Applied ML
Final Project

Detection of Insolubles in Real Ice Core Data

Amalie Mygind, Marcus Nygaard, Michala Jensen, Søren Langkilde & Ulrik Friis-Jensen

UNIVERSITY OF COPENHAGEN

Outline

- Introduction
- Presentation of models
  - K-means
  - NN
  - CNN
  - Ensemble model
  - Mixed model
- Evaluation of models
- Prediction on real data

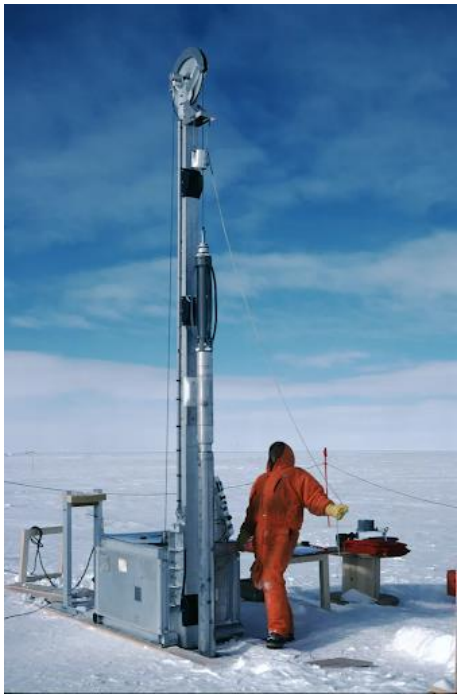
Image:
http://lindseynicholson.org/wp-content/uploads/2018/04/1_Core_mount-hunter-ice-core-preserves-climate-and-environmental-record-eos_org.jpg

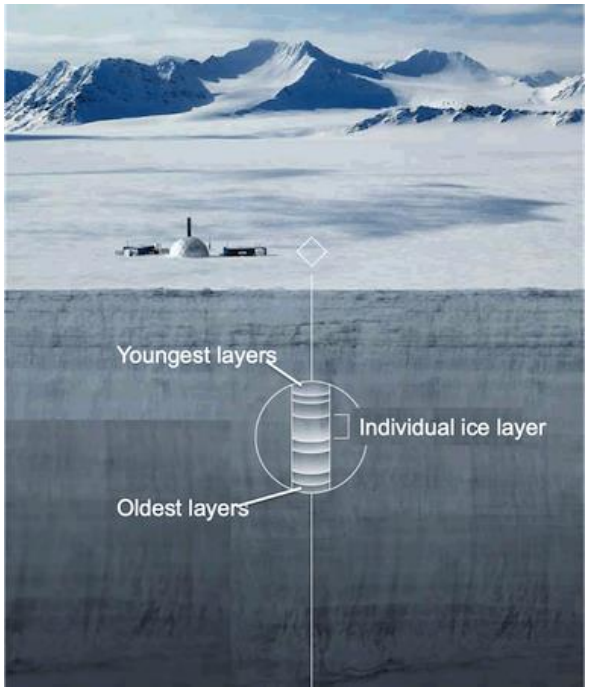
Introduction

- Goal: To classify particles in ice core data using machine learning.

- Approach: Utilizing all available data to create machine learning algorithms for classification and combining them in efforts to improve accuracy.

The ice core data



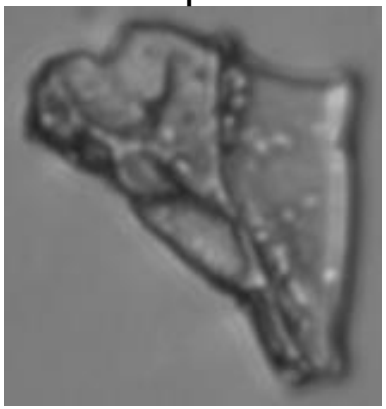


1 datapoint
=
1 image + meta data

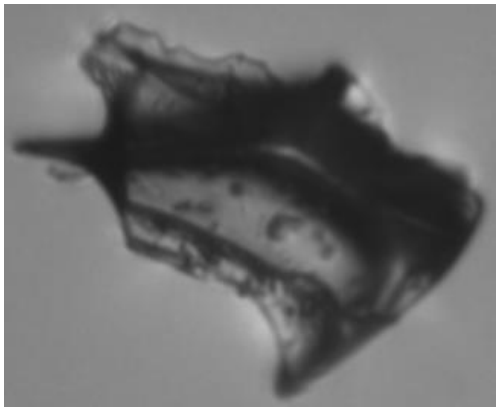
Image:
https://theconversation.com/explainer-what-are-ice-cores-24302
https://www.chemedx.org/article/ice-cores-stable-isotopes-climate-change-and-chemistry
https://www.crowcanyon.org/index.php/ice-core-studies

The data used for training

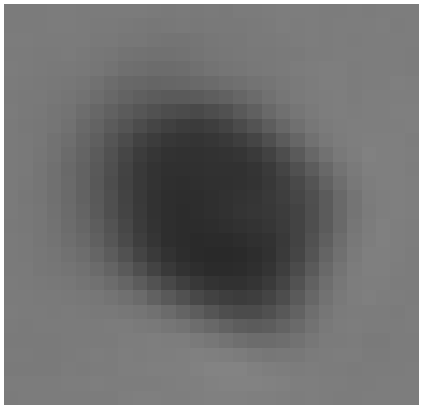
Campanian



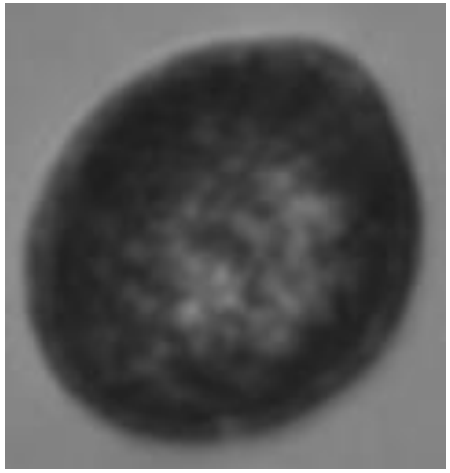
Grimsvotn



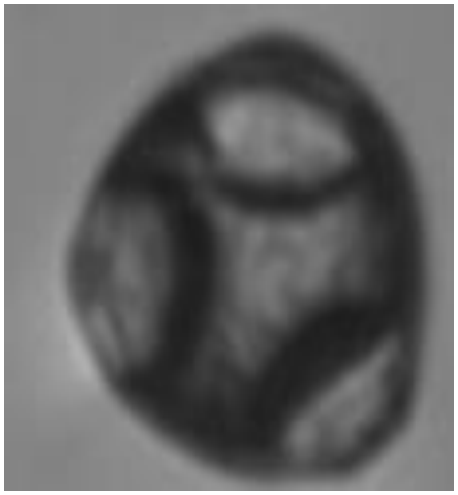
Dust



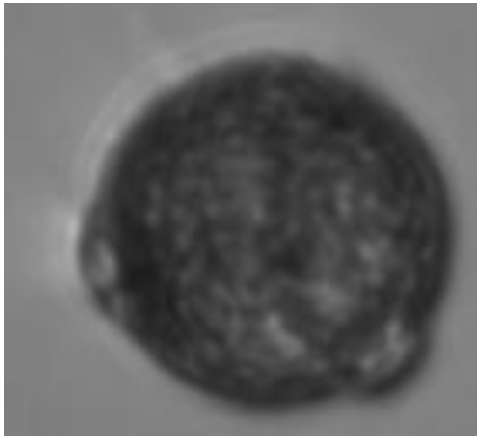
Qsuber



Corylus



Qrobur



Meta data

	#
Total features	63
Useable features	39
Unused feature	24

Microscope used for data capture

Models

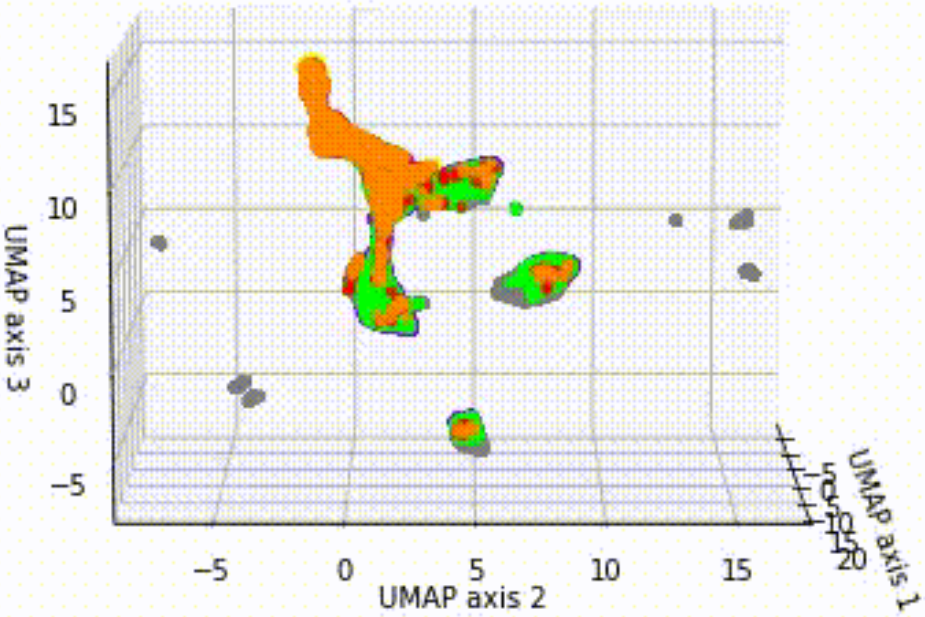
K-means, NN, CNN, ensemble model & mixed model

UNIVERSITY OF COPENHAGEN

UMAP and clustering on meta data

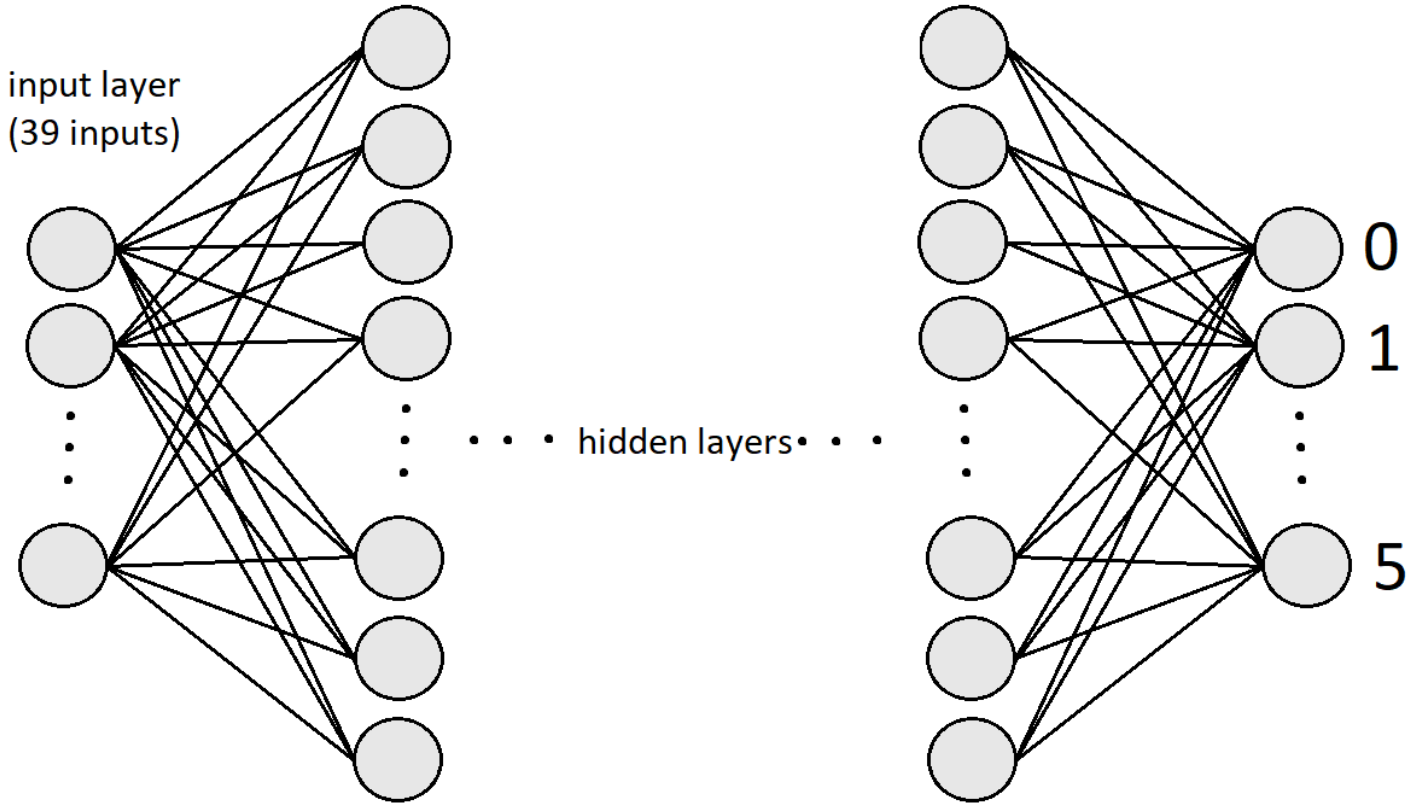
- UMAP
  - Dimensionality reduction
  - 39 feature to 3D
  - Considered 200 neighbors
- Clustering by k-means
- Performance
  - ~ 63%

- = campanian
- = corylus
- = dust
- = grimsvotn
- = qrobur
- = qsuber



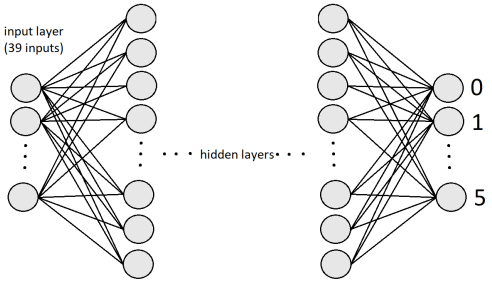
Neural Network on meta data

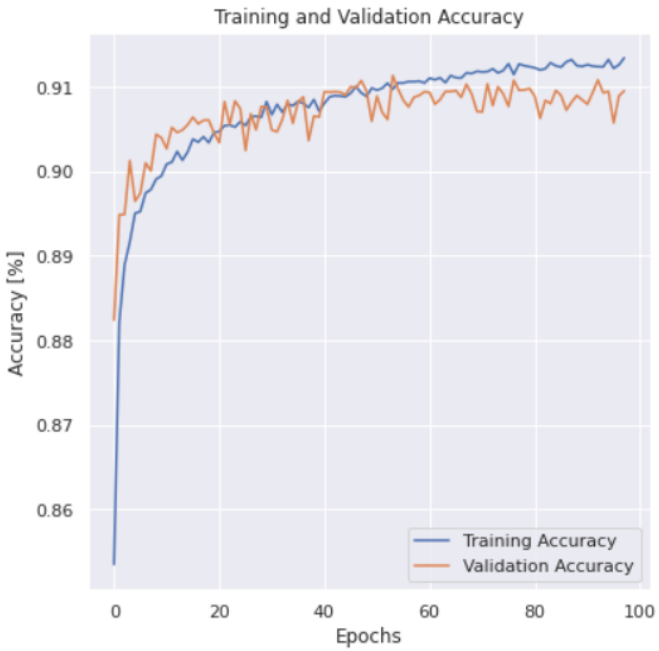
- 39 inputs, 4 hidden layers, 50 neurons in each



Neural Network on meta data

- 39 inputs, 4 hidden layers, 50 neurons in each
- Performance
  - Almost perfect dust predicter
- Loss function: 0.26 (categorical cross entropy)

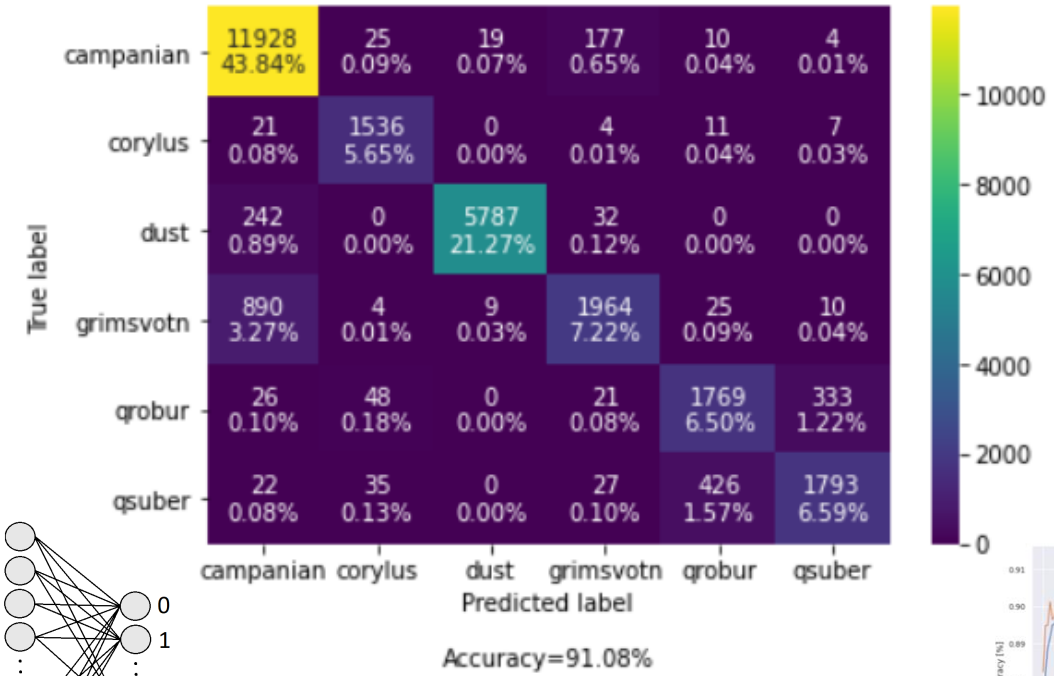


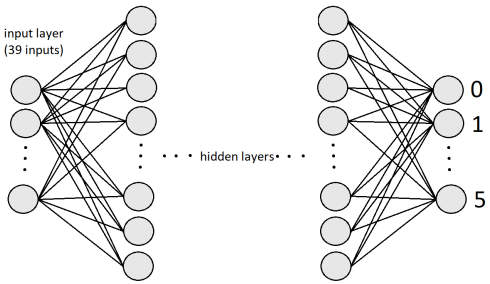


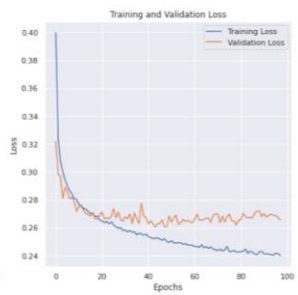
Neural Network on meta data

- 39 inputs, 4 hidden layers, 50 neurons in each
- Performance
  - Almost perfect dust predicter
- Loss function: 0.26 (categorical cross entropy)

NN on meta data

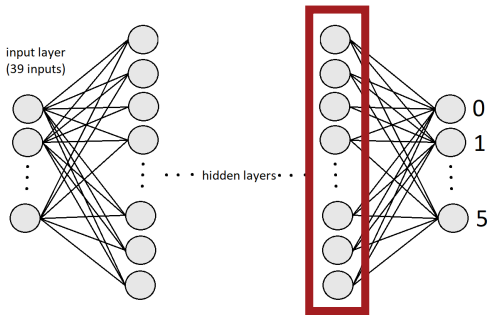




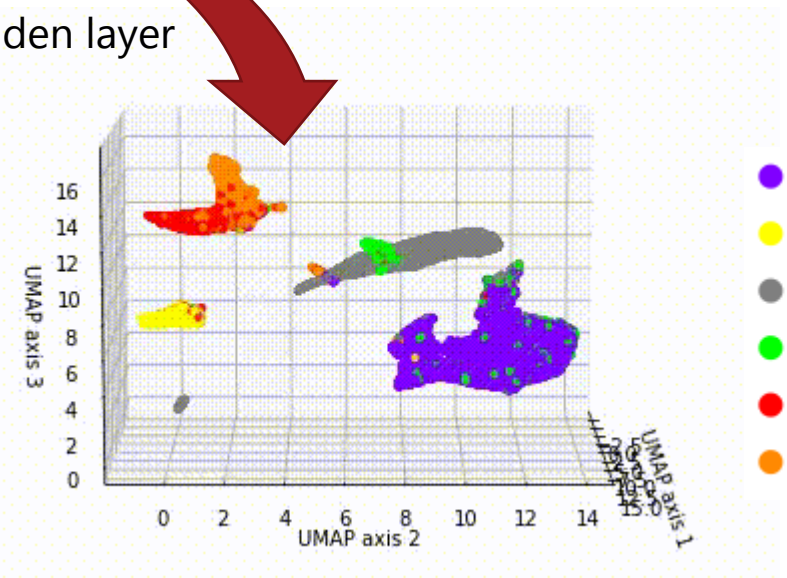


Neural Network on meta data

- 39 inputs, 4 hidden layers, 50 neurons in each
- Performance
  - Almost perfect dust predicter
- Loss function: 0.26 (categorical cross entropy)
- Primary classes group together

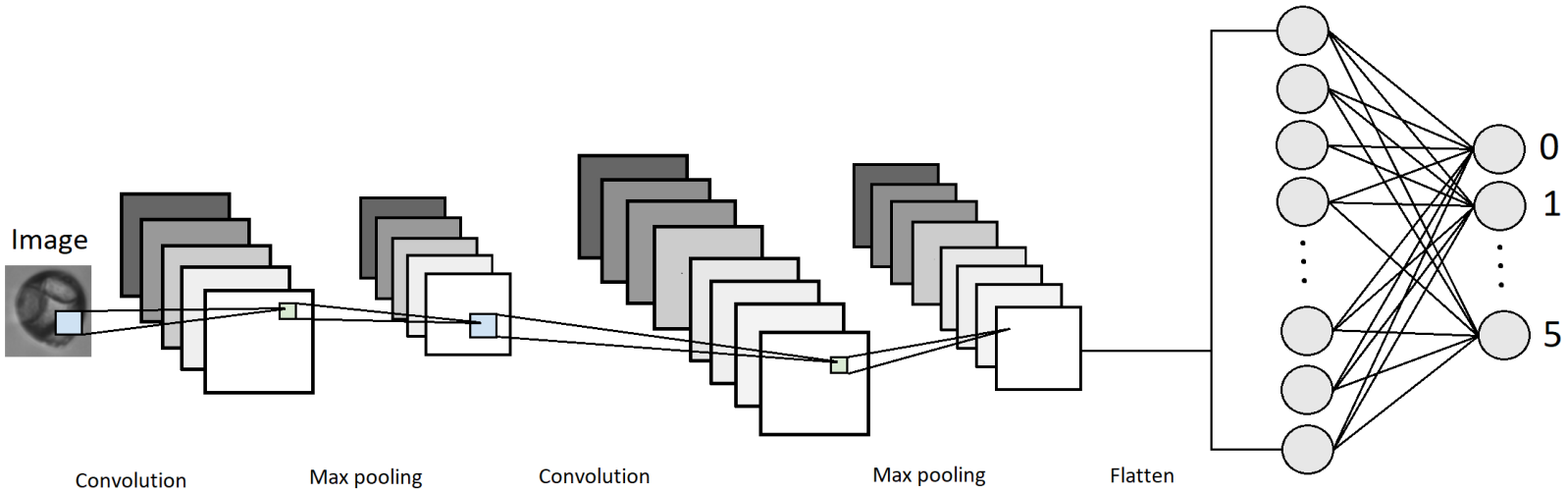


UMAP on last hidden layer

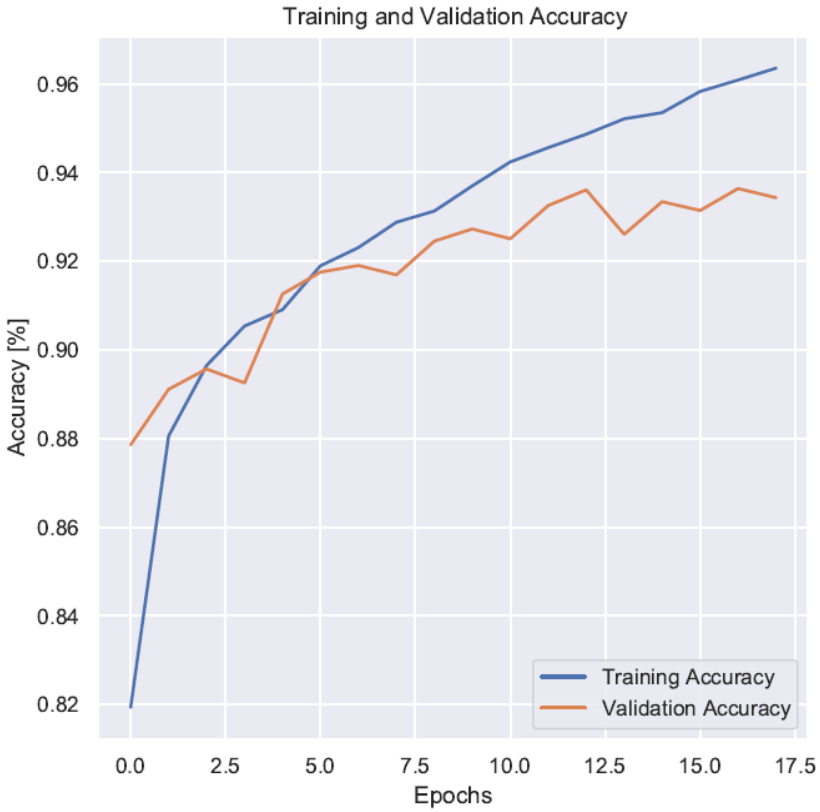


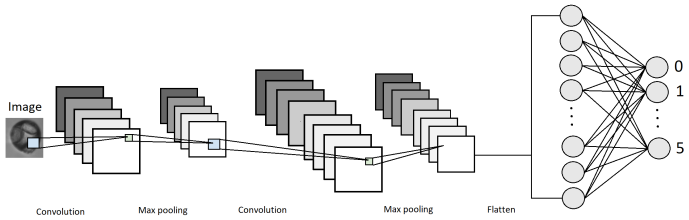
CNN on image data

- Convolutional Neural Network
  - 4 conv layers w/ max pool, 2 drop out layers
- Bayesian optimization of hyperparameters
  - # filters, drop out rate, kernel regularizer factor

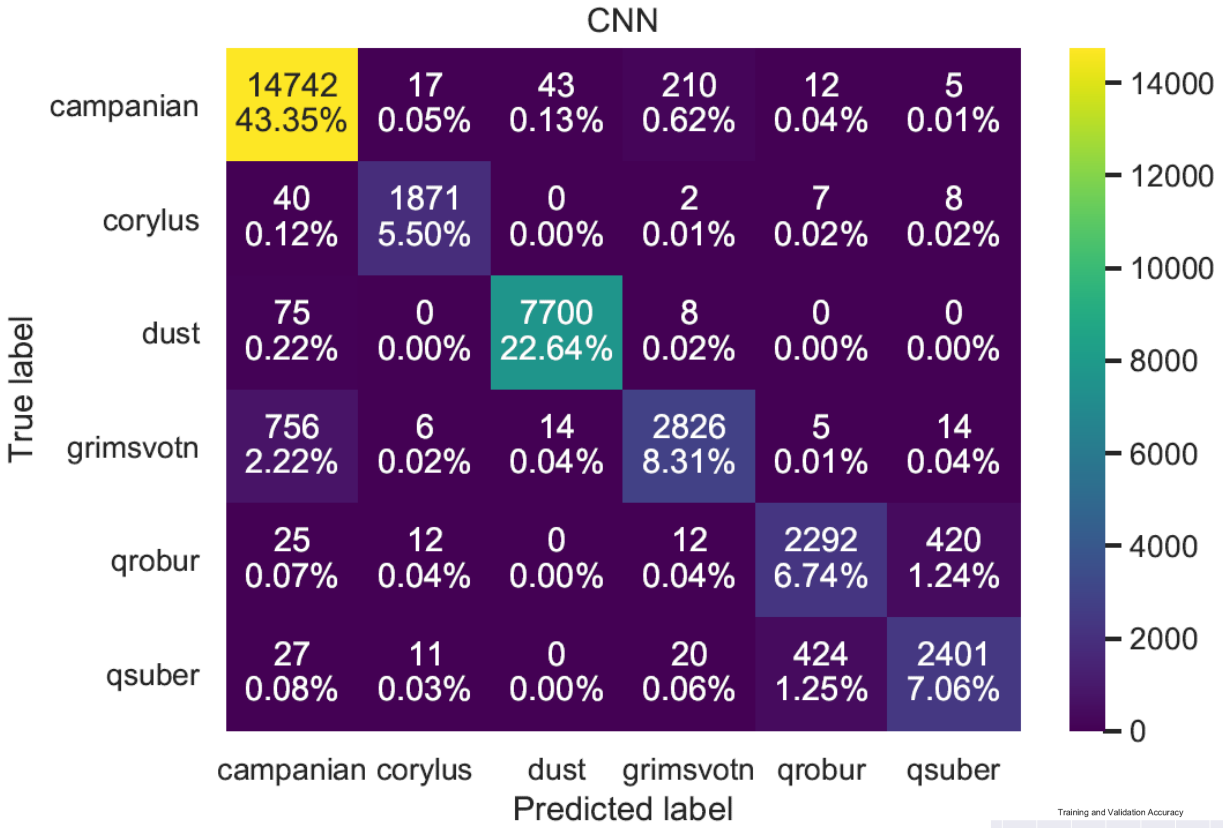


CNN on image data

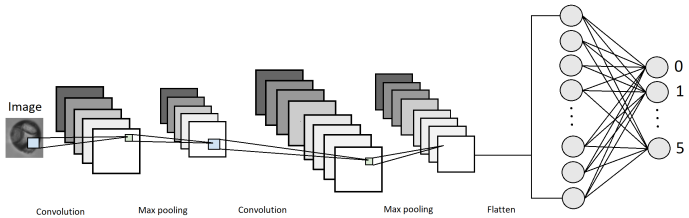




CNN on image data

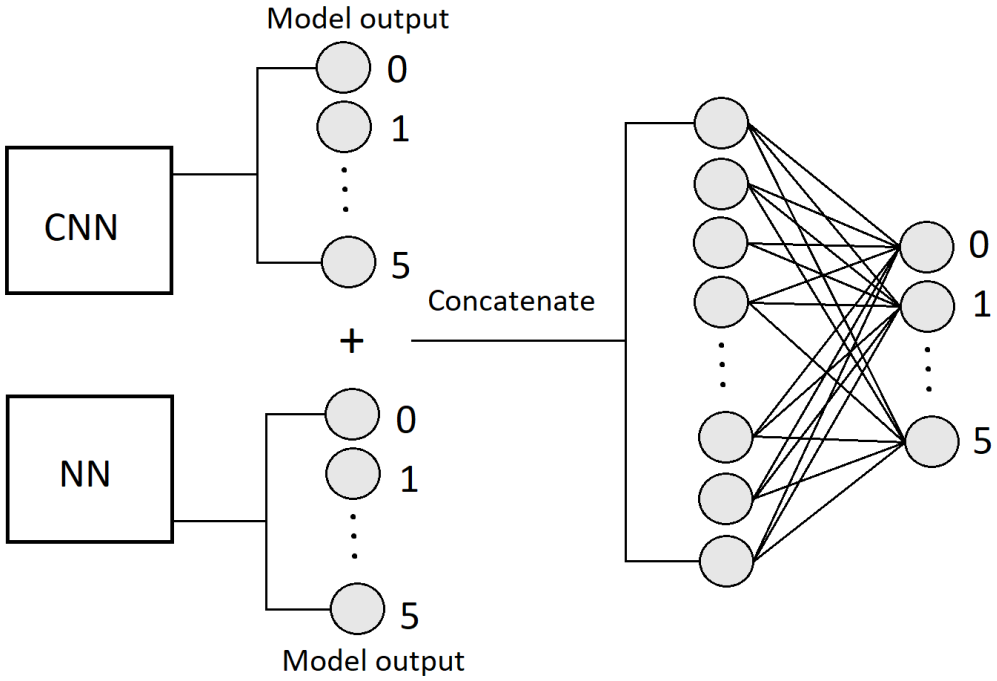


Accuracy=93.61%

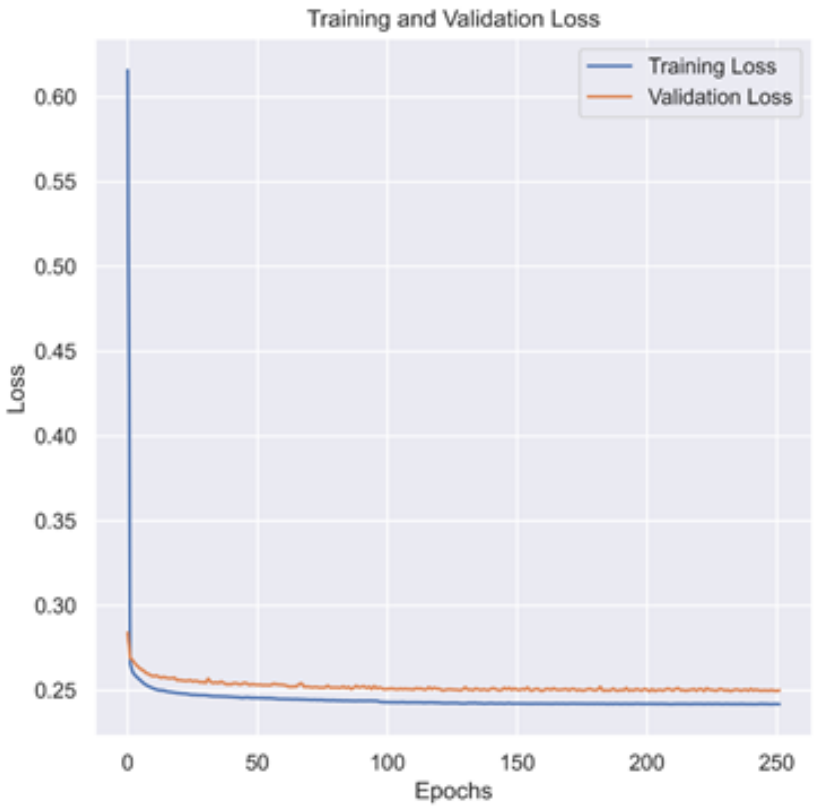


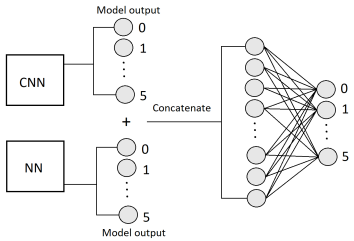
Ensemble model

- Combine output from CNN and NN into a single input vector

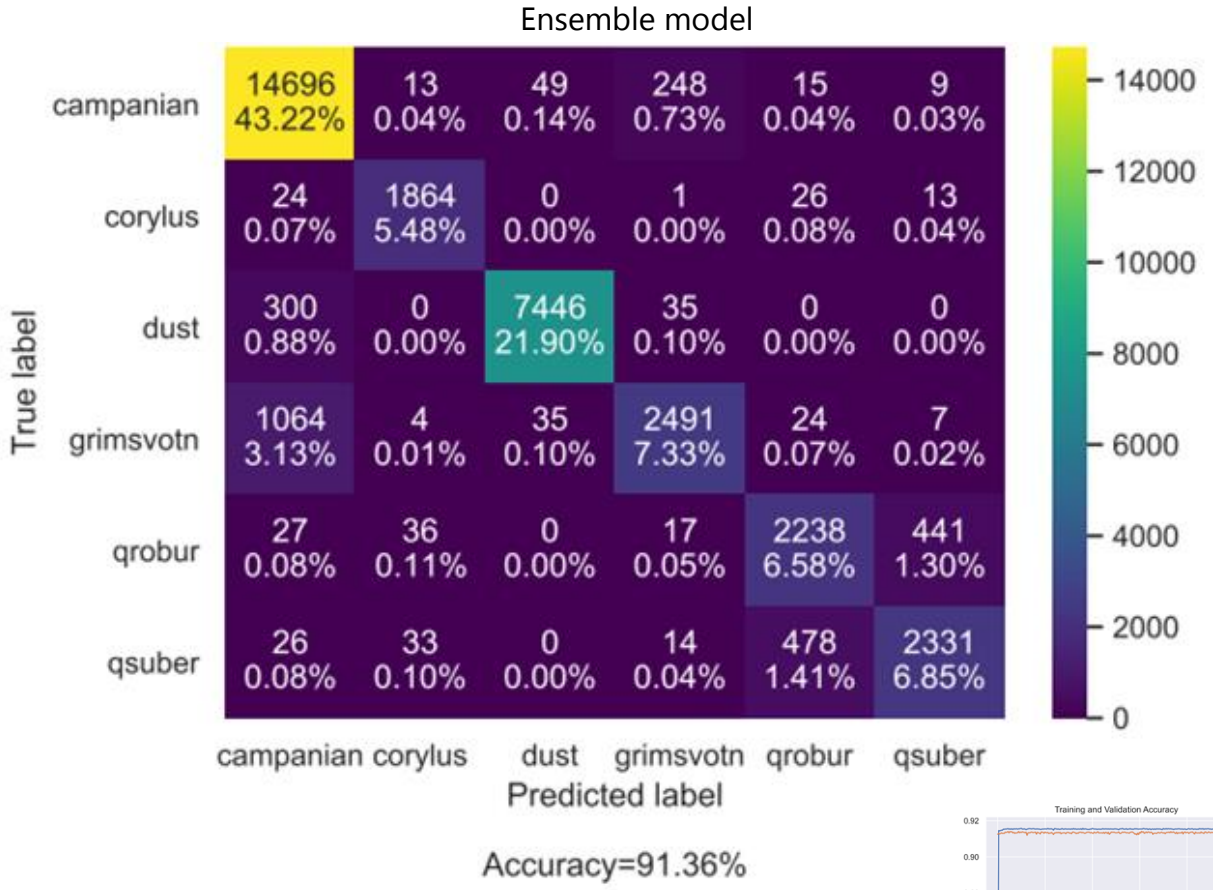


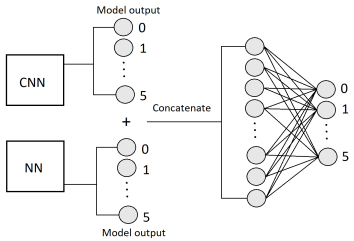
Ensemble model

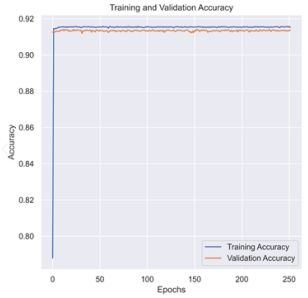


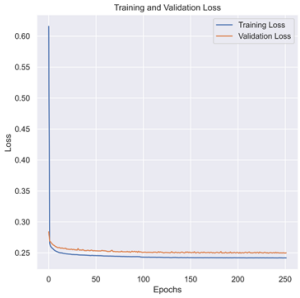


Ensemble model



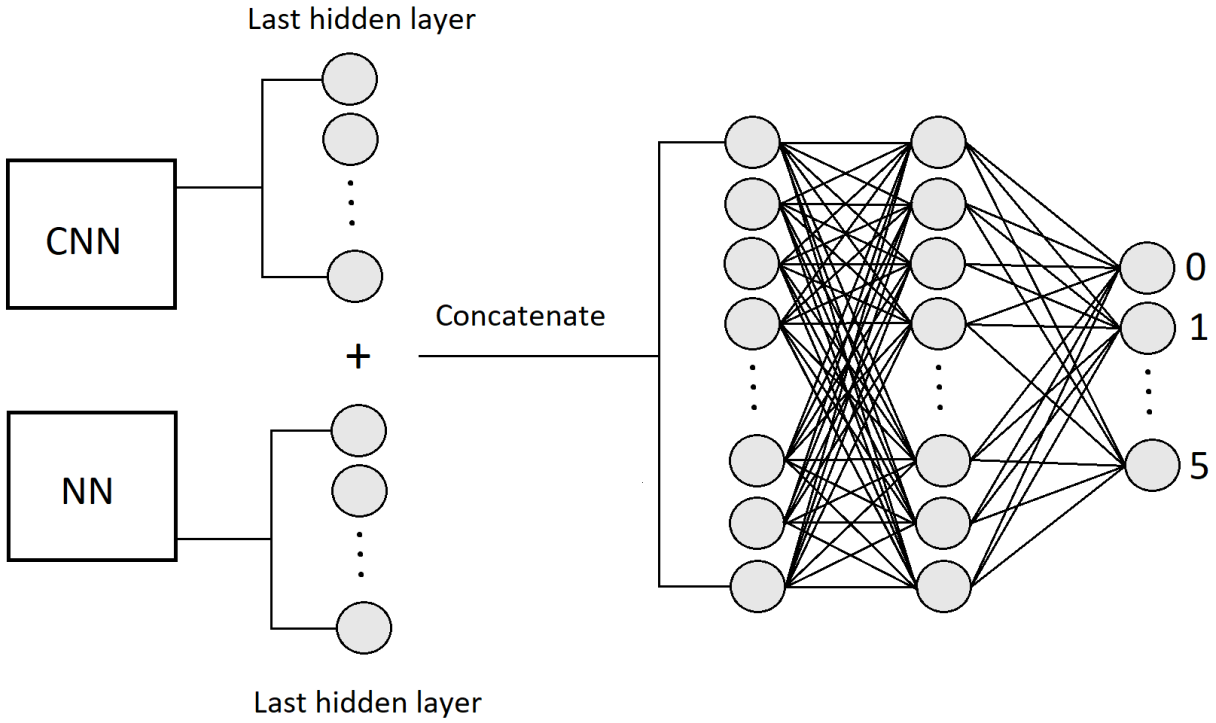






Mixed model

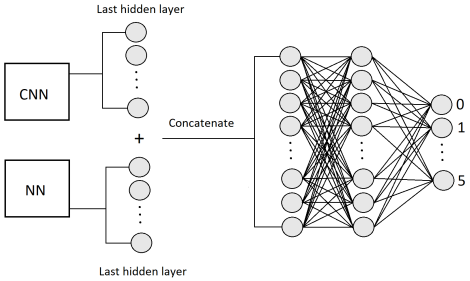
- Two inputs
  - Image data
  - Meta data
- Custom data loader to handle different data types



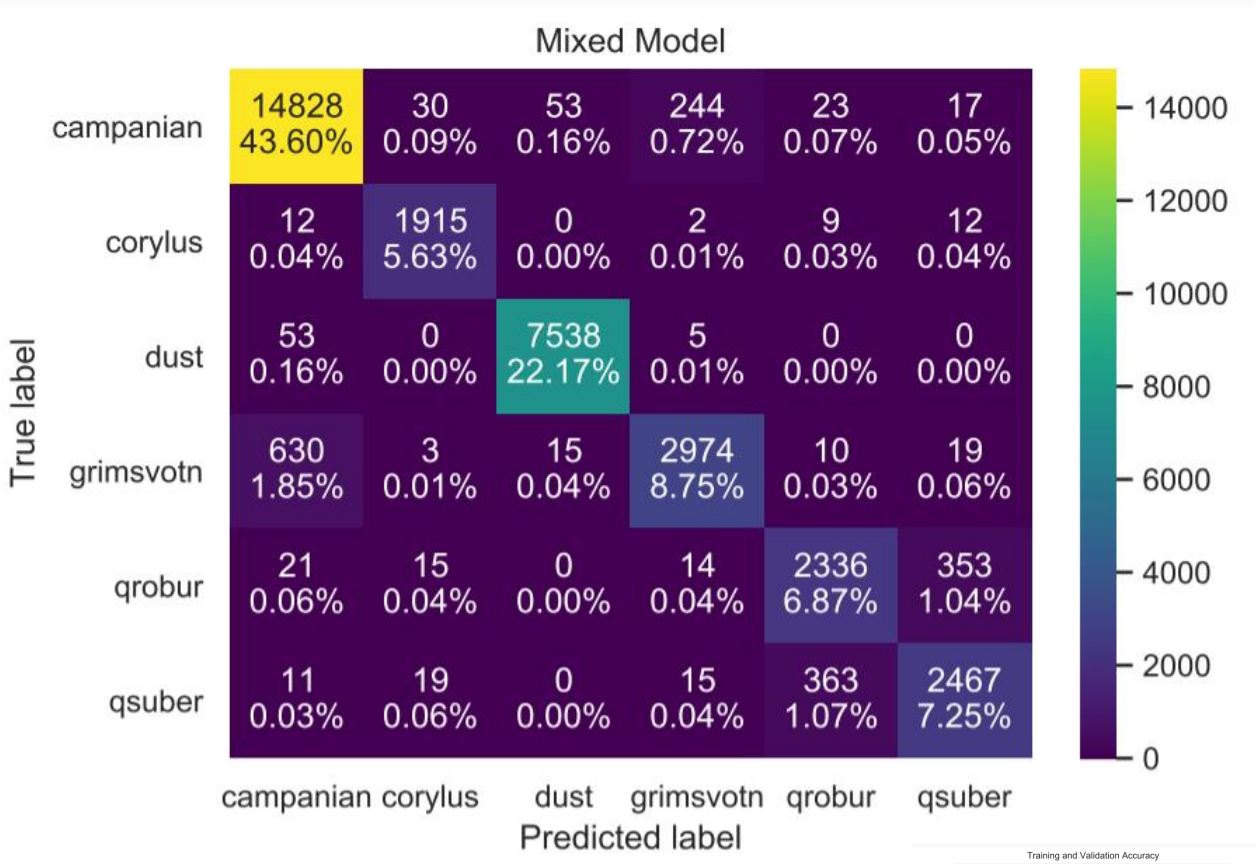
Mixed model

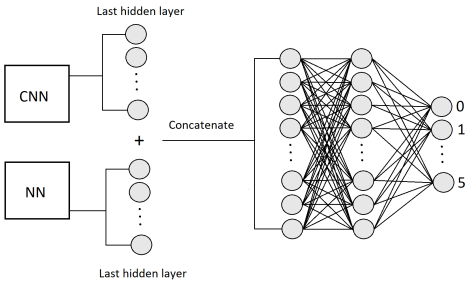

Training and Validation Accuracy
Accuracy
Epochs
Training Accuracy
Validation Accuracy

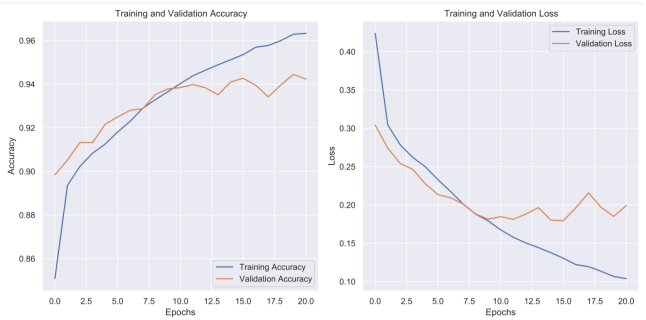

Training and Validation Loss
Loss
Epochs
Training Loss
Validation Loss


Last hidden layer
CNN
+
NN
Concatenate
Last hidden layer
0
1
5

Mixed model


Mixed Model
campanian
corylus
dust
grimsvotn
qrobur
qsuber
True label
14828 43.60%
30 0.09%
53 0.16%
244 0.72%
23 0.07%
17 0.05%
12 0.04%
1915 5.63%
0 0.00%
2 0.01%
9 0.03%
12 0.04%
53 0.16%
0 0.00%
7538 22.17%
5 0.01%
0 0.00%
0 0.00%
630 1.85%
3 0.01%
15 0.04%
2974 8.75%
10 0.03%
19 0.06%
21 0.06%
15 0.04%
0 0.00%
14 0.04%
2336 6.87%
353 1.04%
11 0.03%
19 0.06%
0 0.00%
15 0.04%
363 1.07%
2467 7.25%
14000
12000
10000
8000
6000
4000
2000
0
campanian corylus dust grimsvotn qrobur qsuber
Predicted label
Accuracy=94.27%


Last hidden layer
CNN
+
NN
Concatenate
0
1
5
Last hidden layer

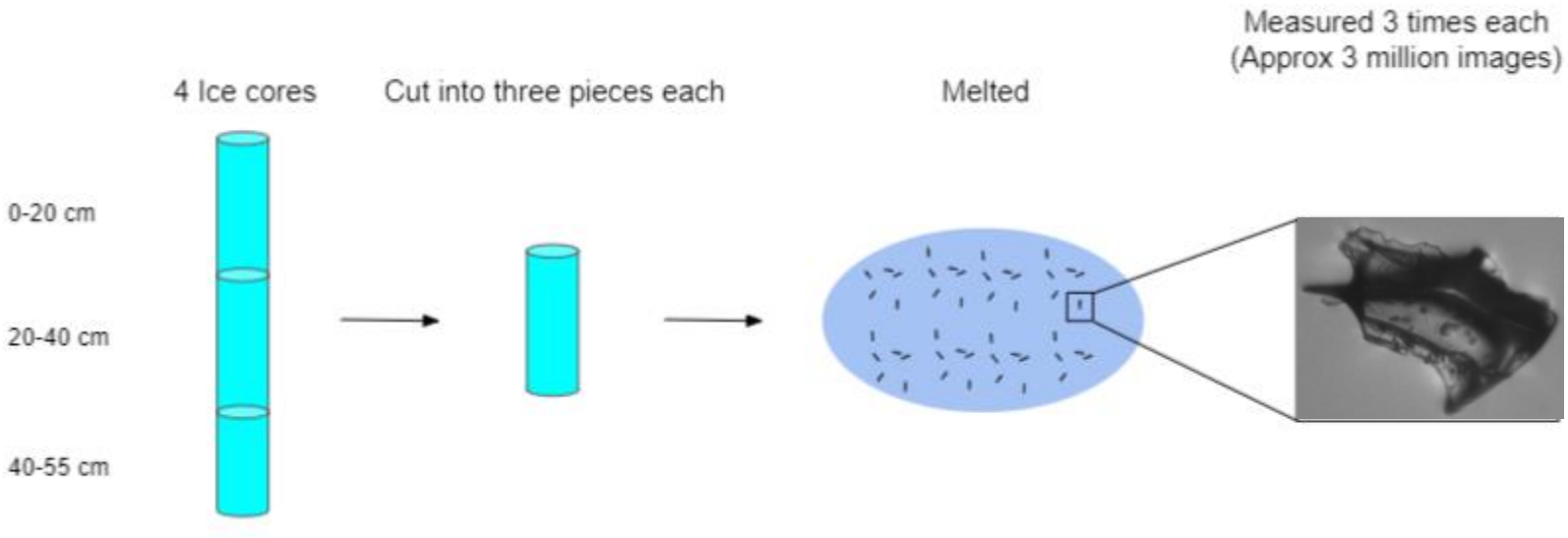

Training and Validation Accuracy
Accuracy
Epochs
Training Accuracy
Validation Accuracy
Training and Validation Loss
Loss
Epochs
Training Loss
Validation Loss

Type of model	Trainable parameters	Cross Entropy Loss	Accuracy	Evaluation
NN	9,956	0.26	91.08%	Simple, Medium performance
CNN	4,022,554	0.19	93.61%	High complexity, Good performance
Mixed model	4,061,098	0.18	94.27%	High complexity, Best performance
Ensemble model	276	0.25	91.36%	Simple (Requires other models), Medium performance

Real ice core data

UNIVERSITY OF COPENHAGEN

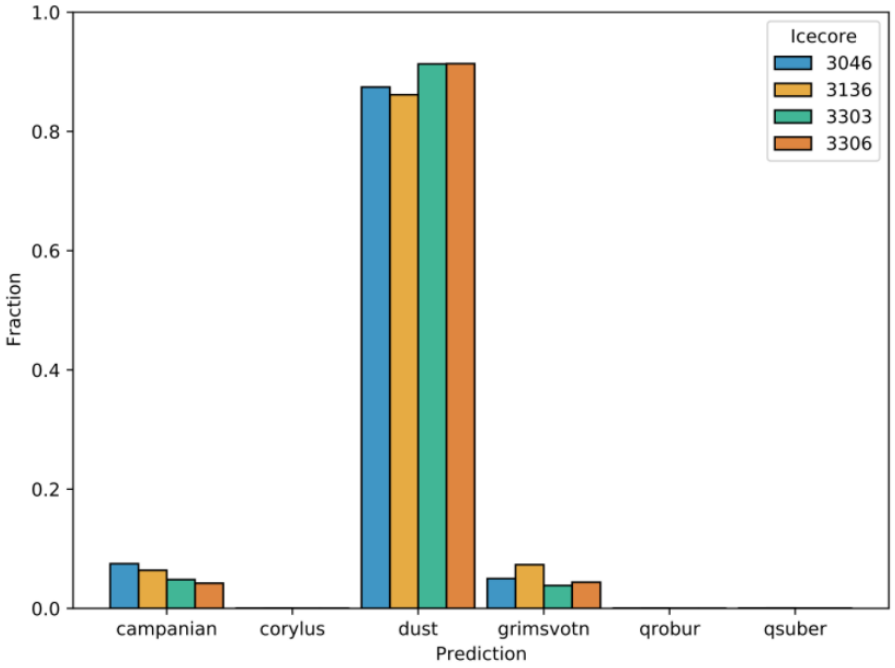
The ice core data used for testing



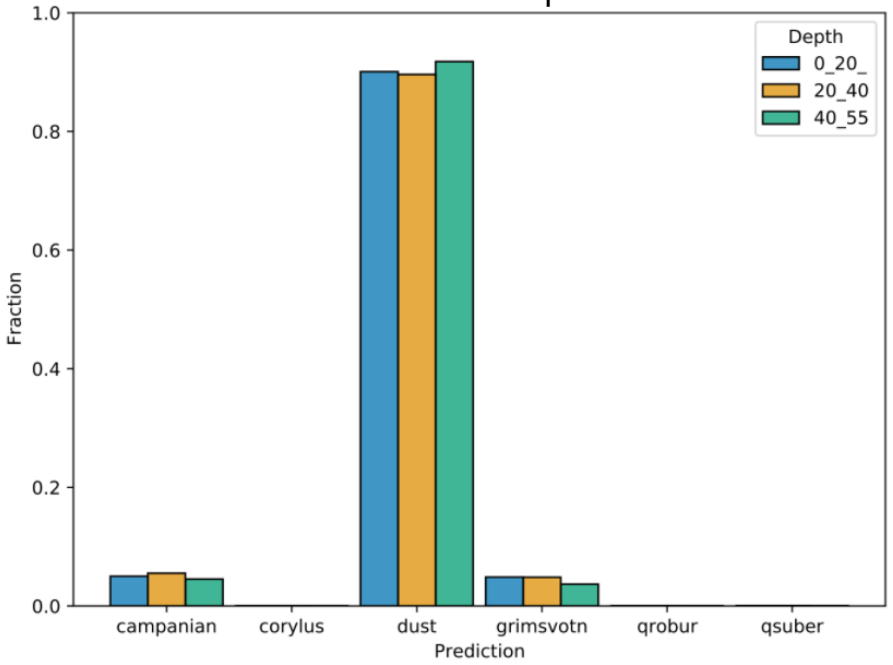
Model predictions on test data

- Predictions using mixed model
- Expected 99% dust, 1% ash/other, 0% pollen
- Predicted 90.4% dust, 9.6% ash, 0.035% pollen

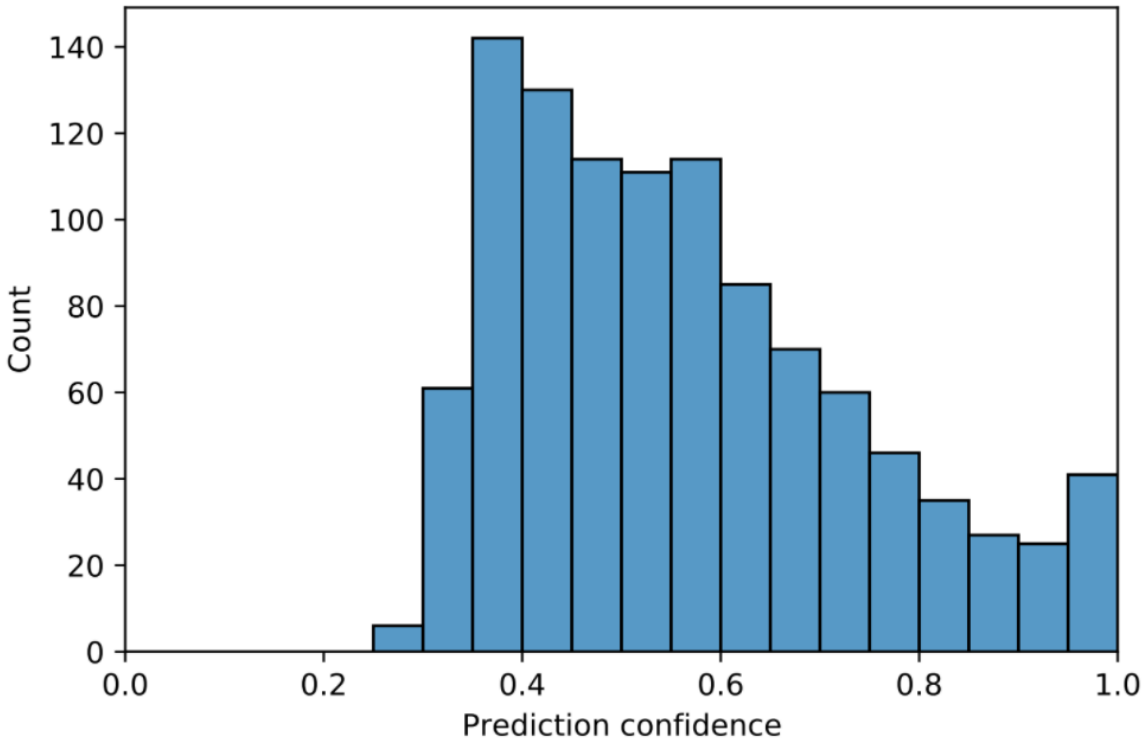
Different ice cores



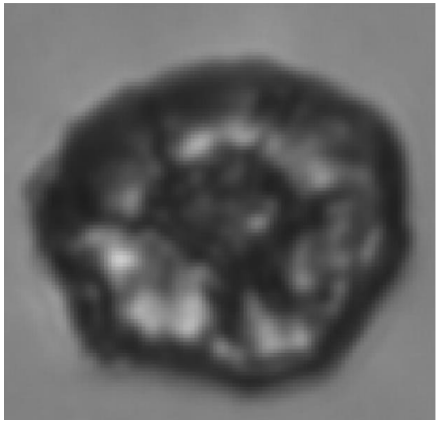
Different depths



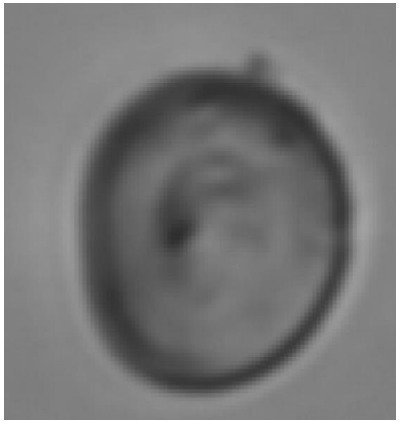
Confidence level on pollen predictions



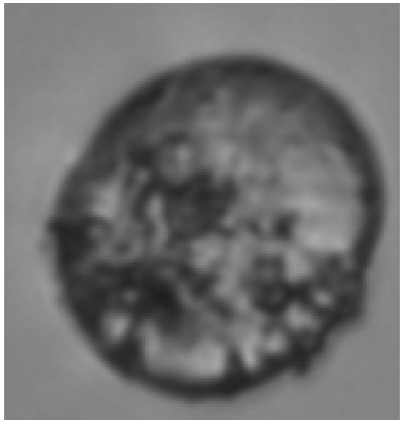
6 most interesting images in test data



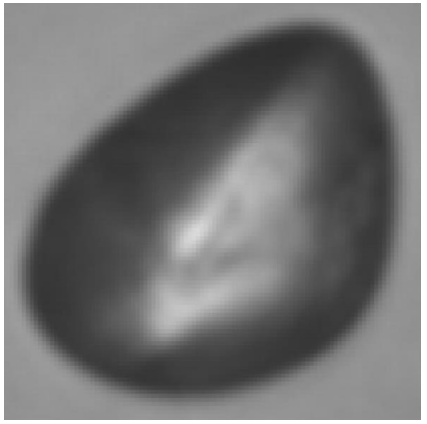
test_GRIP_31may2021/GRIP_raw/GRIP_3136_20_40_3/GRIP_3136_20_40_3_16380.png



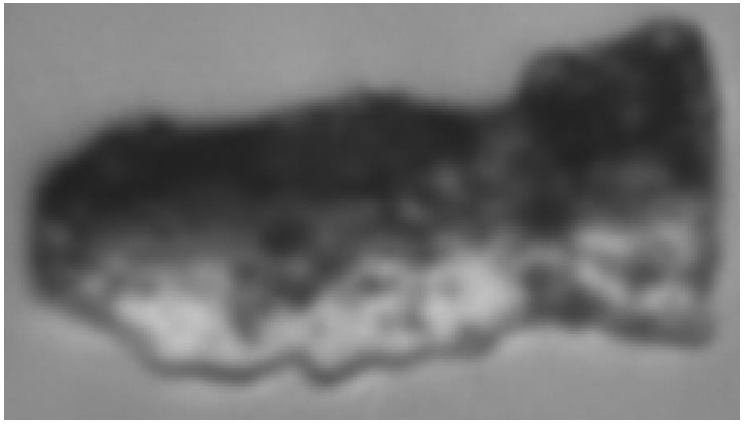
test_GRIP_31may2021/GRIP_raw/GRIP_3306_0_20_3/GRIP_3306_0_20_3_100001.png



test_GRIP_31may2021/GRIP_raw/GRIP_3136_40_55_3/GRIP_3136_40_55_3_912.png



test_GRIP_31may2021/GRIP_raw/GRIP_3046_0_20_1/GRIP_3046_0_20_1_1358.png



test_GRIP_31may2021/GRIP_raw/GRIP_3136_20_40_2/GRIP_3136_20_40_2_856.png

test_GRIP_31may2021/GRIP_raw/GRIP_3136_40_55_2/GRIP_3136_40_55_2_3767.png

Thank you for listening!

Any questions?

Appendix

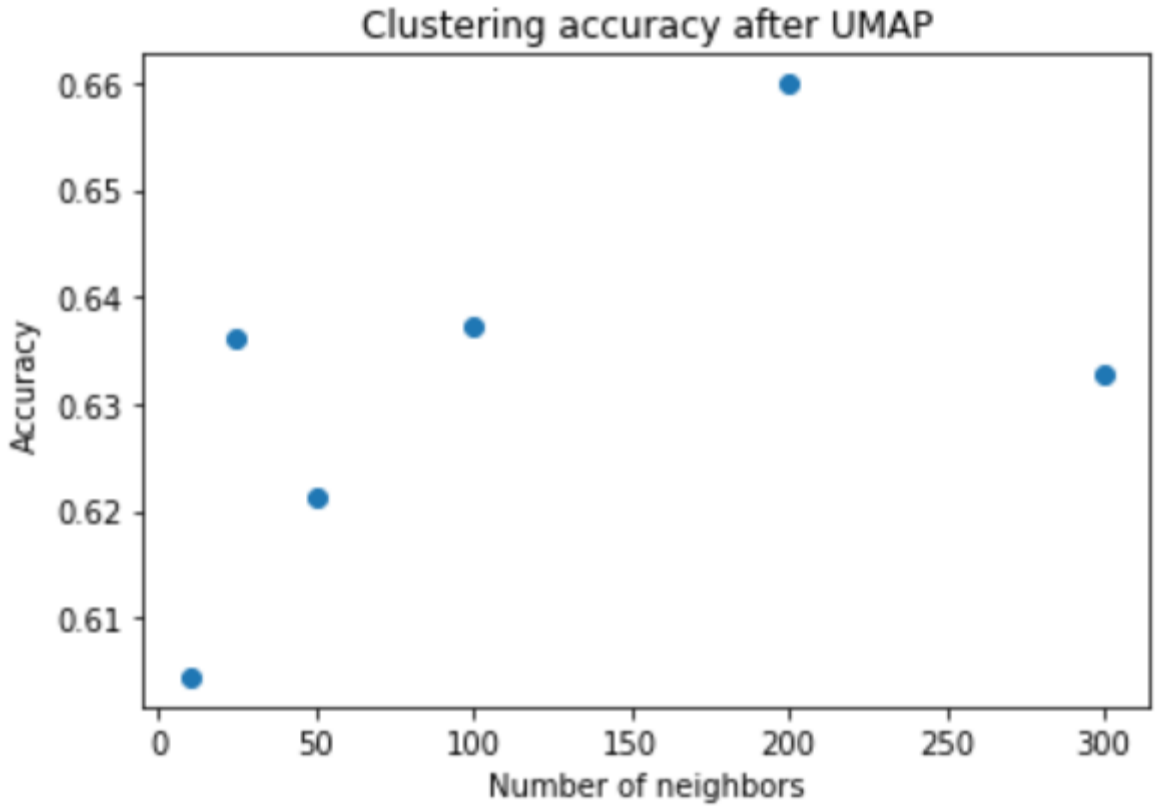
UNIVERSITY OF COPENHAGEN

Which meta data features were dropped/ used

```
dropped_features=['Particle ID', 'Calibration Image', 'Camera', 'Capture X',
'Capture Y', 'Date', 'ImageFile', 'Timestamp', 'imgpaths', 'camp', 'corylus',
'dust', 'grim', 'qrob', 'qsub', 'Time', 'Source Image', 'Image Y', 'Image X',
'Image Height', 'Image Width', 'Filter Score', 'Elapsed Time', 'Calibration
Factor']

used_features=['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']
```

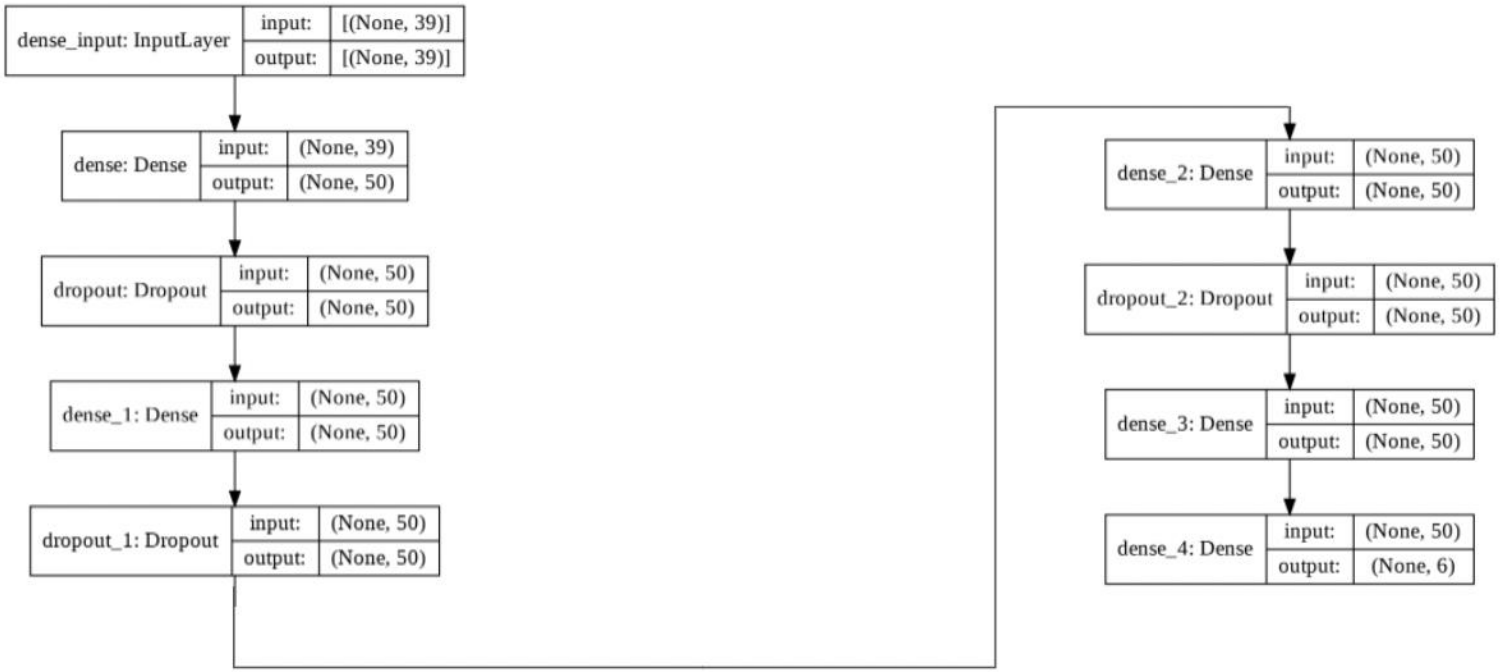
Clustering accuracy for different number of neighbors on training meta data



Neural network on the meta data

Model architecture

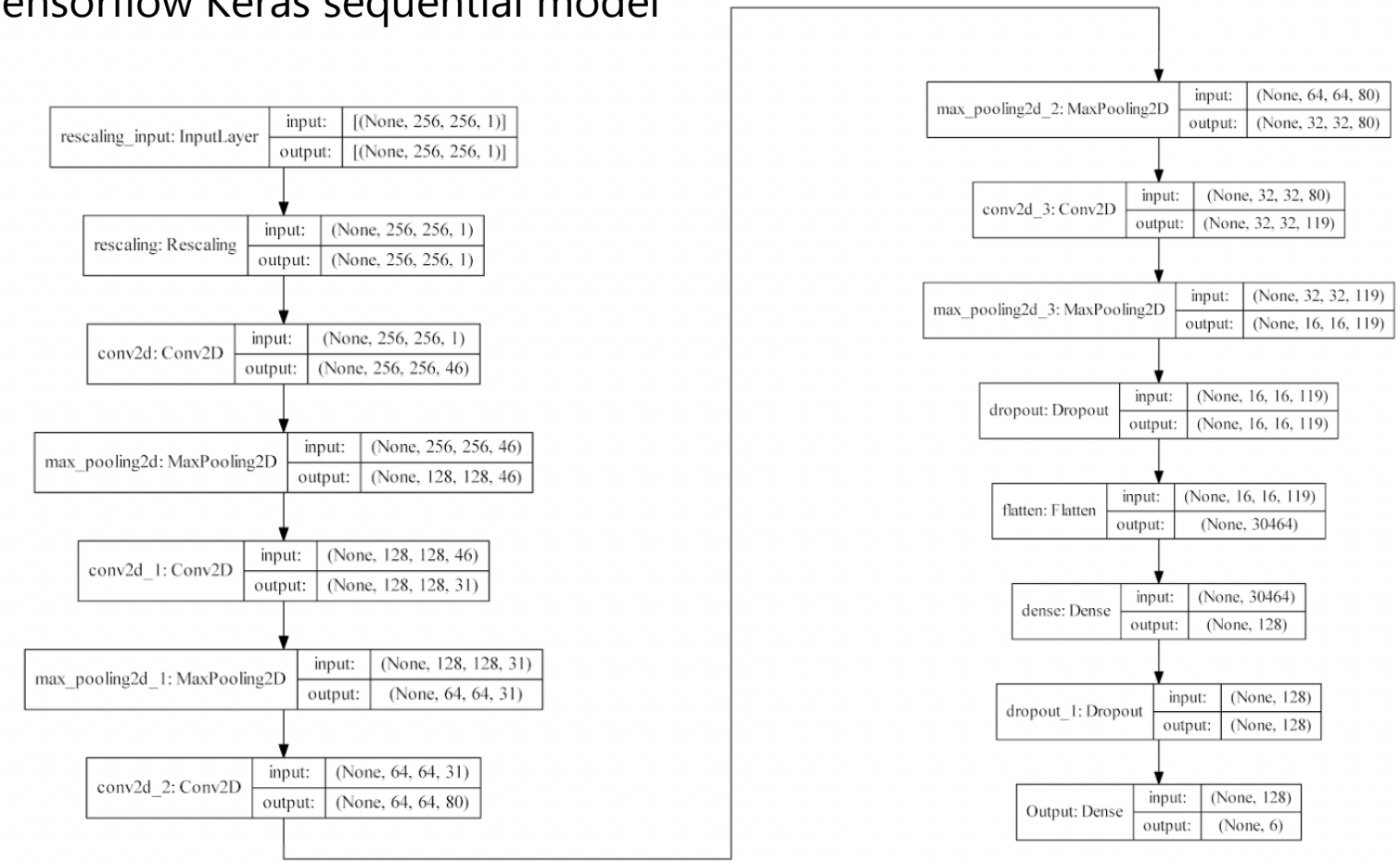
- Tensorflow Keras sequential model



Convolutional Neural Network on images

Model architecture

- Tensorflow Keras sequential model



Bayesian optimisation on CNN model

- Package used: Bayesian-optimiziation 1.2.0

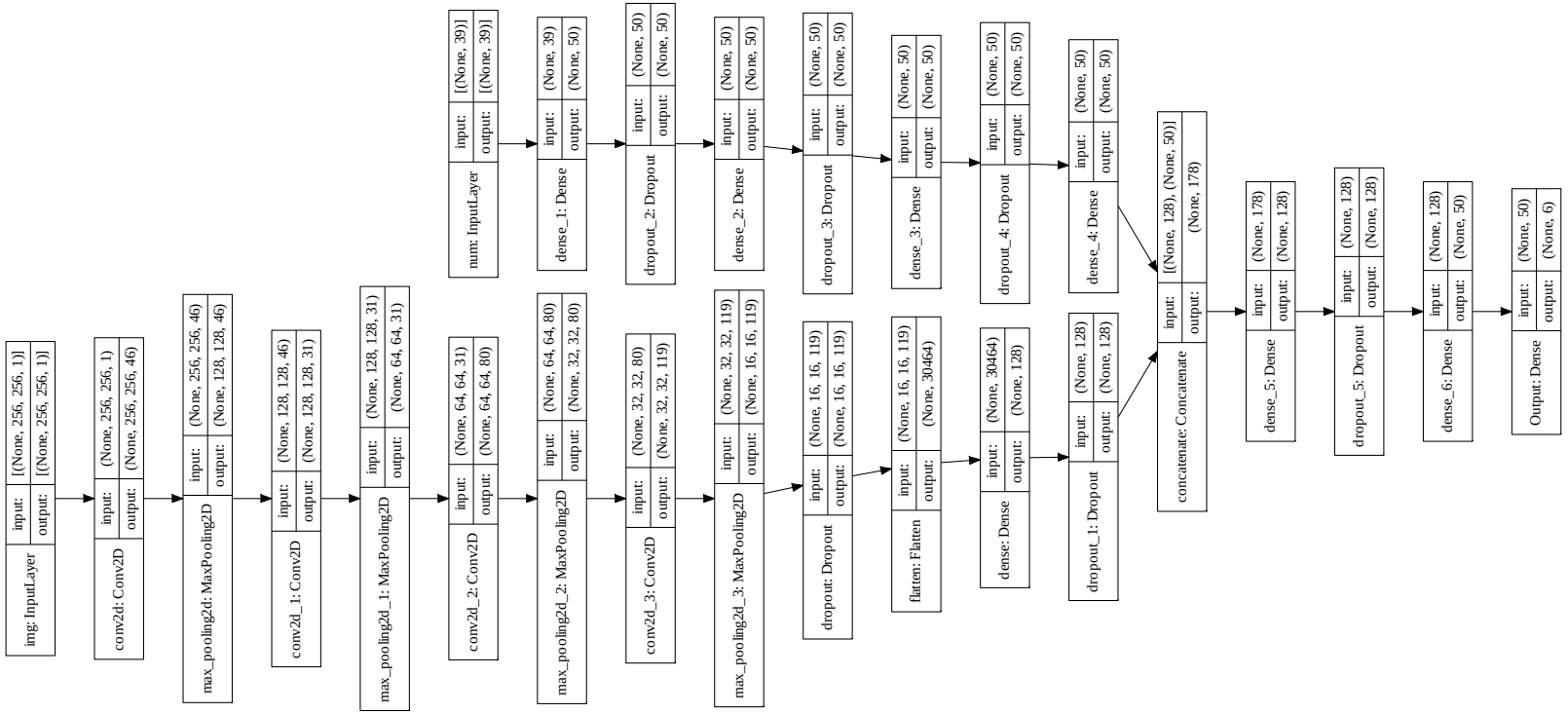
- Parameters tuned
  - Number of filters for each Conv2D layer (n_1, n_2, n_3, n_4)
  - Drop out rate for each Dropout layer (dropout1_rate, dropout_rate)
  - l2 kernel regularization factor for Conv2D layers (kernel_reg_rate)

- Parameter space
  - n_1: (16, 64), n_2: (16,64), n_3: (32, 128), n_4: (32, 128)
  - dropout1_rate: (0.1, 0.5), dropout2_rate: (0.1, 0.5)
  - kernel_reg_rate: (0.0001, 0.001)

