DIMENSION REDUCTION WITH PRINCIPAL COMPONENT ANALYSIS IN HYPERSPECTRAL IMAGE CLASSIFICATION USING MACHINE LEARNING

Saziye Ozge Atik[1]
[1]Gebze Technical University, Cayirova Campus, Turkey
Email: soatik@gtu.edu.tr

KEY WORDS: Hyperspectral imagery, Dimension-reduction, Principal Component Analysis, Machine learning

ABSTRACT: Hyperspectral imagery technologies are one of the future trend investigations. In recent years, numerous studies have been applied to hyperspectral imagery for many purposes. Feature extraction is more complex and time-consuming when compared to other data sources. However, dimension reduction algorithms can help in processing and extracting this manner. Principal Component Analysis (PCA) algorithm has many advantages for hyperspectral imagery. In the study, PCA algorithm was used in the Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) methods for hyperspectral image classification on Pavia University dataset. The dataset includes nine classes asphalt, meadows, gravel, trees, painted metal sheets, bare soil, bitumen, self-blocking bricks, and shadows. The geometric resolution of the images is 1.3 meters. The study conducted different PCA band combinations using proper machine learning algorithm parameters. Different band combinations are used in the experiments as 25 and 50 bands. The results are compared quantitatively in the meaning of accuracy and time. General accuracies have been seen at over % 85 for two band combinations, too.

1. INTRODUCTION

The world has limited resources and human influence is one of the most influential factors in changing land use worldwide. Therefore, the strategic planning of sustainable applications and optimal management concerns require regular monitoring of changes in the environment and heterogeneous areas (Donmez and İpbuker, 2018). Furthermore, cities' use of land is impacted by globalization and rapid urban growth, and there is a growing need for automatic remote sensing image interpretation (Atik et al., 2022). The ability to distinguish and categorize a wide variety of land cover and use classes has been made feasible with the use of satellite technology using hyperspectral sensors. Its applications have become widespread in image processing, analysis and classification using hundreds of bands. In recent years, machine learning algorithms have offered extremely high performance for classification processes. On the other hand, as the size of the data used in the studies increases with the number of bands, the duration of the analysis increases and can be more difficult for the processors. For this reason, steps can be taken to extract as much meaningful information as possible from the data and to reduce the size. In this way, existing size reduction methods are used to improve the analysis of the feature information to be extracted from the data, as well as to increase the ease and efficiency of data processing. In this study, three different machine learning algorithms and Pavia University hyperspectral dataset were classified with support vector machines, random forest, and multilayer perceptron algorithms, using datasets reduced to different dimensions, in order to determine the dimension reduction using feature extraction technique at the appropriate scale. There are various related studies in the literature for image analysis (Rodarmel and Shan, 2002; Prasad and Bruce, 2008; Machidon et al., 2020; Atik and Atik, 2021). In this study, three different machine learning algorithms and Pavia University hyperspectral dataset were classified with support vector machines, random forest, and multilayer perceptron algorithms, using datasets reduced to different dimensions, in order to determine the size reduction at the appropriate scale.

2. DATA AND METHODOLOGY

Pavia University dataset has 103 bands and was acquired by the ROSIS Sensor over Pavia, northern Italy. The dataset is a 610 x 610 pixels image and has nine classes asphalt, meadows, gravel, trees, painted metal sheets, bare soil, bitumen, self-blocking bricks, and shadows. In Figure 1, the hyperspectral image is shown with ground truth. The number of class-labeled samples is presented in Table 1. Pavia scenes belonged to the Telecommunications and Remote Sensing Laboratory, Pavia University by Prof. Paolo Gamba.

Table 1. Class distribution of the Pavia University dataset.

