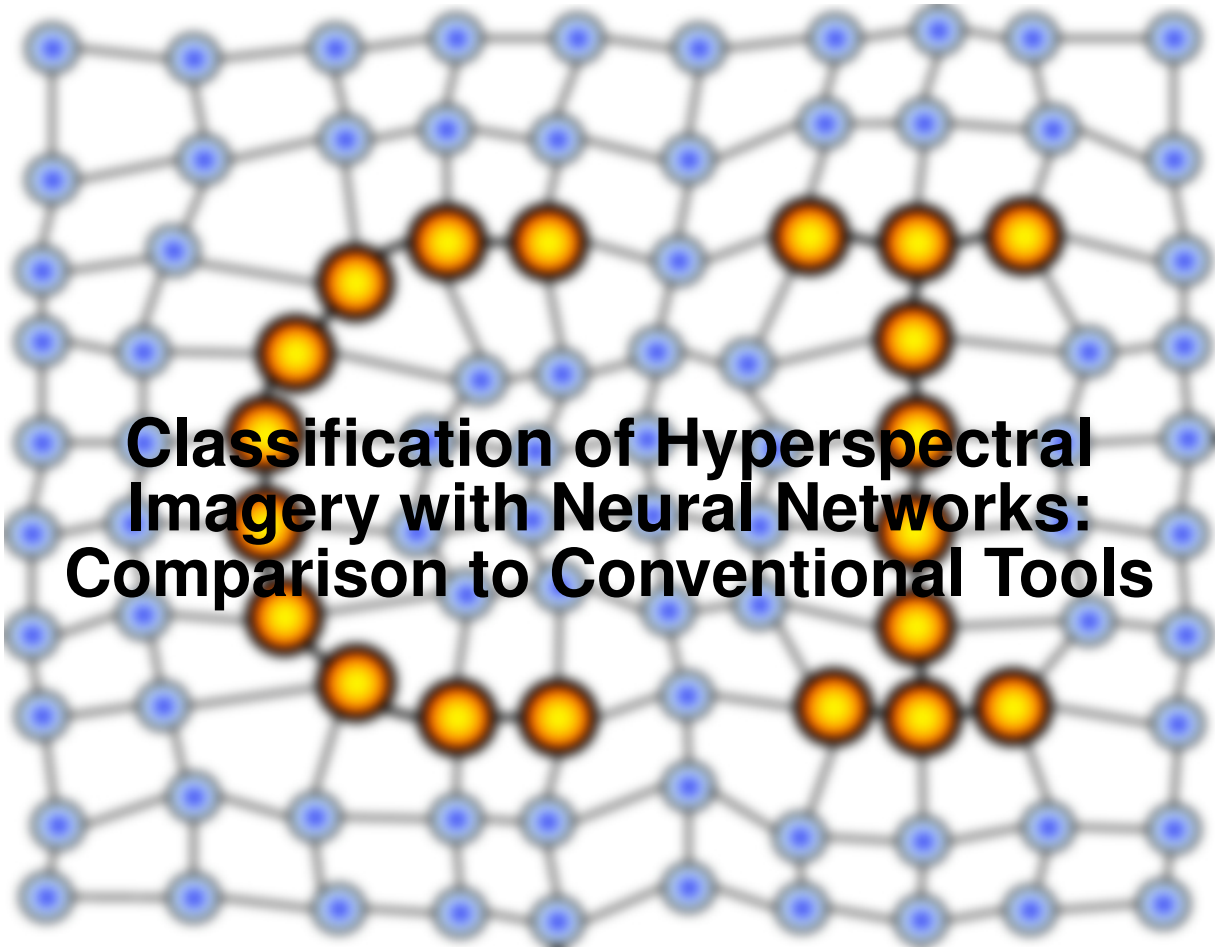
MACHINE LEARNING REPORTS



Classification of Hyperspectral Imagery with Neural Networks: Comparison to Conventional Tools

Report 04/2011
Submitted: 31.05.2011
Published: 06.06.2011

Erzsébet Merényi[1], William H. Farrand[2], James V. Taranik[3], and Timothy B. Minor[4]
(1) Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, U.S.A. {erzsebet@rice.edu}
(2) Space Science Institute, Boulder, CO 80303 {farrand@colorado.edu}
(3) Department of Geological Sciences and Engineering, Mackay School of Earth Science and Engineering, University of Nevada, Reno, NV 89557 {jvtaranik@mines.unr.edu}
(4) Desert Research Institute, 2215 Raggio Parkway, Reno, Nevada 89512-1095 {tim.minor@dri.edu}.

Abstract

Efficient exploitation of hyperspectral imagery is of great importance in remote sensing. Artificial Intelligence approaches have been receiving favorable reviews for classification of hypersepctral 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 an 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 $\approx 90\%$ accuracy on test data.

Classification of Hyperspectral Imagery with Neural Networks: Comparison to Conventional Tools

Erzsébet Merényi, William H. Farrand, James V. Taranik, and Timothy B. Minor

***Abstract*—Efficient exploitation of hyperspectral imagery is of great importance in remote sensing. Artificial Intelligence approaches have been receiving favorable reviews for classification of hypersepctral 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 an 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.**

I. Background, Motivation, Objectives

HIGH spatial and spectral resolution images from advanced remote sensors such as NASA's AVIRIS (*e.g.,* Green, 1996), Hyperion, HyMap, HYDICE (Basedow *et al.,* 1995) 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.

E. Merényi is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, U.S.A. (e-mail: erzsebet@rice.edu).

William H. Farrand is with the Space Science Institute, Boulder, CO 80303 (e-mail: farrand@colorado.edu).

James V. Taranik is with the Department of Geological Sciences and Engineering, Mackay School of Earth Science and Engineering, University of Nevada, Reno, NV 89557 (e-mail: jvtaranik@mines.unr.edu).

Timothy B. Minor is with the Desert Research Institute, 2215 Raggio Parkway, Reno, Nevada 89512-1095 (e-mail: tim.minor@dri.edu).

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.,* Haykin (1999). Many earlier works documented ANN capabilities for remote sensing spectra on relatively modest scales: few (5–12) classes, low-to-moderate number of channels (*e.g.,* Huang and Lippman, 1987, Hepner *et al.,* 1990, Tenorio *et al.,* 1990, Dryer, 1993, Paola and Schowengerdt, 1994, Abuelgasim and Gopal, 1994, Solaiman and Mouchot, 1994, Këpuska and Mason, 1995, Wang and Civco, 1995, Carpenter *et al.,* 1997, Liu *et al.,* 2002, Foody and and Cutler, 2006, Aitkenhead and Dyer, 2007). Several studies for higher spectral resolution (*e.g.,* 60 channels in Benediktsson *et al.,* 1990a, Kim and Landgrebe, 1991) 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.,* Benediktsson *et al.,* 1994, Benediktsson *et al.,* 1995, Monteiro *et al.,* 2007). Some research targeted selected, narrow spectral windows of hyperspectral data, to classify one specific important spectral feature (Gilmore *et al.,* 2004). A small number of ANN works classified hyperspectral data directly, without prior dimensionality reduction (Howell *et al.,* 1994, Merényi *et al.,* 1996, Rudd and Merényi, 2005). 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 30–40 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 (Lee and Landgrebe, 1997, Shahshahani and Landgrebe, 1994) 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 (Jackson and Landgrebe, 2001, Fardanesh and Ersoy, 1998), 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, 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, analysis and results of the comparisons are presented, followed by a discussion of outstanding issues and future directions.

