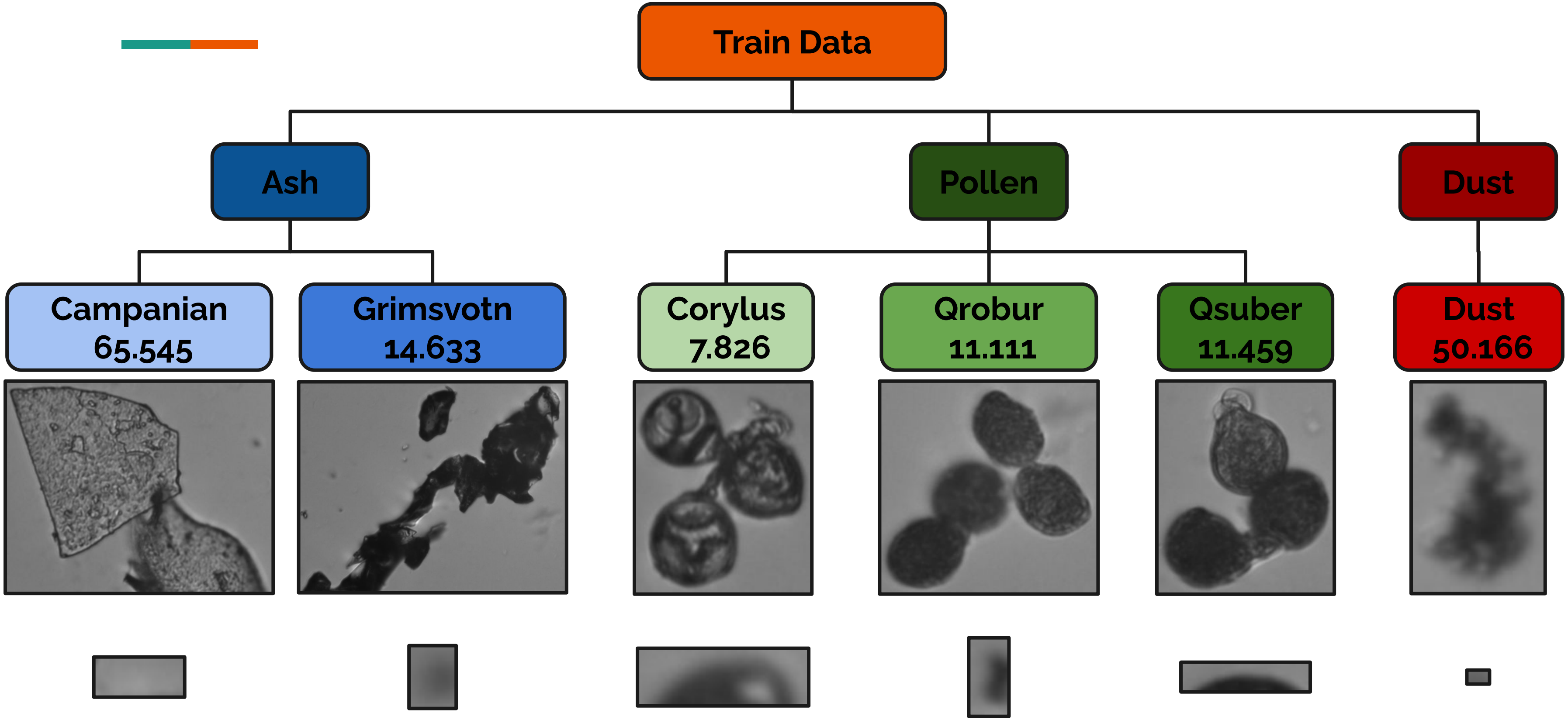
UNIVERSITY OF COPENHAGEN

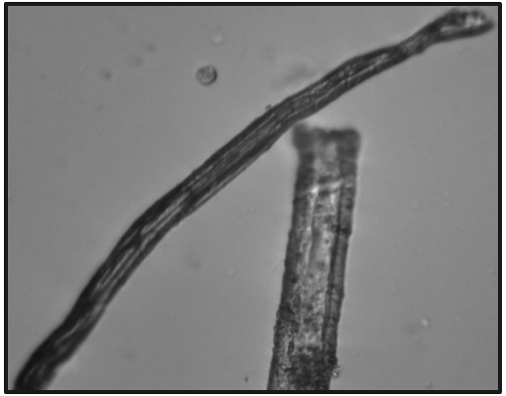
Identifying insolubles in IceCore data

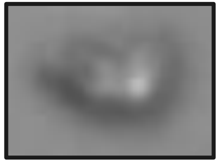
by Kristine Krighaar, Emma Y. Lenander, Jeppe Cederholm, Simon C. Debes, Camilla Okkels and Martin H. Petersen

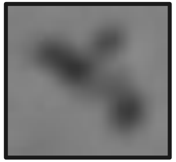
Figure adapted from:
Bulat et. al. (2018). Unknown Widespread Iron- and Sulfur-Oxidizing Bacteria beneath the East Antarctic Ice Sheet. Paleontological Journal. 52. 1196-1203. 10.1134/S0031030118100076.

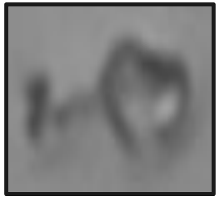

Train Data
Ash
Pollen
Dust
Campanian 65.545
Grimsvotn 14.633
Corylus 7.826
Qrobur 11.111
Qsuber 11.459
Dust 50.166

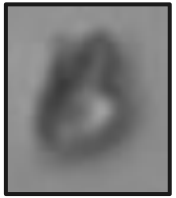
Test Data

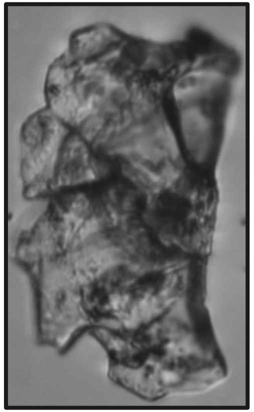


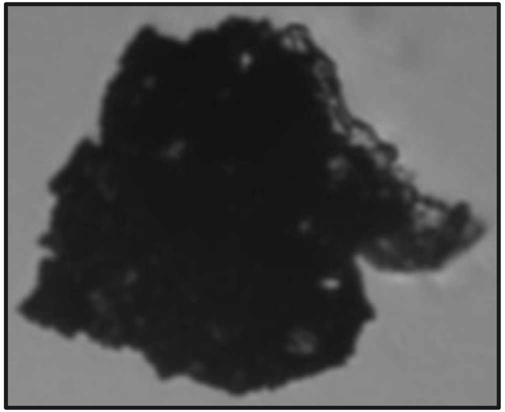










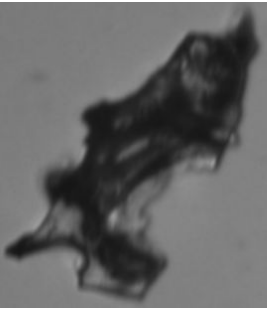


- Area (ABD)
- Area (Filled)
- Aspect Ratio
- Biovolume (Cylinder)
- Biovolume (P. Spheroid)
- Biovolume (Sphere)
- Circle Fit
- Circularity
- Circularity (Hu)
- Compactness
- Convex Perimeter
- Convexity
- Diameter (ABD)
- Diameter (ESD)
- Edge Gradient
- Elongation
- Feret Angle Max
- Feret Angle Min
- Fiber Curl
- Fiber Straightness
- Geodesic Aspect Ratio
- Geodesic Length
- Geodesic Thickness
- Intensity
- Length
- Particles Per Chain
- Perimeter
- Roughness
- Sigma Intensity
- Sphere Complement
- Sphere Count
- Sphere Unknown
- Sphere Volume
- Sum Intensity
- Symmetry
- Transparency
- Volume (ABD)
- Volume (ESD)
- Width

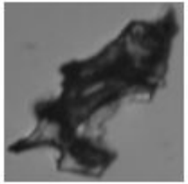
Preprocessing

- Scale pictures, such they can be used in a CNN
- Avoid overfitting by using same amount of data for each class in training

Scale down/up



?? x ??



