Multiclass Classification of Global Water Bodies in ReaLSAT

Tanisha Shrotriya, Master's Student in Computer Science

College of Science and Engineering

University of Minnesota

Driven to Discover®

Introduction

Classify the bodies of water in the ReaLSAT dataset into their type such as farm lake, river etc. and learn hidden structure in the data.

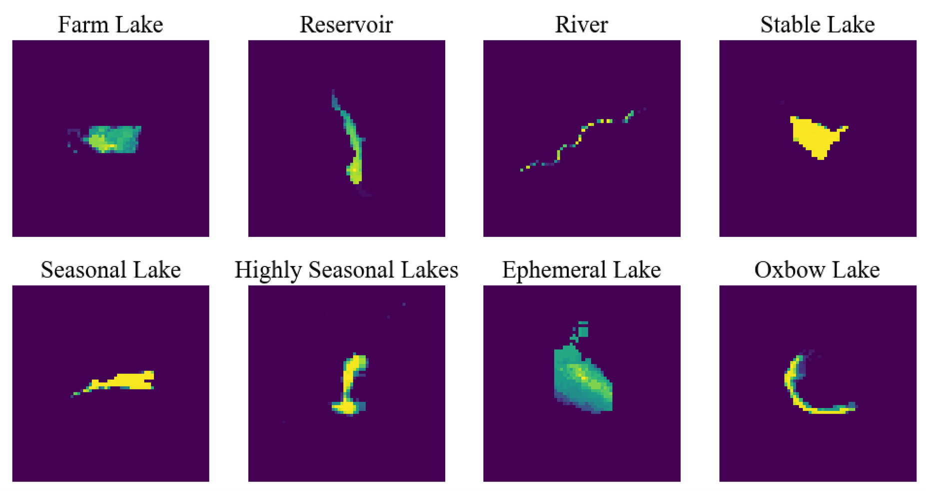
Challenges:

- Available labelled data may not be a good representative of the entire unlabelled dataset.
- Finding hidden structure in the larger unlabelled data of ReaLSAT.
- Remote Sensing Dataset Challenges.

Uses:

- Answer questions related to urbanization and water shortage

Figure 1: Data Samples: The color is a representation of the amount of time each pixel was a water pixel.



Methodology

Objective: Validate if the clusters generated using KMeans algorithm on the feature embeddings of a multi-class classifier, can be used as pseudo-labels for unsupervised curriculum learning.

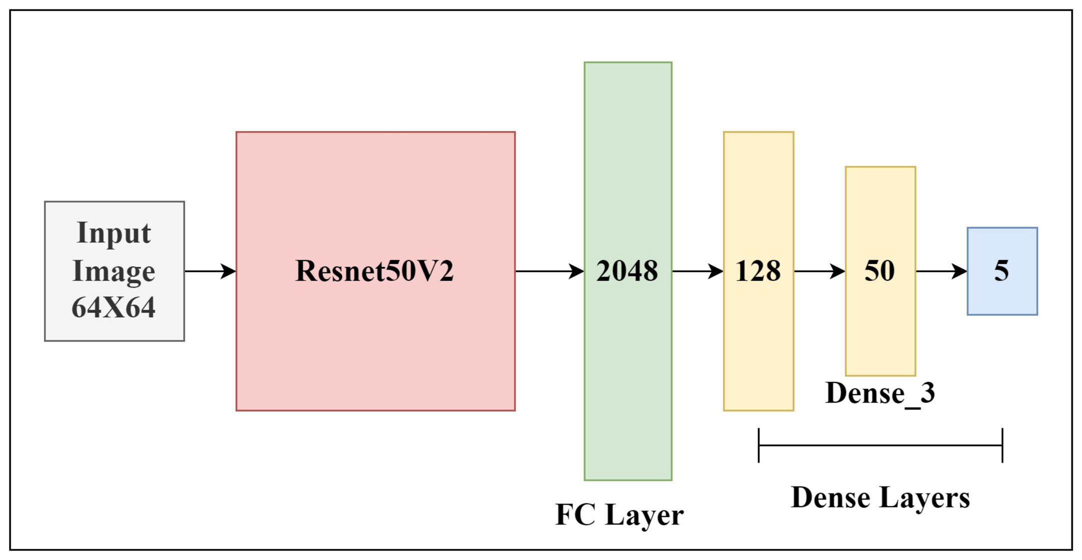
Experiment 1: Train Supervised Classifier using Transfer Learning on the Resnet50V2 pre-trained model.

Input Data: Normalized aggregated fraction maps of size 64X64X3

Data Split: Train:80%, Validation:10%, Test:10%

Process: Freeze Resnet50V2 module weights, train all other layers, test on an unseen set of images.

Stopping Criteria: When validation and training losses converge



Architecture Diagram

Experiment 2: Cluster the feature embeddings obtained at Dense_3 layer of size

Process:

- Fit KMeans for k=2,3,4,5 on feature size of 50.
- Visualize results in 2D using PCA on input features.

Dataset

ReaLSAT: A global dataset that contains the location and surface area variations of 681,137 bodies of water for 32 years.

Data Description: The aggregated fraction map shows the percentage of times a pixel was a water pixel over the timeframe of 32 years. Color in the dataset depicts water-life of a pixel and not depth.

Figure 2: Reservoir Anomalies

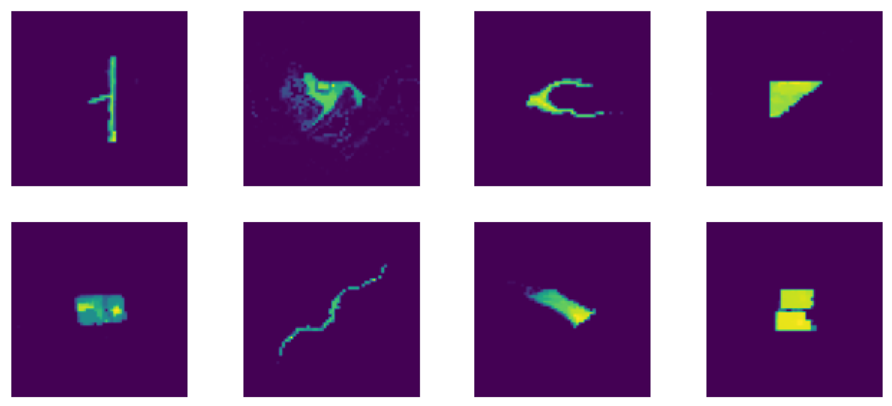


Table 1: Data Analysis

Class	**Reason For Exclusion**
Reservoir	Over-represented, highly variable.
Highly Seasonal Lakes	Low count (52)
Oxbow Lakes	Low count (12)

Table 2: Final Classes

Class	**Count**
Farm Lakes	427
Stable Lakes	143
Seasonal Lakes	288
Ephemeral Lakes	255
Rivers	317

Figure 3: Seasonal Lakes Anomalies

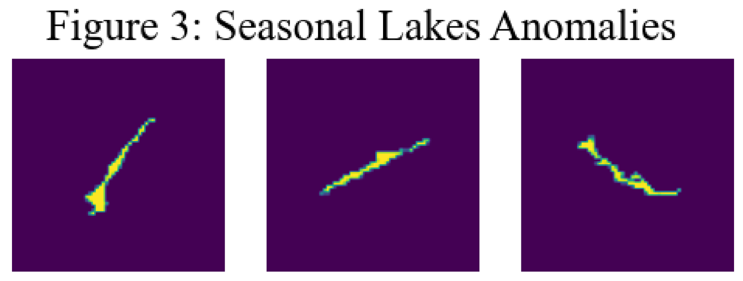
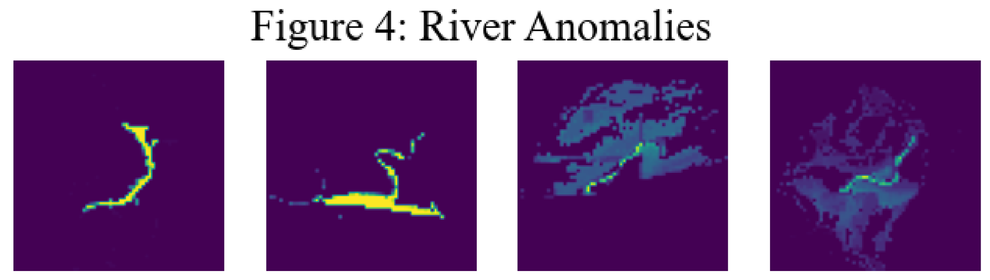
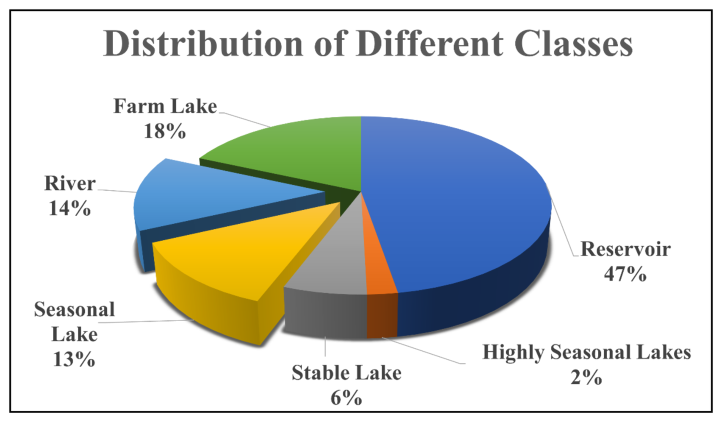


Figure 4: River Anomalies





Results

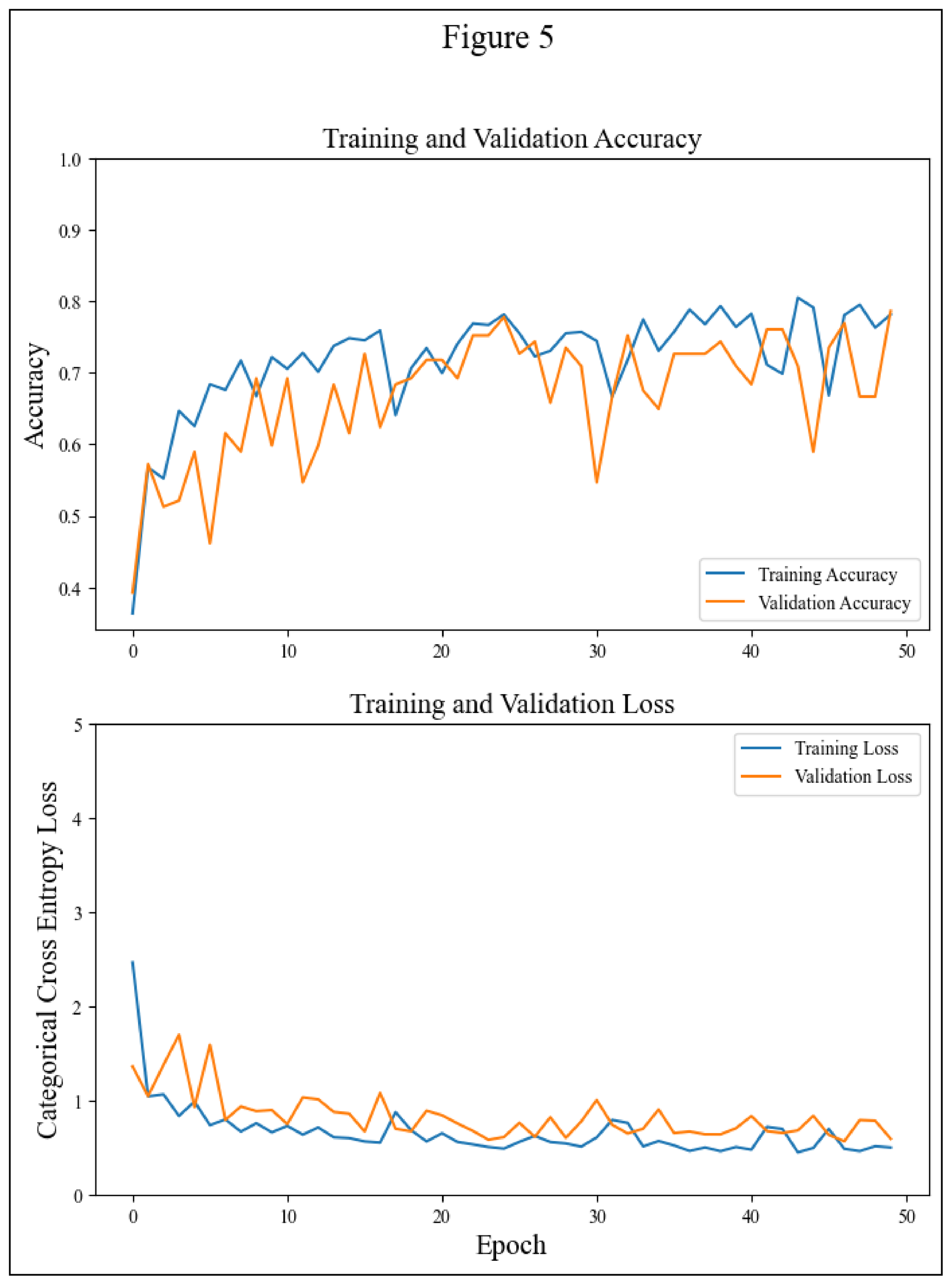
Table 3: Experiment 1 Results

Data	**Categorical Cross Entropy Loss**	**Accuracy**
Train	0.49	0.78
Validation	0.59	0.78
Test	0.46	0.79

