Infinite Latent Feature Selection Technique for Hyperspectral Image Classification

Tajul Miftahushudur [a,b,*], Chaeriah Bin Ali Wael [b], Teguh Praludi [b]

[a] *School of Electrical and Electronics Engineering*
The University of Manchester
Manchester, United Kingdom
[b] *Research Center for Electronics and Telecommunication*
Indonesian Institute of Sciences (P2ET-LIPI)
Komplek LIPI Gedung 20 lantai 4, Jl Sangkuriang Cisitu - 40135
Bandung, Indonesia

Abstract

The classification process is one of the most crucial processes in hyperspectral imaging. One of the limitations in classification process using machine learning technique is its complexities, where hyperspectral image format has a thousand band that can be used as a feature for learning purpose. This paper presents a comparison between two feature selection technique based on probability approach that not only can tackle the problem, but also improve accuracy. Infinite Latent Feature Selection (ILFS) and Relief Techniques are implemented in a hyperspectral image to select the most important feature or band before applied in Support Vector Machine (SVM). The result showed ILFS technique can improve classification accuracy better than Relief (92.21% vs. 88.10%). However, Relief can extract less feature to reach its best accuracy with only 6 features compared with ILFS with 9.

Keywords: classification, feature selection, hyperspectral carbon, Infinite Latent Feature Selection (ILFS), SVM.

I. INTRODUCTION

Hyperspectral imaging (HSI) is one of the remote sensing technique that has been widely used in various application such as military [1], agriculture [2], medical [3], earth observation [4], and others. One of the advantages of hyperspectral images against multispectral images is in its ability to acquire and measure the reflectance from the object material. A multispectral sensor is only able to measure on some wide wavelength bands where there is a gap acquisition which may cause missing some essential measurable data. On the other hand, hyperspectral sensors can acquire and measure reflectance radiation continuously.

The spectrum in a single pixel of hyperspectral images has a similar format to a spectrum that is measured in a spectroscopy laboratory. With this format, the pixel can provide more information. Also, the hyperspectral system has better samples in spectral and spatial domains. Therefore, hyperspectral images have more potential to present more accurate and detailed information than other remote sensing techniques [12].

Another advantage of hyperspectral imaging system compared with another imaging system is its ability to cover hundreds of bands of the electromagnetic wave. This ability makes hyperspectral imaging can store larger information of spectral signature or feature of an object. With this large information, HSI can produce great accuracy to classifying or distinguish object material.
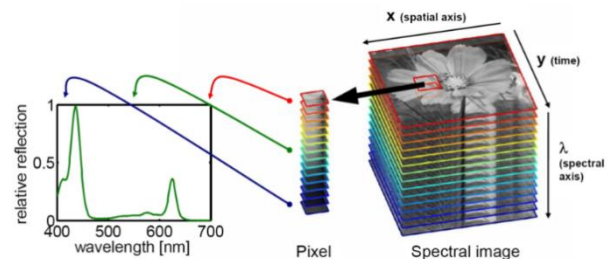
Figure 1. Illustration of Hyperspectral cube.

Figure 1 shows illustration that HSI system consists of a hundred of images corresponding to the certain wavelength on the same surface. Therefore, each pixel in the HSI represents a reflection or radiation on certain wavelength range.

As the consequences of the large data acquisition, computational complexity and data storage issue have made a new challenge to researchers in this field to design new approaches to tackle this problem. With the large number of spectral features that obtained in the HSI, sometimes only several feature channel that have contribution during the classification process. Eliminating the less influencing feature has potential to tackle those issues and reduce the high dimension of HSI. Furthermore, two common techniques were performed by the researcher to reduce dimensions in HSI are feature extraction [5]-[6] and feature selection [7]-[8]. In this experiment, feature selection technique is used because it is easy and relatively safe to reduce dimensionality as it does not change the spectral information like feature extraction method does. Therefore, it can reduce the high dimension of HSI.

The feature selection technique has been widely implemented in the hyperspectral image in recent years. The Steepest Ascent (S.A) search strategy [5] is used by Serpico to find the most important features which represented in discrete binary space. The results of this study show that the accuracy is better than the Sequential Feature Selector (SFS) algorithm, but in the trade-off, this technique generates a higher computation process. In the same paper, Fast Constrained Search (FCS) provides a better tradeoff than SA. Another fast feature selection technique on HSI was presented by [9].