64 x 64

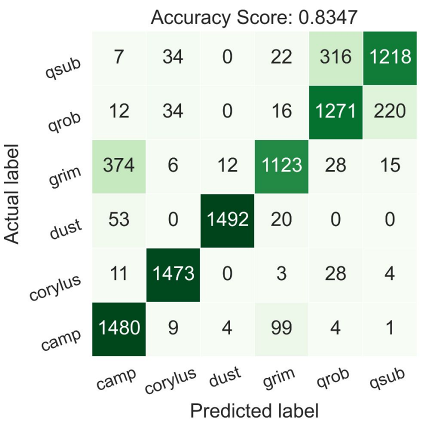
Convert to matrix

```
[[120 121 122 ... 129 129 130]
 [122 122 122 ... 128 131 131]
 [122 123 122 ... 123 131 132]
 ...
 [125 126 125 ... 131 130 131]
 [125 125 125 ... 132 131 131]
 [125 125 125 ... 131 131 131]]
```

Scale down

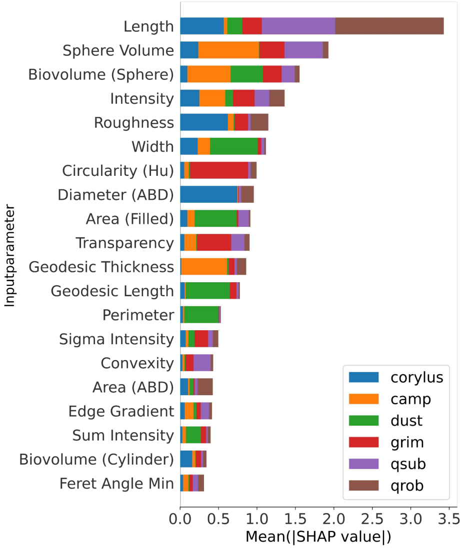
```
[[0.532 0.557 0.583 ... 0.762 0.762 0.788]
 [0.583 0.583 0.583 ... 0.737 0.814 0.814]
 [0.583 0.609 0.583 ... 0.609 0.814 0.839]
 ...
 [0.66  0.686 0.66  ... 0.814 0.788 0.814]
 [0.66  0.66  0.66  ... 0.839 0.814 0.814]
 [0.66  0.66  0.66  ... 0.814 0.814 0.814]]
```

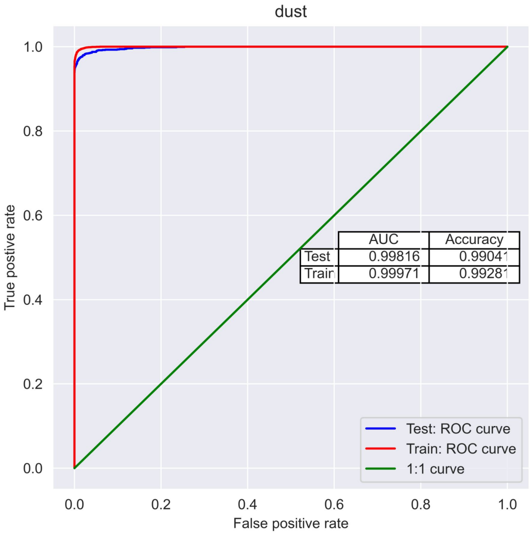
Visualize with imshow

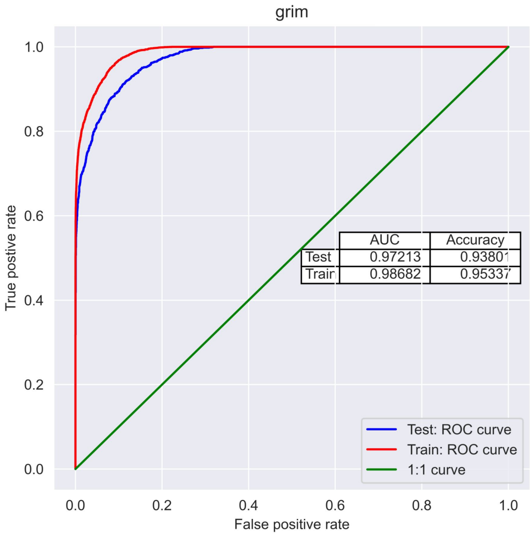


Meta data

- Image analysis data of the pictures
- LightGBM used for classification of the 6 types