II. STUDY AREA, DATA, AND PREPROCESSING

A. 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 (Arvidson *et al.,* 1991). Since 1992, the large playa in the LCVF, Lunar Lake along with the surrounding terrain, has been one of 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 (Farrand, 1991, Farrand and Singer, 1991, Farrand *et al.,* 1994). Fig. 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 I. 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 (Scott and Trask, 1971). 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 (1973). 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 Fig. 1.

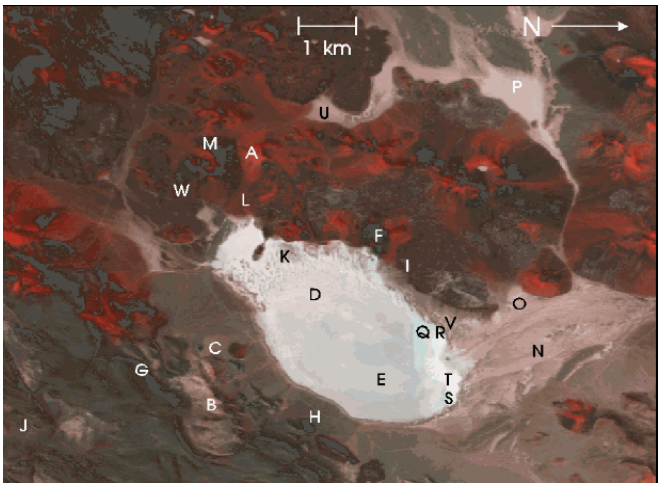


Fig. 1. False color composite of the Lunar Crater Volcanic Field (LCVF) site. Letters stand for the various cover classes used throughout this paper, and mark some of the training locations of distinguishing geologic features for the corresponding classes.

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 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

TABLE I
THE COVER TYPES IN THE LUNAR CRATER VOLCANIC FIELD SITE, WITH CLASS LABELS USED IN THIS STUDY, AND WITH THE NUMBER OF ORIGINAL TRAINING SAMPLES (#TR) IDENTIFIED FOR EACH CLASS.

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

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.

23 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 Fig. 2.

B. Data preprocessing

The LCVF image was atmospherically corrected, and converted to reflectance units, using the Empirical Line Method (*e.g.,* Roberts *et al.,* 1986, Farrand *et al.,* 1994), which produced spectra with fewer noise artifacts for this 1994 image than ATREM (Gao *et al.,* 1993). After exclusion of the excessively noisy and overlapping channels 194 bands remained with excellent signal-to-noise ratio (Swazey *et al.,* 2003). A brightness normalization such as that described in Merényi *et al.,* 1996, (also called the Hyperspherical Directional Cosine Transformation (Pouch and Campagna, 1990)) was also applied 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

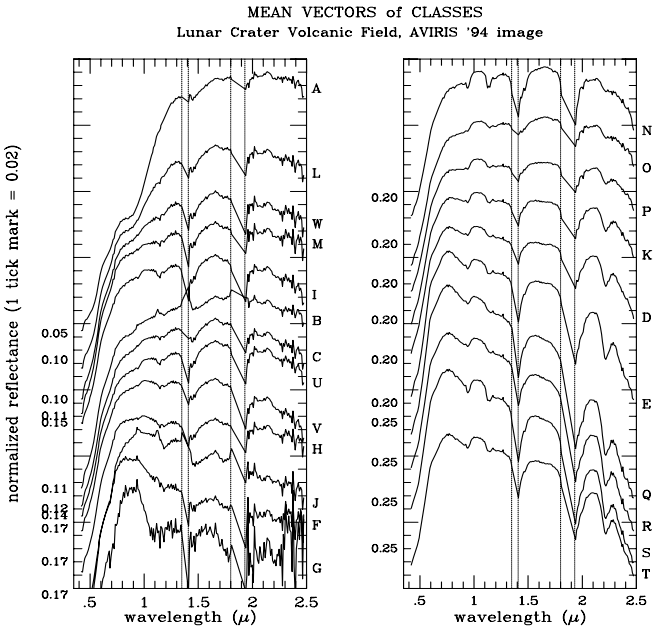


Fig. 2. Representative average spectra of the 23 LCVF classes listed in Table I, vertically offset for clarity. Many signatures (such as the clay bearing series on the right) have subtle variations, others (such as the iron oxide bearing species A, L, W, F, G, on the left) have larger differences. The 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.

spectrally the same but distinguished by albedo, in a post-processing step (as in, *e.g.,* Howell *et al.,* 1994). Fortunately this is not a frequent situation, in our experience. For the present analysis the advantages of this brightness normaliztion 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.

III. CLASSIFIERS AND METHODOLOGY FOR COMPARISON

A. The ANN paradigm and the competing classifiers

Back Propagation (BP) neural networks, which are perhaps the most popular and best known ANN paradigm, can be very 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 is frequently applied to high spectral resolution data to achieve tolerable training time, or training convergence at all with a BP network (*e.g.,* Benediktsson *et al.,* 1994, Paola and Schowengerdt, 1994).

To overcome these dificulties, we used an hybrid ANN architecture, the details of which are given in Merényi *et al.,* 1997 and Howell *et al.,* 1994. Briefly, it consists of an input layer with as many nodes as the number of spectral bands plus one "bias" neuron; a 2-dimensional Self-Organizing

Map (SOM) (Kohonen 1988) as the hidden layer, and the SOM layer is connected to a categorization learning output layer with a Widrow-Hoff (delta) learning rule (Widrow and Smith, 1963, Haykin, 1999). The output layer has one node for each class, and the subsequent layers are fully connected. This network scales up well and trains fast and relatively easily, as demonstrated by previous works (Howell *et al.,* 1994, Merényi *et al.,* 1997, Merényi *et al.,* 2000, Rudd and Merényi, 2005) with up to 200 spectral channels and 20 – 30 classes. It works in two phases. First, in an unsupervised regime (during which the output layer is not allowed to learn), it builds its own view of the manifold structure through forming a cluster map of the data in the hidden SOM layer. Then learning of the output layer is turned on and the weights between the SOM hidden layer and the output layer are trained to recognize labeled data. 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 labeled data in the subsequent supervised phase. This results in better generalization from even a small number of samples, and in higher classification accuracy, than without the SOM stage. Back Propagation, in contrast, is powerful enough to simply "memorize" inconsistent information if the number of training samples is relatively small. This means that the network can learn that the labels of individual training samples are A and B even if the characteristics of the two samples are very similar (for example, B is an A type sample accidentally mislabeled). 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. The delta learning rule (Widrow and Smith, 1963) is much simpler than Back Propagation, which enables successful training even for very high-dimensional input spaces, in this SOM-hybrid architecture. It is also much faster to train the supervised phase of this network than to train a BP network. Once the SOM layer has been trained 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.

In a recent paper Foody and Cutler (2006) 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 Howell *et al.,* 1994.)

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 supporting the delivery of good results. For discussion of related issues, which are beyond the scope of this paper, we refer to vanHulle (2000), Villmann *et al.,* 1997, Merényi *et al.,* 2007, and references therein. We mention here that we use, instead of the basic Kohonen SOM (Kohonen, 1988) a variant called *conscience learning* (DeSieno, 1988), 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 (Merényi *et al.,* 2007). 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 (Merényi, 2000, Villmann *et al.,* 2003).

