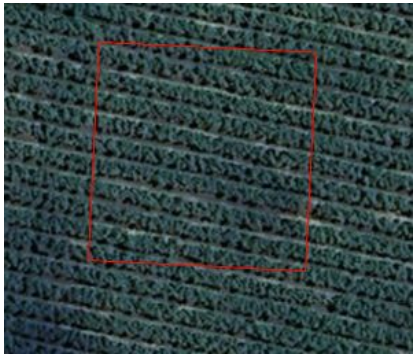
Land Cover Classification (one dimensional analysis)

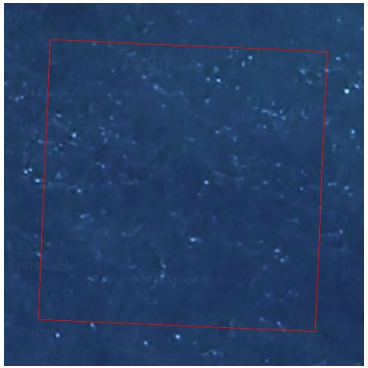
Ben Harris
Ian Swallow
Sam Hobbs
Collins Senaya

What is Land Cover Classification?

- Taking input images and determining what type of geological features are shown
  - Each image consists of pixels that contain 432 “bands” of information



These are “Permanent Crops”



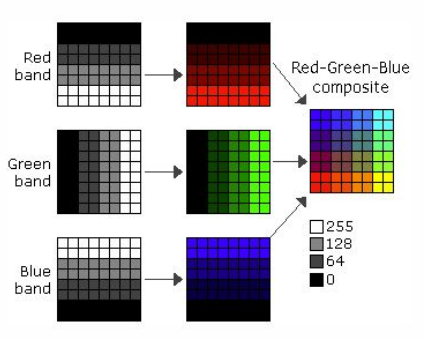
This is a “Waterbody”

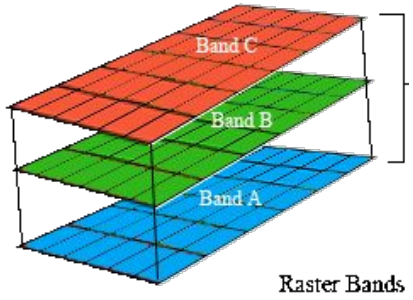
This is “built-up”

etc...

What is a “band” of data?

- Each band represents a particular characteristic of a pixel
  - Each band is a different infrared frequency that acts differently when they collide with different materials like rocks, water, and tree canopies
- Each pixel has 432 bands
  - Put together, that’s A LOT of data! Millions upon millions of data points to consider
    - Our .csv had 115 million+ cells…





Why do we care?

- It’s useful information!
  - Gives insight to regions without having a human needing to label everything
  - Can track trends in landscape
- Broad applications
  - Cuts down on Labor & Time for classification
  - Can be repurposed for other classifications projects.
- BioSCape Project
  - NASA backed Biodiversity research of the South Cape of Africa

