contributions in this Thesis are:

- Medical Images
  - Active Learning of HERF for CTA image segmentation.
  - Enhancing active learning computed tomography image segmentation with domain knowledge.
  - Active learning with bootstrapped dendritic classifier applied to medical image segmentation.
  - Meta-ensembles of classifiers for Alzheimer's disease detection using independent ROI features.
  - Random Forest active learning for retinal image segmentation.
  - Automated segmentation of subcutaneous and visceral adipose tissues from MRI.
- Hyperspectral Data

- Nonlinear unmixing and reconstruction by ELM regression ensembles.
- Spectral-spatial semi-supervised learning with spatial correction using AHERF.
- Spatially regularized semisupervised ensembles of ELM's.

1.3 Publications

The following publications have been achieved during the works of this Thesis. Some of them are not directly related to the main topics presented in this Thesis reports, but are nevertheless referred here for a complete view of the efforts and collaborations carried out in this period.

1. Borja Ayerdi, Manuel Graña, *"Hyperspectral Image Analysis by Spectral-Spatial Processing and Anticipatively Tuned Hybrid Extreme Rotation Forest Classification",* Transactions on Geoscience and Remote Sensing (TGARS). [JCR 2014 - 3.514]

2. Ariadna Besga, Itxaso González-Ortega, Enrique Echeburúa, Alexandre Savio, Borja Ayerdi, Darya Chyzhyk, Jose LM Madrigal, Juan C. Leza, Manuel Graña, Ana González-Pinto, *"Discrimination between Alzheimer's Disease and Late Onset Bipolar Disorder using multivariate analysis",* Frontiers in Aging Neuroscience. [JCR 2014 - 4.000]

3. Ariadna Besga-Basterra, Borja Ayerdi, Guillermo Alcalde-Behold, Antonio Manzano-Ramirez, Pedro Lopetegui-Eraso, Manuel Graña, Ana González-Pinto-Arrillaga, *"Risk factors for emergency department short time readmission in stratified population"*, Biomed Research International. [JCR 2014 - 1.579]

4. Borja Ayerdi, Oier Echaniz, Alexandre Savio, Manuel Graña, *"Automated segmentation of subcutaneous and visceral adipose tissues from MRI"*, In-Med 2015.

5. Borja Ayerdi, Manuel Graña, *"Random Forest Active Learning for Detecting Blood Vessels in Angiography"*, CORES 2015.

6. Ariadna Besga-Basterra, Darya Chyzhyk, Itxaso González-Ortega, Alexandre Savio, Borja Ayerdi, Jon Echeveste, Manuel Graña, Ana González-Pinto, *"White matter anatomical correlates with biomarkers discriminating between Alzheimer's Disease and Late Onset Bipolar Disorder by eigenanatomy on fractional anisotropy imaging"*, Current Alzheimer Research 2015. [JCR 2014 - 3.889]

7. Borja Ayerdi, Manuel Graña, *"Hyperspectral Image Nonlinear Unmixing and Reconstruction by ELM Regression Ensemble"*, Neurocomputing. [JCR 2014 - 2.083]

8. Borja Ayerdi, Manuel Graña *"Hyperspectral Image Nonlinear Unmixing by Ensemble ELM Regression"*, Jiuwen Cao et al. (Eds.): Proceedings of ELM-2014 Volume 2, Proceedings in Adaptation, Learning and Optimization1, pp. 289-297. Springer.

9. Borja Ayerdi, Ion Marqués and Manuel Graña, *"Spatially regularized semisupervised Ensembles of Extreme Learning Machines for hyperspectral image segmentation"*, Neurocomputing. [JCR 2014 - 2.083]

10. Borja Ayerdi, Manuel Grana *"Hybrid Extreme Rotation Forest"*, Neural Networks. [JCR 2014 - 2.708]

11. Pawel Ksieniewicz, Dariusz Jankowski, Borja Ayerdi, Konrad Jackowski, Michal Wozniak and Manuel Graña, *"A Novel Hyperspectral Segmentation Algorithm - Concept and Evaluation"*, Logic Journal of the IGPL. [JCR 2012 - 1.136]

12. J. Maiora , G. García , B. Ayerdi , M. Graña , M. De Blas, *"Active Learning enhanced with Expert Knowledge for Computed Tomography Image Segmentation"*, InMed 2013.

13. Borja Ayerdi, Josu Maiora, Manuel Graña, *"Aplications of Hybrid Extreme Rotation Forests for Image Segmentation"*, Journal of Hybrid Intelligent Systems .

14. Borja Ayerdi, Josu Maiora, Manuel Graña, *"Enhancing Active Learning Computed Tomography Image Segmentation with Domain Knowledge"*, HAIS 2013.

15. Borja Ayerdi, Alexandre Savio, Manuel Graña, *"Meta-ensembles of classifiers for Alzheimer's disease detection using independent ROI features"*, J.M. Ferrandez et al. (Eds.): IWINAC 2013, Part II, LNCS 7931, pp. 122–130. Springer, Heidelberg (2013).

16. Borja Ayerdi, Manuel Graña, *"On Spatial Regularization for Semisupervised Hyperspectral Image Segmentation Using Hybrid Extreme Rotation Forest"*, WHISPERS 2013.

17. Darya Chyzhyk, Borja Ayerdi, Josu Maiora, *"Active Learning with Bootstrapped Dendritic Classifier applied to medical image segmentation"*, Pattern Recognition Letters. [JCR 2013 - 1.062]

18. Josu Maiora, Borja Ayerdi, Manuel Graña, *"Random Forest Active Learning for Computed Tomography Angiography Image Segmentation"*, Neurocomputing. [JCR 2013 - 2.005]

19. Borja Ayerdi, Josu Maiora, Manuel Graña, *"Active learning of Hybrid Extreme Rotation Forests for CTA image segmentation"*, Hybrid Intelligent Systems (HIS), 2012 12th International Conference on (pp. 543-548). IEEE.

20. Ekaitz Zulueta Guerrero, Naiara Telleria Garay, Jose Manuel Lopez-Guede, Borja Ayerdi Vilches, Eider Egilegor Iragorri, David Lecumberri Castaños, Ana Belén de la Hoz Rastrollo and Carlos Pertusa Peña, *"Prediction of Bladder Cancer Recurrences Using Artificial Neural Networks",* Lecture Notes in Computer Science, 2010, Volume 6076, Hybrid Artificial Intelligence Systems, Pages 492-499.

1.4 Structure of the Thesis

The contents of the Thesis are divided in three blocks. The first one is centered on new classifier architectures. The second block encompasses the works done during the Thesis related with hyperspectral data. The third block contains the main applications using medical images. The chapters of the Thesis are organized as follows:

1. Chapter 2 contains the computational contributions of the Thesis in the aspect of the design of new classifier architectures. Specifically we review the basic approaches, namely Extreme Learning Machines and Random Forests, before proceeding to present the Hybrid Extreme Rotation Forest and the Anticipative Hybrid Extreme Rotation Forest. Besides, we present our approach to Active Learning.

2. Chapter 3 refers two main applications of the ensemble classifiers to hyperspectral image processing. We discuss spectral unmixing, and the application of ensembles of Extreme Learning Machines to this task. Also we introduce a novel spatial-spectral process for the pixel based processing the images to obtain thematic maps. We provide extensive experimental results.

3. Chapter 4 provides an account of our works related to medical image analysis applications. We report works on the segmentation of Abdominal Aortic Aneurysm images obtained with 3D Computerized Tomography Angiography using an Active Learning approach and several ensemble classifiers.

The same approach is applied to retinal images. We provide extensive experimental results. Additionaly, we report the application to brain magnetic resonance images for the classification of Alzheimer's Disease patients.

4. Chapter 5 summarizes the contributions and conclusions of this Thesis in terms of new classification ensemble methods and applications in different domains.

5. Appendix A presents the main experimental datasets used in this Thesis.

Chapter 2

Ensemble Classifiers

This Chapter contains the computational contributions of the Thesis reagarding the design of new classifier architectures. We have focused on the develpment and test of ensemble classifiers with a strong random component, essentially elaborated from Extreme Learning Machines (ELM), Random Forest (RF), and Rotation Forests which are among the main trends in the paradigm of randomized classifier architectures. We present the classical architectures in sections 2.1 and 2.2, and two innovative architectures produced by our own work, the Hybrid Extreme Rotation Forest (HERF) in Section 2.3, and the Anticipative HERF (AHERF) in Section 2.4. We also describe the Bootstraped Dendritic Classifiers (BDC) in Section 2.5 developed in collaboration with Darya Chyzhyk. Finally, we introduce Active Learning techniques in Section 2.6, which have been useful to deal with heavily imbalanced datasets in medical image processing. Their applications are discussed in the following chapters.

2.1 Decision Tree Ensemble Classifiers

Decision trees [19, 141] are built by recursive partitioning of the data space. A univariate (single attribute) split is recursively defined for each tree node, from the root to the leaves, of the tree using some criterion (e.g., mutual information, gain-ratio, gini index). The data space and data samples are then partitioned according to the univariate test. Tree leaves correspond to the actual assignment of data samples to classes. Often, a pruning process is applied in order to reduce the tree complexity and the overfitting on the training data.

Random Forest (RF) is a classifier ensemble [18] that encompasses bagging [17] and random decision forests [5, 73], being used in a variety of applications. RF captures complex interaction structures in data, and are argued in [18] to be resistant to both data overfitting happening when individual trees are very deep

and no pruned, and under-fitting happening when individual trees are too shallow. RF became popular due to its simplicity of training and tuning while offering a similar performance to boosting. Consider a RF as a collection of decision tree predictors

$$\{h(\mathbf{x}, \psi_k); k = 1, ..., K\},$$

where ψ_t are independent identically distributed random vectors whose nature depends on their use in the tree construction, and each tree casts a unit vote to find the most popular class of input $\mathbf{x}$.

Given a dataset of N samples, a bootstrapped training dataset is used to grow a tree $h(\mathbf{x}; \psi_k)$ on a randomly selected subset of instances with replacement from the original training sample. RF also employs random feature selection. At each node of the decision tree, $\hat{d}$ variables are selected at random $\hat{d} \ll d$. Each decision tree is grown using CART methodology without pruning. This tree growing approach recursively picks the best data split of each node with the criteria of how well separates the classes contained in the parent node.

The independent identically distributed random vectors ψ_t determine the random dimension selection and data sample bootstrapping prior to tree training, which are the source for individual tree diversity and uncorrelation of their outputs. This uncorrelation between the trees and the strength of the individual trees determine the generalization error of the forest.

The trained RF can be used for classification of a new input $\mathbf{x}$ by majority voting among the class prediction of the RF trees. Note that in RF the Law of Large Numbers insures convergence as $k \to \infty$, therefore avoiding overfitting.

The Rotation Forest [146] tries to enhance the diversity of individual trees in the RF by building and applying a random rotation matrix to the training data previous to the tree training. The process starts with a random partition of the input variables. Separate PCA rotation matrices are computed for each subset of input variables, which are composed into a single rotation matrix by interleaving the separate rotation matrix columns according to the original order of the feature variables.

2.2 Extreme Learning Machines

In this subsection we introduce the basic Extreme Learning Machine (ELM) architecture, and an enhanced architecture before introducing the ensembles of ELMs and the specific Voting ELM that was used for benchmarking in computational experiments.

2.2.1 Basic ELM

The Extreme Learning Machine (ELM) [80] is a very fast training algorithm for single-layer feedforward neural networks (SLFN). The key idea of ELM is the random initialization of the SLFN hidden layer node weights. Consider a set of M data samples $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $y_i \in \mathbb{R}^{d_2}$. Then, a SLFN with N hidden neurons is modeled as the following expression:

$$\mathbf{y} = \Phi(\mathbf{x}) = \sum_{i=1}^{N} \beta_i f(\mathbf{w}_i \cdot \mathbf{x} + b_i), \tag{2.1}$$

where $f(x)$ is the activation function, $\mathbf{w}_i$ the input weights to the i-th neuron in the hidden layer, b_i the hidden layer unit bias and β_i are the output weights. The application of this equation to all available data samples can be written in matrix form as

$$\mathbf{H}\beta = \mathbf{Y}, \tag{2.2}$$

where $\mathbf{H}$ is the hidden layer output matrix defined as the output of the hidden layer for each input sample vector

$$\mathbf{H} = \begin{pmatrix} f(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x}_1 + b_N) \\ \vdots & \ddots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x}_M + b_1) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x}_M + b_N) \end{pmatrix}, \tag{2.3}$$

and $\beta = (\beta_1, \ldots, \beta_N)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)^T$. The way to calculate the output weightsβ from the hidden-layer to the target values is computing the of a Moore–Penrose generalized inverse of the matrix $\mathbf{H}$, denoted as $\mathbf{H}^\dagger$. The least mean square error solution is $\beta = \mathbf{H}^\dagger \mathbf{Y}$.

Algorithm 2.1 ELM.

Input A set of training samples $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $y_i \in \mathbb{R}^{d_2}$, an activation function $f(x)$,and a number of hidden nodes N:

1. Randomly generate bias matrix $b = (b_1, \ldots, b_N)^T$ and weight matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$.
2. Calculate $\mathbf{H}$ as in equation 2.3.
3. Calculate the output weight matrix $\beta = \mathbf{H}^\dagger \mathbf{Y}$.

The orthogonal projection method can be used to obtain the Moore–Penrose

pseudoinverse of $\mathbf{H}$: $\mathbf{H}^{\dagger} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$, where $\mathbf{H}^T$ is the transpose of matrix $\mathbf{H}$. In that case, we can add a ridge parameter $1/\lambda$ to the diagonal of $\left(\mathbf{H}^T\mathbf{H}\right)$, which is the ridge-regression regularization approach, to stabilize the solution. Therefore, the regularized calculation of the output weight matrix β is:

$$\beta = \left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Y}, \tag{2.4}$$

where $\mathbf{I}$ is an identity matrix of the same size as $\mathbf{H}$. This regularized ELM will be called rELM in the text and experiments.

ELM universal approximation has been proven [76], however model selection in ELM has been an active research area, where some incremental strategies [77, 79, 78, 33] provide efficient near-optimal hidden layer unit selection. Also, ensembles of ELM have been reported [24] providing enhanced results. Robust estimation of the output weights for situations where the pseudo-inverse is singular is proposed in [75].

2.2.2 OP-ELM

The Optimally Pruned ELM (OP-ELM) [129] aims to cope with variables irrelevant to the problem at hand. The OP-ELM follows a three-step methodology:

1. Construct an SLFN using ELM.

2. Rank the best neurons using LARS algorithm. This process is akin to a "regularization" of the ELM. It uses Allen's PRESS [46] formula to L_1 regularize the ELM,

$$\varepsilon^{\text{PRESS}} = \frac{y_i - h_i b_i}{1 - h_i P h_i^T}, \tag{2.5}$$

   where $P = \left(H^T H\right)^{-1}$.

3. Select the optimal number of neurons using Leave-One-Out (LOO) error versus the number of neurons.

The LOO error evaluation requires to train the model on the whole data set except one sample for all the samples of the data. However, in the OP-ELM the relation between the hidden and the output layers is linear, hence the LOO error has a closed matrix form [46] allowing a fast computation of the MSE, and therefore the computation of the output weights is still computationally fast, and theoretically more robust than the original ELM to correlated variables.

2.2.3 Ensembles of ELMs

An integration of several ELMs was proposed by Sun et al [162] to predict the future sales amount, where several ELMs were connected in parallel and the average of the ELMs outputs was used as the final predicted sales amount. The empirical studies showed that the adaptive ensemble model achieved better generalization performance. The Voting ELM (V-ELM) [24] is a direct composition by majority voting of a collection of ELMs trained independently. The ensemble approach is shown to enhance the stability of ELM results, by approaching the limit perfect classification performance when the number of training samples grows sufficiently. The V-ELM has been empirically proven to improve over ELM in a number of experiments, and it is one of the ensemble architectures tested in this paper. Recent works explore the effect of output dissimilarity based model selection [91]. The individual ELMs are discarded from the final ensemble if their output is too similar to other ELMs in the ensemble. Ensembles of optimally pruned ELMs have been recently applied to an automotive control problem [132]. The approach includes the use of negative correlation to enforce diversity in the selected individual classifiers.

The ensembles of Online Sequential ELMs (EOS-ELM) [102] are intended to provide stability and reproducibility of results. The approach is applied to regression problems, combining the output of the individual OS-ELM regressors by their average. The underlying justification is that the EOS-ELM is closer to the expected values of the true output than the individual OS-ELM. Model selection is performed empirically. It is claimed that it trains faster and with better accuracy. The addition of a forgetting mechanism improves the EOS-ELM in [196] in situations where the lifetime of the data is limited. Such situations happen in the financial market applications, and similar prediction tasks. The FOS-ELM discards outdated data samples in the training process by filtering process that weights the prediction value of each datum. EOS-ELM and FOS-ELM approaches are better suited for dealing with non-stationary time series than image processing applications. Evolutionary algorithms have been proposed to improve the stability of learning ensembles of ELMs [177, 186]. The evolutionary algorithm is guided by the diversity of classifier outputs as the fitness function. The approach is computationally very heavy as it involves the training of a large number of ELM ensembles in order to produce the desired evolution, with the advantage of providing a tuned ensemble size which can be of interest in some specific applications.

2.2.4 Voting ELM

The mechanism of V-ELM [24] is as follows: Suppose that we build a set of K independent ELMs on the given training data. Then for each testing sample $\mathbf{x}^{test}$,

K class predictions can be obtained from the independent ELMs. The results are stored in a vector $S_{K,x^{test}} \in \mathbb{N}_0^C$, where C is the number of classes and $\mathbb{N}_0$ the natural numbers plus the zero value. Let $y_k(\mathbf{x}^{test}) \in \{1,\ldots,C\}$ denote the output of the k-th ELM, then the values of the components of $S_{K,x^{test}}$ are computed as follows:

$$S_{K,x^{test}}(i) = \left|\left\{k \left| y_k\left(\mathbf{x}^{test}\right) = i\right.\right\}\right|,$$

so that each component counts the number of ELMs giving the corresponding class as output. The final class label is computed as follows:

$$c^{test} = \arg \max_{i=\in\{1,\ldots,C\}} \left\{S_{K,x^{test}}(i)\right\}.$$

It was shown in [24] that that if the correct class prediction by an ELM has the greatest probability, then the V-ELM will produce the correct class prediction with probability one provided a large enough ensemble. In this paper we have tested the same schema using ELM, OP-ELM and rELM as the individual classifiers.

2.2.4.1 ELM regression ensembles

Ensembles of ELM, such as the the Voting ELM [24], where the component SLFNs, each with the same number of hidden neurons, are trained on bootstrapped samples of the training data. Individual SFLN outputs are combined by majority voting. Ensembles have been shown to enhance performance of single ELM classifiers. The V-ELM has already been tested in a number of applications, such as hyperspectral image classification [11], remote sensing data classification [68]and natural gas reservoir characterization [6], wastewater quality index modeling [197], and intrusion detection [56] with the enhancement of multi-kernel learning. This basic architecture has been modified in the literature, for instance soft-class dependent voting schemes [23] provide improved reliability and sparseness of the model. A distributed approach allows to perform classification in P2P networks [161], and delta test strategy for hidden units selection enhances the construction of ensembles in [190]. Proof of the expected improvement in performance relies on the assumption that the probability of correct classification is greater than any misclassification (Prop 2.1 in [24]), so that accuracy converges to 100% as the number of individual classifiers grows. In the case of ensembles of ELM regressors, we can state a similar proposition.

Proposition 1. *For an ensemble of SLFN regressors* $\left\{f_L^{(1)},\ldots,f_L^{(K)}\right\}$ *endowed with an average output combination operator, i.e.* $F(\mathbf{x}) = \frac{1}{K}\sum_{j=1}^{K} f_L^{(j)}(\mathbf{x})$. *It is guaranteed that the regression error decreases arbitrarily with increasing number*

of regressors K if the random step of ELM training achieves a uniform sampling of feature space.

Proof. Proof of this proposition is immediate. The regression value predicted by the j-th SFLN in the ensemble is $\hat{\mathbf{t}}_T^{(j)} = \mathbf{t}_T + \xi_T^{(j)}$ where $\mathbf{t}_T$ is the true value of a testing sample $\mathbf{x}_T$, and $\xi_T^{(j)}$ is the error committed by the j-th SFLN in the ensemble. Taken together, the errors $\left\{\xi_T^{(j)}\right\}_{j=1}^{K}$ are i.i.d. random variables, which are the residuals of the least squares solution of a linear problem, corresponding to a general linear model. They are independent as far as the separate ELM training of individual SLFN are statistically independent process. To ensure this property, they are trained on independent bootstrapped training sets, under the assumption that the random generator underlying the whole process generates truly statistically independent random number values. The only requirement for these residuals to have zero mean is that the projections of the regressors $\mathbf{H}^{(j)}\beta$ follow a distribution centered in the true values. Without any prior information on the target distribution, ensuring a uniform distribution of the feature space provides the best approximation to get the desired zero mean residuals [42]. The estimation provided by the ensemble is therefore of the form:

$$\hat{\mathbf{t}}_T = F(\mathbf{x}_T) = \mathbf{t}_T + \frac{1}{K}\sum_{j=1}^{K} \xi_T^{(j)},$$

whose second term becomes zero in the limit

$$\lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{K} \xi_T^{(j)} = 0,$$

hence the ensemble estimation converges to the true value as the number of ELM regressors increases. □

According to this proposition, we expect that the regression results of ELM multivariate regressor ensembles will be improving single ELMs. However, the searching for the adequate ensemble size must be done empirically.

2.3 Hybrid Extreme Rotation Forest

The Hybrid Extreme Rotation Forest (HERF) algorithm for train and test is summarized in Algorithm 2.2. Let $\mathbf{x} = [x_1, \ldots, x_n]^T$ be a sample described by n feature variables, F is the feature variable set and X is the data set containing N training samples in a matrix of size $N \times n$. Let Y be a vector containing the class labels of the data samples, $Y = [y_1, \ldots, y_N]^T$. Denote by $D_1, \ldots, D_L$ the classifiers in the

ensemble and by. As with most ensemble methods, we need to pick L in advance. All classifiers can be trained in parallel, which is also the case with Bagging and Random Forests. To construct the training set for classifier D_i, we carry out the following steps:

- Partition the set of feature variables F into K subsets of variables.

- For each subset of feature variables,

  - extract the data from the training data set

  - compute the rotation matrix using Quartimax or Principal Component Analysis

- Compose the global rotation matrix reordering columns according to the original data.

- Transform the train and test data applying the same rotation matrix.

Algorithm 2.2 Hybrid Extreme Rotation Forest (HERF)

Training Phase

Given

X : training data set (N x n matrix)

Y : the labels of the training set (N x *1* matrix)

L : the number of classifiers in the ensemble

K : the number of subsets

For $i = 1 \ldots L$

Computation of rotation matrix R_i^{α}:

Partition F (the feature variable set) into K random subsets:

$$F_{i,j}; j = 1 \ldots K$$

For $j = 1 \ldots K$

- Let $X_{i,j}$ be the data set X for features in $F_{i,j}$.
- $X_{i,j}^{'}$ equal to 75% bootstrap sample from $X_{i,j}$ /optional step/
- $C_{i,j}$ obtained from PCA or quartimax on $X_{i,j}^{'}$ /optional step/

Compose $R_{i,j}^{\alpha}$ using matrices $C_{i,j}$.

Decide ifD_i is a tree or an ELM

Train classifier D_i on training set (XR_i^{α}, Y).

Classification Phase

For a given $\mathbf{x}^{test}$,

$d_i = D_i(x^{test} R_i^{\alpha})$

$c^{test} = \max_i \{d_i, i = 1, \ldots, L\}$

The algorithm has been presented in [**?**], where comparative results with state-of-the-art classifiers did show that approach is competitive both in terms of classification performance and time efficiency.

2.3.1 Data Rotation

Denote by $\mathbf{x} = [x_1, \ldots, x_n]^T$ an n-component input random vector. Denote by

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^N \\ \vdots & \ddots & \vdots \\ x_n^1 & \ldots & x_n^N \end{pmatrix}$$

a matrix whose columns are samples of $\mathbf{x}$, that is, x_i^j is an entry of this matrix, where $j = 1, ...N$ is the sample number. The i-th row of X, denoted by

$$X_i = [x_i^1, x_i^2, \ldots, x_i^N]$$

is called the sample vector of the i-th feature variable F_i. Rotations are matrix transformations of the input matrix X. The most popular is the Principal Component Analysis. Below we describe the Quartimax, which is a little less known transformation.

Quartimax In statistics, a varimax rotation [92] is a change of coordinates that maximizes the sum of the variances of the squared coefficients after the transformation. That is, it seeks a basis that is the most economical representation of each individual sample in terms of the latent variables. The rotation matrix is found solving the following optimization problem:

$$R_v = \arg\max_R \left(\sum_{j=1}^{k} \sum_{i=1}^{p} (\Lambda R)_{ij}^4 - \frac{\delta}{p} \sum_{j=1}^{k} (\sum_{i=1}^{p} (\Lambda R)_{ij}^2)^2 \right), \tag{2.6}$$

where $\delta = 0$ for Quartimax, so that

$$R_Q = \arg\max_R \left(\sum_{j=1}^{k} \sum_{i=1}^{p} (\Lambda R)_{ij}^4 \right), \tag{2.7}$$

Specifically, the Quartimax rotation maximizes the variance of the squared transformation coefficients simplifying the the rows of the data transformatin matrix. In each variable the large coefficients are increased and the small ones are decreased so that the number of significative coefficients is minimized.

2.4 Anticipative Hybrid Extreme Rotation Forest

The Anticipative Hybrid Extreme Rotation Forest (AHERF) algorithm training and testing phases are summarized in Algorithm 2.3. We specify the training and test phases of each cross-validation fold. For training, first, a model selection phase is performed, where 30% of the training data is used. For each classifier type described in the previous section, a 5-fold cross-validation is performed on the selected data (line M3). The resulting average accuracies are ranked, so that r_k is the ranking value of the k-th classifier type (line M4). Then (line M5), each classifier is assigned a selection probability according to the expression $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, where $Fib(i)$ is the i-th value of the Fibonacci series. Figure 2.1 shows the plot of the probability distribution computed from the ranking of the classifier types.

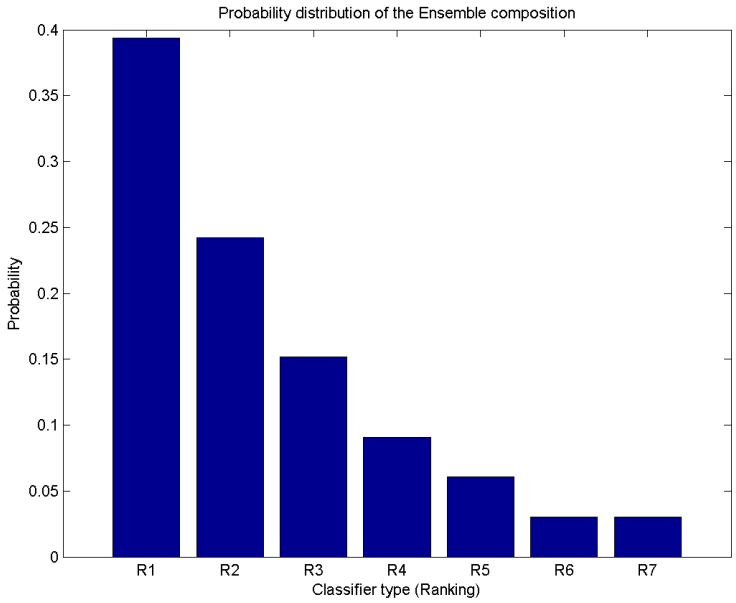


Figure 2.1: The probability distribution from the ranking of the classifiers.

In the training phase, for each classifier D_i in the ensemble, the first step in its training is the construction of the randomized rotation matrix (line 3) which requires the random partition of the set of features into a K subsets (line 4). For each subset of features $F_{i,j}$, the algorithm extracts the corresponding sample values in a matrix $X_{i,j}$ (line 6), used to build a component $C_{i,j}$ rotation matrix (line 7). The randomized rotation matrix R_i^{α} is built by composing the component rotation matrices reordering the columns in order to match the original variable ordering. Next, (line 9), there is a random decision on the type of the classifier, using the selection probabilities $\{p_k\}$ (built in line M5). Finally, the D_i classifier is trained on the rotated data. In the test phase, a new vector $\mathbf{x}^{test}$ is first applied each classifier in the ensemble, obtaining a class hypothesis d_i, (line C2). Majority voting is implemented as follows: the counter c_ω has the number of classifiers that have casted their vote for class ω, (line 3, where $\delta_{i,j}$ is the Kronecker's delta function). Finally, the class with the maximum votes is selected (line C4) and returned as the classification result.

Algorithm 2.3 Anticipative tuning of Hybrid Extreme Rotation Forest (AHERF)

Training Phase

Given

X : z-scores of input dataset ($n \times N$ matrix).

Y : the labels of the dataset ($1 \times N$ matrix)

L : the number of classifiers in the ensemble

K : the number of feature subsets

Begin

Anticipative Model selection

M1 Select 30% of the dataset for model selection

M2 For each classifier type $k = 1, \ldots, M$

M3 Perform 5-fold cross-validation, obtain accuracy A_k

M4 Rank A_k, assigning r_k to the k-th classifier

M5 Assign selection probability $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, $k = 1, \ldots, M$

On the 70% unused data, perform 10-fold cv, at each fold:

Ensemble construction on each training fold

2 For each individual classifier D_i, $i = 1 \ldots L$

3 Computation of rotation matrix R_i^{α}:

4 Partition F into K random subsets: $F_{i,j}; j = 1 \ldots K$

5 For each F_{ij}, $j = 1 \ldots K$

6 - Let $X_{i,j}$ be the subset of X corresponding to features in $F_{i,j}$.

7 - $C_{i,j}$ obtained from PCA on $X_{i,j}$

8 Compose R_i^{α} using matrices $C_{i,j}$.

9 Decide the model of D_i sampling $\{p_k; k = 1, \ldots, M\}$

10 Train classifier D_i on training set $(R_i^{\alpha} X, Y)$ or (X, Y)

End ensemble construction

Test on each testing fold

Let Ω be number of classes

C1 For each unknown $\mathbf{x}^{test}$ z-scores.

C2 $d_i = D_i(R_i^{\alpha} \mathbf{x}^{test}); i = 1, \ldots, L$

C3 $c_{\omega} = \sum_{i=1}^{L} \delta_{d_i,\omega}; i = 1, \ldots, L$

C4 $c^{test} = \arg\max_{\omega} \{c_{\omega}, \omega = 1, \ldots, \Omega\}$

2.4.1 Rationale for AHERF

The work on the design of heterogenous ensembles of classifiers is motivated by the well known no-free lunch theorems [182], which state that no single approach is optimal either for the solution of difficult optimization problems, as well as for machine learning solution of classification and regression problems. Therefore, we

would like to predict which kind of classifier building approach is optimal for the problem domain at hand. If we have no hint, we may try building a heterogeneous ensemble so that we may have some chance of applying the best approach from the pool of available approaches.

The driving idea in AHERF is to build an ensemble where the best fitted classifier types are more frequent. Suppose we have a problem domain characterized by a ground truth classification mapping $\mathscr{C} : \mathscr{X} \to \Omega$, that gives the true class $\omega \in \Omega$ corresponding to each input feature vector $\mathbf{x} \in \mathscr{X}$. Any classifier ${}^tC$ that we may built from a collection of input/output patterns $X = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^N$, where $t \in T$ denotes the classifier type from a collection of methods T, i.e. the classification building type, provides us with its best estimation of the true class $\hat{\omega} =^t C(\mathbf{x})$. We can safely say that this estimation is given somehow as a maximum *a posteriori* estimation, i.e. $\hat{\omega} = \max_{\omega} {}^t\hat{P}(\omega|\mathbf{x})$, where t denotes the type of classifier, and $\left\{{}^t\hat{P}(\omega|\mathbf{x})\right\}_{\omega\in\Omega}$ denotes the data driven estimation of the *a posteriori* probabilities by classifier ${}^tC$. The accuracy of a classifier can be computed as the expectation of the distance between the a posteriori distribution and the ground truth classification: ${}^tA = E_{\mathscr{X}}\left[\left\|\left[{}^t\hat{P}(\omega|\mathbf{x}) - \mathscr{C}(\omega,\mathbf{x})\right]_{\omega}\right\|\right]$, where $E_{\mathscr{X}}[.]$ denotes the expectation over the input space, i.e. over all possible sampling processes providing the training dataset X, and $\mathscr{C}(\omega,\mathbf{x})$ is 1 for the true class, and 0 for the others. The cross-validation experiments are a minimum variance way to provide estimates of the accuracy.

If we have an ensemble of classifiers $\{{}^tC_k\}_{k=1}^M$, then we will have as many *a posteriori* distribution estimations $\left\{\left\{{}^t\hat{P}_k(\omega|\mathbf{x})\right\}_{\omega}\right\}_{k=1}^M$ as classifiers. If the ensemble decision is by majority voting, such as in AHERF, then the ensemble class estimation is given by $\hat{\omega} = \arg\max_{\omega} |\{k\,|\omega = \hat{\omega}_k\}|$, where $\hat{\omega}_k = \max_{\omega} {}^t\hat{P}_k(\omega|\mathbf{x})$. In a broad sense, we can say that the accuracy of the ensemble can be modeled by $A_M \propto E_{\mathscr{X}}\left[\sum_k \left\|\left[{}^t\hat{P}_k(\omega|\mathbf{x}) - \mathscr{C}(\omega,\mathbf{x})\right]_{\omega}\right\|\right]$ in the sense that the increase in closeness of the *a posteriori* distributions of the ensemble constituents to the ground truth will always reflect in an increase in accuracy. It is immediate that $A_M \propto \sum_{k=1}^M ({}^tA_k)$.

Let us assume that there is some accuracy ranking of the classifier types, so that ${}^{t_1}A > {}^{t_2}A > {}^{t_3}A > ....$ Let us denote by n_t the number of ensemble constituents of type t, so that an ensemble is characterized by the vector $\mathbf{n} = [n_t\,|t \in T^*]$, where T^* denotes the identifiers of the classifiers types ordered by accuracy ranking. Then it is immediate that for two ensembles such that $\mathbf{n}' > \mathbf{n}''$ according the lexicographic ordering, i.e. the classifier with the best ranking is more frequent, then the first ensemble will very likely have accuracy greater than the second. The strategy of AHERF is to estimate via cross-validation on a small dataset the classifier type ranking $\widehat{{}^{t_1}A} > \widehat{{}^{t_2}A} > \widehat{{}^{t_3}A} > ...$, using this information to drive the selection of the

classifier type of each individual constituent. In order to have ensembles whose characteristic vector **n** is of the form $n_{t_1} >> n_{t_2} >> n_{t_3} > ...$ we sample an integer random variable whose distribution of probability is an approximation of the exponential distribution built using the Fibbonacci series on the ranking. The anticipatory character of AHERF comes from this prediction of the appropriate distribution of classifier types before building and training the ensemble. This stochastic sampling of the classifier type (with replacement) allows for some flexibility (i.e. when the difference in performance is not so great we would like to have some less sharp difference in number of classifiers, and viceversa) and avoids the need to set specific settings.