Mixed model

Model architecture

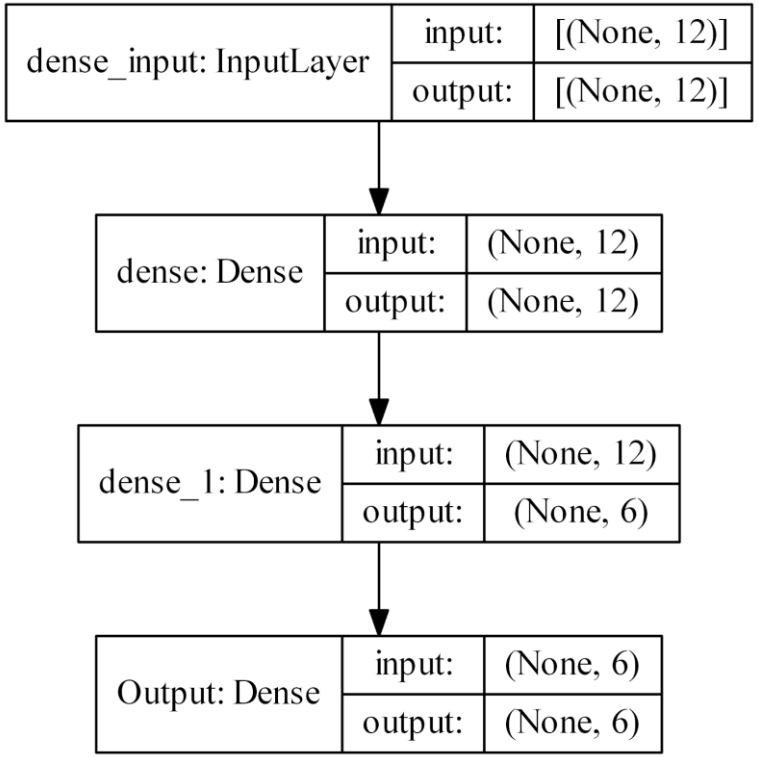
- Tensorflow Keras functional API model



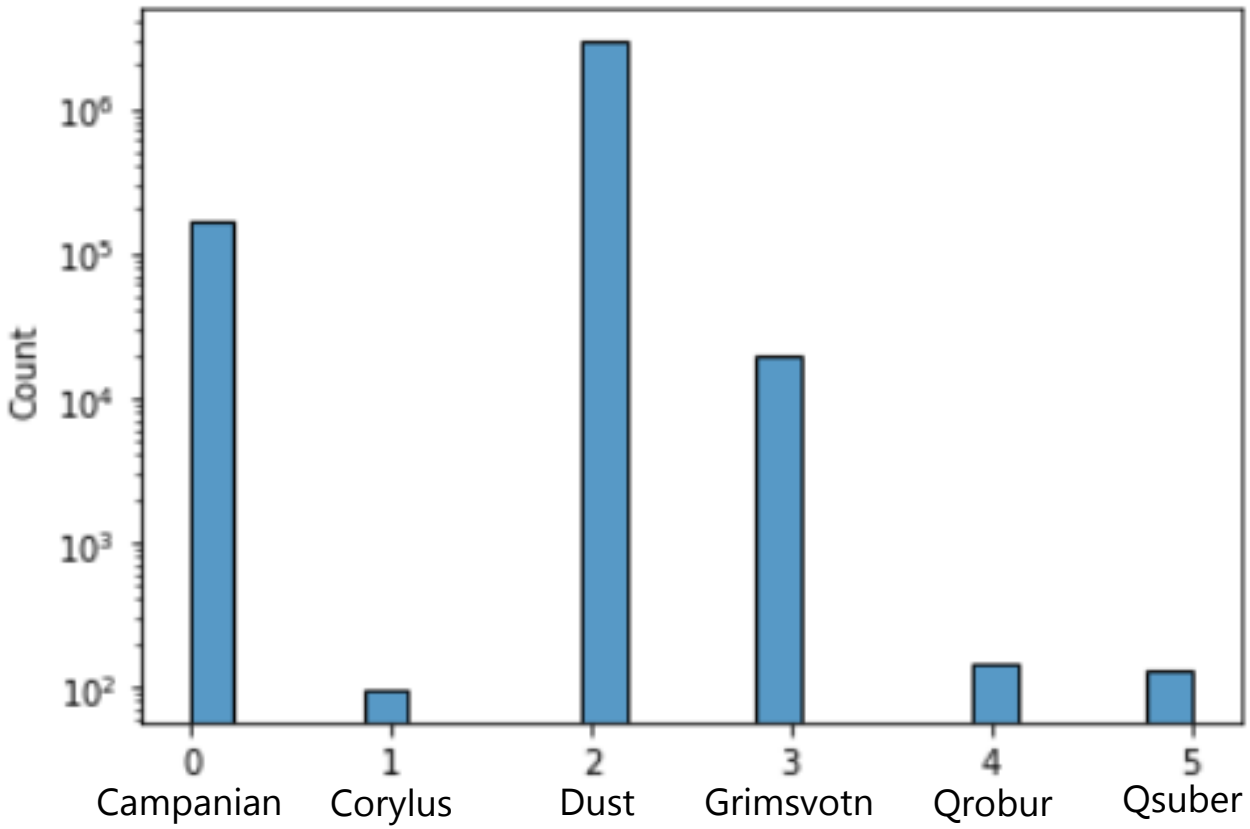
Ensemble model

Model architecture

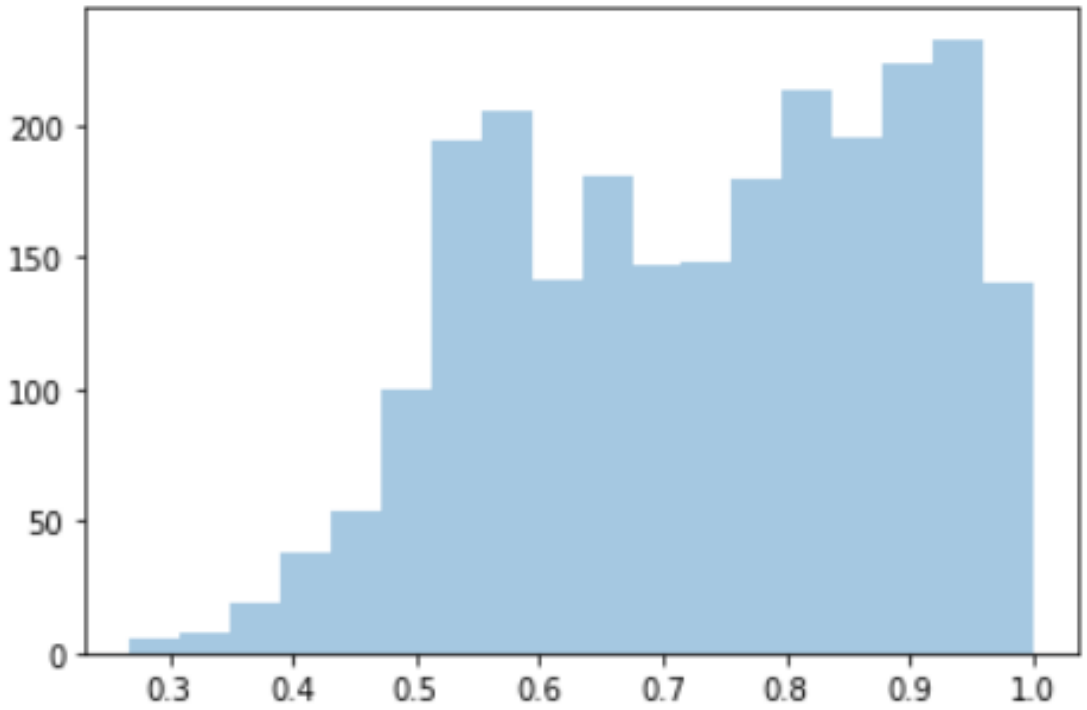
- Tensorflow Keras sequential model



Log histplot of predictions on test ice core meta data

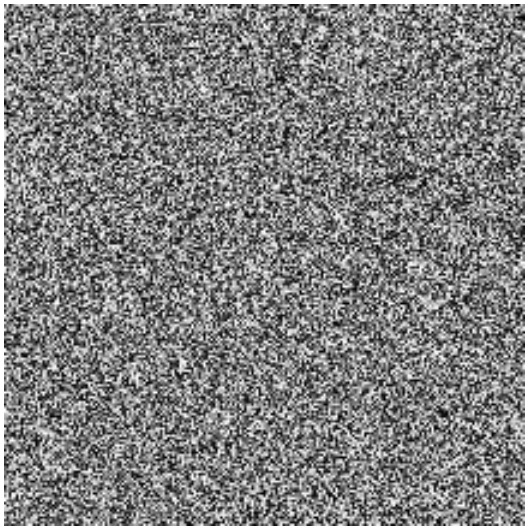
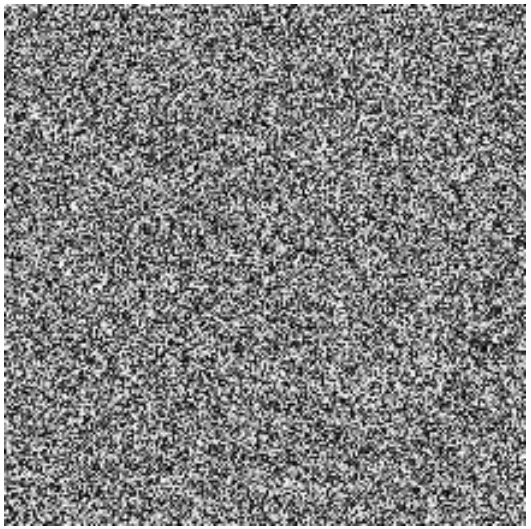
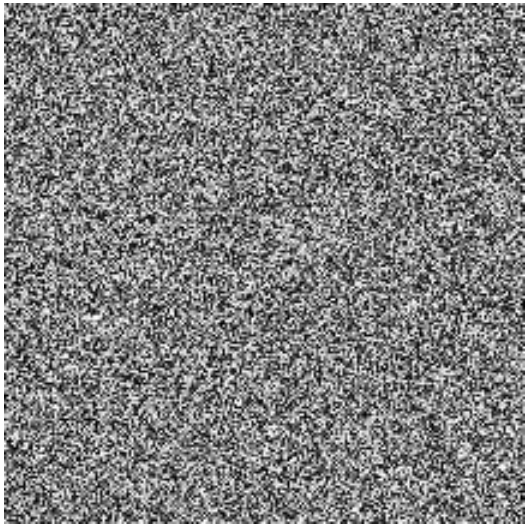


Confidence level on wrong predictions from meta data



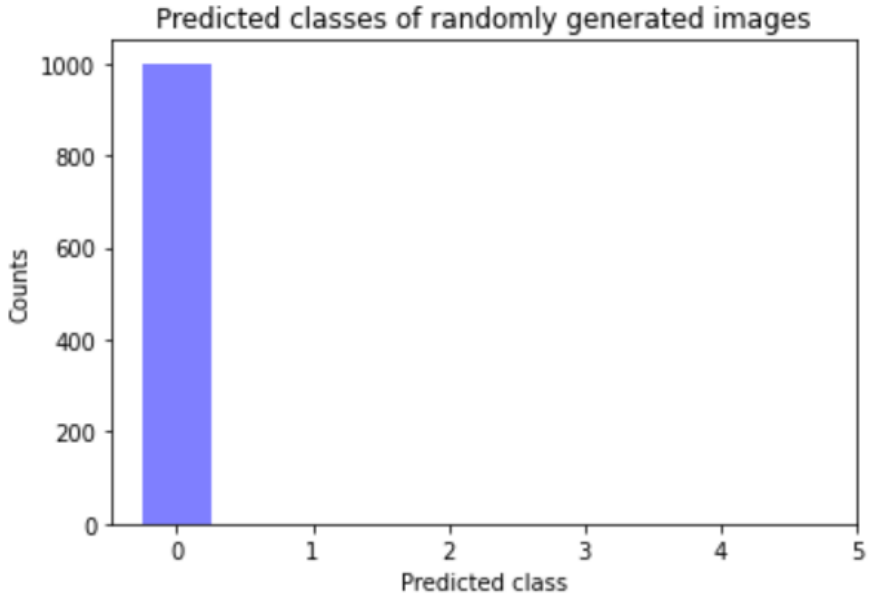
Predictions on randomly generated pictures

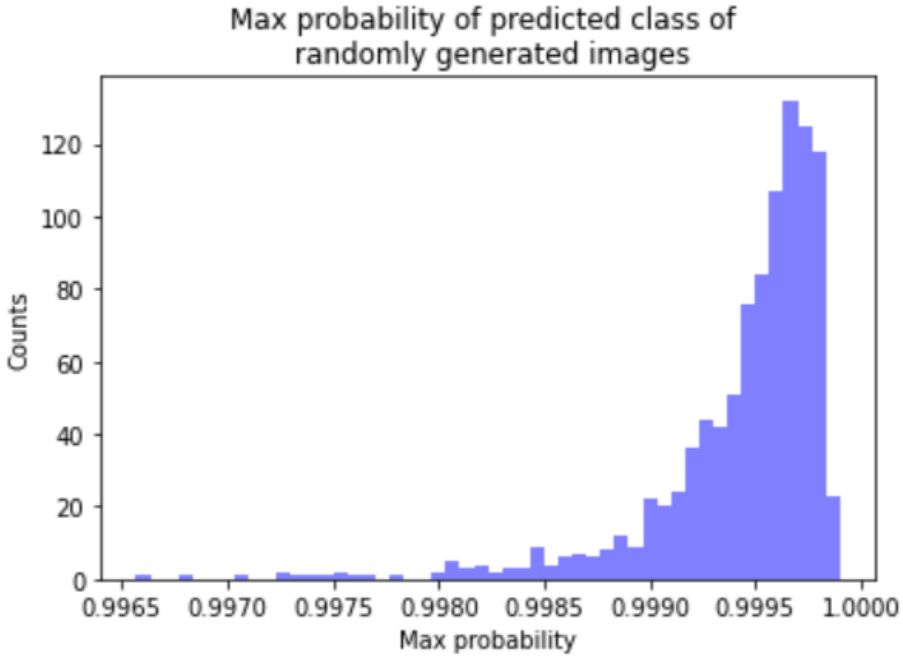
- 1000 randomly generated images run through CNN to predict classes.

Predictions on randomly generated pictures

- All randomly generated images are classified as class 0 i.e. Campanian with high confidence by the CNN.





Predictions on random real images

- 8 random real images run through CNN to predict classes.

Predicted class: Grimsvotn
Prediction confidence: 100%

Predicted class: Grimsvotn
Prediction confidence: 100%

Predicted class: Grimsvotn
Prediction confidence: 100%

Predicted class: Grimsvotn
Prediction confidence: 100%

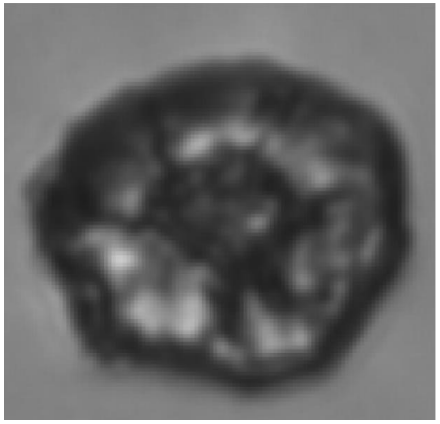
Predicted class: Campanian
Prediction confidence: 79%

Predicted class: Campanian
Prediction confidence: 99%

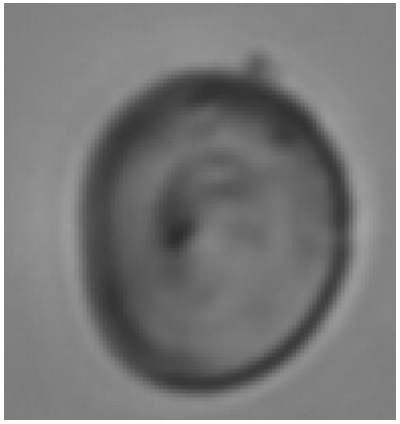
Predicted class: Grimsvotn
Prediction confidence: 100%

Predicted class: Grimsvotn
Prediction confidence: 84%

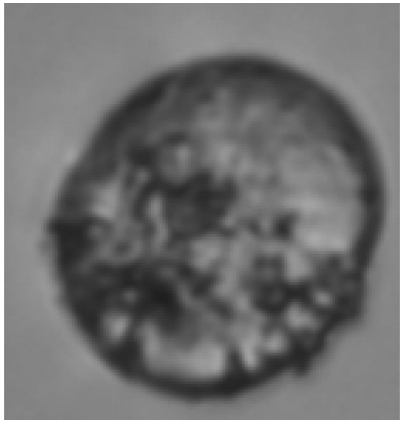
Predictions on 6 most interesting images



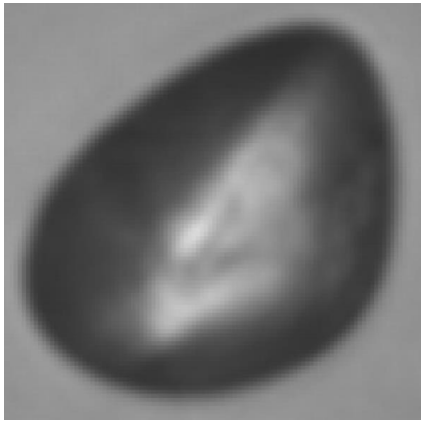
Predicted class: Corylus
Prediction confidence: 100%



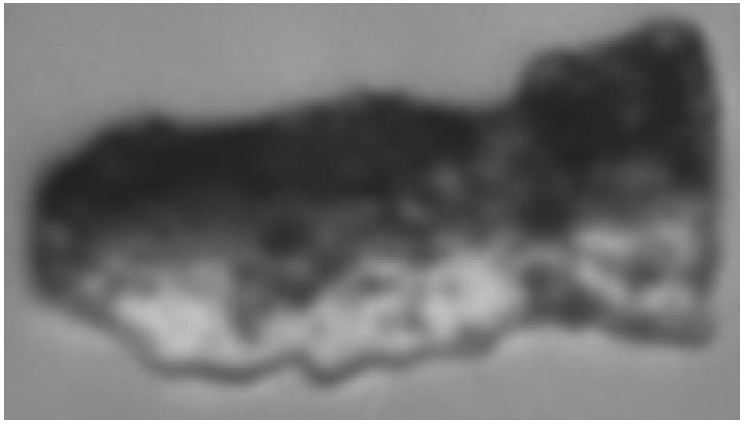
Predicted class: Qrobur
Prediction confidence: 99%



Predicted class: Qsuber
Prediction confidence: 62%



Predicted class: Corylus
Prediction confidence: 100%



Predicted class: Corylus
Prediction confidence: 45%

Predicted class: Qrobur
Prediction confidence: 81%

How much pollen do we expect the model to predict?

- Back of envelope-calculation

observed_confusion * ash_frac_in_real_data = expected_frac_pollen_prediction ⇒

(0.09+0.07+0.05)% *5% + (0.01+0.03+0.06)% *5% ~ 0.016%