Mesopotamian journal of Big Data
Vol. **(2023)**, 2023, **pp**. 53-60
DOI: https://doi.org/10.58496/MJBD/2023/008 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Active learning with Bayesian CNN using the BALD method for Hyperspectral Image Classification

Mahmood Siddeeq Qadir[1,*], Gokhan BİLGİN[1]

[1] *Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey*

ARTICLE INFO

Article History
Received 17 May 2023
Accepted 05 Jun 2023
Published 7 June. 2023

Keywords
Deep learning
Hyperspectral
Active Learning
Big Data
Bayesian
Convolutional Neural Networks

ABSTRACT

Deep learning DL techniques have recently been used to examine the classification of remote sensing data like hyperspectral images HSI. However, DL models are difficult to obtain since they rely largely on a large number of labeled training data. Therefore, a current challenge in the field of HSI classification is how to effectively incorporate DL models in constrained labeled data. The Bayesian Convolutional Neural Networks BCNN method is robust against overfitting on small datasets. One of the key methods for automating data selection is active learning AL, which has gained popularity in recent decades. By choosing the most informative samples, AL aims to reduce the costly data labeling procedure and build a robust training set that is resource-efficient. In this work, we aim to improve the performance of BCNN using AL method to build a competitive classifier considering the Bayesian Active Learning Disagreement BALD acquisition function (Dropout Bayesian Active Learning by Disagreement), which incorporates model uncertainty information. In a previous work, BCNN was built and applied on Pavia datasets giving 99.7% classification accuracy. For comparison traditional BCNN with BALD, The techniques were applied on the Indian Pines dataset. The average accuracy of the classification had increased from 90% to 98% using BALD method.

1. INTRODUCTION

As a result of the rapid advancements in hyperspectral imaging technology, hyperspectral sensors can now collect big data from a wide variety of electromagnetic spectrum bands. A significant quantity of spectral and spatial data indicating the texture, boundary, and shape of ground objects is available in the resulting hyperspectral image HSI. Because ground objects have unique features, HSI can precisely identify each every pixel. A wide range of applications have so far been created using HSI, including those for agriculture [1], anomaly detection [2], and marine monitoring [3].

Deep learning DL performs in feature extraction and learning capacity. Traditional Convolutional Neural Networks CNN still has drawbacks, though, namely its reliance on massively labeled samples and restriction to the local receptive field. Thus, a popular topic in the field of HSI classification is optimizing DL models to work effectively in small labeled sample data [4]. To address this issue, several studies have concentrated on developing DL paradigms with few labeled examples, such as active learning AL. AL is based on an iterative process working on unlabeled data in order to choose the samples, which are most informative. This selection is made depending on the scores that are calculated from a model outcome [5]. The candidates are acquired and added to the training set, after which the classifier is trained using the new training data. Compared to training using randomly obtained samples, training using actively chosen samples is more effective. Machine learning researchers frequently employ AL techniques including, normalized entropy [6, 7], uncertainty sampling [8], and margin sampling [9]. The posterior probability of sample categories predicted by the deep learning model may be used by AL to calculate the uncertainty score before selecting informative samples. A few advanced AL techniques are created for HSI. Costs associated with labeling can be reduced by AL. Actually, a lot of AL techniques combine with posterior probability and rely on a different classifier to query informative samples [10]. As can be seen in Fig. 1, there are theoretically three AL approaches that are frequently discussed in the literature, membership query synthesis, pool-based sampling, and stream-based selective sampling.

