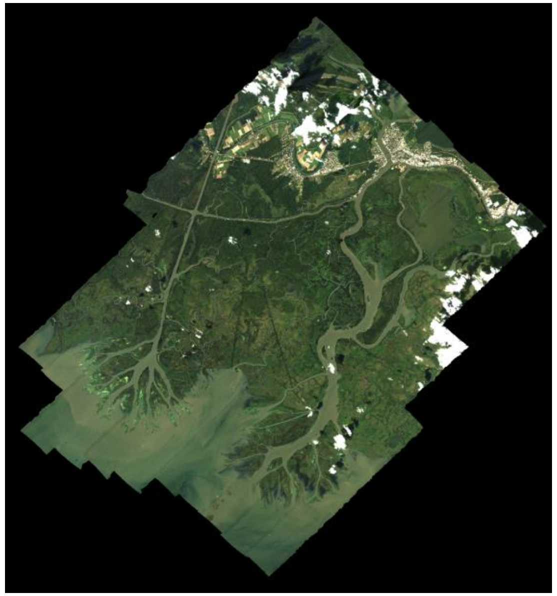
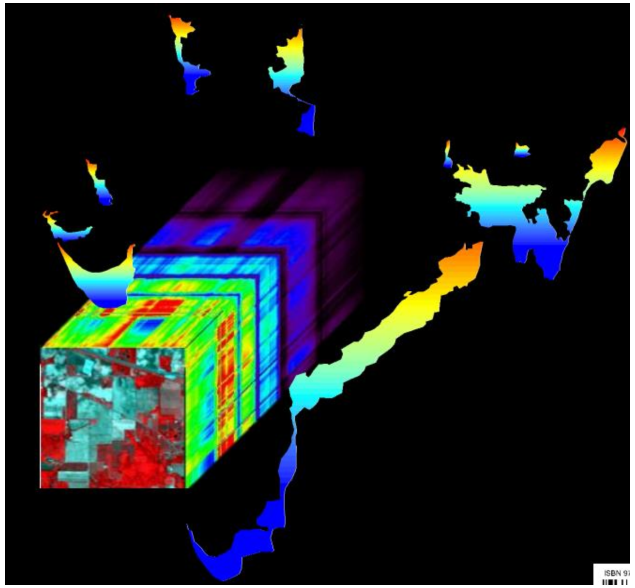
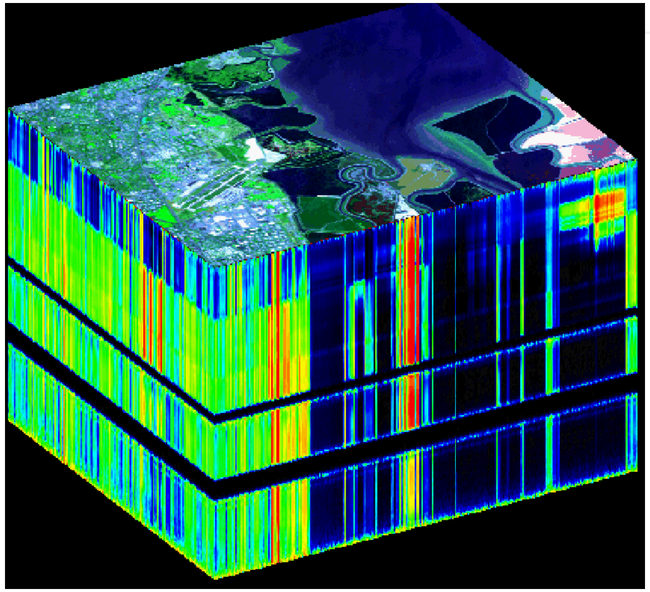
# Active Learning for Machine Learning-based Classification of Hyperspectral Imagery

**Contents**

- Introduction
- Data description
- Methodology
- Results
- Classified Output
- References

**Supervisor:**
**Dr.Deepak Mishra, Professor**

**Presented by**
**R RAMBABU, SC22M075**

# Introduction

- This mini project aims to evaluate the efficacy of the classification of AVIRIS-NG Hyperspectral data using the Support vector machines algorithm in conjunction with active learning through the pool-based uncertainty sampling method.
- In the Image data classification tasks where labelling classes can be a labour-intensive and time-consuming process. It involves manually identifying and labelling different types of classes, such as forests, natural vegetation, agricultural crops, buildings, roads, and water bodies.
- Active learning is a semi-supervised machine learning technique that selects the most informative samples to label and add to the training set, with the SVM machine learning algorithm (RBF kernel trick is a powerful technique for transforming non-linearly separable data into a linearly separable form) as the base estimator thereby reducing labelling effort while maintaining high classification accuracy.
- Active learning has broad applications across computer vision machine learning tasks, including the classification of Images, Object detection, Semantic segmentation, Face recognition, Video analysis etc.