2.5 Bootstrapped Dendritic Classifiers

Algorithm 2.4 Crossvalidation scheme for the training of the BDC

Let be $X = \{x_1, \ldots, x_n\}$ input data $x_i \in \mathbb{R}^d$, and $Y = \{y_1, \ldots, y_n\}$ the input data class labels $y_i \in \{0, 1\}$.
K is the number of DC classifiers

for i=1:10 (cross-validation folds)

1. select disjoint train $X^e = \left\{x^e_{i_1}, \ldots, x^e_{i_{n-n/10}}\right\} \subset X$, $Y^e = \left\{y^e_{i_1}, \ldots, y^e_{i_{n-n/10}}\right\} \subset Y$ and test $X^t = \left\{x^t_{i_1}, \ldots, x^t_{i_{n/10}}\right\} \subset X$, $Y^t = \left\{y^t_{i_1}, \ldots, y^t_{i_{n/10}}\right\} \subset Y$ datasets .
2. For $j = 1 : K$ (construct of classifiers)
   (a) Bootstrap a train dataset $X^{eb} = \left\{x^{eb}_{i_1}, \ldots, x^{eb}_{i_{n-2n/10}}\right\} \subset X^e$, $Y^{eb} = \left\{y^{eb}_{i_1}, \ldots, y^{eb}_{i_{n-2n/10}}\right\} \subset Y^e$. Out-of-bag error may be computed on the remaining training data and test $X^e - X^{eb}$, $Y^e - Y^{eb}$, disjunctions.
   (b) Apply DC to train classifier $C_j : \mathbb{R}^d \to \{0, 1\}$ on $\left(X^{eb}, Y^{eb}\right)$.
3. end for. Optionally compute out of bag error
4. Crossvalidation test, For each $x \in X^t$
   (a) compute $C_1(x), \ldots, C_N(x)$
   (b) Majority voting, class $y = 0$ if $\left|\left\{j \left| C_j(x) = 0\right.\right\}\right| > \left|\left\{j \left| C_j(x) = 1\right.\right\}\right|$
5. compute accuracy, sensitivity and specificity statistics

In this section we introduce an ensemble classifier whose elementary classifiers are specific instances of little known kind of architectures based on Lattice Computing. This work was developed in collaboration with Darya Chyzhyk for an application to medical image analysis described in following chapters. We start by introducing the elementary classifiers.

A single layer morphological neuron endowed with dendrite computation based on lattice algebra was introduced in [145]. The response of the j-th dendrite is as follows:

$$\tau_j(x_i) = p_j \bigwedge_{k \in I_j} \bigwedge_{l \in L_{kj}} (-1)^{1-l} \left(x_{i,k} + w_{kj}^l \right), \tag{2.8}$$

where $l \in L_{kj} \subseteq \{0,1\}$ identifies the existence and inhibitory/excitatory character of the weight, $L_{ij} = \emptyset$ means that there is no synapse from the i-th input neuron to the j-th dendrite; $p_j \in \{-1,1\}$ encodes the inhibitory/excitatory response of the dendrite. The total response of the neuron is given by:

$$\tau(x_i) = f\left(\bigwedge_{j=1}^{J} \tau_j(x_i) \right), \tag{2.9}$$

where $f(x)$ is the Heaviside hard-limiter function. A constructive algorithm [145] obtains perfect classification of the train dataset using J dendrites.

The Bootstrapped Dendritic Classifiers (BDC) is a collection of DC classifiers,

$$C(x; \psi_j), j = 1, ..., N, \tag{2.10}$$

where ψ_j are independent identically distributed random vectors whose nature depends on their use in the classifier construction. Each DC classifier casts a unit vote for the most popular class of input x. Given a dataset of n samples, a bootstrapped training dataset is used to train a DC $C(x; \psi_j)$. The independent identically distributed random vectors ψ_j determine the result of bootstrapping. In conventional RF they also determine the subset of data dimensions $\hat{d}$ such that $\hat{d} \ll d$ on which each tree is grown. However, in this thesis work we are not dealing with this kind of DC randomization, which may be studied elsewhere. The main parameters for the experimental evaluation of the BDC are the number of trees and the maximum depth of each DC given by the maximum number of dendrites allowed. Limiting the number of dendrites is a kind of regularization that weakens the classifier. Finally, Algorithm 2.4 specifies the cross-validation scheme applied in the experiments.

2.6 Active Learning fundamentals

The performance of supervised classifiers strongly depend on the information provided by the data used to train the classifier, so that the appropriate selection and labeling of the training set may be a cumbersome task requiring extensive manual inspection and analysis of the data, typically requiring some visualization tool and labeling of each data sample. Besides, noisy samples may interfere the class statistics, which may lead to poor classification performances and/or over-fitting. For these reasons, a training set must be constructed in a smart way, meaning that it must consists of the minimal set of samples allowing to compute correctly the class boundaries, therefore it must contain the most informative data samples. In the machine learning literature this approach is known as Active Learning.

Active Learning [41, 170] focuses on the interaction between the user and the classifier. Let $X = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ be a training set consisting of labeled samples, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1,\ldots,N\}$. Let be $U = \{\mathbf{x}_i\}_{i=l+1}^{l+u} \in \mathbb{R}^d$ the *pool of candidates*, with $u \gg l$, corresponding to the set of unlabeled samples to be classified. The classifier would be feed with the samples $\mathbf{x}_i$ and the targets y_i in a classic classification scheme. Actively learning on the other hand involves trying to indicate which input vector should be selected from the training set, in order to improve the learning capabilities of the classifier. We can consider that, for a given learning task whose target function is f, there are some areas where the function is more easily learned and some that are more difficult to classify. Measuring how difficult a sample is to classify is not trivial, and is discussed below. The Active Learning approach tries to focus on those difficult to classify regions -similarly to boosting techniques- but interactively.

Generally, an Active Learning process can be summarized as follows: In a given iteration t, the Active Learning algorithm selects from the pool U^t the q candidates that will, at the same time, maximize the gain in performance and reduce the uncertainty of the classification model when added to the current training set X^t. The selected samples $S^t = \{\mathbf{x}_m\}_{m=1}^{q} \subset U$ are labeled with labels $\{y_m\}_{m=1}^{q}$ by an oracle, which can be a human operator in interactive segmentation, or the available ground truth when performing cross-validation experiments. Finally, the set S^t is added to the current training set $(X^{t+1} = X^t \cup S^t)$ and removed from the pool of candidates $(U^{t+1} = U^t \backslash S^t)$. The process is iterated until a stopping criterion is met, such as the achieved accuracy reaching a preset threshold θ_{max}.

2.6.1 Classification uncertainty in ensemble classifiers

Ensemble classifiers allow a committee approach for the estimation of unlabeled sample uncertainty [170]: assume that we have built a committee of k base clas-

sifiers, i.e. a RF with k trees. The output of the committee members provide k labels for each candidate sample $\mathbf{x}_i \in U$. The data sample class label is provided by the majority voting. Our heuristic is that the standard deviation $\sigma(\mathbf{x}_i)$ of the class labels is the measure of the classification uncertainty of $\mathbf{x}_i$. Let us consider an ordering of the pool of candidates $U^* = \left\{\mathbf{x}_{j_i}\right\}_{i=l+1}^{l+u}$, where $\sigma(\mathbf{x}_{j_i}) > \sigma\left(\mathbf{x}_{j_{i+1}}\right)$. The *standard deviation query-by-bagging* heuristic selection of samples to be added to the train set is stated as the following selection:

$$S^t = \left\{\mathbf{x}_{j_m}\right\}_{m=1}^{q} \tag{2.11}$$

Standard deviation of predicted class labels is a natural multiclass heuristic measure of classification uncertainty. A candidate sample for which all the classifiers in the committee agree has a zero prediction standard deviation, thus its inclusion in the training set does not bring additional information. In other words, belongs to an "easy" to classify region of the data. On the contrary, a candidate with maximum disagreement between the classifiers results in maximum standard deviation, so it is "difficult" to classify it. Therefore, its inclusion will be highly beneficial.

Chapter 3

Applications on Hyperspectral Data

This Chapter refers two main applications of ensemble classifiers to hyperspectral image processing. The first is the decomposition of pixel spectra into constituent elementary endmember spectra, called spectral unmixing or subpixel resolution segmentation in the literature referred in Section 3.1. The second is the classification of image pixels to obtain thematic maps. In this task we have developed innovative semi-supervised learning and spatial correction pipeline combined with contributions in ensembles of randomized classifiers, whose general schema is reported in section 3.2.

3.1 Nonlinear Unmixing and Reconstruction

The contribution of this Thesis section is the application of ELM regression ensembles to the construction of the direct and inverse unmixing transformation, based on a training set of pixels whose fractional abundance coefficients are provided by the user. The direct transformation provides the fractional abundances of the pure materials given a pixel spectrum. The inverse transformation provides a reconstruction of the pixel spectrum given an abundance vector. Validation is based on the computation of the reconstruction error over the entire dataset, which is a much larger independent testing dataset than the labeled pixel set, not requiring full or exact labeling of the ground truth. Therefore, the training set can be entirely devoted to model construction. Besides, we have introduced a formal discussion about the convergence of ELM regression ensembles, which justifies the improvements found empirically.

3.1.1 Linear unmixing

We provide in this section some background on the classical linear unmixing approaches, with details about the benchmarking algorithms used for comparison with the proposed approach. We denote $X = \{\mathbf{x}_i\}_{i=1}^{N\times M}$ the collection of pixels in a hyperspectral image, $\mathbf{x}_i \in \mathbb{R}^B$, where B is the number of bands in the hyperspectral image, and $N \times M$ is the total number of pixels in the image. Given a set of endmembers aggregated in a matrix $E = [\mathbf{e}_1, \ldots, \mathbf{e}_d]$, where $d \ll B$ is the number of endmembers, the linear mixing model assumes that each pixel can be expressed as a linear combination of the endmembers:

$$\mathbf{x}_i = \mathbf{a}_i E, \tag{3.1}$$

where $\mathbf{a}_i \in [0,1]^d$ is the vector of fractional abundance coefficients, such that $\sum_{j=1}^{d} a_{ij} = 1$, which can be estimated by computing the inverse of the linear mixing model by least squares:

$$\hat{\mathbf{a}}_i = E^{+}\mathbf{x}_i, \tag{3.2}$$

where E^{+} denotes the Moore-Penrose pseudo-inverse $E^{+} = \left(E^T E\right)^{-1} E^T$. Though this is the most straightforward solution, this estimation of the fractional abundances often breaks the sum to one and the non-negativity constraints. However, we will use it since our performance measure is the reconstruction error:

$$\xi_i^2 = \left\|\mathbf{x}_i - \hat{\mathbf{a}}_i E\right\|,$$

which is a measure of the quality of the unmixing performed that is independent of the training set. Comparisons are done on the relative error:

$$\xi_{REL}^2 = \frac{1}{NM}\sum_{i=1}^{NM}\frac{\left\|\mathbf{x}_i - \hat{\mathbf{a}}_i E\right\|}{\left\|\mathbf{x}_i\right\|}.$$

The linear unmixing process [97] needs some procedure to provide the set of endmembers. It can be a selection from some spectral library, but it is often preferred to extract the endmembers from the actual image data. We have selected three classical endmember extraction algorithms for comparison with the ELM regression ensemble approach proposed below.

Automatic Target Generation Process (ATGP)[28] The ATGP starts selecting the pixel spectrum with maximum norm, denoted $\mathbf{e}_0$. Then an orthogonal subspace projector $P_{\mathbf{U}}^{\perp} = \mathbf{I} - \mathbf{U}\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T$ with $\mathbf{U} = [\mathbf{e}_0]$ is applied to each spectra $\mathbf{x}$ in the training data. The pixel signature with the maximal projection is the next end-

member, i.e. $\mathbf{e}_1 = \arg\max_{\mathbf{x}} \left\{ \mathbf{x} P_{\mathbf{U}}^{\perp} \right\}$. The procedure is repeated recursively adding the new endmember to the projection matrix, i.e. $U = [\mathbf{e}_0, \mathbf{e}_1]$.

Fast Pixel Purity Index (FPPI) [28] The Pixel Purity Index (PPI) is based on the idea that endmembers must repeatedly appear as extrema in random projections of the data, after an initial dimension reduction by the Minimum Noise Fraction (MNF) algorithm. Random direction vectors used to compute the projections are called skewers. The PPI algorithm repeatedly generates skewers and counts the number of times each pixel spectrum projection appears as an extremum point of the collection of pixel spectra projections. Endmembers are selected as the pixel spectra whose count is above some threshold. The FPPI algorithm initially uses as skewers the target spectra found by ATGP. The set of endmembers is incrementally built by adding the new endmembers to the set of skewers until there is no new endmember added to the set of skewers.

N-FINDR [181] The N-FINDR algorithm searches the maximum volume simplex whose vertices are pixel spectra from the training data, i.e. assumes that greater candidate endmember purity corresponds to greater simplex volume. The algorithm carries out a dimensionality reduction by MNF or other suitable method. Initially, a random set of endmembers $E^{(0)} = \left\{ \mathbf{e}_1^{(0)}, \ldots, \mathbf{e}_d^{(0)} \right\}$ is generated by random selection in the training data. The volume enclosed by the simplex defined by a set E is given by

$$V(E) = \frac{\left| \det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_d \end{bmatrix} \right|}{(d-1)!}.$$

For a vector $\mathbf{r}$ in the training data, denote $E(\mathbf{r}, i)$ the set E where r substitutes the i-th endmember in E. The algorithm performs the following iterations: For each $\mathbf{r}$ in the training data, compute $\left\{ V\left(E^{(k)}(\mathbf{r}, i)\right) \right\}_{i=1}^{d}$. If there is no position where $V\left(E^{(k)}(\mathbf{r}, i)\right) > V\left(E^{(k)}\right)$, then $\mathbf{r}$ is not included in the set of endmembers. Otherwise it is included in the position providing maximum volume. Iteration number k is increased at each vector trial. The process is a greedy search, thus providing a local optimum, which is affected by the initial set of endmembers and the sequence of vectors selected for inclusion. Nevertheless, the algorithm has shown good performance in practice.

3.1.2 Non linear unmixing

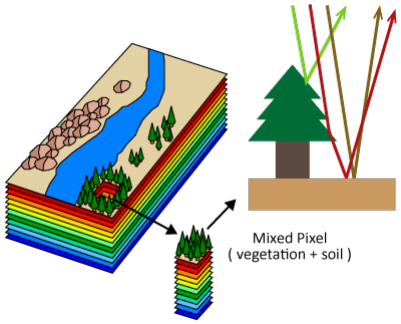


Figure 3.1: Illustration of a physical process that may account for a non-linear spectral mixing at pixel

Figure 3.1 illustrates non-linear mixing process at a given pixel. Reflections and scattering processes produce such nonlinear mixing of material reflectances. In figure 3.1, the ray that goes through the canopy to the soil and then to the sensor suffers this kind of effects. The basic non-linear model accounting for second-order interactions between spectra is the bi-linear model [21]. A more sophisticated approach, still preserving the role of the endmembers as the extrema of the convex polytope enclosing the image data, is the Geodesic Simplex Projection [21] which models the data as a manifold embedded in the simplex defined by the endmembers, computing the abundances as the projection distance into this manifold. The selection of landmark points lying on this manifold helps to reduce the computational burden imposed by the search of the nearest point for projection on the manifold [36, 35]. Other approaches use the kernel trick to model the non-linear manifold of the data [31], proposed as non-linear fluctuations taking into account non-linear interactions among the endmembers. The specification of multiple kernels improve the flexibility of the approach allowing class-dependent tuning of the parameters [65].

The application of Markov Random Field (MRF) modeling provides a graceful way to define Bayesian algorithms to perform non-linear unmixing. For instance, [3] includes a non-linear term in the mixing equation, whose spatial distribution is analyzed by MRF modeling to perform non-linear detection and unmixing. In [4], a second order polynomial on the endmembers is postulated as the non-linear mixing model, and Hamiltonian Monte Carlo algorithm is applied to estimate the parameters of the model at each pixel, profiting from spatial continuity of the MRF. Spatial correlation is also considered in [32] as a regularization of parameter estimation in the non-linear unmixing algorithm based on kernel transformations.

Other approaches consider local windows in order to refine the relevance of the endmembers for each pixel decomposition [43]. These approaches are extended to the use of local NNMF approaches [52, 188].

The approach in this thesis work consists in formulating the problem of unmixing as a non-linear regression problem, where we want to extract the fractional abundances at each pixel by a non-linear ELM regression ensemble model trained over a collection of image pixels with abundances assigned according to a known ground truth. The output of the ELM regression ensemble is a collection of output units with values in the range $[0,1]$, each modeling the fractional abundance of a defined material class. To provide the training data, the easiest way is to select pixel representatives whose abundance vectors have all components equal to zero, but for the one corresponding to the actual pixel class. However, the algorithm works as well with training data labeled with mixed abundances, if the user has this information available. Robustness theoretical results given below ensure that the ELM regression ensemble converges to the true target abundance values when the ensemble size grows.

3.1.3 ELM regression ensembles for spectral unmixing

This section explains our approach to non-linear unmixing by ELM regression ensembles, as well as its validation by reconstruction. The training data is a sample $\{\mathbf{x}_i,\mathbf{a}_i\}_{i=1}^{n}$ from the actual image data, where pixel spectrum $\mathbf{x}_i \in \mathbb{R}^B$ has an associated abundance $\mathbf{a}_i \in [0,1]^d$, B is the number of bands in the image, and d is the number of material classes, equivalent to the number of endmembers in linear unmixing. Spectra in the training dataset are assumed to be spectral representations of the materials, pure or composed, wanted for detection in the image. The abundance vectors correspond to pure material abundances, i.e. all abundances are zero, but for one component equal to one. Estimating abundances is equivalent to a non-crisp classification problem where the pure class representative spectra are encoded by orthogonal binary vectors of unit norm. Therefore, we construct a non-linear map from pixel spectra into their abundances: $\Phi : \mathbb{R}^B \to [0,1]^d$, such that

$$\hat{\mathbf{a}}_i = \Phi(\mathbf{x}_i). \tag{3.3}$$

This map is obtained by training an ELM regressor ensemble, such as $F(\mathbf{x}) = \frac{1}{K}\sum_{j=1}^{K} f_L^{(j)}(\mathbf{x})$, where each individual SLFN has L hidden units. On the same training dataset we can train the inverse model $\Phi^{-1} : \mathbb{R}^d \to \mathbb{R}^B$, such that

$$\hat{\mathbf{x}}_i = \Phi^{-1}(\mathbf{a}_i) \tag{3.4}$$

is built in the same manner. One of the endemic problems in hyperspectral remote sensing is that the number of labeled samples is small, because obtaining each sample ground truth is costly. Therefore, it is desirable to use all the training data for the construction of the direct and inverse transformation. We can compose them to obtain an estimation of the reconstruction of each pixel spectra of the image:

$$\hat{\mathbf{x}}_i = \Phi^{-1}\left(\Phi\left(\mathbf{x}_i\right)\right).$$

The measure of quality of the non-linear unmixing is the relative average reconstruction error

$$\xi_{REL} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left\|\mathbf{x}_{ij} - \Phi^{-1}(\Phi(\mathbf{x}_{ij}))\right\|}{\left\|\mathbf{x}_{ij}\right\|}, \tag{3.5}$$

where $N \times M$ are the image domain dimensions. Therefore, we can use for validation the quality measure computation the entire image, regardless of the pixel labeling.

3.1.4 Experimental Results

All results in this section are averages after 10 repetitions of the experiment with random sampling of the image data.

Sensitivity experiments. The experiments carried out explore the sensitivity of the approach to its critical parameters: number of hidden units of the SLFNs, ensemble size, and training sample size. The default parameters are as follows: number of hidden units is 50, the ensemble size is 5, and training data consists of 25 pixels per class. Figure 3.2 shows the relative error results obtained when the number of hidden units changes. Plots correspond to the different benchmark images. The main conclusion is that the approach is rather insensitive to the number of hidden units, but that the image quality is very important. In fact, Indian Pines provides the worst results in all experiments, while Kennedy Space Center gives almost optimal performance. Figure 3.3 corresponds to the exploration of the effect of ensemble size. It reproduces the pattern found in the previous figure, i.e the variable under test has little effect. Figure 3.4 shows the effect of the number of samples used for training. Again, the quality of the image has a much bigger importance than the number of samples, so that Indian Pines error is well above the others. There are some variations between images, but it is on the average below 10%, which is a very good result if we take into account that we are using a very small sample *per* class to train the ELM. These results confirm also that the ELM ensembles are able to recover from the curse of randomness, ensuring stable realizations of the training despite the random setting of the input-to-hidden weights.

Salinas and Kennedy Space Center datasets provide the best results, sometimes below 1% of relative error. On the other hand, Indian Pines provides the worst results, which is consistent with the fact that this dataset has a large signal-to-noise ratio due to the early generation of the AVIRIS sensor that captured it. A general surprising result is the insensitivity of the approach to the variation in ensemble size, number of hidden neurons and, even, training data size.

Comparison with standard approaches For comparison, we have performed the following experiment: we compute state-of-the-art endmember extraction algorithms on the same training data allowed to the ELM regression ensembles. The endmembers are then used for unmixing and reconstruction of the entire image, as done in the ELM based approach. The only parameter to be explored in this case is the sample size. The results are presented in figures 3.5, 3.6, 3.7 for ATGP, FIPPI, and N-FINDR, respectively. The overall response error is several points higher than for the ELM approach. A t-test carried out on the results of the 10 repetitions, pairing benchmark image and sample size, showed that the differences are statistically significant ($p<0.00001$). Visual inspection of the plots confirms the strong improvement introduced by the ELM regression ensemble approach, which is able to estimate the non-linear mixing model from scarce training data.

Abundance images Figures 3.8, 3.9, 3.10, and 3.11 show the ground truth of the benchmark images, and abundance maps for some of the ground truth classes of the corresponding benchmark images, computed as the average of the output of the individual ELM regressors in the ensemble. For this specific visualization, the number of hidden units was fixed at 50, and the ensembles size was 5. Training data consists of 25 pixels per class. Abundances are visualized with a hot color map, where red corresponds to highest and blue to lowest values. The visual results show some variability due to several factors, such as image noise, labeling noise, and the purity of the selected training pixels. Results are much more crisp in the Salinas image than in Indian Pines, because of the higher signal to noise ratio as well as the greater purity of the pixels. Inside Salinas image, results for the vineyard and grapes classes are less crisp than for Cerlery, Broccoli, lettuce and other grown vegetables, where the approach is even capable of discriminating stages of growth (i.e. weeks of age of lettuce crops). In the Indian Pines image, the steel towers and wheat are quite easy targets, while the soybean is its early growth stages so that its signature is strongly contaminated by the soil signature, hence this class does not include clear pure pixels. The case of KSC and Botswana images is very illustrative of the power of the approach, that is able to generalize results in a quite robust way from very small samples. Both images have relatively small ground truth, though carefully prepared. In both images, water is easily discriminated, despite training

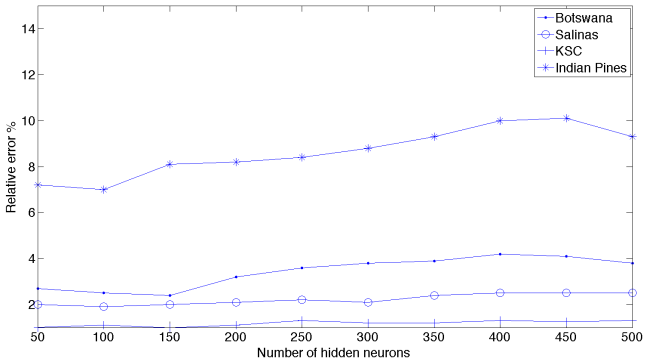


Figure 3.2: Sensitivity of the ELM ensemble reconstruction error to the number of hidden units per individual classifier, for each dataset considered in the test.

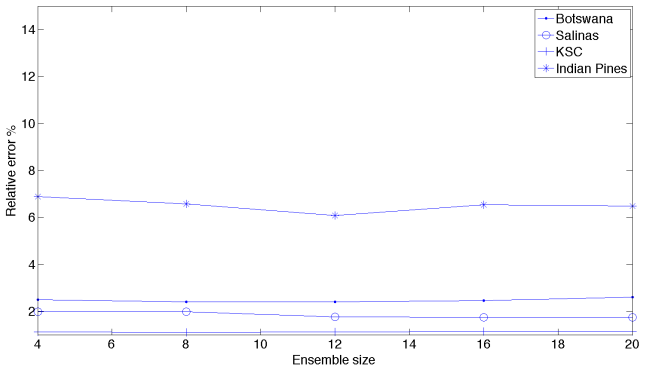


Figure 3.3: Sensitivity of the ELM ensemble reconstruction error to ensemble size, for each dataset considered in the test.

sample sizes are small, and the great variability of water reflectance due to the existence of sand banks or aquatic vegetation. In the Botswana image, firescars are detected very accurately, the acacia grasslands, hippo grass, and bare soil have good detection but with some variability due to low pixel purity, and the floodplain grasses are well detected but with less distance to other classes, due to low pixel purity. In the KSC image, diverse kind of marshes are well detected but they are less differentiated than in other classes (yelowish colors of the background). This may be due to the high content of water in all kinds of terrains. However, in all imagess, the segmentation generalization achieved is remarkable.

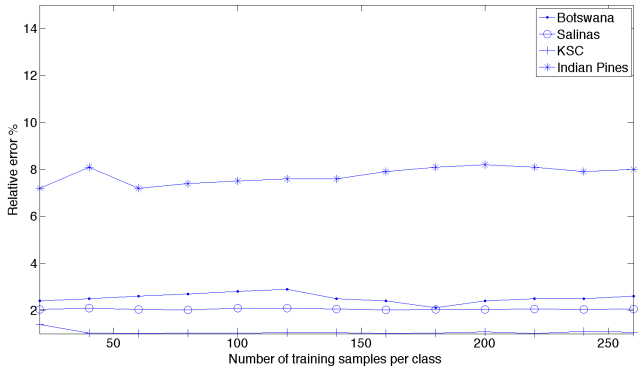


Figure 3.4: Sensitivity of ELM ensemble reconstruction error to size of training set, specified by number of data samples per class.

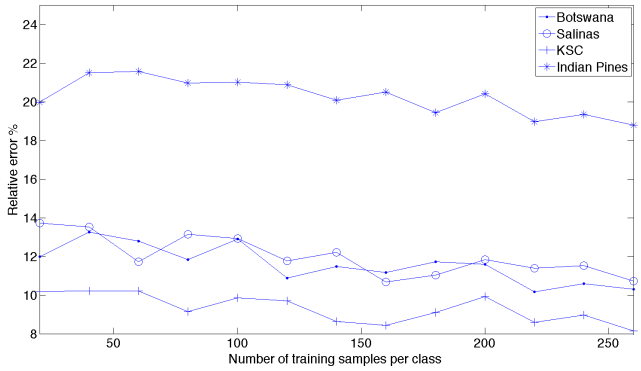


Figure 3.5: Sensitivity of ATGP reconstruction error to size of training set, specified by number of data samples per class.

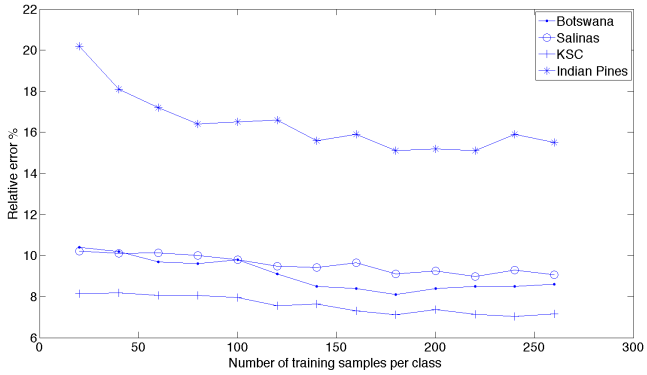


Figure 3.6: Sensitivity of FIPPI reconstruction error to size of training set, specified by number of data samples per class.

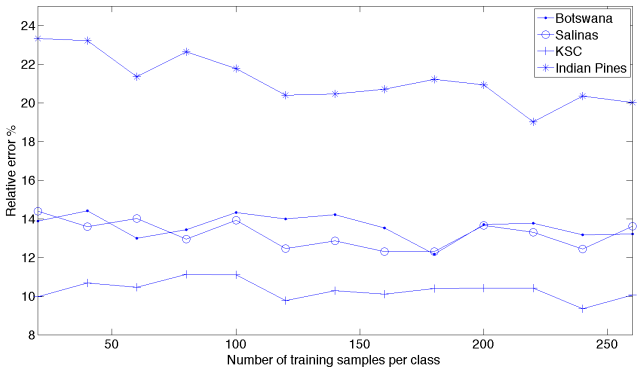


Figure 3.7: Sensitivity of N-FINDR reconstruction error to size of training set, specified by number of data samples per class.

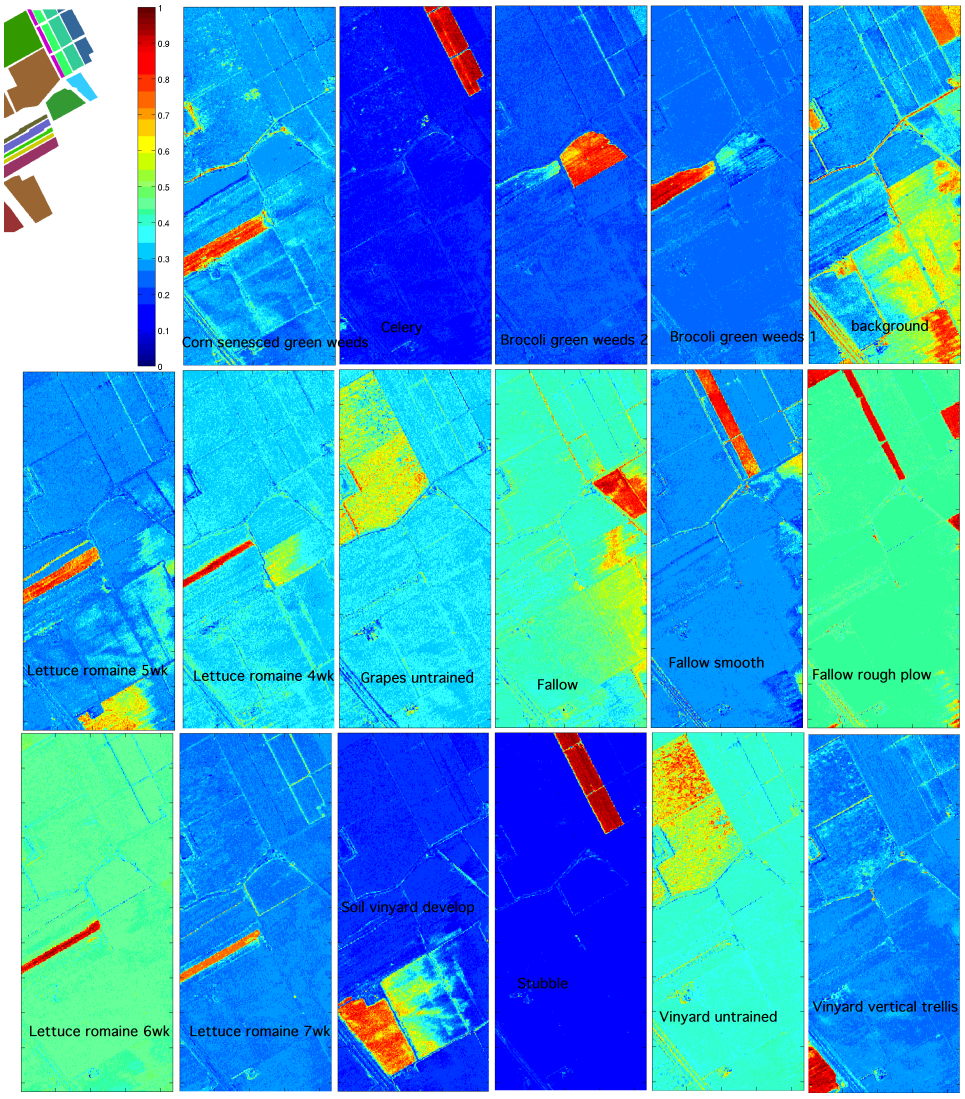


Figure 3.8: Abundances produced by ELM of some of the classes in the Salinas image. Left-up corner is the visualization of the available ground truth and the color map for the abundance visualization. Class name is printed in the figures.

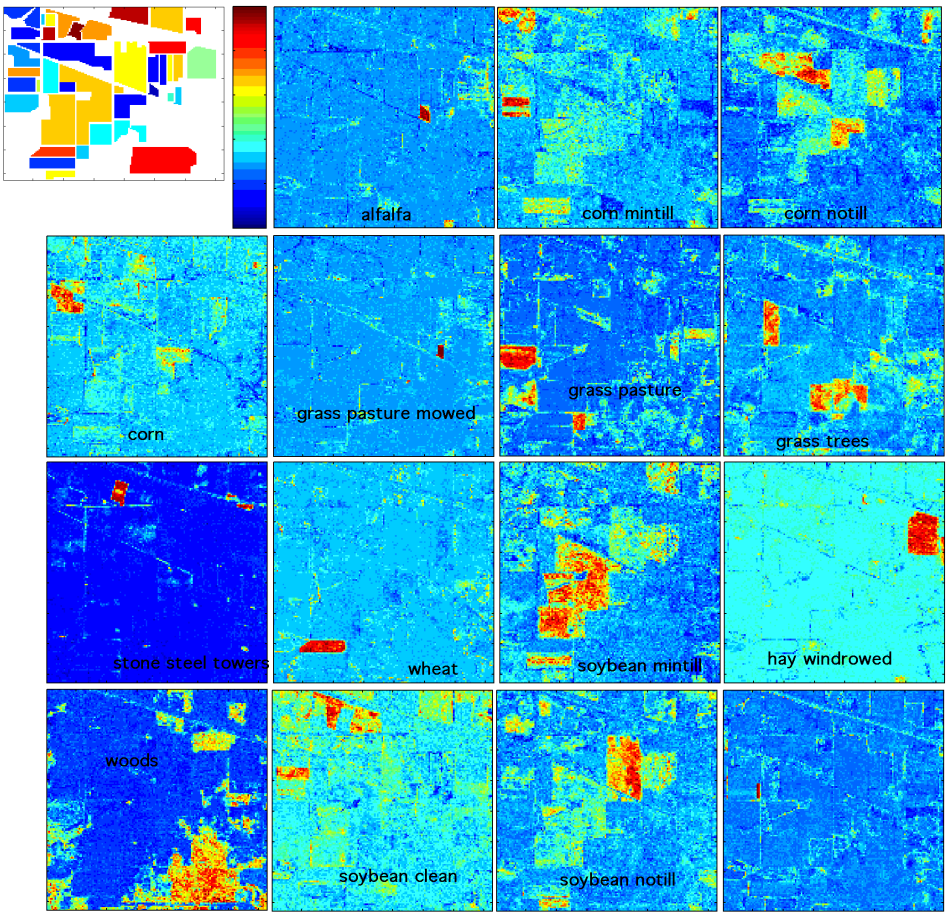


Figure 3.9: Abundances produced by ELM of some of the classes in the Indian Pines image. Left-up corner is the visualization of the available ground truth and the color map for the abundance visualization. Class name is printed in the figures.