In that paper, the Gaussian Mixture Model (GMM) was implemented to select important features before being processed in the forward feature selection. The results of this study are compared with the feature selection technique using SVM and show that the implementation of GMM can produce fewer important features than SVM with the accuracy of the same classification as SVM. Comparison of the three different feature selection approach in HSI is presented by [8]. Those experiment evaluating wrapper approach (combination recursive elimination-SVM technique), filters (correlated based & minimum redundancy technique) and random forest. This study shows that the filter approach has superiority against the wrapper technique. The author proved that the random forest approach has the best level of accuracy rather than the other approaches being tested. In conclusion, all of the previous studies show that the implementation of feature selection not only speeds up computational processing but also increases the accuracy of classification result.

In this study, the use of feature selection using a probabilistic graph approach will be tested to reduce high dimensional of hyperspectral cube. The dataset in a hyperspectral cube format is converted into a map that correspondence between its pixel and the reflectance value to make feature extraction process easier. The next step is to implement a technique feature selection with Infinite Latent Feature Selection (ILFS) [10] to select the most influential feature in the surface recognition process. The results are then used as input for the learning and testing process using a Support Vector Machine (SVM). One of the purposes of this experiment is to implement ILFS into hyperspectral data format, which is 3-dimensional data with very large spectral data. The result is then compared with other techniques if implemented in HSI dataset.

The outline of this paper is organized as follow: section I present an introduction and problem on hyperspectral images. Theoretical background to tackle the problem are explained in section II. Detailed dataset and experiment method are described in section III. Finally, section IV and V contain experiments result and conclusion respectively.

II. THEORETICAL BACKGROUND

A. Dimension Reduction

The dimension reduction technique aims to reduce the number of dimensions of data processing without affecting the processing results. There are two commonly used techniques: feature extraction and feature selection. The difference between the two is that the results in the feature extraction are data with fewer new features, where each new feature is a modified result from the old feature. Whereas in the feature selection, features that have small importance weights will be eliminated and leaving only some features which have the greatest importance weight.

Feature selection technique can be divided into 2 types: filter and wrapper [11]. The difference between the two methods is its procedure to select the best subset. In the filter method, subset is selected independently [12], while in wrapper, an interaction between subset and machine learning algorithm occur recursively, so that makes computation problem. Information gain (IG) [13], chi-square [14], and log like ratio examples of feature selection with the filter types. As for the wrapper type is forward and backward selection [11], relief [15], and infinite latent feature selection [15].

B. Infinite Latent Feature Selection (ILFS)

ILFS technique consists of 3 main processes. The first one is the preprocessing step, then weighting graph and the last one is ranking. The pre-processing stage has a goal to quantify the distribution of features x_i in the matrix format. Then calculate the value for a specific token, so that each feature x_i can be repressed by the token t. This process is called discriminative quantization. The Fisher criterion method is used to calculate vectors from a feature. The next step is graph weighting. Weighting process has purpose to create a fully connected graph in each node that connecting each feature with the other feature. The weights can be calculated with probability co-occurrences between features and tokens using PLSA technique [16] Finally, the weight is optimized using the Expectation-Maximization (EM) algorithm. Further details about ILFS can refer to [10]. The detailed process of ILFS technique is described as follow:

1) Discriminative Quantization Process

The goal of this process is to produce a matrix that represents how well the given feature represents a class before performing many-to-few mapping. First, a modified Fisher criterion that can handle the scoring vector from multi-class framework is given by (1). K is number of features, s is sample, μ is mean, σ is standard deviation, Z is normalization factor and $\emptyset$ is Score vector.

$$\emptyset = \frac{1}{Z}\left[\frac{(s-\mu_1)^2}{\sum_{k=1}^{K}\sigma_k^2}, \dots, \frac{(s-\mu_K)^2}{\sum_{k=1}^{K}\sigma_k^2}\right], \forall_{k \in K} \quad (1)$$

Next step is performing quantization to calculate the value of discriminative feature in particular "token" interval (smaller value represented not well samples vice versa).

2) From co-occurrences to graph weighting

Graph weighting calculated based training data according to the degree of relevance/importance. Conditional probabilities of $P(token|feature)$ and $P(factor|feature)$ are calculated. Furthermore, to optimize this parameter, an Expectation-Maximization (EM): maximum likelihood is performed.

3) Probabilistic Infinite Feature Selection

Matrix that has been obtained from the previous procedure is then calculated its geometric series to expand its path into infinity. In these experiments, Gelfand's formula is used [17].

$$C = (I - rA)^{-1} - I \qquad (2)$$

where A is previous matrix obtained from process 2, I is matrix identity of A and r is obtained by (3).

$$r = \frac{1}{\max(eig(A))} \qquad (3)$$

Finally, energy scores of the path length are calculated by summing the dimension of the matrix, then rank its correspondence with the feature.