# Dataset description

- AVIRIS-NG (Airborne Visible/Infrared Imaging Spectrometer Next Generation) is a next-generation hyperspectral imaging system that operates in the visible to shortwave infrared (VSWIR) with spectral range 380-2510 nano meters.
- It is a powerful tool for earth science research, environmental monitoring, and resource management applications.
- The AVIRIS-NG sensor is capable of collecting high-resolution spectral data with up to 426 spectral bands, enabling detailed analysis of the composition of the Earth's surface.
- The instrument is widely used for environmental monitoring, mineral exploration, agriculture, and other applications that require accurate spectral information

| Parameter | Description |
|---|---|
| Study Area: | 1.2 sq.km, Anand,Gujarat |
| Spatial and spectral resolution | 4 meters and 426 bands |
| Spectral range | 380-2510 nano meters |
| Data format | GeoTIFF, EPSG32643, UTM Projection |
| Data preprocessing | Principal component analysis for reduction of spectral bands |
| Ground truth | 6 classes |

Water [Red] 680 points
Urban [Green] 3879 points
Vegetation [Blue] 1148 points
non crop land [Yellow] 721 points
crops [Cyan] 701 points
urban vacant land [Magenta] 786 points

**Fig. Study Area, Anand,Gujarat**

Fig. True color composite of Hx Image

Fig. False color composite of Hx Image

Fig. Principal component analysis Hx Image (276*276*6)

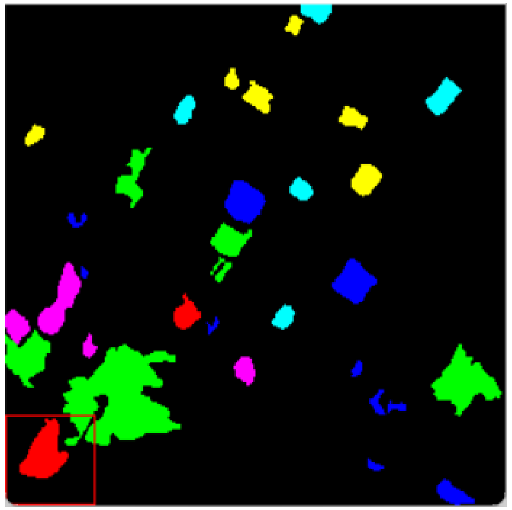


Fig. Ground truth labelled Image

# Methodology

**Input data**
- Data (AVIRIS-NG) Hyperspectral data with PCA
- Labels

**Initialisation**
- Split Labelled, Pool data sets with proportion of 70%,30%
- Initialize the active learner with Support Vector Machines with uncertainty sampling as query strategy with 10 iterations

**Querying and Training**
- Query the learner for most informative instances to labels
- Label the queried instance and add it to the labeled set and remove the queried instance from the pool set

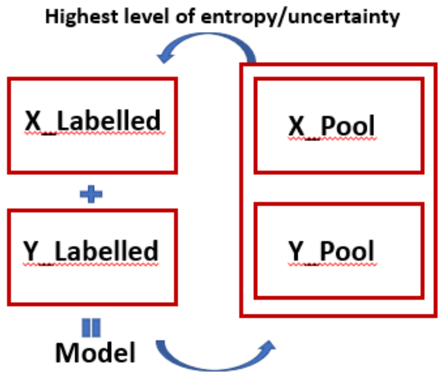
**Training on updated labels**
- Reinitialize the learner with the updated labeled set
- Train the learner on the updated labeled set

**Prediction on remaining labelled data**
- Use the trained learner to predict the labels of the remaining unlabeled data
- Calculate the accuracy of the learner on the remaining unlabeled