*Corresponding author. Email: mahmoodsq@gmail.com

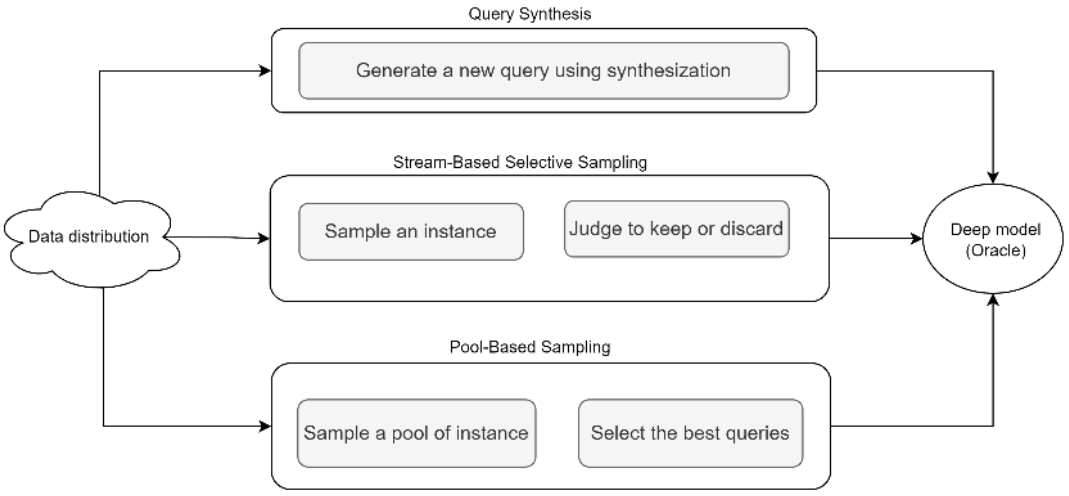


Fig. 1. The three scenarios used in active learning

The number of samples needed for the labeling is not predetermined in the stream-based selective sampling approach since each unlabeled sample is processed individually [11]. For online data that is delivered in sequence, this sampling rule is appropriate. It is challenging to calibrate the threshold in this situation to decide whether or not a sample should be chosen. New data is produced by membership query synthesis that may be annotated [11]. Due to the limited knowledge of data distribution in unexplored areas, it may suffer from low-quality generation even when it is efficient in the interpretable feature space. The best query samples are selected via pool-based sampling from the whole unlabeled set [11]. Each unlabeled sample is often given a score, and the selection is then made based on the score, for example, sending the samples with the highest scores for annotation.

2. LITERATURE REVIEW

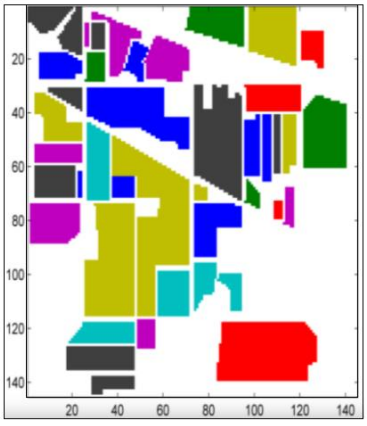
Most deep learning AL techniques described in the literature studies employ pool-based sampling technique and aim to choose the most informative data from a set of unlabeled pool samples. Some deep learning AL techniques combine approaches for deep image creation with membership query synthesis. In the area of hyperspectral image classification, AL has recently received a lot of attention. First, a method using semi-supervised multinomial logistic regression classifier using an active selection approach based on entropy (EP) was reported [12]. Then, the loopy belief propagation methodology and AL methods were examined [13,14], together with the Bayesian classification method. Following that, a Markov random field MRF based AL framework [15] was presented. Additionally, some studies using the AL approach and deep learning in conjunction have been investigated for HSI classification [16, 17, 18]. In particular, a method was developed [16] that combines the stacked auto encoder (SAE)-based neural network with the multiclass level uncertainty (MCLU) AL criteria. A method to combine a weighted incremental dictionary learning criterion with the restricted Boltzmann machine (RBM) was put out in [19]. [7] used the posterior probability evaluation to exclude out samples with comparable data based on the feature's distance from the hyperplane. For the extraction of deep features with uncertainty prediction on a given sample, [20] initially employed an auto encoder using category probability output from a neural network. To assess the inner indeterminacy of a multi-view and to discover the exterior indeterminacy of samples, [10] employed the posterior probability distribution. The combination of AL and DL models has been investigated for HSI classification. Integrating AL technique with DL methods were investigated in many researches for the classification of HIS. [21] combined CNN and BvSB (best versus second best [22]) into a single framework, utilizing AL's efficient labeling and CNN's potent feature extraction capabilities. [23] developed a deep AL approach based on BvSB to evaluate the information diversity with asses the spatial uncertainty for a se of candidate samples

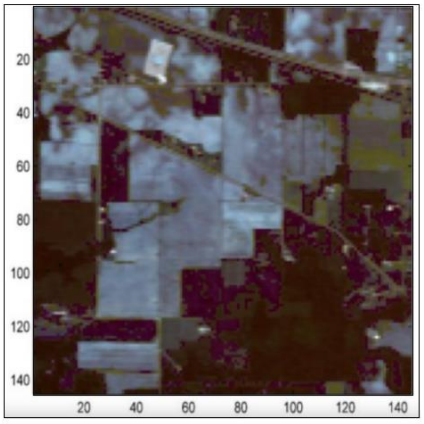
3. METHODOLOGY

3.1 Indian Pines Dataset

Indian Pines dataset was recorded by Aviris sensor. This dataset has an image size of 145 by 145 pixels and a spatial resolution of 20 m/pixel with wavelength range of 0.4 to 2.45 m. 200 effective bands covered 16 crop groups and preserved after processing the data's 224 original bands.