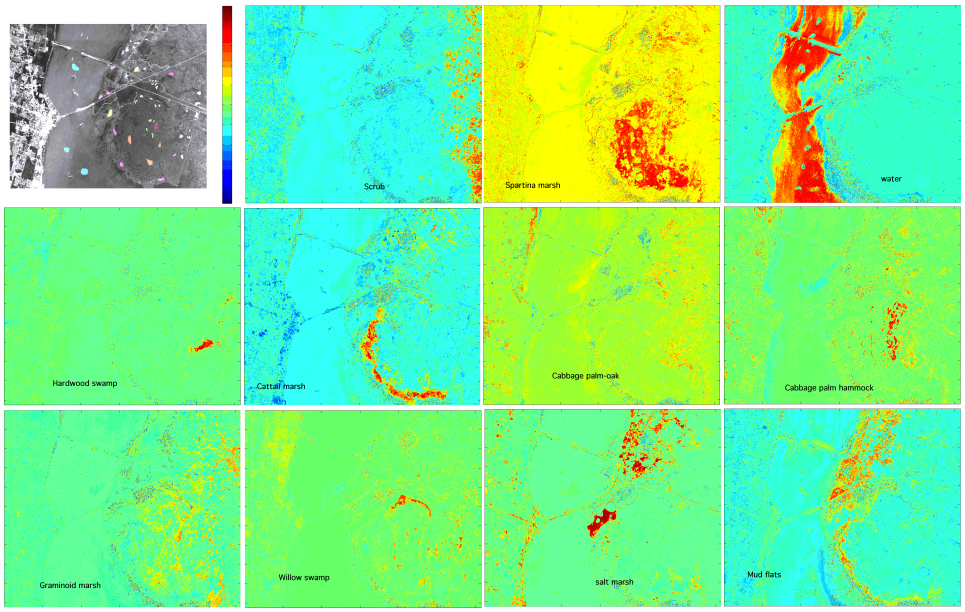


Figure 3.10: Abundances produced by ELM of some of the classes in the Kennedy Space Center image. Left-up corner is the visualization of the available ground truth and the color map for the abundance visualization. Class name is printed in the figures.

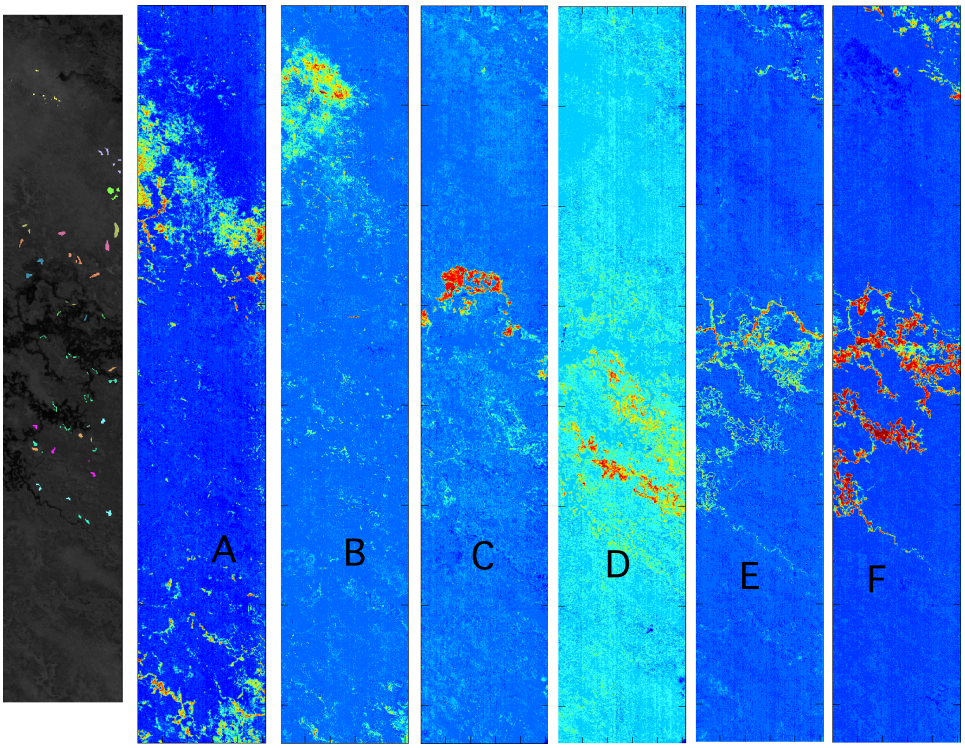


Figure 3.11: Leftmost image is the visualization of the localization of available ground truth overlaid on one the bands. Abundances produced by ELM of some of the classes in the Botwsana image. (A) acacia grasslands, (B) exposed soil, (C) firescar, (D) floodplain grasses 2, (E) Hippo grass, (F) water. The color map is the same as previous figures.

3.2 Spectral-spatial SSL with spatial correction

3.2.1 State of the art

Hyperspectral images pose strong classification challenges inherent to high dimensionality data [90], such as the well-known Hughes effect [83], also known as curse of dimensionality in machine learning literature. The Hughes effect consists in the decrease of classifier performance as data dimensionality increases, for a fixed sample size. Hyperspectral data is characterized by its difficulty to obtain ground truth labels, due to the high cost of the field work needed to obtain them, so researchers on hyperspectral image classification have been facing the Hughes effect since the very beginning of the field.

One approach to mitigate the problem of labeled data scarcity are semi-supervised

learning methods [29, 198], performing unlabeled sample recruitment to enrich the seed training dataset of ground truth labeled samples. Unlabeled data samples are assigned class labels according to their similarity with the ground truth samples. This kind of approaches were explored early in hyperspectral literature [156, 86], and have been extensively pursued since then, including kernel based methods using the semi-supervised samples to improve kernel estimation. Examples are the efficient cost function optimization approach [51], a cluster based estimation of kernel parameters [169], the transductive support vector machine [20], artificial neural networks trained to minimize a regularized objective function [142], or method using the data graph embedding to perform class assignment of unlabeled data [150]. A complementary line of research aiming to overcome the Hughes effect applies dimensionality reduction algorithms. Feature selection looks for a subset of spectral bands maximizing classification accuracy. It is a combinatorial optimization problem, which can be solved by means of heuristics such as the affinity propagation [88] or nature-inspired methods, such as the Darwinian Particle Swarm Optimization enhanced with fractional order [62] or hybridizations of bio-inspired optimization methods [61]. Feature extraction approaches compute a transformation of the pixel data into a lower dimension representation, besides classical approaches such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) [174] or simply averaging of groups of bands [94], recent works exploit combination of several features [112], combinations of class dependent subspace projections [59], or use unmixing as the feature extraction process [50, 49].

Carrying spectral processing and classification independently at each pixel disregards spatial correlation, often very strong, that can be used to improve final classification [54]. For instance, spectral-spatial denoising carried out on wavelet transform coefficients [94, 173] helps improving the final classification. A broad group of approaches aim to correct classification errors enforcing spatial homogeneity. For instance, Markov Random Field (MRF) modeling [87, 184, 166, 131] achieve this correction by approximate relaxation processes, or by belief propagation [108]. Others perform unsupervised image segmentation, using morphological watershed [165] or partitional clustering [163], enforcing pixels in the same region to be of the same class. Some works propose spectral-spatial kernels performing based on suitable similarity measures [84]. A quite different approach performs image decomposition into reflectance and shadow components [94] in order to discard non-informative parts of the image improving classification performance. Combination of multiple spectral-spatial classifiers has been also proposed [164]. Finally, the spatial information has been added as extra features computed from morphological profiles of lower dimension images, such as the first PCA component [81, 53], as an instance of the general approach of attribute profiles for classification [63].

3.2.2 Semi-supervised learning with spatial and spectral similarities

Algorithm 3.1 Semi-supervised learning with spatial correction (SSL-SC) for hyperspectral images

procedure SSL-SC(X, X_L, Ω) ▷ $X = \{\mathbf{x_i}, \mathbf{i} \in I \subset \mathbb{N}^2\}$ hyperspectral image, ▷ $X_L = \{(\mathbf{x_i}, y_\mathbf{i}), \mathbf{i} \in L \subset I\}$ training sample of labeled pixels, ▷ $\mathscr{L}^U$ set of unsupervised classes

Train $C_L : \mathbb{R}^d \to \mathscr{L}$ from X_L ▷ baseline spectral classifier

$\hat{y}_\mathbf{i} = C_L(\mathbf{x_i}), \mathbf{i} \in I$ ▷ baseline class estimations

$C_U : \mathbb{R}^d \to \mathscr{L}^U$, s.t. $\omega_\mathbf{i} = C_U(\mathbf{x_i})$ ▷ unsupervised classification of the image pixel spectra (i.e. K-means)

$$X_U = \left\{ (\mathbf{x_i}, \tilde{y}_\mathbf{i}), \tilde{y}_\mathbf{i} = \arg\max_k |\mathscr{K}_k(\mathbf{i})| \; \mathbf{x_i} \in X_{I-L} \wedge \exists \mathscr{K}_k(\mathbf{i}) \neq \emptyset \right\}$$

$$\mathscr{K}_k(\mathbf{i}) = \left\{ k = y_\mathbf{j} \,\middle|\, \mathbf{i} \in \mathscr{N}_\mathbf{j}(r) \wedge \omega_\mathbf{i} = \omega_\mathbf{j} \wedge \mathbf{x_j} \in X_L \right\}$$

$X_{L+U} = X_L \cup X_U$ ▷ enriched training set

Train $C_{L+U} : \mathbb{R}^d \to \mathscr{L}$ using X_{L+U} ▷ semi-supervised training

$\hat{\mathbf{y}} = \{\hat{y}_\mathbf{i} = C_{L+U}(\mathbf{x_i}), \mathbf{i} \in I\}$. ▷ classify entire image

$$\hat{\hat{\mathbf{y}}} = \left\{ \hat{\hat{y}}_\mathbf{i} = \arg\max_k \left| \left\{ k = \hat{y}_\mathbf{j}, \mathbf{j} \in \mathscr{N}_\mathbf{i}(r) \right\} \right|, \mathbf{i} \in I \right\}.$$ ▷ spatial correction of the class image

end procedure

The proposed spatially regularized semi-supervised classification process is summarized in Algorithm 3.1. We will provide an intuitive description before going into a more formal detailed explanation. Given a small training dataset, which is a very common situation in remote sensing applications, first we enrich it with unlabeled data for which we guess its label according to two conditions: a spatial proximity in the image domain to a labeled pixel belonging to the seed training dataset, and the spectral similarity with this pixel. We say that two pixels are spectrally similar if they belong to the same cluster after an unsupervised clustering process carried out over the entire image data. After training dataset enrichment, the spectral classifier is trained. The entire image is classified with the trained classifier. Finally, a spatial correction step is performed which consists in assigning to each pixel the majority class inside its spatial neighborhood. This spatial correction removes classification errors on the assumption that class spatial regions are compact, i.e. without holes. A more formal description follows.

Let us denote an hyperspectral image as $X = \{\mathbf{x_i}, \mathbf{i} \in I \in \mathbb{N}^2\}$, where $\mathbf{x_i} \in \mathbb{R}^d$ are d-dimensional spectral signatures, and I is the image domain given by a bounded 2D grid of pixel places. The pixel spectra belong to a class whose label is

in the set $\mathscr{L} = \{1, \ldots, \Omega\}$. An image of class labels is specified as $\mathbf{y} = \{y_\mathbf{i}; \mathbf{i} \in I\} \subset \mathscr{L}^{\mathrm{N\ x\ M}}$. The spatial neighborhood of radius r around the pixel site $\mathbf{i}$ is denoted $\mathscr{N}_\mathbf{i}(r)$. Image segmentation aims to obtain a partition of the image domain into a set of regions $R = \{R_k\}$ such that $\cup R_k = I$, and $\cap R_k = \emptyset$. Regions are composed of connected components and pixels falling into one region have some similarity, i.e. they belong to the same class $R_k = \{\mathbf{i} \in I \,|\, y_\mathbf{i} = k\}$.

In Algorithm 3.1, first, we build the baseline spectral classifier $C_L : \mathbb{R}^d \to \mathscr{L}$ from an initial training sample of labeled pixels $X_L = \{(\mathbf{x_i}, y_\mathbf{i}), \mathbf{i} \in L \subset I\}$ (line 2), which will be used to assess the improvement introduced by the ensuing processes from the initial image pixel classification (line 3). Labeled data scarcity means that $|I| \ggg |L|$. Let X_{I-L} denote the set of unlabeled pixel spectra. Next (line 4), we perform the pixel spectra unsupervised classification over all the image $\omega_\mathbf{i} = C_U(\mathbf{x_i})$, which provides the spectral similarity component of the label guessing semi-supervised training set enrichment. Next (line 5), carries the selection of unlabeled pixels $\mathbf{x_i} \in X_{I-L}$ that can be added to the training dataset with a guessed label $\tilde{y}_\mathbf{i}$. To be selected, an unlabeled pixel must be in the spatial neighborhood $\mathbf{i} \in \mathscr{N}_\mathbf{j}(r)$ of a labeled pixel $\mathbf{x_j} \in X_L$. Furthermore, it must belong to the same unsupervised cluster $\omega_\mathbf{i} = \omega_\mathbf{j}$. Then, the label is set according to the labeled sample, with a majority voting if there are more than one (boundary pixels). The obtained X_U is usually a small subset of X_{I-L}, which is used to construct the enriched training dataset $X_{L+U} = X_L \cup X_U$ (line 6). Next (line 7), we build the semi-supervised classifier $C_{L+U} : \mathbb{R}^d \to \mathscr{L}$ using the enriched training set X_{L+U}, obtaining (line 8) the image of class labels $\hat{\mathbf{y}}$. Finally (line 9), we carry out the spatial correction by majority voting in each pixel's neighborhood to obtain the final segmentation image $\hat{\hat{\mathbf{y}}}$.

3.2.3 Spectral-spatial semi-supervised learning rationale

We discuss here foundations for the improvements expected from our novel semi-supervised learning approach, which combines spectral and spatial information to build the set X_U of pixels with "guessed" labels that is added to the training set. This process can be modeled as a label propagation based on spectral and spatial similarity evidences. Let us consider first the spectral aspect. It is natural to assume that the probability of two pixels having the same label is proportional to its spectral distance (either Euclidean or angular), i.e. $p\left(y_\mathbf{i} = y_\mathbf{j} \,|\, \mathbf{x_i}, \mathbf{x_j}\right) \propto \phi\left(\mathbf{x_i} - \mathbf{x_j}\right)$, where ϕ is some monotonic decreasing function of the spectral distance between pixels. This rough statement does not solve the problem of setting the decision thresholds as well as the definition of the appropriate measure. Our proposal to solve this issue is to build an unsupervised classifier $C_U : \mathbb{R}^d \to \mathscr{L}^U$ independently of the given ground truth. Because of this independence we can use all image data without

introducing any bias in the final results. The assignment of the pixel spectra to the latent classes $\omega \in \mathscr{L}^U$ discovered by unsupervised learning, such that $\omega_{\mathbf{i}} = C_U(\mathbf{x}_{\mathbf{i}})$, is done by a MAP decision on the latent class posteriors $\left\{p(\omega|\mathbf{x}); \omega \in \mathscr{L}^U\right\}$. Hence, we can identify the probabilities of having the same label and belonging to the same latent class $p\left(y_{\mathbf{i}} = y_{\mathbf{j}} \left| \mathbf{x}_{\mathbf{i}}, \mathbf{x}_{\mathbf{j}} \right.\right) = p\left(\omega_{\mathbf{i}} = \omega_{\mathbf{j}} \left| \mathbf{x}_{\mathbf{i}}, \mathbf{x}_{\mathbf{j}} \right.\right)$. Moreover, we assume that two pixels have the same class label iff they belong to the same latent class, i.e. $p\left(y_{\mathbf{i}} = y_{\mathbf{j}} \left| \mathbf{x}_{\mathbf{i}}, \mathbf{x}_{\mathbf{j}} \right.\right) \propto \delta\left(\omega_{\mathbf{i}}, \omega_{\mathbf{j}}\right)$, where δ is the Kronecker's delta function. In this thesis work, latent classes are discovered by the classical K-means clustering algorithm [55], however, any other unsupervised classification algorithm may be used. Let us consider now the spatial aspect of the propagation decision, it is natural to assume that the probability of two pixel sites having the same class label decreases with the distance in image domain, i.e. $p\left(y_{\mathbf{i}} = y_{\mathbf{j}} |\mathbf{i}, \mathbf{j}\right) \propto \psi(\mathbf{i} - \mathbf{j})$, where ψ is some monotonic decreasing function. Our approach takes it to the limit

$$p\left(y_{\mathbf{i}} = y_{\mathbf{j}} |\mathbf{i}, \mathbf{j}\right) = \begin{cases} 1 & \mathbf{i} \in \mathscr{N}_{\mathbf{j}}(r) \\ 0 & \text{otherwise} \end{cases}, \tag{3.6}$$

for some radius r which is related with the expected granularity of the regions in the image.

The construction of the semi-supervised training dataset may be modeled as the propagation process of the known labels $y_{\mathbf{i}}$ of pixel sites $\mathbf{i} \in L$ "guessing" labels $\tilde{y}_{\mathbf{j}}$ of unlabeled pixel sites $\mathbf{j} \in I - L$. This propagation combines spectral and spatial information as follows:

$$p\left(y_{\mathbf{i}} = \tilde{y}_{\mathbf{j}}\right) = p\left(y_{\mathbf{i}} = \tilde{y}_{\mathbf{j}} |\mathbf{i}, \mathbf{j}\right) p\left(y_{\mathbf{i}} = \tilde{y}_{\mathbf{j}} \left| \mathbf{x}_{\mathbf{i}}, \mathbf{x}_{\mathbf{j}} \right.\right). \tag{3.7}$$

Therefore, label propagation happens when both spectra belong to the same latent class and to the same spatial neighborhood. To highlight the role of the latent classes, let us consider that we have only one latent class, which is equivalent to not having any spectral discrimination. In such case we have $p\left(y_{\mathbf{i}} = \tilde{y}_{\mathbf{j}} \left| \mathbf{x}_{\mathbf{i}}, \mathbf{x}_{\mathbf{j}} \right.\right) = 1$, so that the decision to add an unlabeled pixel to the enriched training dataset does only depend on the spatial proximity to the labeled pixel in the seed training dataset. This strategy allows quite diverse spectra to enter the same class, being equivalent to introduce noise in the enriched training dataset. Therefore, latent class assignment by clustering acts as a filter for the spatial recruitment of unlabeled pixels.

In Algorithm 3.1, (line 5) we specify the enrichment set X_U as follows:

$$X_U = \left\{ \left(\mathbf{x}_{\mathbf{i}}, \tilde{y}_{\mathbf{i}}\right), \tilde{y}_{\mathbf{i}} = \arg\max_k \left|\mathscr{K}_k(\mathbf{i})\right| \, \mathbf{x}_{\mathbf{i}} \in X_{I-L} \wedge \exists \mathscr{K}_k(\mathbf{i}) \neq \emptyset \right\} \tag{3.8}$$

$$\mathscr{K}_k(\mathbf{i}) = \left\{ k = y_{\mathbf{j}} \left| \mathbf{i} \in \mathscr{N}_{\mathbf{j}}(r) \wedge \omega_{\mathbf{i}} = \omega_{\mathbf{j}} \wedge \mathbf{x}_{\mathbf{j}} \in X_L \right. \right\} \tag{3.9}$$

Therefore, a pixel spectra is added to the enriched training data set when it falls in the spatial neighborhood of a labeled sample in the training dataset and both samples belong to the same latent class. As samples from different classes may belong to the same latent class, it may happen that several classes fill the condition in eq. 3.9. In such event, the most frequent class is selected. The propagation model is a more accurate description of the actual implementation, because we examine the neighbors of the pixels in the training dataset instead of searching the entire image.

3.2.4 Spatial correction rationale

In this subsection we present the reasoning for the proposed spatial correction step, and the source of the improved results that it provides. Following a Bayesian reasoning, the estimation of image pixel labels $\mathbf{y}$ given the observation of the pixel spectra X is carried out by searching for the maximum of the *a posteriori* (MAP) probability $p(\mathbf{y}|X) \propto p(X|\mathbf{y})\,p(\mathbf{y})$, where $p(X|\mathbf{y})$ is the joint conditional distribution of the observed pixel spectra on the hypothesized labels, often Gaussian, and $p(\mathbf{y})$ is the joint *a priori* distribution of image labels. Random Markov Fields [113] are well-known joint *a priori* distribution modeling tools. Given the pixel class label we can assume the spatial independence of the pixel spectra conditional distribution, i.e. $p(X|\mathbf{y}) = \prod_{\mathbf{i}} p(\mathbf{x_i}|y_\mathbf{i})$, so that the joint conditional distribution is decomposed into the local pixel spectra conditional distributions, therefore the MAP classification is specified as

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left\{ \sum_{\mathbf{i}} \left(\log p(y_\mathbf{i}|\mathbf{x_i}) - \log p(y_\mathbf{i})\right) + \log p(\mathbf{y}) \right\}. \tag{3.10}$$

Hence, the MAP estimation of the joint label image is decomposed into local MAP decisions at each pixel site plus an spatial correction term specified by the joint *a priori* distribution. The local MAP decision based on $p(y_\mathbf{i}|\mathbf{x_i})$ is implemented by the AHERS spectral classifier. The *a priori* joint distribution incorporates knowledge that may help to remove classification confusion. The most common and basic assumption is that image regions are compact, so that local correlation of pixel labels is high. We can state in more general terms that images are characterized by a compactness measure $h : \mathscr{L}^{N\times M} \to \mathbb{R}$, such that $h(\mathbf{y}_1) > h(\mathbf{y}_2)$ implies that $p(\mathbf{y}_1) > p(\mathbf{y}_2)$. The compactness measure can be defined in many ways, the simplest one is the number of pixel site pairs with identical class labels. Maximizing compactness is equivalent to enforcing smoothness constraints, such as the minimization of the square differences [113]. Obviously, if we proceed enforcing compactness of sliding windows defined by some neighborhood, we are enforcing

compactness of the entire class regions. This is the case of the label correction given by the most voted class:

$$y_{\mathbf{i}} = \arg\max_{k} \left| \left\{ k = y_{\mathbf{j}} \,|\, j \in \mathscr{N}_{\mathbf{i}}(r) \right\} \right|, \tag{3.11}$$

where $\mathscr{N}_{\mathbf{i}}(r)$ is the local neighborhood of radius r. The radius controls the spread of the smoothing enforced. The improvement achieved by the application of this rule is thus directly rooted to a smooth joint *a priori* class distribution, with compactness related to the size of the neighborhood.

3.2.5 Experimental results with ensembles of ELM's

This section provides the results of the sensitivity exploration on the V-ELM and V-OP-ELM ensemble parameters regarding the number of hidden units. After that, we report an exhaustive comparison of the response of V-ELM and V-rRLM. Finally, we discuss the comparison of the proposed process and other results found in the literature for these datasets. In all computational experiments we have used the same experimental design of [110] to allow for direct comparison of results: spatial neighborhood radius is set to 7, the number of Markov runs is 100, and the same size of the labeled data sample, which is 160 for the Indian Pines dataset, 18 for Salinas A, and 128 for Salinas C, reporting the mean the overall accuracy (OA) in values between 0 and 1, and the standard deviation of the results to assess the stability of the process. The regularization parameter λ of the rELM has been set to 1000. Finally, some visual results of the obtained thematic maps are provided.

3.2.5.1 Exploratory experiments

Sensitivity of V-ELM and V-OP-ELM to the number of hidden units

In figures 3.12, 3.13, 3.14, and 3.15 we present the sensitivity of V-ELM and V-OP-ELM to the number of hidden units computing the segmentation of the Salinas C, Salinas A, Indian Pines and Pavia datasets, respectively. For this experiment we have set the ensemble size to 10, in order to obtain an impression of the influence of the number of hidden units. In all cases, there is a saturation after 300 nodes, which is followed by some oscillations. In Salinas C and Pavia datasets the V-ELM is superior almost for any number of hidden units to V-OP-ELM. In SalinasA dataset, V-OP-ELM improves almost always to V-ELM, finally in the Indian Pines,

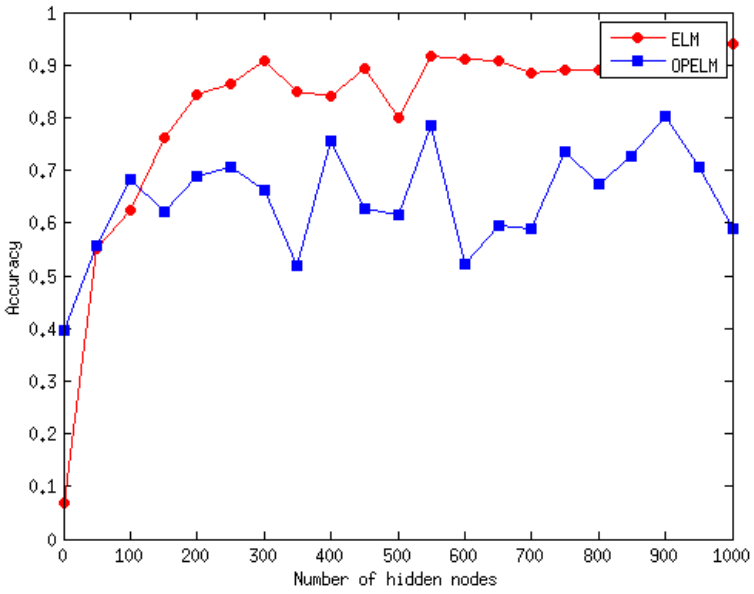


Figure 3.12: Comparison of the sensitivity of V-ELM and V-OP-ELM accuracy values to the number of hidden nodes on the classification of the Salinas C dataset

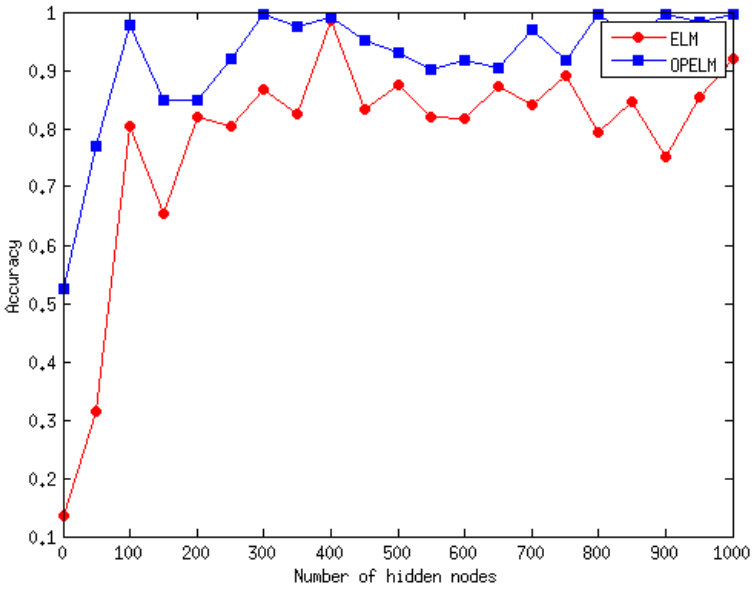


Figure 3.13: Comparison of the sensitivity of V-ELM and V-OP-ELM accuracy values to the number of hidden nodes on the classification of the Salinas A dataset.

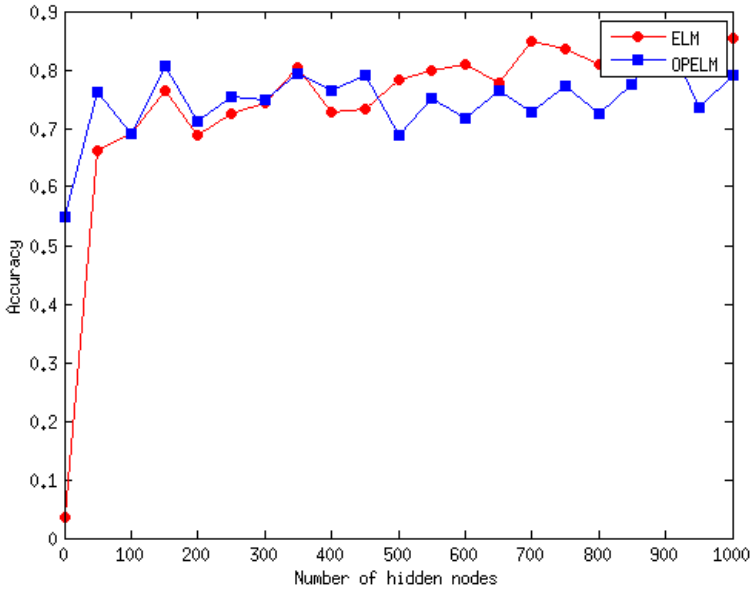


Figure 3.14: Comparison of the sensitivity of V-ELM and V-OP-ELM accuracy values to the number of hidden nodes on the classification of the Indian Pines dataset.

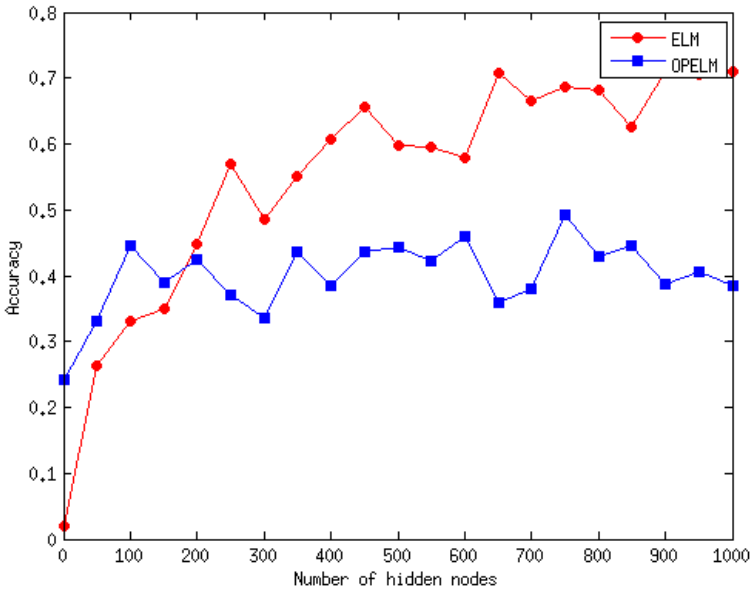


Figure 3.15: Comparison of the sensitivity of V-ELM and V-OP-ELM accuracy values to the number of hidden nodes on the classification of the Pavia dataset.

results are mixed. Notice that there is some relation between the number of spectral bands and the minimal number of hidden units that provides some good results. As a rule of thumb, if the number of hidden units is less than twice the number of spectral bands the results are more erratic and/or lower. V-OP-ELM is much more costly than V-ELM, to the extent that performing exhaustive exploration of the effect of the parameters becomes prohibitive. For this reason, further exploratory experiments have been performed with theV-ELM and V-rELM ensembles.

V-ELM versus V-rELM

Tables 3.1 to 3.4 show the average and variance of the OA obtained after 100

	Salinas C						Salinas A					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5807 (0.0414)	0.6098 (0.0364)	0.6301 (0.0322)	0.6323 (0.0380)	0.6331 (0.0243)	0.6155 (0.0383)	0.7131 (0.0800)	0.7351 (0.0561)	0.7109 (0.0395)	0.6904 (0.0410)	0.7323 (0.0433)	0.7292 (0.0509)
100	0.7240 (0.0242)	0.7369 (0.0294)	0.7511 (0.0123)	0.7448 (0.0208)	0.7586 (0.0084)	0.7621 (0.0107)	0.8506 (0.0554)	0.8591 (0.0352)	0.8591 (0.0480)	0.8699 (0.0452)	0.8540 (0.0416)	0.8710 (0.0290)
150	0.7695 (0.0106)	0.7776 (0.0130)	0.7862 (0.0065)	0.7854 (0.0073)	0.7876 (0.0059)	0.7874 (0.0037)	0.8856 (0.0460)	0.8778 (0.0457)	0.9262 (0.0289)	0.9262 (0.0208)	0.9243 (0.0162)	0.9211 (0.0236)
200	0.7814 (0.0098)	0.8013 (0.0118)	0.8059 (0.0087)	0.8026 (0.0065)	0.8045 (0.0046)	0.8004 (0.0060)	0.9041 (0.0232)	0.9392 (0.0159)	0.9398 (0.0104)	0.9403 (0.0201)	0.9357 (0.0138)	0.9470 (0.0103)
250	0.7994 (0.0049)	0.8057 (0.0051)	0.8152 (0.0045)	0.8164 (0.0055)	0.8172 (0.0032)	0.8154 (0.0049)	0.9389 (0.0222)	0.9413 (0.0140)	0.9458 (0.0066)	0.9517 (0.0045)	0.9530 (0.0040)	0.9491 (0.0045)
300	0.8145 (0.0070)	0.8286 (0.0049)	0.8309 (0.0060)	0.8250 (0.0051)	0.8295 (0.0037)	0.8287 (0.0042)	0.9370 (0.0119)	0.9434 (0.0155)	0.9498 (0.0078)	0.9517 (0.0066)	0.9512 (0.0048)	0.9517 (0.0075)
350	0.8237 (0.0071)	0.8314 (0.0042)	0.8354 (0.0026)	0.8334 (0.0025)	0.8349 (0.0034)	0.8340 (0.0029)	0.9415 (0.0137)	0.9455 (0.0107)	0.9513 (0.0029)	0.9497 (0.0064)	0.9504 (0.0051)	0.9532 (0.0039)
400	0.8308 (0.0052)	0.8358 (0.0064)	0.8389 (0.0036)	0.8407 (0.0055)	0.8408 (0.0034)	0.8402 (0.0036)	0.9371 (0.0143)	0.9453 (0.0108)	0.9494 (0.0049)	0.9522 (0.0047)	0.9515 (0.0049)	0.9518 (0.0049)
450	0.8374 (0.0075)	0.8425 (0.0060)	0.8455 (0.0041)	0.8445 (0.0036)	0.8442 (0.0026)	0.8446 (0.0031)	0.9395 (0.0117)	0.9454 (0.0081)	0.9499 (0.0082)	0.9456 (0.0061)	0.9537 (0.0041)	0.9543 (0.0044)

Table 3.1: Classification results: Average (and variance between brackets) OA results of V-ELM for Salinas A and Salinas C images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Pavia						Indian Pines					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5700 (0.0078)	0.5671 (0.0110)	0.5653 (0.0063)	0.5655 (0.0057)	0.5652 (0.0086)	0.5653 (0.0061)	0.4338 (0.0201)	0.4381 (0.0086)	0.4451 (0.0154)	0.4496 (0.0103)	0.4538 (0.0100)	0.4511 (0.0145)
100	0.5870 (0.0081)	0.5855 (0.0061)	0.5924 (0.0048)	0.5902 (0.0071)	0.5917 (0.0062)	0.5903 (0.0034)	0.4774 (0.0158)	0.4837 (0.0078)	0.4906 (0.0086)	0.4858 (0.0086)	0.4896 (0.0078)	0.4871 (0.0066)
150	0.6055 (0.0082)	0.6052 (0.0084)	0.6084 (0.0028)	0.6068 (0.0040)	0.6089 (0.0058)	0.6056 (0.0039)	0.4783 (0.0175)	0.4964 (0.0068)	0.5050 (0.0117)	0.5097 (0.0054)	0.5074 (0.0048)	0.5090 (0.0039)
200	0.6311 (0.0068)	0.6214 (0.0051)	0.6247 (0.0064)	0.6215 (0.0017)	0.6252 (0.0049)	0.6208 (0.0036)	0.4962 (0.0104)	0.5072 (0.0094)	0.5126 (0.0064)	0.5121 (0.0063)	0.5121 (0.0028)	0.5121 (0.0043)
250	0.6425 (0.0075)	0.6412 (0.0093)	0.6453 (0.0106)	0.6417 (0.0050)	0.6668 (0.0059)	0.6645 (0.0042)	0.5060 (0.0116)	0.5100 (0.0072)	0.5118 (0.0058)	0.5140 (0.0036)	0.5115 (0.0057)	0.5132 (0.0060)
300	0.6695 (0.0060)	0.6707 (0.0054)	0.6824 (0.0048)	0.6783 (0.0042)	0.6815 (0.0038)	0.6817 (0.0056)	0.5050 (0.0087)	0.5134 (0.0053)	0.5150 (0.0063)	0.5160 (0.0057)	0.5163 (0.0039)	0.5150 (0.0043)
350	0.6820 (0.0079)	0.6866 (0.0053)	0.6930 (0.0038)	0.6882 (0.0044)	0.6929 (0.0051)	0.6929 (0.0045)	0.5144 (0.0103)	0.5197 (0.0072)	0.5178 (0.0062)	0.5198 (0.0039)	0.5191 (0.0035)	0.5212 (0.0049)
400	0.6909 (0.0059)	0.6966 (0.0046)	0.7048 (0.0049)	0.7015 (0.0047)	0.7055 (0.0020)	0.7051 (0.0035)	0.5108 (0.0125)	0.5210 (0.0071)	0.5210 (0.0046)	0.5218 (0.0029)	0.5225 (0.0049)	0.5250 (0.0037)
450	0.7021 (0.0070)	0.7035 (0.0055)	0.7121 (0.0029)	0.7124 (0.0050)	0.7129 (0.0027)	0.7128 (0.0034)	0.5220 (0.0112)	0.5237 (0.0050)	0.5267 (0.0040)	0.5265 (0.0034)	0.5247 (0.0033)	0.5267 (0.0038)