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 (NeuralWare, 1993), 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, HyperEye, builds on NeuralWare and Khoros (Rasure and Young, 1992) 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.,* (1993). These non-ANN classifiers are well documented in remote sensing texts (Swain and Davis, 1978, Jensen, 1986, Lillesand and Kiefer, 1987, Campbell, 1996), and are commonly available in commercial image analysis packages such as ENVI (ITT Visual Informaton Systems, 2006, Leica Geosystems, 2006, Earth Resource Mapping, 2006), and others. We chose these non-ANN classifiers for this study because of their widespread use and easy accessibility.

B. 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 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.

IV. ANALYSIS AND RESULTS

A. The classification experiments

Altogether 942 training pixels were identified across all classes ranging from as low as 14, to 160 samples for a class, as shown in Table I. 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 $(194 + 1) * 23 = 4485$ samples 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 increasing number of bands, we created a second, augmented trainig 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. 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 II 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 band selection procedure is described below.

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

TABLE II
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	—	—

40-by-40 2-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 and the outputs were scaled into the [0,1] range in each dimension (spectral band), as usual for neural networks. This scaling was done after the preprocessing described in Section II, and with respect to the global minimum and maximum for all bands. 300,000 unsupervised learning steps were performed to allow the SOM to learn the cluster structure of the input data space. In this phase all image pixels were used (without labels). 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 I, 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 $\approx 20,000$ steps, with a learning rate decreasing from 0.15 to 0.01, the training 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 percent of unclassified pixels was low, $\approx$3.45% for the ground truth test pixels (see below) in the best classifications. In addition, the recorded map of decision strenghs 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.,* (2005) 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 Fig. 2, a window of spatial context may blur

class distinctions. This is especially a danger in the case of hyperspectral data that also has 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 I) were used to classify the 13-band AVIRIS image. The 30-band data set was classified using the above mentioned augmented training set. The MLH classifications were run twice for each of the two subsampled data sets: once with prior probabilities, 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.

B. 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 a domain expert. For both the 13- and 30-band subsamples, band selection concentrated on diagnostic features in the visible and VNIR (0.45–1.0 μm) and the SWIR (2.1–2.35 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 differently 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 above mentioned 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 (Moon and Merényi, 1995) also remained unconvincing for the task.

C. 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 IV.D below, 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. Fig. 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 Fig. 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 Fig. 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 (Fig. 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 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 (Farrand, 1991) as well as previous analyses of this scene by various authors (Farrand, 1991, Farrand and Singer, 1991, Merényi *et al.,* 1993, Shepard, 1991, Shepard *et al.,* 1991) also corroborate these observations.

One important point is that the ANN and SAM maps contain unclassified pixels. In contrast, the other classifiers assigned a

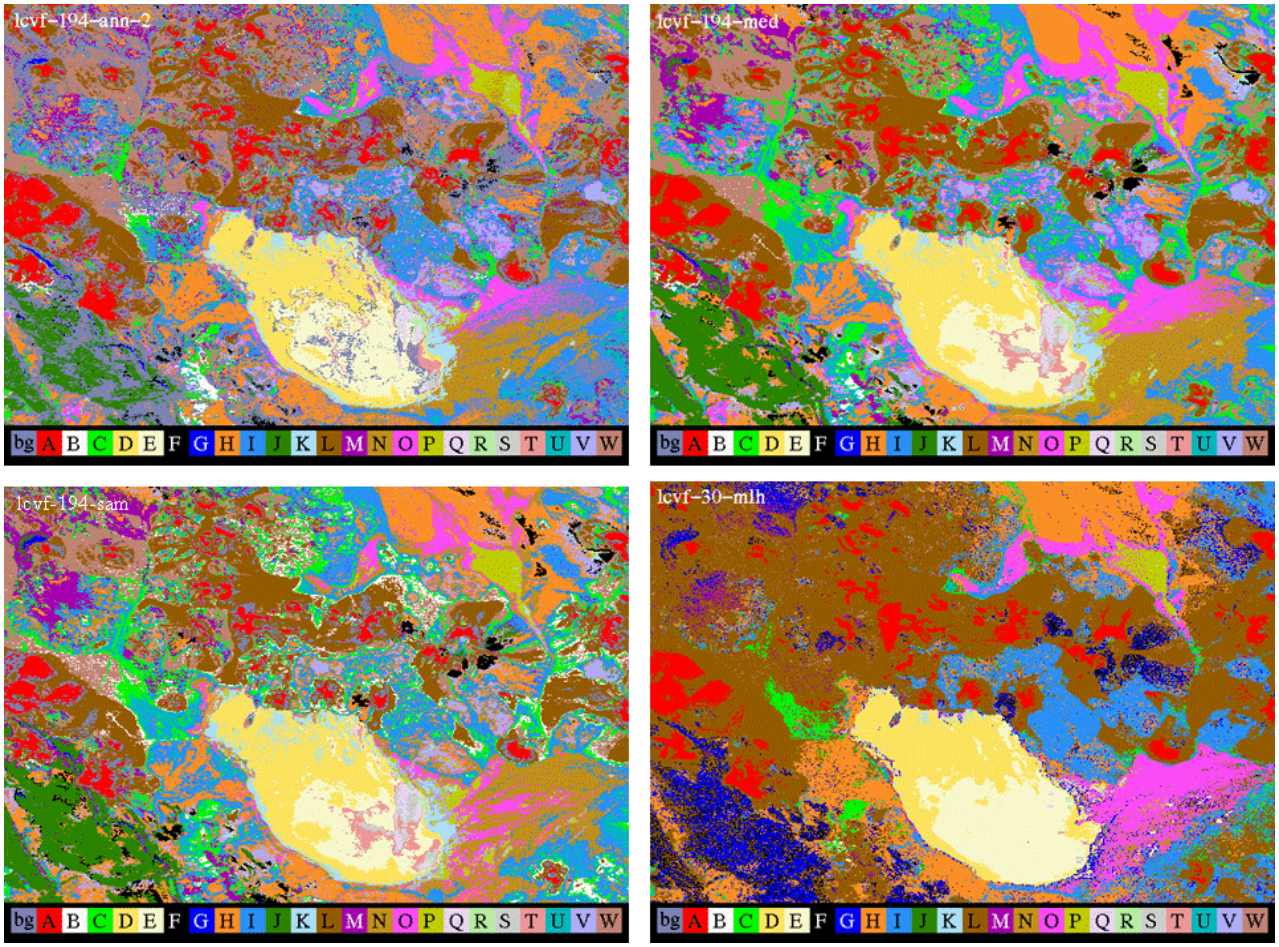


Fig. 3. Comparison of classification maps obtained with four of the classifiers. Clockwise from top left: ANN class map from 194-band data, MED map from 194-band data, MLH map from 30-band data, and SAM map from 194-band data. The class labels are resolved in Table I, and spectral statistics are shown in Fig. 5 and Fig. 6, 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.

