LU-TP 22-04
January 2022

Identification of spectral features differentiating fungal strains in infrared absorption spectroscopic images

Dejan Stancevic

Department of Astronomy and Theoretical Physics, Lund University

Bachelor thesis supervised by Carl Troein

LUND
UNIVERSITY

Abstract

There are many unknowns regarding the interaction between fungi and their surroundings. In this project, we took a closer look at hyperspectral images of several fungal strains on two different substrates. The project mainly consisted of developing a code for the classification of fungal strains and the extraction of information from it. The classifier used hyperspectral images in infrared of four strains of three species that were grown on two different substrates. A total of 192 images were used. Images were processed using software that was already created for the analysis of hyperspectral data. We developed a random forest classifier to classify the samples by fungal strains. The performance of different classifier parameters was determined and the best ones were chosen. Then, spectra and their derivatives were analyzed and their classification performances were compared. As the last step of the project, the developed random forest algorithm was used to identify the most important wavenumbers for discerning different fungal strains. One of the interesting results was an unexpectedly high increase in the accuracy of the classifier when the first derivative of spectra was used instead of plain spectra.

Popular science summary

Fungi are all around us and as such are used in many industries (e.g. farming and medicine). One example is fungi that play a crucial role in tree growth and development. Actually, a tree's root system and a fungus form a strong relationship that is an example of a symbiosis. The fungus supplies the tree with nutrients from the ground, while in return the tree provides carbohydrates obtained through a process of photosynthesis to the fungus. The prime example of this type of symbiosis is the "humongous fungus", a fungus that interconnects the whole Malheur National Forest in Oregon. Grasping the interconnectedness between trees and fungi could teach us more about what we can do to create healthier forests. Healthy forests are of foremost importance, especially now when we are facing an unprecedented number of wildfires and ever-increasing pollution of the air.

Currently, we are not sure exactly how fungi interact with the soil surrounding them. In order to find out, we need to observe a region around a cell wall with sufficiently high spatial resolution. For years it was hard to imagine having fungi on a substrate that is nice enough for imaging and at the same time not harmful for the organism. In recent years that became possible. Still, to get a high spatial resolution image at a certain wavelengths of light takes too much time which makes taking the whole spectrum at high resolution infeasible. This problem is circumvented by taking cruder images for many wavelengths, determining which ones are interesting for further investigation, and then using a higher resolution imaging technique.

One of the main goals of this project was to analyze spectroscopic data of several fungal strains. A machine learning algorithm was used to deduce the most important wavelengths for differentiating among the strains. We hope that those wavelengths will prove useful for future high-resolution spectroscopic analyses.

Contents

1 Introduction

There are still many questions about how fungi interact with their surroundings to which we do not have satisfactory answers. An important example is an interaction between fungi and the soil in which they live. In recent years, experimental procedures [1] have been developed to address this question. In the experiments, Fourier transform infrared spectroscopy was used to gather data (i.e. hyperspectral images). The OCTAVVS software package [2] was developed for processing those images. One of the main benefits of this software is its correction for the Mie scattering.

In this project, we have analyzed images obtained in the above-mentioned experiments. Random forest classifiers were used for that analysis. Firstly, the background of concepts that are important for understanding this thesis is given in the Background section. Then, in the Methods section, specific procedures that were done are explained followed by the Results and Discussion section. Finally, the thesis ends with a Conclusion and Outlook section in which we state the main results and propose some ideas for future research.

2 Background

2.1 Random forest classifier

Random forest classifiers are a class of ensemble machine learning algorithms that use an ensemble of decision trees to predict output based on given inputs and training data. Here we explain basics, while more information can be found in [3]. A decision tree is a set of rules with a structure that resembles a tree-like one. To be more specific, the decision tree contains nodes, branches, and leaves. A schematic diagram of the decision tree is shown in Figure 1.

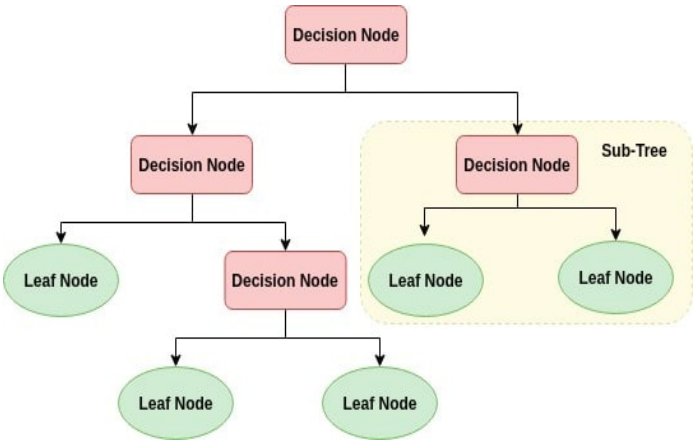

Figure 1: A schematic diagram of a decision tree. Figure taken from [4].