C. Support Vector Machine (SVM)

The idea of SVM is to generate the best hyperplane that can separate two classes. An illustration of class separation using hyperplane is shown in Figure 2. To obtain the most optimum hyperplane, it is needed to calculate hyperplane margin to get its maximum point. Hyperplane margin is the distance between the closest pattern in each class. The process of finding the best hyperplane is the goal of the learning process in the SVM.

To classify data that cannot be completed with a linear hyperplane, the boundary plane needs to be more flexible as shown in figure 2(b). A flexible boundary plane can be obtained using one of the kernel functions as shown in Table 1. Inspired by previous research [18], kernel function using the RBF approach shows the best performance in land cover classification. In this experiment, we set an initial parameter of RBF with *gamma* 10 and penalty factor C 0.1. Whereas $|x_i \, x_j|^2$ is euclidean distance of the two different feature space.

III. DATA AND METHOD

A. Test Area

In this experiment, hyperspectral dataset form instrument NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) is used. Using sample image from Kennedy Space Center Florida from 20 km altitude, it is acquiring 176 bands with wavelength from 400 nm to 25000 nm. For surface classification purpose, 13 different classes are used to representing a class of an object's surface in the image. The detail of thirteen different class is explained in Table 2. Furthermore, the representation of Kennedy space center in RGB format and pixel distribution of training/testing pixel sample is shown in Figure 3. The dataset can be download from [19].

B. Method

To simplify feature analysis of each pixel in the HSI, the HSI cube structure needs to be converted into a 2D map. This process is illustrated in Figure 4. Consider a material on the pixel (x, y) on the truth label represent 'river', then the features or spectral on hyperspectral cube on the same pixel coordinate are extracted to the map array.
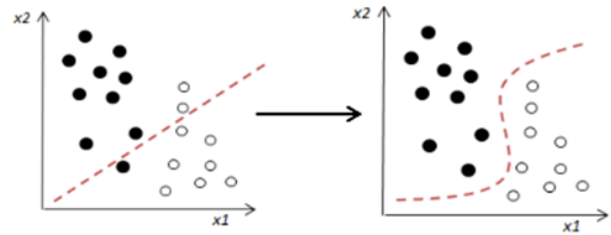

Figure 2. Non-linear problem on SVM

TABLE 1
SVM KERNEL

Kernel	Definition
Linear	$K(\vec{x_i}, \vec{x_j}) = x_i^T \, x_j$
Polynomial	$K(\vec{x_i}, \vec{x_j}) = (\gamma x_i^T \, x_j + r)^d, \gamma > 0$
Radial Basis Function (RBF)	$K(\vec{x_i}, \vec{x_j}) = \exp(-\gamma \, |x_i \, x_j|^2), \gamma > 0$
Sigmoid	$K(\vec{x_i}, \vec{x_j}) = \tanh(\gamma . x_i^T \, x_j + r)$

TABLE 2
CLASSIFICATION AND NUMBER OF PIXEL AMPLE

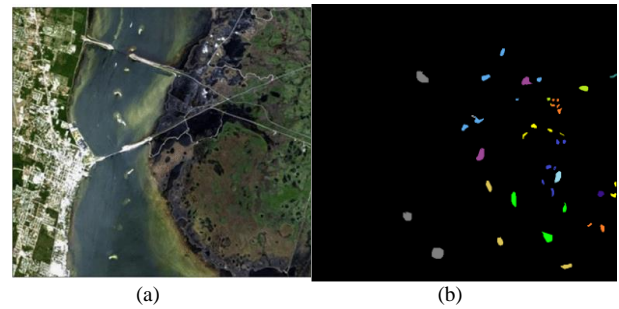| No.	Class	Number of Pixels
1.	Scrub	244
2.	Willow swamp	257
3.	Cabbage palm hammock	253
4.	Cabbage palm/oak hammock	162
5.	Slash pine	230
6.	Oak/broadleaf hammock	106
7.	Hardwood swamp	432
8.	Graminoid marsh	521
9.	Spartina marsh	405
10.	Cattail marsh	420
11.	Saltmarsh	504
12.	Mudflats	928
13.	Water	749
	Total	5211

| (a) | (b) |

Figure 3. (a) Kennedy Space Center in RGB Format, and (b) distribution of the sample class.