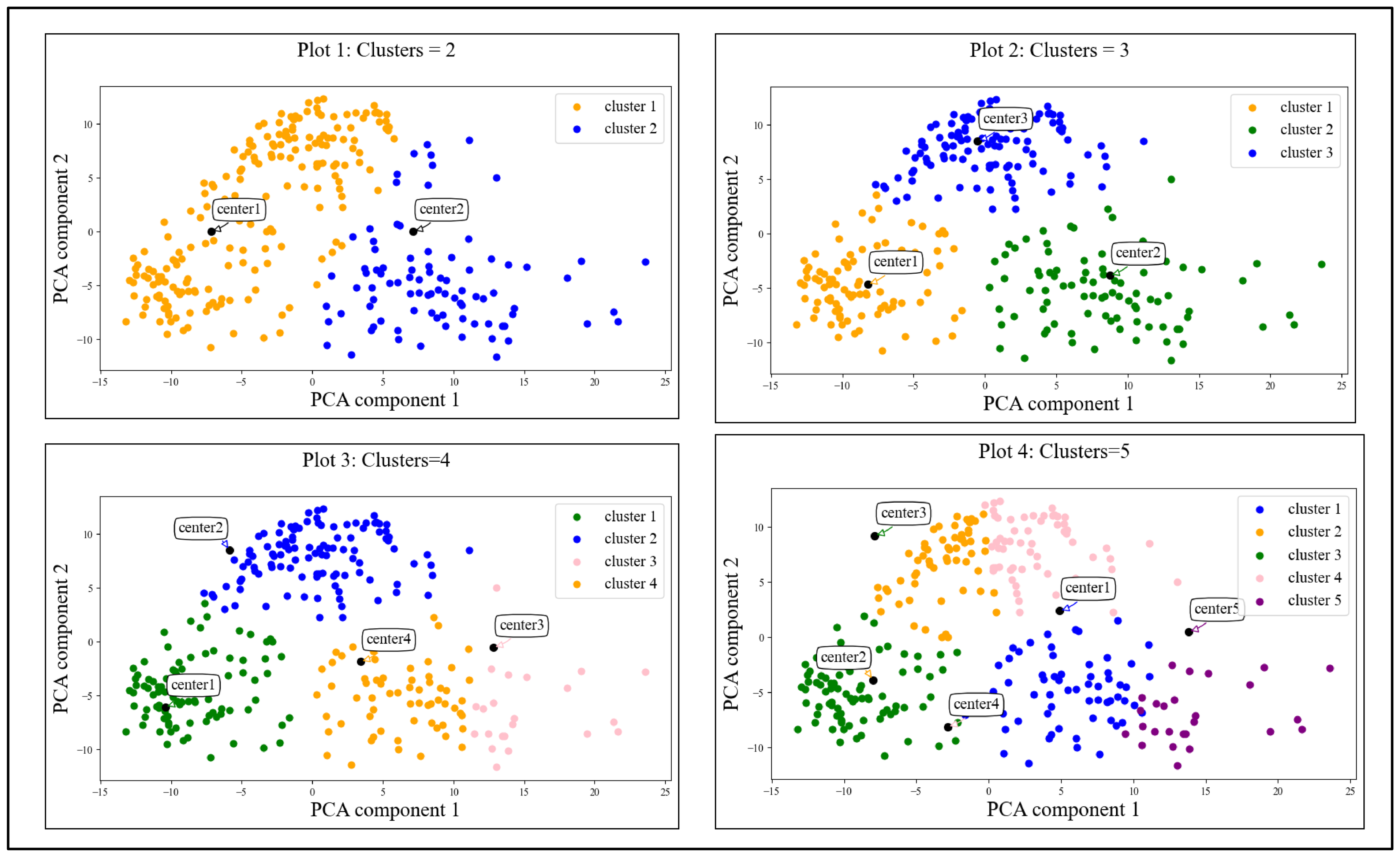
Table 4: Observations Based on Plots 1 to 4

Clusters	**Observations**	**Optimal Cluster**
2	Centroids are centered for each cluster	Yes
3	Centroids are centered for each cluster	Yes
4	Centroids are not centered	No
5	Centroids cross over the cluster borders	**No**

Experiment 1: Train-Validation Convergence



Experiment 2: Results



Conclusion

- The supervised classifier trained with transfer learning works well for the labelled data.
- KMeans does not provide significant insight into the hidden structure of the data.
- There is no one-to-one mapping between KMeans clusters and the discrete classes so the work in [2] cannot be successfully extended to solve this problem for the larger dataset.

Future Scope

- Use additional information such as geographical location and weather data.
- Find ways to check the relationship between labelled and unlabelled data.
- Better clustering algorithm, or self-supervised learning approach.

References

[1] N. Abid *et al.*, "UCL: Unsupervised Curriculum Learning for water body classification from remote sensing imagery,"