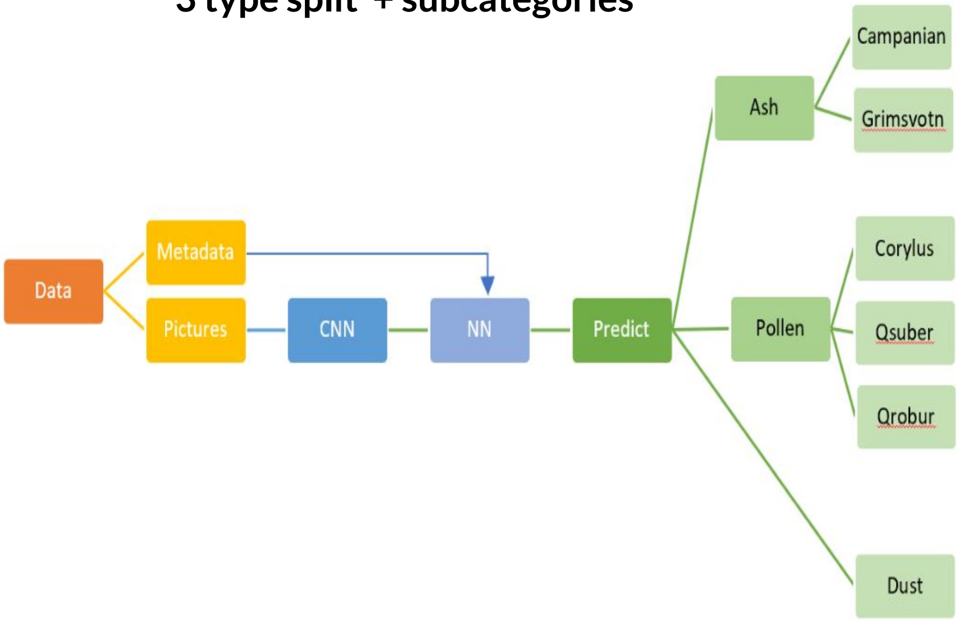




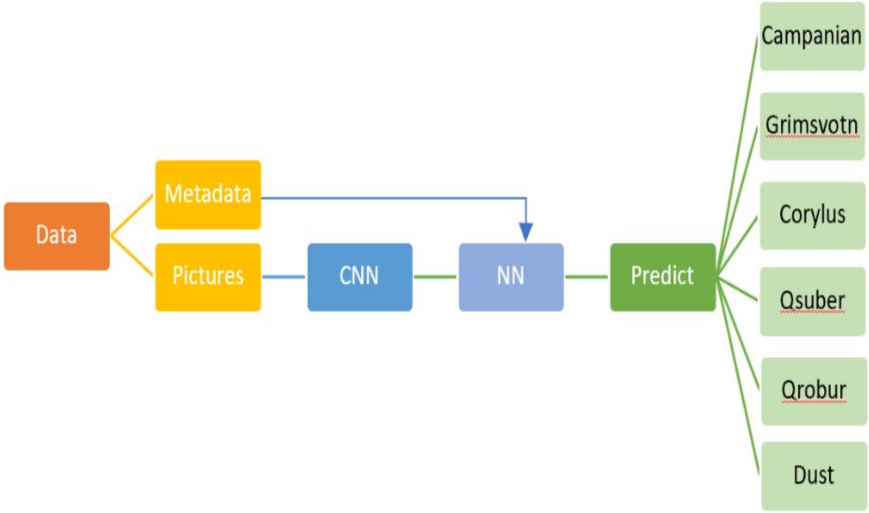


General CNN structure

3 type split + subcategories

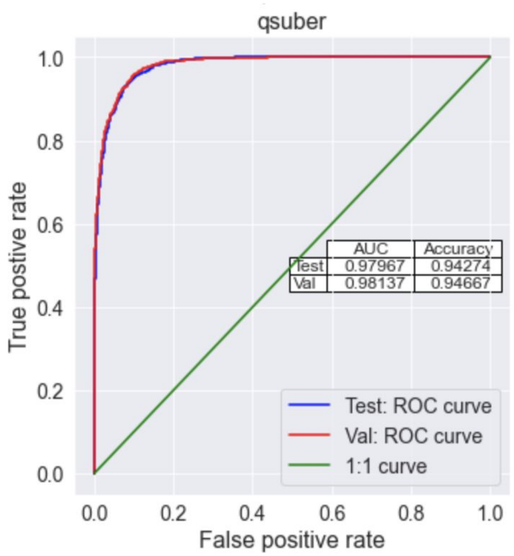


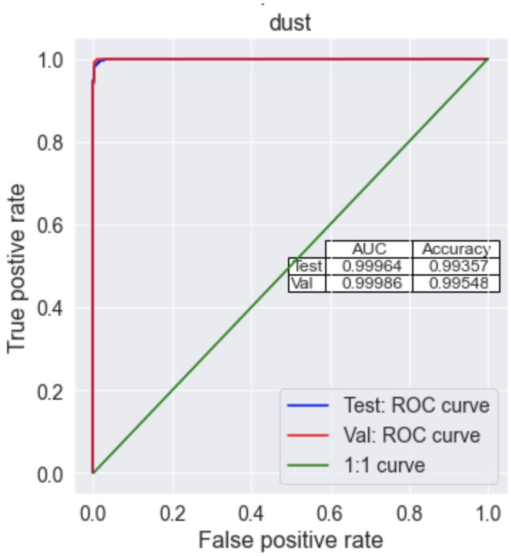
6 type split



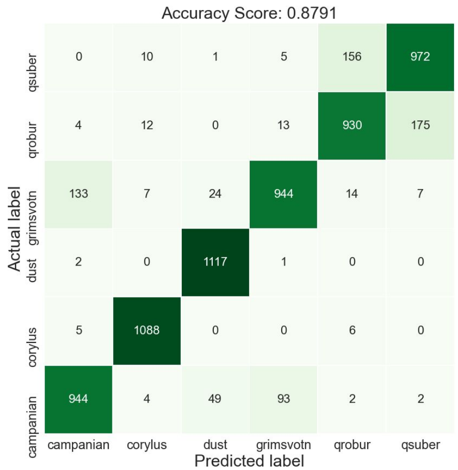
Result: 6 model split

- Confusion matrix with custom loss function

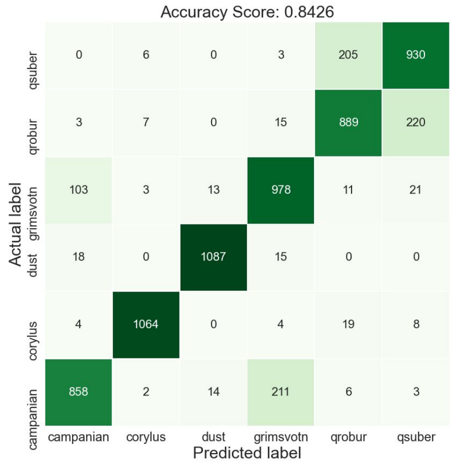




Categorical Crossentropy

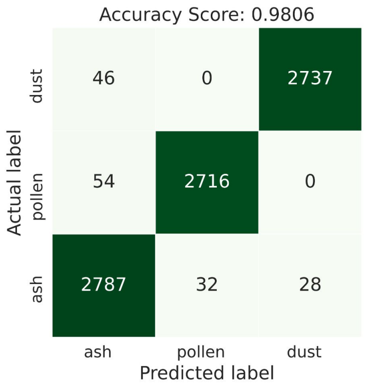


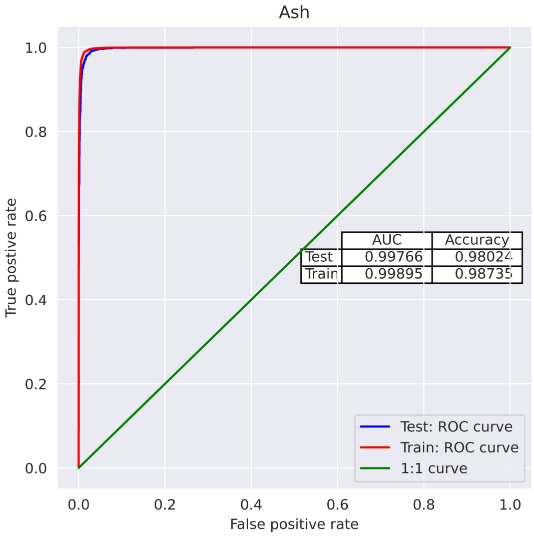
Own loss function

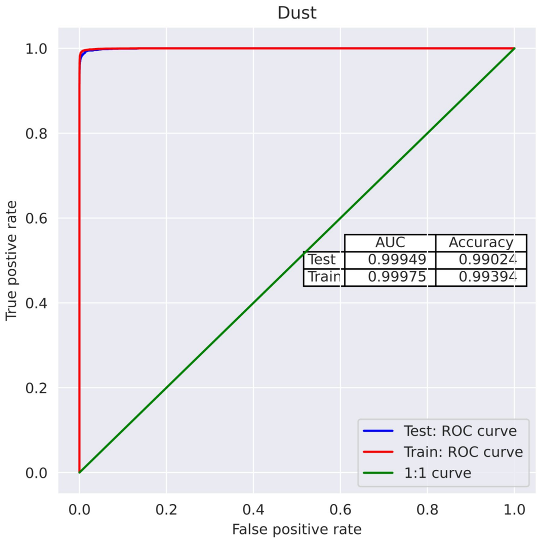


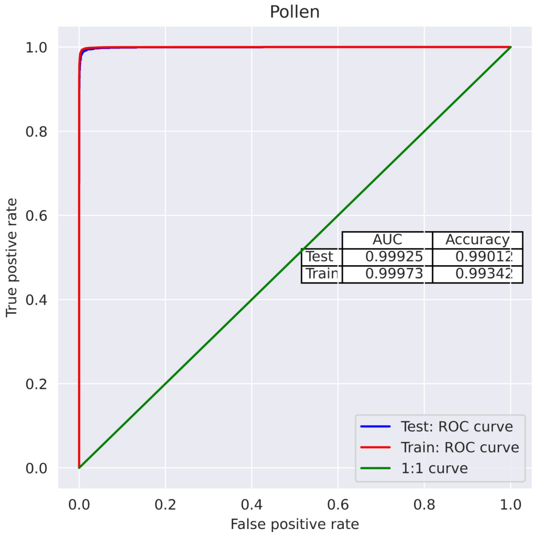
Result: 3 model split

- Bayesian optimization (drop out rate, learning rate, size of layers)



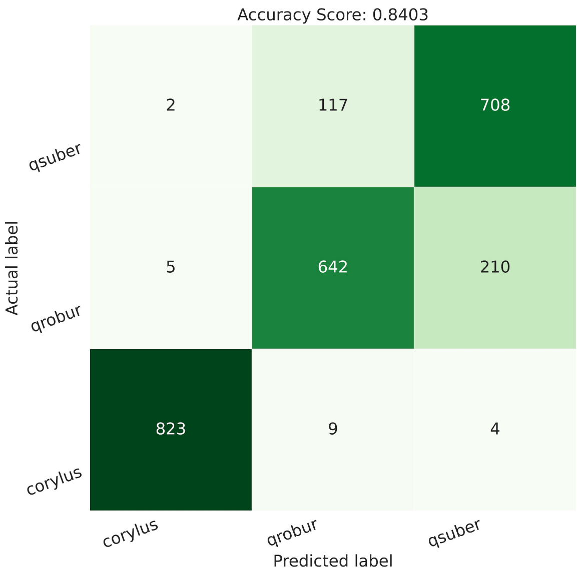




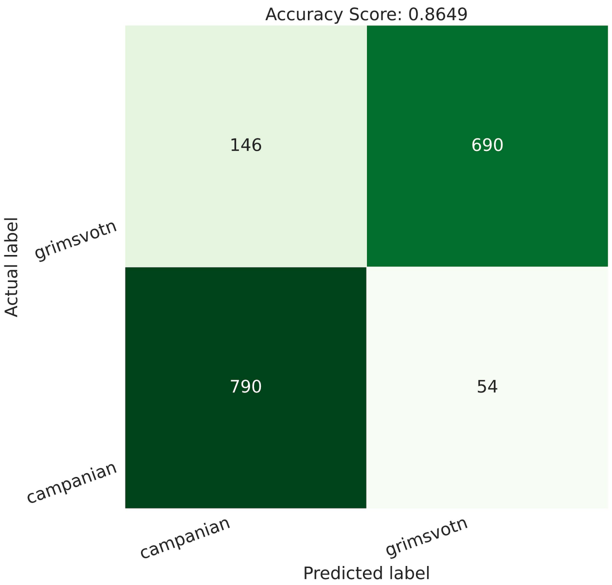


Result: Subcatagorical split

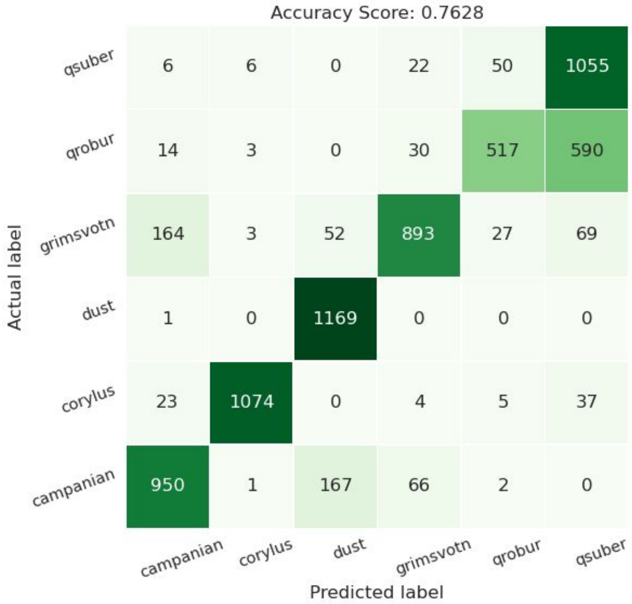
Pollen



Ash



Combined model

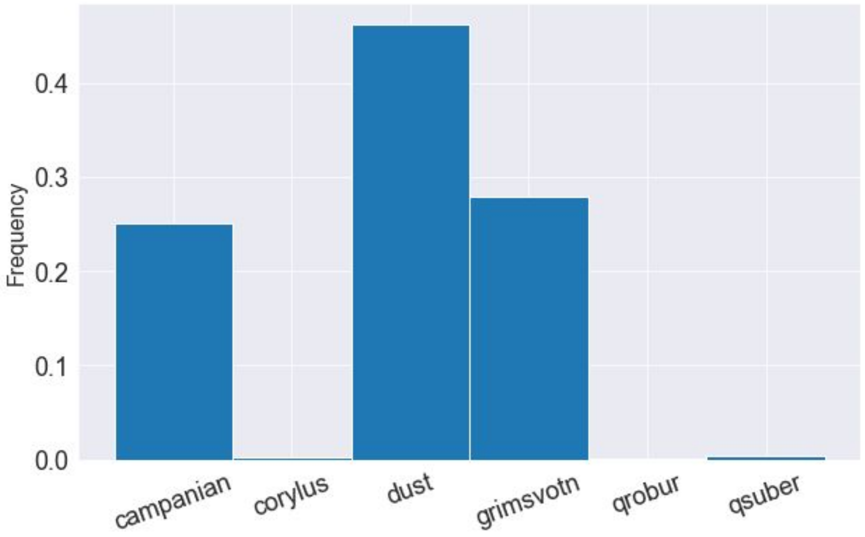


Model performance

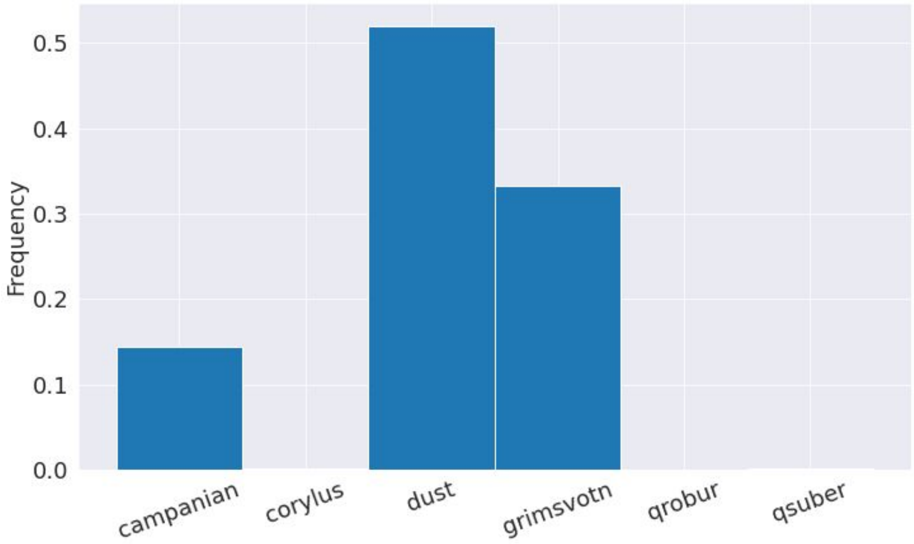
Model	Accuracy
6 split using, only Metadata	0.8347
6 split (Categorical crossentropy loss)	0.8791
6 split (Custom loss function)	0.8426
3 split	0.9806
3 split + subcategories	0.7628

Prediction of the test data

6 splits

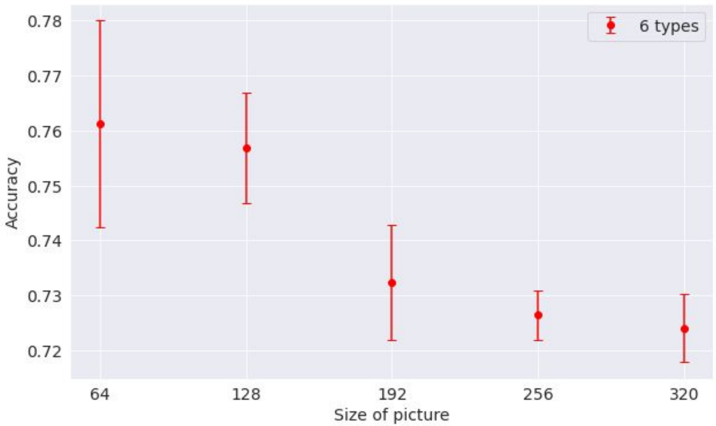


3 split and subcategories

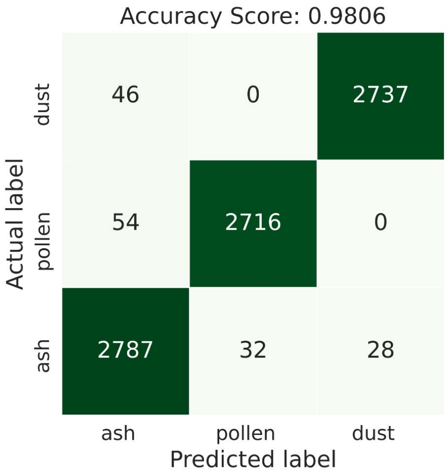


Picture size and inclusion of metadata

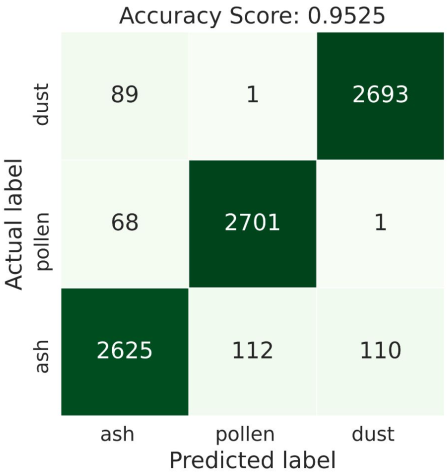
- Chose size 64x64
- With metadata



With metadata

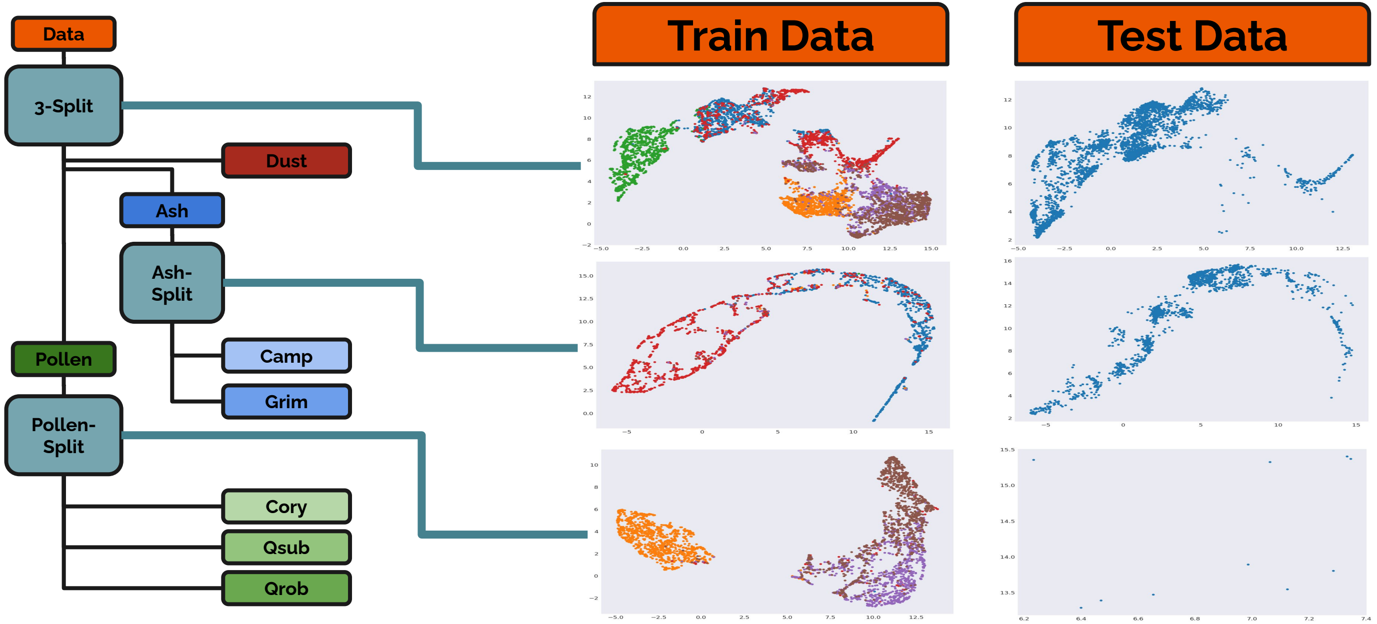


Without metadata



Unsupervised learning (UMAP): 6 model split

Unsupervised learning (UMAP): 3 model split



Conclusion and perspective

What we did:

- Data reduction/ standardization
- LightGBM
- Multiple structures of CNN’s
- Optimization
- Loss function
- Visualization
- Scale of problem

Further ideas:

- Implementation of Loss function both models
- Implementation of optimization
- Handling of unbalanced datasets
- Better data structure so no limitations from hardware.
- Better anomaly detection across models

.

.

.

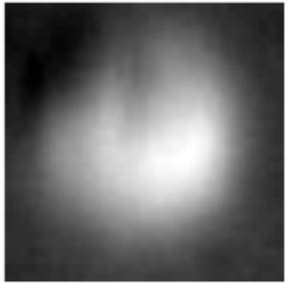
.

(This is a big research project)

6 badly predicted test pictures

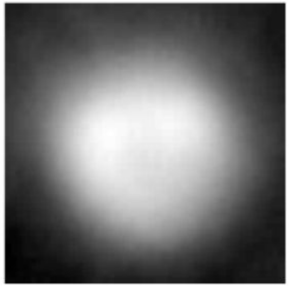
- 3 type classification using only metadata (see appendix)
- 6 picture had no prediction above 50%

Particle ID: GRIP_3046_0_20_1_61191.png
Probability:
Ash: 49.7%
Pollen: 0.69%
Dust: 49.6%

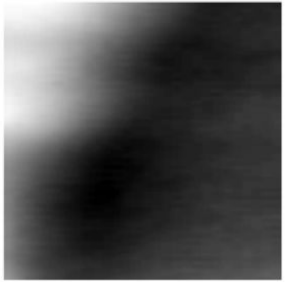


Particle ID: GRIP_3046_0_20_1_55199.png
Probability:
Ash: 49.76%
Pollen: 0.58%
Dust: 49.67%

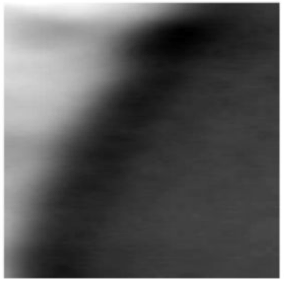
Particle ID: GRIP_3046_0_20_1_21916.png
Probability:
Ash: 46.64%
Pollen: 15.88%
Dust: 37.47%



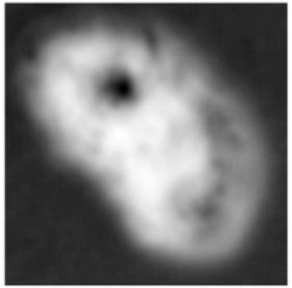
Particle ID: GRIP_3046_0_20_1_24026.png
Probability:
Ash: 49.46%
Pollen: 49.78%
Dust: 0.76%



Particle ID: GRIP_3046_0_20_1_24166.png
Probability:
Ash: 49.99%
Pollen: 49.52%
Dust: 0.49%



Particle ID: GRIP_3046_0_20_1_3247.png
Probability:
Ash: 49.88%
Pollen: 49.73%
Dust: 0.4%



Appendix:

Architecture of used CNN models

3 type classification

Model: "model"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 64, 64, 1)]	0	
conv2d (Conv2D)	(None, 64, 64, 58)	580	input_1[0][0]
max_pooling2d (MaxPooling2D)	(None, 32, 32, 58)	0	conv2d[0][0]
conv2d_1 (Conv2D)	(None, 32, 32, 154)	80542	max_pooling2d[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 16, 16, 154)	0	conv2d_1[0][0]
dropout (Dropout)	(None, 16, 16, 154)	0	max_pooling2d_1[0][0]
flatten (Flatten)	(None, 39424)	0	dropout[0][0]
input_2 (InputLayer)	[(None, 39)]	0	
dense (Dense)	(None, 177)	6978225	flatten[0][0]
dense_1 (Dense)	(None, 220)	8800	input_2[0][0]
concatenate (Concatenate)	(None, 397)	0	dense[0][0] dense_1[0][0]
dense_2 (Dense)	(None, 1263)	502674	concatenate[0][0]
dense_3 (Dense)	(None, 3)	3792	dense_2[0][0]

Total params: 7,574,613
Trainable params: 7,574,613
Non-trainable params: 0

6 type classification

Layer (type)	Output Shape	Param #	Connected to
input_7 (InputLayer)	[(None, 64, 64, 1)]	0	
conv2d_6 (Conv2D)	(None, 64, 64, 32)	320	input_7[0][0]
max_pooling2d_6 (MaxPooling2D)	(None, 32, 32, 32)	0	conv2d_6[0][0]
conv2d_7 (Conv2D)	(None, 32, 32, 64)	18496	max_pooling2d_6[0][0]
max_pooling2d_7 (MaxPooling2D)	(None, 16, 16, 64)	0	conv2d_7[0][0]
dropout_3 (Dropout)	(None, 16, 16, 64)	0	max_pooling2d_7[0][0]
flatten_3 (Flatten)	(None, 16384)	0	dropout_3[0][0]
input_8 (InputLayer)	[(None, 39)]	0	
dense_12 (Dense)	(None, 128)	2097280	flatten_3[0][0]
dense_13 (Dense)	(None, 128)	5120	input_8[0][0]
concatenate_3 (Concatenate)	(None, 256)	0	dense_12[0][0] dense_13[0][0]
dense_14 (Dense)	(None, 1024)	263168	concatenate_3[0][0]
dense_15 (Dense)	(None, 6)	6150	dense_14[0][0]

Total params: 2,390,534
Trainable params: 2,390,534
Non-trainable params: 0

Architecture of used CNN models

Pollen classification

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 64, 64, 1)]	0	
conv2d (Conv2D)	(None, 64, 64, 32)	320	input_1[0][0]
max_pooling2d (MaxPooling2D)	(None, 32, 32, 32)	0	conv2d[0][0]
conv2d_1 (Conv2D)	(None, 32, 32, 64)	18496	max_pooling2d[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 16, 16, 64)	0	conv2d_1[0][0]
dropout (Dropout)	(None, 16, 16, 64)	0	max_pooling2d_1[0][0]
flatten (Flatten)	(None, 16384)	0	dropout[0][0]
input_2 (InputLayer)	[(None, 39)]	0	
dense (Dense)	(None, 128)	2097280	flatten[0][0]
dense_1 (Dense)	(None, 128)	5120	input_2[0][0]
concatenate (Concatenate)	(None, 256)	0	dense[0][0] dense_1[0][0]
dense_2 (Dense)	(None, 1024)	263168	concatenate[0][0]
dense_3 (Dense)	(None, 3)	3075	dense_2[0][0]

Total params: 2,387,459
Trainable params: 2,387,459
Non-trainable params: 0

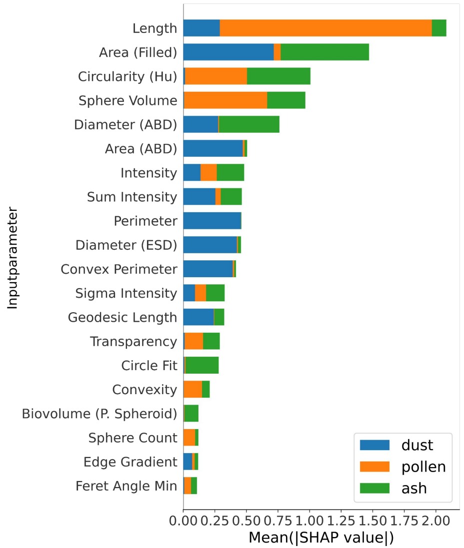
Ash classification

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 64, 64, 1)]	0	
conv2d (Conv2D)	(None, 64, 64, 32)	320	input_1[0][0]
max_pooling2d (MaxPooling2D)	(None, 32, 32, 32)	0	conv2d[0][0]
conv2d_1 (Conv2D)	(None, 32, 32, 64)	18496	max_pooling2d[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 16, 16, 64)	0	conv2d_1[0][0]
conv2d_2 (Conv2D)	(None, 16, 16, 64)	36928	max_pooling2d_1[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 8, 8, 64)	0	conv2d_2[0][0]
conv2d_3 (Conv2D)	(None, 8, 8, 64)	36928	max_pooling2d_2[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 4, 4, 64)	0	conv2d_3[0][0]
dropout (Dropout)	(None, 4, 4, 64)	0	max_pooling2d_3[0][0]
flatten (Flatten)	(None, 1024)	0	dropout[0][0]
input_2 (InputLayer)	[(None, 39)]	0	
dense (Dense)	(None, 128)	131200	flatten[0][0]
dense_1 (Dense)	(None, 128)	5120	input_2[0][0]
concatenate (Concatenate)	(None, 256)	0	dense[0][0] dense_1[0][0]
dense_2 (Dense)	(None, 1024)	263168	concatenate[0][0]
dense_3 (Dense)	(None, 2)	2050	dense_2[0][0]

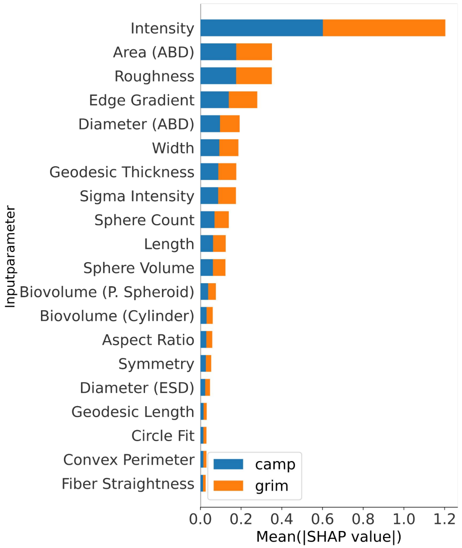
Total params: 494,210
Trainable params: 494,210
Non-trainable params: 0

Feature importance of metadata using SHAP-values

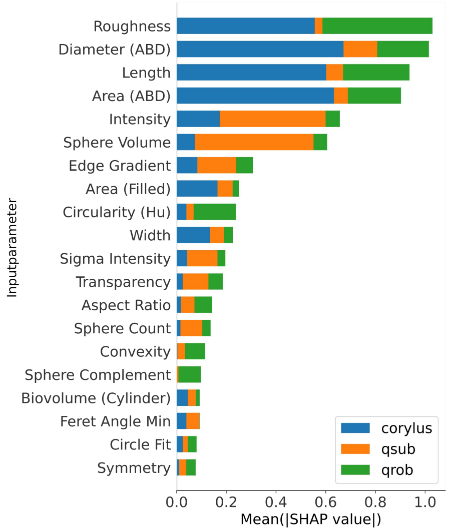
3 type classification



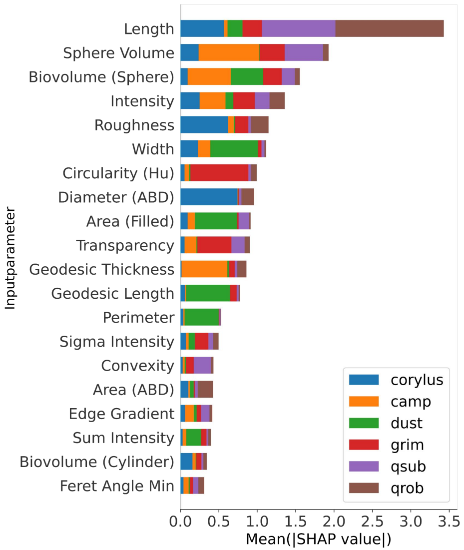
Ash type classification



Pollen type classification

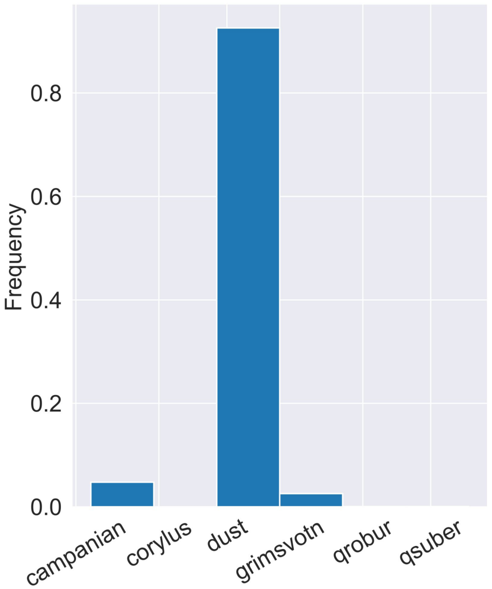


6 type classification

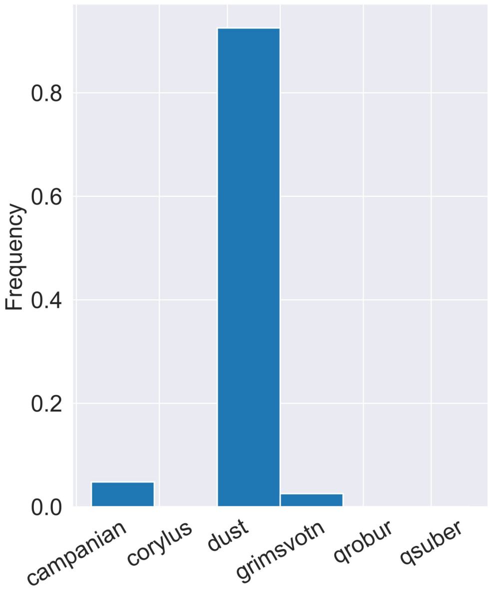


Prediction of the test data using only metadata

3 split and subcategories

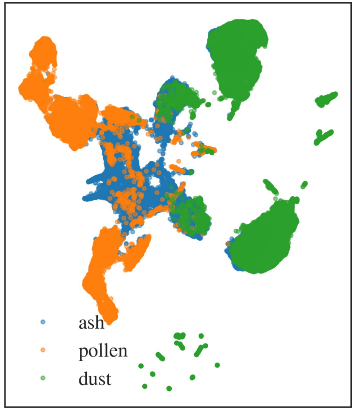


Pure 6 type classification

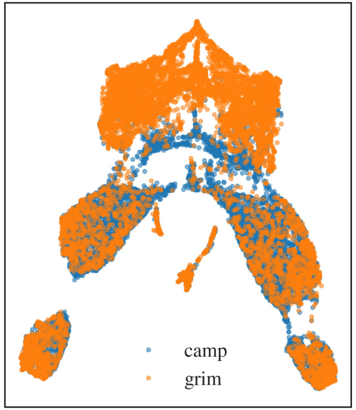


Unsupervised learning on metadata (UMAP)