So that the spectral map will store the label in first column, followed spectral value from particular wavelength range in the next column. Evaluation phase in this experiment consists of two scenarios. Firstly, evaluation of comparing accuracy result using SVM on several different numbers of training dataset condition. And secondly, comparing accuracy result with different feature selection methods. The experiment detail that has been done is shown in Figure 6.

In these experiments, classification result is executed using software data processing called MATLAB with

LIBSVM Library [20] to generate an SVM model. LIBSVM can support several kinds of learning model to solve classification and regression problem like C-SVC, nu-SVC, one-class SVM, epsilon-SVR and nu-SVR.

Furthermore, this library already supports multi-class classification. Some kernel options like linear, polynomial, radial basis function, and sigmoid is also provided in this library. The use of LIBSVM for classification is consist of two steps. Firstly, the training process to create the SVM model. In this step, the SVM model and kernel type can be chosen as well its parameters like C and gamma. Secondly, the testing process to perform the classification process based on the testing input. To perform both steps, training and testing input should be in a standard LIBSVM format and stored in a text file ('.txt'). Dataset format that supports LIBSVM is illustrated in Figure 4. Furthermore, ILFS and Relief are implemented using feature selection library called FSLib [10], [21].

Mapping process forms hyperspectral cube to LIBSVM format is shown in Figure 5. This data structure can be explained as follow: say material on the pixel (x,y) represents 'wall', then feature (reflectance spectrum) of this pixel is extracted on the array with the reflectance at the sampled wavelength as its value. A label 'wall' then attached to the first array. So, the matrix size of the extracted dataset is [sampled pixels x (1+number of sampled wavelength)].

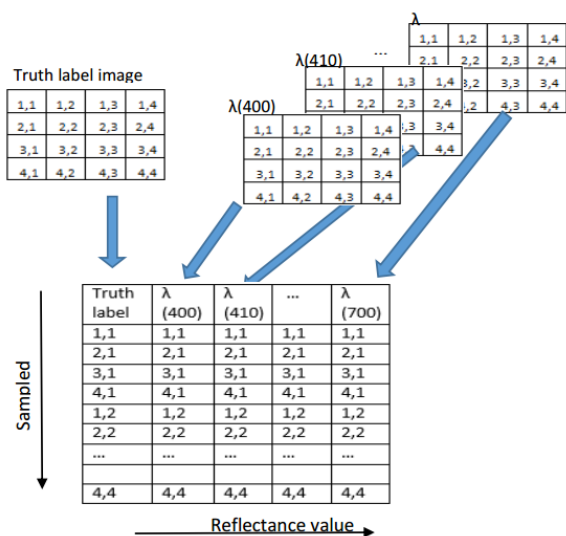Figure 4. LIBSVM structure data

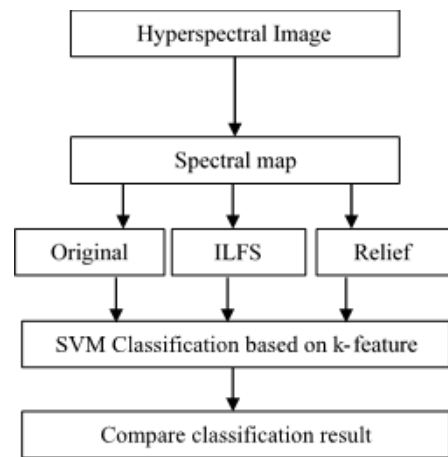Figure 5. Mapping process from Hyperspectral cube into spectral map

Figure 6. General method

IV. RESULTS AND DISCUSSION

The first step needs to be done in this experiment is spectral extraction. Figure 7 shows the extracted spectrum of the 13 sample class. From the figure, it can be seen that every spectral signature has different spectral power. One spectral have very high amplitude and the other has very low spectral power. This phenomenon might occur because of reflectance of the surface of the object material. Light from the sun and shadow from the cloud have important influence in this matter. The brighter reflectance the high reflectance spectral power. Vice versa, darker surface material the lowers spectral amplitude.

After the spectral extraction process is done, the classification results in this experiment were tested in several scenarios to explore the effect of the features number in terms of classification accuracy. The first experiment aims to determine the effect of the number of features used in classification without changing the sequence of the bandwidth. The features chosen in this experiment are ordered from the shortest wavelength to longest wavelength. The results of this experiment can be seen in the following table.