Table 3.2: Classification results: Average (and variance between brackets) OA results of V-ELM for Pavia and Indian Pines images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Salinas C						Salinas A					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5901 (0.0365)	0.6043 (0.0312)	0.6031 (0.0362)	0.6302 (0.0223)	0.6274 (0.0277)	0.6213 (0.0420)	0.7247 (0.0793)	0.7302 (0.0427)	0.7230 (0.0616)	0.7156 (0.0492)	0.6967 (0.0336)	0.7186 (0.0360)
100	0.7181 (0.0264)	0.7460 (0.0148)	0.7593 (0.0191)	0.7607 (0.0117)	0.7517 (0.0092)	0.7538 (0.0085)	0.8455 (0.0606)	0.8416 (0.0456)	0.8745 (0.0373)	0.8537 (0.0342)	0.8540 (0.0503)	0.8320 (0.0130)
150	0.7633 (0.0201)	0.7843 (0.0109)	0.7831 (0.0076)	0.7870 (0.0064)	0.7911 (0.0077)	0.7864 (0.0063)	0.9107 (0.0301)	0.9189 (0.0310)	0.9185 (0.0282)	0.9244 (0.0252)	0.9358 (0.0310)	0.9169 (0.0298)
200	0.7936 (0.0219)	0.7978 (0.0103)	0.8016 (0.0076)	0.8013 (0.0051)	0.8062 (0.0067)	0.8011 (0.0066)	0.9123 (0.0493)	0.9276 (0.0291)	0.9393 (0.0216)	0.9411 (0.0108)	0.9408 (0.0161)	0.9452 (0.0093)
250	0.8009 (0.0132)	0.8136 (0.0074)	0.8163 (0.0082)	0.8165 (0.0050)	0.8169 (0.0057)	0.8158 (0.0043)	0.9371 (0.0248)	0.9428 (0.0073)	0.9516 (0.0061)	0.9444 (0.0090)	0.9470 (0.0054)	0.9526 (0.0044)
300	0.8176 (0.0103)	0.8238 (0.0076)	0.8272 (0.0067)	0.8266 (0.0053)	0.8275 (0.0032)	0.8289 (0.0031)	0.9335 (0.0207)	0.9482 (0.0096)	0.9476 (0.0084)	0.9537 (0.0071)	0.9511 (0.0049)	0.9558 (0.0040)
350	0.8274 (0.0093)	0.8312 (0.0035)	0.8349 (0.0040)	0.8343 (0.0043)	0.8339 (0.0040)	0.8353 (0.0016)	0.9469 (0.0122)	0.9511 (0.0063)	0.9523 (0.0043)	0.9550 (0.0034)	0.9538 (0.0049)	0.9543 (0.0028)
400	0.8281 (0.0071)	0.8354 (0.0036)	0.8385 (0.0027)	0.8395 (0.0019)	0.8390 (0.0038)	0.8377 (0.0029)	0.9488 (0.0058)	0.9475 (0.0063)	0.9547 (0.0036)	0.9531 (0.0027)	0.9545 (0.0040)	0.9547 (0.0034)
450	0.8374 (0.0066)	0.8388 (0.0080)	0.8430 (0.0038)	0.8438 (0.0028)	0.8475 (0.0029)	0.8448 (0.0048)	0.9445 (0.0130)	0.9473 (0.0093)	0.9524 (0.0060)	0.9542 (0.0052)	0.9543 (0.0038)	0.9559 (0.0026)

Table 3.3: Classification results: Average (and variance between brackets) OA results of V-rELM for Salinas A and Salinas C images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Pavia						Indian Pines					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5679 (0.0128)	0.5647 (0.0075)	0.5665 (0.0084)	0.5645 (0.0055)	0.5674 (0.0084)	0.5658 (0.0055)	0.4394 (0.0227)	0.4360 (0.0168)	0.4460 (0.0129)	0.4502 (0.0161)	0.4489 (0.0133)	0.4477 (0.0124)
100	0.5903 (0.0072)	0.5850 (0.0079)	0.5932 (0.0058)	0.5897 (0.0051)	0.5898 (0.0049)	0.5885 (0.0035)	0.4780 (0.0157)	0.4737 (0.0074)	0.4822 (0.0100)	0.4902 (0.0085)	0.4896 (0.0093)	0.4875 (0.0072)
150	0.6079 (0.0102)	0.6063 (0.0056)	0.6081 (0.0029)	0.6043 (0.0034)	0.6084 (0.0037)	0.6075 (0.0031)	0.4865 (0.0149)	0.4971 (0.0108)	0.4981 (0.0118)	0.5002 (0.0062)	0.5084 (0.0076)	0.5059 (0.0061)
200	0.6289 (0.0093)	0.6196 (0.0097)	0.6238 (0.0058)	0.6213 (0.0054)	0.6225 (0.0041)	0.6230 (0.0060)	0.4949 (0.0120)	0.5076 (0.0093)	0.5101 (0.0037)	0.5057 (0.0057)	0.5124 (0.0054)	0.5095 (0.0053)
250	0.6569 (0.0078)	0.6612 (0.0073)	0.6670 (0.0053)	0.6637 (0.0053)	0.6638 (0.0040)	0.6662 (0.0054)	0.5018 (0.0091)	0.5154 (0.0059)	0.5107 (0.0065)	0.5181 (0.0055)	0.5133 (0.0059)	0.5123 (0.0043)
300	0.6726 (0.0103)	0.6680 (0.0077)	0.6793 (0.0058)	0.6764 (0.0055)	0.6813 (0.0038)	0.6780 (0.0036)	0.5035 (0.0150)	0.5145 (0.0078)	0.5155 (0.0051)	0.5174 (0.0051)	0.5172 (0.0045)	0.5149 (0.0033)
350	0.6845 (0.0069)	0.6837 (0.0040)	0.6934 (0.0047)	0.6890 (0.0065)	0.6937 (0.0052)	0.6910 (0.0039)	0.5118 (0.0084)	0.5182 (0.0072)	0.5194 (0.0067)	0.5204 (0.0050)	0.5213 (0.0037)	0.5185 (0.0054)
400	0.6922 (0.0077)	0.6959 (0.0061)	0.7020 (0.0060)	0.7024 (0.0049)	0.7024 (0.0066)	0.7037 (0.0049)	0.5144 (0.0111)	0.5217 (0.0064)	0.5192 (0.0037)	0.5218 (0.0052)	0.5212 (0.0035)	0.5208 (0.0042)
450	0.7040 (0.0088)	0.7005 (0.0064)	0.7119 (0.0052)	0.7105 (0.0047)	0.7164 (0.0034)	0.7137 (0.0037)	0.5210 (0.0112)	0.5241 (0.0065)	0.5249 (0.0043)	0.5241 (0.0032)	0.5258 (0.0028)	0.5269 (0.0039)

Table 3.4: Classification results: Average (and variance between brackets) OA results of V-rELM for Pavia and Indian Pines images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Salinas C						Salinas A					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.6081 (0.0493)	0.6214 (0.0445)	0.6441 (0.0405)	0.6476 (0.0451)	0.6410 (0.0316)	0.6204 (0.0459)	0.7351 (0.1025)	0.7762 (0.0749)	0.7328 (0.0774)	0.7048 (0.0702)	0.7732 (0.0760)	0.7584 (0.0874)
100	0.7779 (0.0306)	0.7877 (0.0365)	0.8099 (0.0206)	0.7926 (0.0363)	0.8102 (0.0180)	0.8145 (0.0151)	0.9059 (0.0620)	0.9022 (0.0385)	0.8980 (0.0562)	0.9155 (0.0411)	0.8973 (0.0494)	0.9080 (0.0368)
150	0.8378 (0.0192)	0.8379 (0.0128)	0.8437 (0.0090)	0.8432 (0.0100)	0.8438 (0.0086)	0.8472 (0.0040)	0.9380 (0.0521)	0.9124 (0.0558)	0.9730 (0.0324)	0.9696 (0.0266)	0.9662 (0.0193)	0.9596 (0.0278)
200	0.8504 (0.0129)	0.8522 (0.0162)	0.8669 (0.0103)	0.8571 (0.0059)	0.8651 (0.0084)	0.8564 (0.0096)	0.9617 (0.0366)	0.9890 (0.0084)	0.9840 (0.0144)	0.9886 (0.0159)	0.9766 (0.0205)	0.9897 (0.0131)
250	0.8714 (0.0115)	0.8651 (0.0075)	0.8771 (0.0083)	0.8741 (0.0095)	0.8784 (0.0068)	0.8724 (0.0068)	0.9867 (0.0186)	0.9924 (0.0029)	0.9938 (0.0018)	0.9945 (0.0014)	0.9945 (0.0013)	0.9939 (0.0016)
300	0.8637 (0.0164)	0.8692 (0.0037)	0.8728 (0.0069)	0.8619 (0.0072)	0.8730 (0.0070)	0.8704 (0.0077)	0.9923 (0.0029)	0.9900 (0.0112)	0.9928 (0.0055)	0.9947 (0.0011)	0.9943 (0.0019)	0.9936 (0.0037)
350	0.8726 (0.0096)	0.8749 (0.0063)	0.8780 (0.0036)	0.8751 (0.0050)	0.8784 (0.0067)	0.8742 (0.0039)	0.9936 (0.0053)	0.9949 (0.0025)	0.9944 (0.0013)	0.9948 (0.0020)	0.9950 (0.0012)	0.9949 (0.0010)
400	0.8835 (0.0102)	0.8788 (0.0120)	0.8841 (0.0063)	0.8852 (0.0099)	0.8828 (0.0046)	0.8809 (0.0055)	0.9929 (0.0065)	0.9947 (0.0023)	0.9944 (0.0012)	0.9952 (0.0012)	0.9952 (0.0008)	0.9946 (0.0017)
450	0.8956 (0.0135)	0.8865 (0.0108)	0.8929 (0.0048)	0.8880 (0.0048)	0.8877 (0.0038)	0.8862 (0.0050)	0.9913 (0.0084)	0.9950 (0.0019)	0.9955 (0.0014)	0.9942 (0.0014)	0.9951 (0.0023)	0.9954 (0.0015)

Table 3.5: Segmentation results: Average (and variance between brackets) OA results of V-ELM for Salinas A and Salinas C images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Pavia						Indian Pines					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5827 (0.0064)	0.5825 (0.0072)	0.5780 (0.0034)	0.5786 (0.0027)	0.5804 (0.0048)	0.5796 (0.0036)	0.4428 (0.0296)	0.4379 (0.0145)	0.4508 (0.0220)	0.4552 (0.0159)	0.4631 (0.0141)	0.4585 (0.0244)
100	0.6036 (0.0172)	0.5960 (0.0097)	0.5988 (0.0034)	0.5969 (0.0060)	0.5973 (0.0039)	0.5960 (0.0036)	0.5015 (0.0312)	0.4952 (0.0080)	0.5026 (0.0118)	0.4965 (0.0088)	0.4949 (0.0085)	0.4937 (0.0082)
150	0.6606 (0.0272)	0.6457 (0.0279)	0.6327 (0.0222)	0.6297 (0.0220)	0.6396 (0.0247)	0.6123 (0.0065)	0.5132 (0.0135)	0.5317 (0.0086)	0.5326 (0.0132)	0.5385 (0.0147)	0.5359 (0.0182)	0.5279 (0.0098)
200	0.7401 (0.0217)	0.7118 (0.0337)	0.7124 (0.0252)	0.7049 (0.0193)	0.7141 (0.0232)	0.6969 (0.0298)	0.5591 (0.0114)	0.5575 (0.0152)	0.5566 (0.0122)	0.5593 (0.0126)	0.5614 (0.0062)	0.5639 (0.0089)
250	0.7633 (0.0133)	0.7588 (0.0155)	0.7599 (0.0167)	0.7542 (0.0160)	0.7410 (0.0056)	0.7427 (0.0057)	0.5777 (0.0138)	0.5752 (0.0109)	0.5648 (0.0131)	0.5761 (0.0102)	0.5751 (0.0112)	0.5743 (0.0097)
300	0.7517 (0.0114)	0.7440 (0.0056)	0.7559 (0.0073)	0.7498 (0.0037)	0.7525 (0.0064)	0.7500 (0.0026)	0.5823 (0.0172)	0.5852 (0.0078)	0.5785 (0.0124)	0.5777 (0.0058)	0.5799 (0.0116)	0.5831 (0.0093)
350	0.7517 (0.0118)	0.7621 (0.0105)	0.7601 (0.0079)	0.7546 (0.0053)	0.7625 (0.0077)	0.7570 (0.0067)	0.5870 (0.0117)	0.5897 (0.0106)	0.5913 (0.0072)	0.5909 (0.0063)	0.5917 (0.0085)	0.5957 (0.0060)
400	0.7599 (0.0123)	0.7654 (0.0072)	0.7712 (0.0140)	0.7658 (0.0058)	0.7686 (0.0053)	0.7706 (0.0095)	0.5910 (0.0162)	0.5986 (0.0099)	0.5947 (0.0085)	0.5931 (0.0059)	0.5984 (0.0079)	0.6000 (0.0041)
450	0.7660 (0.0132)	0.7716 (0.0127)	0.7822 (0.0082)	0.7792 (0.0125)	0.7797 (0.0043)	0.7812 (0.0104)	0.6002 (0.0087)	0.6018 (0.0055)	0.6026 (0.0072)	0.6001 (0.0071)	0.6003 (0.0055)	0.5984 (0.0059)

Table 3.6: Segmentation results: Average (and variance between brackets) OA results of V-ELM for Pavia and Indian Pines images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Salinas A						Salinas C					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.6125 (0.0529)	0.6185 (0.0435)	0.6174 (0.0420)	0.6368 (0.0312)	0.6401 (0.0403)	0.6275 (0.0589)	0.7455 (0.1080)	0.7763 (0.0769)	0.7507 (0.0823)	0.7370 (0.0767)	0.7140 (0.0525)	0.7494 (0.0624)
100	0.7831 (0.0303)	0.8028 (0.0162)	0.8163 (0.0196)	0.8143 (0.0141)	0.8050 (0.0135)	0.8072 (0.0110)	0.8967 (0.0686)	0.8941 (0.0420)	0.9194 (0.0452)	0.8900 (0.0371)	0.8874 (0.0488)	0.8731 (0.0089)
150	0.8342 (0.0246)	0.8365 (0.0152)	0.8431 (0.0123)	0.8435 (0.0098)	0.8479 (0.0078)	0.8432 (0.0057)	0.9709 (0.0371)	0.9574 (0.0452)	0.9560 (0.0428)	0.9654 (0.0343)	0.9729 (0.0438)	0.9551 (0.0449)
200	0.8624 (0.0281)	0.8558 (0.0122)	0.8651 (0.0109)	0.8580 (0.0088)	0.8684 (0.0099)	0.8582 (0.0060)	0.9632 (0.0546)	0.9754 (0.0340)	0.9797 (0.0252)	0.9842 (0.0142)	0.9828 (0.0191)	0.9917 (0.0043)
250	0.8683 (0.0142)	0.8742 (0.0102)	0.8776 (0.0088)	0.8783 (0.0089)	0.8767 (0.0085)	0.8749 (0.0079)	0.9791 (0.0378)	0.9905 (0.0101)	0.9944 (0.0013)	0.9912 (0.0065)	0.9920 (0.0064)	0.9946 (0.0020)
300	0.8711 (0.0155)	0.8649 (0.0101)	0.8720 (0.0107)	0.8700 (0.0093)	0.8698 (0.0069)	0.8685 (0.0054)	0.9811 (0.0257)	0.9945 (0.0022)	0.9939 (0.0018)	0.9952 (0.0007)	0.9941 (0.0013)	0.9954 (0.0010)
350	0.8768 (0.0139)	0.8729 (0.0053)	0.8804 (0.0078)	0.8754 (0.0052)	0.8773 (0.0064)	0.8757 (0.0027)	0.9938 (0.0032)	0.9946 (0.0016)	0.9944 (0.0014)	0.9946 (0.0015)	0.9947 (0.0015)	0.9943 (0.0009)
400	0.8841 (0.0169)	0.8759 (0.0052)	0.8874 (0.0067)	0.8808 (0.0047)	0.8848 (0.0045)	0.8783 (0.0030)	0.9931 (0.0035)	0.9944 (0.0025)	0.9944 (0.0019)	0.9945 (0.0010)	0.9944 (0.0017)	0.9941 (0.0010)
450	0.8925 (0.0139)	0.8813 (0.0123)	0.8906 (0.0063)	0.8865 (0.0047)	0.8945 (0.0056)	0.8878 (0.0083)	0.9929 (0.0055)	0.9944 (0.0012)	0.9948 (0.0014)	0.9948 (0.0017)	0.9942 (0.0017)	0.9947 (0.0011)

Table 3.7: Segmentation results: Average (and variance between brackets) OA results of V-rELM for Salinas A and Salinas C images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

	Pavia						Indian Pines					
	5	10	15	20	25	30	5	10	15	20	25	30
50	0.5841 (0.0116)	0.5826 (0.0070)	0.5797 (0.0044)	0.5802 (0.0048)	0.5801 (0.0051)	0.5791 (0.0032)	0.4465 (0.0285)	0.4407 (0.0253)	0.4494 (0.0216)	0.4571 (0.0233)	0.4535 (0.0230)	0.4539 (0.0200)
100	0.6091 (0.0208)	0.5968 (0.0098)	0.6001 (0.0059)	0.5964 (0.0041)	0.5959 (0.0044)	0.5954 (0.0027)	0.5011 (0.0170)	0.4869 (0.0116)	0.4922 (0.0097)	0.4965 (0.0062)	0.4960 (0.0053)	0.4926 (0.0079)
150	0.6587 (0.0225)	0.6433 (0.0272)	0.6376 (0.0176)	0.6143 (0.0092)	0.6167 (0.0069)	0.6208 (0.0149)	0.5221 (0.0210)	0.5320 (0.0178)	0.5240 (0.0165)	0.5236 (0.0115)	0.5289 (0.0166)	0.5279 (0.0133)
200	0.7398 (0.0282)	0.7060 (0.0362)	0.7165 (0.0272)	0.6922 (0.0329)	0.7054 (0.0222)	0.6946 (0.0350)	0.5533 (0.0252)	0.5547 (0.0078)	0.5593 (0.0120)	0.5586 (0.0139)	0.5591 (0.0081)	0.5531 (0.0108)
250	0.7372 (0.0056)	0.7414 (0.0084)	0.7402 (0.0045)	0.7389 (0.0064)	0.7389 (0.0025)	0.7398 (0.0040)	0.5782 (0.0140)	0.5828 (0.0137)	0.5675 (0.0111)	0.5786 (0.0064)	0.5702 (0.0088)	0.5707 (0.0112)
300	0.7475 (0.0071)	0.7429 (0.0072)	0.7534 (0.0080)	0.7497 (0.0040)	0.7534 (0.0068)	0.7482 (0.0023)	0.5797 (0.0100)	0.5840 (0.0106)	0.5820 (0.0080)	0.5813 (0.0086)	0.5818 (0.0075)	0.5783 (0.0062)
350	0.7534 (0.0063)	0.7574 (0.0152)	0.7593 (0.0096)	0.7546 (0.0058)	0.7613 (0.0092)	0.7583 (0.0063)	0.5896 (0.0114)	0.5954 (0.0124)	0.5914 (0.0075)	0.5941 (0.0083)	0.5906 (0.0066)	0.5920 (0.0061)
400	0.7623 (0.0132)	0.7633 (0.0111)	0.7701 (0.0105)	0.7697 (0.0093)	0.7677 (0.0097)	0.7674 (0.0106)	0.5944 (0.0150)	0.5997 (0.0065)	0.5945 (0.0110)	0.5955 (0.0105)	0.5964 (0.0061)	0.5933 (0.0070)
450	0.7650 (0.0127)	0.7644 (0.0083)	0.7775 (0.0136)	0.7784 (0.0122)	0.7787 (0.0095)	0.7758 (0.0090)	0.6041 (0.0129)	0.6012 (0.0083)	0.5991 (0.0070)	0.5986 (0.0030)	0.6018 (0.0075)	0.5997 (0.0057)

Table 3.8: Segmentation results: Average (and variance between brackets) OA results of V-rELM for Pavia and Indian Pines images. Rows correspond to the number of hidden units per individual classifier, columns correspond to ensemble size.

Markov runs for increasing number of hidden units and ensemble size for the V-ELM and V-rELM after semisupervised classification, while Tables 3.5 to 3.8 show the average and variance of the segmentation OA after performing spatial regularization. The big number of Markov repetitions is aimed to obtain an accurate estimation of the variance of the results. We have varied the size of the ensemble and the number of hidden units as shown in the tables. The experiment leads to several conclusions:

Hidden units versus ensemble size The effect of ensemble size on the OA is not significant for a given number of hidden units ($p>0.5$ in independent two sided t-tests for each number of hidden units), and the spatial regularization introduces a further reduction of this effect. On the other hand, the number of hidden units has a strong effect ($p<0.01$ in independent t-tests for each ensemble size), monotonically increasing the average OA. Both factors influence the variance of the OA: increasing them decreases the OA variance, however the number of hidden units has a stronger effect.

V-ELM versus V-rELM There is not a significant difference between V-ELM and V-rELM results ($p>0.5$ in a two sided t-test on the average OA of the tables). This result is somewhat surprising, since we expect that the regularized ELM would improve results or at least reduce the variance of results. This lack of effect may be due to the compensation of the regularization effects of the ensembles and the rELM, an indirect confirmation of the regularization achieved by ensembles of ELMs.

Improvement introduced by the spatial regularization Results are uniform in this regard, the spatial regularization introduces a significant increase (up to 10% for Pavia and Indian Pines datasets, and 5% for Salinas datasets) regardless of the kind of classifier (V-ELM vs. V-rELM). The improvement is depending on the quality of the classification results. It can be appreciated that for the smallest ensembles with lowest number of hidden units, the improvement is much less than in the case of the best classification results. Nevertheless, the spatial regularization has the effect of reducing the differences of performance in the lower end of the tables.

3.2.5.2 Comparison with other approaches

Table 3.9: Best accuracy results obtained before spatial regularization for the 4 remote sensing images. Last columns are the number of hidden units and ensemble size.

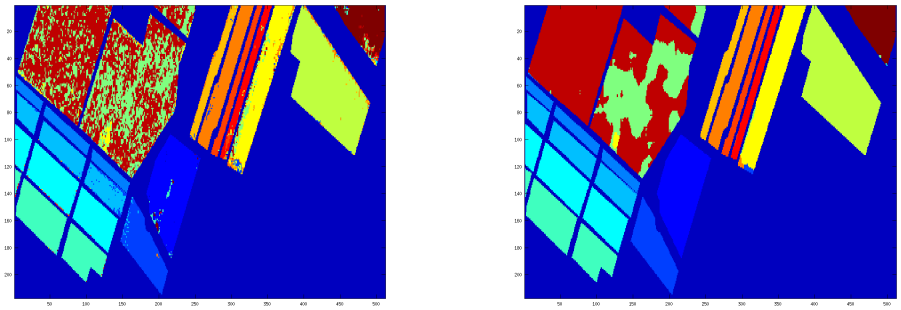
Image	Accuracy	Classifier	n	size
SalinasC	0.8508	V-ELM	1000	10
	0.8240	MLR	-	-
SalinasA	0.9746	V-ELM	1000	25
	0.9086	MLR	-	-
Pavia	0.8002	V-ELM	1000	30
	/	MLR	-	-
Indian	0.6012	V-ELM	1000	30
	0.6344	MLR	-	-

Table 3.10: Best accuracy obtained after spatial regularization for the 4 remote sensing images. Last columns are the number of hidden units and ensemble size.

Image	Accuracy	Classifier	n	size
SalinasC	0.9388	V-ELM	1000	10
	0.8961	MLR	-	-
SalinasA	0.9984	V-ELM	750	10
	0.9674	MLR	-	-
Pavia	0.8918	V-ELM	1000	25
	/	MLR	-	-
Indian	0.7008	V-ELM	950	20
	0.7560	MLR	-	-

Method and reference	OA	% Train
MPM-LBP[111]	0.8578	290 training samples
EMP/SVM[111]	0.8522	290 training samples
SVM + majority vote (WHEDs)[167]	0.8305	8,00%
CENA[37]	0.8847	2,00%
class 2vs4 (Proposed) [37]	0.8847	8,00%
D[178]	0.7370	450 training samples
SVM-CK[159]	0.8711	9,00%
SVA + ICA [133]	0.9447	9,00%
EMP +KPCA[116]	0.9655	9,00%
SVMCK[1]	0.8718	9,00%
KSSP[1]	0.8765	9,00%
KSPCK[1]	0.8319	9,00%

Table 3.11: Reference results for Pavia dataset



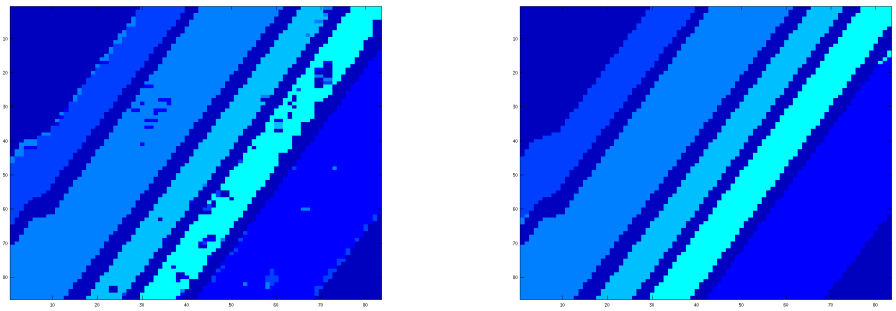
(a) (b)

Figure 3.16: Visualization of classification results on Salinas C using 123 labeled samples, ensemble of 15 basic ELMs and 500 hidden nodes. (a) After supervised classification with OA=80.97%. (b) After spatial regularization with OA=87.02%.

We make a direct comparison with the results reported in [110] for the same size of the labeled sample. Comparative classification and segmentation OA results (i.e. before and after spatial regularization) are given in tables 3.9 and 3.10, respectively, where MRL denotes the Multinomial Logistic Regression approach in [110]. We gather the best results of our experiments to this comparison. The voting ensemble V-ELM improves the MLR approach in the Salinas datasets, however it does not improve it in the Indian Pines dataset. Table 3.11 shows some reference results along with the size of the training set. Only two reported results improve over our approach, using bigger training labeled samples, our results use about 0.5% of the data in the image, while reported results use 9% of the image for training.

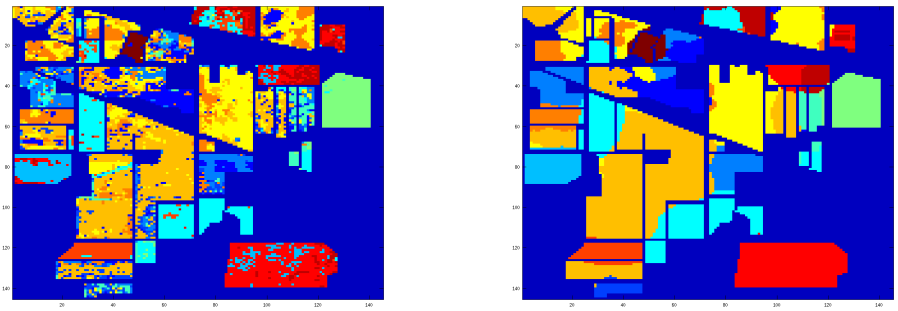
3.2.5.3 Visual results

Figures 3.16, 3.17, 3.18, and 3.19 show the visualization of the thematic map for a V-ELM of size 15, 500 hidden nodes, demonstrating the effect of the spatial regularization. It can be appreciated in all cases that the spatial regularization acts as a smoothing filter of the classification image, removing speckle noise that correspond to erroneous classifications in most cases.



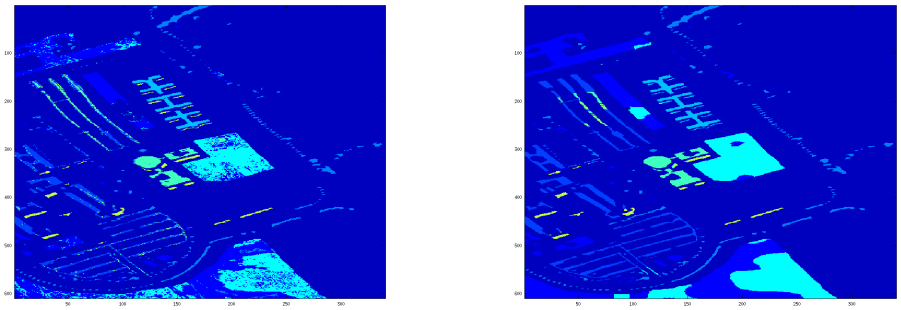
(a) (b)

Figure 3.17: Visualization of classification results on Salinas A using 16 labeled samples, ensemble of 15 basic ELMs and 500 hidden nodes.(a) After supervised classification with OA=95.34%. (b) After spatial regularization with OA=99.86%.



(a) (b)

Figure 3.18: Visualization of classification results on Indian Pines. using 151 labeled samples, ensemble of 15 basic ELMs and 500 hidden nodes. (a) After supervised classification with OA=54.40%. (b) After spatial regularization with OA=62.42%.



(a) (b)

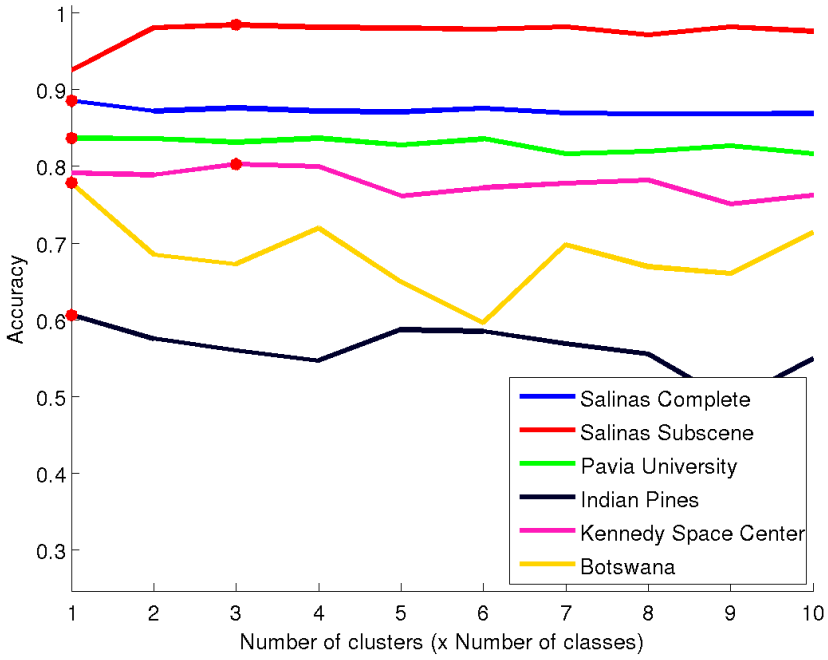
Figure 3.19: Visualization of classification results on Pavia University using 158 labeled samples, ensemble of 15 basic ELMs and 500 hidden nodes. (a) After supervised classification OA=66.94%. (b) After spatial regularization OA=69.93%.

3.2.6 Experimental results with AHERF

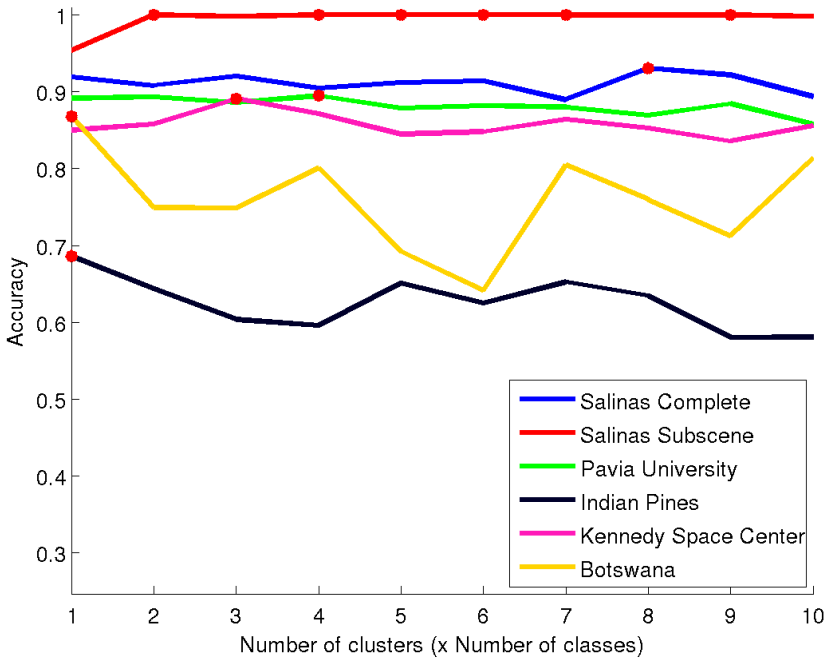
We have carried out sensitivity studies of the proposed method over the different parameters, the best results are then compared with the best ones found in the literature. We have used the average accuracy of the entire image classification as the comparative measure. All computational experiments are repeated 30 times to provide an average result. Each repetition consists on the random selection of the training dataset from the available ground truth, realization of the semi-supervised enrichment, classification and correction. We provide the results before and after spatial correction to show the improvement introduced. The background class is not considered for classification, because of its intrinsic heterogeneity.

Effect of the number of latent classes (clusters) In Figure 3.20 we show the effect of the number of latent classes (clusters) on the average accuracy for all the datasets before and after the spatial correction. The abscissa in the plots is a factor multiplying the number of classes to determine the number of clusters. Star points are the maxima of the corresponding plot. There isn't a big effect of this parameter, with small real improvement on average accuracy before and after spatial correction, with increase the number of clusters.

Effect of the seed training dataset size In Figure 3.21 we can see the effect on the average accuracy of the seed training dataset size, specified as the percentage of ground truth labeled pixels, before and after spatial correction. In this case for both cases there is a definitive improvement when we increase the seed training

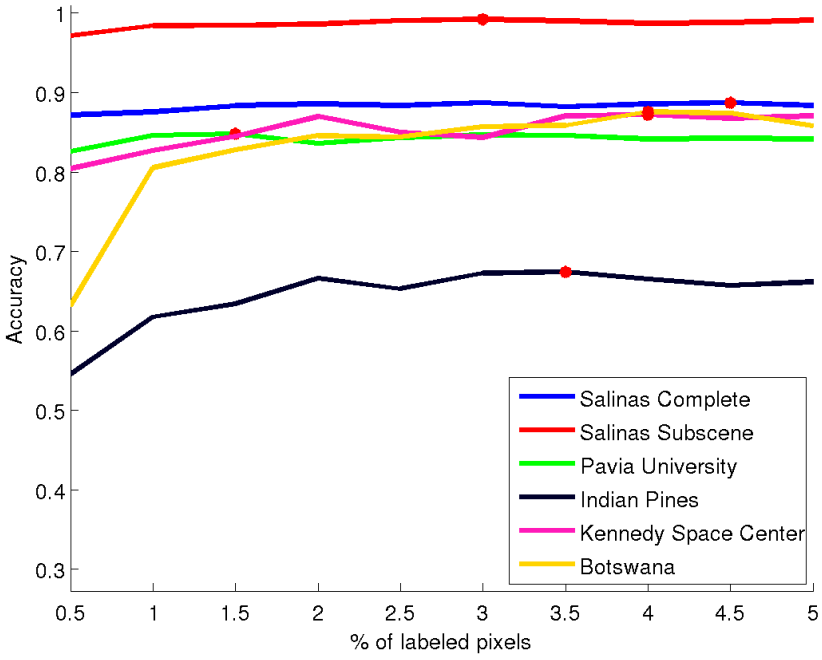


(a)

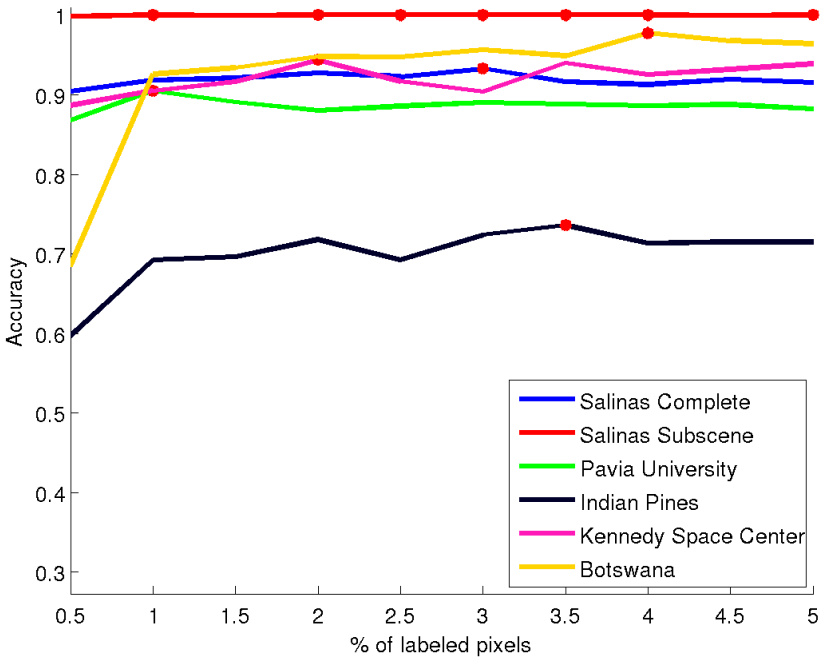


(b)

Figure 3.20: Effect of the number of latent classes (clusters) selected for the spectral-spatial semi-supervised data training. All other parameters set to nominal values. Abscissa is the factor multiplying the number of classes to determine the number of latent classes. Star points are the maxima of the corresponding plot. (a) Accuracy before spatial correction, (b) Accuracy after spatial correction.

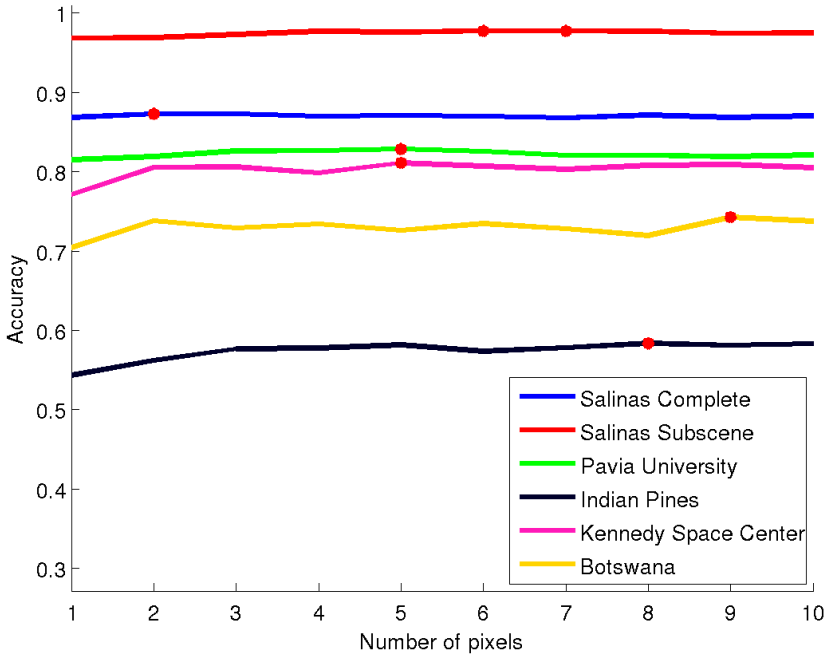


(a)

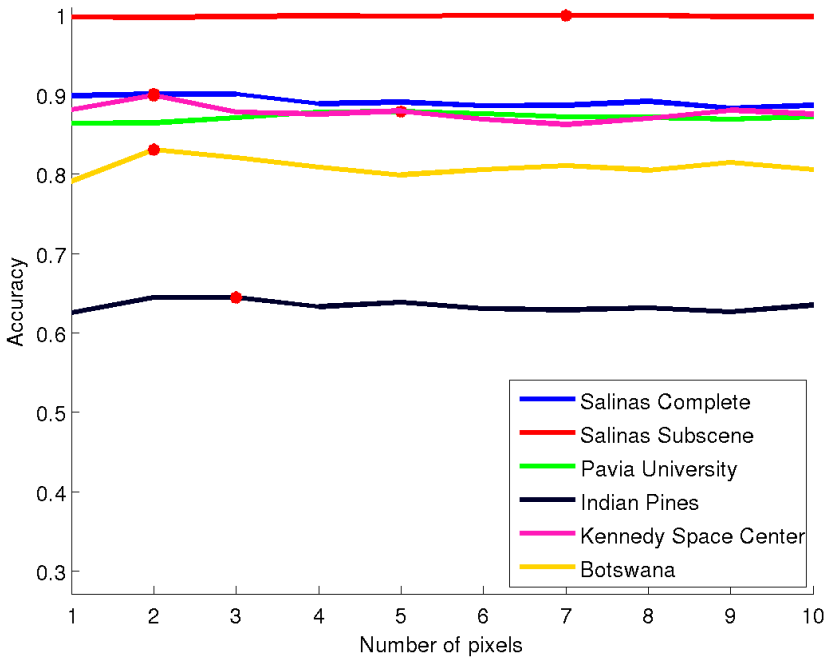


(b)

Figure 3.21: Effect on the average accuracy of the seed training dataset size specified as the ground truth percentage of labeled pixels selected to build it. Star points are the maxima of the corresponding plot. (a) Accuracy before spatial correction, (b) Accuracy after spatial correction.

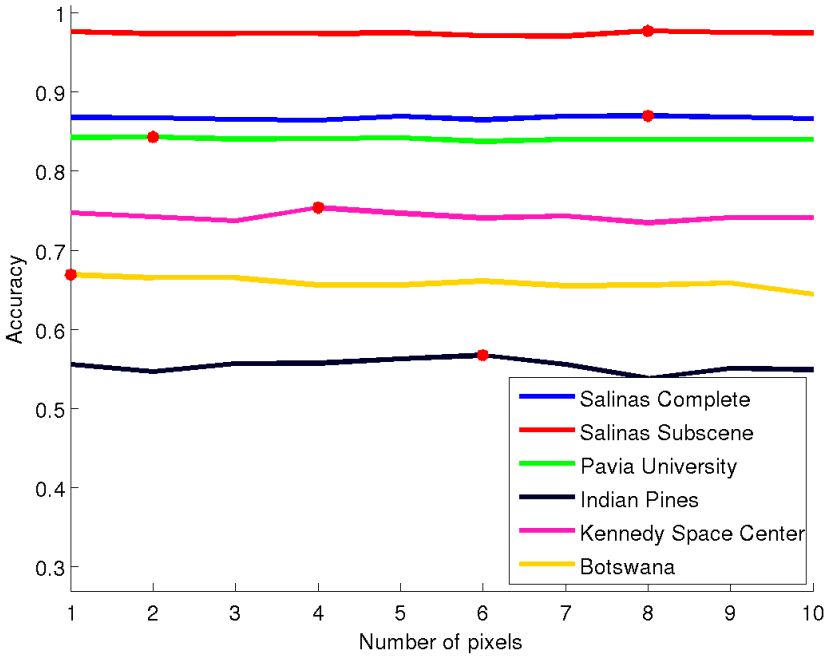


(a)

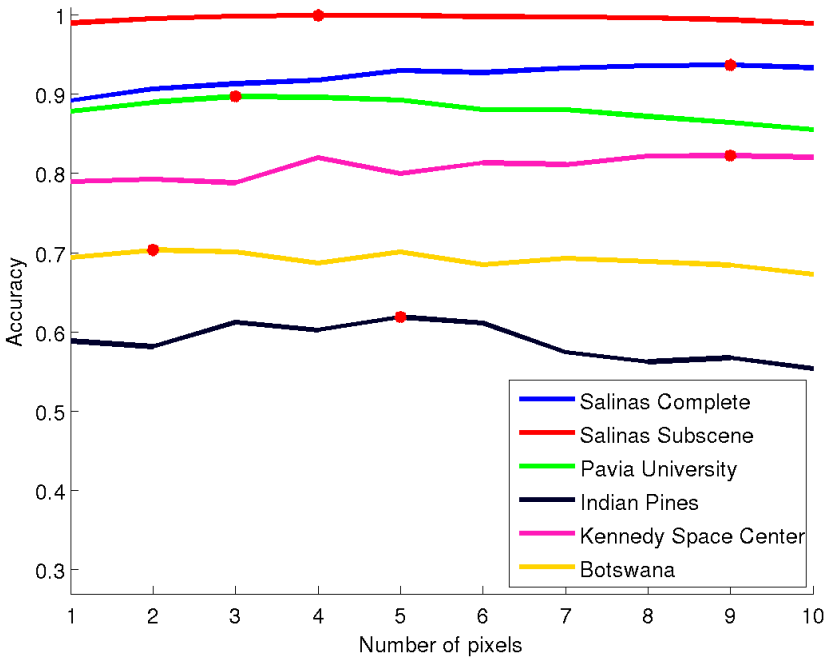


(b)

Figure 3.22: Effect on the average accuracy of the spectral-spatial semi-supervised learning neighborhood radius. Star points are the maxima of the corresponding plot. (a) Accuracy before spatial correction, (b) Accuracy after spatial correction.

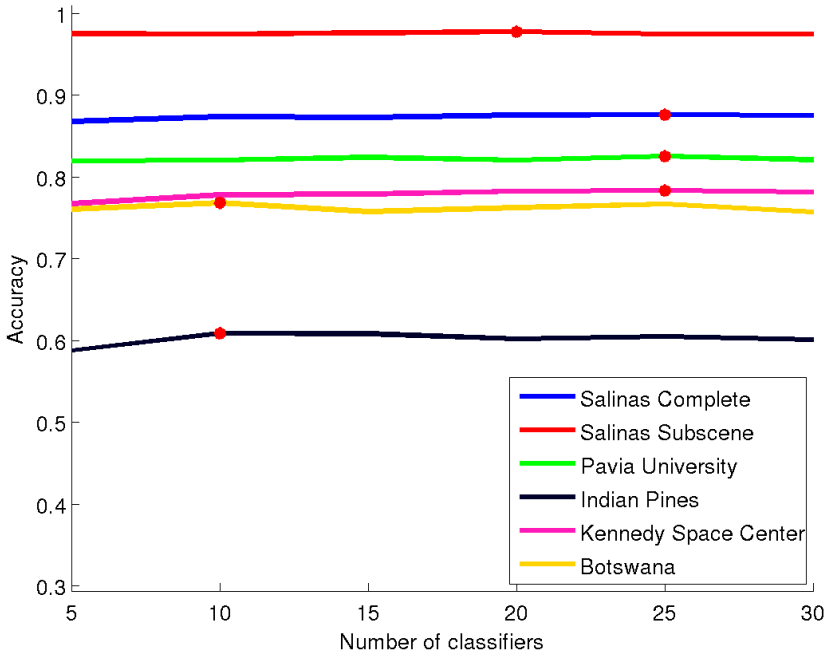


(a)

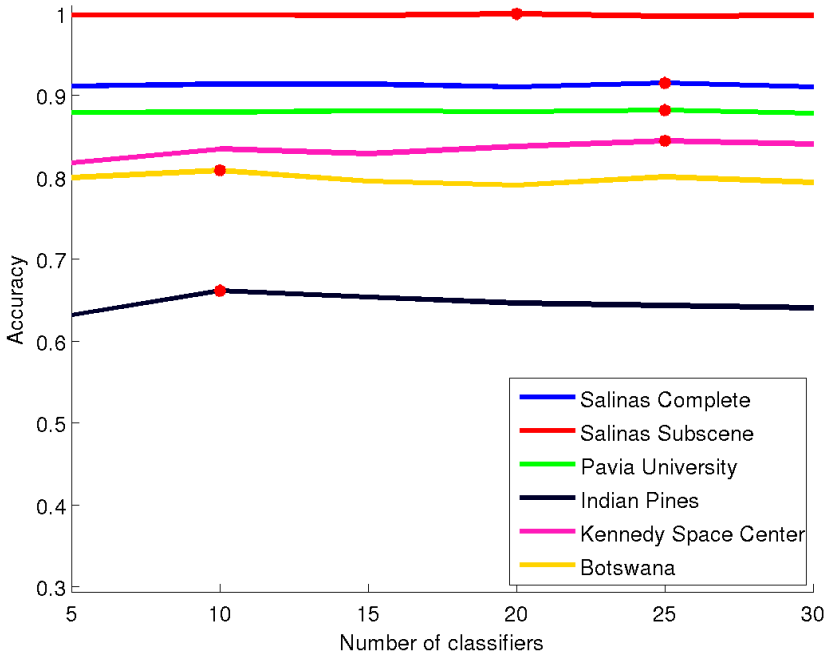


(b)

Figure 3.23: Effect on the average accuracy of the spatial correction neighborhood radius. Star points are the maxima of the corresponding plot. (a) Accuracy before spatial correction, (b) Accuracy after spatial correction.



(a)



(b)

Figure 3.24: Effect on the average accuracy of the AHERF ensemble size. Star points are the maxima of the corresponding plot. (a) Accuracy before spatial correction, (b) Accuracy after spatial correction.

dataset size. It can be observed that using just a 3% of the labeled data may be enough in some cases to achieve really good accuracy performances, comparable with the state-of-the-art ones. The worst case is Indian Pines, which needs larger training datasets to obtain competitive results, though state of the art algorithms also require bigger training datasets.

Effect of the neighborhood radius used in the spectral-spatial semi-supervised correction In Figure 3.22 we show the effect on average accuracy of the size of spatial neighborhood used in the semi-supervised training. Thought the effect of semi-supervised learning with minimal neighborhood radius is a definitive increase of performance, the effect of large radius is small. This may be due to the compactness of the regions, and the effect of boundaries.

Effect of the neighborhood radius used in the a posterior spatial correction In Figure 3.23 we show how the spatial correction neighborhood radius affects the average accuracy. As in previous cases the effect of increasing the neighborhood radius is minimal.

Effect of the ensemble size In Figure 3.24 we plot the average accuracy versus the size of the ensemble AHERF. A small number of classifiers ($\mathscr{C} = 5$) is enough to obtain good results and the effect of increasing the ensemble size is not great.

Summary of the results achieved From the previous sensitivity experiments, we conclude that our proposal is very robust to bad parameter settings. Taking a 1% of the ground truth for training dataset, and setting the parameters to their optimal values we obtain the results in table 3.12, given as the average and standard deviation of the accuracy of 10 repetitions for each dataset before and after spatial correction. The effect of the spatial correction is very strong, in some cases with improvements of more than 15% of accuracy. It is also clear that the Indian Pines dataset is the most difficult one. In fact, results in the literature are obtained with much bigger training datasets.

In Table 3.13, Table 3.14, Table 3.15 and Table 3.16 we show the obtained results before and after spatial correction, respectively, given as the average and standard deviation of the accuracy over 10 repetitions, varying the size of the training dataset, from 0.5% to 5% of the ground truth, with all other parameters set to optimal values. This exploration has been done in order to compare our proposed approach results with other state-of-the-art results, most of them using bigger training datasets.

Table 3.12: Best results achieved over each dataset with optimal parameter settings.

	Before Correction	After Correction
Salinas Complete	92.25 ± 0.59	98.62 ± 0.89
Salinas Subscene	98.78 ± 0.08	99.52 ± 0.02
Pavia University	89.53 ± 0.32	93.86 ± 1.41
Indian Pines	67.23 ± 1.89	79.07 ± 0.76
Kennedy Space Center	84.01 ± 2.48	91.81 ± 0.65
Botswana	86.26 ± 0.85	96.03 ± 0.29

Table 3.13: Best results achieved with optimal parameters for increasing training set size before the spatial correction. Average±standard deviation of the accuracy of 10 repetitions.

	0.5%	1%	1.5%
Salinas Complete	90.43 ± 0.69	92.25 ± 0.59	93.26 ± 0.46
Salinas Subscene	97.77 ± 0.36	98.78 ± 0.08	98.89 ± 0.09
Pavia University	85.89 ± 0.23	89.53 ± 0.32	90.41 ± 0.43
Indian Pines	62.01 ± 2.45	67.23 ± 1.89	70.49 ± 1.81
Kennedy Space Center	77.65 ± 4.24	84.01 ± 2.48	86.33 ± 0.45
Botswana	68.70 ± 0.35	86.26 ± 0.85	86.52 ± 0.46

Table 3.14: Best results achieved with optimal parameters for increasing training set size before the spatial correction. Average±standard deviation of the accuracy of 10 repetitions.

	2%	2.5%	3%
Salinas Complete	93.90 ± 0.23	94.85 ± 0.26	94.80 ± 0.59
Salinas Subscene	98.85 ± 0.08	98.80 ± 0.11	98.77 ± 0.28
Pavia University	91.30 ± 0.45	91.63 ± 0.28	93.00 ± 0.64
Indian Pines	75.75 ± 1.92	75.69 ± 0.96	77.58 ± 1.40
Kennedy Space Center	88.91 ± 0.36	89.83 ± 1.12	91.76 ± 1.86
Botswana	88.10 ± 1.60	88.25 ± 2.61	90.46 ± 2.67

Table 3.15: Results with optimal parameter values and different training set size after the spatial correction. Average±standard deviation of the accuracy of 10 repetitions.

	0.5%	1%	1.5%
Salinas Complete	94.84 ± 0.58	98.62 ± 0.89	99.06 ± 0.71
Salinas Subscene	99.49 ± 0.04	99.52 ± 0.02	99.50 ± 0.02
Pavia University	91.01 ± 1.16	93.86 ± 1.41	94.36 ± 0.19
Indian Pines	64.70 ± 3.42	79.07 ± 0.76	79.29 ± 1.50
Kennedy Space Center	86.51 ± 5.10	91.81 ± 0.65	91.23 ± 0.12
Botswana	76.08 ± 0.86	96.03 ± 0.29	92.96 ± 0.97

Table 3.16: Results with optimal parameter values and different training set size after the spatial correction. Average±standard deviation of the accuracy of 10 repetitions.

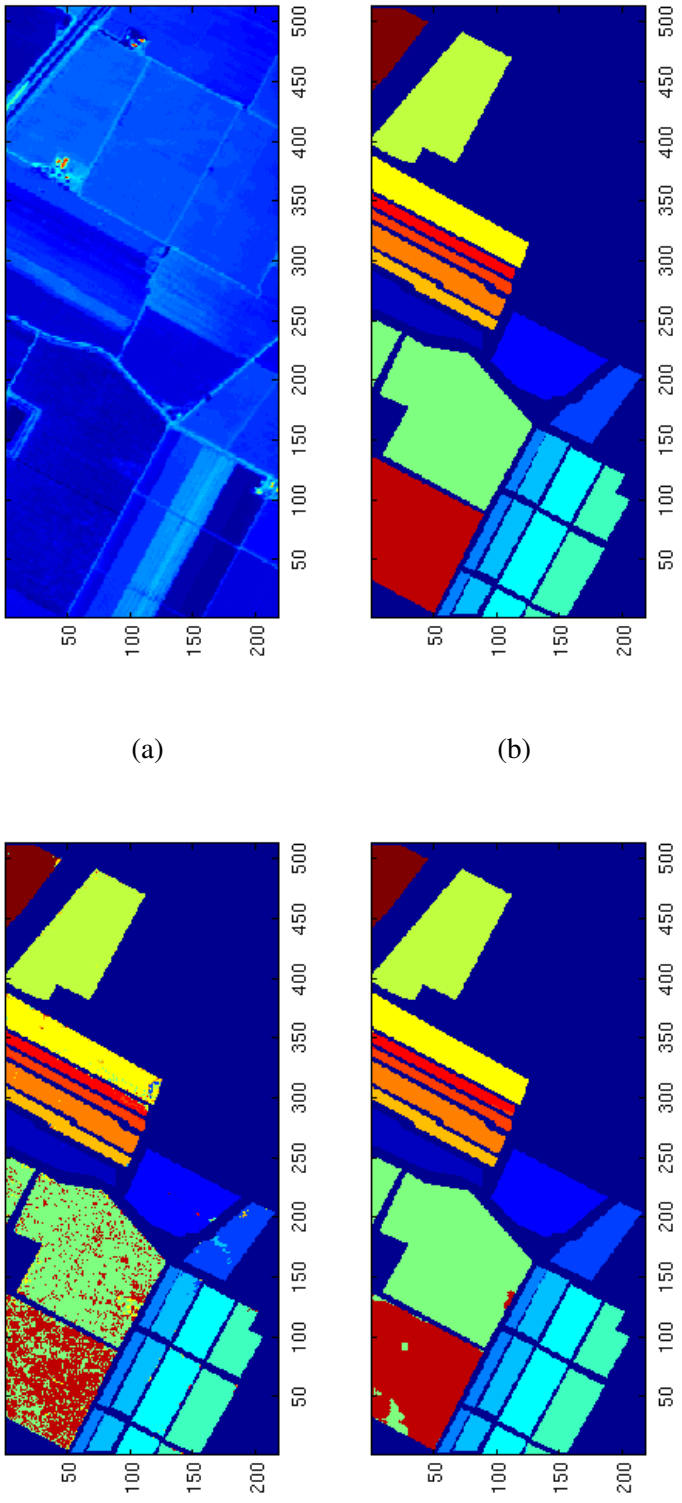
	2%	2.5%	3%
Salinas Complete	99.31 ± 0.22	99.84 ± 0.04	99.34 ± 0.62
Salinas Subscene	99.52 ± 0.02	99.54 ± 0.01	99.47 ± 0.03
Pavia University	95.21 ± 0.98	95.28 ± 0.67	96.19 ± 0.38
Indian Pines	83.10 ± 0.22	81.68 ± 0.84	81.09 ± 2.78
Kennedy Space Center	95.82 ± 0.58	96.84 ± 0.42	97.94 ± 0.51
Botswana	97.43 ± 0.90	94.52 ± 1.42	97.70 ± 1.31

Comparison with other approaches in the literature As mentioned before when comparing to the state-of-the-art methods we found that the % of training data used by them was much bigger in most of the cases. Anyway as we show in this subsection it depends a lot on the dataset. In Table 3.17 we gather overall accuracy (OA) results obtained by several authors on the same benchmark datasets. The AHERF refers to the approach proposed in this thesis work. We give also the training dataset size in order to stress the dependence of results to this parameter. In the case of Indian Pines, we extended the exploration of the effect of training dataset size in order to compare with the literature. Exception made of the Indian Pines dataset, our approach ranks first or second with very small training datasets. The reason of the bad results on Indian Pines may be related to the noise and the not very compact regions of some classes.

Visual Results Besides the quantitative evaluation of the results, we provide visual results of the classification and spatial correction of some images. We show (a) a sample band image, (b) the available ground truth, (c) regions segmented before spatial correction, and (d) regions segmented after spatial correction. Figures 3.25, 3.27, 3.28, and 3.29 present these images for the Salinas, Pavia University, Indian Pines and KSC, respectively. The effect of the spatial correction is very clear in all the images.

Table 3.17: Comparison of results in the literature. OA Overall accuracy, %Train percentage of ground truth for seed training dataset.

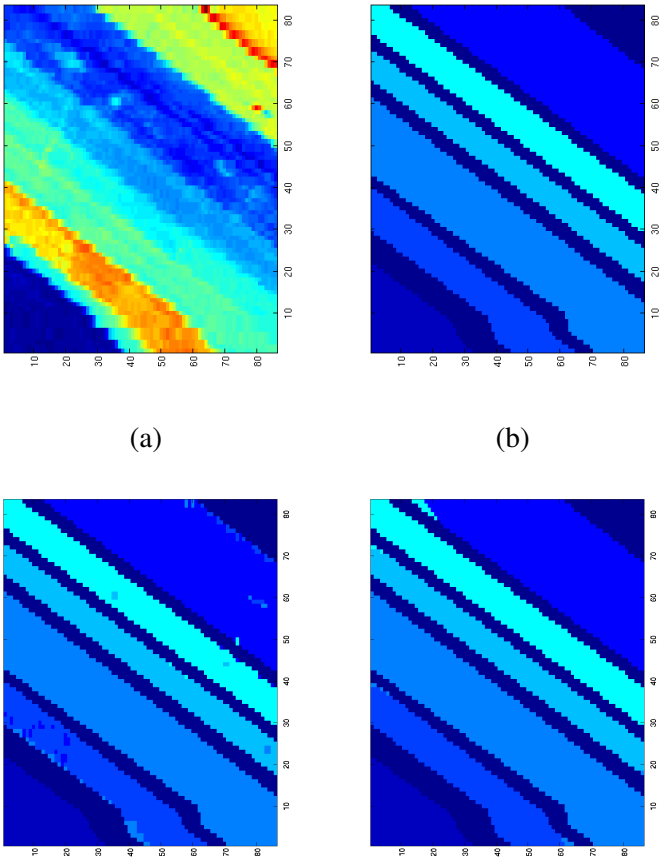
Botswana			Kennedy Space Center			Indian Pines			Univ. Pavia		
	OA	**% Train**		**OA**	**% Train**		**OA**	**% Train**		**OA**	**% Train**
AHERF	97.70	3,00%	**AHERF**	97.94	3,00%	**KSPCK[1]**	98.47	10,00%	**EMP +KPCA[116]**	96.55	9,00%
AHERF	97.43	2,00%	**AHERF**	96.84	2,50%	**KOMPCK[1]**	98.33	10,00%	**AHERF**	96.19	3,00%
PC-SVM[154]	96.26	50,00%	**BHC-soft[101]**	96.80	50,00%	**KSOMP[1]**	97.33	10,00%	**AHERF**	95.28	2,50%
AHERF	96.03	1,00%	**AHERF**	95.82	2,00%	**MPM-LBP[111]**	94.76	10,00%	**AHERF**	95.21	2,00%
PCA-SVM[154]	95.55	50,00%	**RF-BHC[67]**	93.75	20,00%	**LORSAL-MLL[111]**	92.72	10,00%	**SVA + ICA [133]**	94.47	9,00%
SFS-SVM[154]	94.72	50,00%	**RS-BHC[67]**	92.00	20,00%	**AHERF**	91.37	20,00%	**AHERF**	94.36	1,50%
AHERF	94.52	2,50%	**AHERF**	91.81	1,00%	**AHERF**	91.32	15,00%	**BoostELM[149]**	94.30	15,00%
RF-BHC[67]	94.00	20,00%	**AHERF**	91.23	1,50%	**6-SVM[25]**	90.80	25,00%	**AHERF**	93.86	1,00%
DBFE-SVM[154]	93.80	50,00%	**BB-BHC[67]**	88.50	20,00%	**AHERF**	90.32	10,00%	**AHERF**	91.01	0,50%
AHERF	92.96	1,50%	**RF-CART[67]**	88.00	20,00%	**6-SVM[25]**	86.70	12,50%	**CENA[37]**	88.47	2,00%
RS-BHC[67]	92.00	20,00%	**PCA-SVM[154]**	87.78	50,00%	**SVM[1]**	84.52	10,00%	**class 2vs4 [37]**	88.47	8,00%
BB-BHC[67]	88.50	20,00%	**AHERF**	86.51	0,50%	**AHERF**	83.10	2,00%	**KSSP[1]**	87.65	9,00%
RF-CART[67]	87.00	20,00%	**K-NN[139]**	85.69	15,00%	**LORSAL[111]**	82.60	10,00%	**SVMCK[1]**	87.18	9,00%
AHERF	76.08	0,50%	**SSNN[142]**	84.74	19,19%	**6-SVM[25]**	82.20	6,25%	**SVM-CK[159]**	87.11	9,00%
						AHERF	81.68	2,50%	**MPM-LBP[111]**	85.78	0,68%
						AHERF	81.08	3,00%	**LORSAL-MLL[111]**	85.57	0,68%
						SVM[111]	80.56	10,00%	**Watershed[111]**	85.42	0,68%
						AHERF	79.29	1,50%	**EMP/SVM[111]**	85.22	0,68%
									SVM +SC[167]	83.05	8,00%
									SVM[37]	81.73	2,00%
									SVM[111]	80.99	0,68%
									LORSAL[111]	80.11	0,68%
									D[178]	73.70	1,05%



(a) (b)

(c) (d)

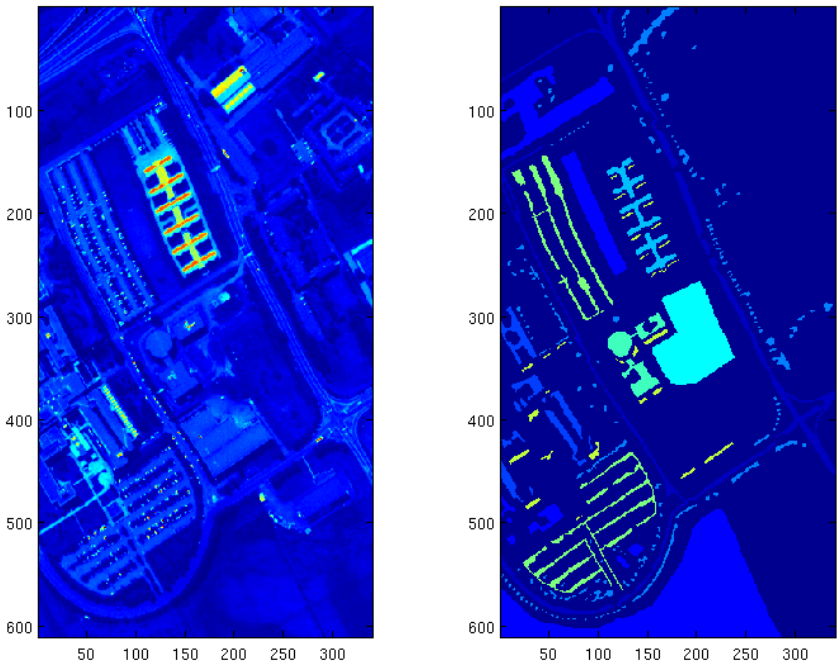
Figure 3.25: Salinas complete scene visual results: (a) sample image band, (b) ground truth, (c) classification results before spatial correction. (d) classification results after a posteriori spatial correction.



(a) (b)

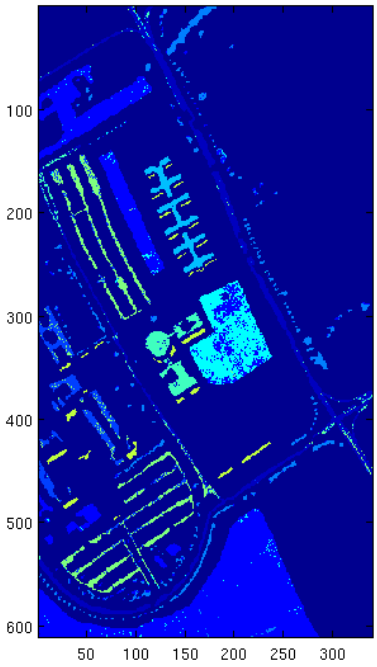
(c) (d)

Figure 3.26: Salinas Subscene A visual results: (a) sample image band, (b) ground truth, (c) classification results before a posteriori spatial correction. (d) (c) classification results before a posteriori spatial correction.

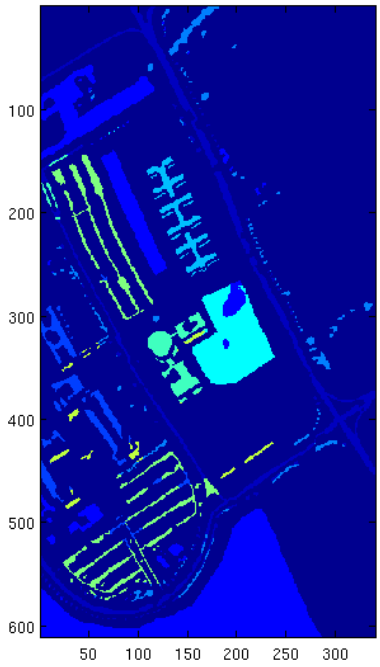


(a)

(b)

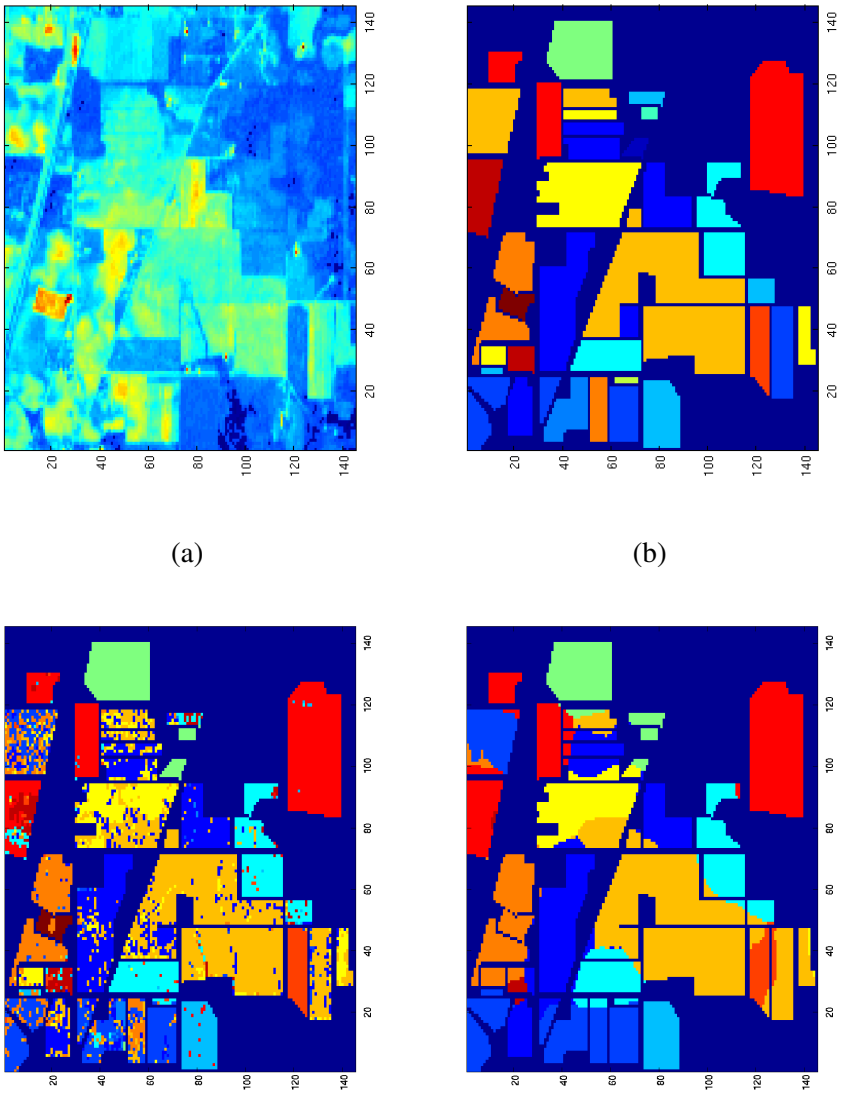


(c)



(d)

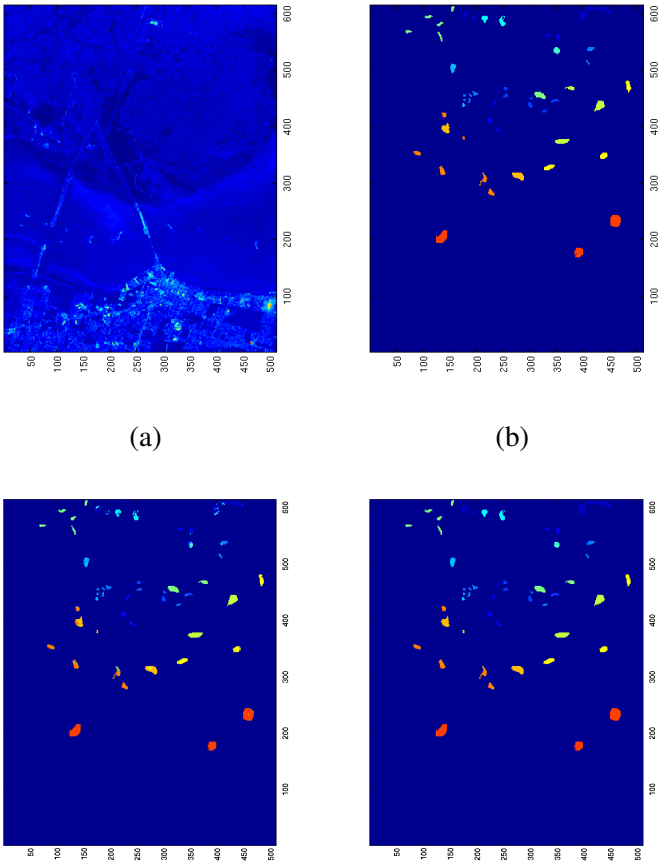
Figure 3.27: Pavia University visual results: (a) sample image band, (b) ground truth, (c) classification results before spatial correction. (d) classification results after spatial correction.



(a) (b)

(c) (d)

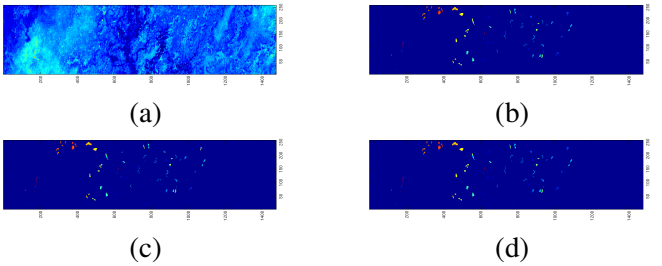
Figure 3.28: Indian Pines visual results: (a) sample image band, (b) ground truth, (c) classification results before spatial correction. (d) classification results after spatial correction.



(a) (b)

(c) (d)

Figure 3.29: Kennedy Space Center visual results: (a) sample image band, (b) ground truth, (c) classification results before spatial correction. (d) classification results after spatial correction.



(a) (b)

(c) (d)

Figure 3.30: Botswana visual results: (a) sample image band, (b) ground truth, (c) classification results before a posteriori spatial correction. (d) (c) classification results before a posteriori spatial correction.

Chapter 4

Medical Image Analysis Applications

In this Chapter we provide an account of our works related to medical image analysis applications. We have covered in essence two kind of problems: blood vessel segmentation and classification of brain MRI images for computer aided diagnosis. The first problem is highly unbalanced, so that we have favored an Active Learning approach. The second is very high dimension, so that we have focus on image feature extraction that preserves effect localization. First we provide some state of the art views on each problem tacked in Section 4.1. Section 4.2 reports results on various classifiers trained by Active Learning to solve the AAA segmentation problem. Section 4.3 refers to retinal image segmentation by Active Learning. Finally, Section 4.4 reports the experiments on brain image classification.

4.1 Introduction

In this Section we review the state of the art related to the three kind of medical image analsyis applications that have been approached along the Thesis works. They cover topics which have some overlapping with the works developed by other members of the research group, hence some publications are naturally shared. We give a brief description of the data in the Appendix of this documen. Here we offer some hints of the state of the art on these images at the time the works were carried out.

4.1.1 Abdominal Aortic Aneurysm (AAA)

Classification based approaches to AAA thrombus segmentation allow to learn the optimal segmentation from the data. A classification approach that needs an initial

manual segmentation of the Aorta lumen [136]; an active shape model that uses the classification of grayscale profiles to move the active contour [47]; Support Vector Machine (SVM) is used to drive a level set segmentation [199]; a Gaussian Mixture Model probability map drives a deformable NURBS model in [48]; finally, a segmentation based on voxel classification using Random Forests is proposed in [121, 120].

Abdominal Aortic Aneurysm (AAA) is the concrete medical domain of application of the AL based CTA image segmentation. AAA is a local dilation of the Aorta happening between the renal and iliac arteries. The weakening of the aortic wall leads to its deformation and the generation of a thrombus. 3D Contrast CTA is the preferred imaging method allowing minimally invasive visualization of the Aorta's lumen, thrombus and calcifications. Even though several segmentation methods for vascular structures have been developed [107] [118, 119], the segmentation of the AAA thrombus is still a challenging task due to the low contrast of signal intensity values between the aneurysm thrombus and its surrounding tissue as can be appreciated in Fig. 4.1. Furthermore, prior information is useless because of its great shape variability. Several AAA thrombus segmentation methods have been recently developed. The method by De Bruijne et al.[47] is an interactive contour tracking method for axial slices; Olabarriaga et al. [136] employ a deformable model approach based on a nonparametric statistical grey-level appearance model to determine the deformable model adaptation direction starting from a lumen contour shape interactive segmentation; Zhuge et al. [199] present a level-set segmentation based on a parametric statistical model; Demirci et al. [48] propose a deformable B-spline parametric model based on a nonparametric intensity distribution model and; Freiman et al. [58] apply a an iterative model-constrained graph-cut algorithm. All these methods involve at least as much user interaction as the AL approach, either for initialization or during the algorithm evolution.

In this thesis, AAA thrombus segmentation on CTA data volumes is tackled as a voxel classification problem into two classes: aortic thrombus or background. This kind of data poses ver badly balanced problems [151]. Specifically, in this thesis work the following generalization validation approach has been followed on each patient dataset: First, train a HERF classifier on the volume's axial central slice using AL. Second, apply the resulting classifier to the remaining slices of the volume, computing the classification accuracy for each slice independently.

4.1.2 Retinal images

Thought some taxonomy of methods can be outlined, attending to the main claim by the authors, most of the approaches combine several procedures. Here we summarize some of the procedures found in the literature a as an example. A matched

filter is a image pattern of the expected appearance of some image structure. The response by the matched filters can then be processed by morphological processes, such as thinning, to obtain estimations of the vessel branches which are combined heuristically. [74]. Kande et al. [93] compute a collection of responses by rotated matched filters which are then aggregated as features for fuzzy clustering to obtain the thresholding needed for detection. A connected components filter is applied to remove isolated detections. Ng et al. [134] compute a bank of second derivative Gaussian filters at different scales. The inverse model tries to find the maximum likelihood estimates of the parameters of the vessels in the image, including a noise term which is estimated from the image. Post-processing is done on the basis of the estimation that the result can be obtained by pure noise sampling. Morphological processes, such as multi-directional top-hat filters, have been also been combined with centerline detection in order to obtain a map of vessel detections in [57]. Final cleaning removes isolated non-vessel pixels.

The approach of supervised classification needs some features extracted from the spatial information in the image. The adaboost classifier in [117] uses a battery of 41 features that include gaussian based vesselness and ridgeness features, 2D Gabor features and curvature measures. The Support Vector Machine in [185] uses the residuals of wavelet and curvelet multi-scale transforms after a thresholding as the classification features. The result of the classification is subject to thinning and line tracking to find the vessel network structure.

In our approach, we compute several simple local spatial features over the green band of the image, which contains most of the contrast information. The size of the local window is related to the detection scale. These features are then presented to a classifier to obtain image segmentation. The classifier is built following an active learning approach, where the most uncertain unlabeled pixels are presented to a human operator which labels them for addition to the training dataset and classifier retraining.

4.1.3 Alzheimer's Disease (AD)

Machine learning methods have become very popular to classify functional or structural brain images to discriminate them into two classes: healthy control or suffering a specific neurodegenerative disorder [38]. The development of automated detection procedures based in Magnetic Resonance Imaging (MRI) and other medical imaging techniques [45] is of high interest to help the disgnosis in clinical medicine.

Studies using ensemble of classifiers on brain anatomical MRI can be found in the literature. In [34] they propose a classification method via aggregation of regression algorithms fed with histograms of deformations generated from the Open

Access Series of Imaging Studies (OASIS) database obtaining a 0.04 test error rate. Another study shows a local patch-based subspace ensemble method which builds multiple individual classifiers based on different subsets of local patches with the sparse representation-based classifier obtaining an accuracy of 90.8% on the ADNI database [115]. In [172] subsets of ranked features from neuroimaging data are used to in an ensemble of linear Support Vector Machine (SVM) classifiers obtaining 0.94 of Area Under the Receiver Operating Characteritic (ROC) Curve (AUC) when detecting a AD patients vs. control subjects.

We have used modulated Grey-Matter (GM) maps partitioned according to the regions from the Automatic Anatomical Labeling (AAL) atlas to create datasets of statistical features of these GM maps within each of these regions for each subject. These datasets are put into a leave-one-out with grid search process for classifier validation. After that, an ensemble collective decision is made in order to obtain a classification result. We report the results of an ensemble of linear Support Vector Machine (SVM) classifiers.

4.2 Active Learning for AAA Segmentation in CTA Images

Computerized Tomography Angiography (CTA) image segmentation is posed as a classification problem, where each pixel is assigned to an anatomical structure on the basis of its intensity and other features extracted from the image. CTA variability due to noise, changes in subject, calibration of the hardware and others, mean that often a classifier has little generalization capabilities outside the dataset used for training and validation. Therefore, CTA segmentation fits in the paradigm of Active Learning (AL), where a human operator must label the misclassified data for retraining. Of course to provide cross-validation results in this thesis, we resort to the use a given image ground truth to obtain the labels for the data samples added to the train set.

Segmentation problem: We are looking for the segmentation of the thrombus in the AAA formed after the placement of the endo-protesis. Figure 4.1 shows the localization of the thrombus, aorta lumen in an axial slice of the abdomen. Therefore, we deal with a two-class problem, which is extremely umbalanced because the thrombus is a very small percentage of the whole image. We have been applying several classification approaches withing an Active Learning environment.

The experimental setup is illustrated in Fig. 4.2. We load the complete CTA volume data, computing first the feature vectors on each voxel. The feature selection and feature extraction processes are the same as in [121, 39, 120]. A single axial slice situated approximately at the center of the thrombus is selected to perform

the Active Learning construction of the voxel classifier. Next, the voxel classifier is applied to all remaining slices of the CTA volume, obtaining an identification of the regions detected as thrombus by this classifier. Expert Domain Knowledge can be applied to post-process the detection results, removing spurious detections, as done in Section 4.2.3. Finally, we perform a volume rendering showing the quality of the thrombus detection. The Active Learning oracle in the experiments is the ground truth provided by manual segmentation.

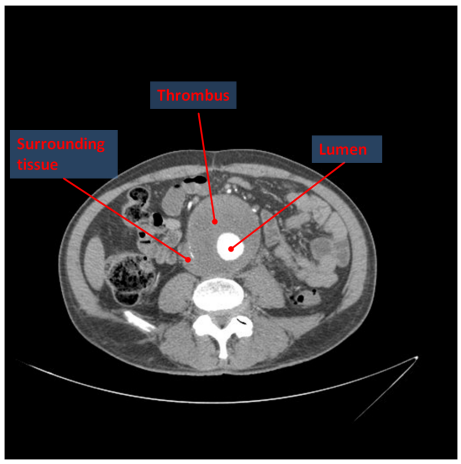


Figure 4.1: Axial view of thrombus and lumen in a CTA orthoslice using the contrast agents, blood in lumen is highlighted but thrombus intensity levels are similar to other surrounding tissue.

Next we proceed with the description of the feature selection process carried out previously to the classification experiments, which is common to all tested approaches.

4.2.1 Feature selection based on variable importance

Let us denote x_j the j-th feature of the feature vector. The HERF variable importance of x_j is defined as follows. For each classifier $h(\mathbf{x};\psi_t)$ of the HERF ensemble, consider the associated out-of-box OOB_t dataset [18] constituted by data samples not included in the bootstrap sample used to construct $h(\mathbf{x};\psi_t)$. Denote e_{OOB_t} the error corresponding to the miss-classification rate for classification of the single tree $h(\mathbf{x};\psi_t)$ on the OOB_t dataset.

Next, randomly permute the values of x_j in OOB_t to get a perturbed sample denoted by $\widetilde{OOB_t}^j$ and compute $e_{\widetilde{OOB_t}^j}$, the error of $h(\mathbf{x};\psi_t)$ on the perturbed sample.

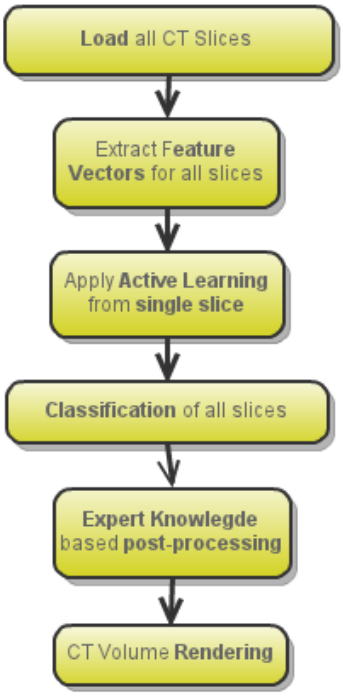


Figure 4.2: Pipeline of the experimental setup for the Active Learning enhanced with Domain Knowledge segmentation process

Table 4.1: Feature importance ranking for the first 10 features selected, specifying the operator used (O), neighborhood radius (R) and the variable importance (VI) value. Max, Med, GA correspond to Maximum, Median and Gaussian weighted average, respectively

	O	R	VI
#1	Max	16	1.277
#2	Max	4	0.953
#3	Max	8	0.9531
#4	Med	8	0.803
#5	Max	2	0.762
#6	Max	1	0.759
#7	Med	1	0.741
#8	Med	4	0.740
#9	Med	16	0.732
#10	GA	4	0.725

Variable importance of x_j is then equal to:

$$VI(x_j) = \frac{1}{T}\sum_t (e_{\widetilde{OOB_t}^j} - e_{OOB_t}), \tag{4.1}$$

where T denotes the number of classifiers of the ensemble. For the feature selection, we order the features by decreasing value of $VI(x_j)$, we compute the total accumulated variable importance of the features,

$$TI = \sum_{j=1}^{N} VI(x_j). \tag{4.2}$$

We discard features falling behind the 95% of the TI value in the ordered list of features. We select 10 features given in Table 4.1.

4.2.2 Experimental results on AAA Thrombus segmentation by HERF

Parameter tuning: After performing a complete sensitivity of the HERF classifier training over a single slice, we have set the system parameters as defined in Table 4.2. The data rotation method used is the Quartimax.

Experiment definition: To test the generalization of a single slice classifier, we build just one HERF classifier from the data of the central slice of the aneurysm, and we test its generalization to the remaining slices of the CT volume.

Image features before selection: In this thesis work the features initially associated with CTA voxels are: its coordinates in the data domain grid, the voxel

Table 4.2: Parameter settings of the HERF for AAA thrombus segmentation by Active Learning (AL).

Percentage of ELMs in the ensemble	33%
Percentage of Decision Trees in the ensemble	67%
Number of hidden nodes in each ELM	14
activation function in ELM	sigmoid
Number of samples added per AL iteration	5

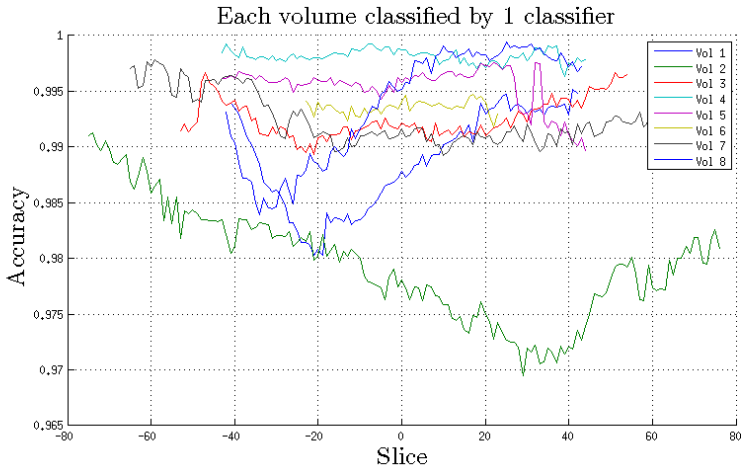


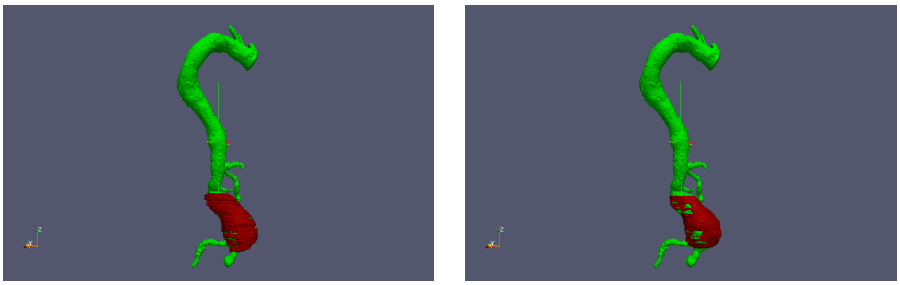
Figure 4.3: Generalization result on all the datasets. The classifier learned on the central slice is applied to the remaining slices.

intensity, the mean, variance, maximum and minimum of the voxel neighborhood, for the different values of the neighborhood radius in the set $\{1,2,4...2^n\}$.

Validation measure. The performance measure results of the experiments is the classification accuracy. We show also the 3D reconstruction of one of the segmented volumes.

Segmentation accuracy: Fig. 4.3 shows the overlaid accuracy plots at each CTA volume slice applying the HERF classifier trained on the thrombus central slice for each of the 8 CTA volumes in the experiment. The abscissa's zero value corresponds to the central slice, the negative abscissa values correspond to slices above the central slice, the positive values correspond to slices below the central slice. There is some variability of the plots' span, due to the different sizes of the thrombus in each patient. As can be expected, the drop in classification accuracy is symmetric in most cases. The generalization results are very good: the worst accuracy is above 0.97 in almost all cases. This results comes with a hundred-fold reduction of learning complexity involving computer and human operator time.

Segmentation visualization: A 3D volume rendering of the Aorta's lumen (green)



(a) (b)

Figure 4.4: Volume rendering of aortic lumen (green) and thrombus (red) obtained from the segmentation of one CTA volume. (a) manual segmentation defining the ground truth, (b) result of applying the HERF classifier trained by Active Learning on the central slice to the remaining slices.

and thrombus (red) of one patient is shown in Fig.4.4 in three situations. Fig.4.4(a) shows the rendering of the ground truth given by volume manual segmentation. Fig.4.4(b) shows the result of the segmentation based on the Active Learning performed on each slice. The high accuracy of the segmentation is evident from the comparison with the rendering of the manual segmentation. Fig. 4.4(c) shows the result of the segmentation based on the HERF classifier built from the thrombus' central slice.

4.2.3 Enhancing Segmentation using Domain Knowledge

This Section enhances the previously proposed Active Learning image segmentation system with Domain Knowledge. Active Learning iterates the following process: first, a classifier is trained on the basis of a set of image features extrated for each training labeled voxel; second, a human operator is presented with the most uncertain unlabeled voxels to select some of them for inclusion in the training set assigining corresponding label. Finally, image segmentation is produced by voxel classification of the entire volume with the resulting classifier. The approach has been applied to the segmentation of the thrombus in CTA data of Abdominal Aortic Aneurysm (AAA) patients. The Domain Knowledge referring to the expected shape of the target structures is used to filter out undesired region detections in a post-processing step. We report computational experiments over 6 abdominal CTA datasets consisting. The performance measure is the true positive rate (TPR). Surface rendering provides a 3D visualization of the segmented thrombus. A few Active Learning iterations achieve accurate segmentation in areas where it is difficult to distinguish the anatomical structures due to noise conditions and similarity of gray levels between the thrombus and other structures.

4.2.3.1 Domain Knowlegde

The use of Domain Knowledge allows to post-processing the results of the classification in order to remove spurious detections. This Domain Knowledge consists in the following rules for the specific detection of the thrombus in AAA images:

- At each axial slice, the thrombus is composed of only one connected component. We can remove all connected components disconnected from the one that is more likely to be the thrombus, which is identified by the following rules.
- Thrombus has a roughly circular shape in any axial cut of the volume.
- In succesive slices moving away from the thrombus middle slice the radius of the thrombus region decreases.
- In successive axial slices, the thrombus region overlap is large (between 80% and 90% of the area).
- The 2D coordinates of the centroid of the thrombus region have a small (smooth) variation between successive slices.

These rules allow us to perform a heuristic post-processing of the classification results which show a dramatic increase in detection in some cases. These rules do not need any specific parameter tuning and are easily implementable.

4.2.3.2 Experimental setup

Datasets. We have performed computational experiments over 6 datasets to test the proposed Active Learning enhanced with Expert Knowlegde based image classification approach. Each dataset consists in real human contrast-enhanced datasets of the abdominal area with 512x512 pixel resolution on each slice. Each dataset consists of between 216 and 560 slices and 0.887x0.887x1 mm spatial resolution corresponding to patients who suffered Abdominal Aortic Aneurysm. The datasets show diverse sizes and locations of the thrombus. Some of them have metal streaking artifacts due to the stent graft placement. Ground truth segmentations of the thrombus for each dataset that simulates the human oracle providing the labels for the voxels, was obtained manually by a clinical radiologist.

Parameter tuning. We train a RF classifier with a single slice a to test the sensitivity of the forest parameters: the number of the trees T and their depth D. The increase in performance stabilizes around number of trees = 80 and depth = 20. Once we get the optimal parameters and feature set, we perform the experiment to test our method in the patients CT volumes as illustrated in figure 4.2.

Validation. The performance measure results of the experiments are the post-processing average True Positive Rate (TPR).

4.2.3.3 Experimental Results with domain knowledge enhancement

We have performed computational experiments over 6 datasets to test the proposed approach. Each dataset consists in real human contrast-enhanced datasets of the abdominal area with 512x512 pixel resolution on each slice. Each dataset consists of a number of slices between 216 and 560, and 0.887x0.887x1 mm spatial resolution corresponding to patients who suffered Abdominal Aortic Aneurysm. The datasets show diverse sizes and locations of the thrombus. Fig. 4.5 shows the performance of the Active Learning based image segmentation algorithm for CT volumes of AAA patients, plotting the average True Positive Rate (TPR) versus the slice number (relative to the middle slice of the thrombus used for training) of the RF classifiers trained with Active Learning without (red) and with (blue) the application of the heuristic postprocessing rules derived from Domain Knowledge. In most of the cases, the Domain Knowledge based post-processing provides some improvement, mostly in the slices that fall far away from the middle slice.

A 3D volume rendering of the Aorta's lumen (green) and thrombus (red) of one patient is shown in Fig.4.6. Fig.4.6(a) shows the rendering of the ground truth given by volume manual segmentation. Fig.4.6 (b) shows the result of the segmentation based on the Active Learning enhanced with Domain Knowledge classifier built from the thrombus' central slice. The structure of the thrombus is well delineated and fits almost perfectly to the ground truth.

4.2.4 Active Learning with Bootstrapped Dendritic Classifier applied to medical image segmentation

The basic classifier in this section is the Bootstrapped Dendritic Classifier (BDC), which combine the output of an ensemble of weak Dendritic Classifiers by majority voting. Weak Dendritic Classifiers are trained on bootstrapped samples of the train data setting a limit on the number of dendrites. We validate the approach on the segmentation of the thrombus in 3D Computed Tomography Angiography (CTA) data of Abdominal Aortic Aneurysm (AAA) patients simulating the human oracle by the provided ground truth. The generalization results in terms of accuracy and true positive ratio of the classification of the entire volume by the classifier trained on one slice confirm that the approach is worth its consideration for clinical practice.

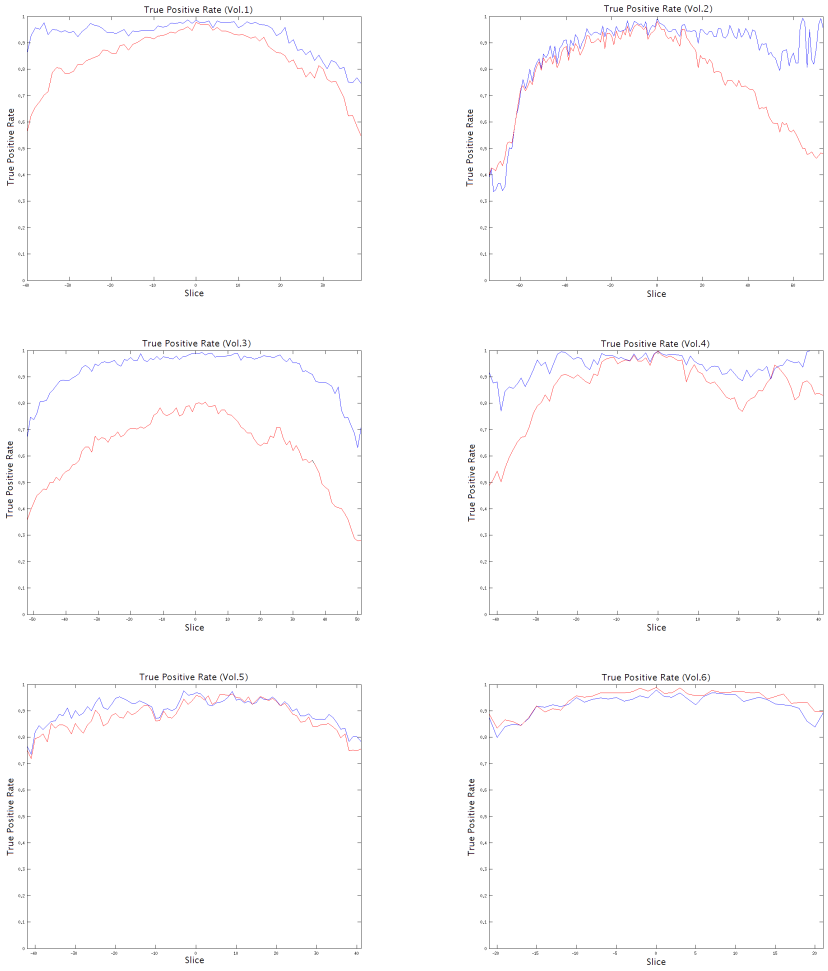
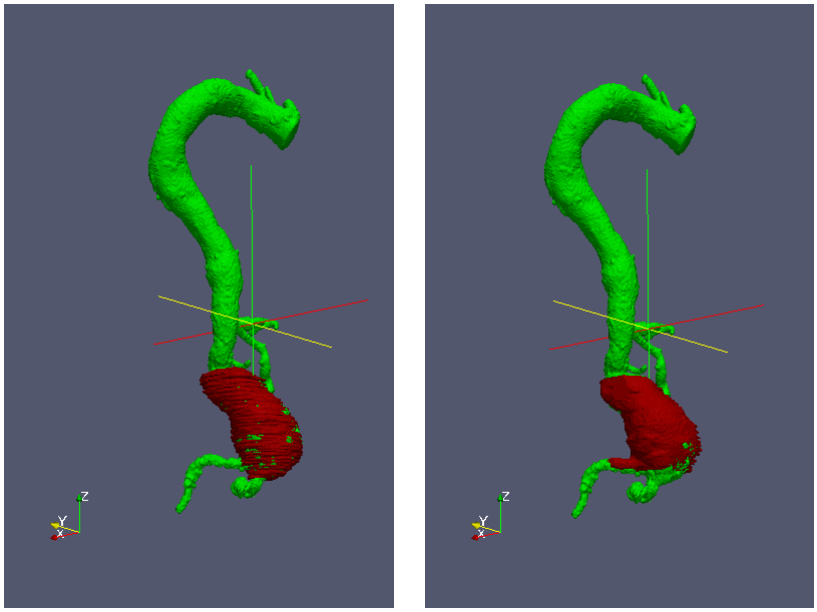


Figure 4.5: True Positive Rates for all volumes treated. Red curves corresponds to RF results trained with Active Learning, and blue curves to the Domain Knowledge post-processing.



(a) (b)

Figure 4.6: Volume rendering of aortic lumen (green) and thrombus (red) obtained from the segmentation of one CT volume. (a) manual segmentation of the ground truth, (b) result of Active Learning trainning of RF classifier, enhanced with Domain Knowledge post-processing rules.

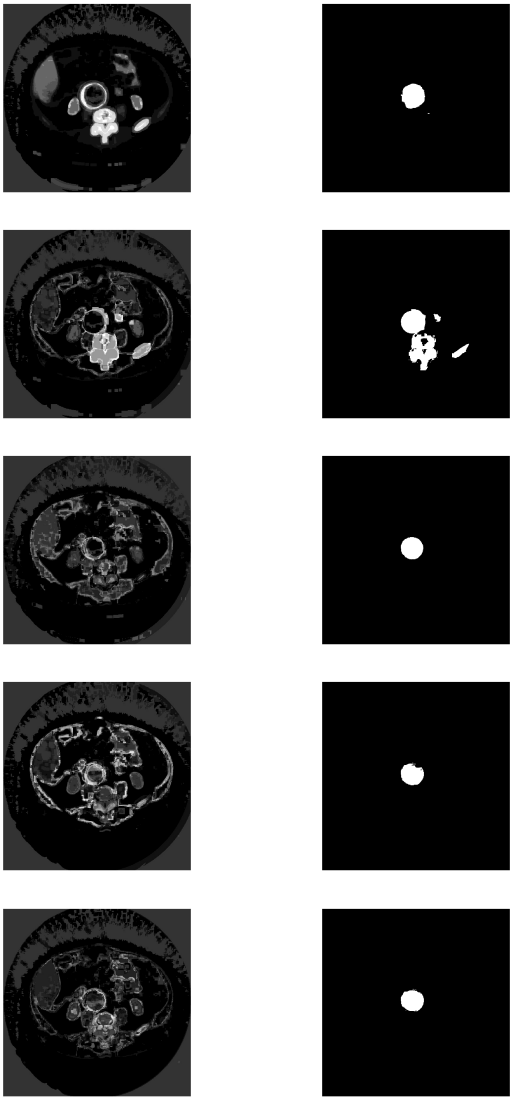


Figure 4.7: Evolution of the active learning process in the central slice of one of the experimental volumes under study, shown at learning iterations 1, 5, 10, 15, and 20. Left column corresponds to the uncertainty value of each voxel. Right column shows the actual thrombus segmentation obtained with the classifier built at this iteration.

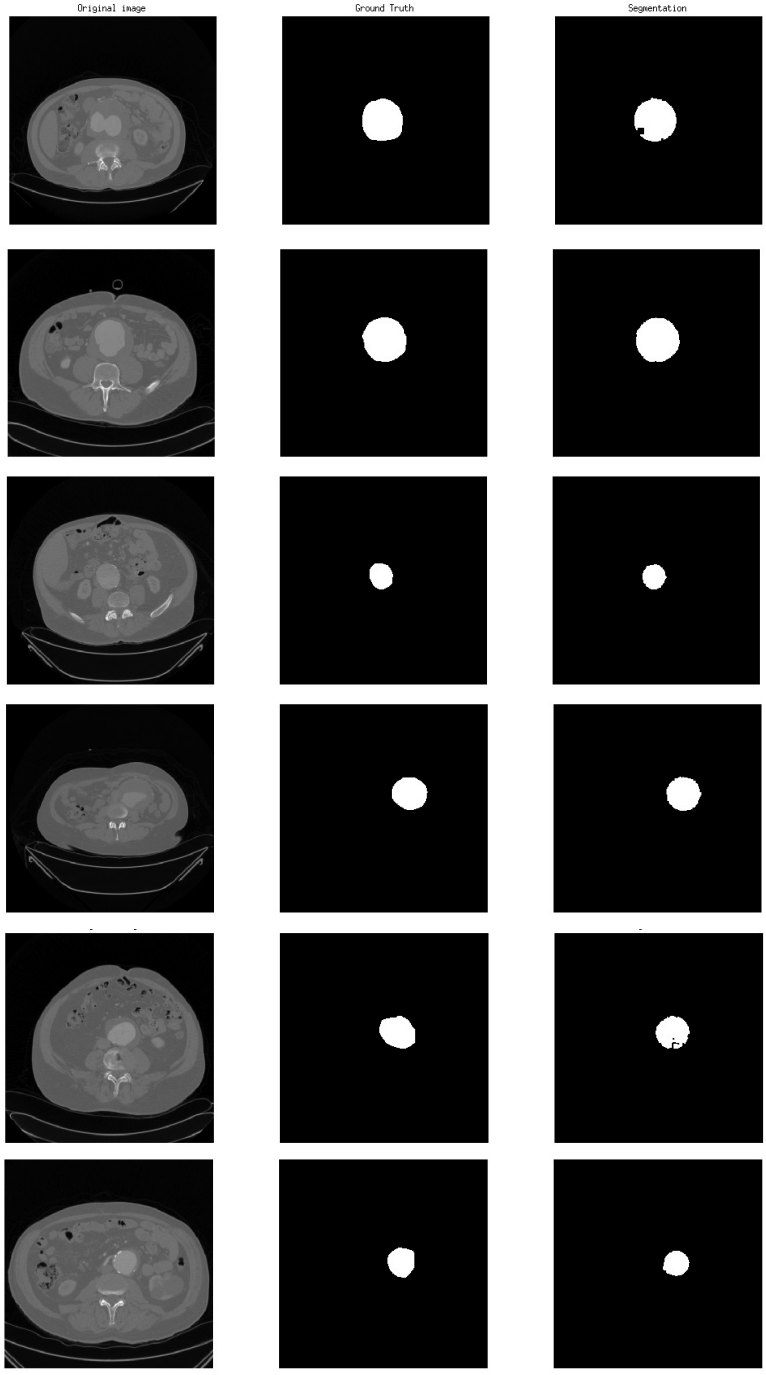


Figure 4.8: Segmentation results in the central slice of the CTA volumes under study after active learning construction of the classifiers. Left column original slice, middle column provided ground truth, right column segmentation achieved by the classifier.

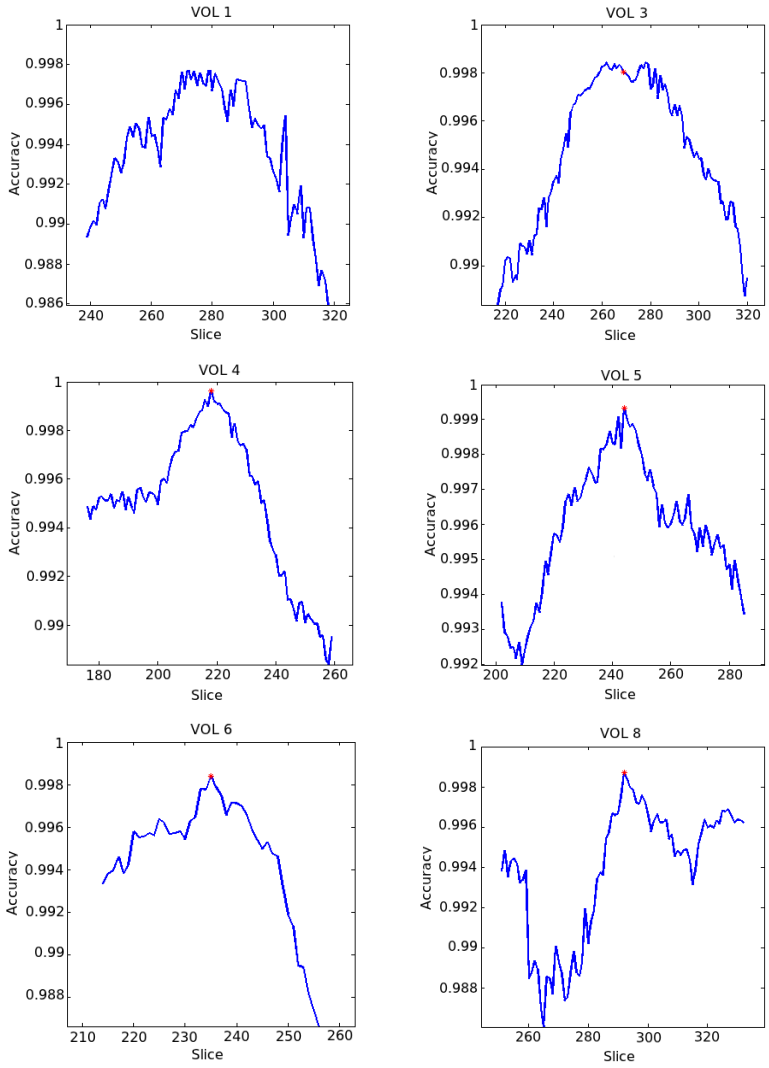


Figure 4.9: Accuracies obtained on the remaining axial slices by the BDC classifier trained on the central axial slice of each of the CTA volumes. Slice numbers are the actual numbers in the volume. The red asterisk identifies the central slice result.

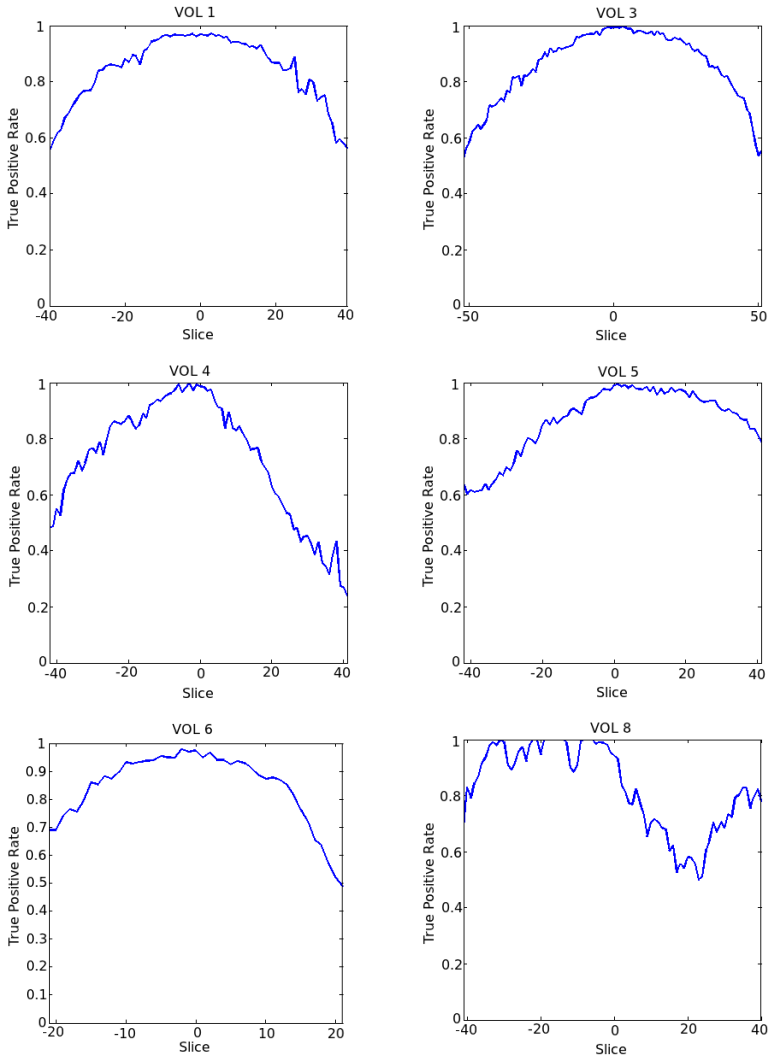


Figure 4.10: True positive rate of the thrombus detection on the 6 CTA volumes when applying the BDC learnt on the central axial slice to the remaining axial slices. Slices are numbered relative to the central slice, positive below it, negative above it.

4.2.4.1 Experimental Results

Figure 4.7 shows snapshots of the evolution of the Active Learning on one of the volumes. The interval between each snapshot is five iterations. We show the classification uncertainty of each voxel in the images of the left column, white corresponding to the maximum value. The right column shows the actual segmentation obtained by the classifier trained at the corresponding iteration. It can be appreciated that the uncertainty map evolves towards the boundaries of the thrombus, which are the natural places of maximal uncertainty. The segmentation has some alterations until reaching the final result. The classifier learns, forgets and relearns the target class along the learning and sample generation process.

Figure 4.8 shows the segmentation results in the central axial slice of each of the volumes considered. The left column shows the original CTA slice. This visualization helps to highlight the difficulties of the segmentation process: AAAs are of different sizes, with different placements and surrounded by different spatial layouts of structures due to anatomical differences between subjects. Moreover, in some cases the actual lumen situated in the middle of the thrombus is hiperintense, but not always. The middle column shows the ground truth segmentation provide by the expert human delineation. The right column shows the segmentation obtained after the Active Learning process in this slice.

Notice that if the BDC classifier obtained on one slice can be applied to the remaining slices without loss of accuracy, the human operator would only need to perform once the Active Learning process to obtain the whole volume segmentation. Fig. 4.9 shows the plots of the accuracy obtained at each CTA volume slice applying the BDC classifier trained on the thrombus central slice for each of the 6 CTA volumes treated in the experiment. The abscissa values correspond to the actual slice numbers in the volume. Obviously, slices where there is no thrombus detected by the ground truth are not included. There is some variability of the plots' span, due to the different sizes of the thrombus in each patient. As can be expected, the drop in classification accuracy is mostly symmetric, but not completely so. The generalization results are very good: the worst accuracy is above 0.98 in almost all cases.

Figure 4.10 plots the true positive ratios obtained at each CTA volume slice applying the BDC classifier trained on the thrombus central slice for each of the 6 CTA volumes treated in the experiment. In those plots, the abscissa's zero value corresponds to the central axial slice, the negative abscissa values correspond to slices above the central slice, the positive values correspond to slices below the central slice. The decrease of the values of true positive ratios are symmetric and show that the system is able to maintain a high sensitivity near the train slice, but that this sensitivity decreases greatly at the extremes of the thrombus. It is feasible

Table 4.3: Comparative average accuracy results published in the literature. Classifiers trained with Active Leraning over one central slice and tested over the remaining data. Classifiers tested are: Random Forest (RF) [120] and Hybrid ELM Rotation Forest (HERF) [9]. Bold values are the maximum for the corresponding dataset.

	RF	HERF	BDC
Vol. 1	0.993	0.992	**0.994**
Vol. 2	0.980	0.981	**0.985**
Vol. 3	0.992	0.993	**0.995**
Vol. 4	0.995	**0.996**	**0.996**
Vol. 5	0.991	**0.995**	**0.995**
Vol. 6	**0.996**	0.993	0.994
Vol. 7	**0.994**	0.993	0.979
Vol. 8	0.990	0.898	**0.994**

to propose additional Active Learning processes in slices where the sensitivity has decreased too much. Detection of such slices can be done based on the growing accumulated uncertainty that can be computed along the classification/segmentation of the slices. Finally, we provide the comparison of average accuracy results published in the literature over the same datasets applying the same training and testing methodology in Table 4.3: training is performed on the central slice of the thrombus, test on the remaining slices, and the accuracy reported is the average over all test slices in the volume. It can be appreciated that BDC provides the best global result in most volumes.

4.3 Active Learning for Retinal Image Segmentation

Computer-assisted detection and segmentation of blood vessels in retinal images of pathological subjects is difficult problem due to the great variability of the images. In this thesis work we propose an interactive image segmentation system using active learning which will allow quick volume segmentation requiring minimal intervention of the human operator. The advantage of this approach is that it can cope with large variability in images with minimal effort. The collection of image features used for this approach are simple statistics and undirected morphological operators computed on the green component of the image. Image segmentation is produced by classification by a Random Forest (RF) classifier. An initial RF classifier is built from seed set of labeled points. The human operator is presented with the most uncertain unlabeled voxels to select some of them for inclusion in the training set, retraining the RF classifier. We apply this approach to a well know benchmarking dataset achieving results comparable to the state-of-the-art in the

literature.

The goal is to classify image pixels into at least two classes, the interest region corresponding to blood vessels and the background [187]. Image intensity is not a discriminant value, because often many unrelated regions have similar pixel intensity values. Therefore, a feature vector is computed for each pixel location using information extracted from its neighboring pixels. This information comes from the result of linear and/or non-linear filtering performed on the pixel neighborhood. In this thesis work the features are: the voxel intensity, the mean, variance, maximum and minimum of the voxel neighborhood, for different values of the neighborhood radius (1,2,4...2^n). The definition of these features increases the data dimensionality and the complexity of the classifiers built on them.

4.3.1 Experiments

Datasets. We have performed computational experiments over 40 images of the DRIVE[1] dataset to test the proposed Active Learning based image classification approach. We are using 20 image for training set and the remaining 20 ones for testing purposes.

Segmentation problem. We are looking for the segmentation of the blood vessels. Therefore, we deal with a two-class problem.

Validation. The performance measure results of the experiments is the classification accuracy, overall error, specificity and sensibility. We provide summary comparison of results from state-of-the-art approaches in the literature. Apart from that we are also including some testing images visual results for a visual validation.

Table 4.4 shows the comparative results of our approach with state-of-the-art algorithms. It can be appreciated that specificity and accuracy results are close to the best in the literature. Also, sensitivity is comparable to some of them. In order to have a fair comparison, the reader must take into account that there is no post-processing or cleaning step in our algorithm, that the number of features (24) is lower than many other approaches, and the computation of these features is very straightforward, much more than the ones reported in the literature. Regarding the number of pixel samples used in our approach is much less than the other approaches in the literature due to the active learning strategy.

[1] `http://www.isi.uu.nl/Research/Databases/DRIVE/`

Table 4.4: Comparison of summary results of our approach with approaches reported in the literature

Method	Sensitivity	Specificity	Accuracy
Human rater	0.7763	0.9723	0.9470
Proposed method	0.6499	0.9803	0.9501
Abramoff et al. [135]	0.7145	-	0.9416
Staal et at. [160]	-	-	0.9442
Soares et al. [158]	-	-	0.9466
Ricci and Perfetti [143]	-	-	0.9563
Lupascu et al. [117]	0.72	-	0.9597
Xu and Luo [185]	0.7760	-	0.9328
You et al. [189]	0.7410	0.9751	0.9434
Marin et al. [123]	0.7067	0.9801	0.9452
Ng et al. [134]	0.7000	0.9530	-
Kande et at. [93]	-	-	0.8911
Salem et al. [148]	0.8215	0.9750	-
Chaudhuri et al. [30]	-	-	0.8773
Hoover et al. [74]	0.6751	0.9567	0.9267
Xiaoji and Mojon [89]	-	-	0.9212
Al-Rawi et al. [2]	-	-	0.9535
Zhanget al. [193]	0.7120	0.9724	0.9382
Cinsdikici and Aydin [40]	-	-	0.9293
Zana and Klein [191]	0.6971	-	0.9377
Mendonca and Campilho [127]	0.7344	0.9764	0.9452
M.M Fraz et al. [57]	0.7152	0.9769	0.9430
Miri and Mahloojifar [130]	0.7352	0.9795	0.9458
Martinez-Perez et al.[124]	0.6389	-	0.9181
Martinez-Perez et al. [125]	0.7246	0.9655	0.9344
Perez et al. [126]	0.6600	0.9612	0.9220
Anzalone et al. [7]	-	-	0.9419
Vlachos and Dermatas [175]	0.747	0.955	0.929

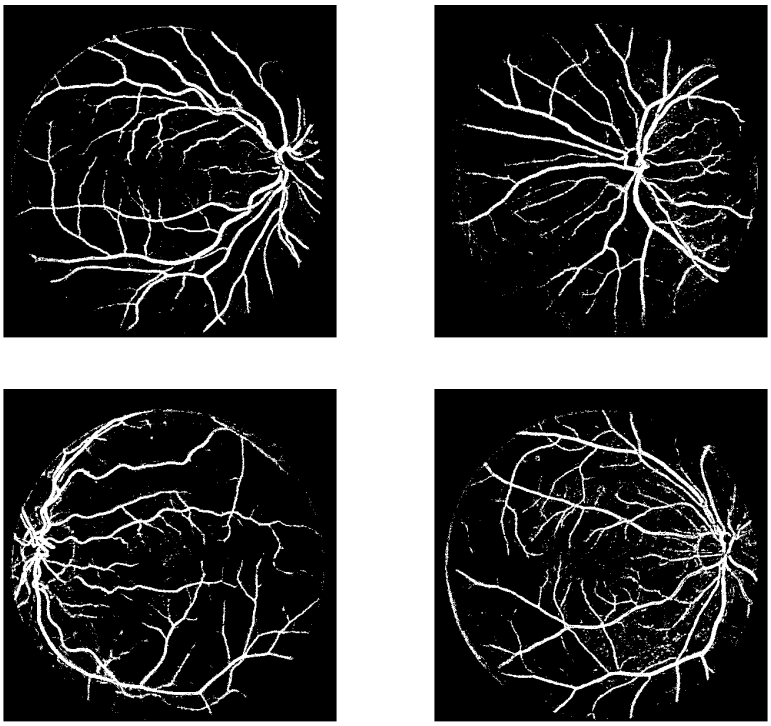


Figure 4.11: Segmentation result

Visual results

4.4 Meta-ensembles of Classifiers for Alzheimer's Disease Detection

Due to its growing social impact, prodromal detection of Alzheimer's disease is of paramount importance. Biomarkers based on Magnetic Resonance Imaging (MRI) are one of the most sought results in the neuroscience community. In this thesis work we evaluate several ensembles of classifiers trained and tested in a two level ensemble scheme as follows: the 116 regions of interest (ROI) of the Anatomical Auto- matic Labeling (AAL) brain atlas are used to compute disjoint feature sets from the Grey-matter probability maps from the segmentation of the T1 weighted MRI of each subject; ROI features are the summary statistics inside this ROI; one ensemble of classifiers is trained on each independent ROI feature data set; the final classification of each sub- ject is given by the combination of the classifications of each ROI, as meta-ensemble classifier. Experiments are performed on the 416 subjects (316 controls and 100 patients) of the OASIS database. We perform

a hold-out of the 20% of the data for model selection, computing a leave- one-out validation on the 80% remaining data. Results are computed without circularity. Tested classifiers are the Extreme Learning Machines (ELM), Bootstrapped Dendritic Computing (BDC), Hybrid Extreme Random Forest (HERF) and Random Forest (RF). We also report the most discriminant ROIs obtained in the model selection phase.

4.4.1 MRI Data Preprocessing

The spatial normalization of each subject of the database is performed with the FMRIB Software Library (FSL) FNIRT [157]. A four step registration process with increasing resolution and a scaled conjugate gradient minimization method has been performed using the default parameters, nearest neighbour interpolation and the standard MNI brain template. The Jacobian matrix at each voxel site describes the speed of change of the deformation in the neighboring area of each voxel. The determinant of the Jacobian matrix $\mathbf{J}_i$ (aka Jacobian) is commonly used scalar measure of the amount of distortion necessary to register the images. A value $\det(\mathbf{J}_i) > 1$ implies that the neighborhood adjacent to the displacement vector in voxel i was stretched to match the template (i.e., local volumetric expansion), while $\det(\mathbf{J}_i) < 1$ is associated with local shrinkage.

Apart, we segment the subjects with FSL FAST [194] into 3 volume brain tissue probability maps: grey (GM), white (WM) matter, and cerebral-spinal fluid (CSF). In this study we are interested in the GM maps, which we multiply by the Jacobians from the non-linear registration in order to get a modulated GM map in the standard MNI space. Subsequently, these maps, after smoothing with a 2mm Full-Width Half-Maximum (FWHM) Gaussian filter, are the basis for feature extraction. A visual check has been performed for all images in every processing step carried out in this experiment.

4.4.2 Feature extraction

The Automatic Anatomical Labeling (AAL) atlas [171] is used to partition the GM maps into 116 brain anatomical regions. In this study we compute for each AAL anatomical region from each subject GM map 7 statistical measures: the maximum voxel value, the minimum, the mean, the variance, the median, the kurtosis and the skewness. Resulting in 116 sub-datasets of 416 subjects with 7 features each.

4.4.3 Meta-ensemble classification

The overall classification is a meta-ensemble process. For each ROI we train a separate classifier, which can be a single or ensemble classifier. The meta-ensemble

classification performs a majority voting on the results of the ROI classifiers. To care for the effect of class unbalance we assume a majority threshold, that is the minimum number of ROIs that must agree in order to decide on the most frequent class. This majority threshold acts as a correction for the *a priori* probability (which is unknown to the model selection algorithm).

4.4.4 Model selection and validation

Algorithm 4.1 specifies the experimental setting including the model selection and validation phases. Model selection is performed on a 20% of the data holding out 80% for validation. Model selection aims to select the subset of ROIs giving best classification performance and the majority threshold value. Model selection proceeds as follows, first the classifier performance on each ROI features is estimated by a leave-one-out (LOO) procedure. Second, the ROIs are ordered according to their independent performances. Third, a greedy search for the optimal subset of ROIs is performed, adding them according to the previous order and testing their LOO performance for varying majority threshold values. The maximum performance gives the optimal ROI subset and majority threshold value.

Classifier parameter settings Each ensemble (i.e. BDC, RF and HERF) is composed of 5 classifiers. The number of ELMs in HERF ensembles is either one third or two thirds, this decision is taken on the model selection phase. For ELMs, the number of hidden nodes is fixed to 14 nodes. In the case of BDC the depth given by the number of dendrites D is fixed to 31 and the box size to 0.8.

4.4.5 Results

We report accuracy $((TP+TN)/N)$, sensitivity $(TP/(TP+FN))$, specificity $(TN/(FP+TN))$, for each ensemble classifier. In table 4.5 we show the classification performance of the ensembles. Best results correspond to the BDC, but in general results are below other conventional approaches in the literature. The presented results do use very few ROIs for the classification, so that the number of features is very small compared with other methods in the literature. Figures 4.12 and 4.13 show the ROIs selected from the model selection with each of the classifiers. The RF has selected the largest number of ROI (49), while the BDC selected the smallest number (10), meaning that BDC achieves its results with only 70 features. The only ROI which has been selected for all the classifiers is the left Parahippocampal gyrus, which is known to show atrophies together with the entorhinal cortex in an early stage of visible anatomical degradation of the brain with Alzheimer's disease.

Algorithm 4.1 Experiment cross-validation procedure

Let be $X = \{x_1, \ldots, x_n\}$ input data $x_i \in \mathbb{R}^d$, and $Y = \{y_1, \ldots, y_n\}$ the input data class labels $y_i \in \{0, 1\}$.
N is the number of samples.
K is the number of classifiers in the ensemble.
R is the number of ROIs.

1. Perform dataset partition $X = X^m \cup X^v$, $Y = Y^m \cup Y^v$, where (X^m, Y^m) contains the 20% of the dataset.

model selection using (X^m, Y^m)

1. for $i = 1 : R$
   (a) Select from X^m the i-th ROI data, denoted X_i^m .
   (b) Compute by LOO, the ROI accuracy, sensitivity and specificity.(a_i, ss_i, sp_i) of the ensemble classifier on X_i^m
2. end for
3. sort the ROIs according to decreasing value of $ss_i + sp_i$
4. for $i = 1 : R$
   (a) Select from X^m the i-th ROI data, denoted X_i^m , adding it to the incremental model selection feature set $X_i^{inc} \leftarrow X_{i-1}^{inc} \cup X_i^m$, with $X_0^{inc} = \emptyset$.
   (b) for $\theta = 1 : i$
      i. Compute by LOO, the ROI accuracy, sensitivity and specificity.$\left(a_i^\theta, ss_i^\theta, sp_i^\theta\right)$ of the *meta-ensemble* classifier on X_i^{inc} applying majority threshold θ.
   (c) endfor
5. Find the set of ROIs (*selectedROIs*) and majority threshold θ^* giving the highest performance $ss_i^\theta + sp_i^\theta$

validation using (X^v, Y^v)

1. for i in *selectedROIs*
   (a) Select from X^v the i-th ROI data, denoted X_i^v , adding it to the final feature set $X^f \leftarrow \cup X_i^v$ used for testing.
2. endfor
3. Compute final accuracy, sensitivity and specificity by LOO on X^f using θ^* as the majority threshold.

	Accuracy	Specificity	Sensitivity
RF	79.0	73.1	97.5
HERF	79.3	73.9	96.3
ELM	70.8	43.7	92.5
BDC	**80.8**	**77.1**	**92.5**

Table 4.5: Results

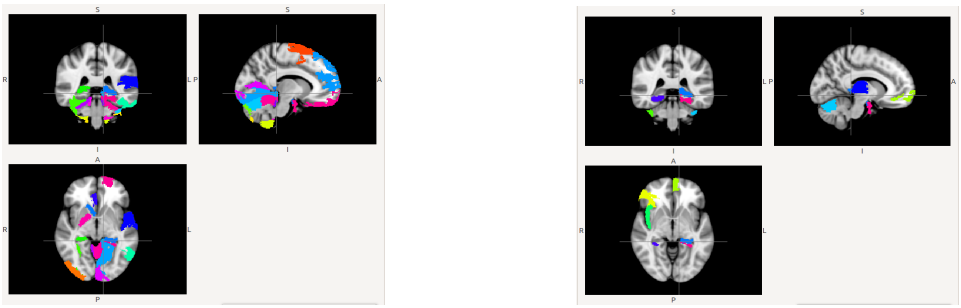


Figure 4.12: Slices of the MNI standard template where the ROIs selected by ELM (left) and BDC (right) are colored.

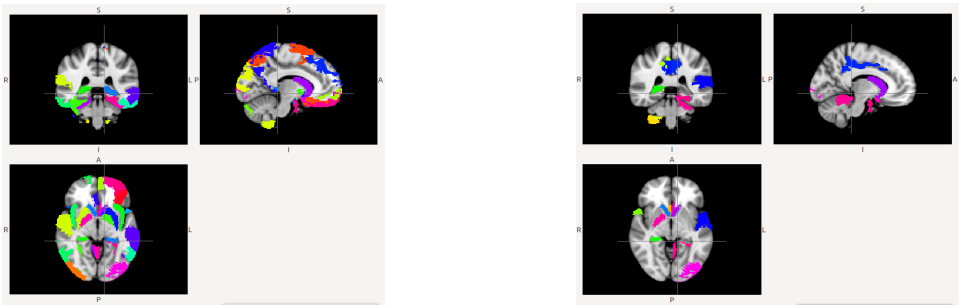


Figure 4.13: Slices of the MNI standard template where the ROIs selected by RF (left) and HERF (right) are colored.

Chapter 5

Conclusions

The works carried out in this Thesis have covered aspects of research in Computational Intelligence ranging from the proposition of new ensembles of classifiers up to their application in topics and domains of high impact. This chapter summarizes the contributions achieved in the diverse issues tackled and conclusions derived of the works.

5.1 Computational contributions in the Thesis

In this Thesis we have dealt with the proposal of new architectures of ensembles of classifiers based on randomized architectures, namely Extreme Learning Machines (ELM) and Random Forests (RF) have been the main building blocks. In this regard we have worked with ensembles of ELM regressors, providing also a formal proof of their convergence. Such proof is of critical importance to ensure reproducibility of results. Assured convergence means that results can be reproduced independently by diverse researchers, so that the proposal could be trusted as tool in the engineering of new systems. It is dependent on the uniform and exhaustive sampling of the feature space, in order to obtain unbiased zero mean error estimation of the target values. Therefore, combination of ELM individual outputs by averaging them removes the error as the number of components in the ensemble grows. We have discussed that ELM regression ensemble convergence requires the uniform exhaustive sampling of the feature space, in order to obtain unbiased zero mean error estimation of the target values. Then, combination of ELM individual outputs by averaging them removes the error as the number of components in the ensemble grows. Detailed analysis of convergence conditions for ELM regression ensembles is an open research field.

In a different line of work, we have proposed innovative inhomogeneous ensemble architectures, i.e. composed of diverse elemental architectures, including

data rotation to achieve greater diversity of the individual classifiers. The first new ensemble proposition is a called HERF including ELM, and RF as elementary classifier architectures. The second proposal, called AHERF, is an improvement of HERF that features anticipative selection of model architecture frequency inside the ensemble. Both architectures have been shown to be competitive against state of the art algorithms both in general benchmarking datasets and in specific problems, such as hyperspectral image classification. Moreover, we provide a formal justification for the AHERF model selection.

5.2 Hyperspectral image processing contributions

We have considered two aspects of hyperspectral image processing. One is the subpixel resolution analysis via non-linear unmixing, and the other is the thematic map building via pixel spectra classification. We give separate conclusions for each topic.

5.2.1 Conclusions on non-linear unmixing

We have proposed multivariate ELM regression ensembles to perform non-linear unmixing of hyperspectral images, learning the non-linear mapping between spectra and abundance vectors, as well as the inverse mapping returning spectra reconstructions from the abundance vectors. The inverse mapping serves the purpose of evaluation, by allowing to compute the reconstruction error, so that the approach is comparable to others. It is important to notice that ELMs are efficient to be trained even with a multidimensional output space, which is very infrequent in the literature. An specific advantage of this approach is that it does not compute explicitly the endmembers in the image. Experimental results on well known benchmark images, and comparison with stat-of-the-art approaches for endmember extraction and linear unmixing show the advantage of ELM regression ensembles both in terms of reconstruction error and visual segmentation of the images.

5.2.2 Conclusions on spectral classification

We have contributed a new spectral-spatial thematic map construction method with uses a spatial-spectral semi-supervised learning approach, and a spatial correction on the classification image. The method attacks several issues plaguing hyperspectral image processing. For instance, the semi-supervised learning component is very robust regarding labeled data scarcity, achieving state-of-the-art results from very little labeled data. For pixel spectral classification we propose ensembles of ELM, specially our own specific propositions HERF and AHEF. Results on several

well-known benchmark databases are encouraging, because the proposed approach competes well with methods reported in the literature with very straightforward parameter settings. We have provided an exhaustive sensitivity exploration of the effect of system parameters such as the number of hidden nodes and the ensemble size. The approach is quite insensitive to the ensemble size, but the number of hidden nodes of each individual ELM classifier has a definitive effect on performance, though some fluctuations are appreciated. A recommendation that can be ascertained from the results is that the number of hidden units must be at least of the order of the spectral dimension to ensure good results. Also, there is some effect of the underlying elementary classifier training algorithm (ELM, rELM, OP-ELM), but results seem to be greatly image dependent. Future work may be addressing the effect of the unsupervised classification algorithm. Post-processing spatial regularization algorithms may be based on anisotropic smoothing filters, which provide enhanced preservation of edges between regions.

5.3 Medical Image Analysis contributions

On the field of medical image analysis we have tackled three kinds of problems, two of them related to blood vessel extraction from images of different kinds. Both problems have been tackled with Active Learning approaches, in order to build a semi-automated image segmentation system, which may be helpful for the radiologist work. The third problem is related to brain imaging and the detection of neurodegenerative diseases. We present our detailed conclusions below.

5.3.1 Conclusions on Active Learning for CTA segmentation

The Computer Tomography Angiogram (CTA) produces large data volumes, whose segmentation is tedious. Automated systems require manual labeling of data for training. Moreover, CTA data may have large geometrical and intensity variations of the objects of interest between image captures. For these reason, we choose Active Learning as the most appropriate mean to provide the optimal training dataset extracted from each volume for the construction of the classification system by Machine Learning approaches providing the best segmentation. Therefore, the proposed systems are interactive semi-automated. We have tested several classifiers, i.e. Random Forest, HERF, the Bootstrapped Dendritic Classifiers (BDC). Training dataset are optimal in two senses: it is composed of the most informative pixels, and it has the smallest size required for the task.

To obtain an assessment of the approach, we compute the whole volume classification obtained applying the classifier (HERF, BDC) trained on one slice. We present experimental validation of the approach on a collection of CTA volumes

of AAA patients using the provided ground truth to simulate the human oracle. This approach means a hundred-fold reduction of the human intervention compared with building classifiers, for each slice. Classification performance results are very good. Besides, false positives can be easily filtered out by the application of area conditioned morphological operators. Results can be further improved by using specific Domain Knowledge of the structure being segmented. In this Thesis, we have transformed such rules in heuristic post-processing rules. The results show that in some cases, the application of such heuristics can provide dramatic increase in performance, maintaining all the advantages of the Active Learning approach. We propose the use in the clinical domain of this tool in order to enhance the working conditions of the radiologists.

5.3.2 Conclusions on retinal image segmentation

We proposed an Active Learning approach, such as discussed above, for training RF classifiers for the segmentation of the blood vessels in retinal images. Our approach is very efficient in time, requiring 10% of the training samples used by other approaches, while the image features are very simple to compute and much more efficient that those reported in the literature. Results have been provided without a cleaning or post-processing step, which is often used to increase sensitivity. The results on a standard benchmarking dataset are comparable with state of the art approaches.

5.3.3 Conclusions on MRI classification of AD patients

In this thesis work we report classification results of an ensemble of different types of classifiers on features extracted from modulated gray matter probability maps partitioned with the AAL atlas. The sample is the complete cross-sectional OASIS database of Alzheimer's Disease patients and controls. For each subject, the modulated gray matter data is partitioned into 116 regions and 7 statistical values from each are used as feature vectors. The results are in agreement with most of our previous classification experiments [38, 153, 152]. It was our aim to assess the performance of features built using *a priori* ROI maps against those experiments with supervised methods of feature selection. Unsupervised feature selection methods will most probably show worse classification performance than those which are supervised, but lead to more general systems which are less over-fitted to the given training database. We are aware that registration and segmentation errors can lead to biases in the accuracy of the classifiers. In addition, atrophy in brain structures can be interpreted as a late stage of Alzheimer's Disease and functional MRI could be used instead to detect previous stages of the disease.

Appendix A

Experimental Datasets

A.1 Hyperspectral experimental datasets

All these datasets are available at the research group wiki[1].

Salinas Complete and Subscene Salinas hyperspectral dataset was collected by the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor over the Valley of Salinas, Southern California, in 1998. AVIRIS acquires data in 224 bands from 0.4 to 2.5 μm, with nominal spectral resolution of 0.10 μm. The image contains 217 × 512 pixels. It was taken at low altitude, so that pixel spatial resolution is 3.7 m. Data classes include vegetables, bare soils, and vineyard fields. The Salinas A sub-scene comprises 83 × 86 pixels and is known as a difficult classification scenario with highly mixed pixels.

Pavia University Is a scene acquired by the ROSIS sensor during a flight campaign over Pavia, Northern Italy. It contains 610x610 pixels of 103 spectral bands. The pixel spatial resolution is 1.3 meters. Image ground truth differentiate 9 classes. Pavia scenes were provided by Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, Pavia university (Italy).

Indian Pines Indian Pines scene was collected by the AVIRIS sensor over Northwestern Indiana in June of 1992[103], it is the most widely used benchmark for testing the accuracy of hyperspectral data classification and segmentation algorithms. This scene of 145 × 145 pixels was acquired over a mixed agricultural/forest area early in the growing season, so that there is a strong mixture of soil and veg-

[1] `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes`

etal signatures. The ground truth available for the scene has 16 mutually exclusive classes.

Kennedy Space Center The scene was collected by the AVIRIS instrument over the Kennedy Space Center (KSC), Florida, on March 23, 1996. The KSC data was acquired from an altitude of approximately 20 km, so that its pixel spatial resolution is 18 m. After removing water absorption and low SNR bands, 176 bands remain for analysis. Training labeled data were selected using land cover maps derived from color infrared photography provided by the Kennedy Space Center and Landsat Thematic Mapper (TM) imagery. The vegetation classification scheme was developed by KSC personnel in an effort to define functional types that are discernible at the spatial resolution of Landsat and these AVIRIS data. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site.

Botswana The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001-2004. The Hyperion sensor on EO-1 acquires data at 30 m pixel resolution over a 7.7 km strip in 242 bands covering the 0.4 to 2.5 μm portion of the spectrum in 0.10 μm bands. Preprocessing of the data was performed by the UT Center for Space Research removing noisy bands, 145 bands were useful for analysis. The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta.

A.2 Medical Image Datasets

AAA We have performed computational experiments on 6 CTA datasets to test the proposed Active Learning based image classification approach. Each dataset consists of real human contrast-enhanced datasets of the abdominal area with 512x512 voxel resolution on each slice. Each dataset consists of 216 to 560 slices and 0.887x0.887x1 mm spatial resolution corresponding to patients with Abdominal Aortic Aneurysm. The dataset collection shows a wide diversity of sizes and locations of the thrombus. Some of them have metal streaking artifacts due to the stent graft placement. Ground truth segmentations of the thrombus for each dataset, that simulate a human oracle providing the labels for the voxels, were obtained manually by a clinical radiologist.

Retina The images for the DRIVE database were obtained from a diabetic retinopathy screening program in The Netherlands. The screening population consisted of 400 diabetic subjects between 25-90 years of age. Forty photographs have been randomly selected, 33 do not show any sign of diabetic retinopathy and 7 show signs of mild early diabetic retinopathy. Each image has been JPEG compressed.

The images were acquired using a Canon CR5 non-mydriatic 3CCD camera with a 45 degree field of view (FOV). Each image was captured using 8 bits per color plane at 768 by 584 pixels. The FOV of each image is circular with a diameter of approximately 540 pixels. For this database, the images have been cropped around the FOV. For each image, a mask image is provided that delineates the FOV.

The set of 40 images has been divided into training and test sets, both containing 20 images. For the training images, a single manual segmentation of the vasculature is available. For the test cases, two manual segmentations are available; one is used as gold standard, the other one can be used to compare computer generated segmentations with those of an independent human observer. All human observers that manually segmented the vasculature were instructed and trained by an experienced ophthalmologist. They were asked to mark all pixels for which they were for at least 70% certain that they were vessel.

AD We have worked on data of all the subjects from a public available brain MRI database, the first Open Access Series of Imaging Studies (OASIS) [122]. These subjects were selected from a larger database of individuals who had participated in MRI studies at Washington University, they were all right-handed and older adults had a recent clinical evaluation. Older subjects with and without dementia were obtained from the longitudinal pool of the Washington University Alzheimer Disease Research Center (ADRC). This release of OASIS consists of a cross-sectional collection of 416 male (119 controls and 41 patients) and female (197 controls and 59 patients) subjects aged 18 to 96 years (218 aged 18 to 59 years and 198 subjects aged 60 to 96 years). Further demographic and image acquisition details can be found in [122]. The data we are using are the skull-stripped and corrected for intensity inhomogeneity volumes.