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 (Ekren, 1973). This scarp spans across the label G in Fig. 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, 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 Fig. 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 0.9–1.2 μm, 1.4–1.6, and 1.95–2.2 μm windows, which may remain less resolved with the 13- and 30-band selections than with the full (194-band) resolution, as seen in Figs. 5 and 6. 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 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 (Fig. 4) that will be described, and referred to later, in subsection "Assessment of classification accuracies".

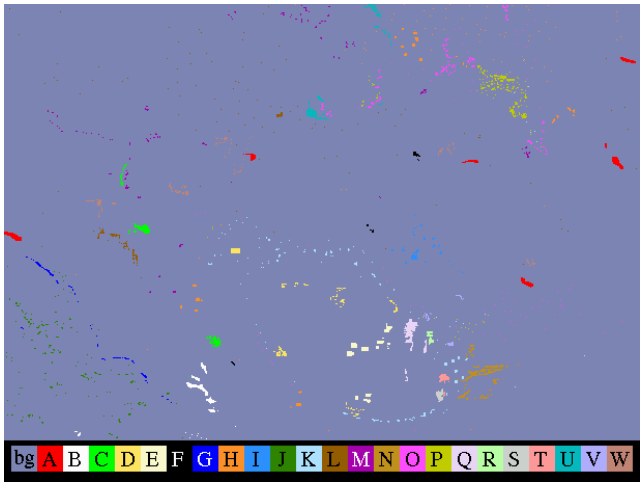


Fig. 4. The ground truth image, containing at least a 100 test pixels for each of the 23 classes, with the exception of those classes that consisted of less than 100 samples. A total of 4332 test pixels were collected, using stratified random sampling.

The spectral plots in Fig. 5 and Fig. 6 show training and class statistics, for the 194-band ANN and for the 30-band MLH classification, respectively. The averages of the training spectra (sold 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 Fig. 6. 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 classification accuracies. It can alert the analyst to poor performance without having to do a full κ-statistics. More importantly, if the number of test samples used in confusion matrices is small, the κ-statistics may not reflect the effect of many misclassified pixels (commission errors). In contrast, in plots like Fig. 5 and Fig. 6 the statistics includes 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 Fig. 6 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 Fig. 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 mach 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, by Fig. 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 Fig. 3. While we are aware of potential uncertainties of this method, as well as of alternative ways of accuracy comparisons (*e.g.,* Demšar, 2006, Foody, 2004, Wilcoxon, 1947) we chose this form of accuracy assessment because of its common current use in remote sensing.

D. Assessment of classification accuracies

The evaluation of classification accuracies followed that outlined in several standard texts on the subject (*e.g.,* Fitzpatrick-Lins, 1981, Curran and Williamson, 1986, Congalton, 1988, Congalton and Green, 1999). Statistical estimation of classification accuracy has been a long studied and established subject that has a vast literature. However, most theoretical considerations 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 (Fitzpatrick-Lins, 1981), the number of test samples needed for map accuracy assessment is $ntest = x^2 * p * (1-p)/E^2$ 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 2700 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.,* Congalton and Green, 1999). 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.

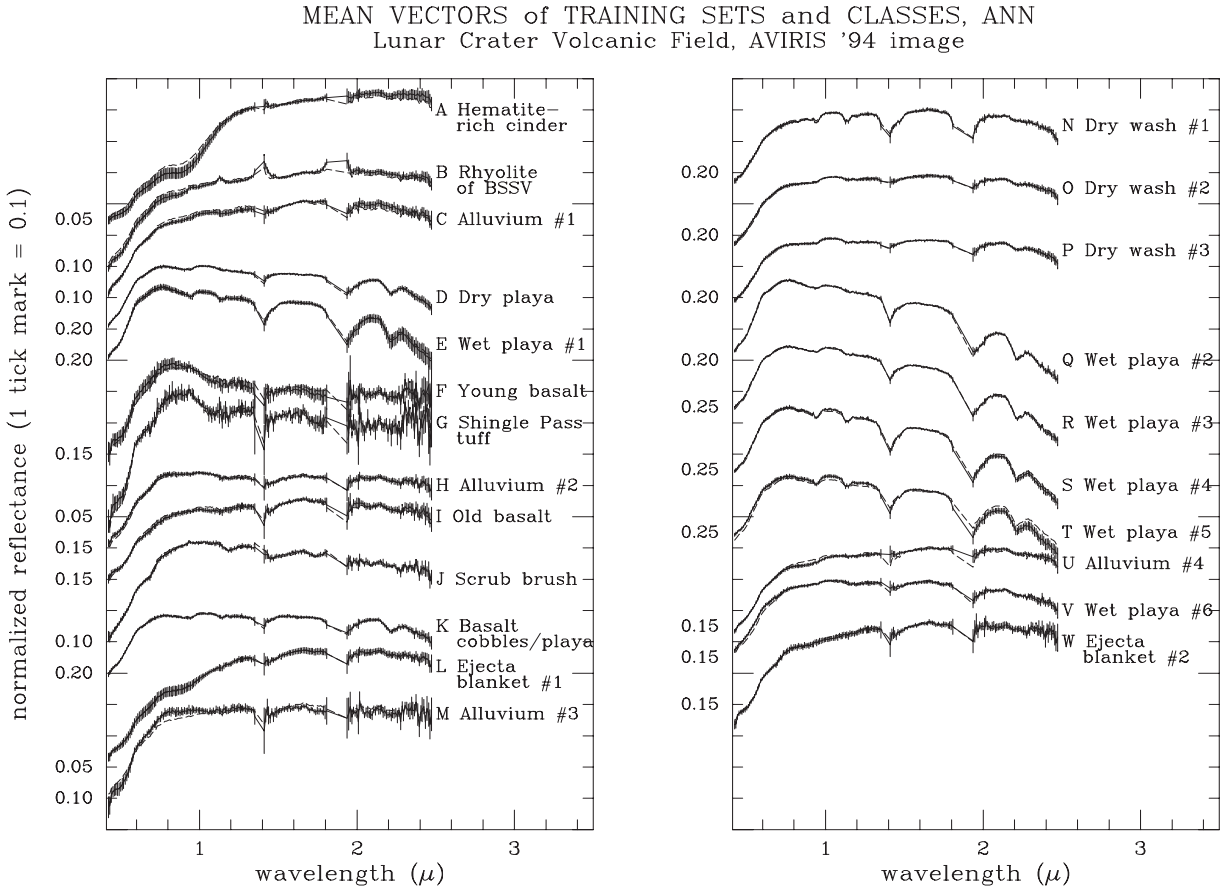


Fig. 5. The mean spectra of the training samples (solid lines) for each class, and the mean of the pixels classified by the SOM-hybrid ANN into the respective classes (dashed lines), using the full, 194-band image cube. The vertical bars show the 1 standard deviation of the training samples. The mean 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.

The literature also strongly recommends strtified a random sampling in 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 a 100 test pixels for each class, altogether 4332 samples (individually verified by one of us (WHF) with extensive field 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 Fig. 4.

Tables III - VII 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 VIII. The numbers in Table VIII support the visual and semi-quantitative evaluations that we made above.