From Figure 8 and Table 3, it can be seen that reducing features number from 176 to 11 can produce better accuracy rather than using the whole features. Besides, the classification accuracy looks constant after implements 21 or more features. The next experiment compares the results of the feature selection after the feature is sorted by its importance with the ILFS and Relief algorithm. Similar to the previous experiment, the testing phase was done with a different number of training data. The results of this experiment are shown in the following table.

Figure 9 and Table 4 show the fact that the ranking features using the feature selection technique can improve classification accuracy even with a smaller number of features than the previous experiment. Furthermore, the ILFS algorithm shows better performance than Relief because from the graph it can be seen that ILFS can find features with importance weight better than relief. The graph shows the first feature selected by ILFS produces better performance than Relief, and it increases constantly without a significant jump which shows a large distance of the importance of weight value.
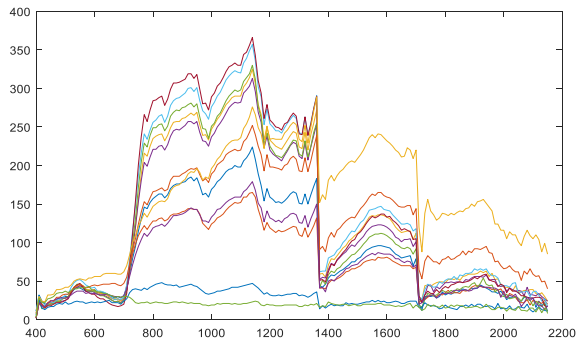
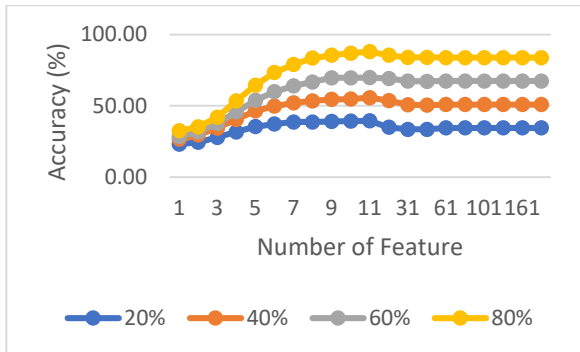Figure 7. spectral signature of the sampled training data

Figure 8. Classification result with increment feature of a number without feature selection
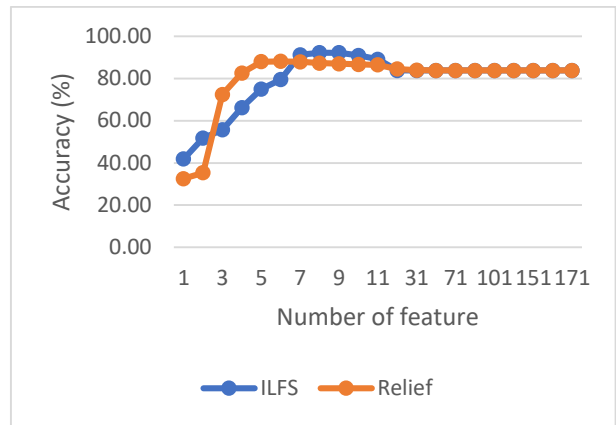

Figure 9. Comparison of Feature selection using ILFS and Relief

Overall, ILFS shows its superiority against Relief because of its ability to produce a better combination of most important feature rather than the Relief in the top rank. Besides, the accuracy graph tends to be constant after more than 11 features used, this phenomena shows there is an overfitting problem while using a lot of feature in HSI. Overfitting occurs due to huge number of parameter in training phase to fitting the machine learning model. This problem may cause a random noise instead of fit relationship between the feature and the desired output. Consequently, the model performs worse rather than using a smaller number of feature but have significant discriminant.

TABLE 3
COMPARISON RESULT WITH DIFFERENT NUMBER OF TRAINING SAMPLE

	20%	40%	60%	80%
Max accuracy	39.45 %	55.68%	69.7 %	87.95
(number of features)	(11)	(11)	(11)	(11)
Min accuracy	23.20 %	26.6 %	28.5 &	32.51
(number of feature)	(1)	(1)	(1)	(1)
Difference	16.25 %	29.05 %	41.16	55.44
Accuracy with full feature	34.45 %	50.89 %	67.35 %	83.78

TABLE 4
COMPARISON OF CLASSIFICATION RESULT

Method	20%		40%		60%		80%	
	Acc(%)	# feature	Acc(%)	# feature	Acc(%)	# feature	Acc(%)	# feature
ILFS	39,59	5	67,45	6	76,67	6	92,21	9
Relief	39,63	5	58,05	5	74,38	5	88,10	6
No	39,45	11	55,69	11	69,76	11	87,95	11