Two-thirds of the Indian Pines landscape is made up of agricultural land, while the other one-third is either covered with forest or other kinds of perennial forest flora. Minor roads, a rail line, a few low-density homes, various constructed elements, and two important dual-lane highways are all present.
The sixteen classes of available ground truth do not all mutually exclusive. Figure 1 displays the classes and the sample sizes for each class.





	Class	Samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Fig. 2. Ground reference map of Indian Pines dataset, Indian Pines image, classes and samples for each class

Noticing that there are four classes with low number of samples, which are the classes (1, 7, 9, 16). They represent areas with a small number of pixels. This low number of training samples will form a challenge for the classification model.

3.2 BCNN with active learning

With the least amount of training data possible, AL aims to create the best machine learning model. When using AL, the learner looks for the most informative data for the training. This is especially beneficial for HSI hyperspectral images. It might be difficult to train an effective classifier with little cost in labeling. The pool-based AL assesses the informativeness of instances with the highest levels of uncertainty that is most crucial to the label. The uncertainty estimates obtained in a DL environment are used in AL's query strategy [24]. Acquisition functions depend on evaluating the expected informativeness of the pool points. Based on these results they actively choose the following data point and add it to the training set. Those acquisition functions can be employed within a Bayesian AL framework

For the purpose of developing machine learning models, AL algorithms concentrate on choosing their own training data. As mentioned earlier, AL can be used in three different situations. We look at the scenario of pool-based AL, where the learner can choose points for annotation from a pool of unlabeled data. AL algorithms must give a score or utility to each position in the input space that can be queried in order to determine which locations in the input space are the most informative for the learner to pick for the training data. Every point in the pool set is assessed using this utility function. An information theoretic approach can be used to construct these functions. Probabilistic AL generally refers to methods based on information theory.

3.3 Methodology

In Bayesian AL considering the HSI pool-based active learning. For a collection of N samples, where each one is a member of one of the L classes. Separating the training set into three categories: train, validation, and pool sets. Assuming that the class labels are unknown for the pool set samples. The active learner can choose points for annotation from a pool set of unlabeled data. The active learner selects one or more of the N samples based on an acquisition function, and we will assign the appropriate class labels to these images. At each iteration of the algorithm, the AL selects more samples from the unlabeled pool that would be highly informative if their labels were known. In other words, after collecting data, we must query points from a pool set in a way that minimizes the posterior entropy. Conditioned upon the input and observed data, these points are queried using the predicted information gain. This gain is comparable to the mutual information between the parameters and the unobserved output. It is give by:

$$U(x) = H[p(\theta|D)] - E_{p(y|x,D)}H[p(\theta|D,x,y) \quad (1)$$

$$= I(\theta, y|D, x)$$

$$= H[p(y|x,D)] - E_{p(\theta,D)}H[p(y|x,\theta)]$$

Where $E_{p(y|x,D)}H[p(\theta|D,x,y)$ is the conditional entropy, and $H[p(\theta|D)]$ is the entropy which is a measure of the uncertainty in a distribution (average information):

$$H[P(x)] = -\sum_{x} P(x) log P(x)$$

Where $-logP(x) = J(x)$ is known as the information content. The relation of the mutual information between two different random variables is expressed according to

$$I[X,Y] = H[p(X)] - E_{p(Y)}H[p(X|Y)]$$

The gain in equation (1) is the acquisition function by [25]. It is known as the Bayesian Active Learning by Disagreement (BALD). The first term searches for the input x with a high degree of the model uncertainty regarding the output y. The second term searches for a data point with a low level of expected conditional uncertainty $E_{p(y|x,D)}H[p(\theta|D,x,y)$.

It can be explained that this acquisition function penalizes parameter uncertainty caused by inherent noise which is modeled by the likelihood $p(y|x,\theta)$. It will rewards data points with high output entropy caused by the parameter uncertainty. The marginal predictive distribution $p(y|x,D)$ is used to capture the parameter uncertainty.