In Table VIII, 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 2–3% 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%

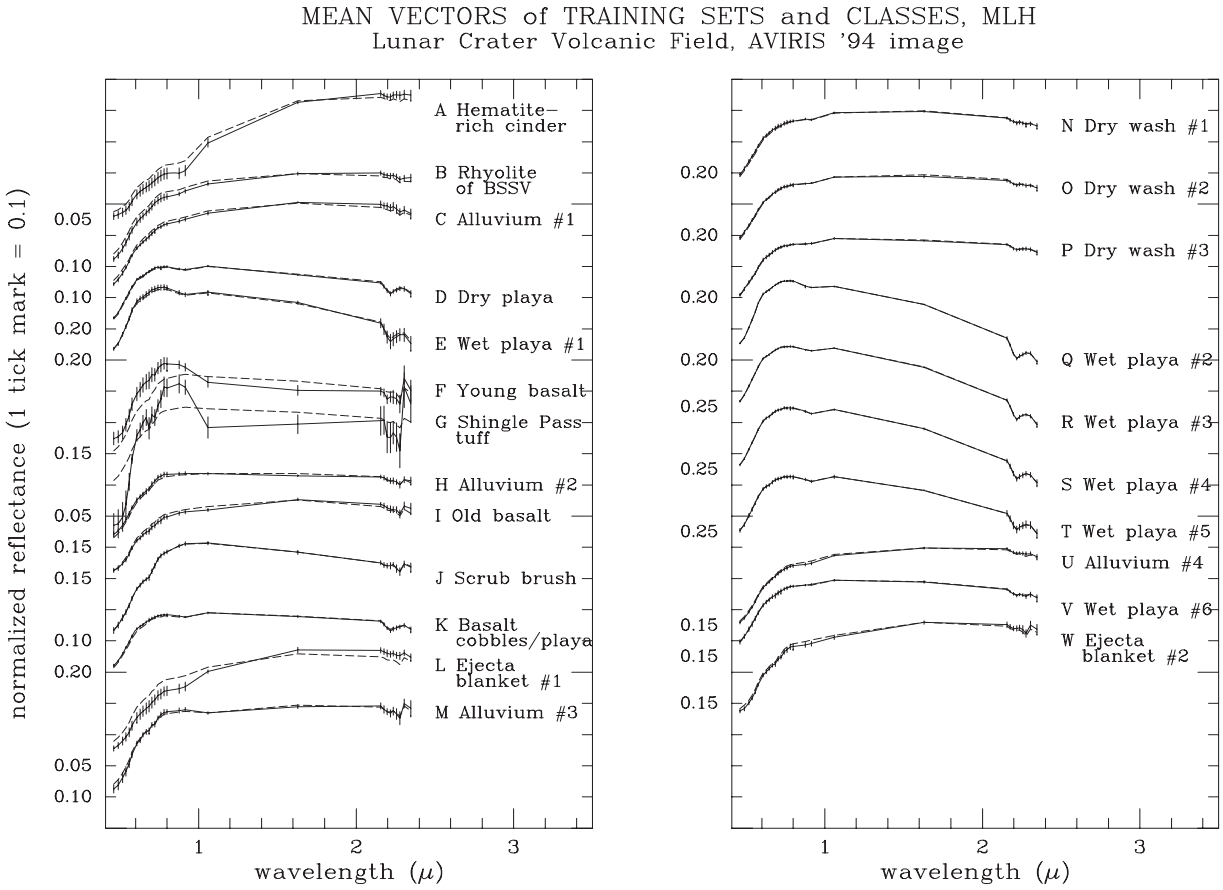


Fig. 6. The mean spectra of the training samples (solid lines) for each class, and the mean of the pixels classified by the MLH classifier into the respective classes (dashed lines), using the 30-band subsampled image cube. The vertical bars show the 1 standard deviation of the training samples. The mean of the predicted classes departs considerably from the mean of the training sampes 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.

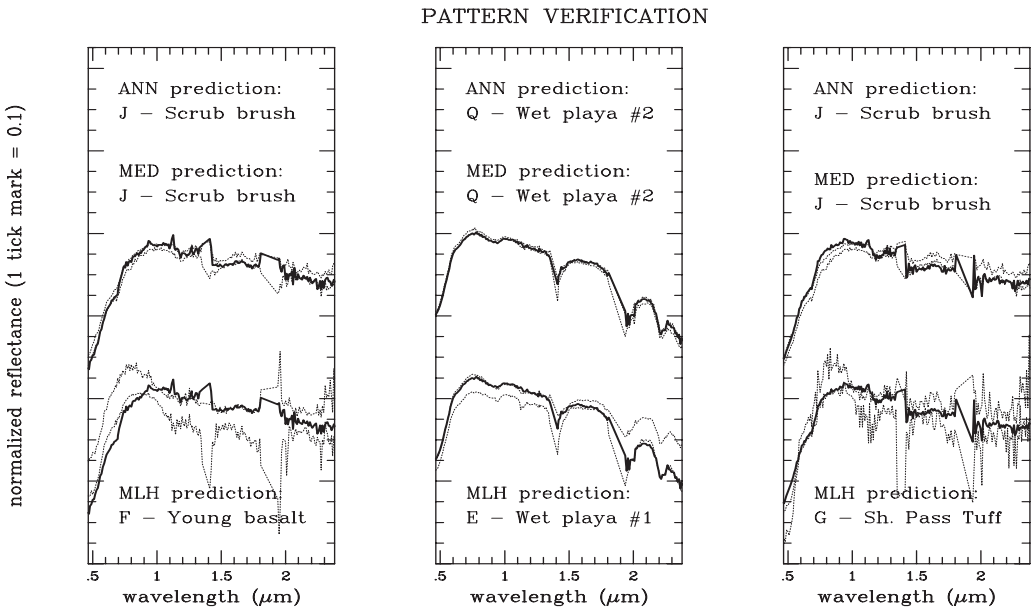


Fig. 7. Examples of typical misclassifications by the MLH classifier, from left to right, for a class J, a class Q, and another class J sample. In each panel, the top plots show how the correctly classified sample pattern (thick line) fits into the envelop of the training samples (thin dotted lines) 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.

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 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 than it increases the "burden" on the classifier for the discrimination of classes. One previous work that seems to support this thought is Moon and Merényi, 1995, 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 (Mendenhall and Merényi, 2006, 2007).

TABLE III
CLASSIFICATION ACCURACIES FOR THE ANN CLASSIFICATIONS

	194-band		30-band		13-band	
	user	prod	user	prod	user	prod
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
	Acc(%)=88.71		Acc(%)=75.02		Acc(%)=76.39	
	κ = .8811		κ = .7380		κ = .7522	

TABLE IV
CLASSIFICATION ACCURACIES FOR THE MED CLASSIFICATIONS

	194-band		30-band		13-band	
	user	prod	user	prod	user	prod
class	accuracy		accuracy		accuracy	
A	98.39	98.39	98.39	98.71	98.39	98.71
B	96.05	36.50	84.00	21.00	66.67	23.00
C	54.63	56.50	26.50	55.00	41.37	51.50
D	91.44	96.67	92.38	96.67	90.67	97.14
E	88.20	67.62	89.05	70.48	87.10	64.29
F	41.05	97.50	-32.50	90.00	41.30	95.00
G	100.00	67.05	100.00	65.32	100.00	65.32
H	98.14	95.91	97.73	85.91	98.48	88.64
I	33.49	70.00	-99.00	65.00	24.33	64.00
J	100.00	100.00	98.00	98.80	98.81	99.60
K	97.22	84.00	99.20	84.00	97.98	77.60
L	93.69	97.20	96.99	96.73	93.27	97.20
M	98.18	27.00	98.50	16.50	88.24	22.50
N	97.66	83.33	78.33	9.33	61.88	33.00
O	88.86	95.16	47.74	78.71	71.79	81.29
P	80.00	96.00	50.00	97.67	65.31	96.00
Q	99.07	96.82	97.73	95.91	96.76	95.00
R	100.00	94.00	100.00	92.00	93.33	84.00
S	81.94	98.33	53.33	95.00	65.48	91.67
T	45.61	80.00	32.31	81.54	48.21	83.08
U	43.21	78.00	12.50	68.00	33.89	51.00
V	71.43	70.00	28.00	66.00	43.06	62.00
W	95.40	83.00	97.00	73.00	94.70	71.50
	Acc(%)=82.04		Acc(%)=72.85		Acc(%)=73.29	
	κ = .8126		κ = .7140		κ = .7187	