Legend Code	Class Names	Labeled Samples
0	Background	
1	Asphalt	6 852
2	Meadows	18 686
3	Gravel	2 207
4	Trees	3 436
5	Metal sheets	1 378
6	Bare soil	5 104
7	Bitumen	1 356
8	Bricks	3 878
9	Shadows	1 026

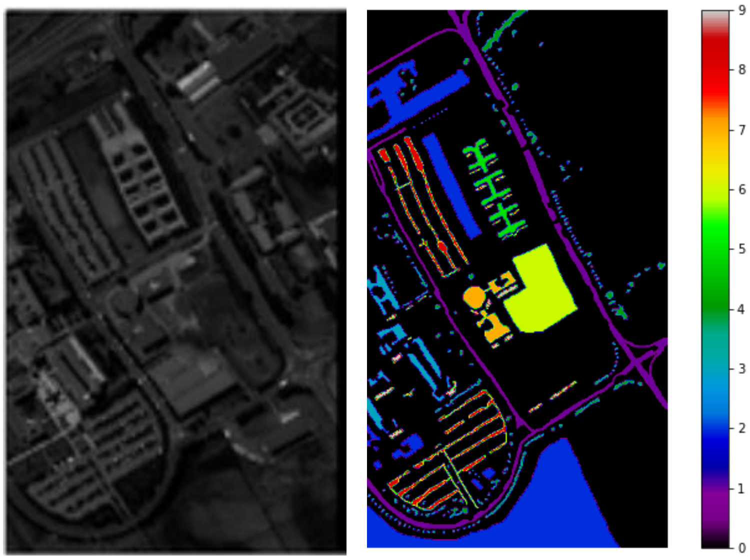

Figure 1. Sample image from Pavia University Dataset. Image on the left and ground truth on the right.

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) (Cunningham, 2008) is a widely used size reduction technique. The dimensions of hyperspectral images consist of hundreds of bands, making it difficult to process all the data for controlled classification. PCA uses the statistical properties of hyperspectral bands to examine the correlation of bands with each other (Rodarmel and Shan, 2002). Therefore, size reduction is an essential preprocessing step in hyperspectral image analysis (Deepa and Thilagavathi, 2015). Various analyzes of hyperspectral images can be performed with size reduction methods (Agarwal et al., 2007) and the benefit and effectiveness of using this technique as a preprocessing step are recommended (Cunningham, 2008).

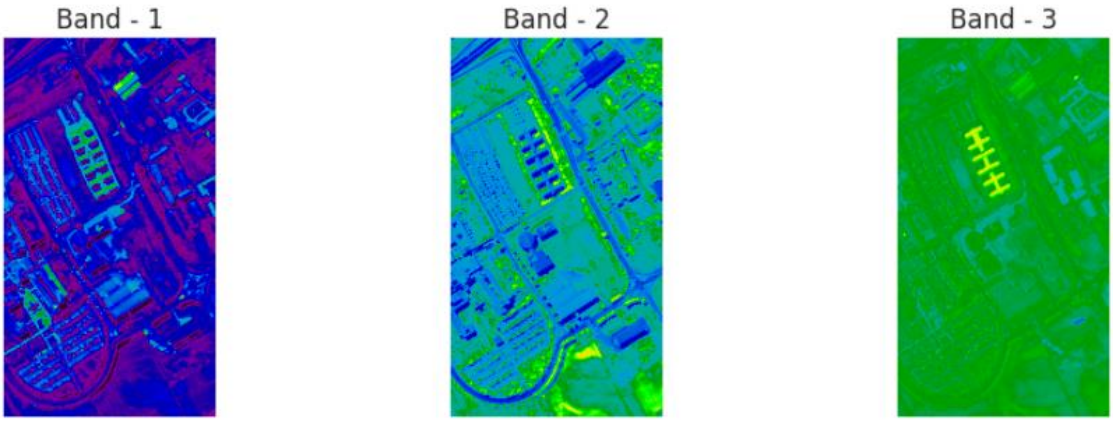

Figure 2. PCA Band Samples of Pavia University Dataset

2.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) is a controlled machine learning algorithm used for regression and classification operations. In SVM, the goal is to create a hyperplane that classifies the points in the data separately in an area of n features. There are many hyperplanes that can be used to separate multiple classes of data points. The SVM algorithm is aimed to find the plane with the maximum margin and the distance between the data points of both classes. Thus, a very accurate classification will be realized in the classification of new future data. The input variable x is shown in Equation 1.

$$x_i(i = 1,2, \ldots , n) \tag{1}$$

Mathematical expression of the hyperplane can be calculated with the help of the function in Equation 2, where the variable w corresponds to an n-dimensional vector and b the parameter of the hyperplane is a scalar number.

$$f(x) = w^T x + b = \sum\nolimits_{j=1}^{N} w_j x_j + b = 0 \tag{2}$$

2.3 Random Forest (RF)

Random forest (Breiman, 2001) is a non-parametric classification and regression algorithm. In the algorithm, the input space is subdivided into portions and calculated parameters for each portion (Atik et al., 2021). Each tree in the random forest predicts a class, and the class with the highest votes determines the prediction made by the model. The trees created are independent of each other and are voted, used for prediction in the algorithm that gets the largest vote value (Akar and Güngör, 2012). To generate a tree with the RF classifier, 2 user-defined parameters are required. According to the values obtained, the tree is developed without pruning (Akar and Güngör, 2012). The error obtained as a result of this part is named the generalized error. The generalized error calculation is shown in Equation 3:

$$PE^* = P_{X,Y}(mg(X, Y) < 0) \tag{3}$$

In the equation, mg() refers to the margin function. Margin measures how much the average number of votes in (X, Y) for the correct class exceeds the average number of votes for any other class. The larger the margin, the more reliable the classification can be (Breiman, 2001).

2.4 Multilayer Perceptron (MLP)

A sensor with only one weight layer can only process linear input functions. Nonlinear functions cannot be successfully used to produce results. Multi-layer perceptrons, if used for classification, can solve such nonlinear problems. (Alpaydin,2010). There is an input layer, a hidden layer, and an output layer in the MLP algorithm, which can also be used as a deep learning method.

3. RESULTS AND DISCUSSION

In this study, the number of features in hyperspectral data is reduced by using PCA, which is a size reduction technique, as a preprocessing step. At this stage, classification tests were carried out to create a data set with three different band numbers. The data, consisting of 103 bands in total, were classified separately with the help of different set sizes, SVM, RF and MLP machine learning algorithms, 50 and 75 bands (Figure 5). On the data set, the variance graphs explained cumulatively with the selection of the number of bands using the PCA technique were observed (Figure 3).

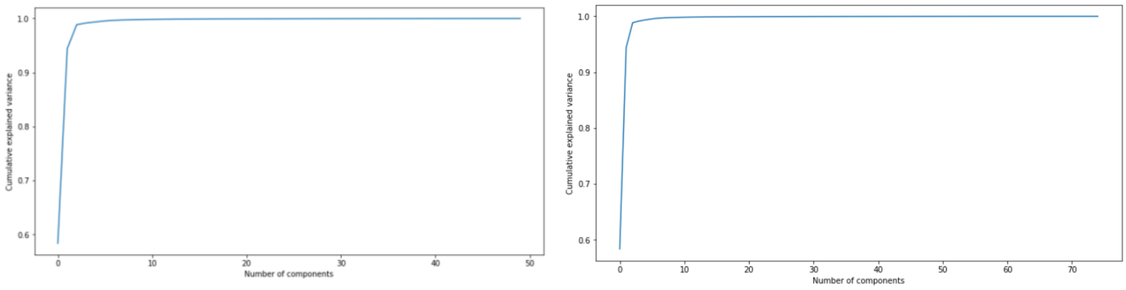

Figure 3. Cumulative curve of variance with PCA in 50 AND 75 band data set

In this study, 20% of the data set was randomly allocated as test data. Since the data set used has an unbalanced class weight, the parameters of the fit SVM algorithm were applied following the class weights while the SVM algorithm was used. In the SVM algorithm, the radial basis function is used, at this stage, the parameter C is set as 100. In the application of the RF method, the number of estimation parameters n was chosen as 200, the depth was determined as 100 and the minimum number of trees was determined as 25. In the classification made by the MLP method, the alpha value was optimized as 0.0001 and the maximum iteration was applied as 1500.