**Output Results**
- Calculate the accuracy of the learner on the entire dataset
- Deriving the performance metrics

Fig. Flow Chart

- ModAL is a Python library that provides a framework for implementing pool-based active learning. One of the most common pool-based active learning strategies is uncertainty sampling.
- A pool-based active learning strategy, which involves iteratively selecting the most informative data points from an unlabeled pool and adding them to a labeled set, the pool-based active learning strategies tend to be more computationally efficient, as they only require a subset of the unlabeled data to be evaluated at each iteration
- the most informative samples are typically those with the highest uncertainty or the greatest potential to improve the classifier's performance added to the training set. One way to measure uncertainty is by using entropy, which is the measure of the amount of uncertainty in a probability distribution
- The labeled set is used to train a machine learning model, and the model's performance is evaluated on the remaining unlabeled data with 10 iterations




```
Iteration 1: Accuracy = 0.987
Iteration 2: Accuracy = 0.987
Iteration 3: Accuracy = 0.986
Iteration 4: Accuracy = 0.986
Iteration 5: Accuracy = 0.987
Iteration 6: Accuracy = 0.987
Iteration 7: Accuracy = 0.987
Iteration 8: Accuracy = 0.988
Iteration 9: Accuracy = 0.988
Iteration 10: Accuracy = 0.989
```

Fig. Accuracy obtained during the iterations

# Results and Discussions

- From the derived performance metrics, it can be inferred that the classification model has performed well with high precision, recall, and F1-score values for most of the classes. The accuracy of the model is also high at 0.99. The water, urban, vegetation, crop land classes has achieved a perfect score of 1.0 in all performance metrics, indicating that the model is very good at identifying this classes.
- The non crop land, urban vacant land classes has a slightly lower F1-score at 0.99, which could be due to the complexity of the class.
- Receiver Operating Characteristic (ROC) curve for a multiclass classification problem. The micro-average ROC curve has an area under the curve of 1.0, indicating a perfect classifier.
- The ROC curve for each individual class shows varying levels of classification accuracy, with some classes achieving a perfect score and others having a lower area under the curve. Overall, the classification model shows high accuracy in correctly identifying positive cases (TPR of 1.0) with relatively low false positives (FPR of 0.99 to 1.0) for most classes

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| water | 1.00 | 1.00 | 1.00 | 679 |
| urban | 1.00 | 1.00 | 1.00 | 3879 |
| vegetation | 1.00 | 1.00 | 1.00 | 1146 |
| non crop land | 0.99 | 0.99 | 0.99 | 713 |
| crop land | 1.00 | 1.00 | 1.00 | 701 |
| urban vacant land | 0.98 | 0.99 | 0.99 | 786 |
| accuracy | | | 1.00 | 7904 |
| macro avg | 1.00 | 1.00 | 1.00 | 7904 |
| weighted avg | 1.00 | 1.00 | 1.00 | 7904 |

Fig. Confusion matrix

```
Test Accuracy: 0.996584008097166
Confusion Matrix:
[[ 679     0     0     0     0     0]
 [   0  3868     2     2     0     7]
 [   0     1  1145     0     0     0]
 [   0     0     0   707     0     6]
 [   0     1     0     0   700     0]
 [   0     6     0     2     0   778]]
```

Fig. Performance metrics of classified data

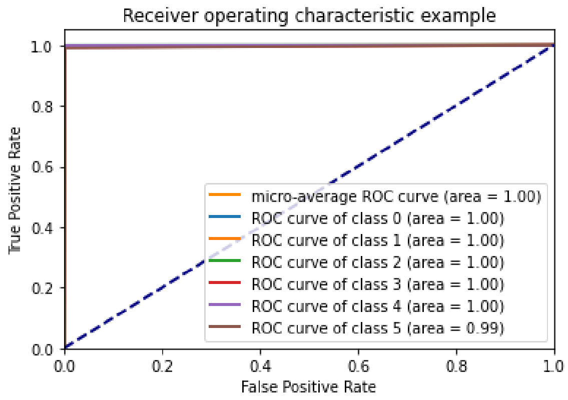



Fig. ROC curve

# Output: Classified Hyperspectral Image

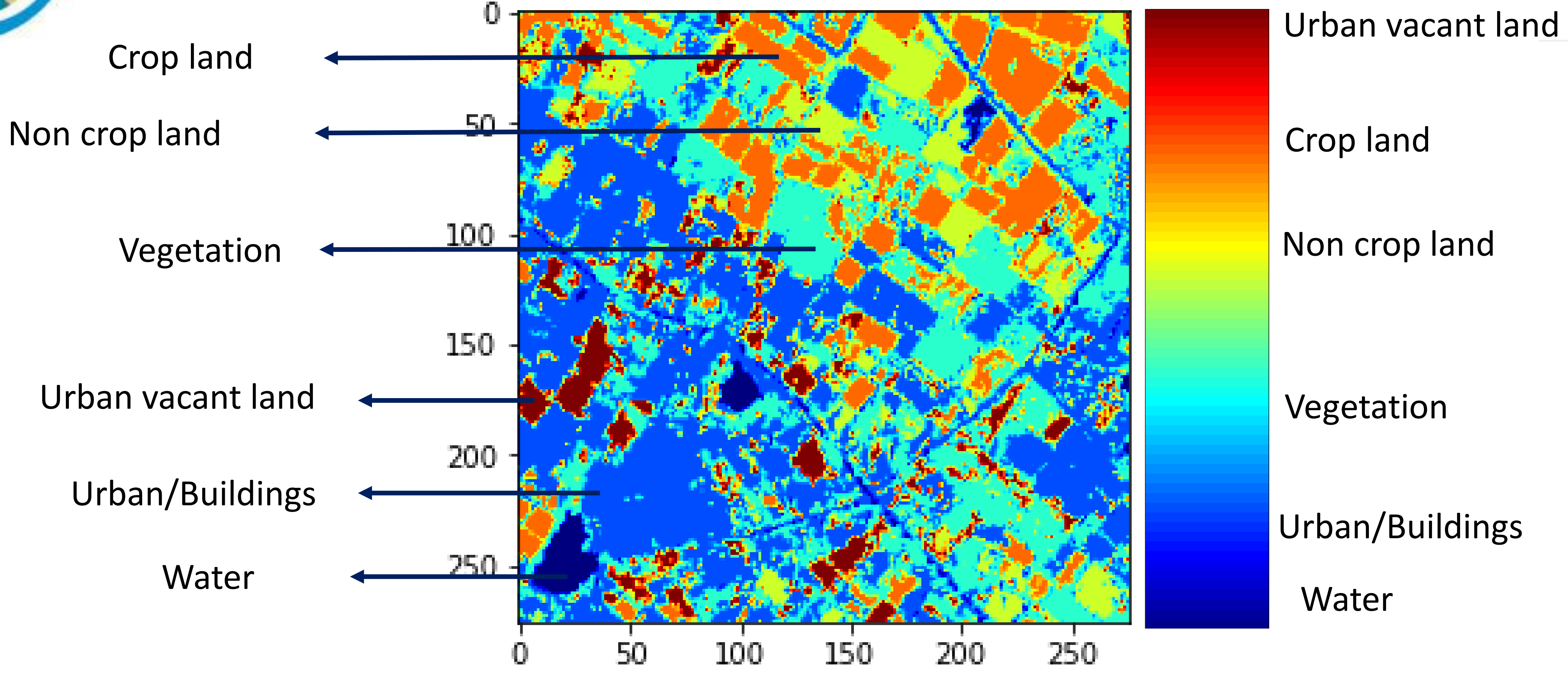



Fig. classified out put

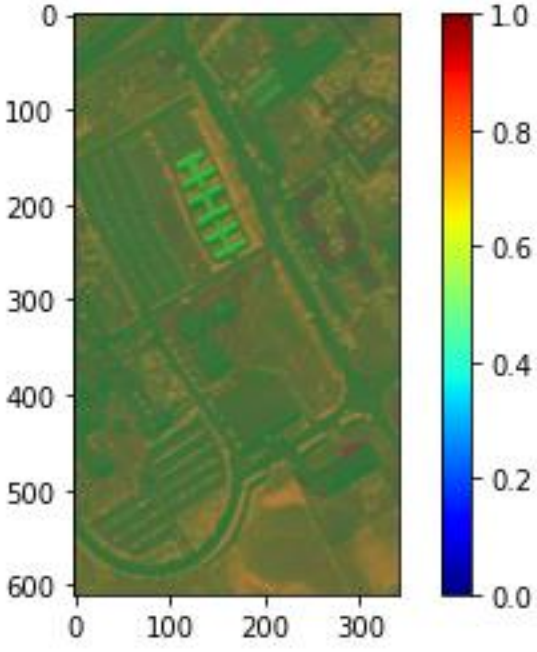



Fig.Pavia_PCA

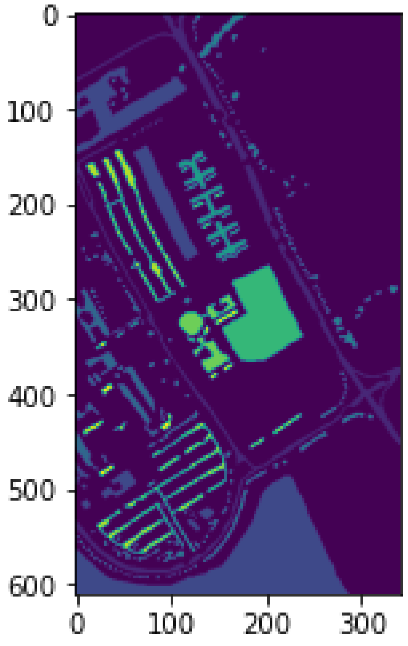



Fig. Pavia_ground truth

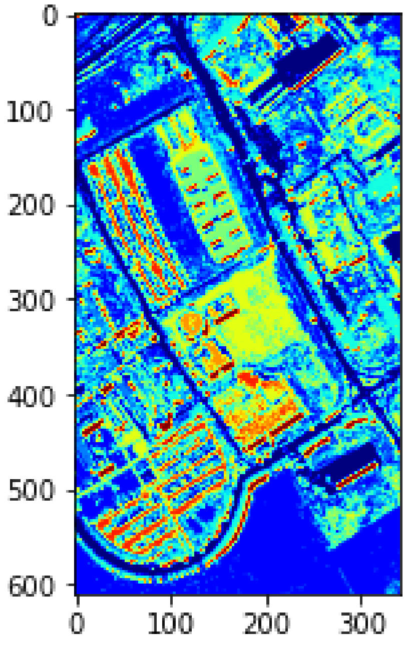



Fig. Pavia_Classified ouput

```
Test Accuracy: 0.9720170188890966
Confusion Matrix:
[[ 6510    12    21     0     0     5    32    51     0]
 [    0 18480     1    46     0   122     0     0     0]
 [   47     3  1925     0     0     1     0   123     0]
 [    1    91     0  2972     0     0     0     0     0]
 [    0     0     0     0  1345     0     0     0     0]
 [    1   409     0     0     0  4618     0     1     0]
 [   93     0     2     0     0     0  1235     0     0]
 [   38     7    79     0     0     9     0  3549     0]
 [    2     0     0     0     0     0     0     0   945]]
```

Fig. Confusion matrix

```
Iteration 1: Accuracy = 0.908
Iteration 2: Accuracy = 0.906
Iteration 3: Accuracy = 0.908
Iteration 4: Accuracy = 0.907
Iteration 5: Accuracy = 0.906
Iteration 6: Accuracy = 0.907
Iteration 7: Accuracy = 0.906
Iteration 8: Accuracy = 0.907
Iteration 9: Accuracy = 0.909
Iteration 10: Accuracy = 0.907
```

Fig. Accuracy during Iterations

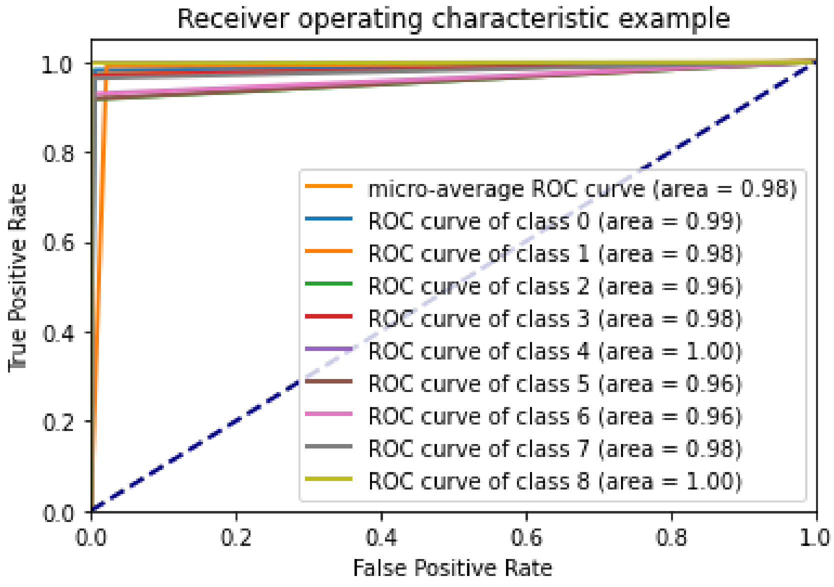



Fig. ROC curve

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Asphalt | 0.97 | 0.98 | 0.98 | 6631 |
| Meadows | 0.97 | 0.99 | 0.98 | 18649 |
| Gravel | 0.95 | 0.92 | 0.93 | 2099 |
| Trees | 0.98 | 0.97 | 0.98 | 3064 |
| Painted metal sheets | 1.00 | 1.00 | 1.00 | 1345 |
| Bare Soil | 0.97 | 0.92 | 0.94 | 5029 |
| Bitumen | 0.97 | 0.93 | 0.95 | 1330 |
| Self-Blocking Bricks | 0.95 | 0.96 | 0.96 | 3682 |
| Shadows | 1.00 | 1.00 | 1.00 | 947 |
| accuracy | | | 0.97 | 42776 |
| macro avg | 0.98 | 0.96 | 0.97 | 42776 |
| weighted avg | 0.97 | 0.97 | 0.97 | 42776 |

Fig. Performance metrics

# References


- [1] Zhang, L., & Huang, W. (2013). Hyperspectral image classification based on SVM with optimized RBF kernel. Journal of Sensors, 2013.
- [2] Sun, J., Chen, Y., Zhang, Y., & Song, B. (2019). A hyperspectral image classification approach based on SVM optimized by the fruit fly optimization algorithm. Remote Sensing, 11(1), 77.
- [3] Zhang, X., Li, X., Zhang, W., Li, W., & Li, D. (2021). Hyperspectral image classification based on SVM-RBF optimized by quantum particle swarm optimization algorithm. International Journal of Remote Sensing, 42(4), 1344-1362.
- [4] Lin, Y., Li, Q., Ma, H., & Zhang, X. (2019). Hyperspectral image classification based on kernel SVM with extended multi-feature optimization. Remote Sensing, 11(3), 293.
- [5] Liu, Y., Li, W., Ghamisi, P., & Plaza, A. (2018). Active learning with SVM for hyperspectral image classification: a study on updating strategies. Remote Sensing, 10(6), 919.
- [6] Zhang, Y., Chen, X., Liu, Y., & Chen, Y. (2018). Active learning for SVM classification of hyperspectral remote sensing imagery using adaptive hyperbox partition. International Journal of Remote Sensing, 39(18), 5919-5940.
- [7] Zhang, Y., Chen, X., Liu, Y., Chen, Y., & Zhang, J. (2019). Active learning for SVM classification of hyperspectral remote sensing images based on minimum spanning tree. International Journal of Remote Sensing, 40(17), 6743-6763.
- [8] Chen, X., Zhang, Y., & Liu, Y. (2020). Active learning for SVM classification of hyperspectral remote sensing images using adaptive sample size strategy. Remote Sensing, 12(2), 302.
- [9] https://github.com/modAL-python/modAL
- [10] https://modal-python.readthedocs.io/en/latest/content/examples/pool-based_sampling.html
- [11] Sheykhmousa, M. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review.


Questions?
Suggestions!
THANK YOU

# Literature Survey

- Thoreau, R., Achard, V., Risser, L., Berthelot, B., & Briottet, X. (2015). Active learning for hyperspectral image classification. this paper highlights the importance of active learning (AL) techniques in optimizing the training data sets for hyperspectral image classification. The paper suggests that the Core Set, Hierarchical, Breaking Tie, and BALD (batch Bayesian active learning disagreement) heuristics are effective in achieving accuracy metrics and discovering new classes quickly
- while autoencoders have been explored for solving the sample problem in hyperspectral image classification, their performance still has room for improvement. Future development should focus on few-shot learning, transfer learning, and active learning, and a fusion of these paradigms could lead to better results. Additionally, the use of RNNs and transformers in combination with learning paradigms could improve classification accuracy. Graph convolution networks have also shown promise in HSI classification, as they are capable of processing non-euclidean data directly. Finally, the construction of light-weight models to reduce the need for large amounts of labeled samples is another important direction for future research.