TABLE V
CLASSIFICATION ACCURACIES FOR THE SAM CLASSIFICATIONS

	194-band		30-band		13-band	
	user	prod	user	prod	user	prod
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
	Acc(%)=79.18		Acc(%)=66.37		Acc(%)=66.81	
	κ = .7808		κ = .0.6466		κ = .6510	

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 VIII (≈92% and ≈82%, respectively), while the accuracies of the other classifiers remain the same as those have no unclassified samples. While the statistics 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 IX lists the overall accuracies when calculated without the unclassified pixels, *i.e.,* the percent of correctly classified test samples which were assigned

TABLE VI
CLASSIFICATION ACCURACIES FOR THE MLH CLASSIFICATIONS

	194-band		30-band		13-band	
	user	prod	user	prod	user	prod
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
	Acc(%)=N/A		Acc(%)=49.72		Acc(%)=63.23	
	κ = N/A		κ = .0.4713		κ = .6136	

TABLE VII
CLASSIFICATION ACCURACIES FOR THE MHD CLASSIFICATIONS

	194-band		30-band		13-band	
	user	prod	user	prod	user	prod
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
	Acc(%)=N/A		Acc(%)=72.53		Acc(%)=59.34	
	κ = N/A		κ = .0.7108		κ = .5726	

a label by the classifier. Table IX also shows that the number of unclassified test pixels is a small (approx. 3.4%) for 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

TABLE VIII
SUMMARY OF OVERALL CLASSIFICATION ACCURACIES AND κ VALUES FOR TEST DATA.

	194-band		30-band		13-band	
	%	κ	%	κ	%	κ
ANN	88.72	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 IX
SUMMARY OF OVERALL CLASSIFICATION ACCURACIES FOR TEST DATA, COMPUTED WITH EXCLUSION OF UNCLASSIFIED SAMPLES. NUMBERS UNDER #UC INDICATE THE NUMBER OF UNCLASSIFIED TEST SAMPLES (OUT OF 4332 TEST SAMPLES) FOR EACH CASE.

	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

pixels from the accuracy calculation effects a reversal of the accuracy ranking between the 13- and the 30-band maps.

V. CONCLUSIONS AND FUTURE DIRECTIONS

The main thrust of this paper was to compare the performances of classifiers for hyperspectral data under realistic circumstances. We used real 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 resluted in best case map accuracies of 63% and 73% for these classifiers, respectively.

It would be valuable to extend the range of non-ANN classifiers in a future comparative work to relatively new ones that have been gaining recognition. Some of the candidates are Constrained Energy Minimization (CEM) (Farrand and Harsanyi, 1997), the "Tetracorder" algorithm (Clark *et al.,* 2003), and the n-dimensional Probability Density Function (n-dPDF) (Cetin and Levandowski, 1991). Further interesting comparisons would be with Bayesian classifiers (*e.g.,* Ramsey *et al.,* 2002), rule based AI classifiers (*e.g.,* Gazis and Roush, 2001), or some of those (variants of Bayesian, neural net, minimum distance classifiers) in the data mining environment of ADaM (Rushing *et al.,* 2005). However, not all of these (and other emerging) algorithms are commonly available or straighforward to use, thus a comparison study would need more extensive collaboration with their authors.

Equipped with the capability to produce a good benchmark

classification with the full spectral resolution of hyperspectral data, 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 (Howell *et al.,* 1994, Moon and Merényi, 1995, and as discussed in this paper). Non-linear dimensionality reduction approaches, especially with AI, neural network techniques such as by Benediktsson *et al.,* 1994, Mendenhall and Merényi, 2006, retain more of the relevant information and can improve classification accuracy at the same time.

A systematic investigation of noise sensitivity to test how the classifiers' performance changes with known amounts of various forms of noise 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, and the above described SOM-hybrid classifier in particular, are directly applicable to fused disparate data (such as spectral, elevation or geophysical measurements together), 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 (Benediktsson *et al.,* 1990b). 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 the ANN 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 for menaningful statistics. One possibility to overcome this, for the purpose of comparing various classifiers, is to use synthetically created hyperspectral imagery where each pixel is labeled. This is becoming a realistic choice through rigorous simulation work (Schott *et al.,* 1999, Ientilucci and Brown, 2003). 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 results as the κ-statistics or other widely accepted measure, for known test data (such as we constructed for this study, or for synthetic data), while relying on less test data points and perhaps using more of the internal characteristics of the data.

ACKNOWLEDGMENTS

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 software support by James Winburn and NASA Space Grant Interns Michael Shipman and Trevor Laing at PIRL/LPL, are gratefully acknowledged.

REFERENCES

[1] A. Abuelgasim and S. Gopal. Classification of multiangle and multispectral asas data using a hybrid neural network model. In *Proc. Int'l Geosci. and Remote Sensing Symposium*, volume III, pages 1670–1672, Caltech, Pasadena, CA, August 8–12 1994.

[2] M. J. Aitkenhead and R. Dyer. Improving land-cover classification using recognition threshold neural networks. *Photogrammetric Engineering & Remote Sensing*, 73(4):413–421, April 2007.

[3] Arvidson, R. E. and M. Dale-Bannister and. et al. 1991. Archiving and distribution of Geologic Remote Sensing Field Experiment data. *EOS, Transactions of the American Geophysical Union*, 72(17):176.

[4] J.A. Benediktsson, J.A. Palmason, and J.R. Sveinsson. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):480–491, 2005.

[5] J.A. Benediktsson, J.R. Sweinsson, and K. Arnason. Classification and feature extraction of AVIRIS data. *IEEE Transactions on Geoscience and Remote Sensing*, 33(5):1194–1205, September 1995.

[6] J. A. Benediktsson J. R. Sveinsson and et al. Classification of very-high-dimensional data with geological applications. In *Proc. MAC Europe 91*, pages 13–18, Lenggries, Germany, 1994.

[7] J. A. Benediktsson, P. H. Swain, and et al. Classification of very high dimensional data using neural networks. In *IGARSS'90 10th Annual International Geoscience and Remote Sensing Symp.*, volume 2, page 1269, 1990a.

[8] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE. Trans. Geosci. and Remote Sens.*, 28(4):540, 1990b.

[9] J. Campbell *Introduction to Remote Sensing* The Guilford Press, U.S.A., 1996.