The overall accuracy (OA) formula is shown in equation 4. An estimate of OA is the proportion of correctly identified pixels (Atik and İpbüker, 2021). In Table 2 overall accuracy values are shown as %.

$$OA = \Sigma_{i=1}^{K} \frac{Nii}{N} \tag{4}$$

Table 2. Overall Accuracy of Models with Different Band Numbers

Model	50 PCA Bands	75 PCA Bands
SVM	**95,86**	**95,87**
RF	**88,52**	**85,94**
MLP	93,22	92,05

SVM algorithm produced the highest classification accuracy with 95.86% in the 50-band dataset and 95,87% in the 75-band dataset. Then, the MLP method reached a higher classification accuracy of 92.05% in the 75-band dataset compared to 50-band dataset with 93.22%. The method with the lowest classification accuracy is the RF algorithm. In the RF method, the lowest classification accuracy among the tests performed on the 75-band dataset was 85.94%. The highest accuracy of the RF algorithm was obtained as 88.52% in the 50-band data set.

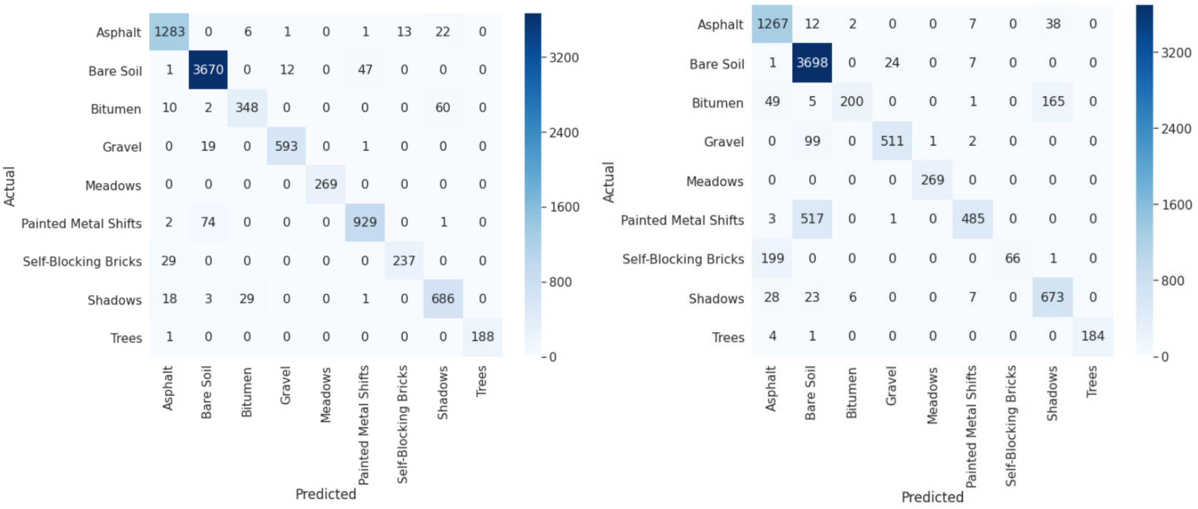

Figure 4. Error matrix of SVM method (left) and RF method (right) experiments in 75 band dataset.

In the applications, the highest classification accuracy was found to be 95.87% with the SVM method with 75 bands (Figure 4). The lowest classification accuracy was obtained with the RF method with 85.94% with 75 bands (Figure 4). According to these results, different machine learning algorithms have different classification accuracies in the hyperspectral data set used. It has been observed that the size difference of the data set and the effect of the feature information on the classification accuracy according to the machine learning algorithm are not at the same rate. In this case, while determining the ideal data size, tests should be done according to the machine learning algorithm to be used. In future projections, several selected classes can also be studied for specific purposes.

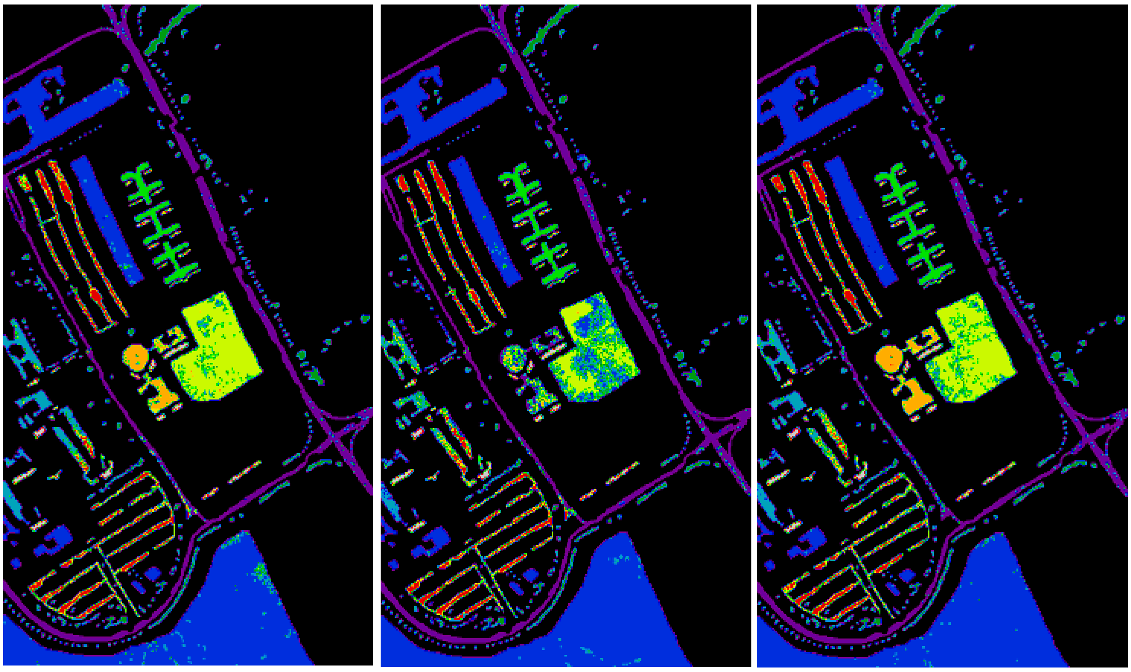

Figure 5. PCA 50 Bands of SVM, RF and MLP algorithms by order

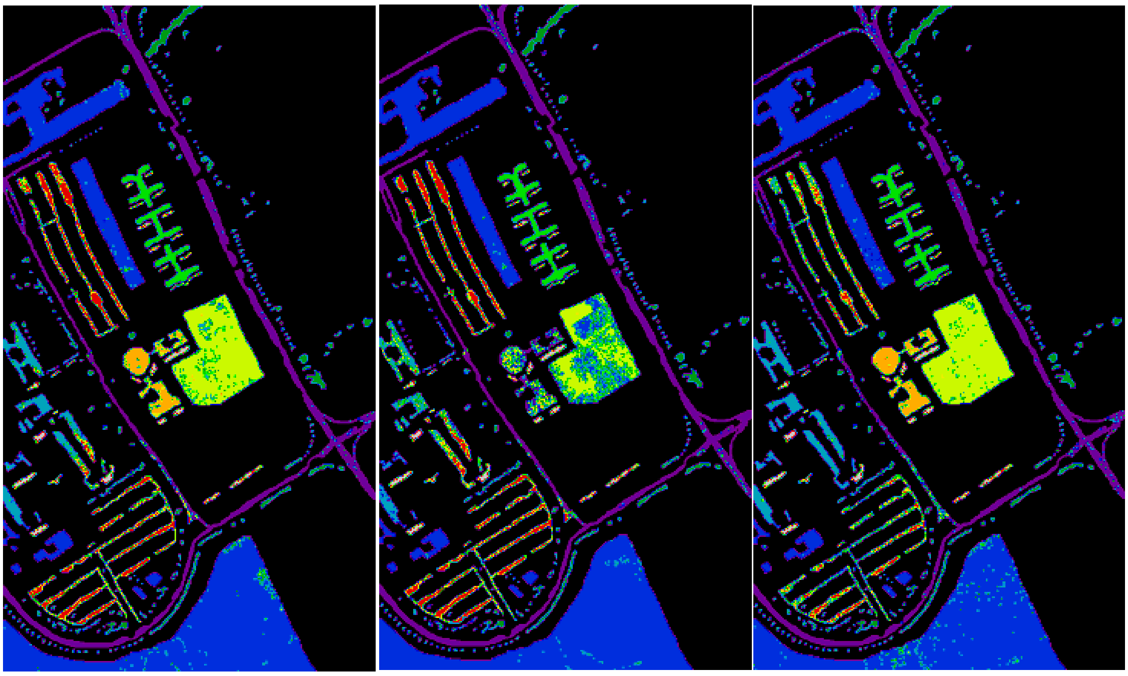

Figure 5. PCA 75 Bands of SVM, RF and MLP algorithms by order

REFERENCES

Agarwal A., El-Ghazawi T., El-Askary H., Le-Moigne, J. 2007. Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral Imagery", Signal Processing and Information Technology 2007 IEEE International Symposium on, pp. 353-356.

Akar, Ö., Güngör, O. 2012. Classification of multispectral images using Random Forest algorithm. Journal of Geodesy and Geoinformation, 1(2), pp. 105-112.

Alpaydin, E. 2010. Introduction to machine learning (2nd Edition). MIT press.

Atik, M. E., Duran, Z., Seker, D. Z. 2021. Machine learning-based supervised classification of point clouds using multiscale geometric features. ISPRS International Journal of Geo-Information, 10(3), 187.

Atik, S. O., Atik M. E. 2021. PRINCIPAL COMPONENT ANALYSIS APPROACH IN HYPERSPECTRAL IMAGE CLASSIFICATION WITH MACHINE LEARNING METHODS. International Symposium on Applied Geoinformatics. Riga Latvia, 2021.

Atik, S. O., Atik, M. E., Ipbuker, C. 2022. Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. Journal of Applied Remote Sensing, 16(2), 024510.

Atik, S. O., Ipbuker, C. 2021. Integrating convolutional neural network and multiresolution segmentation for land cover and land use mapping using satellite imagery. Applied Sciences, 11(12), pp. 5551.