Machine Learning

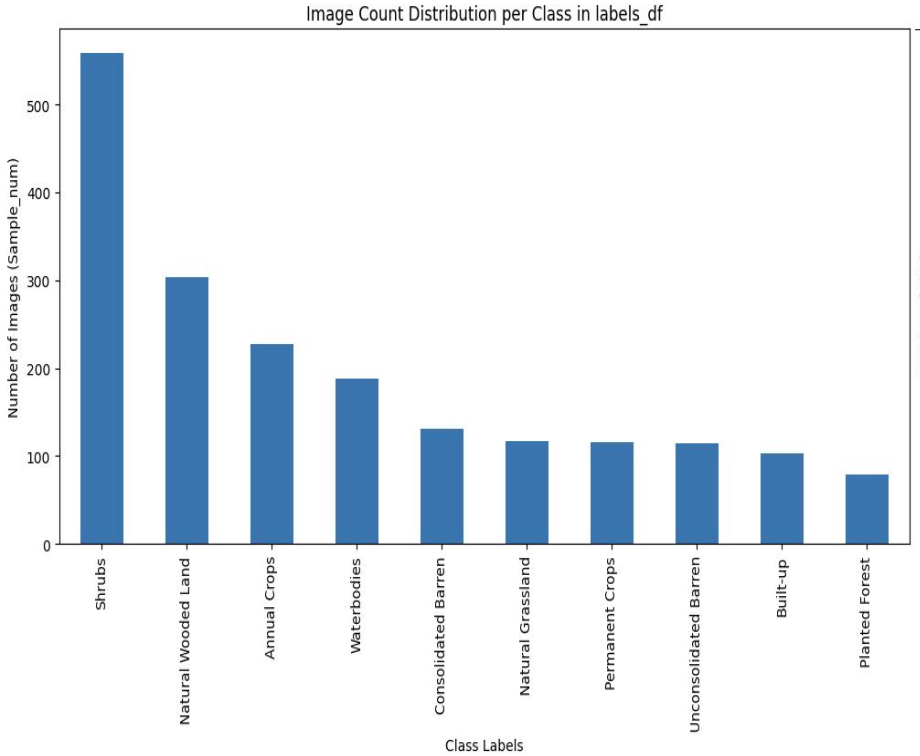
- Convolutional Neural Network (CNN)
  - Learns trends from previous data, and performs predictions on new input data based on those trends
- PCA Analysis
  - Reduces the dimensionality of the data
    - In this case, reduce the usage of 373 bands down to the most important bands
    - Useful for KNN and Logistic Regression
      - K-Nearest-Neighbors (KNN)
        - A method of classification where each point is categorized based on similar characteristics to other data points
      - Logistic Regression (LR)
        - Predicts class probabilities using a curve that maps values between 0 and 1.
- Linear Discriminant Analysis (LDA)
  - Reducing the feature separability between classes

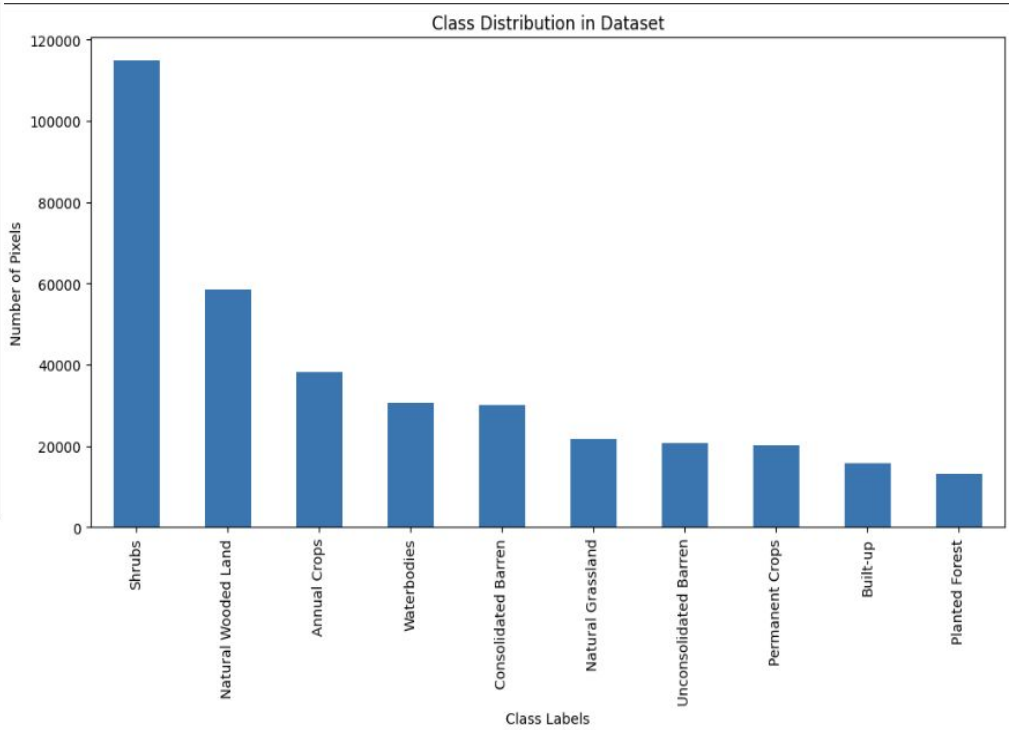
Pre-Processing

- ~2,200 images to train and test on
  - We labelled each one a minimum of three times to get a majority vote on classification
- Create a CSV file so the information is usable
  - Rasterio library
- Filtering of data
  - Not all data given from the images is useable (Noisy, Flight patterns not in images)
- Split data
  - Set aside some of the data to test our models, ensuring they are predicting accurately
  - Splitting data on an Image-level

Data Imbalance

- Oversampling was performed with no improvements
- Class Weights were also implemented to counteract class imbalance, also with no improvements.



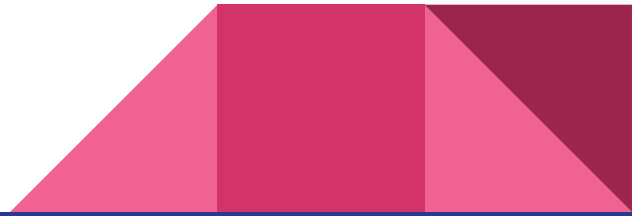


Initial Results

- All of these were done with varying degrees of incorrect data and formatting
  - Gives insight to where we began; **none** of these were correct in the end
- CNN
  - 86% initially
- PCA
  - Highest was 42% using KNN
- LDA
  - 76% accuracy

Improvements

- Further filtering of data
  - Removing elements that were noisy
  - Re-formatting data
- Grid-Search
  - A mechanized method to optimize parameters
- Random Forest
  - Improve class separability using advanced techniques like:
    - Balancing the dataset
    - Adding or transforming features
- Reformatting
  - Using the algorithms in different ways to achieve slightly different but more applicable results
    - This reduced our accuracy, but gave us more applicable information



Final Product

- CNN
  - 69% accuracy
- PCA
  - Highest accuracy 67% using KNN
- LDA
  - Highest accuracy 64%

	precision	recall	f1-score	support
Annual Crops	0.74	0.74	0.74	8240
Built-up	0.79	0.85	0.82	3453
Consolidated Barren	0.53	0.64	0.58	5664
Natural Grassland	0.43	0.23	0.30	4720
Natural Wooded Land	0.69	0.61	0.65	10891
Permanent Crops	0.84	0.66	0.74	3816
Planted Forest	0.75	0.57	0.65	2320
Shrubs	0.65	0.80	0.72	24045
Unconsolidated Barren	0.52	0.32	0.40	3868
Waterbodies	1.00	0.91	0.95	7161
accuracy			0.69	74178
macro avg	0.69	0.63	0.65	74178
weighted avg	0.69	0.69	0.68	74178

Confusion Matrices

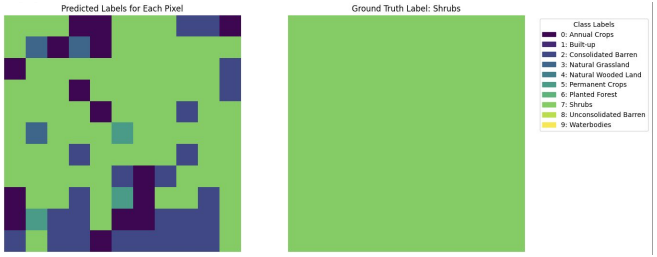
- Confusion Matrices give insight to how a model is performing
  - Shown here is our confusion matrix for CNN

Recognizing Trends

- We do extremely well in the “built-up” and “waterbodies” categories
  - This is likely because they are easily distinguishable compared to other categories
    - Water is a solid blue with consistent waves, buildings have rigid lines
- We perform worst on “natural grassland”
  - This is likely due to how seemingly mixed a lot of the grassland images were
    - We classified these images as “shrubs” more often than not
      - Consequently, we perform very well on shrubs!

Example of grassland

Example of shrubs



Further Improvements and Closing Thoughts

- Further optimization of data
  - There's a LOT of small things to optimize and tune
- More samples
  - We had to remove the "wetlands" category because we simply didn't have enough data
  - More samples gives more insight to the model