Monte Carlo approach can approximate the equation (1) using samples from the posterior in order to estimated U(x) according to the relation

$$U(x) \approx H[\frac{1}{N}\sum_{i=1}^{N} p(y|x,\theta_i)] - \frac{1}{N}\sum_{i=1}^{N} H(p(y|x,\theta))] \qquad (2)$$

Using the Monte-Carlo samples of the predicted distribution generated by the test-time dropout of the Bayesian CNN implementation, equation (2) is then used to compute dropout BALD. The Bayesian CNN implementation may be done using dropout after each parameter layer to yield the expected class probabilities $p(y|x)$. Following the Bayesian interpretation of CNNs, the model is averaged over T stochastic forward passes to produce MC dropout samples of predicted class probabilities.

By averaging stochastic running over the model using Monte Carlo, the approach of using dropout at test time can be used [26]. We receive noisy estimates from the MC dropout testing performed on CNNs, with the possibility of varying test outcomes across multiple runs. This allows for the construction of the Dropout BALD acquisition function as following:

$$U(x) \approx H[\frac{1}{k}\sum_{i=1}^{k} p(y_i|x_i)] - \frac{1}{k}\sum_{i=1}^{k} H(p(y_i,x_i))] \qquad (3)$$

Where k represent the number of Monte-Carlo samples approximations.

The first term in (3) given by $H[\frac{1}{k}\sum_{i=1}^{k} p(y_i|x_i)]$ is the entropy of the average predicted probability. The AL searches for the locations where the model's average projected probability is marginally most uncertain. The AL searches for the location where the model's average output is the most uncertain. The second term in (3) given by $\frac{1}{k}\sum_{i=1}^{k} H(p(y_i,x_i))]$ seeks the point for which the average uncertainty is low.

The interpretation of the Dropout BALD acquisition can be formed as that the learner queries samples based on the predicted information gain. This gain is formed by the difference between the uncertainty of the resulted average output and the resulted average uncertainty of the output. Figure (3) summarizes the flowchart of the applied AL method employing the Dropout BALD function.

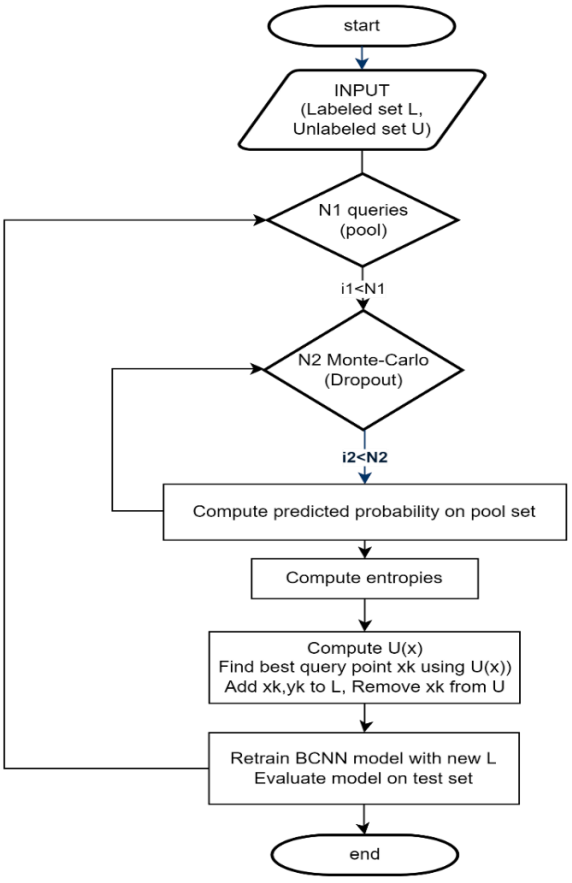


Fig. 3. Flowchart of the method of work using active learning with Bald

4. EXPERIMENTAL RESULTS

Considering the Indian Pines dataset, the work included two stages, data preparation and model building. First, Principal Components Analysis PCA was applied on the data for features extraction and dimensions reduction. The dimensions were reduced from 200 to 80. In order to apply the BCNN, two steps were implemented. First, the channel wise-shift to highlight more important spectral bands and suppressing less useful ones by relocating the spectral bands that are comparatively more significant to a more central location for more convolutions operations. Second, patches creating of size 7-by-7 with 80 channels.

A previous work using traditional BCNN is in [27][28], where the model outperforms for Pavia University and Pavia Center datasets with 99% accuracy. Applying our Bayesian model on the new dataset Indian Pines, the results in table (1) were obtained. Where the accuracy of classification is shown for each class.

TABLE I BCNN CLASSIFICATION RESULTS

		precision	recall f1	f1-score	support
1	Alfalfa	0.857	0.857	0.857	14
2	Corn-notill	0.962	0.886	0.922	428
3	Corn-mintill	0.803	0.900	0.848	249
4	Corn	0.781	0.704	0.741	71
5	Grass-pasture	0.925	0.931	0.928	145
6	Grass-trees	0.972	0.968	0.970	219
7	Grass-pasture-mowed	0.667	1.000	0.800	8
8	Hay-windrowed	0.972	0.986	0.979	143
9	Oats	0.500	0.500	0.500	6
10	Soybean-notill	0.847	0.836	0.841	292
11	Soybean-mintill	0.885	0.946	0.914	737

12	Soybean-clean	0.910	0.685	0.782	178
13	Wheat	0.935	0.951	0.943	61
14	Woods	0.969	0.976	0.972	380
15	Buildings-Grass-Trees-Drives	0.874	0.836	0.855	116
16	Stone-Steel-Towers	0.903	1.000	0.949	28
	Accuracy			0.904	3075
	Macro Average	0.860	0.873	0.863	3075
	Weighted Average	0.906	0.904	0.904	3075

Implementing the technique of BALD using the algorithm in figure (3), we considered N1=100 acquisition iterations and N2=50 dropout iterations. The method was applied on the 7*7*80 samples obtained from the previous steps. Table II. shows the results of the classification for each class.

TABLE II BALD CLASSIFICATION RESULTS

		precision	recall f1	f1-score	support
1	Alfalfa	1.000	0.929	0.963	14
2	Corn-notill	0.959	0.974	0.966	428
3	Corn-mintill	0.968	0.964	0.966	249
4	Corn	0.944	0.958	0.951	71
5	Grass-pasture	1.000	0.966	0.982	145
6	Grass-trees	0.991	0.995	0.993	219
7	Grass-pasture-mowed	1.000	1.000	1.000	8
8	Hay-windrowed	0.993	0.986	0.989	143
9	Oats	1.000	0.500	0.667	6
10	Soybean-notill	0.938	0.979	0.958	292
11	Soybean-mintill	0.989	0.959	0.974	737
12	Soybean-clean	0.947	0.994	0.970	178
13	Wheat	1.000	1.000	1.000	61
14	Woods	0.997	0.995	0.996	380
15	Buildings-Grass-Trees-Drives	0.966	0.983	0.974	116
16	Stone-Steel-Towers	0.966	1.000	0.982	28
	Accuracy			0.975	3075
	Macro Average	0.979	0.949	0.958	3075
	Weighted Average	0.976	0.975	0.975	3075

Comparing with the results of the BCNN, the improvement in accuracy can be seen for each class especially (1, 7, 9, 16) where the amount of data is small. The experimental results show the effectiveness of the BALD acquisition function based on Bayesian CNN architecture. Applying BALD can significantly solve the problem of overfitting for training CNN classifiers on small dataset. Due to the efficiency and characteristics of the acquisition function BALD, this technique outperforms.

5. CONCLUSION

In this study, we applied the BALD approach with BCCN for HSI classification. The challenge of required large amounts of training data was resolved by this method. Considering the pool-based active learning with the BALD acquisition function, which is computed using the approximating predictive distribution as an indication of uncertainty. A Bayesian CNN architecture's predictive distribution is obtained via BALD using Monte-Carlo (MC) dropout. A significant improvement in the classification performance of HIS Indian Pines dataset had achieved based on the training of BCNN on a small set of labelled train data using BALD. About 8% improvement in the classification’s averaged accuracy comparing with the traditional BCNN. One benefit of BALD is that the necessary computations are not inextricably linked to a certain model. A future work could consider more complicated HSI datasets in order to evaluate the performance of this method upon more difficult classification tasks.

Funding

Non.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to express their gratitude to the Department of Computer Engineering, Yildiz Technical University for their moral support. Please accept my sincere gratitude for the useful recommendations and constructive remarks provided by the anonymous reviewers.

References

[1] F. Lacar, M. Lewis, I. Grierson, "Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley, South Australia," in Proceedings of the IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217), Sydney, NSW, Australia, 9–13 July 2001, vol. 6, pp. 2875–2877, 2001.

[2] S. Matteoli, M. Diani, G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," IEEE Aerosp. Electron. Syst. Mag., vol. 25, pp. 5–28, 2010.

[3] F. Foglini et al., "Underwater Hyperspectral Imaging for seafloor and benthic habitat mapping," in Proceedings of the 2018 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea), Bari, Italy, 8–10 October 2018, pp. 201–205.

[4] S. Jia et al., "A survey: Deep learning for hyperspectral image classification with few labeled samples," Neurocomputing, vol. 448, pp. 179–204, 2021.

[5] P. Liu, H. Zhang, K. B. Eom, "Active Deep Learning for Classification of Hyperspectral Images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 2, pp. 712–724, February 2017.

[6] T. Luo et al., "Active learning to recognize multiple types of plankton," J. Mach. Learn. Res., vol. 6, pp. 589–613, 2005.

[7] D. Tuia et al., "Active learning methods for remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 47, pp. 2218–2232, 2009.

[8] J. Zhu et al., "Active learning with sampling by uncertainty and density for data annotations," IEEE Trans. Audio Speech Lang. Process., vol. 18, pp. 1323–1331, 2009.

[9] B. Demir, C. Persello, L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 49, pp. 1014–1031, 2010.

[10] J. Hu et al., "3D-Gabor inspired multiview active learning for spectral-spatial hyperspectral image classification," Remote Sens., vol. 10, p. 1070, 2018.

[11] R. Takezoe, X. Liu, S. Mao, M. T. Chen, Z. Feng, S. Zhang, X. Wang, "Deep Active Learning for Computer Vision: Past and Future," APSIPA Transactions on Signal and Information Processing, 2023.

[12] J. Li, J. M. Bioucas-Dias, A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 11, pp. 4085–4098, 2010.

[13] J. Li, J. M. Bioucas-Dias, A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 10, pp. 3947–3960, 2011.

[14] J. Li, J. M. Bioucas-Dias, A. Plaza, "Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 3, pp. 809–823, 2011.

[15] S. Sun, P. Zhong, H. Xiao, et al., "An MRF model-based active learning framework for the spectral-spatial classification of hyperspectral imagery," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 6, pp. 1074–1088, 2015.

[16] J. Li, "Active learning for hyperspectral image classification with a stacked autoencoders based neural network," pp. 1–4, 2015.

[17] C. Li, L. Zhang, "Hyperspectral image classification by combination of active learning and extended multi-attribute profile," pp. 541–544, 2016.

[18] J. M. Haut, M. E. Paoletti, J. Plaza, et al., "Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 11, pp. 6440–6461, 2018.

[19] P. Liu, H. Zhang, K. B. Eom, "Active deep learning for classification of hyperspectral images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 2, pp. 712–724, 2016.

[20] J. Li, "Active learning for hyperspectral image classification with a stacked autoencoders based neural network," in Proceedings of the 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–4.

[21] X. Cao, J. Yao, Z. Xu, D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, pp. 4604–4616, 2020.

[22] A. J. Joshi, F. Porikli, N. Papanikolopoulos, "Multi-class active learning for image classification," in Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 2372–2379.

[23] Z. Xue, S. Zhou, P. Zhao, "Active learning improved by neighborhoods and superpixels for hyperspectral image classification," IEEE Geosci. Remote Sens. Lett., vol. 15, pp. 469–473, 2018.

[24] D. A. Cohn, Z. Ghahramani, M. I. Jordan, "Active learning with statistical models," 1995.

[25] N. Houlsby, F. Huszar, Z. Ghahramani, M. Lengyel, "Bayesian active learning for classification and preference learning," CoRR, abs/1112.5745, 2011.

[26] R. Islam, "Active Learning for High Dimensional Inputs using Bayesian Convolutional Neural Networks," Department of Engineering, University of Cambridge, M.Phil in Machine Learning, Speech and Language Technology. This dissertation is submitted for the degree of Master of Philosophy, St John's College, August 2016.

[27] M. Qadir, G. Bilgin, "Hyperspectral Images Classification With Deep Bayesian Neural Networks," in Proceedings ACN International conference, Association with World Research Forum for advances in Science and Engineering (WRFASE) International Conference on Computer Science, Machine Learning and Big Data, Istanbul, Turkey, 5 March 2023, pp. 20-28.

[28] H. M. Fadhil, M. Abdullah, M. Younis, "Innovations in T-way test creation based on a hybrid hill climbing-greedy algorithm," IAES International Journal of Artificial Intelligence, vol. 12, no. 2, p. 794, 2023.