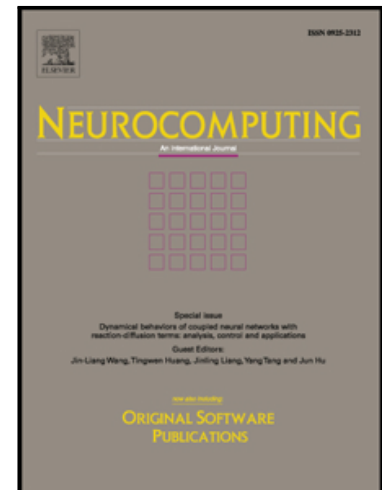


Accepted Manuscript

Ensemble of Extreme Learning Machines with Trained Classifier Combination and Statistical Features for Hyperspectral Data

Paweł Ksieniewicz, Bartosz Krawczyk, Michał Woźniak

PII: S0925-2312(17)31219-5
DOI: [10.1016/j.neucom.2016.04.076](https://doi.org/10.1016/j.neucom.2016.04.076)
Reference: NEUCOM 18678



To appear in: *Neurocomputing*

Received date: 27 December 2015
Accepted date: 25 April 2016

Please cite this article as: Paweł Ksieniewicz, Bartosz Krawczyk, Michał Woźniak, Ensemble of Extreme Learning Machines with Trained Classifier Combination and Statistical Features for Hyperspectral Data, *Neurocomputing* (2017), doi: [10.1016/j.neucom.2016.04.076](https://doi.org/10.1016/j.neucom.2016.04.076)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Ensemble of Extreme Learning Machines with Trained Classifier Combination and Statistical Features for Hyperspectral Data

Paweł Ksieniewicz^a, Bartosz Krawczyk^{a,*}, Michał Woźniak^a

^a*Department of Systems and Computer Networks, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland*

Abstract

Remote sensing and hyperspectral data analysis are areas offering wide range of valuable practical applications. However, they generate massive and complex data that is very difficult to be analyzed by a human being. Therefore, methods for efficient data representation and data mining are of high interest to these fields. In this paper we introduce a novel pipeline for feature extraction and classification of hyperspectral images. To obtain a compressed representation we propose to extract a set of statistical-based properties from these images. This allows for embedding feature space into fourteen channels, obtaining a significant dimensionality reduction. These features are used as an input for the ensemble learning based on randomized neural networks. We introduce a novel method for forming ensembles of extreme learning machines based on randomized feature subspaces and a trained combiner. It is based on continuous outputs and uses a perceptron-based learning scheme to calculate weights assigned to each classifier and class independently. Extensive experiments carried on a number of benchmarks images prove that using proposed feature extraction and extreme learning ensemble leads to a significant gain in classification accuracy.

Keywords: Ensemble learning, Extreme learning machines, Hyperspectral imaging, Computer vision, Feature extraction, Dimensionality reduction,

*Corresponding author

Email addresses: `pawel.ksieniewicz@pwr.edu.pl` (Paweł Ksieniewicz),
`bartosz.krawczyk@pwr.edu.pl` (Bartosz Krawczyk), `michal.wozniak@pwr.edu.pl`
(Michał Woźniak)

Image classification

1 **1. Introduction**

2 Because we are living in big data century, therefore the efficient analyt-
 3 ical tools which can analyze the huge volume of multidimensional data are
 4 still focus of intense research. One of the example of such a data is hyper-
 5 spectral imaging, which is widely used in agriculture, mineralogy etc. One
 6 can say that *if a picture is worth 1000 words, a hyperspectral image is worth*
 7 *almost 1000 pictures*¹. Hyperspectral cameras are able to capture hundreds
 8 of monochrome images correlated with a particular spectrum, nevertheless
 9 they still need to be analyzed manually, what is highly time consuming and
 10 requires very expensive manual labeling. Therefore methods which can use
 11 partially labelled data are desirable tools for hyperspectral image classifi-
 12 cation [6]. One of the very promising direction is active learning paradigm
 13 [35, 27], which employs an iterative data labeling and classifier training strat-
 14 egy with as small as possible set of training examples. A complementary
 15 approach proposes an efficient data representation of hyperspectral images
 16 which could be used by a classification system. As we deal with multi-class
 17 and high-dimensional problem, we require a highly effective pattern classifi-
 18 cation system to be able to analyze such data.

19 Classifier ensembles are nowadays recognized as the on of the most promis-
 20 ing direction in pattern classification [42]. This approach exploit the conclu-
 21 sions from so-called Wolpert's *no free lunch* theorem, that there is not a
 22 single classifier, which is the best one for all decision tasks, but each model
 23 has its own, specific domain of competence [41] where it may outperform
 24 other competing algorithms. Let's formulate the main presumptions of using
 25 such a classification model [13]

- 26 • Classifier ensembles behave well both in the case when a learning set
 27 is very small and when we have a huge amount of learning examples at
 28 our disposal. In the first case, classifier ensemble can exploit methods
 29 based on bootstrapping [30], while for the second case it allows to train
 30 individuals on partitions of dataset.

¹J.P.Ferguson, An Introduction to Hyperspectral Imaging, Photonics and Analytical Marketing Ltd.

- 31 • Classifier ensemble may outperform the best individual classifier [10]
32 and under some conditions (e.g., majority voting by a group of indi-
33 vidual classifiers committed error independently) this improvement has
34 been proven analytically [26].
- 35 • Many classifier training methods, as decision tree [34], are heuristic
36 search algorithms which usually suffer from local optima. Therefore,
37 the ensemble learning approach approach is equivalent to a multi-start
38 local random search which increases the probability of finding an opti-
39 mal model.
- 40 • Classifier ensemble may be easy implemented in efficient computing en-
41 vironments such as parallel and multithreaded computer architectures
42 [40].

43 In this work, we propose a novel ensemble dedicated to analysis of hyper-
44 spectral data. Its base classifiers are being built on the basis of decomposed
45 color channels. This assures their initial diversity, as every color channel car-
46 ries different information. We further augment this idea by using a trained
47 fuser, based on perceptron learning. This allows us to assign higher weights
48 to more competent classifiers. As not all of the channels carry equally useful
49 features, we boost the influence of the most relevant ones on the final decision
50 of the ensemble.

51 As the basis of our ensemble we propose to use Extreme Learning Ma-
52 chines (ELMs), a popular branch of randomized neural networks. Due to
53 their efficacy and low training complexity they have been reported to display
54 high usefulness for the hyperspectral data analysis task [28, 31]. However,
55 methods for constructing efficient ELMs ensembles still require development
56 [4].

57 The main contributions of the paper are as follows:

- 58 1. A novel proposition of the statistical-based feature extraction from hy-
59 perspectral images.
- 60 2. An efficient ELMs ensemble architecture based on trained combiner.
- 61 3. Application of the proposed features and ensemble structure together
62 with Random Subspaces method to the problem of hyperspectral image
63 classification.
- 64 4. Experimental evaluation of the proposed approach.

In section 2 we shortly introduce into hyperspectral image analysis, then in section 3 presents the proposition how to extract the valuable features from hyperspectral data. The section 4 describes the classification methods based on ensemble approach. The experimental evaluation is presented in section 5. At the end, in section 6 shows conclusions and possible usages of the proposed approach.

2. Hyperspectral Image Analysis

Natural perception of electromagnetic waves is limited to only four features of information spectrum. Each of them is a single chrome channel, which composed together by a human brain brings its owner an chemical illusion called color vision. The color can be interpreted as short vector, most often builded by three values. Its most popular representation is based on human perception on daylight, described by Svaetichin in 1956 [37] RGB model. Place of S M and L cone cells is taken there for channels of particular light impressions.

Hyperspectral image is a collection of high-resolution monochromatic pictures covering large spacial region for broad range of wavelengths. Structurally it is a three-dimensional matrix of reflectance. First two dimensions are standard lengths of a flat projection. The third is a spectral depth. Main idea of hyperspectral imaging is minimization of range covered by every band with maximization of band number. The current industrial standard, AVIRIS spectrometer, captures images with 224 channels in range $0.4 - 2.5\mu m$.

A slice taken from hyperspectral cube provides us information of reflectance of the area for a given spectral band. Taking a vector alongside the spectral band axis provides us spectral signature, which carries information about reflectance of one particular pixel for every covered spectral band. Example slice and signature are presented in Figure 1.

Signatures are used to detect type of material represented by pixel on an image. It is possible to distinguish type of ground, vegetation, used building material, rock strata or many other.

Method of separation of an hyperspectral image into channels is based on human perception of colorful images. Its main base is to replace a *reading* from *photoreceptors* with statistical measurement, doing e.g. elementary statistical operations on signature vector. Monochromatic image from this kind of metric can turn into channel used to construct colorful picture or,

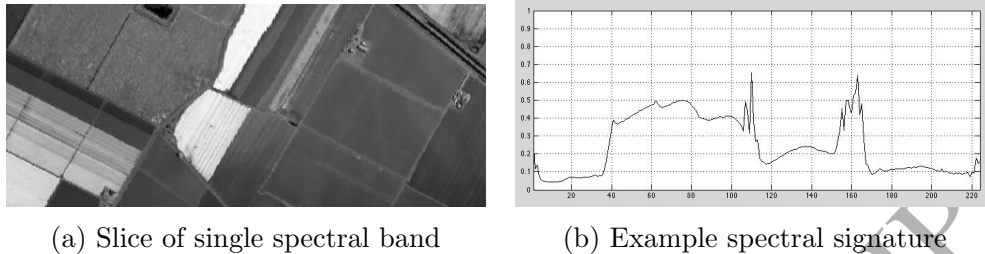


Figure 1: Hyperspectral image elements

100 after posterization, set of labels. It also implements a method of separation
 101 of homogenous areas on image, used also to filter noisy ranges of spectrum.

102 After the image color decomposition, we need to apply machine learn-
 103 ing algorithms in order to conduct segmentation or classification. Among a
 104 plethora of classification methods, ensembles has gained a significant interest
 105 of researchers over the last decade [21]. Combining multiple classifiers can
 106 lead to a significant improvement of the accuracy in comparison to single
 107 learner. There are many different methods for forming efficient ensembles
 108 [42], but they all share several fundamental ideas. In order for the ensemble
 109 to work, we need to have more than one classifiers at our disposal. They
 110 can be trained on the given dataset, or supplied by heterogeneous sources.
 111 A special attention should be paid to the properties of used classifiers. For
 112 an ensemble to work properly, it must consist of classifiers that at the same
 113 time display a high individual accuracy and are mutually complementary
 114 with each other. As, in most cases, not all of the available classifiers satisfy
 115 this condition, one needs to discard the irrelevant models. This step has
 116 a crucial impact on the quality of the formed committee and is known as
 117 classifier selection or ensemble pruning [11]. Another important part of en-
 118 semble design is the combination rule. It will fuse the individual outputs of
 119 base classifiers into a single committee decision. This task can be tackled in
 120 two different ways: with untrained or trained fuser. Untrained fusers (such
 121 as voting) [39] are simple and straightforward to use, but can be subject of
 122 performance limitations. Trained fusers adapt their behavior to the analyzed
 123 data, but require some time to establish their rules and a dedicated training
 124 set [25].

125 Most common method of generating false-color pictures from hyperspec-
 126 tral data is mapping three bands from a wide signature into RGB channels. For
 127 case of spectral depth reduction, the most popular standard is PCA(*Principal*

128 *Components Analysis*) [1]. Three, richest in information, principal compo-
 129 nents from hyperspectral cube are mapped to various color models channels
 130 (RGB, HSL, HSV) [38].

131 Some works suggest to balance S/N (*Signal-to-noise ratio*) to enhance con-
 132 trast of an image [15] and reduce noise impact.

133 3. Proposed Statistical Features for Hyperspectral Images

134 We propose a novel method for simultaneous feature extraction and low-
 135 dimensional embedding of hyperspectral images. It is based on the idea of
 136 creating a new representation, evolving from human color perception, pre-
 137 serving as much information as possible, with simple, time efficient compu-
 138 tations.

139 We are interpreting the matrix of cone cells reacting on same wave-
 140 lengths as a transformation, projecting three-dimensional input onto two-
 141 dimensional result. Hyperspectral imaging is there a discrete form of this
 142 three-dimensional input, which provides enough data to acquire other trans-
 143 formation functions. So created artificial cone cells matrixes will generate
 144 our statistical features.

145 New proposition is a significant extension of our previous proposal [24],
 146 introducing procedure of spatial blurring, normalization and histogram equal-
 147 ization. Also the extended set of statistical features is proposed.

148 The procedure run as follows. At the start, the class edges are recognized
 149 and calculated. Next, the collection of side information produced during
 150 first step, lets us to generate a filter for noisy bands of image. Next the
 151 features of filtered image are computed and, at the end, they are prepared
 152 for classification.

153 3.1. Noise detection

154 A value denivelation in finite neighborhood of every pixel can be used to
 155 detect borders between non-texture areas of picture[12]. A side effect of this
 156 method is the measurement of entropy (\bar{H}), calculated from amount of all
 157 values (ρ) divided by calculation of pixels per layer (ppl).

$$\bar{H} = \frac{\sum \rho}{ppl} \quad (1)$$

158 While every hyperspectral cube contains wavelengths with high noise ra-
 159 tio, adequate threshold to drain most of them would be a mean value of

160 entropy. To separate hills of entropy changes we are using information about
 161 its dynamics. An vector of dynamics was made in a way analogous to edge
 162 detection, by calculating discrepancy between actual (\bar{H}) and next value (\bar{H}')
 163 on the vector of entropy.

$$\bar{D}H = |\bar{H} - \bar{H}'| \quad (2)$$

164 Mean dynamics filter was generated in an analogous way as the one for
 165 entropy. Concluding filter was the *blend* of mean entropy and mean dynamics
 166 filters. Figure 2 presents difference between unfiltered and filtered mask of
 167 region borders.

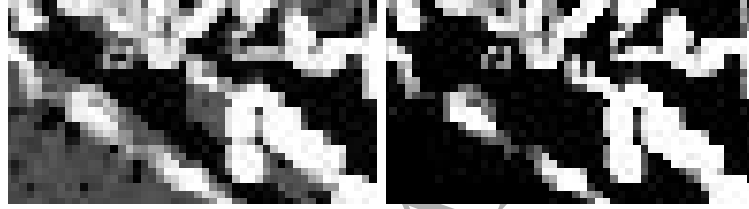


Figure 2: Mask of region borders before (left) and after filtering (right).

168 3.2. Feature computation

169 Filtering out the noise makes possible an effective usage of simple sta-
 170 tistical operations like maximum or minimum, and improve the precision
 171 of average, mean, mode or median value. We have proposed a set of four-
 172 teen features. Figure 3 provides HSV visualization three example features.
 173 Complete collection of statistical features is presented in Table 1.

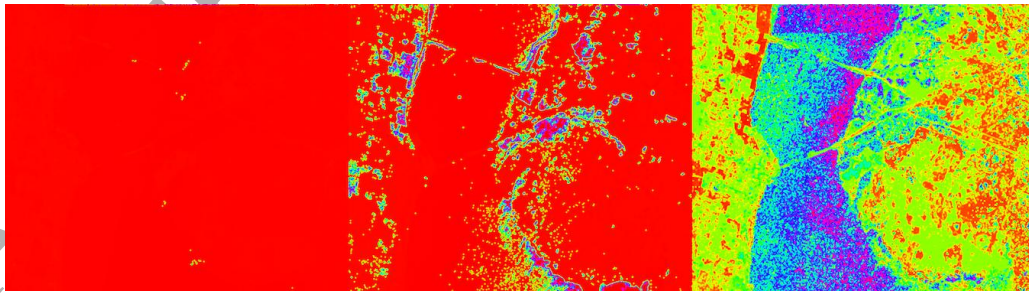


Figure 3: Color visualization of example metrics.

Abbr.	Description
<i>hsvr</i>	Red channel from pseudocolor HSV2RGB conversion.
<i>hsvg</i>	Green channel from pseudocolor HSV2RGB conversion.
<i>hsvb</i>	Blue channel from pseudocolor HSV2RGB conversion.
<i>min</i>	Lowest value in signature.
<i>min_idx</i>	Index of lowest value in signature
<i>max</i>	Highest value in signature.
<i>max_idx</i>	Index of highest value in signature.
<i>mean</i>	Mean value of signature.
<i>median</i>	Median value of signature.
<i>maxmin</i>	Difference between highest and lowest value in signature
<i>maxmin_dist</i>	Distance between indexes of highest and lowest value in signature.
<i>std</i>	Standard deviation from set of signature values.
<i>var</i>	Variance from set of signature values.
<i>mode</i>	Mode of quantified set of signature values.

Table 1: Statistical features implemented in algorithm

174 3.3. Preparing features for classification

175 As we can see in Figure 3, while some statistical features are giving us
176 clear information, enough to distinguish classes in data, some of them seems
177 completely useless. To extract, boost and stabilize data coming from them,
178 we have added three more steps of processing.

179 To stabilize information, we used the anisotropic diffusion [33]. Normal-
180 ization paired with histogram equalization brought us more contrast and
181 extraction of sparse values. Figure 4 shows the same set of metrics, after
182 these three steps.

183 Result of this three-staged process is presented in Figure 4.

184 4. Ensemble of Extreme Learning Machines

185 In this section we will present a brief description of the ELMs approach
186 and present details of the proposed ensemble approach.

187 4.1. Extreme Learning Machines

188 Extreme Learning Machines [14] are a family of algorithms designed for
189 fast, random-based training of single-layer feedforward neural networks. In

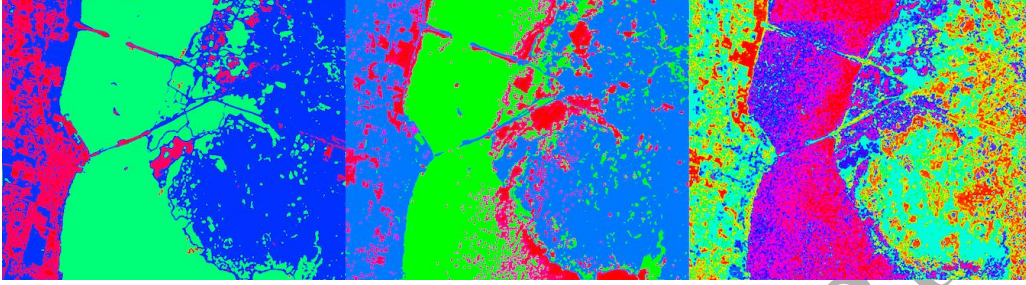


Figure 4: Color visualization of example metrics after anisotropic diffusion, normalization and histogram equalization.

last decades there were significant developments reported on methods designed for training accurate neural classifiers [22]. However, most of these approaches suffered from the extended computational time required for effective execution and a large number of parameters to be set. ELMs are one of recently emerging trends in neural-based classification that aims at alleviating the training complexities of its predecessor methods by using random weights assigned to hidden layer in a neural network. One must note here that despite the emerging popularity of ELMs-based approaches this concept can be traced further down in the literature to the proposals of Randomized Neural Networks [36] and Random Vector Functional Link [32].

Let us describe now the basic concept of ELMs. We assume that we have n labeled objects described by d features and a set of M labels. A single-layer feedforward neural network with N hidden neurons can be described by the following equation:

$$\mathbf{y} = \sum_{i=1}^N \mathbf{B}_i f(\mathbf{w}_i \cdot \mathbf{x} + b_i), \quad (3)$$

where $f()$ is the activation function, \mathbf{x} is the analyzed object, \mathbf{w}_i are the input weights for i -th hidden neuron, b_i is the bias of i -th hidden neuron and \mathbf{B}_i are weights assigned to outputs.

This equation with respect to all n points can be written in matrix form:

$$\mathbf{Y} = \mathbf{HB}, \quad (4)$$

where \mathbf{H} is the matrix consisting of outputs of hidden layer for each input object:

$$\mathbf{H} = \begin{pmatrix} f(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & f(\mathbf{w}_2 \cdot \mathbf{x}_1 + b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x}_1 + b_N) \\ f(\mathbf{w}_1 \cdot \mathbf{x}_2 + b_1) & f(\mathbf{w}_2 \cdot \mathbf{x}_2 + b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x}_2 + b_N) \\ \vdots & \vdots & \ddots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x}_n + b_1) & f(\mathbf{w}_2 \cdot \mathbf{x}_n + b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x}_n + b_N) \end{pmatrix}, \quad (5)$$

and $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_N)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$. To calculate the output weights \mathbf{B} is to compute the Moore-Penrose generalized inverse of the matrix \mathbf{H} , which we denote as \mathbf{H}^{-1} .

Basic ELM algorithm proceeds in three main steps:

1. Generate randomly the bias matrix $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N)^T$ and weight matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)^T$.
2. Calculate \mathbf{H} according to the Eq. 5.
3. Calculate the matrix of output weights $\mathbf{B} = \mathbf{H}^{-1}\mathbf{Y}$

However, there is a need for regularization in ELMs, which was reported as one of the crucial factors affecting their performance. To obtain it we can use an orthogonal projection to get the Moore-Penrose pseudoinverse of \mathbf{H} :

$$\mathbf{H}^{-1*} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \quad (6)$$

where \mathbf{H}^T is transposed matrix \mathbf{H} . This allows use to add a ridge parameter $\frac{1}{\lambda}$ to the diagonal of $(\mathbf{H}^T \mathbf{H})$. This is known as ridge-regression regularization approach [7] that results in a more stable solution. After applying this you calculate the matrix of output weights in step 3 as follows:

$$\mathbf{B} = \left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T \mathbf{H} \right)^{-1} \mathbf{H}^T \mathbf{Y}, \quad (7)$$

where \mathbf{I} is an identity matrix of equal size to \mathbf{H} .

4.2. Proposed Ensemble Architecture

ELMs can be considered as unstable classifiers due to their random nature. As the entire learning process relies on the randomly set weights in the first layer one cannot assure that given model will return highly efficient performance for every initialization. Therefore ensemble learning paradigm [42] started to attract the attention of ELMs community in recent years [3, 8, 29].

Most of the solutions proposed in this area uses the random generation of input weights as a diversification procedure to obtain a pool of base classifiers. This solution is based on the concept that different random initializations of ELMs will be sufficient to provide mutually complementary classifiers and that this will reduce the probability of using a weak model for classification. However, one may discuss the efficacy of such a method, as methods based on varying the input [19] or output [17] spaces were reported to deliver superior performance in varying multi-class scenarios.

Ensembles of ELMs use mainly voting procedures to combine individual outputs of base classifiers, usually assuming that each base classifier is equally important to the final decision making process (majority voting approach) [9]. One should note that such combination methods cannot take advantage of local specialization of its base classifiers and do not assume a varied quality of its base members.

In this paper we address these two important issues in forming ensembles of ELMs: how to create a pool of base classifiers with high individual quality and mutual diversity, and how to combine their individual outputs in the most efficient manner.

4.2.1. Forming Pool of ELM Classifiers

We propose to investigate the possibility of constructing ELMs committees on the basis of Random Subspace method (RSM) [18].

This method assumes that in the training set \mathcal{TS} we have at our disposal n objects, where x_j is the j -th training sample described as a d -dimensional feature vector in given feature space \mathcal{F} . Our aim is to construct an ensemble consisting of L classifiers. In RSM each base classifier is constructed using r features, where $r < d$ and features in r are selected randomly from \mathcal{F} .

One may see that RSM allows us to train a given number of classifiers, here each is based on a randomly reduced feature space. We assume that all of feature subspaces are of identical size r and that there is a possible overlap between these subspaces.

RSM is especially efficient for high-dimensional data, where it brings a benefit of both diversification of the committee members and simplification of their individual training procedures (as each base classifier work in a reduced space). However there is no clear indicator how many classifiers we should construct using RSM, therefore often the overproduce-and-select approach is used [23].

RSM seems as a highly attractive method for hyperspectral data analysis,

as here we deal with high-dimensional datasets (equivalent to the number of bands used) [43]. However high number of features can on one hand be beneficial to RSM (higher potential for diversification), but on the other will lead to very high number of classifiers being trained to actually cover the entire original feature space and to obtain good accuracy.

Therefore, we propose to combine RSM method with our statistical features described in Section 3. They allow to obtain a highly compressed feature space, extracting 14 different statistical channels. Such a new feature space offers a more compact representation, while still being able to benefit from RSM. Below we present justification for a good behavior of such a reduced space as an input for RSM.

The first issue that must be taken into consideration when designing a RSM-based ensemble is the complete coverage of the original feature space. That is a situation in which every single original feature is used by at least one classifier in the ensemble. As RSM does not rank the importance of features, then it should use all of available information to prevent when one or more features are randomly discarded and actually never used in RSM. One may calculate the probability of complete coverage of given RSM-based ensemble as follows:

$$P(\text{coverage}) = 1 - \left(1 - \frac{r}{d}\right)^L. \quad (8)$$

Here we can see that probability of full coverage decreases with the growing original space dimensional d and increases with the size of the ensemble L . This shows that for high-dimensional data we will require large ensembles when forming them on the basis of RSM.

Second issue important for RSM is the diversity of base classifiers. As we create each feature subspace in a random way there is some chance that certain classifiers will be trained on identical or fairly similar set of features. This of course does not contribute to the efficacy of ensemble being constructed, but only increases its overall computational complexity. One may calculate the probability of RSM-based ensemble consisting of classifiers with nonidentical subset of features as:

$$P(\text{non-id}) = \left(1 - \frac{1}{\binom{d}{r}}\right)^{L(L-1)/2}. \quad (9)$$

Here we can see that the probability of having a pool of non-identical

classifiers in RSM increases with the dimensionality d of the original feature space and decreases with the size of the ensemble L .

Therefore, one can see that these two goals are contradictory with the respect to the size of the ensemble.

We propose to create compact ensembles with RSM method based on our extracted statistical features. When transforming the original feature space into 14-dimensional one it is easier to obtain a small ensemble with full coverage, which at the same time increases the probability of classifiers within it being non-identical.

To further boost the quality of proposed compact ensemble we introduce a trained classifier combination approach for ELMs.

4.2.2. Trained Combination of ELMs

ELMs output continuous values for each of classes being considered. Therefore, we may consider such outputs as support values in form $F_m(x)$ which represents classifiers' support that object x belongs to m -th class. According to this the final class outputted by a single ELM classifier will be established according to maximum rule (winner-takes-all).

We propose to consider a weighted classifier combination based continuous outputs of ELMs for each of considered classes.

Assume that we have a pool of L classifiers $\Pi = \{\Psi^{(1)}, \Psi^{(2)}, \dots, \Psi^{(L)}\}$. For a given object x , each individual classifier decides whether it belongs to class $m \in \mathcal{M} = \{1, \dots, M\}$ based on the values of discriminants. Let $F_m^{(l)}(x)$ denote a function that is assigned to class m for a given value of x and that is used by the l -th classifier Ψ^l . The combined classifier Ψ uses the following decision rule [20]:

$$\Psi(x) = m \iff \hat{F}_m(x) = \max_{k \in \mathcal{M}} \hat{F}_k(x). \quad (10)$$

There is a number of proposal on how to assign weights to classifiers. We propose to use an approach in which weights dependent on the classifier and class number: Weight w_m^l is assigned to the l -th classifier and the i -th class:

$$\hat{F}_m(x) = \sum_{l=1}^L w_m^l F_m^l(x). \quad (11)$$

with the respect to constraint:

$$\sum_{l=1}^L w_m^l = 1. \quad (12)$$

Here, the given classifier weights assigned to different classes may differ, which allows us to obtain a highly flexible ensemble structure. This way we can exploit the local competencies of each base ELM classifier. A single ELM may have assigned high weights for classes which are recognized accurately by it and low weights for classes in which it is deemed as non-competent. This is in line with the idea of ELMs ensembles, as we may obtain different base models due to the random initialization process. Additionally, it may counter the drawbacks of RSM, as it controls the degree of importance of each base ELM and can reduce the negative effects produced by base models trained on weak feature subspaces.

We need an efficient method to compute the weights assigned to each class and classifier, thus obtaining a trained combiner. We propose to use a highly efficient perceptron-based combiner. Here we delegate a single perceptron for each class as an aggregation function, which may be trained with any standard procedure used in neural network learning. The input weights established during the learning process are then used as the weights assigned to each of the base classifiers. The implementation of the proposed combiner is presented in Figure 5.

5. Experimental Study

The experimental study was designed to provide answers to the three following questions:

- Is the proposed 14-channel statistical representation superior to using pixel-based one?
- Is there any benefit from using RSM-based ensembles of ELMs.
- Is the proposed trained combiner more efficient for combination of ELMs than popular voting approach?

In the following subsections we will present details about datasets used, set-up of our experiments, obtained results and their meaning.

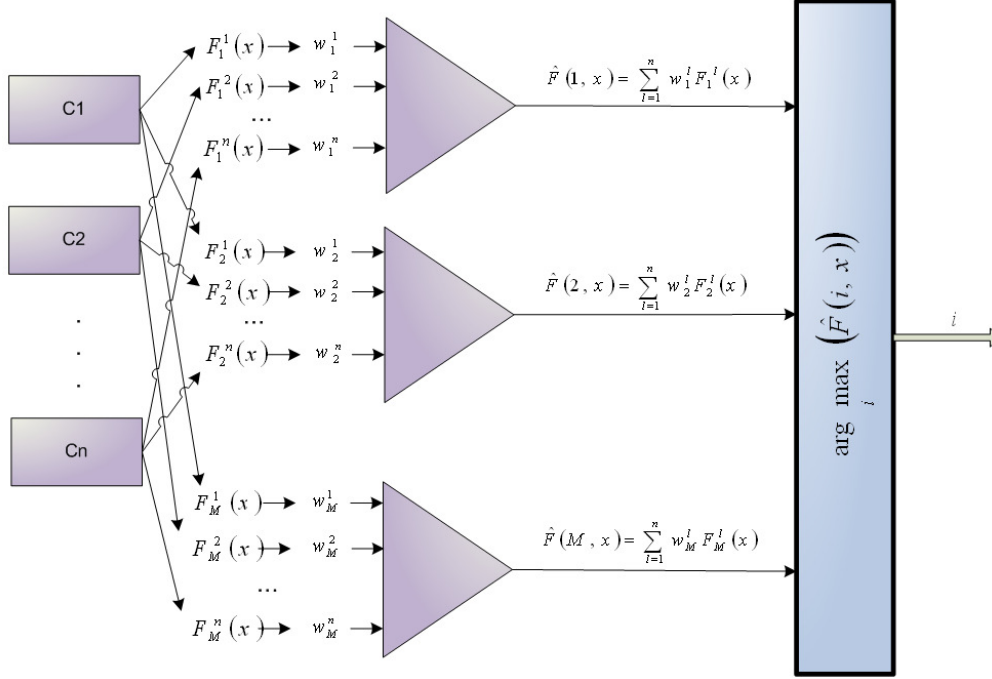


Figure 5: The idea of the trained combiner, which is a linear combination of the support functions returned by the individual classifiers. It is implemented as an one-layer perceptron, where one perceptron fuser is constructed for each of the classes under consideration.

5.1. Datasets

In experiments we are using hyperspectral imaging database provided by Group of the Computational Intelligence from *Universidad del Pais Vasco* (UPV/EHU)². It consists of seven images described by ground truth maps.

- **Salinas** scene, collected by the AVIRIS sensor over Salinas Valley, California, with high spatial resolution (3.7-meter pixels). It includes vegetables, bare soils, and vineyard fields.
- **Salinas A**, which is an small sub-scene of Salinas image.
- **Indian Pines**, gathered by AVIRIS sensor over the Indian Pines test site in North-western Indiana. Scene contains two-thirds agriculture,

²http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

and one-third forest or other natural perennial vegetation. There are two major dual lane highways, a rail line, as well as low density housing, other built structures, and smaller roads. Since the scene is taken in June some of the crops present, corn, soybeans, are in early stages of growth with less than 5% coverage.

- **Pavia Centre and Pavia University**, acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The geometric resolution is 1.3 meters. Pavia scenes were provided by Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, Pavia university (Italy).
- **Botswana**, acquired by the NASA EO-1 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 400-2500 nm portion of the spectrum in 10 nm windows. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Identified classes are representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta.
- **Kennedy Space Center (KSC)**, acquired by AVIRIS sensor over the Kennedy Space Center, Florida, on March 23, 1996. Data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. Training 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 discernable at the spatial resolution of Landsat. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types.

Detailed informations about images are included in Table 2, and cropped previews are shown in Figure 6.

5.2. Set-up

We propose to compare the proposed method with widely used single-model ELMs and their voting ensembles. Additionally, we present results

Name	Spatial res.	Depth	Classes	Sensor
Salinas	512 × 217	224	16	AVIRIS
Salinas A	86 × 83	224	6	AVIRIS
Indian Pines	145 × 145	224	16	AVIRIS
Pavia University	610 × 610	103	9	ROSIS
Pavia Centre	1096 × 1096	102	9	ROSIS
Botswana	256 × 1476	242	14	Hyperion
KSC	614 × 512	224	14	AVIRIS

Table 2: Parameters of datasets

obtained by Support Vector Machine (SVM) [5] in one-versus-all (OVA) and one-versus-one (OVO) settings [16].

We use a 5x2 fold CV combined F-test [2] for simultaneous training/testing and pairwise statistical analysis. It repeats two times five-fold cross-validation. The combined F-test is conducted by comparison of all versus all. As a test score the probability of rejecting the null hypothesis is adopted, i.e. that classifiers have the same error rates. As an alternative hypothesis, it is conjectured that tested classifiers have different error rates. A small difference in the error rate implies that the different algorithms construct two similar classifiers with similar error rates; thus, the hypothesis should not be rejected. For a large difference, the classifiers have different error rates and the hypothesis should be rejected.

For the combiner training purposes we utilize 10% of the training data.

Classifier parameters are optimized using internal 3 fold CV. Details regarding their parameter setting are presented in Table 3.

5.3. Results and Discussion

Detailed experimental results with respect to obtained accuracy and statistical significance are depicted in Table 4, while Table 5 presents the averaged sizes of formed ensembles.

Let us now take a closer look on the obtained results.

When comparing data representations approaches we can see a significant improvement obtained when using the proposed statistical-based feature extraction. Regardless of the type of classifier used (single or ensemble models) we can observe a significant accuracy gain in five datasets. For the remaining two (Salinas A and Botswana) results were similar to using full feature space.

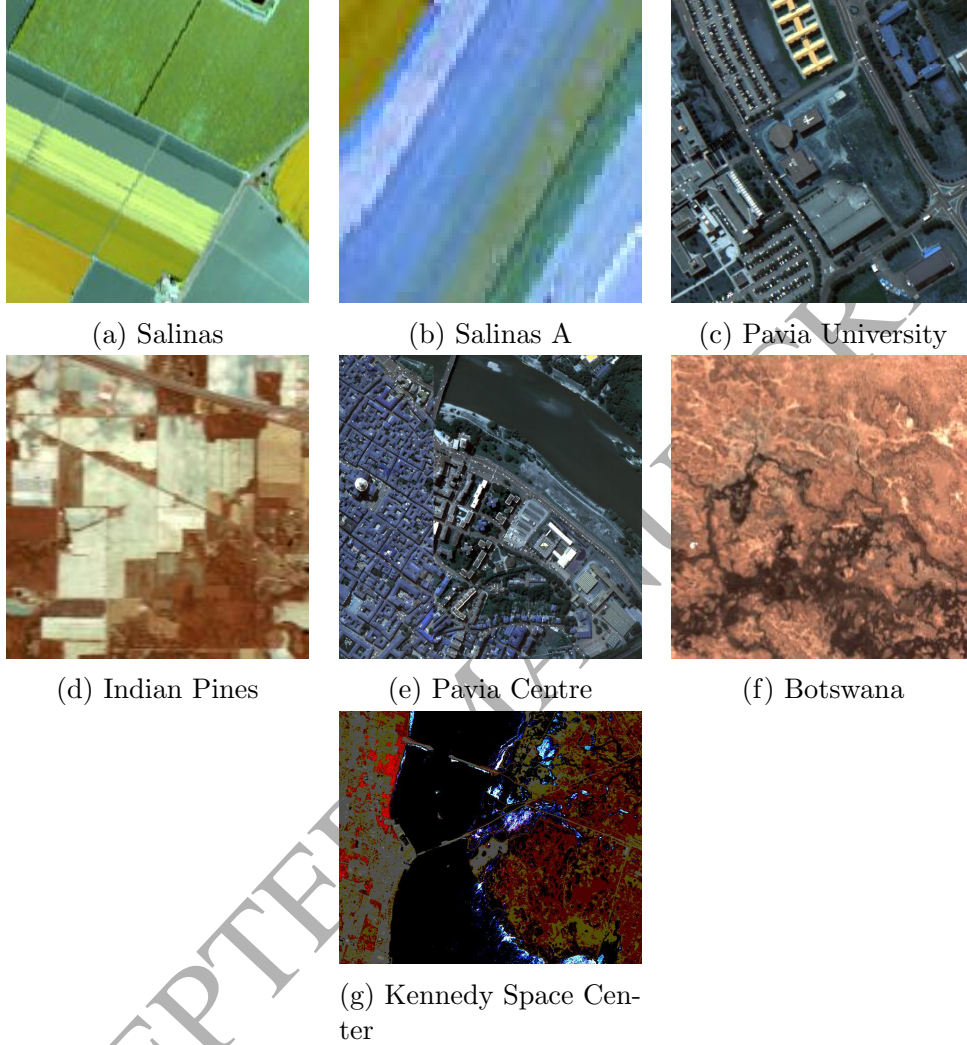


Figure 6: Cropped preview of datasets

426 However, we must note that for all seven datasets using the proposed feature
 427 set did not lead to a decrease of classifiers performance. Thus we are able
 428 to conclude that the proposed approach offers an effective low-dimensional
 429 embedding that reduces the complexity of trained classifiers and offers highly
 430 discriminating features that can be used as an efficient input regardless of
 431 the base classifier selected.

432 When taking into account used classifiers we can see that SVM in OVO

Table 3: Details of classifier parameters used in the experiments. Presented are ranges of parameter values. The final values were established using internal 3 fold CV.

Algorithm	Parameters
ELM	no. of hidden neurons $\in [10;100]$ activation function = sigmoid $\lambda = 10$
SVM	$C \in [0.1;1.0]$ tolerance parameter = 0.001 $\epsilon \in [1.0E-12;1.0E-8]$ kernel = polynomial polynomial degree $\in [1;3]$ OVA = maximum confidence strategy OVO = weighted voting strategy
RSM	no. of subspaces $\in [10;50]$ size of subspaces = $\lfloor 0.4d \rfloor$
Perceptron combiner	activation function = sigmoid learning rule = MADALINE

Table 4: Accuracies [%] of examined methods in hyperspectral image classification task. Small symbols under each result indicate the indexes of methods from which the considered one was statistically significantly better according to the combined 5x2 CV F-test.

Dataset	Pixel-based					Statistical-based		
	SVM ¹ _{OVA}	SVM ² _{OVO}	ELM ³	RSM-ELM ⁴ _{vot}	RSM-ELM ⁵ _{per}	ELM ⁶	RSM-ELM ⁷ _{vot}	RSM-ELM ⁸ _{per}
Salinas	68.83	73.04	70.46	72.98	74.40	71.03 1	73.45 1,3	76.92 ALL
Salinas A	98.20	99.64	97.61	98.04	98.51	97.92 —	98.45 3	99.02 3,4,6
Pavia U	88.62	90.76	88.93	91.58	93.99	90.93 1,3	93.64 1,2,3,4	96.51 ALL
I Pines	61.94	64.73	62.86	63.34	65.83	65.18 1,3,4	67.06 1,2,3,4,5,6	68.94 ALL
Pavia C	89.67	90.70	90.18	91.18	93.65	93.24 1,2,3,4	95.38 1,2,3,4,5,6	96.90 ALL
Botswana	96.04	97.98	93.18	94.14	94.97	93.21 —	94.09 —	94.91 3,6
KSC	86.16	90.88	87.59	90.27	92.60	87.97 1	90.81 1,3,6	92.89 1,2,3,4,6,7

Table 5: Average ensemble sizes (rounded) with standard deviations for examined ELMs committees.

Dataset	Pixel-based		Statistical-based	
	RSM-ELM _{vot}	RSM-ELM _{per}	RSM-ELM _{vot}	RSM-ELM _{per}
Salinas	36±4.78	29±3.18	27±3.46	20±2.66
Salinas A	25±5.32	20±2.74	24±5.16	20±2.70
Pavia U	17±3.60	16±1.72	15±2.62	14±1.18
I Pines	38±5.91	30±2.99	34±3.58	26±1.98
Pavia C	43±6.71	35±3.88	27±6.02	23±2.63
Botswana	31±4.52	28±3.81	32±4.74	28±3.70
KSC	33±6.92	27±2.89	31±5.34	21±2.02

mode always outperforms single ELM. This can be explained by the fact, that while ELM tries to fit a single model for multi-class problem, SVM actually trains a number of binary classifiers on decomposed pairwise subtasks. Therefore, it can offer superior performance for hyperspectral data with high number of classes. This situation changes however when using ensemble learning. Ensembles of ELM, regardless of their combination method, offer significantly improved performance and are able to outperform SVMs. This shows that our Random Subspace-based ensemble has two beneficial properties: by using several models we reduce the variance caused by randomized neural network architectures and by training each model on a subset of features we further increase the ensemble diversity. Both these factors directly translate to improved accuracy.

When comparing combination methods we observe that the proposed trained combiner is able to greatly improve the efficacy of Extreme Learning Ensembles. By using a weighted combination based on support functions and assigning weights for each classifier and class we are able to vary the importance of each base model and exploit their local competencies. This offers much more flexibility than using voting strategy (which is popular for combining ELMs) as we do not assume that each classifier is equally competent for each of classes. This allows us to further counter the randomness embedded in ELMs training procedure and Random Subspaces method by reducing the importance of weaker or non-diverse models.

Finally, let us analyze the obtained ensemble sizes (please refer to Table 5). We can see that using our proposed feature extraction method leads

to smaller ensembles with lower variance, as by using smaller feature space we reduce the number of classifiers needed for full coverage. Additionally, the usage of trained combiner allowed to further reduce the needed number of base classifiers, as we are now able to much more efficiently exploit the given set of learners than in case of voting procedures (where due to their randomness and equal importance we needed larger committees). This also leads to more stable ensembles, as variance in their size over folds of CV is greatly reduced.

6. Conclusions and Future Works

In this paper we have addressed the problem of hyperspectral data classification from the perspective of data representation and learning schemes. A new feature extraction method based on statistical channels was proposed. It allowed for a low-dimensional embedding of the original feature space, thus reducing the complexity of analyzed data. We showed how to compute 14 diverse metrics from any hyperspectral image and process them to improve their discriminatory power.

On the basis of these features we proposed to form a novel ensemble of randomized neural networks. It used Extreme Learning Machines as base classifier and used Random Subspaces method to improve the diversity among ensemble members. This architecture was augmented with a trained classifier combination step that used a perceptron-based training method to compute weights. These were assigned to each classifier and class individually, thus resulting in a flexible exploitation of local competencies of base classifiers and efficient multi-class pattern recognition. Additionally, we showed that such an ensemble method combined with proposed feature representation not only boosts the accuracy, but also leads to forming more sparse committees.

Obtained results encourage us to work with statistical-based data representation and Extreme Learning Ensembles. In future we plan to apply our framework to semi-supervised scenario with limited access to class labels and self-labeling mechanisms.

Acknowledgment

This was supported by the Polish National Science Center under the grant no. DEC-2013/09/B/ST6/02264.

All experiments were carried out using computer equipment sponsored by EC under FP7, Coordination and Support Action, Grant Agreement Number 316097, ENGINE - European Research Centre of Network Intelligence for Innovation Enhancement (<http://engine.pwr.wroc.pl/>).

References

- [1] A. Agarwal, T. El-Ghazawi, H. El-Askary, and J. Le-Moigne. Efficient hierarchical-PCA dimension reduction for hyperspectral imagery. In *2007 IEEE International Symposium on Signal Processing and Information Technology*, pages 353–356, December 2007.
- [2] Ethem Alpaydin. Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. *Neural Computation*, 11(8):1885–1892, 1999.
- [3] B. Ayerdi and M. Graña. Hybrid extreme rotation forest. *Neural Networks*, 52:33–42, 2014.
- [4] B. Ayerdi and M. Graña. Hyperspectral image nonlinear unmixing and reconstruction by ELM regression ensemble. *Neurocomputing*, 174:299–309, 2016.
- [5] Y. Bazi and F. Melgani. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE T. Geoscience and Remote Sensing*, 44(11-2):3374–3385, 2006.
- [6] Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, pages 368–374. MIT Press, 1998.
- [7] P. Buteneers, K. Caluwaerts, J. Dambre, D. Verstraeten, and B. Schrauwen. Optimized parameter search for large datasets of the regularization parameter and feature selection for ridge regression. *Neural Processing Letters*, 38(3):403–416, 2013.
- [8] J. Cao, S. Kwong, R. Wang, X. Li, K. Li, and X. Kong. Class-specific soft voting based multiple extreme learning machines ensemble. *Neurocomputing*, 149:275–284, 2015.

- [9] J. Cao, Z. Lin, G.-B. Huang, and N. Liu. Voting based extreme learning machine. *Inf. Sci.*, 185(1):66–77, 2012.
- [10] R.T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989.
- [11] Q. Dai. A competitive ensemble pruning approach based on cross-validation technique. *Knowledge-Based Systems*, 37:394–414, 2013.
- [12] E. R. Davies. *Machine Vision: Theory, Algorithms, Practicalities*. Elsevier, December 2004.
- [13] Thomas Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin, Heidelberg, 2000.
- [14] S. Ding, Zhao. H., Y. Zhang, X. Xu, and R. Nie. Extreme learning machine: algorithm, theory and applications. *Artif. Intell. Rev.*, 44(1):103–115, 2015.
- [15] J.M. Durand and Y.H. Kerr. An improved decorrelation method for the efficient display of multispectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 27(5):611–619, September 1989.
- [16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776, 2011.
- [17] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera. DRCW-OVO: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. *Pattern Recognition*, 48(1):28–42, 2015.
- [18] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:832–844, August 1998.
- [19] K. Jackowski, B. Krawczyk, and M. Woźniak. Improved adaptive splitting and selection: the hybrid training method of a classifier based on a feature space partitioning. *Int. J. Neural Syst.*, 24(3), 2014.

- [20] Robert A. Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, 1995.
- [21] A.K. Jain, R.P.W. Duin, and Jianchang Mao. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, jan 2000.
- [22] L. C. Jain, M. Seera, C. P. Lim, and P. Balasubramaniam. A review of online learning in supervised neural networks. *Neural Computing and Applications*, 25(3-4):491–509, 2014.
- [23] A. H.-R. Ko, R. Sabourin, L. E. S. de Oliveira, and A. de Souza Britto Jr. The implication of data diversity for a classifier-free ensemble selection in random subspaces. In *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pages 1–5, 2008.
- [24] B. Krawczyk, P. Ksieniewicz, and M. Woźniak. Hyperspectral image analysis based on color channels and ensemble classifier. In *Proceedings of the 9th International Conference on Hybrid Artificial Intelligence Systems - Volume 8480*, HAIS 2014, pages 274–284, New York, NY, USA, 2014. Springer-Verlag New York, Inc.
- [25] L.I. Kuncheva and L.C. Jain. Designing classifier fusion systems by genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 4(4):327 – 336, nov 2000.
- [26] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.
- [27] 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.
- [28] W. Li, C. Chen, H. Su, and Q. Du. Local binary patterns and extreme learning machine for hyperspectral imagery classification. *IEEE T. Geoscience and Remote Sensing*, 53(7):3681–3693, 2015.
- [29] N. Liu and H. Wang. Ensemble based extreme learning machine. *IEEE Signal Process. Lett.*, 17(8):754–757, 2010.

- [30] Gian Luca Marcialis and Fabio Roli. Fusion of face recognition algorithms for video-based surveillance systems. *G.L. Foresti, C. Regazzoni, P. Varshney Eds*, pages 235–250, 2003.
- [31] R. Moreno, F. Corona, A. Lendasse, M. Graña, and L. S. Galvão. Extreme learning machines for soybean classification in remote sensing hyperspectral images. *Neurocomputing*, 128:207–216, 2014.
- [32] Y.-H. Pao, G. H. Park, and D. J. Sobajic. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 6(2):163–180, 1994.
- [33] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, July 1990.
- [34] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann Publishers, 1993.
- [35] Suju Rajan, Joydeep Ghosh, and Melba M. Crawford. An active learning approach to hyperspectral data classification. *IEEE T. Geoscience and Remote Sensing*, 46(4):1231–1242, 2008.
- [36] W.F. Schmidt, M.A. Kraaijveld, and R.P.W. Duin. Feedforward neural networks with random weights. In *Pattern Recognition, 1992. Vol.II. Conference B: Pattern Recognition Methodology and Systems, Proceedings, 11th IAPR International Conference on*, pages 1–4, 1992.
- [37] G SVAETICHIN. Spectral response curves from single cones. *Acta physiologica Scandinavica. Supplementum*, 39(134):1746, 1956. PMID: 13444020.
- [38] J.S. Tyo, A. Konsolakis, D.I. Diersen, and R.C. Olsen. Principal-components-based display strategy for spectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 41(3):708–718, March 2003.
- [39] M. van Erp, L. Vuurpijl, and L. Schomaker. An overview and comparison of voting methods for pattern recognition. In *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, pages 195 – 200, 2002.

- [40] T. Wilk and M. Woźniak. Complexity and multithreaded implementation analysis of one class-classifiers fuzzy combiner. In Emilio Corchado, Marek Kurzynski, and Michal Wozniak, editors, *Hybrid Artificial Intelligent Systems*, volume 6679 of *Lecture Notes in Computer Science*, pages 237–244. Springer Berlin / Heidelberg, 2011.
- [41] David H. Wolpert. The supervised learning no-free-lunch theorems. In *In Proc. 6th Online World Conference on Soft Computing in Industrial Applications*, pages 25–42, 2001.
- [42] M. Woźniak, M. Graña, and E. Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.
- [43] J. Xia, M. D. Mura, J. Chanussot, P. Du, and X. He. Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles. *IEEE T. Geoscience and Remote Sensing*, 53(9):4768–4786, 2015.



Paweł Ksieniewicz is a research assistant at Wrocław University of Technology, where he achieved BSc in 2012, MSc and started doctoral studies on computer science in 2013. His research focuses on multidimensional data representation and image processing. Most of his papers concerns the

hyperspectral imaging in context of data segmentation and visualization.



633

634 **Bartosz Krawczyk** received an MSc degree in 2012 and PhD degree in 2015,
 635 both with distinctions from Wroclaw University of Technology, Poland. He
 636 is an Assistant Professor at the Department of Systems and Computer Net-
 637 works of the same university. His research is focused on machine learning,
 638 multiple classifier systems, one-class classifiers, class imbalance, data streams
 639 and interdisciplinary applications of these methods. He has authored 26
 640 international journal papers and over 75 contributions to conferences. Dr
 641 Krawczyk was awarded with numerous prestigious awards for his scientific
 642 achievements like PRELUDIUM and ETIUDA grants from Polish National
 643 Science Center, Scholarship of Polish Minister of Science and Higher Educa-
 644 tion or START award from Foundation for Polish Science among others. He
 645 served as a Guest Editor in four journal special issues devoted to ensemble
 646 learning and data stream classification. He is a member of Program Com-
 647 mittee for over 40 international conferences and a reviewer for 22 journals.



648

649 **Micha Woźniak** is a professor of computer science at the Department of
650 Systems and Computer Networks, Wrocław University of Technology, Poland.
651 He received M.Sc. degree in biomedical engineering from the Wrocław Uni-
652 versity of Technology in 1992, and Ph.D. and D.Sc. (habilitation) degrees in
653 computer science in 1996 and 2007, respectively, from the same university.
654 His research focuses on compound classification methods, hybrid artificial
655 intelligence and medical informatics. Prof. Woźniak has published over 260
656 papers and three books. His recent one Hybrid classifiers: Method of Data,
657 Knowledge, and Data Hybridization was published by Springer in 2014. He
658 has been involved in several research projects related to the above-mentioned
659 topics and has been a consultant of several commercial projects for wellknown
660 Polish companies and public administration. Prof. Woźniak is a senior mem-
661 ber of the IEEE and a member of the International Biometric Society.