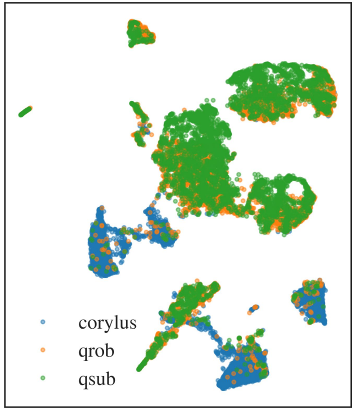
3 type classification



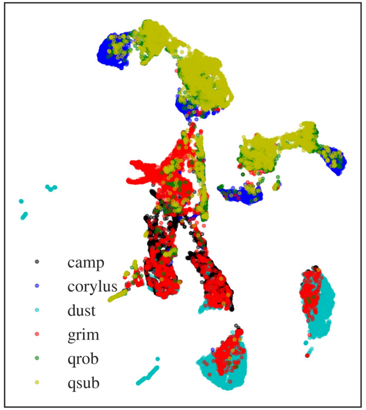
Ash classification



Pollen type classification

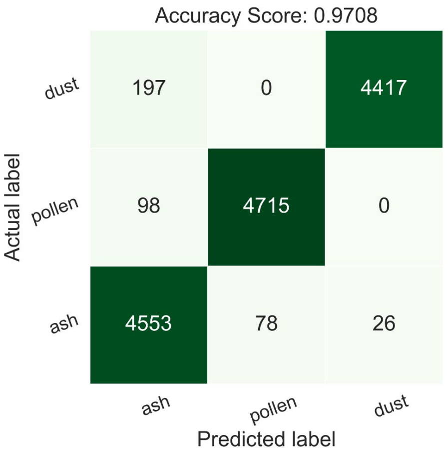


6 type classification

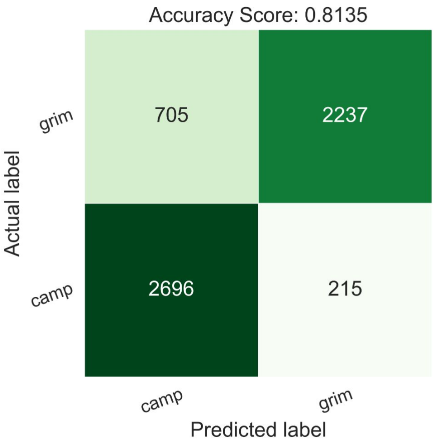


Confusion matrices from training on only the metadata

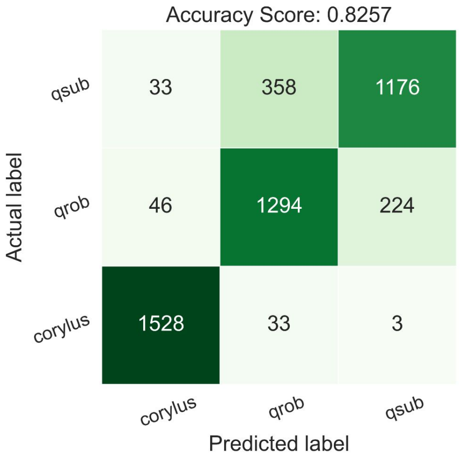
3 type classification



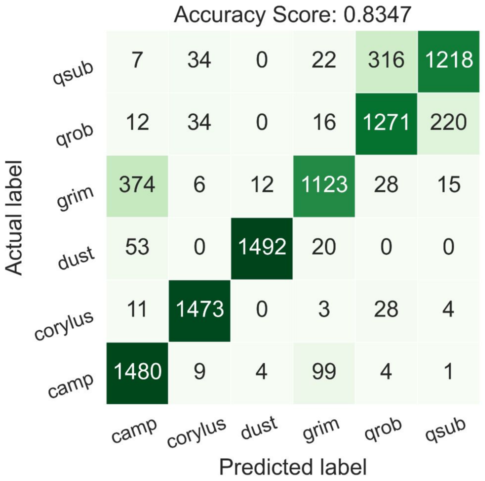
Ash classification



Pollen type classification

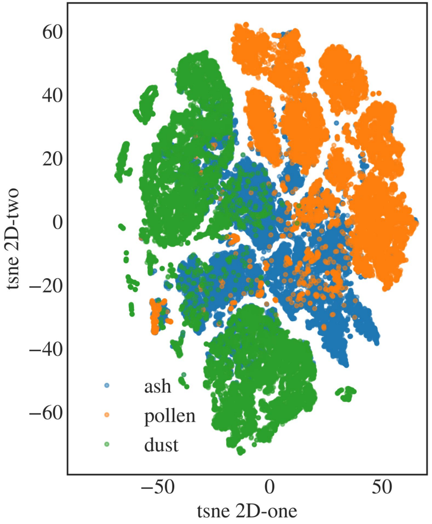


6 type classification

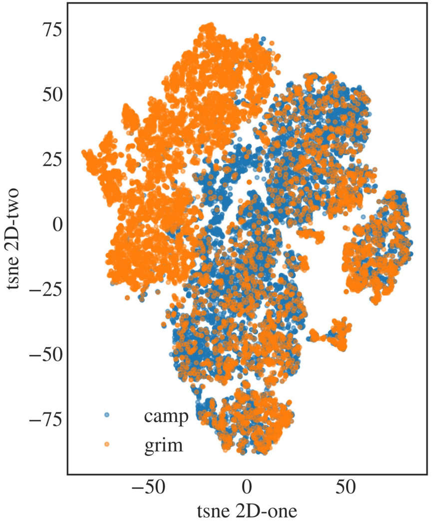


Unsupervised learning on metadata (t-SNE)

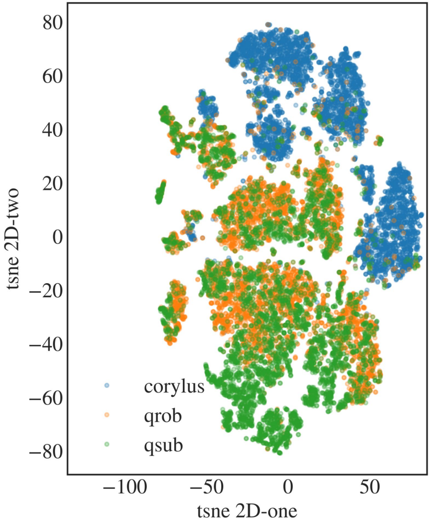
3 type classification



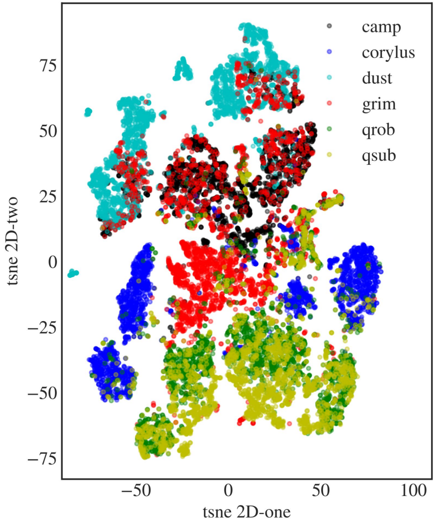
Ash type classification



Pollen type classification

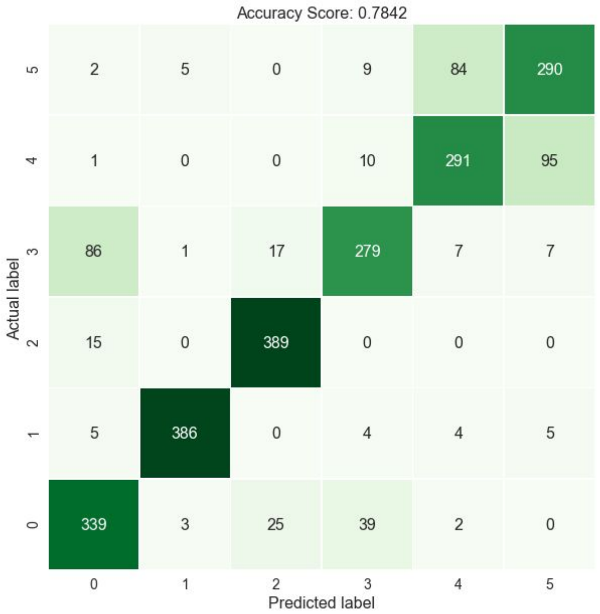


6 type classification



Contrast and Sharpness on pictures

Before:



After: