Classification of microorganisms

Jing Lee

Metis Module 4

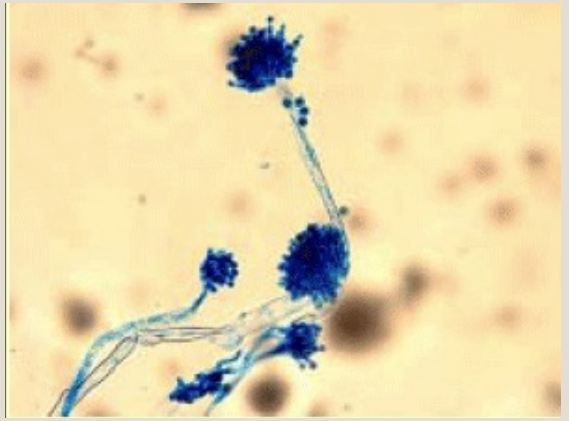
Machine Learning Classification

June 15, 2022

Introduction

- Motivation
  - Microscope image analysis: laborious and tedious
  - Prone to human error and bias
    - morphological features
- Objective
  - Classifying species of microscope images using machine learning models
  - Binary class identification
- Goal
  - Multi-class classification of each species s
  - Add-on to microscope software
  - Improve accuracy of image analysis

Fungi plate image

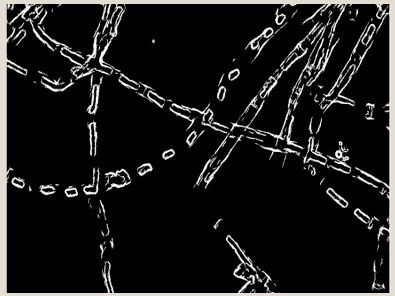


Fungi under microscope

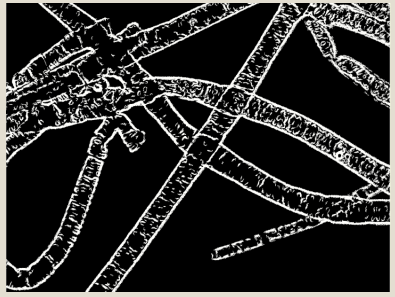
Image source: Dibe, K, et. al “Identification of aspergillus species using morphological characteristics”

Microscope images (10 types)