Bibliography

[1] Yi Chen 0001, Nasser M. Nasrabadi, and Trac D. Tran. Hyperspectral image classification via kernel sparse representation. In Benoit Macq and Peter Schelkens, editors, *ICIP*, pages 1233–1236. IEEE, 2011.

[2] Mohammed Al-Rawi, Munib Qutaishat, and Mohammed Arrar. An improved matched filter for blood vessel detection of digital retinal images. *Comput. Biol. Med.*, 37(2):262–267, February 2007.

[3] Y. Altmann, N. Dobigeon, S. McLaughlin, and J.-Y. Tourneret. Residual component analysis of hyperspectral images - application to joint nonlinear unmixing and nonlinearity detection. *Image Processing, IEEE Transactions on*, 23(5):2148–2158, May 2014.

[4] Y. Altmann, N. Dobigeon, and J. Tourneret. Unsupervised post-nonlinear unmixing of hyperspectral images using a hamiltonian monte carlo algorithm. *Image Processing, IEEE Transactions on*, 23(6):2663–2675, June 2014.

[5] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

[6] Fatai Adesina Anifowose, Jane Labadin, and Abdulazeez Abdulraheem. Ensemble model of non-linear feature selection-based extreme learning machine for improved natural gas reservoir characterization. *Journal of Natural Gas Science and Engineering*, (0):–, 2015.

[7] Andrea Anzalone, Federico Bizzarri, Mauro Parodi, and Marco Storace. A modular supervised algorithm for vessel segmentation in red-free retinal images. *Comp. in Bio. and Med.*, 38(8):913–922, 2008.

[8] B Ayerdi, J Maiora, A d'Anjou, and M Graña. Applications of hybrid extreme rotation forests for image segmentation. *International Journal of Hybrid Intelligent Systems*, 11(1):13–24, 2014.

[9] B. Ayerdi, J. Maiora, and M. Grana. Active learning of hybrid extreme rotation forests for cta image segmentation. In *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*, pages 543–548, 2012.

[10] Borja Ayerdi and Manuel Graña. Hybrid extreme rotation forest. *Neural Networks*, 52(0):33 – 42, 2014.

[11] Borja Ayerdi, Ion Marqués, and Manuel Graña. Spatially regularized semisupervised ensembles of extreme learning machines for hyperspectral image segmentation. *Neurocomputing*, 149, Part A(0):373 – 386, 2015. Advances in neural networks Advances in Extreme Learning Machines Selected papers from the Tenth International Symposium on Neural Networks (ISNN 2013) Selected articles from the International Symposium on Extreme Learning Machines (ELM 2013).

[12] George Azzopardi and Nicolai Petkov. Automatic detection of vascular bifurcations in segmented retinal images using trainable cosfire filters. *Pattern Recognition Letters*, page in press, 2012.

[13] J.A. Benediktsson, J. Chanussot, and W.M. Moon. Very high-resolution remote sensing: Challenges and opportunities [point of view]. *Proceedings of the IEEE*, 100(6):1907–1910, June 2012.

[14] J.M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N.M. Nasrabadi, and J. Chanussot. Hyperspectral remote sensing data analysis and future challenges. *Geoscience and Remote Sensing Magazine, IEEE*, 1(2):6–36, June 2013.

[15] J.M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N.M. Nasrabadi, and J. Chanussot. Hyperspectral remote sensing data analysis and future challenges. *Geoscience and Remote Sensing Magazine, IEEE*, 1(2):6–36, June 2013.

[16] J.M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Qian Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(2):354–379, April 2012.

[17] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[18] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[20] L. Bruzzone, Mingmin Chi, and M. Marconcini. A novel transductive SVM for semisupervised classification of Remote-Sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363 –3373, November 2006.

[21] D. Burazerovic, R. Heylen, B. Geens, S. Sterckx, and P. Scheunders. Detecting the adjacency effect in hyperspectral imagery with spectral unmixing techniques. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(3):1070–1078, June 2013.

[22] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. Atli Benediktsson. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *Signal Processing Magazine, IEEE*, 31(1):45–54, Jan 2014.

[23] Jingjing Cao, Sam Kwong, Ran Wang, Xiaodong Li, Ke Li, and Xiangfei Kong. Class-specific soft voting based multiple extreme learning machines ensemble. *Neurocomputing*, 149, Part A(0):275 – 284, 2015. Advances in neural networks Advances in Extreme Learning Machines Selected papers from the Tenth International Symposium on Neural Networks (ISNN 2013) Selected articles from the International Symposium on Extreme Learning Machines (ELM 2013).

[24] Jiuwen Cao, Zhiping Lin, Guang-Bin Huang, and Nan Liu. Voting based extreme learning machine. *Information Sciences*, 185(1):66 – 77, 2012.

[25] Xavier Ceamanos, Björn Waske, Jón Atli Benediktsson, Jocelyn Chanussot, Mathieu Fauvel, and Johannes R. Sveinsson. A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data. *Int. J. Image Graphics*, 1(4):293–307, December 2010.

[26] Jonathan Cheung-Wai Chan, Pieter Beckers, Toon Spanhove, and Jeroen Vanden 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*, 18(0):13–22, August 2012.

[27] Jonathan Cheung-Wai Chan and Desiré Paelinckx. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6):2999–3011, June 2008.

[28] Chein-I Chang and A. Plaza. A fast iterative algorithm for implementation of pixel purity index. *Geoscience and Remote Sensing Letters, IEEE*, 3(1):63–67, Jan 2006.

[29] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[30] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum. Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE transactions on medical imaging*, 8(3):263–269, 1989.

[31] Jie Chen, C. Richard, and P. Honeine. Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model. *Signal Processing, IEEE Transactions on*, 61(2):480–492, Jan 2013.

[32] Jie Chen, C. Richard, and P. Honeine. Nonlinear estimation of material abundances in hyperspectral images with ell_1-norm spatial regularization. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(5):2654–2665, May 2014.

[33] Lei Chen, Guang-Bin Huang, and Hung Keng Pung. Systemical convergence rate analysis of convex incremental feedforward neural networks. *Neurocomputing*, 72(10–12):2627 – 2635, 2009.

[34] T. Chen, A. Rangarajan, et al. CAVIAR: Classification via aggregated regression and its application in classifying OASIS brain database. *Proceedings / IEEE International Symposium on Biomedical Imaging: from nano to macro.*, 2010:1337–1340, April 2010.

[35] J. Chi and M.M. Crawford. Active landmark sampling for manifold learning based spectral unmixing. *Geoscience and Remote Sensing Letters, IEEE*, 11(11):1881–1885, Nov 2014.

[36] Junhwa Chi and M.M. Crawford. Selection of landmark points on nonlinear manifolds for spectral unmixing using local homogeneity. *Geoscience and Remote Sensing Letters, IEEE*, 10(4):711–715, July 2013.

[37] Mingmin Chi, Qun Qian, and Jon Atli Benediktsson. Cluster-based ensemble classification for hyperspectral remote sensing images. In *IGARSS (1)*, pages 209–212. IEEE, 2008.

[38] D. Chyzhyk, M. Graña, et al. Hybrid dendritic computing with kernel-LICA applied to Alzheimer's disease detection in MRI. *Neurocomputing*, 75(1):72–77, January 2012.

[39] Darya Chyzhyk, Borja Ayerdi, and Josu Maiora. Active learning with bootstrapped dendritic classifier applied to medical image segmentation. *Pattern Recognition Letters*, submitted, 2013.

[40] Muhammed Gokhan Cinsdikici and Dogan Aydin. Detection of blood vessels in ophthalmoscope images using mf/ant (matched filter/ant colony) algorithm. *Computer Methods and Programs in Biomedicine*, 96(2):85–95, 2009.

[41] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[42] R.D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982.

[43] Jiantao Cui, Xiaorun Li, and Liaoying Zhao. Nonlinear spectral mixture analysis by determining per-pixel endmember sets. *Geoscience and Remote Sensing Letters, IEEE*, 11(8):1404–1408, Aug 2014.

[44] Michele Dalponte, Lorenzo Bruzzone, Loris Vescovo, and Damiano Gianelle. The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. *Remote Sensing of Environment*, 113(11):2345–2355, November 2009.

[45] C. Davatzikos, Y. Fan, et al. Detection of prodromal alzheimer's disease via pattern classification of MRI. *Neurobiology of aging*, 29(4):514–523, April 2008. PMID: 17174012 PMCID: 2323584.

[46] M.A. David. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

[47] Marleen de Bruijne, Bram van Ginneken, Max A Viergever, and Wiro J Niessen. Interactive segmentation of abdominal aortic aneurysms in cta images. *Med Image Anal*, 8(2):127–138, Jun 2004.

[48] Stefanie Demirci, Guy Lejeune, and Nassir Navab. Hybrid deformable model for aneurysm segmentation. In *ISBI'09*, pages 33–36, 2009.

[49] I. Dopido, A. Villa, A. Plaza, and P. Gamba. A quantitative and comparative assessment of unmixing-based feature extraction techniques for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):421 –435, April 2012.

[50] I. Dopido, M. Zortea, A. Villa, A. Plaza, and P. Gamba. Unmixing prior to supervised classification of remotely sensed hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 8(4):760 –764, July 2011.

[51] M.M. Dundar and D.A. Landgrebe. A cost-effective semisupervised classifier approach with kernels. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(1):264–270, Jan 2004.

[52] O. Eches and M. Guillaume. A bilinear-bilinear nonnegative matrix factorization method for hyperspectral unmixing. *Geoscience and Remote Sensing Letters, IEEE*, 11(4):778–782, April 2014.

[53] M. Fauvel, J.A. Benediktsson, J. Chanussot, and J.R. Sveinsson. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804 –3814, November 2008.

[54] M. Fauvel, Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C. Tilton. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, 101(3):652–675, March 2013.

[55] EW Forgy. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[56] John M. Fossaceca, Thomas A. Mazzuchi, and Shahram Sarkani. Mark-elm: Application of a novel multiple kernel learning framework for improving the robustness of network intrusion detection. *Expert Systems with Applications*, 42(8):4062 – 4080, 2015.

[57] M. M. Fraz, Sarah Barman, Paolo Remagnino, Andreas Hoppe, A. Basit, Bunyarit Uyyanonvara, Alicja R. Rudnicka, and Christopher G. Owen. An approach to localize the retinal blood vessels using bit planes and centerline detection. *Computer Methods and Programs in Biomedicine*, 108(2):600–616, 2012.

[58] Moti Freiman, Steven J. Esses, Leo Joskowicz, and Jacob Sosna. An Iterative Model-Constraint Graph-cut Algorithm for Abdominal Aortic Aneurysm Thrombus Segmentation. In *Proc. of the 2010 IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro (ISBI'10)*, pages 672–675, Rotterdam, The Netherlands, April 2010. IEEE.

[59] Lianru Gao, Jun Li, M. Khodadadzadeh, A. Plaza, Bing Zhang, Zhijian He, and Huiming Yan. Subspace-based support vector machines for hyperspec-

tral image classification. *Geoscience and Remote Sensing Letters, IEEE*, 12(2):349–353, Feb 2015.

[60] Nicolás García-Pedrajas and César García-Osorio. Constructing ensembles of classifiers using supervised projection methods based on misclassified instances. *Expert Systems with Applications*, 38(1):343 – 359, 2011.

[61] P. Ghamisi and J.A. Benediktsson. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *Geoscience and Remote Sensing Letters, IEEE*, 12(2):309–313, Feb 2015.

[62] P. Ghamisi, M.S. Couceiro, and J.A. Benediktsson. A novel feature selection approach based on fodpso and svm. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(5):2935–2947, May 2015.

[63] P. Ghamisi, M. Dalla Mura, and J.A. Benediktsson. A survey on spectral-spatial classification techniques based on attribute profiles. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(5):2335–2353, May 2015.

[64] B. Gokaraju, S.S. Durbha, R.L. King, and N.H. Younan. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the gulf of mexico. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(3):710 –720, September 2011.

[65] Yanfeng Gu, Shizhe Wang, and Xiuping Jia. Spectral unmixing in multiple-kernel hilbert space for hyperspectral imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(7):3968–3981, July 2013.

[66] Olle Hagner and Heather Reese. A method for calibrated maximum likelihood classification of forest types. *Remote Sensing of Environment*, 110(4):438–444, October 2007.

[67] Jisoo Ham, Yangchi Chen, Melba M. Crawford, and Joydeep Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE T. Geoscience and Remote Sensing*, 43(3):492–501, 2005.

[68] Min Han and Ben Liu. Ensemble of extreme learning machine for remote sensing image classification. *Neurocomputing*, 149, Part A(0):65 – 70, 2015. Advances in neural networks Advances in Extreme Learning Machines Selected papers from the Tenth International Symposium on Neural Networks (ISNN 2013) Selected articles from the International Symposium on Extreme Learning Machines (ELM 2013).

[69] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, October 1990.

[70] M. Heinl, J. Walde, G. Tappeiner, and U. Tappeiner. Classifiers vs. input variables - the drivers in image classification for land cover mapping. *International Journal of Applied Earth Observation and Geoinformation*, 11(6):423–430, December 2009.

[71] R. Heylen and P. Scheunders. Hyperspectral intrinsic dimensionality estimation with nearest-neighbor distance ratios. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(2):570–579, April 2013.

[72] A.T. Hirsch, Z.J. Haskal, N.R. Hertzer, C.W. Bakal, M.A. Creager, J.L. Halperin, L.F. Hiratzka, W.R.C. Murphy, J.W. Olin, J.B. Puschett, et al. Acc/aha 2005 practice guidelines for the management of patients with peripheral arterial disease. *Circulation*, 113(11):e463, 2006.

[73] T.K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[74] Adam Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19:203–210, 2000.

[75] P. Horata, S. Chiewchanwattana, and K. Sunat. Robust extreme learning machine. *Neurocomputing*, on line, 2012.

[76] G.-B. Huang, L. Chen, and C.-K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17(4):879–892, 2006.

[77] Guang-Bin Huang and Lei Chen. Convex incremental extreme learning machine. *Neurocomputing*, 70(16–18):3056 – 3062, 2007.

[78] Guang-Bin Huang and Lei Chen. Enhanced random search based incremental extreme learning machine. *Neurocomputing*, 71(16–18):3460 – 3468, 2008.

[79] Guang-Bin Huang, Ming-Bin Li, Lei Chen, and Chee-Kheong Siew. Incremental extreme learning machine with fully complex hidden nodes. *Neurocomputing*, 71(4–6):576 – 583, 2008.

[80] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3):489 – 501, 2006.

[81] X. Huang, X. Guan, J.A. Benediktsson, L. Zhang, J. Li, A. Plaza, and M. Dalla Mura. Multiple morphological profiles from multicomponent-base images for hyperspectral image classification. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(12):4653–4669, Dec 2014.

[82] L. Hubert-Moy, A. Cotonnec, L. Le Du, A. Chardin, and P. Perez. A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing of Environment*, 75(2):174–187, February 2001.

[83] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory*, 14(1):55–63, 1968.

[84] Lian-Zhi Huo, Ping Tang, Zheng Zhang, and D. Tuia. Semisupervised classification of remote sensing images with hierarchical spatial similarity. *Geoscience and Remote Sensing Letters, IEEE*, 12(1):150–154, Jan 2015.

[85] Jungho Im and John R. Jensen. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sensing of Environment*, 99(3):326–340, November 2005.

[86] Qiong Jackson and D.A. Landgrebe. An adaptive classifier design for high-dimensional data analysis with a limited training data set. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(12):2664–2679, Dec 2001.

[87] Qiong Jackson and D.A. Landgrebe. Adaptive bayesian contextual classification based on markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on*, 40(11):2454–2463, Nov 2002.

[88] Sen Jia, Zhen Ji, Yuntao Qian, and Linlin Shen. Unsupervised band selection for hyperspectral imagery classification without manual band removal. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):531 –543, April 2012.

[89] Xiaoyi Jiang and Daniel Mojon. Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(1):131–137, January 2003.

[90] L.O. Jimenez and D.A. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 28(1):39–54, Feb 1998.

[91] Hui juan Lu, Chun lin An, En hui Zheng, and Yi Lu. Dissimilarity based ensemble of extreme learning machine for gene expression data classification. *Neurocomputing*, online(0):–, 2013.

[92] Henry Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, September 1958.

[93] Giri Babu Kande, P. Venkata Subbaiah, and Satya Savithri Tirumala. Unsupervised fuzzy based vessel segmentation in pathological digital fundus images. *J. Medical Systems*, 34(5):849–858, 2010.

[94] Xudong Kang, Shutao Li, Leyuan Fang, and J.A. Benediktsson. Intrinsic image decomposition for feature extraction of hyperspectral images. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(4):2241–2253, April 2015.

[95] CD Karkos, U. Mukhopadhyay, I. Papakostas, J. Ghosh, GJL Thomson, and R. Hughes. Abdominal aortic aneurysm: the role of clinical examination and opportunistic detection. *European Journal of Vascular and Endovascular Surgery*, 19(3):299–303, 2000.

[96] T. Kavzoglu and I. Colkesen. A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11(5):352–359, October 2009.

[97] N. Keshava and J.F. Mustard. Spectral unmixing. *Signal Processing Magazine, IEEE*, 19(1):44–57, Jan 2002.

[98] Jens Keuchel, Simone Naumann, Matthias Heiler, and Alexander Siegmund. Automatic land cover analysis for tenerife by supervised classification using remotely sensed data. *Remote Sensing of Environment*, 86(4):530–541, August 2003.

[99] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung-Yang Bang. Constructing support vector machine ensemble. *Pattern Recognition*, 36:2757–2767, 2003.

[100] N. Kosaka, Kuniaki Uto, and Yukio Kosugi. Ica-aided mixed-pixel analysis of hyperspectral data in agricultural land. *Geoscience and Remote Sensing Letters, IEEE*, 2(2):220–224, April 2005.

[101] Shailesh Kumar, Joydeep Ghosh, and Melba M. Crawford. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. *Pattern Analysis and Applications*, 5(2):210–220, 2002.

[102] Yuan Lan, Yeng Chai Soh, and Guang-Bin Huang. Ensemble of online sequential extreme learning machine. *Neurocomputing*, 72(13–15):3391 – 3395, 2009. <ce:title>Hybrid Learning Machines (HAIS 2007) / Recent Developments in Natural Computation (ICNC 2007)</ce:title>.

[103] D. A. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, Hoboken, NJ, 2003.

[104] D.A. Landgrebe. Analysis technology for land remote sensing. *Proceedings of the IEEE*, 69(5):628–642, May 1981.

[105] D.A. Landgrebe. Multispectral land sensing: where from, where to? *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):414–421, March 2005.

[106] F.A. Lederle, S.E. Wilson, G.R. Johnson, D.B. Reinke, F.N. Littooy, C.W. Acher, D.J. Ballard, L.M. Messina, I.L. Gordon, E.P. Chute, et al. Immediate repair compared with surveillance of small abdominal aortic aneurysms. *New England Journal of Medicine*, 346(19):1437–1444, 2002.

[107] David Lesage, Elsa D. Angelini, Isabelle Bloch, and Gareth Funka-Lea. A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical Image Analysis*, 13(6):819 – 845, 2009.

[108] J. Li, J. M. Bioucas-Dias, and A. Plaza. Spectral -Spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1 –13, 2012.

[109] Jun Li, J.M. Bioucas-Dias, and A. Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4085 –4098, November 2010.

[110] Jun Li, J.M. Bioucas-Dias, and A. Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4085 –4098, November 2010.

[111] Jun Li, José M. Bioucas-Dias, and Antonio Plaza. Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE T. Geoscience and Remote Sensing*, 51(2):844–856, 2013.

[112] Jun Li, Xin Huang, P. Gamba, J.M. Bioucas-Dias, Liangpei Zhang, J. Atli Benediktsson, and A. Plaza. Multiple feature learning for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(3):1592–1606, March 2015.

[113] SZ Li. *Markov Random Field Modeling in Image Analysis*. Springer Verlag, 2009.

[114] Sin-Jin Lin, Chingho Chang, and Ming-Fu Hsu. Multiple extreme learning machines for a two-class imbalance corporate life cycle prediction. *Knowledge-Based Systems*, 39(0):214 – 223, 2013.

[115] M. Liu, D. Zhang, et al. Ensemble sparse classification of alzheimer's disease. *NeuroImage*, 60(2):1106–1116, April 2012. PMID: 22270352.

[116] Bin Luo and Jocelyn Chanussot. Supervised hyperspectral image classification based on spectral unmixing and geometrical features. *Signal Processing Systems*, 65(3):457–468, 2011.

[117] Carmen Alina Lupascu, Domenico Tegolo, and Emanuele Trucco. Fabc: retinal vessel segmentation using adaboost. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1267–1274, 2010.

[118] I. Macia, M. Grana, J. Maiora, C. Paloc, and M. de Blas. Detection of type ii endoleaks in abdominal aortic aneurysms after endovascular repair. *Computers in Biology and Medicine*, 41(10):871–880, 2011.

[119] I Macia, M. Graña, and C. Paloc. Knowledge management in image-based analysis of blood vessel structures. *Knowledge and Information Systems*, 30(2):457–491, 2012.

[120] J. Maiora, B. Ayerdi, and M. Graña. Random forest active learning for computed tomography angiography image segmentation. *Neurocomputing*, inpress, 2013.

[121] Josu Maiora and Manuel Graña. Abdominal cta image analisys through active learning and decision random forests: Aplication to aaa segmentatio. In *IJCNN*, 2012.

[122] D. S. Marcus, T. H. Wang, et al. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, September 2007. PMID: 17714011.

[123] Diego Marin, Arturo Aquino, Manuel Emilio Gegundez-Arias, and Jose Manuel Bravo. A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Trans. Med. Imaging*, 30(1):146–158, 2011.

[124] M. Elena Martinez-Perez, Alun D. Hughes, Alice V. Stanton, Simon A. Thom, Anil A. Bharath, and Kim H. Parker. Retinal blood vessel segmentation by means of scale-space analysis and region growing. In Chris J. Taylor 0002 and Alan C. F. Colchester, editors, *MICCAI*, volume 1679 of *Lecture Notes in Computer Science*, pages 90–97. Springer, 1999.

[125] M. Elena Martinez-Perez, Alun D. Hughes, Simon A. Thom, Anil A. Bharath, and Kim H. Parker. Segmentation of blood vessels from red-free and fluorescein retinal images. *Medical Image Analysis*, 11(1):47–61, 2007.

[126] ME Martinez-Perez, AD Hughes, SA Thom, and KH Parker. Improvement of a retinal blood vessel segmentation method using the insight segmentation and registration toolkit (itk). pages 892–895. IEEE Computer Society, 2007.

[127] Ana Maria Mendonca and Aurelio C. Campilho. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Trans. Med. Imaging*, 25(9):1200–1213, 2006.

[128] Lidan Miao and Hairong Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(3):765–777, March 2007.

[129] Yoan Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse. Op-elm: Optimally pruned extreme learning machine. *Neural Networks, IEEE Transactions on*, 21(1):158 –162, jan. 2010.

[130] Mohammad Saleh Miri and Ali Mahlooji Far. Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction. *IEEE Trans. Biomed. Engineering*, 58(5):1183–1192, 2011.

[131] G. Moser and S. B. Serpico. Combining support vector machines and markov random fields in an integrated framework for contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1 –19, 2012.

[132] Ahmad Mozaffari and Nasser L. Azad. Optimally pruned extreme learning machine with ensemble of regularization techniques and negative correla-

tion penalty applied to automotive engine coldstart hydrocarbon emission identification. *Neurocomputing*, (0):–, 2013.

[133] Mauro Dalla Mura, Alberto Villa, Jon Atli Benediktsson, Jocelyn Chanussot, and Lorenzo Bruzzone. Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. *IEEE Geosci. Remote Sensing Lett.*, 8(3):542–546, 2011.

[134] J. Ng, S. T. Clay, S. A. Barman, A. R. Fielder, M. J. Moseley, K. H. Parker, and C. Paterson. Maximum likelihood estimation of vessel parameters from scale space analysis. *Image Vision Comput.*, 28(1):55–63, January 2010.

[135] M. Niemeijer, J.J. Staal, B. van Ginneken, M. Loog, and M.D. Abramoff. Comparative study of retinal vessel segmentation methods on a new publicly available database. In J. Michael Fitzpatrick and M. Sonka, editors, *SPIE Medical Imaging*, volume 5370, pages 648–656. SPIE, SPIE, 2004.

[136] SD Olabarriaga, JM Rouet, M Fradkin, M Breeuwer, and WJ Niessen. Segmentation of thrombus in abdominal aortic aneurysms from CTA with nonparametric statistical grey level appearance modeling. *IEEE Transactions On Medical Imaging*, 24(4):477–485, APR 2005.

[137] Nikunj C. Oza and Kagan Tumer. Classifier ensembles: Select real-world applications. *Information Fusion*, 9(1):4 – 20, 2008. Special Issue on Applications of Ensemble Methods.

[138] C. Persello and L. Bruzzone. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4468 – 4483, 2012.

[139] K Perumal and Dr. R Bhaskaran. An efficient k-nearest neighbors based approach for classifying land cover regions in hyperspectral data via nonlinear dimensionality reduction. *International Journal on Signal and Image Processing*, 1(2):8, July 2010.

[140] Antonio Plaza, Jon Atli Benediktsson, Joseph W. Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, Mattia Marconcini, James C. Tilton, and Giovanna Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113, Supplement 1(0):S110–S122, September 2009.

[141] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[142] F. Ratle, G. Camps-Valls, and J. Weston. Semisupervised neural networks for efficient hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(5):2271–2282, May 2010.

[143] Elisa Ricci and Renzo Perfetti. Retinal blood vessel segmentation using line operators and support vector classification. *IEEE Transaction on Medical Imaging*, 26(10):1357–1365, 2007.

[144] J. A. Richards. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2nd edition, 1993.

[145] G.X. Ritter and G. Urcid. Lattice algebra approach to single-neuron computation. *Neural Networks, IEEE Transactions on*, 14(2):282 – 295, mar. 2003.

[146] J.J. Rodriguez, L.I. Kuncheva, and C.J. Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619 –1630, oct. 2006.

[147] V.F. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P.M. Atkinson, and C. Jeganathan. Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 121(0):93–107, June 2012.

[148] Sameh A. Salem, Nancy M. Salem, and Asoke K. Nandi. Segmentation of retinal blood vessels using a novel clustering algorithm (racal) with a partial supervision strategy. *Med. Biol. Engineering and Computing*, 45(3):261–273, 2007.

[149] A Samat, P Du, S Liu, J Li, and L Cheng. E2lms:ensemble extreme learning machines for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (99):1–1, 2014.

[150] Q. Sami ul Haq, Linmi Tao, Fuchun Sun, and Shiqiang Yang. A fast and robust sparse approach for hyperspectral data classification using a few labeled samples. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2287 –2302, June 2012.

[151] B.K. Sarkar, S.S. Sana, and K.S. Chaudhuri. A combined approach to tackle imbalanced data sets. *International Journal of Hybrid Intelligent Systems*, 9(4):185–202, 2013.

[152] A. Savio, M.T. García-Sebastián, et al. Neurocognitive disorder detection based on feature vectors extracted from VBM analysis of structural MRI. *Computers in Biology and Medicine*, 41(8):600–610, August 2011.

[153] A. Savio and M. Graña. Deformation based feature selection for computer aided diagnosis of alzheimer's disease. *Expert Systems with Applications*, 40(5):1619–1628, April 2013.

[154] Anne Schistad Solberg AsbjÃžrn Berg and Are F. C. Jensen. Robust classification of hyperspectral images, 2007.

[155] B. Settles. Active learning literature survey. *SciencesNew York*, 15(2), 2010.

[156] B.M. Shahshahani and D.A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *Geoscience and Remote Sensing, IEEE Transactions on*, 32(5):1087–1095, Sep 1994.

[157] S. M. Smith, M. Jenkinson, et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23 Suppl 1:S208–219, 2004. PMID: 15501092.

[158] JoÃ£o V. B. Soares, Jorge J. G. Le, Roberto M. Cesar, Herbert F. Jelinek, Michael J. Cree, and Senior Member. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Trans. on Medical Imaging*, 25:1214–1222, 2006.

[159] Umamahesh Srinivas, Yi Chen 0001, Vishal Monga, Nasser M. Nasrabadi, and Trac D. Tran. Exploiting sparsity in hyperspectral image classification via graphical models. *IEEE Geosci. Remote Sensing Lett.*, 10(3):505–509, 2013.

[160] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *Medical Imaging, IEEE Transactions on*, 23(4):501–509, April 2004.

[161] Yongjiao Sun, Ye Yuan, and Guoren Wang. An os-elm based distributed ensemble classification framework in {P2P} networks. *Neurocomputing*, 74(16):2438 – 2443, 2011. Advances in Extreme Learning Machine: Theory and Applications Biological Inspired Systems. Computational and Ambient Intelligence Selected papers of the 10th International Work-Conference on Artificial Neural Networks (IWANN2009).

[162] Zhan-Li Sun, Tsan-Ming Choi, Kin-Fan Au, and Yong Yu. Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, 46(1):411–419, 2008.

[163] Y. Tarabalka, J.A. Benediktsson, and J. Chanussot. Spectral -Spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2973 –2987, August 2009.

[164] Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C. Tilton. Multiple spectral-spatial classification approach for hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(11):4122–4132, Nov 2010.

[165] Y. Tarabalka, J. Chanussot, and J.A. Benediktsson. Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, 43(7):2367 – 2379, 2010.

[166] Y. Tarabalka, M. Fauvel, J. Chanussot, and J.A. Benediktsson. Svm- and mrf-based method for accurate classification of hyperspectral images. *Geoscience and Remote Sensing Letters, IEEE*, 7(4):736–740, Oct 2010.

[167] Yuliya Tarabalka, Jocelyn Chanussot, Jon Atli Benediktsson, Jesus Angulo, and Mathieu Fauvel. Segmentation and classification of hyperspectral data using watershed. In *IGARSS (3)*, pages 652–655. IEEE, 2008.

[168] M. Termenon and M. Graña. A two stage sequential ensemble applied to the classification of alzheimer's disease based on mri features. *Neural Processing Letters*, 35(1):1–12, 2012.

[169] D. Tuia and G. Camps-Valls. Semisupervised remote sensing image classification with cluster kernels. *Geoscience and Remote Sensing Letters, IEEE*, 6(2):224–228, April 2009.

[170] D. Tuia, E. Pasolli, and W.J. Emery. Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment*, 115(9):2232–2242, 2011.

[171] N Tzourio-Mazoyer, B Landeau, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289, January 2002. PMID: 11771995.

[172] E. Varol, B. Gaonkar, et al. Feature ranking based nested support vector machine ensemble for medical image classification. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 146 –149, May 2012.

[173] S. Velasco-Forero and V. Manian. Improving hyperspectral image classification using spatial preprocessing. *IEEE Geoscience and Remote Sensing Letters*, 6(2):297–301, April 2009.

[174] A. Villa, J.A. Benediktsson, J. Chanussot, and C. Jutten. Hyperspectral image classification with independent component discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12):4865 –4876, December 2011.

[175] Marios Vlachos and Evangelos Dermatas. Multi-scale retinal vessel segmentation using line tracking. *Comp. Med. Imag. and Graph.*, 34(3):213–227, 2010.

[176] Dhaval Vyas, N.S.R. Krishnayya, K.R. Manjunath, S.S. Ray, and Sushma Panigrahy. Evaluation of classifiers for processing hyperion (EO-1) data of tropical vegetation. *International Journal of Applied Earth Observation and Geoinformation*, 13(2):228–235, April 2011.

[177] Dianhui Wang and Monther Alhamdoosh. Evolutionary extreme learning machine ensembles with size control. *Neurocomputing*, (0):–, 2012.

[178] Xin-Lu Wang, Bjorn Waske, and Jon Atli Benediktsson. Ensemble methods for spectral-spatial classification of urban hyperspectral data. In *IGARSS (4)*, pages 944–947. IEEE, 2009.

[179] Timothy G. Whiteside, Guy S. Boggs, and Stefan W. Maier. Comparing object-based and pixel-based classifications for mapping savannas. *International Journal of Applied Earth Observation and Geoinformation*, 13(6):884–893, December 2011.

[180] T.J. Wilt, Minnesota Evidence based Practice Center, United States. Agency for Healthcare Research, and Quality. *Comparison of endovascular and open surgical repairs for abdominal aortic aneurysm*. Agency for Healthcare Research Quality, 2006.

[181] Michael E. Winter. N-findr: an algorithm for fast autonomous spectral endmember determination in hyperspectral data, 1999.

[182] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Trans. on Evol. Comp.*, 1(1):67 – 82, 1997.

[183] Michal Wozniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16(0):3 – 17, 2014.

[184] Junshi Xia, J. Chanussot, Peijun Du, and Xiyan He. Spectral-spatial classification for hyperspectral data using rotation forests with local feature extraction and markov random fields. *Geoscience and Remote Sensing, IEEE Transactions on*, 53(5):2532–2546, May 2015.

[185] Lili Xu and Shuqian Luo. A novel method for blood vessel detection from retinal images. *BioMedical Engineering OnLine*, 9(1):14, 2010.

[186] Xiaowei Xue, Min Yao, Zhaohui Wu, and Jianhua Yang. Genetic ensemble of extreme learning machine. *Neurocomputing*, on line, 2013.

[187] M. Yaqub, M. Javaid, C. Cooper, and J. Noble. Improving the classification accuracy of the classic rf method by intelligent feature selection and weighted voting of trees with application to medical image segmentation. *Machine Learning in Medical Imaging*, pages 184–192, 2011.

[188] N. Yokoya, J. Chanussot, and A. Iwasaki. Nonlinear unmixing of hyperspectral data using semi-nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(2):1430–1437, Feb 2014.

[189] Xinge You, Qinmu Peng, Yuan Yuan, Yiu-ming Cheung, and Jiajia Lei. Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. *Pattern Recogn.*, 44(10-11):2314–2324, October 2011.

[190] Qi Yu, Mark van Heeswijk, Yoan Miche, Rui Nian, Bo He, Eric Séverin, and Amaury Lendasse. Ensemble delta test-extreme learning machine (dt-elm) for regression. *Neurocomputing*, 129(0):153 – 158, 2014.

[191] Frederic Zana and Jean-Claude Klein. Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation. *IEEE Transactions on Image Processing*, 10(7):1010–1019, 2001.

[192] A. Zare, P. Gader, and G. Casella. Sampling piecewise convex unmixing and endmember extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(3):1655–1665, March 2013.

[193] Bob Zhang, Lin Zhang, Lei Zhang, and Fakhri Karray. Retinal vessel extraction by matched filter with first-order derivative of gaussian. *Comput. Biol. Med.*, 40(4):438–445, April 2010.

[194] Y. Zhang, M. Brady, et al. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, January 2001. PMID: 11293691.

[195] Yuhang Zhang, H.L. Yang, S. Prasad, E. Pasolli, Jinha Jung, and M. Crawford. Ensemble multiple kernel active learning for classification of multisource remote sensing data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 8(2):845–858, Feb 2015.

[196] Jianwei Zhao, Zhihui Wang, and Dong Sun Park. Online sequential extreme learning machine with forgetting mechanism. *Neurocomputing*, 87(0):79 – 89, 2012.

[197] L.J. Zhao, D.C. Yuan, T.Y. Chai, and J. Tang. {KPCA} and {ELM} ensemble modeling of wastewater effluent quality indices. *Procedia Engineering*, 15(0):5558 – 5562, 2011. {CEIS} 2011.

[198] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[199] F. Zhuge, G. D. Rubin, S. H. Sun, and S. Napel. An abdominal aortic aneurysm segmentation method: Level set with region and statistical information. *Medical Physics*, 33(5):1440–1453, 2006.