TABLE 5
FEATURE SELECTION RESULT

Method	Training Set Size			
	20%	40%	60%	80%
ILFS	1,11,2,12,42	99,100,101,103,98, 137	163,158,164,159, 161,162	18,27,29,26,31,30, 162
Relief	1,2,3,15,9	1,2,3,34,35	1,2,3,34,35,15	1,2,34,3,35,33

CONCLUSION

The classification results in the experiment show that choosing the correct features for classification purpose using machine learning can improve the performance and accuracy of the classification results. Two feature selection methods (ILFS and Relief) has been tested in this experiment. Both techniques show better performance than the general classification technique using SVM with original feature. Furthermore, ILFS shows its superiority towards Relief since it can provide greater maximum accuracy of 92.21% rather than relief which only gives 88.10%. However, the weakness of this algorithm compared with relief is that the ILFS need slightly more feature to reach its maximum accuracy. For further development, ILFS need to be optimized so that the feature selection can produce maximum accuracy with fewer feature numbers.

Table 3 shows that utilizes 1042 training data only produce 34.45% compared with 4168 training data can boost the accuracy to 83.78%. From this phenomena, it can be seen that the lack of performance of machine learning is affected by the number of training data. Therefore, further research to multiply the number of limited training data needs to be done. Moreover, the higher and the lowest spectral amplitude has a significant effect on machine learning performance. The high amplitude of the spectral makes the spectral feature dominant. Otherwise, spectral feature with low amplitude has a less significant effect on the training phase. How to make a range of spectral power uniform need to without affecting the most significant wavelength feature is interesting to be explored. The non-uniform spectral power has significant effect on machine learning.

REFERENCES

[1] X. Briottet *et al.*, "Military applications of hyperspectral imagery," vol. 6239, p. 62390B, 2006.

[2] L. M. Dale *et al.*, "Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: A review," *Appl. Spectrosc. Rev.*, vol. 48, no. 2, pp. 142–159, 2013.

[3] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *J. Biomed. Opt.*, vol. 19, no. 1, p. 010901, 2014.

[4] J. Transon, R. d'Andrimont, A. Maugnard, and P. Defourny, "Survey of hyperspectral Earth Observation applications from space in the Sentinel-2 context," *Remote Sens.*, vol. 10, no. 2, pp. 1–32, 2018.

[5] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, 2007.

[6] J. F. R. Rochac and N. Zhang, "Feature Extraction in Hyperspectral Imaging Using Adaptive Feature Selection Approach," pp. 36–40, 2016.

[7] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, 2001.

[8] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, 2010.

[9] M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, "Fast forward feature selection for the nonlinear classification of hyperspectral images," Jan. 2015, [online] Available: https://arxiv.org/abs/1501.00857.

[10] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *Proc. of 2016 Int. Conf. Eng. MIS, (ICEMIS 2016)*, 2016.

[11] N. Sánchez-Maroño, A Alonso-Betanzos, M Tombilla-Sanromán, "Filter methods for feature selection - A comparative study," in *Proc. of the Eighth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2007)*, 2007, pp. 178-187.

[12] S. Lei, "A feature selection method based on information gain and genetic algorithm," in *Proc. of 2012 International Conference on Computer Science and Electronics Engineering*, 2012, vol. 2, pp. 355-358.

[13] P. Meesad, P. Boonrawd, and V. Nuipian, "A chi-square-test for word importance differentiation in text classification," in *Proc. of International Conference on Information and Electronics Engineering*, 2011, pp. 110–114.

[14] R.J. Urbanowicz, M. Meeker, R.S. Olsona, L.C. William, and H.M. Jason, "Relief-based feature selection: introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189-203, September 2018.

[15] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite latent feature selection : a probabilistic latent graph-based ranking approach," in *Proc. of International Conference on Computer Vision*, 2017, pp. 1398-1406.

[16] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proc. of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289-296.

[17] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, Aug. 2004.

[18] C. Chang and C. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, 2013.

[19] L. Wenzhi, "Feature Extraction and Classification for Hyperspectral Remote Sensing Images," Doctoral dissertation, Universiteit Gent, 2012.

[20] G. Roffo, "Ranking to Learn and Learning to Rank: On the Role of Ranking in Pattern Recognition Applications," June 2017, [online] Available: https://arxiv.org/pdf/1706.05933.

[21] Hyperspectral Remote Sensing Scenes http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes