EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Classification of hyperspectral imagery with neural networks: comparison to conventional tools

Erzsébet Merényi[1,2*], William H Farrand[3], James V Taranik[4^] and Timothy B Minor[4]

Abstract

Efficient exploitation of hyperspectral imagery is of great importance in remote sensing. Artificial intelligence approaches have been receiving favorable reviews for classification of hyperspectral data because the complexity of such data challenges the limitations of many conventional methods. Artificial neural networks (ANNs) were shown to outperform traditional classifiers in many situations. However, studies that use the full spectral dimensionality of hyperspectral images to classify a large number of surface covers are scarce if non-existent. We advocate the need for methods that can handle the full dimensionality and a large number of classes to retain the discovery potential and the ability to discriminate classes with subtle spectral differences. We demonstrate that such a method exists in the family of ANNs. We compare the maximum likelihood, Mahalonobis distance, minimum distance, spectral angle mapper, and a hybrid ANN classifier for real hyperspectral AVIRIS data, using the full spectral resolution to map 23 cover types and using a small training set. Rigorous evaluation of the classification accuracies shows that the ANN outperforms the other methods and achieves ≈90% accuracy on test data.

Keywords: Classification; Hyperspectral imagery; Neural networks; High-dimensional data

1 Introduction

High spatial and spectral resolution images from advanced remote sensors such as NASA's AVIRIS (e.g., [1]), Hyperion, HyMap, HYDICE [2], and others provide abundant information for the understanding and monitoring of the Earth. At the same time, they produce data of unprecedented volume and complexity. Unraveling important processes such as the evolution of the solid earth, global cycling of energy, oxygen, water, etc., the responses of the biosphere to disturbances, and others mandates the best possible exploitation of the data. The challenge is to develop methods that are powerful enough to make use of the intricate details in hyperspectral data and are fast, robust, noise tolerant, and adaptive. While the growing number of spectral channels enables discrimination among a large number of cover classes, many conventional techniques fail on these data because of mathematical or practical limitations. For example, the maximum likelihood and other covariance-based classifiers require, on the minimum, as many training samples per class as the number of bands plus one, which creates a severe problem of field sampling for AVIRIS 224-channel data with many classes. Dimensionality reduction is frequently accepted to accomodate data for traditional methods, but this can result in an undesirable loss of information. Covariance-based methods, in particular, often fail to detect subtle but discriminating features in the spectra even when enough training samples are available, because they are limited to working with first and second order statistics, while hyperspectral imagery is typically far from being Gaussian.

The use of artificial neural networks (ANNs) for complex classification tasks is motivated by their power in pattern recognition. For a review, see, e.g., [3]. Many earlier works documented ANN capabilities for remote sensing spectra on relatively modest scales: few (5 to 12) classes, low-to-moderate number of channels (e.g.,

*Correspondence: erzsebet@rice.edu
^Deceased
[1]Department of Statistics, Rice University, 6100 Main Street MS-138, Houston, TX 77005, USA
[2]Department of Electrical and Computer Engineering, Rice University, 6100 Main Street, Houston, TX 77005, USA
Full list of author information is available at the end of the article

[4-17]). Several studies for higher spectral resolution (e.g., 60 channels in [18,19]) used synthetic data which often favor a particular (such as maximum likelihood) classifier, by virtue of (Gaussian) data construction. Others offered some principled dimensionality reduction and showed high accuracies with the reduced number of bands for a moderate number of classes (e.g., [20-22]). Some research targeted selected narrow spectral windows of hyperspectral data to classify one specific important spectral feature [23]. A small number of ANN works classified hyperspectral data directly, without prior dimensionality reduction [24-26]. Experience suggests that the difference in quality between the performance of classical methods and ANN classifiers increases in favor of the ANNs with increasing number of channels. However, this has not yet been quantified for *large-scale* classification of many cover types with subtle differences in complex, noisy hyperspectral patterns. Assessment of ANN performance versus conventional methods for realistic, advanced remote sensing situations requires comparisons using the full spectral resolution of real hyperspectral data with many cover classes because conventional techniques are most likely to reach their limitations in such circumstances. Systematic evaluation is needed to ensure powerful, reliable, automated applications of ANNs or any other classifiers. The present paper is a step toward filling this gap.

We are comparing popular and easily accessible standard classifiers with a neural network paradigm. One aspect that we want to demonstrate in particular is that by using all (or nearly all of the 224) AVIRIS bands more, geologically meaningful spectral variations can be detected than from the same AVIRIS cube reduced to 30 to 40 or less bands; that hyperspectral imagery is highly complex and detailed surface cover information can be extracted with sensitive enough methods.

Another point we wish to highlight is that a sophisticated ANN paradigm can perform well with a small training set, which is always a concern for remote sensing tasks. There have been studies to mitigate the effect of a small training set [27,28] by iteratively labeling unlabeled data with the classifier under training and adding newly labeled samples to the training set. These studies, however, were done mostly on synthetic data or low-dimensional real data [29,30], and the relative benefits decreased with increasing dimensionality. While these methods are very interesting and statistically well founded, they often favor particularly distributed (Gaussian) data and need prior probabilities, and it is unclear how well they would do on the full spectral resolution of real hyperspectral data.

The methods and analysis presented here provide a quantitative comparison between ANN and traditional covariance-based classifers using an AVIRIS data set. The data and classification algorithms utilized in this study are described, and analysis and results of the comparisons are presented, followed by a discussion of outstanding issues and future directions.

2 Study area, data, and preprocessing

2.1 The geologic area and data

The Lunar Crater Volcanic Field (LCVF) was the primary focus of the NASA-sponsored Geologic Remote Sensing Field Experiment (GRSFE) conducted in the summer of 1989 [31]. Since 1992, the large playa in the LCVF, Lunar Lake along with the surrounding terrain, has been one of the several standard sites used as a calibration location by the AVIRIS team and imaged yearly by AVIRIS. We selected this site because it has been studied extensively and independently by other workers, and because one of the authors (WHF) has directly been involved in field measurements and field mapping of cover types through GRSFE and other projects [32-34]. Figure 1 shows a false color composite of the Lunar Lake area analyzed in this paper, with locations representative of various cover types marked by their respective class labels used in this study. The full list of classes is given in Table 1. The data considered here are a 614 samples by 420 lines subsection of the image collected by AVIRIS on April 5, 1994 at 18:22 GMT. The LCVF, which lies roughly halfway between the towns of Ely and Tonopah in northern Nye County, Nevada consists of over 100 square miles of Quaternary basaltic pyroclastic and flow deposits [35]. These deposits lie atop ignimbrites and silicic lava flows of Tertiary age. The basaltic volcanics are in turn overlain by Quaternary alluvial and playa deposits. Also included with the analyzed subsection are the Lunar Lake playa and outcrops of the Rhyolite of Big Sand Spring Valley (label B) mapped by Ekren [36]. Vegetation within the LCVF is sparse, but locally abundant within washes (label C) and atop the plateau (J) that makes up the lower left part of the scene, bordered by 'The Wall,' a prominent NE-SW trending scarp straddled by the label G in Figure 1.

The reflectance signatures of surface materials within the LCVF have variations that range from subtle to significant. Oxidized basaltic cinders (label A) are associated with many of the cinder cones in the LCVF. These cinders are rich in hematite and thus have the prominent absorption band at 0.86 μm caused by crystal field effects and also the diagnostic UV-visible absorption edge attributable to the Fe^{3+} - O^{2-} charge transfer absorption centered in the UV. Hematite also has a high reflectance in the near IR, and these oxidized cinders show up as bright aprons (classes L, W) about the cinder cones in the longer wavelength AVIRIS channels. The Rhyolite of Big Sand Spring Valley that is exposed in the lower left portion of the subsection of the AVIRIS image (label B) contains enough iron so that it too displays the Fe^{3+} - O^{2-} charge transfer edge. It also displays a 2.2-μm absorption

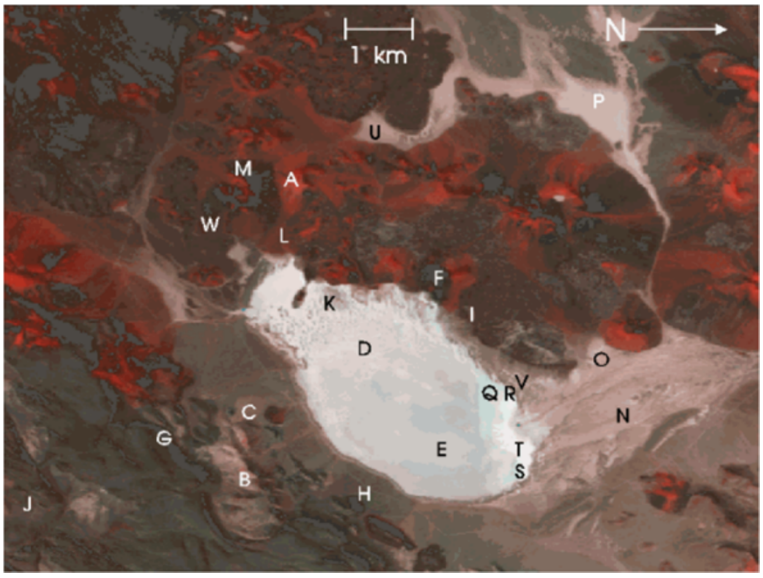

Figure 1 False color composite of the Lunar Crater Volcanic Field (LCVF) site. Letters stand for the various cover classes used throughout this paper and mark some of the training locations of distinguishing geologic features for the corresponding classes.

Table 1 The cover types in the Lunar Crater Volcanic Field site

Class	Cover type description	# tr
A	Hematite-rich cinders	72
B	Rhyolite of Big Sand Spring Valley	22
C	Alluvium #1	50
D	Dry playa	160
E	Wet playa #1	115
F	Young basalt	21
G	Shingle Pass Tuff	14
H	Alluvium #2 (with mixed scrub brush, rocks, and soil)	50
I	Old basalt	36
J	Dense scrub brush stands	14
K	Basalt cobbles on playa	37
L	Ejecta blankets #1 (mixed hematite-rich and unoxidized cinders)	78
M	Alluvium #3 (iron rich)	14
N	Dry wash #1	15
O	Dry wash #2	54
P	Dry wash #3	45
Q	Wet playa #2	15
R	Wet playa #3	14
S	Wet playa #4	15
T	Wet playa #5	18
U	Alluvium #4 (also iron rich)	36
V	Wet playa #6	14
W	Ejecta blankets #2 (primarily unoxidized cinders with smaller percentage of hematite-rich cinders)	33
	Total number of training samples	942

Cover types with class labels used in this study and with the number of original training samples (# tr) identified for each class.

feature indicative of the incipient development of dioctahedral clay minerals. Lunar Lake, which at first glance might appear to be compositionally homogenous, in fact displays several spectrally distinct surface units. These surface cover units ('wet playa' classes E, Q, R, S, T, V) are distinguished primarily on the basis of their clay content and on the basis of their adsorbed and, perhaps, structurally bound hydroxyl and molecular water content. (Higher water content means deeper absorption features at approximately 1.4 and 1.9 μm and a consequent depression of the spectral continuum at longer wavelengths.) Many of the alluvial, or 'dry wash', units (D, N, O, P) are distinguished in a similar fashion by subtle variations in the spectral continuum caused by clay and water content.

Twenty-three known, different geologic units were chosen for this study based on field knowledge, geologic meaning, and spectral properties. The pattern recognition challenge posed by the spectral variations across these 23 classes is illustrated in Figure 2.

2.2 Data preprocessing

The LCVF image was atmospherically corrected and converted to reflectance units, using the empirical line method (e.g., [34,37]), which produced spectra with fewer noise artifacts for this 1994 image than ATREM [38]. After exclusion of excessively noisy channels, as well as duplicates among overlapping channels at the detector interfaces, 194 bands remained with excellent signal-to-noise ratio [39]. A brightness normalization such as that described in [40] (also called the hyperspherical directional cosine transformation [41]) was also applied

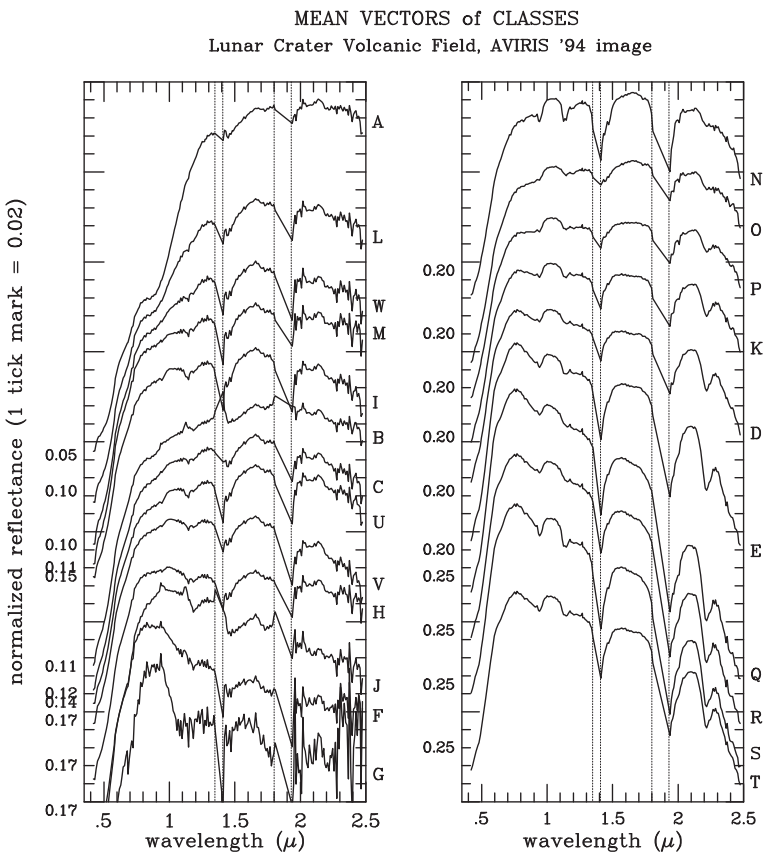

Figure 2 Representative average spectra of the 23 LCVF classes listed in Table 1, vertically offset for clarity. Many signatures (such as the clay-bearing series on the right) have subtle variations; others (such as the iron oxide-bearing species, A, L, W, F, G, on the left) have larger differences. The vertical dotted lines near 1.4 and 1.9 μm indicate data fallout where saturated bands in the water vapor windows were eliminated, after atmospheric correction.

in order to eliminate linear illumination geometry effects. This normalization divides all data vectors by their Euclidean norm, producing unit vector length while preserving the spectral angle relations of the bands. Unfortunately, geometric albedo (any linear effect) is also eliminated in this process; therefore, one may need to separate classes that are spectrally the same but distinguished by albedo, in a post-processing step (as in, e.g., [24]). Fortunately, this is not a frequent situation, in our experience. For the present analysis, the advantages of this brightness normalization outweighed the disadvantages in that the separation among spectral groups increased (due to the enhanced spectral contrast between different species) more than differences were masked by the loss of albedo variations. Classes distinguished only by albedo were not present among the LCVF units.

3 Classifiers and methodology for comparison

3.1 The ANN paradigm and the competing classifiers

Back propagation (BP) neural networks, which are perhaps the most popular and best known among ANN paradigms, can be difficult to train with high-dimensional data as their complexity increases non-linearly with the number of input dimensions and the possibility for the gradient descent learning to get stuck in local minima increases dramatically. Dimensionality reduction prior to classification is frequently applied to high spectral resolution data to achieve tolerable training time, or training convergence at all with a BP network (e.g., [14,20]), or to apply other methods.

To make use of the full spectral resolution, we used a hybrid ANN architecture, the details of which are given in [24,25]. Briefly, it consists of an input layer with as many nodes as the number of spectral bands plus one 'bias' neuron, a two-dimensional self-organizing map (SOM) [42] as the hidden layer, and the SOM layer is fully connected to a categorization learning output layer. Inputs to the output layer are the responses of all SOM neurons with the three largest responses normalized to sum to one, the rest set to zero, before passing them to the output layer. The output layer has one node for each class, i.e., the 1-in-c encoding is used for class labels. (The cth output is expected

to produce 1, the others to produce 0, for the *c*th class.) It learns with a Widrow-Hoff (delta) learning rule [3,43] and uses a linear activation function. This hybrid network learns in two phases. First, in an unsupervised regime, it builds its own view of the manifold structure by forming a topology-preserving feature map of the data (clusters, if they exist) in the hidden SOM layer (while the output layer remains idle). In a subsequent supervised learning phase, the weights between the SOM hidden layer and the output layer are trained to recognize class labels. The pre-formed clusters - the model of the data manifold - in the SOM help prevent the learning of inconsistent labels and thus greatly support accurate learning of class labels in the supervised phase. This results in better generalization from a small number of samples and leading to higher classification accuracy, than without the SOM stage. Back propagation, in contrast, is powerful enough to simply 'memorize' inconsistent labeling if the number of training samples is small. For example, the network can learn to assign labels A and B to two individual training samples even if their characteristics are very similar (for example, B is a mislabeled sample from class A). In this case, no reasonable prediction can be expected because the network does not derive general class properties. This situation can be avoided with the hybrid ANN paradigm we described. It is also much faster to train the supervised output layer of this network than to train a BP network, since the output layer only learns the labeling of the classes (based on the cluster boundaries internally identified by the SOM). The delta learning rule with linear activation function is much simpler than back propagation, which helps relatively easy training even with very high dimensional data, in this SOM-hybrid architecture. The investment of training the SOM layer has the additional benefit that it can be reused in different supervised training sessions, for example, to train for various sets of classes, since the cluster structure of the manifold is the same regardless of how many classes are labeled. The good scale-up properties and high classification accuracies of this network have been demonstrated by previous hyperspectral analyses [24-26,44] for up to 200 spectral channels and 20 to 30 classes.

In a recent paper, Foody and Cutler [9] apply a similar concept: they examine the data structure through SOM clustering and manually evaluate how well the clusters correspond to known cover classes, in order to assess the potential of the particular data for the discrimination of the known classes. The SOM-hybrid network that we use helps accomplish the same, in an implicit and integrated fashion. Misclassified labeled training samples (assuming that the SOM learned correctly, the overall network learned well, and that the training samples were labeled consistently) will alert the analyst to discrimination problems. Conversely, in the case of labeling uncertainties (for example, at the boundaries of similar materials or in the case of data that was labeled on the basis of some other attributes than the given data contains), such misclassified labeled data can guide a revision of the labeling. (An example of this is in [24].)

The quality of SOM learning (including topology preservation, completion of ordering, optimal placement of quantization prototypes (the SOM weights) in the data space, and convergence) is important for ensuring good results. For discussion of related issues, which are beyond the scope of this paper, we refer to [45-48] and the references therein. We mention here that we used, instead of the basic Kohonen SOM [42], a variant called *conscience learning* [49], which encourages all SOM neurons to win with equal frequency through a biasing 'conscience' and thereby maximizes information theoretical entropy of the mapping. This leads to the best possible representation of the data distribution with the given number of quantization prototypes and thus facilitates the most faithful learning of the cluster structure [47]. An additional benefit of conscience learning is that it only needs to update the immediate SOM neighbors, which makes it computationally efficient. Even though we did not use the extracted clusters for establishing labeled classes in this work (we used the determination of a domain expert for class designations), we know that the SOM in this study learned the cluster structure of the LCVF data extremely well. This was demonstrated by another study where the clusters extracted from the SOM showed striking correspondence with the supervised classes [44,50].

One important feature of this ANN is that the class predictions are characterized by a membership strength, and below a predefined threshold of the membership strength, the data sample is labeled 'unclassified.' In addition, we can record the membership strengths that each output node predicts on the 0 to 1 scale, which can be used for assessing the confidence in the class predictions.

This SOM-hybrid ANN was built and tested in NeuralWare's NeuralWorks Professional II/PLus [51], then deployed using NeuralWare's Designer Pack, and embedded in our own software environment that has specifically been developed for the exploitation of high-dimensional data such as large hyperspectral images. Our own algorithm research and data analysis environment builds on NeuralWare and Khoros [52] functions and extends standard neural network capabilities.

The established classifiers that we compared with the above SOM-hybrid ANN are maximum likelihood (MLH), Mahalonobis distance (MHD), minimum Euclidean distance (MED), and spectral angle mapper (SAM) by Kruse et al. [53].

These non-ANN classifiers are well documented in remote sensing texts [54-57] and are commonly available in commercial image analysis packages such as ENVI and

others [58-60]. We chose these non-ANN classifiers for this study because of their widespread use and easy accessibility. Two of them (MED and SAM) can process high-dimensional input signatures without prior dimensionality reduction, while MLH and MHD suffer from input dimension limitations when the training set is smaller than the number of spectral channels. This, however, is a remote sensing reality that needs to be considered when selecting classifiers for a task.

3.2 Evaluation criteria

Performance evaluation criteria were derived from the requirements or necessities dictated by real-life tasks:

1. *Classification accuracy*
2. *The capability of using the full spectral resolution*
3. *Dependence on the number of training samples*. This is of special interest in remote sensing as the minimum necessary number of training samples in the case of a covariance-based classifier (such as maximum likelihood) for AVIRIS class data is over 200 per class, a prohibitively large number for a dozen or more classes.

Sensitivity to uneven class representation and to noise are two of several other important issues. While we do not address these systematically in this paper, the experiments we describe involve uneven class representation as well as noisy data. Learning from unevenly represented class samples is another strength of ANNs, compared to parametric classifiers, and it is an advantage in remote sensing since even sampling across cover types is often impossible.

4 Analysis and results

4.1 The classification experiments

Altogether 942 training pixels were originally identified across all classes ranging from as low as 14, to 160 samples for a class, as shown in Table 1. This limited the application of the covariance-based classifiers (MLH and MHD) to a 13-band subsampled version of the data, with the original training set. For training of the MLH and MHD classifiers with 194-band data for 23 classes, a minimum of $(194 + 1) \times 23 = 4,485$ would be required. The other three classifiers were not limited by the number of spectral bands.

Since we also wanted to see if an increasing difference in the quality of performance manifests with the inclusion of increasing number of bands, we created a second augmented training set. We were able to increase the minimum number of training samples for each class to 31, which allowed us to employ the MLH and MHD classifiers on 30-band data. For this augmentation, we carefully hand-selected additional samples based on prior knowledge of the surface cover types and on spectral similarity to the original samples. Details of the band selection approach are given in Section 4.2 below. Further augmentation, to include the MLH and MHD classifiers in the 194-band experiment, was not possible, partly because the known occurrences of some of the classes (such as B, the rhyolitic outcrop) are smaller or much smaller (classes Q, R, S, T, for example) than 195 pixels. The classifications were performed, after preprocessing, for 13-, 30-, and 194-band cases, as applicable. Table 2 shows a summary of classification runs performed on the 1994 LCVF AVIRIS data using the full 194-band normalized AVIRIS data set as well as the spectrally subsampled data sets containing 30 and 13 bands, respectively.

The SOM-hybrid ANN we used for this work had a configuration of 194 input nodes (30 and 13, respectively, for the subsampled cases) plus one bias node, 23 output nodes, and a 40-by-40 two-dimensional rectangular SOM in the hidden layer. The class labels were encoded as 23-element unit vectors, with a 1 at the position of the output neuron corresponding to the given class and zeros elsewhere. The input samples were scaled into the [0,1] range using the global minimum and maximum of the data. This scaling preserves the relative proportions of the values in the different dimensions, in this case the spectral angles across spectral bands. The target output values did not need scaling since each was already in the range {0,1} because of the 1-in-c class label encoding. With normalized SOM outputs and no bounded activation function in the output layer, there was no need to scale the inputs and the target output values into the range of a particular activation function, or to scale them at all. We performed this internal scaling for the convenience of easier tractability of the training. This scaling was done after the preprocessing described in Section 2.2. To allow the SOM to learn the cluster structure of the input data space, 300,000 unsupervised learning steps were performed. In this phase, all image pixels were used (without labels). The 300,000 may appear as a low number of training steps for nearly the same number of data points (614×420); however, many pixels have similar spectral signatures thus each spectral type was shown to the SOM many times. The subsequent supervised training was performed with the training set shown in Table 1 or with the augmented training set for the MLH and MHD 30-band cases. Because of the support from the SOM hidden layer, the supervised training converged very fast. After ≈20,000 steps, with a learning rate decreasing from 0.15 to 0.01, the training

Table 2 Classifications performed in this study and the number of spectral bands used for each run

# bands	Classification runs				
13	ANN	MED	SAM	MLH	MHD
30	ANN	MED	SAM	MLH	MHD
194	ANN	MED	SAM	-	-

accuracy stabilized at 99.9%. In the recall phase, class predictions with larger than 0.1 decision strength (on a scale of approximately 0 to 1) were accepted, leaving pixels with less than 0.1 decision strength on all output nodes unclassified. In the experiments we conducted with the LCVF image, the percentage of unclassified pixels was low, ≈3.45% for the ground truth test pixels (see below) in the best classifications. In addition, the recorded map of decision strengths associated with each image pixel contains very few instances where the class membership assignment had to rely on less than 0.5 decision strength. There are cases where a pixel had assignment into two (or sometimes three) competing classes, with significant decision strengths (for example, 0.6 and 0.4). For the purpose of this study we accepted the strongest class membership in such cases.

One input sample to the ANN consisted of one image pixel (one 194-element spectrum). No spatial context was considered for input, for two reasons. If a $k * k$ window is selected automatically around the current pixel, the input may contain contamination by spectral signatures that do not belong to the given spectral class. In certain circumstances taking input from a window rather than from a single pixel can be helpful and works well. For example, Benediktsson et al. [61] construct feature vectors from morphological attributes of a single image band. However, when one works with high spectral resolution and with many classes, some of which may have subtle discriminating differences such as seen in Figure 2, a window of spatial context may blur class distinctions. This is especially a danger in the case of hyperspectral data that also have high spatial resolution. Additionally, omitting context in the input allows one to use the spatial coherence, or lack thereof, to help judge the resulting classification.

The MED, MLH, MHD, and SAM classification results were generated in the ENVI image processing software. MED classifications were run for the 194-, 30-, and 13-band cases using the following three sets of minimum distance parameters: (1) no maximum standard deviation around the training class means, which classifies all pixels to the closest class; (2) one standard deviation; and (3) two standard deviations from the training sample means for each class. The latter two can leave many pixels unclassified. For the SAM classifications, a threshold of the spectral angle was applied. This threshold specified the spectral angle between an input and a target spectrum, beyond which the input sample remained unclassified. The default value of 0.1 radians was used. MLH and MHD classifications were run for the 30- and 13-band data sets. The original number of training data (Table 1) were used to classify the 13-band AVIRIS image. The 30-band data set was classified using the abovementioned augmented training set. The MLH classifications were run twice for each of the two subsampled data sets: once with prior probabilities and once without (i.e., with default, equal prior probabilities). Prior probabilities were assigned to each land cover class using area weighted estimates from the MED and ANN results of the 194-band data set, as well as from existing geological field knowledge of the study site.

4.2 Selection of bands for covariance-based classifiers

The 13- and 30-band spectral subsamples of the 194-band AVIRIS data set were constructed using a qualitative assessment process by one of us (TBM), a domain expert. For both the 13- and 30-band subsamples, band selection concentrated on diagnostic features in the visible and VNIR (0.45 to 1.0 μm) and the SWIR (2.1 to 2.35 μm). Visible bands were selected to identify iron oxide mineralogy reflectance features, primarily hematite, but also goethite and jarosite. A band near 0.86 μm was also selected for both the 13- and 30-band subsamples to identify the ferric iron absorption feature of hematite. VNIR bands were included to identify the reflectance of vegetation. A total of eight spectral bands between 0.46 and 1.05 μm were selected for the 13-band subsample; 20 bands in the same range were selected for the 30-band subsample. Bands selected from the SWIR portion of the spectra focus on diagnostic features related to clay minerals, micas, and other hydroxyl-bearing minerals, and were centered around the absorption feature at 2.2 μm. Four bands between 2.15 and 2.34 μm were selected for the 13-band subsample. Nine bands in the same range were selected for the 30-band case, sampled slightly different than in the 13-band case to provide a more even distribution of the bands across the region. A single band at 1.62 μm was selected for both the 13- and 30-band subsamples to identify a hydroxyl-bearing reflectance feature which was present in all the playa and wash classes. The 30-band selection included all the bands selected for the 13-band case, with the exception of the abovementioned slight difference for two bands.

Other approaches to band selection that we tried included uniform subsampling and PCA, but neither produced better results than the band selection by the domain expert. In a different study, wavelets [62] also remained unconvincing for the task. We note that non-linear methods such as non-linear PCA (NLPCA, e.g., [63-65]) may do a better job in selecting the most informative bands than linear PCA or wavelets. Non-linear methods that also take into account the classification goal can be especially useful in the case of supervised tasks and may compete with the human expert.

4.3 Classification results

All of the MED, MLH, and MHD classification results were evaluated in terms of overall accuracies and κ-statistics, as detailed in Section 4.4, and in terms of the

largest number of spectral bands used, to determine which variant of the respective algorithm produced the best results within its category. The best variants were then compared with the SAM and ANN classifications. Of the various MED classification runs, the 194-band run, with no maximum standard deviation specified as a distance constraint, provided the best map. When a distance constraint was imposed, the resulting class maps contained too few classified pixels for the map to be useful. Of the four MLH runs, the 13-band run with no prior probabilities had the highest accuracy. There was little difference between the MLH runs with and without prior probabilities, probably due to the relatively large number of classes and the resultant small probability values. Figure 3 presents a comparison of the best class maps produced by four of the classifiers for the highest applicable number of bands: The ANN and SAM 194-band maps, the 194-band MED map with 0 standard deviation as distance threshold, and the 30-band MLH map, computed without prior probabilities. The 13- and 30-band MLH maps were visually very similar in their tendencies of the misclassifications, in spite of the higher (albeit still quite low) accuracy of the 13-band MLH map. The observations we make based on the 30-band MLH classification in Figure 3 are generally valid for the 13-band MLH map too. The best MHD classification produced the least interesting differences with any of the others; therefore, it was not included in Figure 3, for space considerations. It is easy to see by visual inspection that there are obvious differences among these class maps. Comparison of the classification maps to each other and to the color composite of the site (Figure 1) reveals that the ANN and the MED produced much more detailed class maps than the MLH, and that they are also more detailed than the SAM map, although the differences with the SAM map are more subtle. One example is the almost complete omission of class B (white, rhyolitic

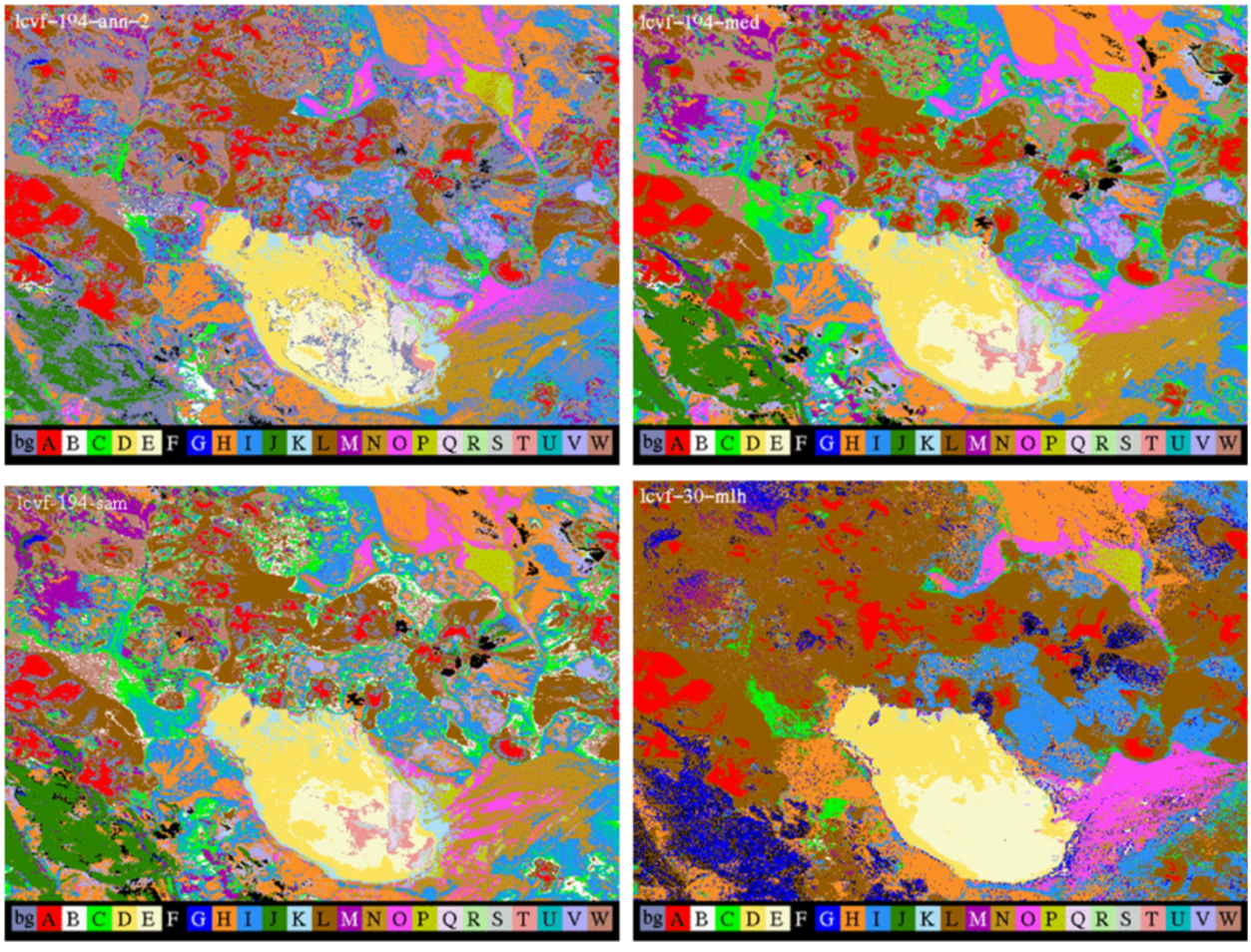

Figure 3 Comparison of classification maps obtained with four of the classifiers. Clockwise from top left: ANN class map from 194-band data, MED map from 194-band data, MLH map from 30-band data, and SAM map from 194-band data. The class labels are resolved in Table 1, and spectral statistics are shown in Figures 4 and 5, for the ANN and MLH cases. The label 'bg' stands for unclassified pixels (background color). The ANN and SAM class maps contain a considerable number of unclassified pixels, while the other class maps do not.

outcrop) in the SAM map, another is the poor delineation of the Shingle Pass Tuff unit (class G), as class F. ANN and MED are, in addition, very similar to one another, which is a strong support for these maps to be more accurate than MLH. Detailed field knowledge [33] as well as previous analyses of this scene by various authors [15,32,33,66,67] also corroborate these observations.

One important point is that the ANN and SAM maps contain unclassified pixels. In contrast, the other classifiers assigned a class label to all pixels in the cases shown. Unclassified pixels in coherent patches may indicate a potential new class. Along the borders of two cover types, it may suggest that those two classes were not represented to the full extent by the training samples. This can be determined by examination of the spectra at such unclassified locations. Although it looks esthetically more pleasing, the MED map is not more accurate than the ANN map, as shown later, and does not leave spectral units to be discovered.

Large areas are dominated by the L class in the MLH map where the MED and ANN classifications display considerable variability in accordance with the color site composite. Class G also seems unreasonably extensive for the cover type, Shingle Pass Tuff, which occurs along The Wall, a NW-SE trending scarp that represents the remaining trace of the Lunar Lake caldera [36]. This scarp spans across the label G in Figure 1 and is more accurately traced by the ANN map. F (young basalt) is another class overestimated by the MLH. Several classes are almost entirely missing from the MLH map. Of Q, R, S, and T, only the rectangular training areas are classified. Class N appears at a few miniscule spots and O overwhelms the wash area where both the ANN and MED classifiers predicted N. As seen from Figure 2, which displays the mean training spectrum for each spectral type, the Q, R, S, and T classes have very fine distinctions among themselves. The subtle differences mainly occur between the 0.9 to 1.2, 1.4 to 1.6, and 1.95 to 2.2 μm windows, which may remain less resolved with the 13- and 30-band selections than with the full (194-band) resolution, as seen in Figures 4 and 5. However, we point out here that the 30-band and even the 13-band cases of MED and ANN resolved more classes than either of the MLH cases, including clear distinction among the classes Q, R, S, and T as well as mapping N and O more similarly to that of the 194-band cases. Class J, which is abundant in the lower left corner of the ANN

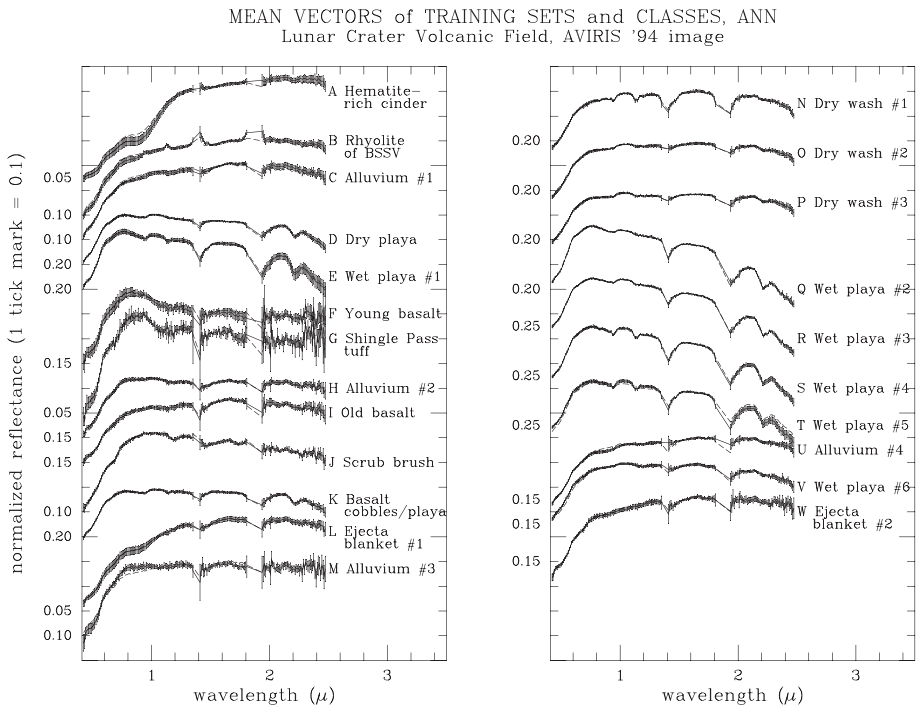

Figure 4 Mean spectra of training samples and mean of the pixels classified by the SOM-hybrid ANN. The mean spectra of the training samples (solid lines) for each class and the mean of the pixels classified by the SOM-hybrid ANN into the respective classes (dashed lines), using the full, 194-band image cube. The vertical bars show the 1 standard deviation of the training samples. The mean of the predicted classes is well within this standard deviation for almost all classes. For many - such as the clay-rich series on the right - the training and class means are virtually indistinguishable (and therefore the dashed line of the class mean may not be easy to see), indicating a good match between the known training representatives and the predicted members of the classes. The bare line segments centered at 1.4 and 1.9 μm indicate data fallout regions in the atmospheric water vapor windows.

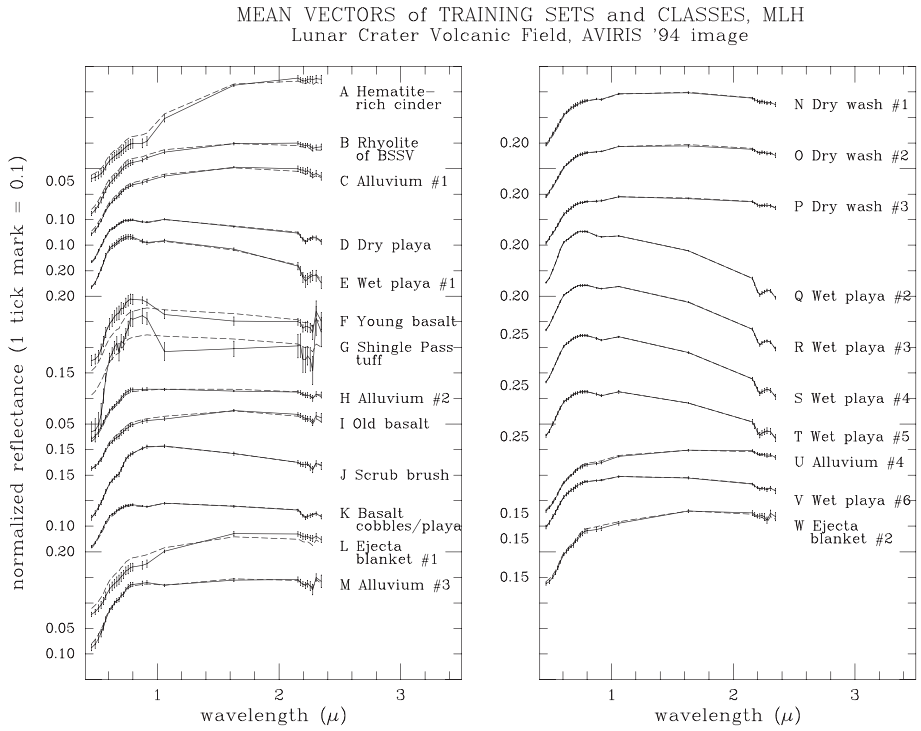

Figure 5 Mean spectra of training samples and mean of the pixels classified by the MLH classifier. The mean spectra of the training samples (solid lines) for each class, and the mean of the pixels classified by the MLH classifier into the respective classes (dashed lines), using the 30-band subsampled image cube. The vertical bars show the 1 standard deviation of the training samples. The mean of the predicted classes departs considerably from the mean of the training samples for a number of classes on the left, indicating a poor match between the known training representatives and the predicted members of the classes. In contrast, the match is very good for the classes on the right: the training and class means are virtually indistinguishable (and therefore the dashed line of the class mean may not be easy to see). This, however, does not mean excellent classification for all classes here, because a number of them have barely more than the training pixels classified into them.

and MED maps is also missing. Further visual comparison of these class maps is left to the reader.

For convenience of visual comparison with the class maps, we also show here a ground truth image (Figure 6) that will be described, and referred to later, in Section 4.4.

The spectral plots in Figures 4 and 5 show training and class statistics, for the 194-band ANN and for the 30-band MLH classification, respectively. The averages of the training spectra (solid lines) are overlain with the averages of the predicted classes (dashed lines). Large deviations of the means, especially when the general shape of the class mean shows a different characteristics, indicate poor pattern recognition. Class G is an example of this in Figure 5. One standard deviation of the training data is also plotted (vertical bars) for each class to show the spread of the training set. The training sets of the various classes are inherently different in their 'tightness' because some materials such as the cinder cones (class A) may be represented by a more broadly varying spectral set than others (e.g., the playa and wash units). These and additional statistics (for example, overlaying also the training and class envelopes), summarized in similar plots, provide a quick and easy semi-quantitative partial assessment of

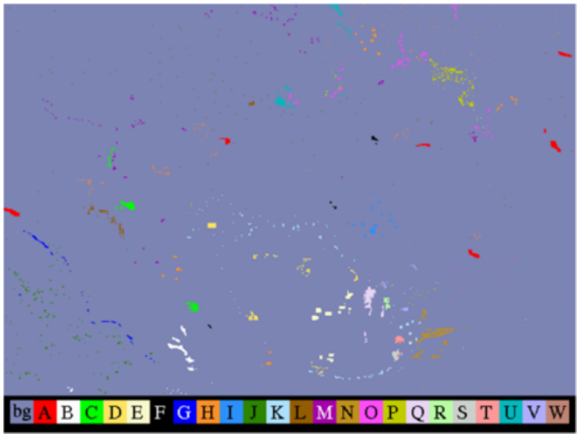

Figure 6 The ground truth image. Containing at least 100 test pixels for each of the 23 classes, with the exception of those classes that consisted of less than 100 samples. A total of 4,332 test pixels were collected, using stratified random sampling.

classification accuracies. It can alert the analyst to poor performance without having to do a full κ-statistic. More importantly, if the number of test samples used in confusion matrices is small, the κ-statistic may not reflect the effect of many misclassified pixels (commission errors). In contrast, in plots like Figures 4 and 5, the statistics include all pixels classified into any class. We note, however, that a tight match of the training and class means does not necessarily mean excellent classification, because this representation does not include omission errors. For example, in the right panel of Figure 5, all classes exhibit very precise match of the means, however, the statistics for a number of those classes (notably Q, R, S, T) includes barely more than the training samples. This can be seen from the class map in Figure 3 as well. In contrast, the ANN classification has many pixels in all of these, as well as other, classes and still exhibits a precise match between training and class means. In this case (and similarly for the MED), one can be more confident in the overall high quality of the classification.

Some insight is provided into the generalization capabilities of the MLH classifier in comparison to ANN and MED (Figure 7). We show three examples of pattern mismatch between the MLH prediction and the true class, chosen from many similar cases observed. The top plots in each panel show how the sample pattern (thick line) fits into the envelope of the training samples of the class into which both ANN and MED classified them - apparently correctly. The bottom plot shows how the same sample fits with the envelope of the class into which MLH misclassified the pattern. The patterns in each case are not simply misclassified by some small difference, but they follow poorly the general shape of the class predicted by MLH.

In the next section, rigorous accuracy evaluation is given, through confusion matrices and κ-statistics, for the best classifications in each category, including the ones in Figure 3. While we are aware of potential uncertainties of the κ-statistic, as well as alternative methods of accuracy comparisons (e.g., [68-71]), we provide the κ-statistic (in addition to confusion matrices and simple overall accuracy) because of its current use in common remote sensing applications.

4.4 Assessment of classification accuracies

The evaluation of classification accuracies followed that outlined in several standard texts on the subject (e.g., [72-75]). Statistical estimation of classification accuracy has been a long studied and established subject that has a vast literature. However, most theoretical considerations

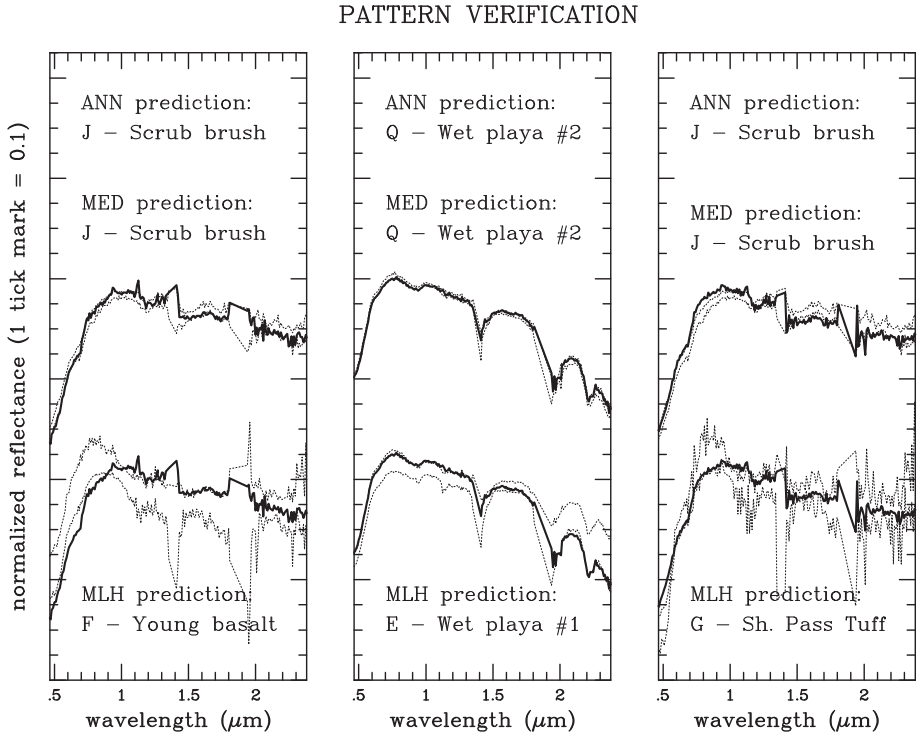

Figure 7 Typical misclassifications. Examples of typical misclassifications by the MLH classifier, from left to right, for a class J, a class Q, and another class J sample. In each panel, the top plots show how the correctly classified sample pattern (thick line) fits into the envelope of the training samples (thin dotted lines) of the corresponding class. The bottom plot shows how the same sample fits with the envelope of the class into which MLH misclassified the pattern.

for sample sizes and other factors for the assessment of map accuracy were developed and verified on low-to-moderate dimensionality data (e.g., Landsat TM, MSS, SPOT HRV) and allowing relatively large errors. With hyperspectral data, the map accuracy is expected to increase; therefore, the sample size required for rigorous assessment of the accuracy within a meaningful error limit and confidence level may become prohibitively large. The works cited above, and others in the literature, offer recommendations for accuracy assessments. According to the formula derived from binomial distribution [72], the number of test samples needed for map accuracy assessment is $\text{ntest} = x^2 \times p \times (1-p) / E^2$, where E is the allowed maximum error in the accuracy assessment, x defines the confidence level (confidence level corresponding to x 'sigma'), and p is the desired map accuracy. As an example, for assessment of the classification at the 95% confidence level within 4% error, this requires at least 2,700 test samples for the 23 classes in this study.

Computation of test sample size based on binomial formulation has been criticized as inadequate for assessment of the confusion among a large number (more than a dozen) of classes and for a large number of image pixels (e.g., [75]). The various studies seem to agree, however, in their conclusion that in such cases (as is also our present study), a minimum of 75 to 100 test pixels per class are necessary for a statistically significant accuracy assessment. The literature also strongly recommends stratified random sampling for identifying the test samples. This is especially important for our study as the LCVF image contains a number of very small but geologically interesting classes. (The rhyolitic outcrop, class B, or the Shingle Pass Tuff unit, class G, are good examples.)

A ground truth image that meets these requirements was painstakingly constructed and used for accuracy assessment in this investigation. It contains at least 100 test pixels for each class, altogether 4,332 samples (individually verified by one of us (WHF) with extensive field

Table 3 Classification accuracies for the ANN classifications

	194-band		30-band		13-band	
	User	Producer	User	Producer	User	Producer
Class	Accuracy		Accuracy		Accuracy	
A	98.36	96.45	97.69	95.48	98.23	89.35
B	95.33	71.50	67.95	53.00	88.29	49.00
C	95.05	86.50	79.65	90.00	77.78	87.50
D	92.49	93.81	81.63	95.24	81.67	93.33
E	93.88	87.62	92.68	90.48	94.82	87.14
F	40.48	85.00	41.98	85.00	39.00	97.50
G	100.00	64.74	87.72	57.80	100.00	64.16
H	96.92	100.00	97.44	69.09	95.24	90.91
I	50.81	63.00	79.05	83.00	48.08	75.00
J	100.00	96.00	99.04	82.40	100.00	94.40
K	93.25	94.00	97.11	67.20	95.24	64.00
L	97.66	97.66	85.46	90.65	94.81	93.93
M	91.67	82.50	84.21	56.00	93.20	48.00
N	88.32	98.33	70.37	19.00	58.33	30.33
O	94.90	96.13	62.68	86.13	80.86	66.77
P	97.25	82.67	77.52	66.67	70.03	92.67
Q	98.60	96.36	95.67	90.45	97.20	94.55
R	88.46	92.00	95.45	84.00	100.00	78.00
S	90.57	80.00	89.09	81.67	78.67	98.33
T	90.91	92.31	92.65	96.92	91.80	86.15
U	77.78	73.50	73.50	86.00	75.86	66.00
V	100.00	62.00	43.21	70.00	23.27	74.00
W	94.36	92.00	85.29	72.50	79.49	77.50
	Accuracy (%) = 88.71		Accuracy (%) = 75.02		Accuracy (%) = 76.39	
	$\kappa = 0.8811$		$\kappa = 0.7380$		$\kappa = 0.7522$	

knowledge of the test area). Exceptions are only those classes that consist of less than 100 pixels. The ground truth image was created in the ENVI software by first selecting regions of interest (ROIs) that represented the best examples of the classes, in order to achieve a stratification for sampling. Application of stratified random sampling is non-trivial because it is hard to know in advance where all the classes are. For that purpose, a mask of the cover types was created from the ANN and MED classifications to provide the above ROIs, and test pixels were randomly selected from each of these class ROIs. The randomly picked pixels were then examined and were selected to be used in the ground truth image only if their reflectance spectra matched what was expected for those surface cover classes and if the locations of the pixels accorded with what was known for the site from one of the authors' (WHF) knowledge of the field site. The spatial distribution of the resulting test pixels is shown in Figure 6.

Tables 3,4,5,6,7 present summaries of user and producer accuracies from the costumary confusion matrices and κ-statistics for each classification. These were computed using the ENVI software. The overall accuracies are summarized in Table 8. The numbers in Table 8 support the visual and semi-quantitative evaluations that we made above.

In Table 8, an increase in the difference of accuracy with growing number of bands can be seen in favor of the ANN, compared to the runner-up MED. While in the 13-band and 30-band cases, only a 2% to 3% difference shows between the ANN and the MED; for the 194-band case, the difference is a more impressive $\approx$7%. The comparison with the SAM does not show this trend; however, the SAM remains $\approx$9.5% below the accuracy of the ANN, for all cases. The increase in accuracy between the 13- and 194-band versions of the individual classifiers is greater than 12% for the ANN, less than 9% for the MED, and greater than 12% for the SAM. This underlines

Table 4 Classification accuracies for the MED classifications

	194-band		30-band		13-band	
	User	Producer	User	Producer	User	Producer
Class	Accuracy		Accuracy		Accuracy	
A	98.39	98.39	98.39	98.71	98.39	98.71
B	96.05	36.50	56.76	21.00	66.67	23.00
C	54.63	56.50	42.80	55.00	41.37	51.50
D	91.44	96.67	92.69	96.67	90.67	97.14
E	88.20	67.62	86.55	70.48	87.10	64.29
F	41.05	97.50	40.45	90.00	41.30	95.00
G	100.00	67.05	100.00	65.32	100.00	65.32
H	98.14	95.91	97.42	85.91	98.48	88.64
I	33.49	70.00	24.62	65.00	24.33	64.00
J	100.00	100.00	98.02	98.80	98.81	99.60
K	97.22	84.00	99.06	84.00	97.98	77.60
L	93.69	97.20	93.24	96.73	93.27	97.20
M	98.18	27.00	91.61	16.50	88.24	22.50
N	97.66	83.33	30.11	9.33	61.88	33.00
O	88.86	95.16	60.10	78.71	71.79	81.29
P	80.00	96.00	66.14	97.67	65.31	96.00
Q	99.07	96.82	97.69	95.91	96.76	95.00
R	100.00	94.00	100.00	92.00	93.33	84.00
S	81.94	98.33	67.06	95.00	65.48	91.67
T	45.61	80.00	54.64	81.54	48.21	83.08
U	43.21	78.00	43.73	68.00	33.89	51.00
V	71.43	70.00	47.83	66.00	43.06	62.00
W	95.40	83.00	96.05	73.00	94.70	71.50
	Accuracy (%) = 82.04		Accuracy (%) = 72.85		Accuracy (%) = 73.29	
	κ = 0.8107		κ = 0.7140		κ = 0.7187	

Table 5 Classification accuracies for the SAM classifications

	194-band		30-band		13-band	
	User	Producer	User	Producer	User	Producer
Class	Accuracy		Accuracy		Accuracy	
A	98.34	95.48	98.70	73.55	98.77	77.42
B	96.05	36.50	54.05	20.00	55.26	21.00
C	54.63	56.00	35.42	42.50	38.14	45.00
D	91.44	96.67	91.44	96.67	91.86	96.67
E	88.20	67.62	89.16	70.48	89.40	64.29
F	41.94	65.00	40.26	77.50	42.17	87.50
G	100.00	8.09	100.00	18.50	100.00	24.86
H	98.14	95.91	97.56	90.91	98.04	90.91
I	33.49	70.00	24.80	63.00	26.52	61.00
J	100.00	100.00	98.41	99.20	98.80	99.20
K	97.22	84.00	99.04	82.40	97.51	78.40
L	93.69	97.20	94.84	94.39	94.50	96.26
M	100.00	27.00	94.44	17.00	92.98	26.50
N	97.66	83.33	5.51	2.33	12.16	6.00
O	88.86	95.16	49.24	62.90	49.24	62.90
P	80.00	96.00	72.84	95.67	70.05	94.33
Q	99.07	96.82	97.71	96.82	96.77	95.45
R	100.00	94.00	79.63	86.00	78.43	80.00
S	81.94	98.33	63.01	76.67	61.04	78.33
T	45.61	80.00	53.54	81.54	48.21	83.08
U	43.21	78.00	39.21	64.50	36.67	60.50
V	71.43	70.00	41.94	78.00	40.79	62.00
W	95.40	83.00	93.46	71.50	94.70	71.50
	Accuracy (%) = 79.18		Accuracy (%) = 66.37		Accuracy (%) = 66.81	
	$\kappa = 0.7808$		$\kappa = 0.6466$		$\kappa = 0.6510$	

that for high-dimensional data the sophistication of the classifier can make a significant difference. It also shows that even for low-dimensional data the difference in performance can be considerable (such as in the 13-band case). Somewhat puzzling is the fact that the accuracy of most of the classifiers is lower for the 30-band case than for the 13-band selection. We have not investigated the reason of this seeming contradiction, but we can speculate that certain combinations of the subselected bands (such as in our 30-band case) may not add more information while it increases the burden on the classifier for the discrimination of classes. One previous work that seems to support this thought is [62], where an inconclusive trend of classification accuracies was observed as a function of band selections made with increasing number of highest-magnitude wavelet coefficients. Another, more recent, work showed that selection of bands based on intelligent understanding of the data structure combined with taking the classification goals into account produces better results and a consistent trend with the number of bands [76].

We add for completeness that if we exclude the unclassified pixels from the confusion matrices as 'neutral' (neither wrong nor correct), the accuracy of the ANN and the SAM classifiers are higher than shown in Table 8 ($\approx$92% and $\approx$82%, respectively), while the accuracies of the other classifiers remain the same as those have no unclassified samples. While the statistic without the unclassified pixels provides less than a complete picture of the classification quality, it is a valuable measure of the classifier's pattern matching capability and its sensitivity to uncertainty. Table 9 lists the overall accuracies when calculated without the unclassified pixels, i.e., the percentage of correctly classified test samples which were assigned a label by the classifier. Table 9 also shows that the number of unclassified test pixels is small (approximately 3.4%) for

Table 6 Classification accuracies for the MLH classifications

	194-band		30-band		13-band	
	User	Producer	User	Producer	User	Producer
Class	Accuracy		Accuracy		Accuracy	
A			96.81	97.74	96.87	99.68
B			96.97	16.00	87.23	61.50
C			58.08	84.50	80.44	90.50
D			62.86	94.29	87.34	95.24
E			35.52	98.10	41.50	100.00
F			54.00	67.50	10.93	100.00
G			14.87	83.82	48.89	50.87
H			100.00	5.91	77.51	73.64
I			82.08	87.00	36.72	94.00
J			100.00	1.20	0.00	0.00
K			80.00	8.00	98.55	54.40
L			32.47	99.53	66.88	97.20
M			54.55	6.00	100.00	2.50
N			92.73	17.00	100.00	6.33
O			56.20	65.81	49.31	69.03
P			95.42	76.33	71.32	90.33
Q			100.00	12.27	100.00	17.73
R			100.00	68.00	100.00	32.00
S			96.88	51.67	93.33	46.67
T			100.00	30.77	85.71	92.31
U			93.24	34.50	80.39	82.00
V			51.47	70.00	100.00	26.00
W			89.66	13.00	65.70	79.50
	Accuracy (%) = N/A		Accuracy (%) = 49.72		Accuracy (%) = 63.23	
	κ = N/A		κ = 0.4713		κ = 0.6136	

the 194-band cases, and it remains below 10% for all cases. It is interesting though that for both ANN and SAM the 30-band cases have more unclassified pixels than either the 194- or the 13-band cases and that the exclusion of the unclassified test pixels from the accuracy calculation results in a reversal of the accuracy ranking between the 13- and the 30-band maps.

Customarily, classification accuracies are assessed for several cross-validation folds, and their average and standard deviation are reported. Here, we show the results from single runs, for largely historical reasons. We performed these experiments years ago when computing power was too limited to do lengthy ANN training (including the training of an SOM) with 23 classes of nearly 200-dimensional data, multiple times. However, one of us (EM) recently conducted cross-validation runs with the same hybrid ANN and data as part of a hyperspectral data compression study (WHISPERS [77]), applying a different random cut of the labeled data via stratified random sampling, using the same number of training samples in each fold as shown in Table 1 of this paper, and the rest for test samples. The average accuracy remained very close to that reported here (89.8% for overall (weighted), which is what we report in this paper), and 93.10% for unweighted), with less than 1% variation (STD = 0.46%) across the folds. Since the ANN classifier involves the most random elements of all methods used in this work, the variance would be even smaller for the other - deterministic - classifiers across the same cross-validation folds. Thus, re-doing all the conventional classifications for these cross-validation folds as well as all κ-tables would not produce additional value for this paper, while it would be a difficult and time-consuming exercise.

5 Conclusions

The main thrust of this paper was to compare the performances of classifiers for hyperspectral data under realistic circumstances, for a large number of classes. We used real

Table 7 Classification accuracies for the MHD classifications

	194-band		30-band		13-band	
	User	Producer	User	Producer	User	Producer
Class	Accuracy		Accuracy		Accuracy	
A			93.31	90.00	97.75	98.06
B			75.00	60.00	100.00	18.50
C			58.49	46.50	68.36	87.50
D			75.74	84.76	83.47	93.81
E			99.44	84.76	39.11	100.00
F			35.87	82.50	10.03	100.00
G			100.00	50.29	36.75	24.86
H			72.91	83.18	78.80	77.73
I			63.50	87.00	23.54	93.00
J			97.50	93.60	100.00	0.40
K			80.81	64.00	93.94	49.60
L			74.09	85.51	71.08	95.33
M			40.38	10.50	00.00	00.00
N			85.25	52.00	100.00	5.00
O			57.49	60.65	48.41	59.03
P			82.43	81.33	78.19	83.67
Q			87.76	94.55	100.00	11.82
R			82.46	94.00	100.00	24.00
S			89.06	95.00	100.00	33.33
T			89.71	93.85	94.83	84.62
U			48.48	95.50	81.60	66.50
V			25.53	72.00	100.00	12.00
W			51.53	59.00	84.92	76.00
	Accuracy (%) = N/A		Accuracy (%) = 72.53		Accuracy (%) = 59.34	
	κ = N/A		κ = 0.7108		κ = 0.5726	

Table 8 Summary of overall classification accuracies and κ values for test data

	194-band		30-band		13-band	
	%	κ	%	κ	%	κ
ANN	88.71	0.88	75.02	0.74	76.36	0.75
MED	82.04	0.81	72.85	0.71	73.29	0.72
SAM	79.18	0.78	66.37	0.65	66.81	0.65
MLH			49.72	0.47	63.23	0.61
MHD			72.53	0.71	59.34	0.57

Table 9 Summary of overall classification accuracies for test data, computed with exclusion of unclassified samples

	194-band		30-band		13-band	
	%	#uc	%	#uc	%	#uc
ANN	91.89	150	82.13	375	81.24	259
MED	82.04	0	72.85	0	73.29	0
SAM	81.92	145	69.34	186	69.30	156
MLH			49.72	0	63.23	0
MHD			72.53	0	59.34	0

Numbers under #uc indicate the number of unclassified test samples (out of 4,332 test samples) for each case.

AVIRIS data with real noise. The SOM-hybrid ANN classifier produced the most accurate map from the 194-band hyperspectral data with 23 cover classes (89%), followed by the minimum Euclidean distance algorithm (82%) and the spectral angle mapper (79%). The two covariance-based methods, maximum likelihood and Mahalonobis distance, could not be applied to the full spectral resolution, which resulted in the best case map accuracies of 63% and 73% for these classifiers, respectively, on reduced dimensionality versions of the data.

We plan to extend the range of comparative classification algorithms in subsequent work to more advanced classifiers such as SVMs, tree-based methods, boosting, and relatively new ones that have been gaining recognition. Some candidates are constrained energy minimization (CEM) [78], 'Tetracorder' [79], and the n-dimensional probability density function (n-dPDF) [80]. Further interesting comparisons would be with Bayesian classifiers (e.g., [81]), rule-based AI classifiers (e.g., [82]), or some of those (variants of Bayesian, neural net, minimum distance classifiers) in the data mining environment of ADaM [83]. However, not all of these (and other emerging) algorithms are commonly available or straighforward to use; thus, a comparative study would need more extensive collaboration with their authors.

Equipped with the capability to produce a good benchmark classification with full spectral resolution, one can do systematic dimensionality reduction and rigorously assess the effect. We note that dimensionality reduction is most frequently performed by PCA or wavelets, or by selection of important bands by domain experts. We found undesirable loss of class distinction with all of these approaches ([24,62] and as discussed in this paper). Non-linear dimensionality reduction approaches, especially with AI, neural network techniques such as by [20,63-65,76], retain more of the relevant information and can improve classification accuracy at the same time.

A systematic investigation of the classifiers' noise sensitivity is a desirable subject of a future study. While one classifier may outperform others in a low-noise situation, another could prove more robust under noisy circumstances even if the classification accuracy is lower. These and other properties of classifiers should make up a more complete picture of the suitability of different methods for different purposes.

ANN classifiers, including the above SOM-hybrid classifier, are directly applicable to fused disparate data (such as stacked spectral, elevation, or geophysical measurements), which could improve classification, but processing such data with traditional methods is admittedly a problem because estimation of the relative contributions of the different components is difficult [84]. Neural approaches, in contrast, can derive those contribution weightings during supervised learning from labeled data samples.

Economy of computation is another important aspect by which methods could and should be compared. We did not do it here because ANN learning is an inherently massively parallel procedure which, when run on sequential computers, is very slow. A training session for this LCVF image, including the concurrent monitoring of the training, can take hours on low-end Sun workstations (depending on the CPU speed of the given machine). Real, large-scale applications will need to invest in appropriate massively parallel hardware in order to utilize the full power of ANNs.

Finally, we want to suggest that accuracy assessment will need to be dealt with differently for hyperspectral classifications than for lower-dimensionality data. Hyperspectral dimensionality poses a difficult challenge for rigorous performance evaluations because of the unavailability of the number of test samples required by theories. One possibility to overcome this, for the purpose of comparing various classifiers, is to use synthetic hyperspectral imagery where each pixel is labeled. This is becoming a realistic choice through rigorous simulation work [85,86]. We want to stress, however, the need for research that can yield new, innovative measures of performance for accuracy evaluation of class maps obtained from real data, for which it is not possible to obtain the requisite number of test samples. Such new measures will have to produce the same accuracies as the κ-statistic or other widely accepted measures for test data that meets the theoretical sampling requirements (such as the test data we constructed for this study), while relying on less test samples and perhaps using more of the internal characteristics of the data.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was partially supported by the Applied Information Systems Research Program (AISRP) of NASA, Science Dierectorate, grants NAG5-9045, NAG5-10432, NNG05GA94G. Part of this work was carried out at the Planetary Image Research Laboratory (PIRL) of the Lunar and Planetary Laboratory (LPL) when the first author was at the University of Arizona. The use of the computing facilities of PIRL/LPL and of the Laboratory for Spatial Image Analysis at the Desert Research Institute, as well as the software support by James Winburn and NASA Arizona Space Grant Interns Michael Shipman and Trevor Laing at PIRL/LPL, are gratefully acknowledged.

Author details
[1]Department of Statistics, Rice University, 6100 Main Street MS-138, Houston, TX 77005, USA. [2]Department of Electrical and Computer Engineering, Rice University, 6100 Main Street, Houston, TX 77005, USA. [3]Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA. [4]Division of Earth and Ecosystem Sciences, Desert Research Institute, 2215 Raggio Parkway, Reno, NV 89512, USA.

Received: 26 June 2013 Accepted: 3 April 2014
Published: 16 May 2014

References

1. RO Green (ed.), *Summaries of the 6th Annual JPL Airborne Geoscience Workshop*, vol. 1 (AVIRIS Workshop, Pasadena, 4–6 Mar 1996)
2. RW Basedow, DC Carmer, ML Anderson, HYDICE: an airborne system for hyperspectral imaging, in *Proc. SPIE*, vol. 2480 (Orlando, 17–18 April 1995), pp. 258–267
3. S Haykin, *Neural Networks. A Comprehensive Foundation* (Prentice Hall, Inc., Simon & Schuster/A Viacom Company, Upper Saddle River, 1999)
4. B Solaiman, MC Mouchot, A comparative study of conventional and neural network classification of multispectral data, in *Proc. Int'l Geosci. and Remote Sensing Symposium*, vol. 3 (Caltech, Pasadena, 8–12 August 1994), pp. 1413–1415
5. A Abuelgasim, S Gopal, Classification of multiangle and multispectral asas data using a hybrid neural network model, in *Proc. Int'l Geosci. and Remote Sensing Symposium*, vol. 3 (Caltech, 8–12 August 1994), pp. 1670–1672
6. MJ Aitkenhead, R Dyer, Improving land-cover classification using recognition threshold neural networks. Photogramm Eng. Rem. Sens. **73**(4), 413–421 (2007)
7. GA Carpenter, MN Gjaja, S Gopal, CE Woodcock, ART neural networks for remote sensing: vegetation classification from Landsat TM and terrain data. IEEE. Trans. Geosci. Rem. Sens. **35**(2), 308–325 (1997)
8. P Dryer, Classification of land cover using optimized neural nets on SPOT data. Photogramm Eng. Rem. Sens. **59**(5), 617–621 (1993)
9. GM Foody, MEJ Cutler, Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. Ecol. Model. **195**, 37–42 (2006)
10. GF Hepner, T Logan, N Ritter, N Bryant, Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. Photogramm. Eng. Rem. Sens. **56**(4), 469–473 (1990)
11. W Huang, R Lippman, Comparisons between neural net and conventional classifiers, in *IEEE First International Conference on Neural Networks* (San Diego, 1987), pp. 485–494
12. SO Këpuska, VZ Mason, A hierarchical neural network system for signalized point recognition in aerial photographs. Photogramm. Eng. Rem. Sens. **61**(7), 917–925 (1995)
13. X-U Liu, AK Skidmore, H Van Oosten, Integration of classification methods for improvement of land-cover map accuracy. ISPRS J. Photogram. Rem. Sens. **56**, 257–268 (2002)
14. JD Paola, RA Schowengerdt, Comparison of neural network to standard techniques for image classification and correlation, in *Proc. Int'l Geosci. and Remote Sensing Symposium*, vol. 3 (Caltech, 8–12 Aug 1994), pp. 1404–1405
15. MK Shepard, RE Arvidson, EA Guinness, DW Deering, Scattering behavior of Lunar Lake playa determined from PARABOLA bidirectional reflectance data. J. Geophys. Res. **18**, 2241–2244 (1991)
16. MF Tenorio, SR Safavian, J Kassebaum, A comparative study of conventional and neural network classification of multispectral data, in *Proc. 10th Annual Int'l Geosci. and Remote Sensing Symposium*, vol. 2, (1990), pp. 1289–1292
17. Y Wang, DL Civco, Artificial neural networks in high dimensional spatial data classification: A performance evaluation, in *ACMS/ASPRS Annual Convention and Exposition Technical Papers*, vol. 3 (Charlotte, 27 Feb–2 Mar 1995), pp. 662–671
18. JA Benediktsson, PH Swain, OK Ersoy, D Hong, Classification of very high dimensional data using neural networks, in *IGARSS'90 10th Annual International Geoscience and Remote Sensing Symp*, vol. 2 (College Park, 20–24 May 1990), pp. 1269–1272
19. B Kim, DA Landgrebe, Hierarchical classifier design in high-dimensional, numerous class cases. IEEE Trans. Geosci. Rem. Sens. **29**(4), 518–528 (1991)
20. JA Benediktsson, JR Sveinsson, K Arnason, Classification of very-high-dimensional data with geological applications, in *Proc. MAC Europe 91* (Lenggries, 1994), pp. 13–18
21. JA Benediktsson, JR Sweinsson, K Arnason, Classification and feature extraction of AVIRIS data. IEEE Trans. Geosci. Rem. Sens. **33**(5), 1194–1205 (1995)
22. ST Monteiro, Y Minekawa, Y Kosugi, T Akazawa, K Oda, Prediction of sweetness and amino acid content in soybean crops from hyperspectral imagery. ISPRS J. Photogram. Rem. Sens. **62**, 2–12 (2007)
23. MS Gilmore, MD Merrill, R Castaño, B Bornstein, J Greenwood, Effect of Mars analogue dust deposition on the automated detection of calcite in visible/near-infrared spectra. Icarus. **172**, 641–646 (2004)
24. ES Howell, E Merényi, LA Lebofsky, Classification of asteroid spectra using a neural network. J. Geophys. Res. **99**(E5), 10847–10865 (1994)
25. E Merényi, ES Howell, LA Lebofsky, AS Rivkin, Prediction of water in asteroids from spectral data shortward of 3 microns. ICARUS. **129**, 421–439 (1997)
26. L Rudd, E Merényi, Assessing debris-flow potential by using AVIRIS imagery to map surface materials and stratigraphy in Cataract Canyon, Utah, in *Proc. 14th AVIRIS Earth Science and Applications Workshop*. ed. by RO Green (Pasadena, 24–27 May 2005)
27. C Lee, DA Landgrebe, Decision boundary feature extraction for neural networks. IEEE Trans. Geosci. Rem. Sens. **8**(1), 75–83 (1997)
28. BM Shahshahani, DA Landgrebe, The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Trans. Geosci. Rem. Sens. **32**(5), 1087–1095 (1994)
29. Q Jackson, DA Landgrebe, An adaptive classifier design for high-dimensional data analysis with a limited training data set. IEEE Trans. Geosci. Rem. Sens. **39**(12), 2664–2679 (2001)
30. MT Fardanesh, OK Ersoy, Classification accuracy improvement of neural network classifiers by using unlabeled data. IEEE. Trans. Geosci. Rem. Sens. **36**(3), 1020–1025 (1998)
31. RE Arvidson, M Dale-Bannister, Archiving and distribution of geologic remote sensing field experiment data. EOS, Trans. Am. Geophysical Union. **72**(17), 176 (1991)
32. WH Farrand, RB Singer, Analysis of altered volcanic pyroclasts using AVIRIS data, in *Proceedings of the Third Airborne Visble/Infrared Imaging Spectrometer (AVIRIS) Workshop* (Pasadena, 20–21 Mar 1991)
33. WH Farrand, VIS/NIR reflectance spectroscopy of tuff rings and tuff cones. PhD thesis, University of Arizona, 1991
34. WH Farrand, RB Singer, E Merényi, Calibration of AVIRIS data to reflectance: a comparison of empirical line, radiative transfer and spectral mixture methods. Rem. Sens. Environ. **47**, 311–321 (1994)
35. DH Scott, NJ Trask, Geology of the Lunar Crater volcanic field, Nye County, Nevada. Technical report, USGS, 1971
36. EB Ekren, EN Hinrichs, GL Dixon, Geologic map of the wall quadrangle, Nye County. Nevada. Misc. Geol. Inv. Map I-179, scale. **1**, 48,000 (1973). USGS
37. DA Roberts, Y Yamagushi, RJP Lyon, Comparison of various techniques for calibration of AIS data, in *Proc. of the 2nd Airborne Imaging Spectrometer Data Analysis Workshop*. ed. by G Vane, AFH Goetz (Pasadena, 1986), pp. 21–30
38. B-C Gao, KB Heidebrecht, AFH Goetz, Derivation of scaled surface reflectances from AVIRIS data. Remote Sens. Environ. **44**, 165–178 (1993)
39. GA Swayze, RN Clark, AFH Goetz, TG Chrien, NS Gorelick, Effects of spectrometer band pass, sampling, and signal-to-noise ratio on spectral identification using the Tetracorder algorithm. J. Geoph. Res. (Planets). **108**(E9), 5105 (2003). doi:10.1029/2002JE001975
40. E Merényi, RB Singer, JS Miller, Mapping of spectral variations on the surface of Mars from high spectral resolution telescopic images. ICARUS. **124**, 280–295 (1996)
41. GW Pouch, DJ Campagna, Hyperspectral direction cosine transformation for separation of spectral and illumination information in digital scanner data. Photogramm. Eng. Rem. Sens. **56**, 475–479 (1990)
42. T Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, New York, 1988)
43. B Widrow, FW Smith, Pattern-recognizing control systems, in *Computer and Information Science Symposium Proceedings* (Spartan Books, Washington, D.C., 1963)
44. E Merényi, Precision mining of high-dimensional patterns with self-organizing maps: interpretation of hyperspectral images, in *Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence (Studies in Fuzziness and Soft Computing, vol. 54),* ed. by P Sincak, J Vascak (Physica Verlag, Heidelberg, 2000)
45. MM Van Hulle, *Faithful Representations and Topographic Maps*. Wiley Series and Adaptive Learning Systems for Signal Processing, Communications, and Control (Wiley, New York, 2000)
46. T Villmann, R Der, M Herrmann, TM Martinetz, Topology preservation in self-organizing feature maps: exact definition and measurement. IEEE Trans. Neural Network. **8**(2), 256–266 (1997)
47. E Merényi, A Jain, T Villmann, Explicit magnification control of self-organizing maps for "forbidden" data. IEEE Trans. Neural Netw. **18**(3), 786–797 (2007)

48. E Merényi, K Tasdemir, L Zhang, ed. by M Biehl, B Hammer, M Verleysen, and T Villmann, Learning highly structured manifolds: harnessing the power of, SOMs, in *Similarity Based Clustering, Lecture Notes in Comnputer Science* (Springer-Verlag, Heidelberg, 2009), pp. 138–168. LNAI 5400
49. D DeSieno, Adding a conscience to competitive learning, in *Proc. IEEE Int'l Conference on Neural Networks (ICNN)*, vol. 1 (New York, July 1988), pp. 117–124
50. T Villmann, E Merényi, B Hammer, Neural maps in remote sensing image analysis. Neural Netw. **16**, 389–403 (2003)
51. NeuralWare, Neural Computing, NeuralWorks Professional II/PLUS v5.40 (2003). www.neuralware.com
52. J Rasure, M Young, An open environment for image processing software development, in *Proceedings of the SPIE/IS&T Symposium in Electronic Imaging*, vol. 1659 (Pasadena, 14 Feb 1992)
53. FA Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, AFH Goetz, The spectral image processing system (SIPS) - interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. **44**, 145–163 (1993)
54. PH Swain, SM Davis (eds.), *Remote Sensing: the Quantitative Approach* (McGraw-Hill, New York, 1978)
55. TM Lillesand, RW Kiefer, *Remote Sensing and Image Interpretation* (Wiley, New York, 1987)
56. J Jensen, *Introductory Digital Image Processing* (Prentice-Hall, Englewood Cliffs, 1986)
57. J Campbell, *Introduction to Remote Sensing* (Guilford, New York, 1996)
58. Leica Geosystems, ERDAS Imagine v.9.1 (2006). www.leica-geosystems.com.
59. ITT Visual Information Systems, ENVI v.4.3 (2006). www.exelisvis.com.
60. Earth Resources Mapping, ER Mapper, v.7.1 (2006). erdas-er-mapper.software.informer.com.
61. JA Benediktsson, JA Palmason, JR Sveinsson, Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans. Geosci. Rem. Sens. **43**(3), 480–491 (2005)
62. T Moon, E Merényi, Classification of hyperspectral images using wavelet transforms and neural networks, in *Proc. Annual SPIE Conf.* (San Diego, 1995), p. 2569
63. MA Kramer, Nonlinear principal component analysis using autoassociative neural networks. Am. Inst. Chem. Eng. **37**, 233–243 (1991)
64. G Licciardi, F Del Frate, Pixel unmixing in hyperspectral data by means of neural networks. IEEE Trans. Inform. Theor. **49**(11), 4163–4172 (2011)
65. G Licciardi, PR Marpu, J Chanussot, JA Benediktsson, Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. IEEE Trans. Geosci. Rem. Sens. **9**(3), 447–451 (2012)
66. E Merényi, RB Singer, WH Farrand, Classification of the, LCVF AVIRIS test site with a Kohonen artificial neural network, in *Summaries of the Fourth Airborne JPL Geoscience Workshop, JPL Publication 93-26*, vol. 1 (Washington, D.C., 25–29 Oct 1993), pp. 117–120
67. MK Shepard, Application of cosmogenic exposure age dating and remote sensing to studies of surficial processes. PhD thesis, Washington University, St. Louis, 1991
68. RG Pontius, M Millones, Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Rem. Sens. **32**(15), 4407–4429 (2011)
69. J Demsår, Statistical comparison of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
70. GM Foody, Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogramm. Eng. Rem. Sens. **70**(5), 627–633 (2004)
71. F Wilcoxon, Probability tables for individual comparisons by ranking methods. Biometrics. **3**(3), 119–122 (1947)
72. K Fitzpatrick-Lins, Comparison of sampling procedures and data analysis for a land-use and land-cover map. Photogramm. Eng. Rem. Sens. **47**(3), 343–351 (1981)
73. PJ Curran, HD Williamson, Sample size for ground and remotely sensed data. Photogramm. Eng. Rem. Sens. **20**, 31–41 (1986)
74. RG Congalton, A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. Photogramm. Eng. Rem. Sens. **54**(5), 593–600 (1988)
75. RG Congalton, C Green, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (CRC, Boca Raton, 1999)
76. MJ Mendenhall, E Merényi, Relevance-based feature extraction for hyperspectral images. IEEE Trans. Neural Netw. **19**(4), 658–672 (2008)
77. B Xie, T Bose, E Merényi, A novel scheme for the compression and classification of hyperspectral images, in *Proc. First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2009)* (Grenoble, 26–28 Aug 2009)
78. WH Farrand, JC Harsanyi, Mapping the distribution of mine tailings in the coeur d'alene river valley, idaho through the use of a constrained energy minimization technique. Rem. Sens. Environ. **59**, 64–76 (1997)
79. RN Clark, GA Swayze, KE Livo, RF Kokaly, SJ Sutley, JB Dalton, RR McDougal, CA Gent, Imaging spectroscopy: earth and planetary remote sensing with the USGS tetracorder and expert systems. J. Geophys. Res. **108**(E12, 5131), 5–1–5–44 (2003)
80. H Cetin, DW Levandowski, Interactive classification and mapping of multi-dimensional remotely sensed data using n-dimensional probability density functions (nPDF). Photogramm. Eng. Rem. Sens. **57**(12), 1579–1587 (1991)
81. J Ramsey, P Gazis, T Roush, P Spirtes, C Glymour, Automated remote sensing with near infrared reflectance spectra: carbonate recognition. Data Min. Knowl. Discov. **6**(3), 277–293 (2002)
82. PR Gazis, T Roush, Autonomous identification of carbonates using near-IR reflectance spectra during the February 1999 Marsokhod field tests. J. Geophys. Res. **106**(E4), 7765–7773 (2001)
83. J Rushing, R Ramachandran, U Nair, S Graves, R Welch, H Lin, ADaM: a data mining toolkit for scientists and engineers. Comput. Geosci. **31**, 607–618 (2005)
84. JA Benediktsson, PH Swain, OK Ersoy, Neural network approaches versus statistical methods in classification of multisource remote sensing data. IEEE. Trans. Geosci. Rem. Sens. **28**(4), 540 (1990)
85. J Schott, S Brown, R Raqueño, H Gross, G Robinson, An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development. Can. J. Rem. Sens. **25**(2), 99–111 (1999)
86. E Ientilucci, S Brown, Advances in wide-area hyperspectral image simulation, in *Proceedings of SPIE*, vol. 5075, (5–8 May 2003), pp. 110–121

doi:10.1186/1687-6180-2014-71
Cite this article as: Merényi *et al.*: **Classification of hyperspectral imagery with neural networks: comparison to conventional tools.** *EURASIP Journal on Advances in Signal Processing* 2014 **2014**:71.