[10] G. A. Carpenter, M. N. Gjaja, and et al. ART neural networks for remote sensing: Vegetation classification from Landsat TM and terrain data. *IEEE. Trans. Geosci. and Remote Sens.*, 35(2):308–325, 1997.

[11] H. Cetin and D.W. Levandowski. Interactive classification and mapping of multi-dimensional remotely sensed data using n-dimensional probability density functions (nPDF). *Photogrammetric Engineering & Remote Sensing*, 57(12):1579–1587, 1988.

[12] R.N. Clark, G.A. Swayze, K.E. Livo, R.F. Kokaly, S.J. Sutley, J.B. Dalton, R.R. McDougal, and C.A. 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.

[13] R. G. Congalton and C. Green. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Lewis Publishers, U.S.A., 1999.

[14] R. G. Congalton. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 54(5):593–600, 1988.

[15] P. J. Curran and H. D. Williamson. Sample size for ground and remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 20:31–41, 1986.

[16] J. Demsăr. Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 2006.

[17] D. DeSieno. Adding a conscience to competitive learning. In *Proc. IEEE Int'l Conference on Neural Networks (ICNN), July 1988*, volume I, pages I–117–124, New York, 1988.

[18] P. Dryer. Classifcation of land cover using optimized neural nets on SPOT data. *Photogrammetric Engineering & Remote Sensing*, 59(5):617–621, 1993.

[19] Earth Resources Mapping. *ER Mapper, v.7.1*, 2006.

[20] E. B. Ekren, E. N. Hinrichs, and G. L. Dixon. Geologic map of the Wall quadrangle, Nye County, Nevada. Misc. Geol. Inv. Map I-179, scale 1:48,000, USGS, 1973.

[21] W.H. Farrand and J.C. Harsanyi. Mapping the distribution of mine tailings in the coeur dalene river valley, idaho through the use of a constrained energy minimization technique. *Remote Sensing of Environment*, 59:64–76, 1997.

[22] W. H. Farrand, R. B. Singer, and E. Merényi. Calibration of AVIRIS data to reflectance: A comparison of empirical line, radiative transfer and spectral mixture methods. *Remote Sensing of Environment*, 47:311–321, 1994.

[23] W. H. Farrand. *VIS/NIR Reflectance Spectroscopy of Tuff Rings and Tuff Cones*. PhD thesis, University of Arizona, 1991.

[24] W. H. Farrand and R. B. Singer. Analysis of altered volcanic pyroclasts using AVIRIS data. In *Proceedings of the Third Airborne Visble/Infrared Imaging Spectrometer (AVIRIS) Workshop*, Pasadena, CA, March 20–21 1991.

[25] M. T. Fardanesh and O. K. Ersoy. Classification accuracy improvement of neural network classifiers by using unlabeled data. *IEEE. Trans. Geosci. and Remote Sens.*, 36(3):1020–1025, 1998.

[26] K. Fitzpatrick-Lins. Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering & Remote Sensing*, 47(3):343–351, 1981.

[27] G.M. Foody and M.E.J. Cutler. Mapping the species richness and composition of tropical forests from remotely sensed data with neural networks. *Ecological Modelling*, 195:37–42, 2006.

[28] G.M. Foody. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, 70(5):627–633, May 2004.

[29] B.-C. Gao, K. B. Heidebrecht, and A. F. H. Goetz. Derivation of scaled surface reflectances from AVIRIS data. *Remote Sens. Environ.*, 44:165–178, 1993.

[30] P.R. Gazis and 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, April 25 2001.

[31] M.S. Gilmore, M.D. Merrill, R. Casta no, B. Bornstein, and J. Greenwood. Effect of Mars analogue dust deposition on the automated detection of calcite in visible/near-infrared spectra. *Icarus*, 172:641–646, 2004.

[32] S. Haykin. *Neural Networks. A Comprehensive Foundation*. McMillan, New Jersey, 1999.

[33] G. F. Hepner, T. Logan N., and Ritter N. Bryant. Artificial neural network classification using a minimal training set: Comparision to conventional supervised classification. *Photogrammetric Engineering & Remote Sensing*, 56(4):469–473, 1990.

[34] E. S. Howell, E. Merényi, and L. A. Lebofsky. Classification of asteroid spectra using a neural network. *Jour. Geophys. Res.*, 99(E5):10,847–10,865, 1994.

[35] W. Huang and R. Lippman. Comparisions between neural net and conventional classifiers. In *IEEE First International Conference on Neural Networks*, pages 485–494, San Diego, CA, 1987.

[36] M. M. Van Hulle. *Faithful Representations and Topographic Maps*. Wiley Series and Adaptive Learning Systems for Signal Processing, Communications, and Control. Wiley & Sons, New York, 2000.

[37] E. Ientilucci and S. Brown. Advances in wide-area hyperspectral image simulation. In *Proceedings of SPIE*, volume 5075, pages 110–121, May 5–8 2003.

[38] ITT Visual Information Systems. *ENVI v.4.3*, 2006.

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

[40] J. Jensen. *Introductory digital image processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[41] V. Z. Këpuska and S. O. Mason. A hierarchical neural network system for signalized point recognition in aerial photographs. *Photogrammetric Engineering & Remote Sensing*, 61(7):917–925, 1995.

[42] B. Kim and D.A. Landgrebe. Hierarchical classifier design in high-dimensional, numerous class cases. *IEEE Transactions on Geoscience and Remote Sensing*, 29(4):518–528, July 1991.

[43] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, New York, 1988.

[44] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. The spectral image processing system (SIPS) - interactive visualization and analysis of imaging spectrometer data. *Remote Sens. Environ.*, 44:145–163, 1993.

[45] C. Lee and D.A. Landgrebe. Decision boundary feature extraction for neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 8(1):75–83, January 1997.

[46] Leica Geosystems. *ERDAS Imagine v.9.1*, 2006.

[47] T. M. Lillesand and R. W. Kiefer. *Remote Sensing and Image Interpretation*. John Wiley & Sons, Inc, New York, New York, 1987.

[48] X-U. Liu, A.K. Skidmore, and H. Van Oosten. Integration of classification methods for improvement of land-cover map accuracy. *ISPRS Journal of Photogrammetry & Remote Sensing*, 56:257–268, 2002.

[49] M.J. Mendenhall and E. Merényi. Generalized relevance learning vector quantization for classification driven feature extraction from hyperspectral data. In *Proc. ASPRS 2006 Annual Conference and Technology Exhibition*, page 8, Reno, Nevada, May 5–8 2006.

[50] M.J. Mendenhall and E. Merényi. Relevance-based feature extraction for hyperspectral images. *IEEE Trans. on Neural Networks*, 19(4):658–672, Aprl 2008.

[51] E. Merényi, A. Jain, and T. Villmann. Explicit magnification control of self-organizing maps for "forbidden" data. *IEEE Trans. on Neural Networks*, 18(3):786–797, May 2007.

[52] 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, P. Sincak and J. Vascak Eds.)*. Physica Verlag, 2000.

[53] E. Merényi, W. H. Farrand, L.E. Stevens, T.S. Melis, and K. Chhibber. Mapping Colorado River ecosystem resources in Glen Canyon: Analysis of hyperspectral low-altitude AVIRIS imagery. In *Proc. ERIM, 14th Int'l Conference and Workshops on Applied Geologic Remote Sensing, 4–6 November, 2000, Las Vegas, Nevada*, 2000.

[54] E. Merényi, E. S. Howell, and et al. Prediction of water in asteroids from spectral data shortward of 3 microns. *ICARUS*, 129:421–439, 1997.

[55] E. Merényi, R. B. Singer, and J. S. Miller. Mapping of spectral variations on the surface of Mars from high spectral resolution telescopic images. *ICARUS*, 124:280–295, 1996.

[56] E. Merényi, R. B. Singer, and W. H. 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*, volume 1, pages 117–120, Washington, D.C., Oct. 25-29 1993.

[57] S.T. Monteiro, Y. Minekawa, Y. Kosugi, T. Akazawa, and K. Oda. Prediction of sweetness and amino acid content in soybean crops from hyperspectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, page in print, 2007.

[58] T. Moon and E. Merényi. Classification of hyperspectral images using wavelet transforms and neural networks. In *Proc. Annual SPIE Conf.*, page 2569, San Diego, CA, 1995.

[59] NeuralWare. *Neural Computing, NeuralWorks Professional II/PLUS*, 1993.

[60] J. D. Paola and R. A. Schowengerdt. Comparison of neural network to standard techniques for image classification and correlation. In *Proc. Int'l Geosci. and Remote Sensing Symposium*, volume III, pages 1404–1405, Caltech, Pasadena, CA, August 8–12 1994.

[61] G. W. Pouch and D. J. Campagna. Hyperspectral Direction Cosine Transformation for separation of spectral and illumination information in digital scanner data. *Photogrammetric Engineering & Remote Sensing*, 56:475–479, 1990.

[62] Joseph Ramsey, Paul Gazis, Ted Roush, Peter Spirtes, and Clark Glymour. Automated remote sensing with near infrared reflectance spectra: Carbonate recognition. *Data Min. Knowl. Discov.*, 6(3):277–293, 2002.

[63] J. Rasure and M. Young. An open environment for image processing software development. In *Proceedings of the SPIE/IS&T Symposium in Electronic Imaging*, volume 1659, Pasadena, CA, February 14 1992.

[64] D.A. Roberts, Y. Yamagushi, and R.J.P. Lyon. Comparison of various techniques for calibration of AIS data. In G. Vane and A.F.H. Goetz, editors, *Proc. of the 2nd Airborne Imaging Spectrometer Data Analysis Workshop*, pages 21–30, Pasadena, CA, 1986.

[65] L. Rudd and E. Merényi. Assessing debris-flow potential by using AVIRIS imagery to map surface materials and stratigraphy in cataract canyon, Utah. In R.O. Green, editor, *Proc. 14th AVIRIS Earth Science and Applications Workshop*, Pasadena, CA, May 24–27 2005.

[66] J. Rushing, R. Ramachandran, U. Nair, S. Graves, R. Welch, and H. Lin. ADaM: a data mining toolkit for scientists and engineers. *Computers and Geosciences*, 31:607–618, 2005.

[67] J. Schott, S. Brown, R. Raqueo, H. Gross, and G. Robinson. An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development. *Canadian Journal of Remote Sensing*, 25(2), June 1999.

[68] D. H. Scott and N. J. Trask. Geology of the Lunar Crater Volcanic Field, Nye County, Nevada. Technical report, USGS, 1971.

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

[70] M. K. Shepard. *Application of cosmogenic exposure age dating and remote sensing to studies of surficial processes*. PhD thesis, Washington University, St. Louis, MO, 1991.

[71] M. K. Shepard, R. E. Arvidson, E. A. Guinness, and D. W. Deering. Scattering behavior of Lunar Lake playa determined from PARABOLA bidirectional reflectance data. *J. Geophys. Res.*, 18:2241–2244, 1991.

[72] B. Solaiman and M. C. Mouchot. A comparative study of conventional and neural network classification of multispectral data. In *Proc. Int'l Geosci. and Remote Sensing Symposium*, volume III, pages 1413–1415, Caltech, Pasadena, CA, August 8–12 1994.

[73] P. H. Swain and S. M. Davis Eds. *Remote Sensing: The Quantitative Approach*. McGraw-Hill, New York, 1978.

[74] Swayze,G. A., R. N. Clark, A. F. H. Goetz, T. G. Chrien, and N. S. Gorelick, Effects of spectrometer band pass, sampling, and signal-to-noise ratio on spectral identification using the Tetracorder algorithm *J. Geoph. Research (Planets)*, 108(E9), 5105, doi: 10.1029/2002JE001975, 2003.

[75] M. F. Tenorio, S. R. Safavian, and 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*, volume II, pages 1289–1292, 1990.

[76] T. Villmann, E. Merényi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16:389–403, 2003.

[77] Th. Villmann, R. Der, M. Herrmann, and Th. Martinetz. Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.

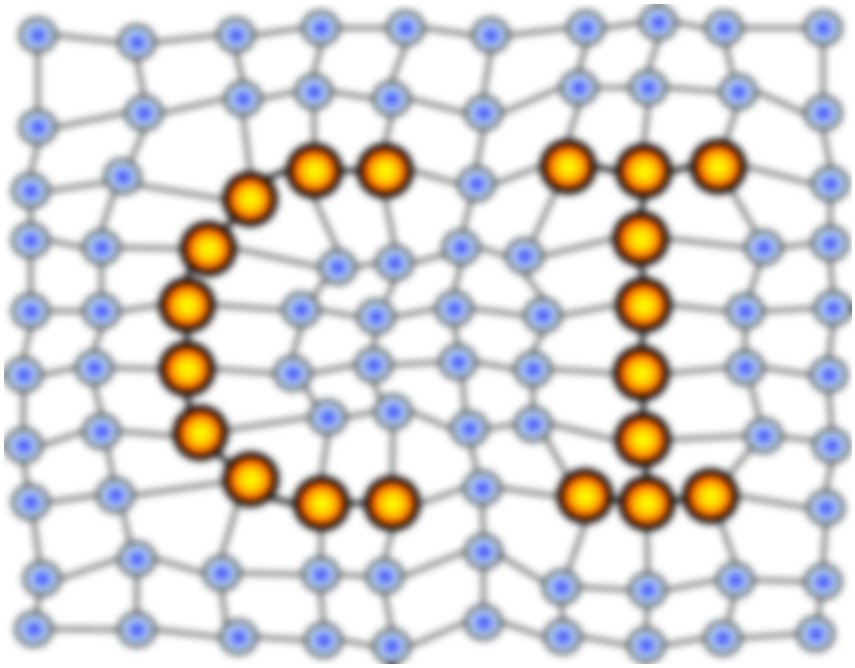
[78] Y. Wang and D. L. Civco. Artificial neural networks in high dimensional spatial data classification: A performance evaluation. In *ACMS/ASPRS Annual Convention & Exposition Technical Papers*, volume 3, pages 662–671, Charlotte, North Carolina, February 27 – March 2 1995.

[79] B. Widrow and F. W. Smith. Pattern-recognizing control systems. In *Computer and Information Science Symposium Proceedings*, Washington, D. C., 1963. Spartan Books.

[80] F. Wilcoxon. Probabilty tables for individual coparisons by ranking methods. *Biometrics*, 3(3):119–122, September 1947.

MACHINE LEARNING REPORTS

Report 04/2011



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Copyright & Licence

Copyright of the articles remains to the authors.

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.