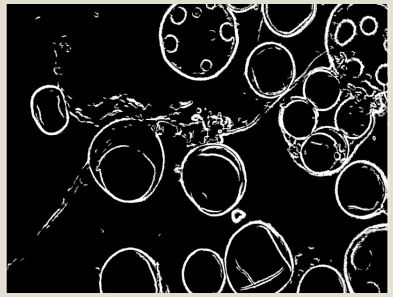
Algae



Ulothrix

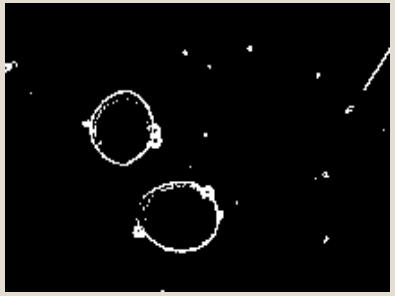


Spirogyra

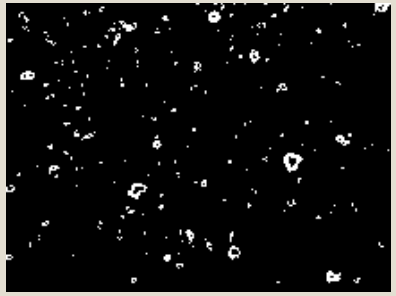


Volvox

Pithophora



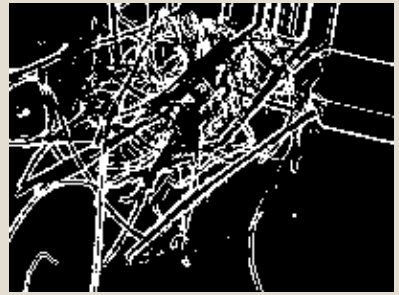
Protozoa



Diatom

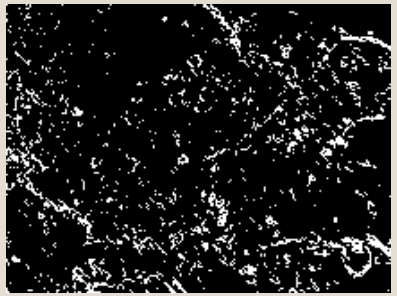
Fungi

Yeast



Raizopus

Penicillum

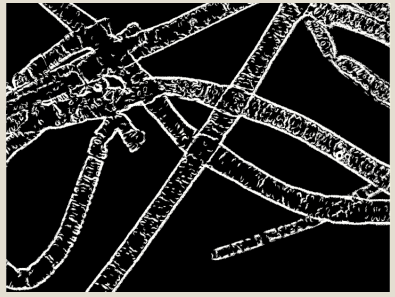


Aspergillus sp

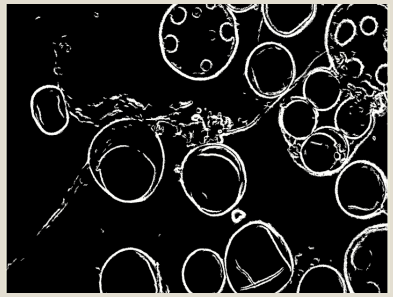
Target: Ulothrix vs Others

Algae

Ulothrix

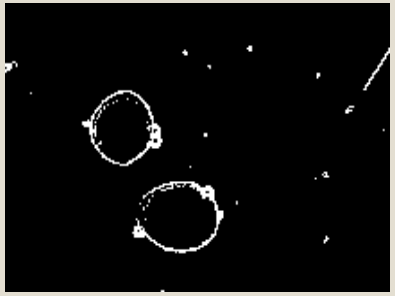


Spirogyra

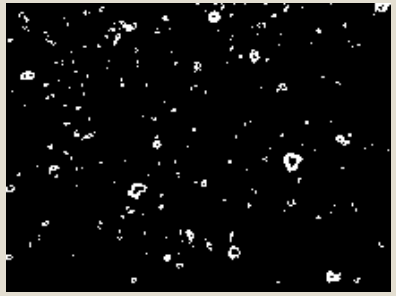


Volvox

Pithophora



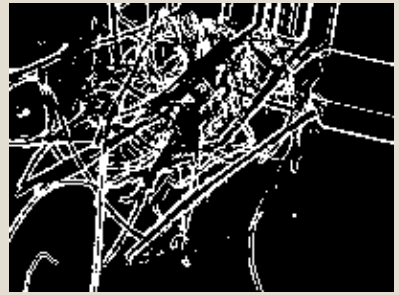
Protozoa



Diatom

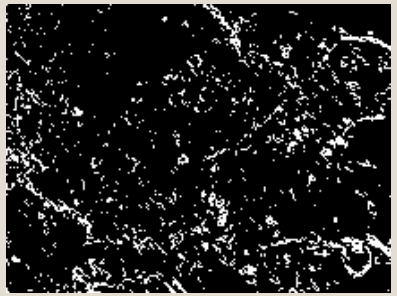
Fungi

Yeast



Raizopus

Penicillum

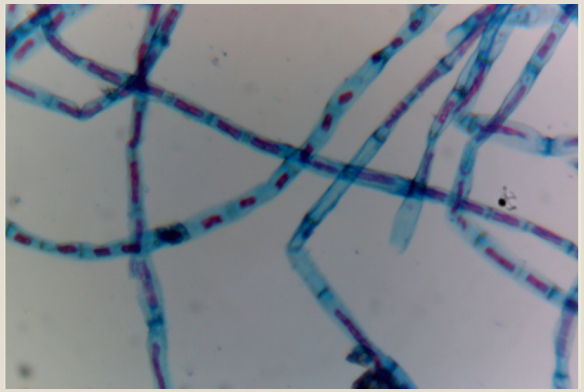


Aspergillus sp

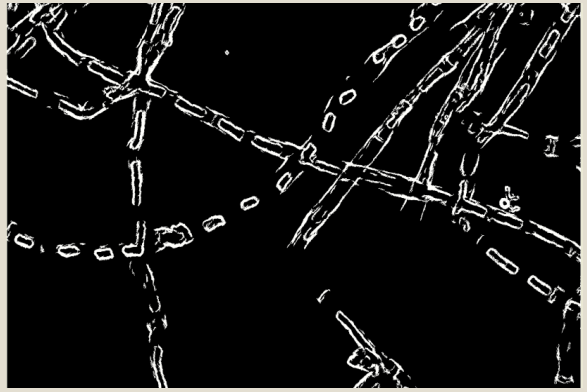
Source: Dhindsa, et. al. (2021), “Dataset for Efficient Microbes Classification System ”

Methodology

- Dataset - Kaggle
- Microscope images (30,527)
  - Dropped many duplicates
- Target
  - 1 = Ulothrix, 0 = Other
- Slightly imbalanced → oversampling
  - 30% (1s, Ulothrix)
  - 70% (0s, Other)



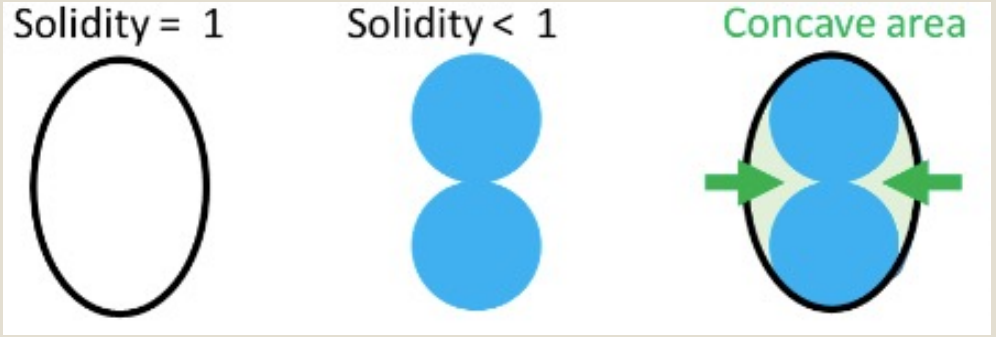
Microscope image: Ulothrix

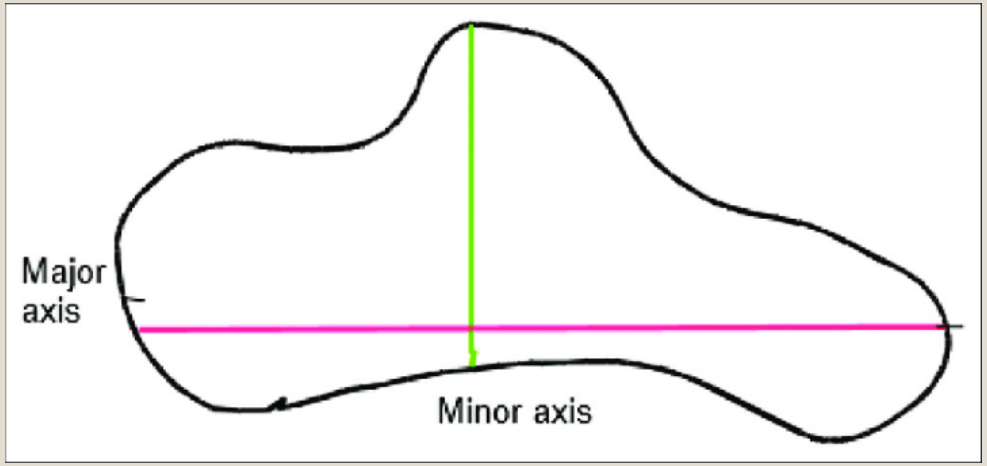


Processed with MATLAB for features extraction

Features

- Total: 24 features
  - Used: 16 features
- Some important features:
  1. Raddi
  2. Perimeter
  3. Solidity (area/convex area)
     1. shape of cell
  4. Eccentricity (major axis/minor axis)
  5. Extent (ration pixel area/ bounding box)





Model training

- 80% train, 20 % test
- Cross-validation
- Models trained:
  1. K-nearest neighbor (knn)
  2. Logistic regression
  3. Random Forests
  4. Gradient boosting
- Metrics: optimize f1 score

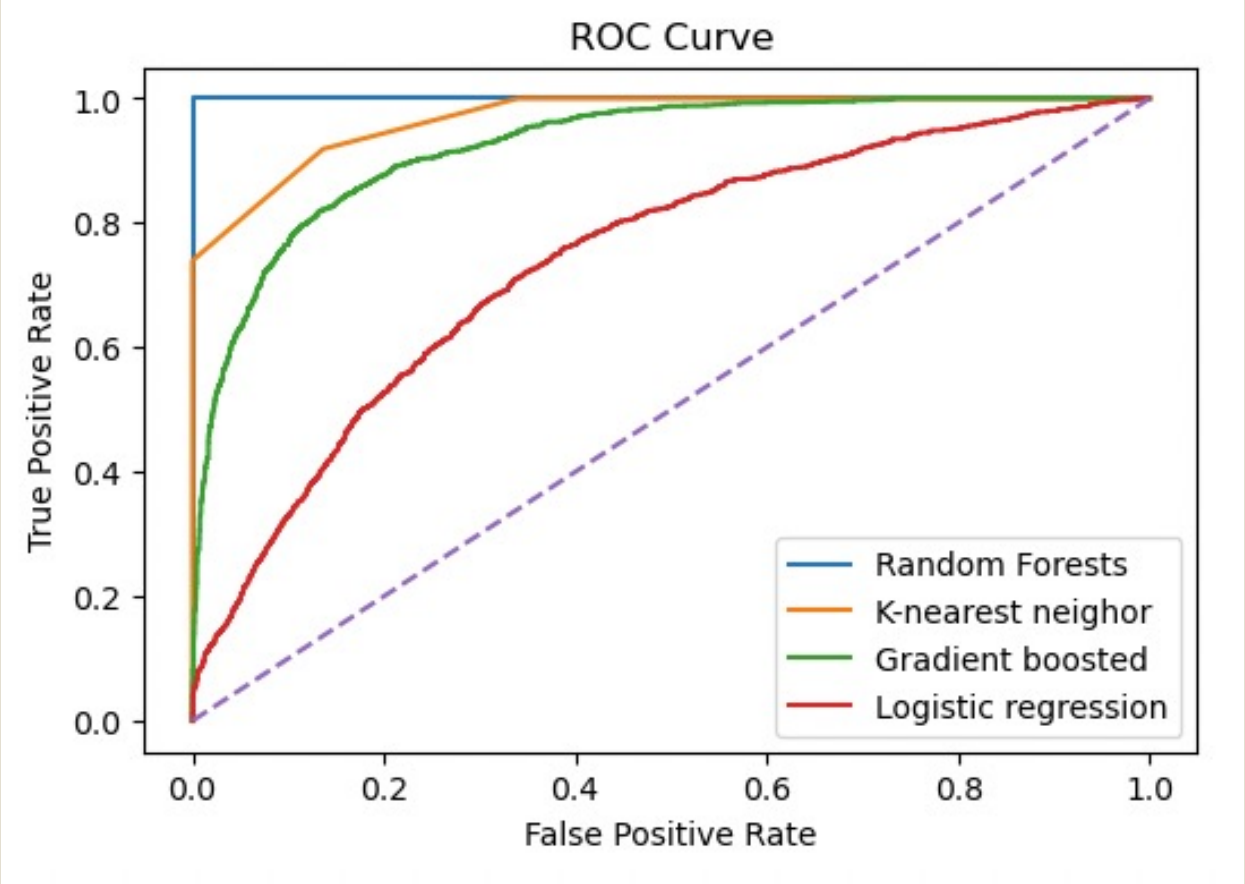
Model – training scores

Model	Accuracy	Precision	Recall	f1	auc
1. k-nearest neighbor	0.788	0.758	0.849	0.801	0.862
2. Logistic Regression	0.678	0.683	0.668	0.675	0.738
3. Random Forests	0.897	0.879	0.923	0.900	0.964
4. Gradient boosting	0.807	0.788	0.841	0.814	0.893

Best model: Random forests (best overall score)

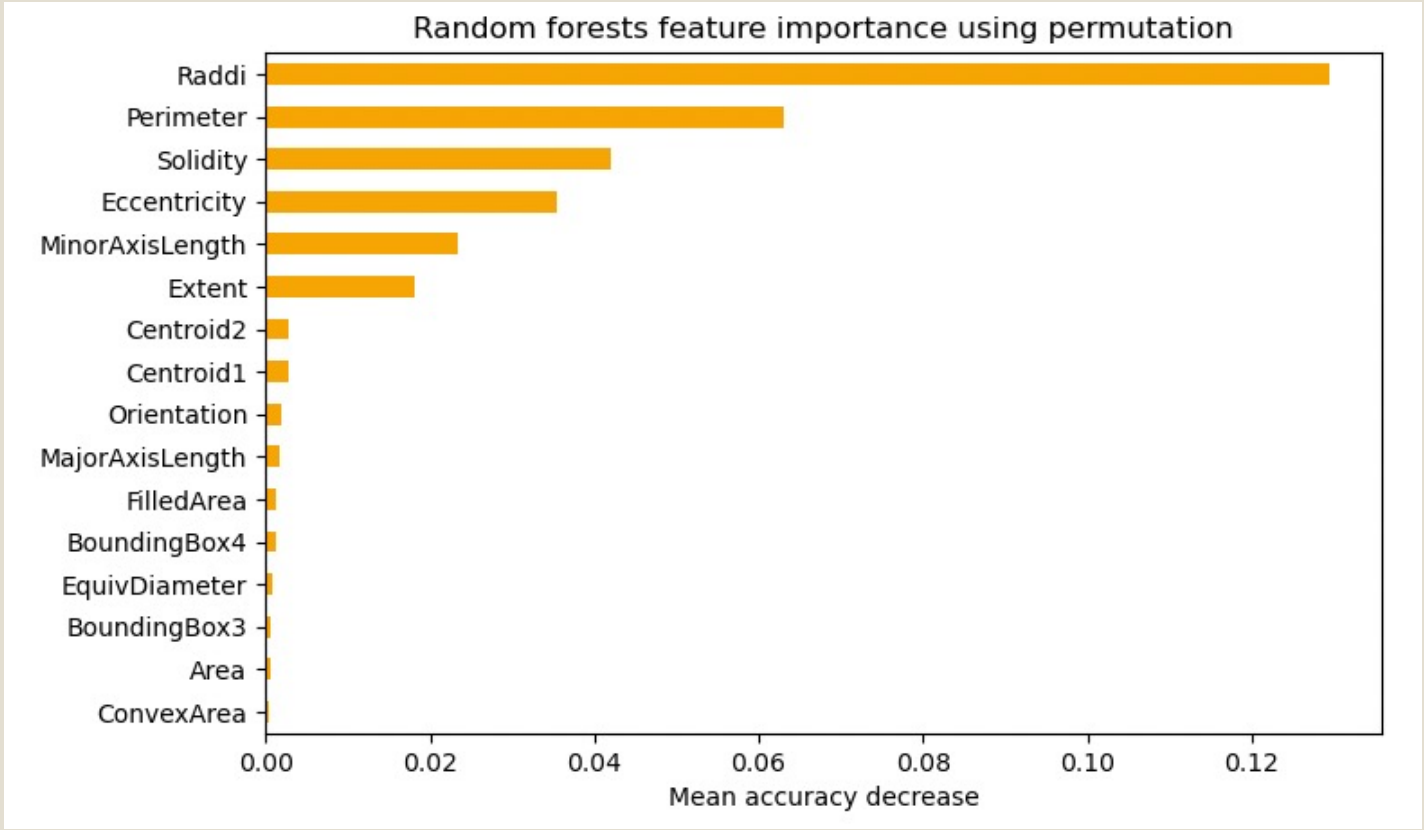
Model comparisons

- Best model (roc curve)
  - Random forests



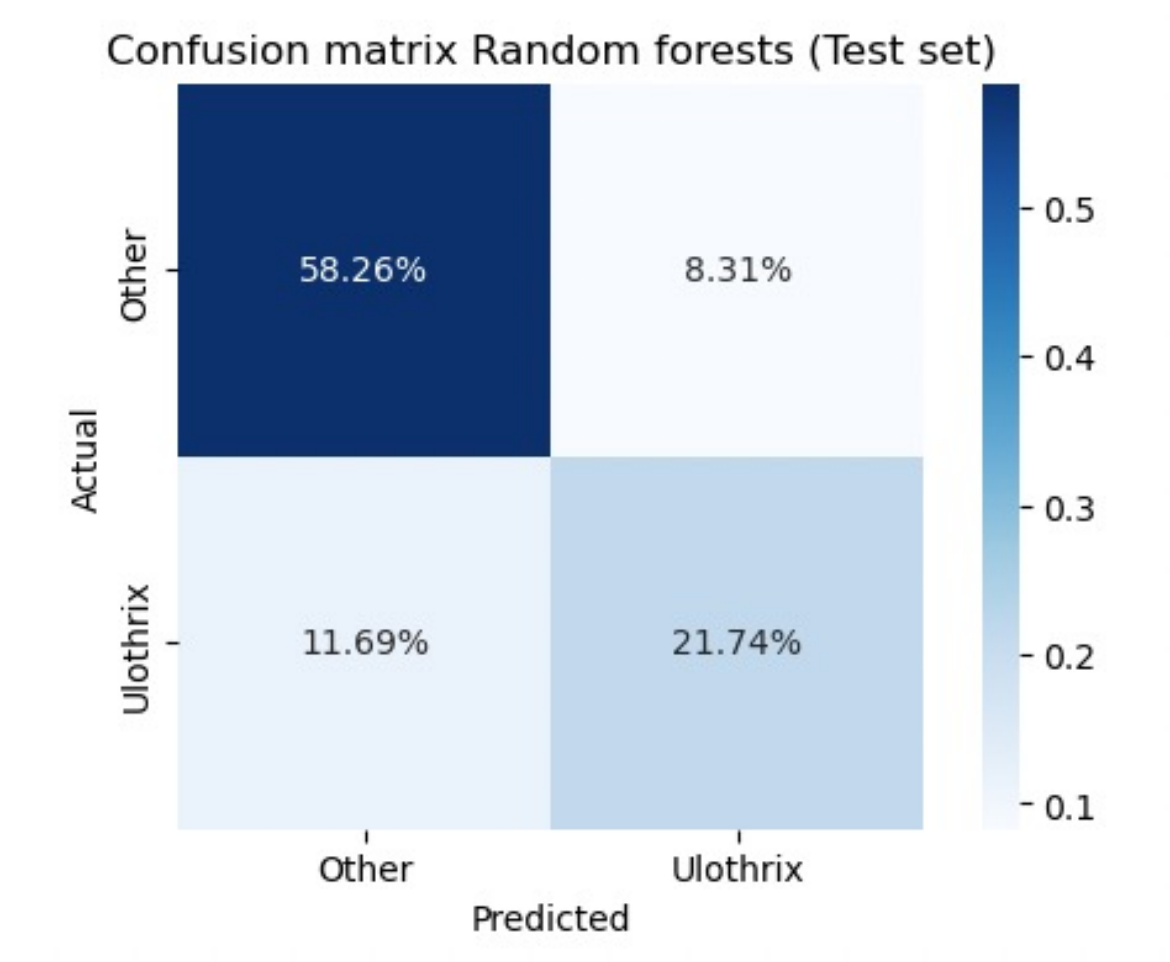
Features importance – Random forests

Top 5 important features
1. Raddi
2. Perimeter
3. Solidity
4. Eccentricity
5. MinorAxisLength



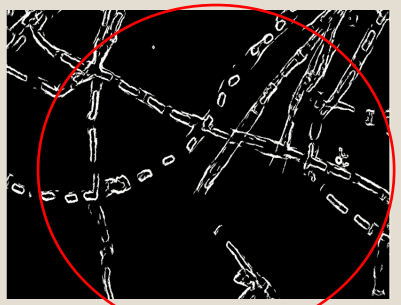
Random forests: test set

- Random forests (test scores)
  - accuracy 0.800
  - precision: 0.724
  - recall: 0.650
  - f1: 0.685
- Model is better at predicting Other class
- Model is overfitting!

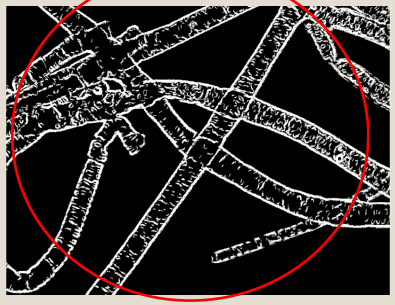


Microscope images (10 types)

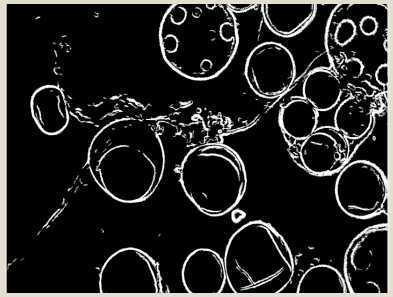
Algae



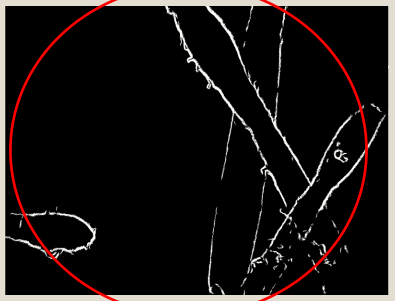
Ulothrix



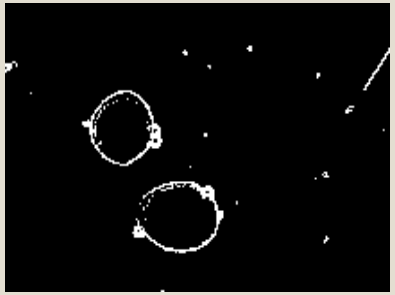
Spirogyra



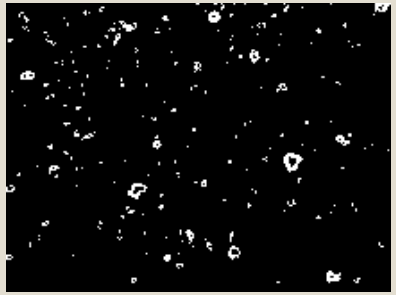
Volvox



Pithophora



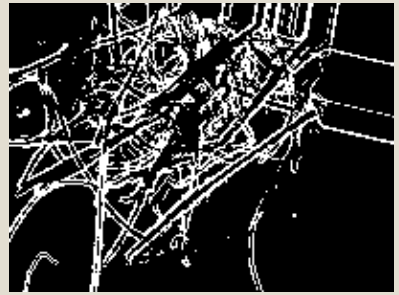
Protozoa



Diatom

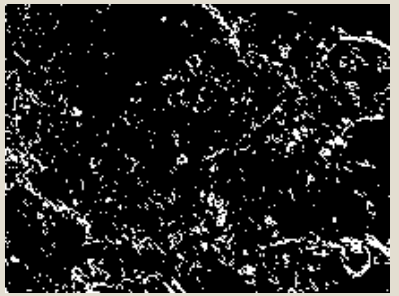
Fungi

Yeast



Raizopus

Penicillum



Aspergillus sp

Conclusion

- Best model selection: Random forests
- Model overfitting
- Test data set f1: 0.685
- Model is good at predicting other (58%) while Ulothrix (20%)
- Overall model correctly classified classes 78%

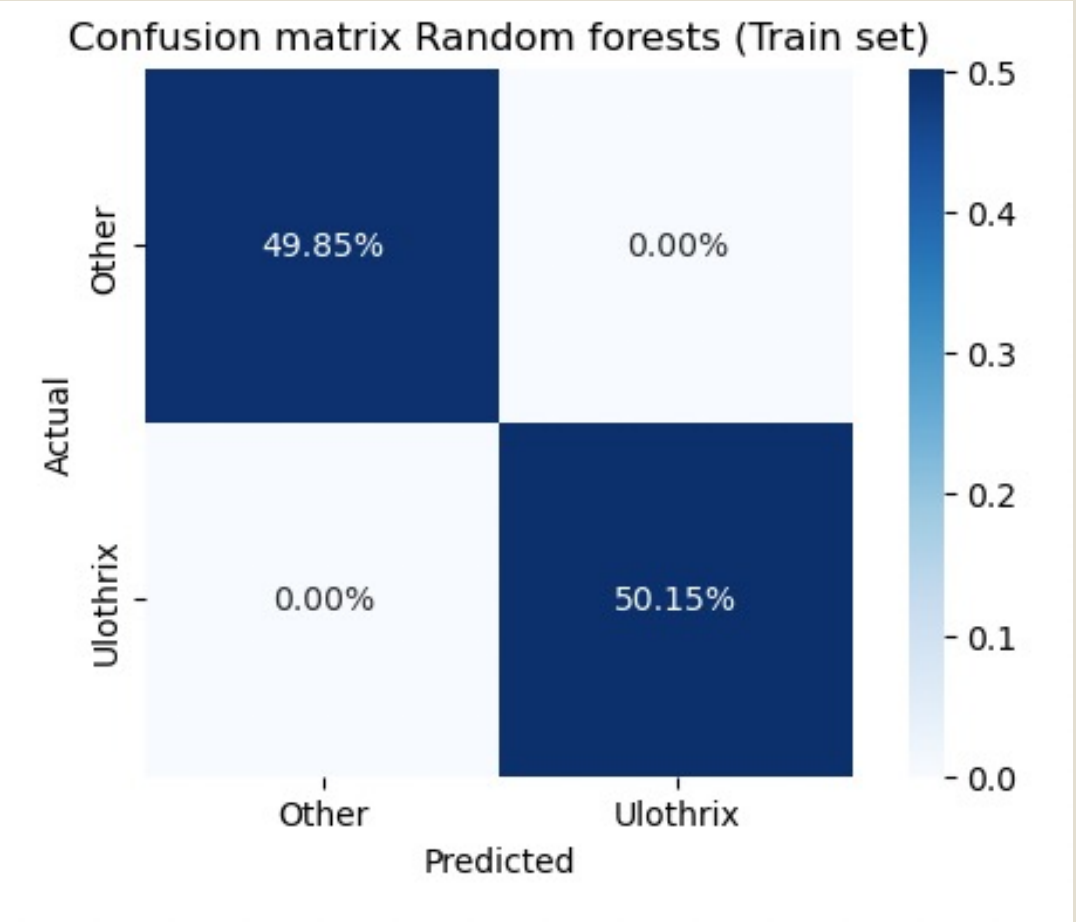
Future

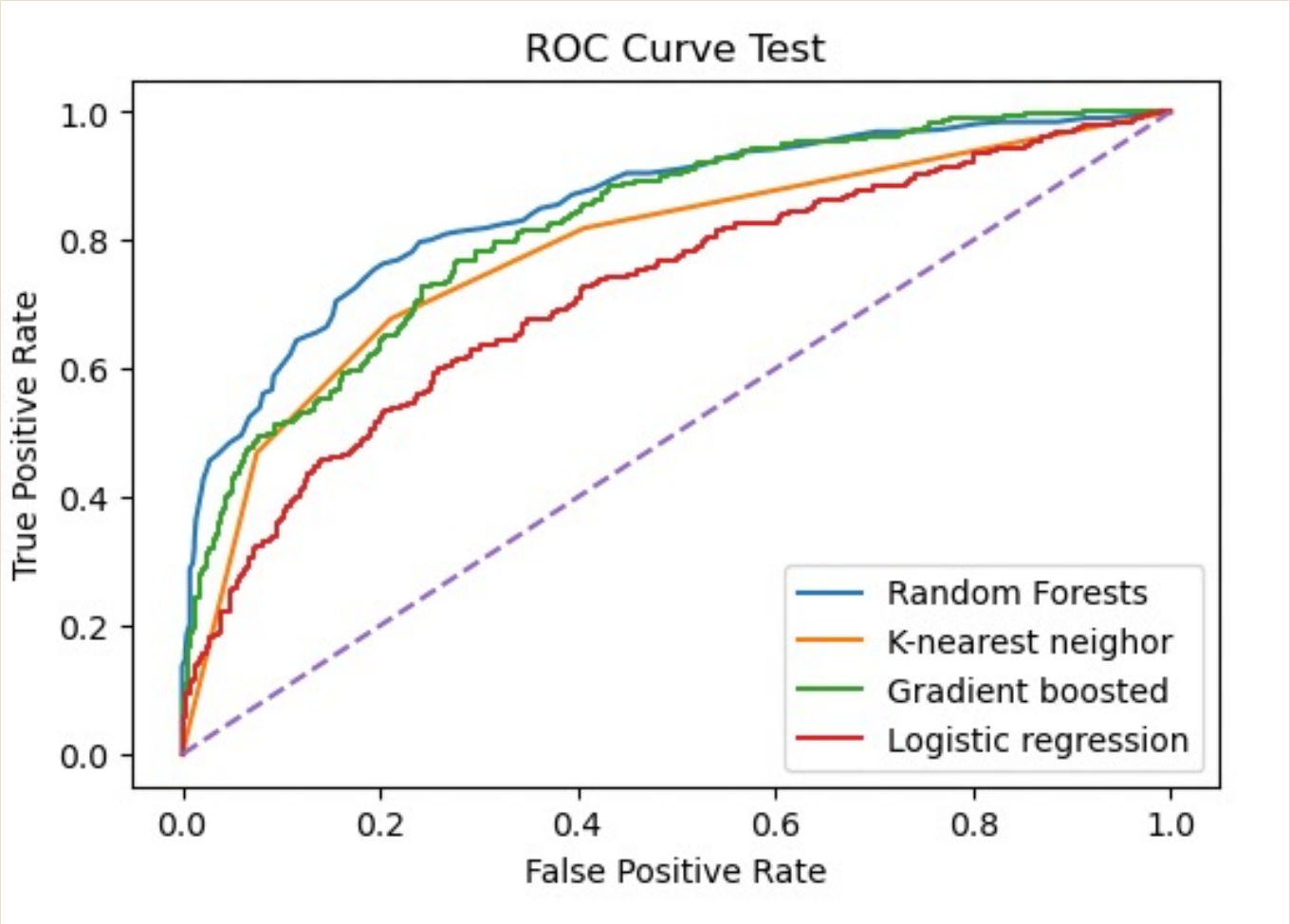
- Tune hyperparameters random forests to reduce overfitting
- Optimize F1 score
- Lower FP and FN
- Multi-class classification to identify each microorganism type

Thank you

Questions?

Appendix


Confusion matrix Random forests (Train set)
Other
Ulothrix
Actual
49.85%
0.00%
0.00%
50.15%
Other
Ulothrix
Predicted
0.5
0.4
0.3
0.2
0.1
0.0


ROC Curve Test
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forests
K-nearest neighor
Gradient boosted
Logistic regression