Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Ensemble of Extreme Learning Machines with trained classifier combination and statistical features for hyperspectral data

Paweł Ksieniewicz[a], Bartosz Krawczyk[b,*], Michał Woźniak[a]

[a] Department of Systems and Computer Networks, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50–370 Wrocław, Poland
[b] Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Article history:
Received 27 December 2015
Accepted 25 April 2016
Available online 6 July 2017

Keywords:
Ensemble learning
Extreme Learning Machines
Hyperspectral imaging
Computer vision
Feature extraction
Dimensionality reduction
Image classification

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.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

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

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

- Classifier ensembles behave well both in the case when a learning set is very small and when we have a huge amount of learning examples at our disposal. In the first case, classifier ensemble can exploit methods based on bootstrapping [30], while for the second case it allows to train individuals on partitions of dataset.
- Classifier ensemble may outperform the best individual classifier [10] and under some conditions (e.g., majority voting by a group of individual classifiers committed error independently) this improvement has been proven analytically [26].
- Many classifier training methods, as decision tree [34], are heuristic search algorithms which usually suffer from local

* Corresponding author.
E-mail addresses: pawel.ksieniewicz@pwr.edu.pl (P. Ksieniewicz), bkrawczyk@vcu.edu (B. Krawczyk), michal.wozniak@pwr.edu.pl (M. Woźniak).

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

http://dx.doi.org/10.1016/j.neucom.2016.04.076
0925-2312/© 2017 Elsevier B.V. All rights reserved.

optima. Therefore, the ensemble learning approach approach is equivalent to a multi-start local random search which increases the probability of finding an optimal model.

- Classifier ensemble may be easy implemented in efficient computing environments such as parallel and multithreaded computer architectures [40].

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

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

The main contributions of the paper are as follows:

1. A novel proposition of the statistical-based feature extraction from hyperspectral images.
2. An efficient ELMs ensemble architecture based on trained combiner.
3. Application of the proposed features and ensemble structure together with Random Subspaces method to the problem of hyperspectral image classification.
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. 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 Fig. 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, after posterization, set of labels. It also implements a method of separation of homogenous areas on image, used also to filter noisy ranges of spectrum.

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

Most common method of generating false-color pictures from hyperspectral data is maping three bands from a wide signature into RGB channels. For case of spectral depth reduction, the most popular standard is PCA (*Principal Components Analysis*) [1]. Three, richest in information, principal components from hyperspectral cube are mapped to various color models chanels (RGB, HSL, HSV) [38].

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

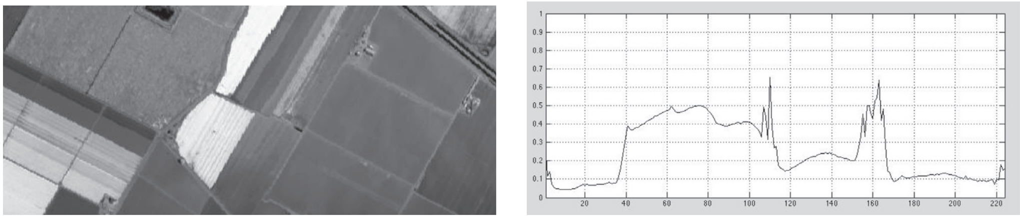
3. Proposed statistical features for hyperspectral images

We propose a novel method for simultaneous feature extraction and low-dimensional embedding of hyperspectral images. It is based on the idea of creating a new representation, evolving from human color perception, preserving as much information as possible, with simple, time efficient computations.

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

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

The procedure run as follows. At the start, the class edges are recognized and calculated. Next, the collection of side information



(a) Slice of single spectral band

(b) Example spectral signature

Fig. 1. Hyperspectral image elements.

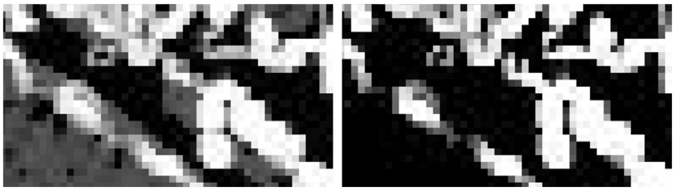


Fig. 2. Mask of region borders before (left) and after filtering (right).

produced during first step, lets us to generate a filter for noisy bands of image. Next the features of filtered image are computed and, at the end, they are prepared for classification.

3.1. Noise detection

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

$$\bar{\mathrm{H}} = \frac{\sum \rho}{ppl} \tag{1}$$

While every hyperspectral cube contains wavelengths with high noise ratio, adequate threshold to drain most of them would be a mean value of entropy. To separate hills of entropy changes we are using information about its dynamics. An vector of dynamics was made in a way analogous to edge detection, by calculating discrepancy between actual ($\bar{H}$) and next value ($\bar{H}'$) on the vector of entropy.

$$\mathrm{D\bar{H}} = |\bar{H} - \bar{H}'| \tag{2}$$

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

3.2. Feature computation

Filtering out the noise makes possible an effective usage of simple statistical operations like maximum or minimum, and improve the precision of average, mean, mode or median value. We have proposed a set of fourteen features. Fig. 3 provides HSV visualization three example features. Complete collection of statistical features is presented in Table 1.

3.3. Preparing features for classification

As we can see in Fig. 3, while some statistical features are giving us clear information, enough to distinguish classes in data, some of them seems completely useless. To extract, boost and stabilize data coming from them, we have added three more steps of processing. To stabilize information, we used the anisotropic diffusion [33]. Normalization paired with histogram equalization brought us more contrast and extraction of sparse values. Fig. 4 shows the same set of metrics, after these three steps.

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

4. Ensemble of Extreme Learning Machines

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

4.1. Extreme Learning Machines

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

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

where $f\,()$ is the activation function, $\mathbf{x}$ is the analyzed object, $\mathbf{w}_i$ are the input weights for ith hidden neuron, b_i is the bias of ith 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:

$$\mathbf{Y} = \mathbf{HB}, \tag{4}$$

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

$$\mathbf{H} = \begin{pmatrix} f(\mathbf{w}_1 \cdot \mathbf{x_1}+b_1) & f(\mathbf{w}_2 \cdot \mathbf{x_1}+b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x_1}+b_N) \\ f(\mathbf{w}_1 \cdot \mathbf{x_2}+b_1) & f(\mathbf{w}_2 \cdot \mathbf{x_2}+b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x_2}+b_N) \\ \vdots & \vdots & \ddots & \vdots \\ f(\mathbf{w}_1 \cdot \mathbf{x_n}+b_1) & f(\mathbf{w}_2 \cdot \mathbf{x_n}+b_2) & \cdots & f(\mathbf{w}_N \cdot \mathbf{x_n}+b_N) \end{pmatrix}, \tag{5}$$

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

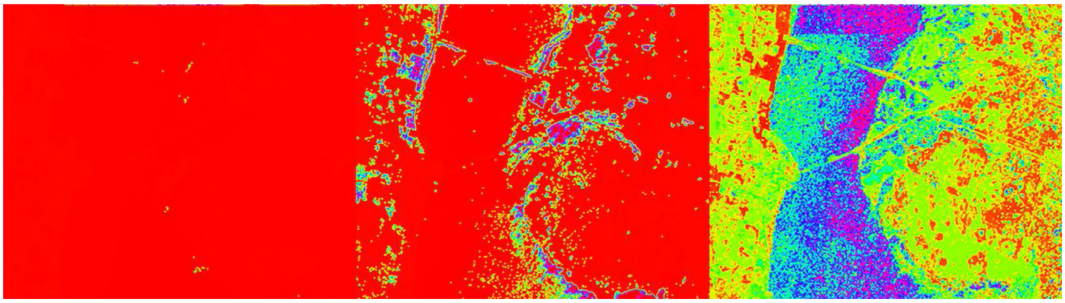


Fig. 3. Color visualization of example metrics.

Table 1
Statistical features implemented in algorithm.

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

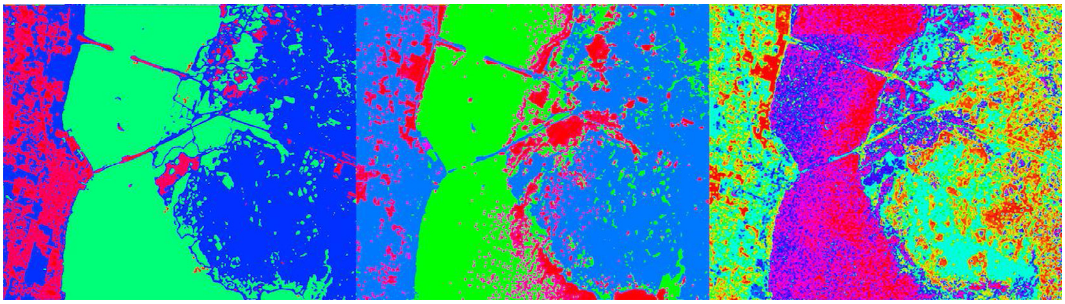


Fig. 4. Color visualization of example metrics after anisotropic diffusion, normalization and histogram equalization.

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 $\mathbf{H}$:

$$\mathbf{H}^{-1^*} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T \tag{6}$$

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:

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

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 jth 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:

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

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 it overall computational complexity. One may calculate the probability of RSM-based ensemble consisting of classifiers with nonidentical subset of features as:

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

Here we can see that the probability of having a pool of nonidentical 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 $\Pi = \{\Psi^{(1)}, \Psi^{(2)}, \ldots, \Psi^{(L)}\}$. For a given object x, each individual classifier decides whether it belongs to class $m \in \mathcal{M} = \{1, \ldots, M\}$ 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 lth classifier Ψ^l. The combined classifier Ψ uses the following decision rule [20]:

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

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 lth classifier and the ith class:

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

with the respect to constraint:

$$\sum_{l=1}^{L} w_m^l = 1. \tag{12}$$

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

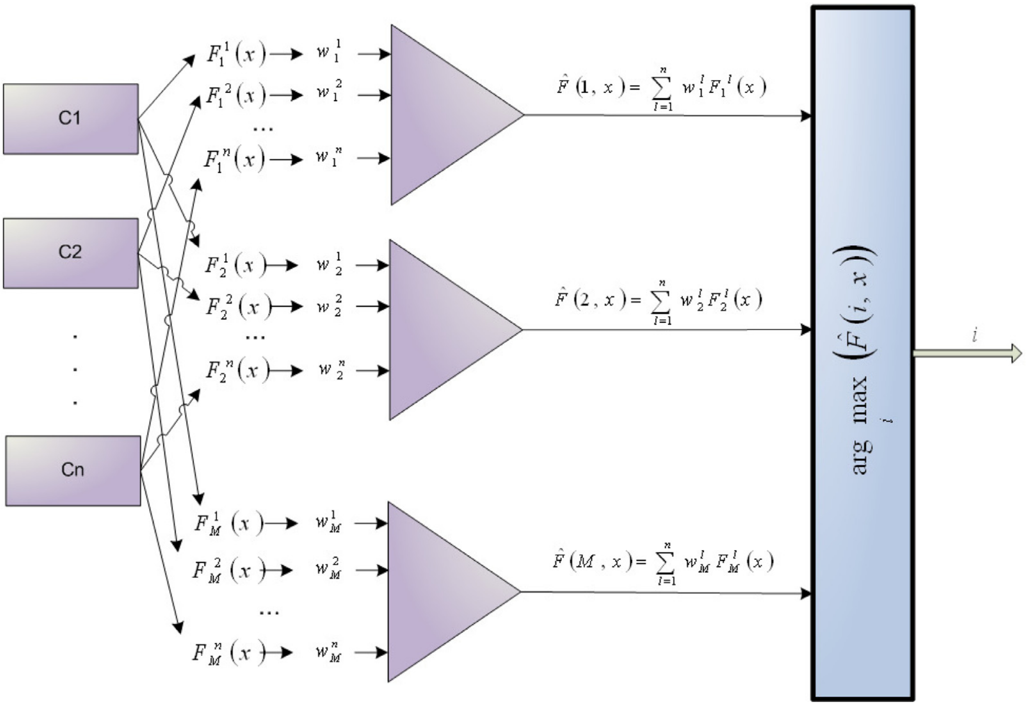


Fig. 5. The idea of the trained combiner, which is a linear combination of the support functions returned by the individual classifiers. It is implemented as an one-layer perceptron, where one perceptron fuser is constructed for each of the classes under consideration.

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 Fig. 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.

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, 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, nothern Italy. The geometric resolution is 1.3 m. 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 Fig. 6.

5.2. Set-up

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

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

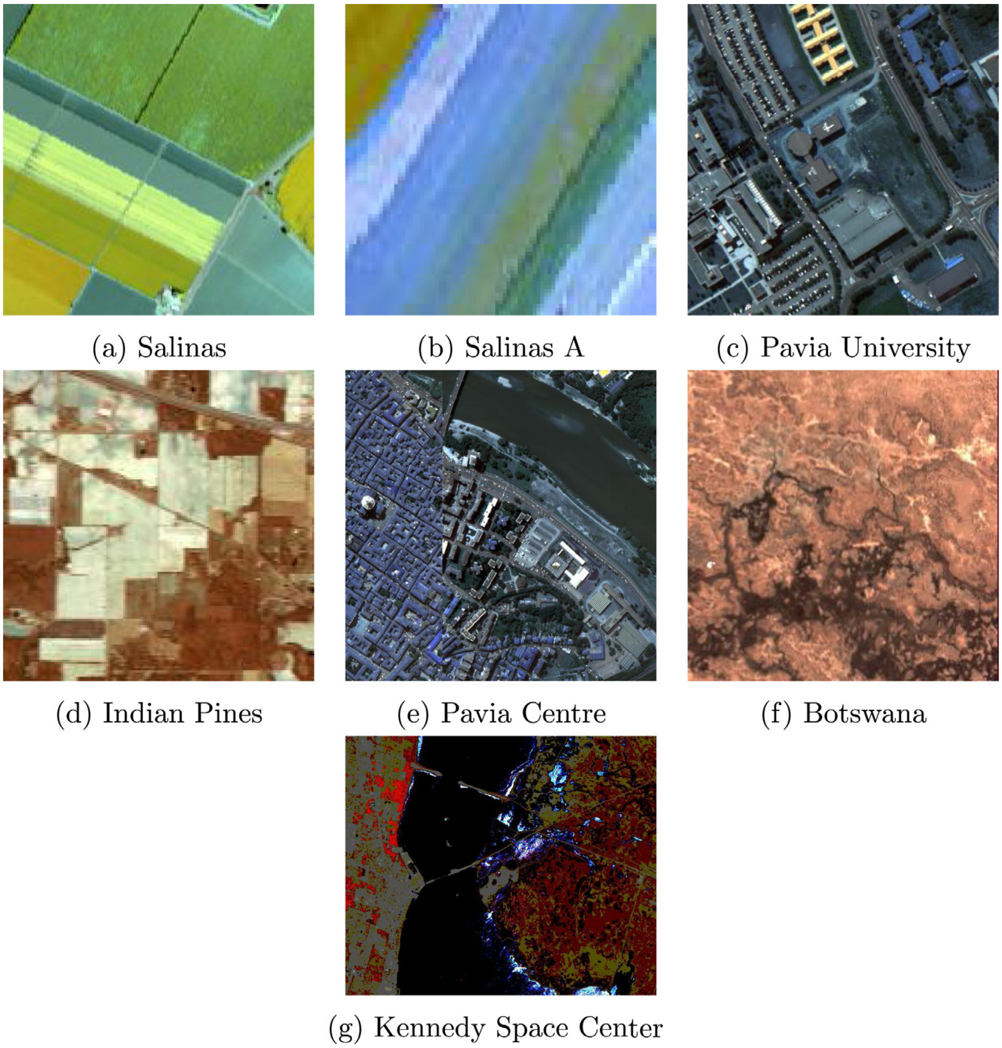


Fig. 6. Cropped preview of datasets.

Table 2
Parameters of datasets.

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

present results 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.

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

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.

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	$\mathrm{RSM{-}ELM}^4_{vot}$	$\mathrm{RSM{-}ELM}^5_{per}$	ELM^6	$\mathrm{RSM{-}ELM}^7_{vot}$	$\mathrm{RSM{-}ELM}^8_{per}$
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	
	$\mathrm{RSM{-}ELM}_{vot}$	$\mathrm{RSM{-}ELM}_{per}$	$\mathrm{RSM{-}ELM}_{vot}$	$\mathrm{RSM{-}ELM}_{per}$
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

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

When taking into account used classifiers we can see that SVM in OVO 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 he 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, J. Le-Moigne, Efficient hierarchical-PCA dimension reduction for hyperspectral imagery, in: Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, 2007, pp. 353–356, doi:10.1109/ISSPIT.2007.4458191.

[2] E. Alpaydin, Combined 5 x 2cv f test for comparing supervised classification learning algorithms, Neural Comput. 11 (8) (1999) 1885–1892.

[3] B. Ayerdi, M. Graña, Hybrid extreme rotation forest, Neural Netw. 52 (2014) 33–42.

[4] B. Ayerdi, M. Graña, Hyperspectral image nonlinear unmixing and reconstruction by ELM regression ensemble, Neurocomputing 174 (2016) 299–309.

[5] Y. Bazi, F. Melgani, Toward an optimal SVM classification system for hyperspectral remote sensing images, IEEE Trans. Geosci. Remote Sens. 44 (11–2) (2006) 3374–3385.

[6] K.P. Bennett, A. Demiriz, Semi-supervised support vector machines, in: Proceedings of the Advances in Neural Information Processing Systems, MIT Press, 1998, pp. 368–374.

[7] P. Buteneers, K. Caluwaerts, J. Dambre, D. Verstraeten, B. Schrauwen, Optimized parameter search for large datasets of the regularization parameter and feature selection for ridge regression, Neural Process. Lett. 38 (3) (2013) 403–416.

[8] J. Cao, S. Kwong, R. Wang, X. Li, K. Li, X. Kong, Class-specific soft voting based multiple extreme learning machines ensemble, Neurocomputing 149 (2015) 275–284.

[9] J. Cao, Z. Lin, G. Huang, N. Liu, Voting based extreme learning machine, Inf. Sci. 185 (1) (2012) 66–77.

[10] R. Clemen, Combining forecasts: a review and annotated bibliography, Int. J. Forecast. 5 (4) (1989) 559–583.

[11] Q. Dai, A competitive ensemble pruning approach based on cross-validation technique, Knowl. Based Syst. 37 (2013) 394–414.

[12] E.R. Davies, Machine Vision: Theory, Algorithms, Practicalities, Elsevier, 2004.

[13] T. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, in: Lecture Notes in Computer Science, 1857, Springer, Berlin, Heidelberg, 2000, pp. 1–15.

[14] S. Ding, H. Zhao, Y. Zhang, X. Xu, R. Nie, Extreme learning machine: algorithm, theory and applications, Artif. Intell. Rev. 44 (1) (2015) 103–115.

[15] J. Durand, Y. Kerr, An improved decorrelation method for the efficient display of multispectral data, IEEE Trans. Geosci. Remote Sens. 27 (5) (1989) 611–619, doi:10.1109/TGRS.1989.35944.

[16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, 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 Recognit. 44 (8) (2011) 1761–1776.

[17] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, DRCW-OVO: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems, Pattern Recognit. 48 (1) (2015) 28–42.

[18] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.

[19] K. Jackowski, B. Krawczyk, 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] R.A. Jacobs, Methods for combining experts' probability assessments, Neural Comput. 7 (5) (1995) 867–888.

[21] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37, doi:10.1109/34.824819.

[22] L.C. Jain, M. Seera, C.P. Lim, P. Balasubramaniam, A review of online learning in supervised neural networks, Neural Comput. Appl. 25 (3–4) (2014) 491–509.

[23] A.H. Ko, R. Sabourin, L.E.S. de Oliveira, A. de Souza Britto Jr., The implication of data diversity for a classifier-free ensemble selection in random subspaces, in: Proceedings of the Ninteenthth International Conference on Pattern Recognition, Tampa, Florida, USA, 2008, pp. 1–5.

[24] B. Krawczyk, P. Ksieniewicz, M. Woźniak, Hyperspectral image analysis based on color channels and ensemble classifier, in: Proceedings of the Ninth International Conference on Hybrid Artificial Intelligence Systems, in: HAIS 2014, 8480, Springer-Verlag New York, Inc., New York, NY, USA, 2014, pp. 274–284.

[25] L. Kuncheva, L. Jain, Designing classifier fusion systems by genetic algorithms, IEEE Trans. Evolut. Comput. 4 (4) (2000) 327–336.

[26] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 2004.

[27] J. Li, J.M. Bioucas-Dias, A. Plaza, Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 844–856.

[28] W. Li, C. Chen, H. Su, Q. Du, Local binary patterns and extreme learning machine for hyperspectral imagery classification, IEEE Trans. Geosci. Remote Sens. 53 (7) (2015) 3681–3693.

[29] N. Liu, H. Wang, Ensemble based extreme learning machine, IEEE Signal Process. Lett. 17 (8) (2010) 754–757.

[30] G.L. Marcialis and F.Roli. Fusion of face recognition algorithms for video-based surveillance systems. G.L. Foresti, C. Regazzoni, P. Varshney Eds, 235--250, 2003.

[31] R. Moreno, F. Corona, A. Lendasse, M. Graña, L.S. Galvão, Extreme learning machines for soybean classification in remote sensing hyperspectral images, Neurocomputing 128 (2014) 207–216.

[32] Y. Pao, G.H. Park, D.J. Sobajic, Learning and generalization characteristics of the random vector functional-link net, Neurocomputing 6 (2) (1994) 163–180.

[33] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 629–639, doi:10.1109/34.56205.

[34] J. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, 1993.

[35] S. Rajan, J. Ghosh, M.M. Crawford, An active learning approach to hyperspectral data classification., IEEE Trans. Geosci. Remote Sens. 46 (4) (2008) 1231–1242.

[36] W. Schmidt, M. Kraaijveld, R. Duin, Feedforward neural networks with random weights, in: Proceedings of the Eleventh IAPR International Conference on Pattern Recognition, II, 1992, pp. 1–4. Conference B: Pattern Recognition Methodology and Systems, Proceedings

[37] G. SVAETICHIN, Spectral response curves from single cones, Acta Physiol. Scand. 39 (134) (1956) 17–46. PMID: 13444020

[38] J. Tyo, A. Konsolakis, D. Diersen, R. Olsen, Principal-components-based display strategy for spectral imagery, IEEE Trans. Geosci. Remote Sens. 41 (3) (2003) 708–718, doi:10.1109/TGRS.2003.808879.

[39] M. van Erp, L. Vuurpijl, L. Schomaker, An overview and comparison of voting methods for pattern recognition, in: Proceedings. Eighth International Workshop on Frontiers in Handwriting Recognition, 2002, pp. 195–200.

[40] T. Wilk, M. Woźniak, Complexity and multithreaded implementation analysis of one class-classifiers fuzzy combiner, in: E. Corchado, M. Kurzynski, M. Wozniak (Eds.), Hybrid Artificial Intelligent Systems, Lecture Notes in Computer Science, 6679, Springer, Berlin, Heidelberg, 2011, pp. 237–244.

[41] D.H. Wolpert, The supervised learning no-free-lunch theorems, in: Proceedings of the Sixth Online World Conference on Soft Computing in Industrial Applications, 2001, pp. 25–42.

[42] M. Woźniak, M. Graña, E. Corchado, A survey of multiple classifier systems as hybrid systems, Inf. Fus. 16 (2014) 3–17.

[43] J. Xia, M.D. Mura, J. Chanussot, P. Du, X. He, Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles, IEEE Trans. Geosci. Remote Sens. 53 (9) (2015) 4768–4786.

Paweł Ksieniewicz is a research assistant at Wroclaw University of Technology, where he achieved M.Sc. degree in 2013 and Ph.D. degree in 2017. 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.

Bartosz Krawczyk is an assistant professor in the Department of Computer Science, Virginia Commonwealth University, Richmond VA, USA, where he heads the Machine Learning and Stream Mining Lab. He obtained his MSc and PhD degrees from Wroclaw University of Science and Technology, Wroclaw, Poland, in 2012 and 2015 respectively. His research is focused on machine learning, data streams, ensemble learning, class imbalance, one-class classifiers, and interdisciplinary applications of these methods. He has authored 35+ international journal papers and 80+ contributions to conferences. Dr Krawczyk was awarded with numerous prestigious awards for his scientific achievements like IEEE Richard Merwin Scholarship and IEEE Outstanding Leadership Award among others. He served as a Guest Editor in four journal special issues and as a chair of ten special session and workshops. He is a member of Program Committee for over 40 international conferences and a reviewer for 30 journals.

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