Breiman, L. 2001. Random forests, Machine learning, 45(1), pp. 5-32.

Cortes, C., Vapnik, V. 1995. Support-vector networks, Machine learning, 20(3), pp. 273-297.

Cunningham, P. 2008. Dimension reduction. In Machine learning techniques for multimedia. Springer, Berlin, Heidelberg, pp. 91-112.

Deepa, P., and K. Thilagavathi. 2015. Feature extraction of hyperspectral image using principal component analysis and folded-principal component analysis. 2nd International Conference on Electronics and Communication Systems (ICECS). IEEE.

Donmez, S.O.; Ipbuker, C. 2018. Investigation on Agent Based Models for Image Classification of Land Use and Land Cover Maps. In Proceedings of the 39th Asian Conference on Remote Sensing (ACRS): Remote Sensing Enabling Prosperity, Kuala Lumpur, Malaysia, 15–19 October 2018; pp. 2005–2008.

Machidon, A. L., Del Frate, F., Picchiani, M., Machidon, O. M., Ogrutan, P. L. 2020. Geometrical approximated principal component analysis for hyperspectral image analysis. Remote Sensing, 12(11), 1698.

Prasad, S., Bruce, L. M. 2008. Limitations of principal components analysis for hyperspectral target recognition. IEEE Geoscience and Remote Sensing Letters, 5(4), pp. 625-629.

Rodarmel, C., Shan J. 2002. Principal component analysis for hyperspectral image classification. Surveying and Land Information Science 62.2, 2002, pp. 115-122.