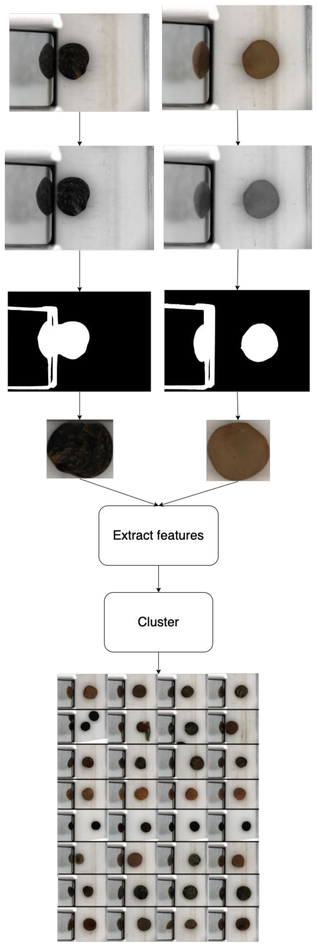
Classifying lentil testa (seedcoat) phenotypes using unsupervised learning

Amanda Ewen

BACKGROUND: Lentils have different seed coat colours and patterns. Accurate classification of lentils by testa patterns helps plant scientists understand the genetics of seed coats in lentils. Computers can be used for this analysis.

METHODS

- Segment images around lentil seed
- Extract information/features using:
  - Image processing methods
  - Machine learning methods
  - Hybrid methods
  - Deep neural networks
- Cluster using K-means
- Get numerical score and visually inspect each grouping from different methods to compare
- Repeat feature extraction on dataset of new lentils and compare results



RESULTS

- Features extracted using hybrid methods get best visual clustering result
- Convolutional Neural Networks get high scores but group by size and position
- Patterns are confused based on how much of lentil is covered

Combining image processing and machine learning algorithms improves classification of lentil seedcoat patterns

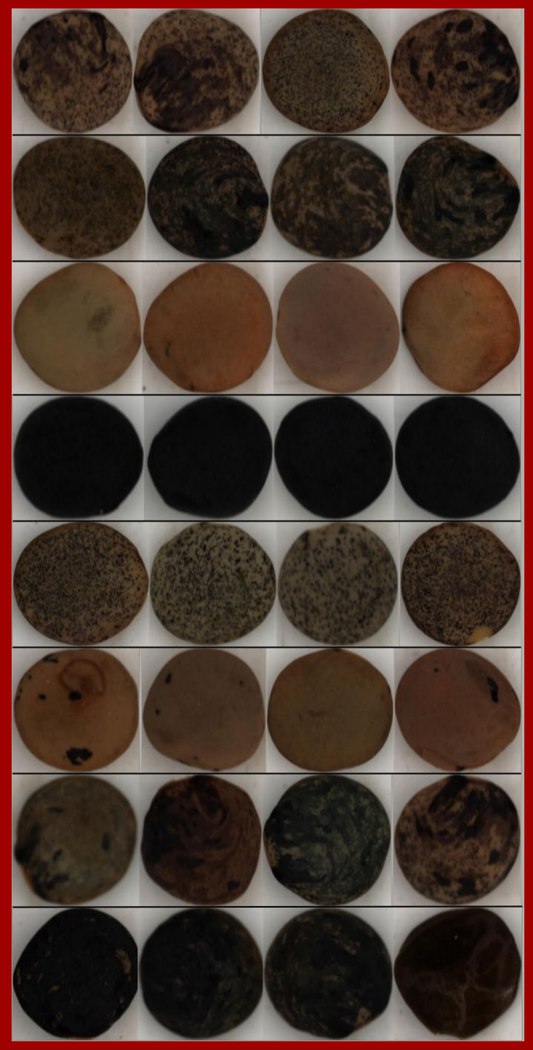


Figure 1: Feature extraction using image processing and machine learning algorithms

Results		
Feature Extraction Method	**K-value**	**Silhouette Score**
Image processing methods with PCA on original images	16	0.17751
Image processing methods with PCA on regions of interest	16	0.23380
Image processing methods with PCA on regions of interest	8	0.29237
LBP features only from regions of interest	8	0.41582
Features from image processing methods on regions of interest fed into autoencoder	8	0.45718
Regions of interest given to simple CNN autoencdoer	8	0.53255
Regions of interest given to CNN autoencoder with VGG16	8	0.40439
Regions of interest given to CNN autoencoder with VGG19	8	0.73681

Comparison of Results on Original and New Dataset			
Feature Extraction Method	**Silhouette Score on Dataset of Aged Lentils**	**Silhouette Score on Dataset of New Lentils (after fine tuning)**	**Completeness Score**
Image processing methods with PCA on regions of interest	0.29237	0.34007	**0.43888**
LBP features only from regions of interest	0.41582	0.41979	0.43728
Features from image processing methods on regions of interest fed into autoencoder	0.45718	**0.56415**	0.36878
Regions of interest given to simple CNN autoencdoer	0.53255	0.19368	0.31337
Regions of interest given to CNN autoencoder with VGG16	0.40439	0.56177	0.37887
Regions of interest given to CNN autoencoder with VGG19	**0.73681**	0.54517	0.36976

Future Work

- Include K-means clustering as evaluation step of model development
- Try advanced autoencoders (i.e. a masked autoencoder)
- Develop a self-supervised model to classify images

Amanda Ewen, Samuel Horovatin
Keyhan Najafian, Dr. Kirstin Bett,
Derek Wright, Dr. Ian Stavness,
Dr. Lingling Jin

University of Saskatchewan

NSERC CRSNG