Connections between nodes, branches, and leaves are based on training data. At each node,

there is a "question" regarding a particular data feature, which, based on a sample, splits data in such a way to minimize a loss function. Usually, the Gini index is used as the loss function for classification trees. The Gini index, G, is calculated according to

$$G = \sum_{n=1}^{N} r_n(1 - r_n), \tag{1}$$

where r_n is ratio of samples belonging to the nth out of N classes compared to the number of all the samples at the branched node after the "question". Branches, based on an answer to the node's "question", divide data into new nodes. The first node, also called a root node, has to go through all the training data, while the following nodes have to deal only with the data that has reached them via branches. Once a stopping criterion is met, the node stops branching and becomes a leaf. The leaf contains information about which class a specific sample belongs to. The criterion for stopping branching could be that the accuracy of the node is high enough, meaning that the samples that belong to a single class comprise a high enough percentage of all samples at that node. Here, high enough means that there is no new information gained by branching.

However, a single decision tree is prone to over-fitting. In order to resolve the over-fitting problem of decision trees, a method of bootstrap aggregating, or bagging for short, is introduced. Bagging decreases the dependence of the algorithm on the training data by generating several new data sets from the old one and training many decision trees based on those data sets. It creates new data sets by choosing with replacement samples at random from the original training data. New data sets do not have to be the same size as the original data set (but they usually are). This implies that the new data sets usually contain several copies of the same sample. After creating new data sets, decision trees are trained based on each one of the newly created data sets. Also, at each node instead of using all possible features to determine the "question", only several randomly selected features are used. Commonly, the number of randomly selected features is the square root of the total number of features. Another useful parameter that can be changed is the number of decision trees used. The suggestion is to increase the number until the accuracy of predictions stops changing. The final output of the random forest classifier is the class that has been predicted by the most decision trees. This process is known as "voting". One fact worth mentioning is that there is a similar algorithm that instead of classifying the data predicts continuous variables by finding the average from all decision trees (e.g. price of a car based on its manufacturer and age).

Once classification is done, feature importance can be calculated. The most common method, which was also used in this project, is the mean decrease of impurity. It is obtained by calculating the total amount by which the Gini index is decreased from one node to the following ones, finding an average value for all decision trees, and normalizing it to 1. Hence, every feature will have importance equal to some number between 0 and 1, with higher values implying higher importance.

The accuracy of a classifier can be determined using k-fold cross-validation, which is performed by dividing data into k equally sized sets and training k random forest classifiers on data

excluding one of the k sets. Then, the accuracy is calculated as the average accuracy of those k random forests. Cross-validation can be repeated several times to get better precision in the estimation of the accuracy.

2.2 Fourier transform infrared spectroscopy

Fourier transform infrared spectroscopy (FTIR) is a nondestructive spectroscopic method that utilizes the mathematical technique of Fourier transforms to extract information from acquired spectra. Here we discuss basic ideas behind FTIR, while more information can be found in [5]. The technique was developed during the 1960s as an opponent method to dispersive infrared spectroscopy. Some of the main advantages of FTIR are signal-to-noise ratio, the time it takes to complete a measurement, and precision. As technology progressed, FTIR replaced dispersive spectroscopy by the 1980s. FTIR is well suited for studying biological specimens because the most of molecules found in them are infrared active (due to vibrational modes). However, FTIR is used to differentiate between different chemical groups (saying anything more specific is difficult).

FTIR consists of a light source, a Michelson interferometer, a sample, and an infrared spectrometer. The light source emits a black body radiation with a peak in infrared part of the spectrum. Next, that light goes through the Michelson interferometer where only a certain set of wavenumbers (inverse of wavelengths) interfere constructively and is shined on a sample. Changing the length of one arm of the interferometer, a set of constructively interfering wavenumbers is changed and results are recorded. As a result, a plot of the intensity versus the arm length is acquired and it is called an interferogram. After employing the Fourier transform, the interferogram is converted to a graph of the intensity against the wavenumber.

The limiting factor of FTIR is its spatial resolution. Since infrared light is used, it is not possible to obtain a better resolution than a few microns. For better resolution, optical photothermal infrared spectroscopy (OPTIR) can be used.

2.3 Optical photothermal infrared spectroscopy

OPTIR is a nondestructive, pump-probe spectroscopic method. It is a technique that has emerged in recent years as a complement to FTIR. Already, it has found many applications in various industries (e.g. high-tech and biology).

OPTIR uses infrared light as a pump and visible light as a probe. Visible light is focused on a specific part of a sample and shines continuously, while infrared laser shines infrared light with different wavenumbers onto the same region of the sample. Because of a photothermal effect, that part of the sample expands at a specific infrared wavenumbers. Visible light reflects differently as a result of photothermal expansion and an infrared spectrum for that part of the sample is acquired. More information about OPTIR can be found on [6].

Since OPTIR uses visible light to probe a sample it has much better resolution than FTIR. Also, OPTIR requires almost no sample preparation, which is not the case for FTIR. One of

the main drawbacks of OPTIR is that it probes only a small part of the sample at the time, causing the whole procedure to take more time than FTIR. Hence, OPTIR and FTIR are best used together.

3 Methods

Hyperpectral images used in this project were generated by Michiel Op De Beeck from Lund University, Department of Biology. In the experiment, four strains of three species were grown on two different substrates. The four strains were *Paxillus involutus* (PAI), *Hydnomerulius pinastri* (HYP), *Neurospora crassa* (NC4200), and *Neurospora crassa* with a loss-of-function mutation that affects cell wall composition and makes cell wall proteins diffuse away (NC16862). The two substrates were casein (protein) and mix of lignin (organic polymer found in cell walls in wood) and casein. Each set-up was reproduced three times, giving a total of twenty-four fungal colonies. From each colony eight different hyphae were recorded using FTIR spectroscopy. Therefore, a total of 192 images were recorded. However, only the part from 900cm^{-1} to 1800cm^{-1} of the spectra was used since it seemed that all of organic compounds had clear peaks in that range (other parts were either full of noise or without peaks). OCTAVVS was used to process images and to divide pixels to several categories according to their spectra. We used those categories to group pixels as background (substrate) and foreground (fungus). Average spectra for each group of pixels of each obtained image was calculated and was exported in csv files.

New files were created that contained labels for all samples and their spectroscopic data (i.e. in total two files, one containing foreground data and the other one containing background data). Furthermore, difference spectra were calculated by subtracting the background from the foreground data. The first derivatives of the difference spectra, in further text referred as the first derivative spectra, were calculated as

$$I'[\lambda_i] = \frac{I[\lambda_{i+1}] - I[\lambda_i]}{\lambda_{i+1} - \lambda_i}, \tag{2}$$

where $I[\lambda_i]$ and $I'[\lambda_i]$ represent an absorbance and its derivative, respectively, of difference spectrum of a sample at λ_i wavenumber (i denotes wavenumber's position in the data set). Similarly, the second derivatives of the difference spectra, the second derivative spectra in further text, were obtained from the first derivatives.

The code was written using Python and its libraries SKlearn [7], NumPy [8], Pandas [9], seaborn [10], and Matplotlib [11]. The accuracy for every classifier used was calculated by repeating 10-fold cross-validation 10 times. We will refer to it as a validation accuracy.

In order to see whether some basic expectations regarding the data were true, classifications of substrate were done. Classification of the substrate based on the background data was performed since it was expected to see the validation accuracy of the classifier close to 100%. The expectation was justified because the chemical composition of the two substrates was vastly different (one substrate contains large quantities of lignin which was absent in the

other substrate). Also, validation accuracies of substrate classifications from the foreground and difference spectra were determined.

Next, validation accuracies of classifications of fungal strains were determined based on the difference, the first derivative, and the second derivative spectra, separately. The best parameters for each classifier were obtained as well. Furthermore, confusion matrices were acquired in order to get a better grasp of what species were being misclassified.

Two different normalization procedures on the difference spectra were used to increase the validation accuracy of the classifier. One normalization method divided all of the data points of every spectrum by the maximum value of that spectrum. The other normalization method used as a divisor the mean value of the spectrum instead of its maximum. The same normalizations were applied to the first derivative and the second derivative spectra. Classification of species using difference, first derivative, and second derivative data for each substrate separately was performed as well. It was done to check whether there were some unknown problems with one of the two substrates. Since the first derivative data that was divided by the maximum value showed the best results, it was used throughout the latter part of the project.

A list of misclassified images was obtained by finding images that were misclassified the most times in five runs of the classifier. The number of runs is low since it was obvious that only a certain set of images gets misclassified. In order to get a better feeling for misclassified samples, the first derivative data of correctly and incorrectly classified samples for each of the species were plotted.

In the last part of the project, the most important wavenumbers were determined by three different methods. The first and the second methods start with a set of all wavenumbers and try to distill the most important ones. In contrast, the third method starts from an empty set and adds more and more wavenumbers.

The first method consists of starting with all of the features and at each iteration only features whose importance is higher than $\frac{1}{n}$ (where n is a number of features used in that iteration), survives to the next one. Once number of features is below a wanted number, in this project that number was 2, iterations would stop. The second method starts the same as the first method. However, instead of choosing wavenumbers with importance higher than $\frac{1}{n}$, it chooses the top 90% of them, according to the feature importance. From that set the wavenumber with the lowest feature importance is deleted in order for this method to be useful even when number of wavenumbers is less than 10. Again, it stops once the number of features is less than a wanted number (2 for this project). The third method begins with an empty set of features and adds more features as it runs. More precisely, it adds only a single new feature in every iteration while preserving the old ones. Every possible feature is added to a set of wavenumbers from the previous step and the set with newly added feature that has the highest validation accuracy is saved for next step. This process is continued until a wanted number of features is obtained (in this project that number was 20).

4 Results and Discussion

In this section we present our results in chronological order as they were acquired and described in the Methods section. Alongside stating the results, we give our thoughts on their meaning and importance.

In Table 1 and Figure 2 are given validation accuracies of classifications of substrates based on the background, foreground, and difference spectra. As expected, classification of substrates from the background and the foreground data gave a perfect score. Taking the difference between them was supposed to erase some of the information about substrates and that is exactly what the result tells us. However, we did not expect to see the validation accuracy much worse than for the background and the foreground since the concentration of the substrate is not the same for the foreground and the background (because a fungus alters it). Hence, information about substrates is still there but the signal to noise ratio decreases.

Table 1: The mean and the standard deviation of validation accuracies of classifications of substrates based on the background, foreground, and difference spectra.

Accuracy / Type of spectra	Mean	Standard deviation
Background	100.0 %	0.0 %
Foreground	100.0 %	0.0 %
Difference	97.8 %	3.1 %

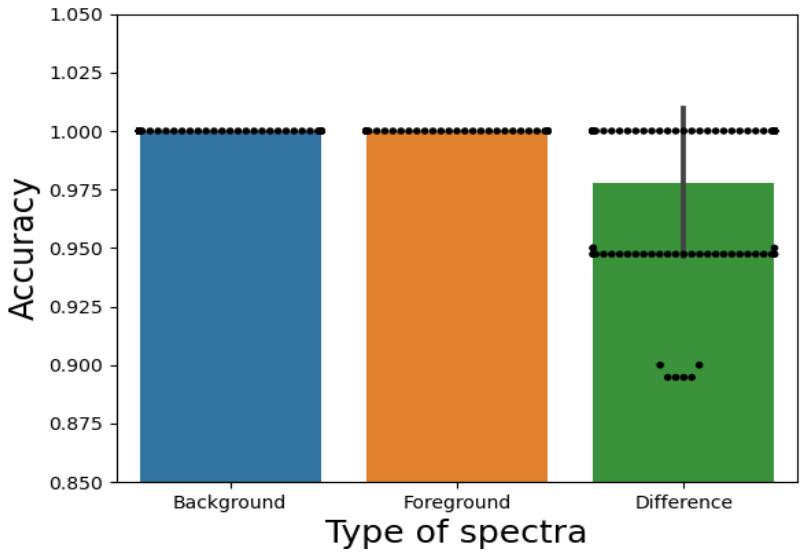

Figure 2: The mean and the standard deviation of validation accuracies of classifications of substrates based on the background, foreground, and difference spectra. Points represent distributions of validation accuracies. In favor of clarity, not all points are shown.

Figure 4 shows validation accuracies of classifiers with respect to the number of decision trees and the number of features considered at each node. Other parameters were changed as well, but they always gave the same or lower validation accuracy. Hence, all of the parameters were set to the default values. In Table 2 and Figure 3 are given validation accuracies of those classifiers that classified fungal strains based on the type of spectra. We can see that taking derivatives drastically increased the validation accuracy of the classifier. Also, we can see that the validation accuracy is smaller when the second derivative data was used compared to the first derivative data. One possible cause for this is that a higher derivatives of a spectrum focus more on smaller regions of the spectrum, increasing a noise to signal ratio, and decreasing the accuracy of the classifier [12].

Confusion matrices of the above-mentioned classifications are given in Figures 5, 6, and 7. Rows represent the correct species of samples, while columns show how those samples were classified. Interestingly, the classifier got better results for all species except for *Neurospora crassa* (NC4200) when the first derivative data was used. We can see that the pattern of misclassification stayed the same. We are not sure exactly why this is the case.

Table 2: The mean and the standard deviation of validation accuracies of classifications of fungal strains based on the difference, the first derivative, and the second derivative spectra.

Accuracy / Type of spectra	Mean	Standard deviation
Difference	68.3 %	11.6 %
First derivative	85.3 %	7.0 %
Second derivative	80.7 %	9.2 %

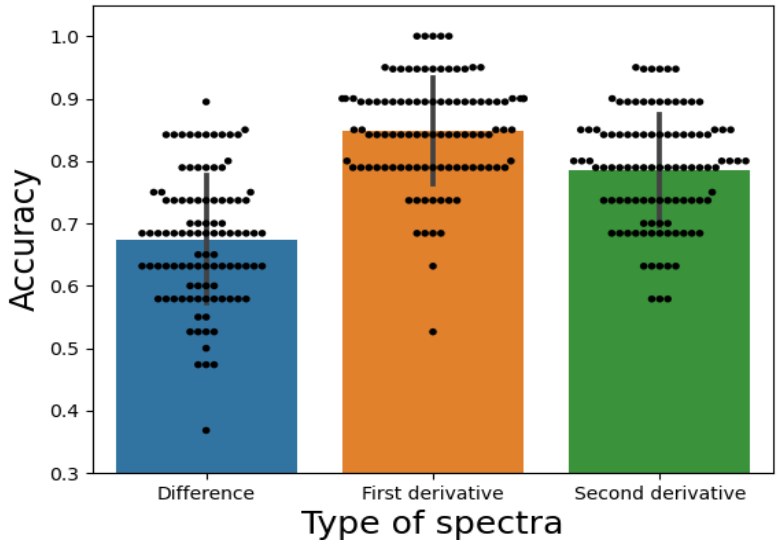

Figure 3: The mean and the standard deviation of validation accuracies of classifications of fungal strains based on difference, the first and the second derivative spectra. Points represent distributions of validation accuracies. In favor of clarity, not all points are shown.

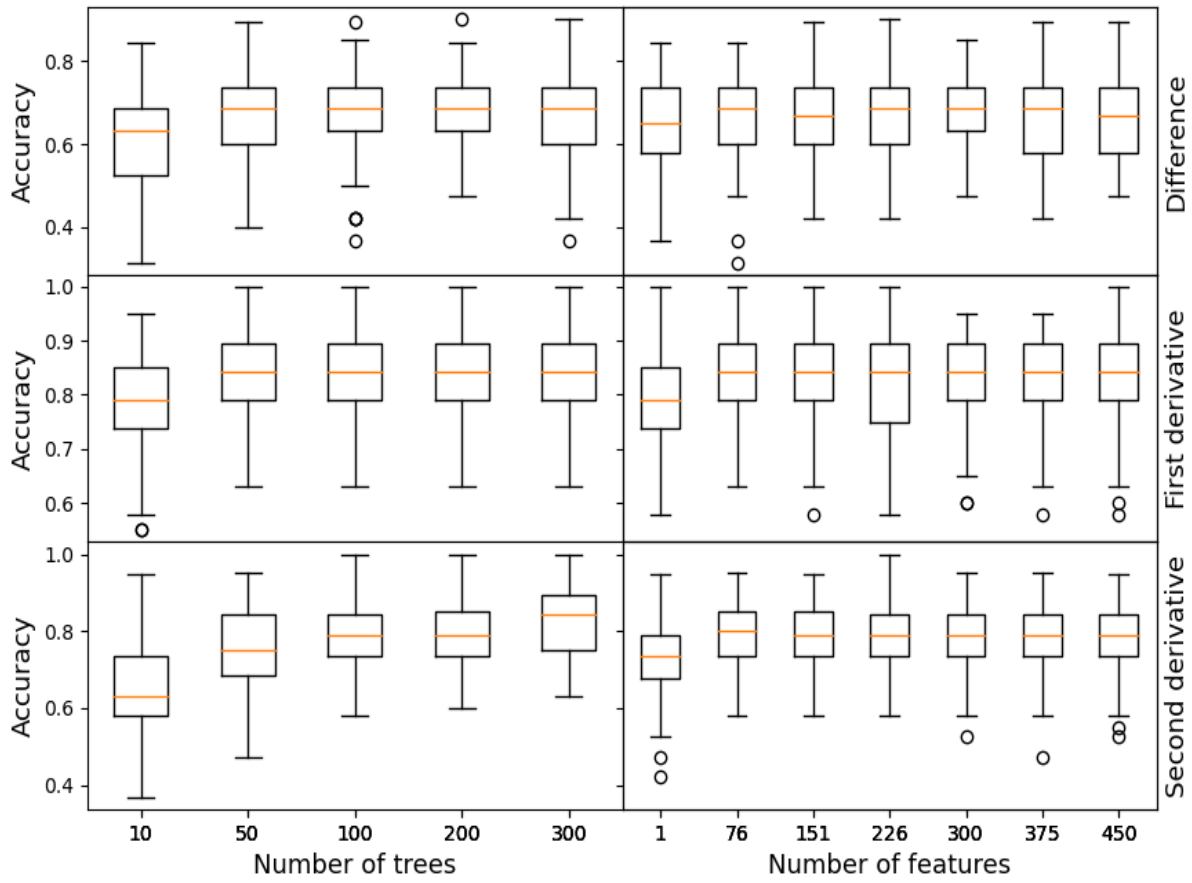

Figure 4: validation accuracies of classifications of fungal strains based on the difference, the first derivative, and the second derivative spectra are plotted with respect to the number of decision trees and the number of features considered at each node.

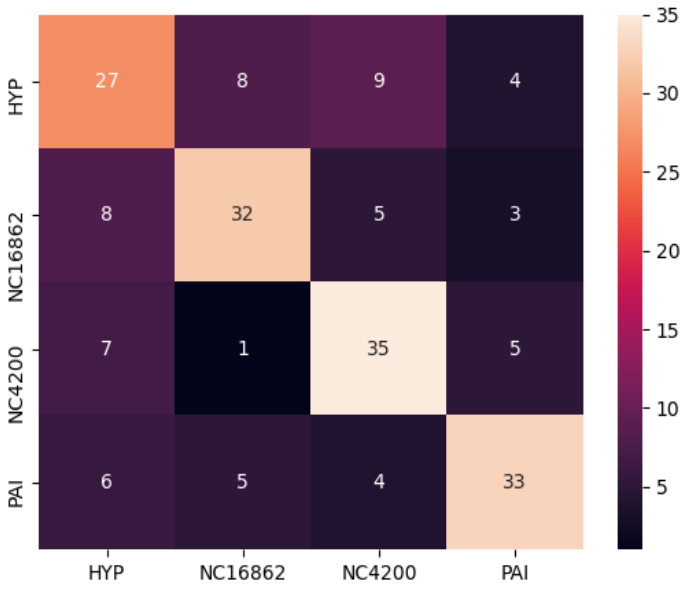

Figure 5: Confusion matrix for classification of fungal strains based on the difference spectra. Rows represent correct species of samples, while columns show how those samples were classified.

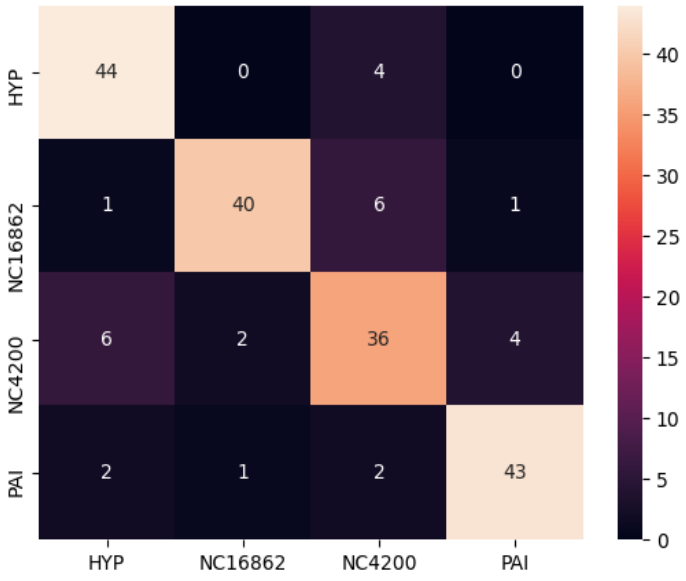

Figure 6: Confusion matrix for classification of fungal strains based on the first derivative of the difference spectra. Rows represent correct species of samples, while columns show how those samples were classified.

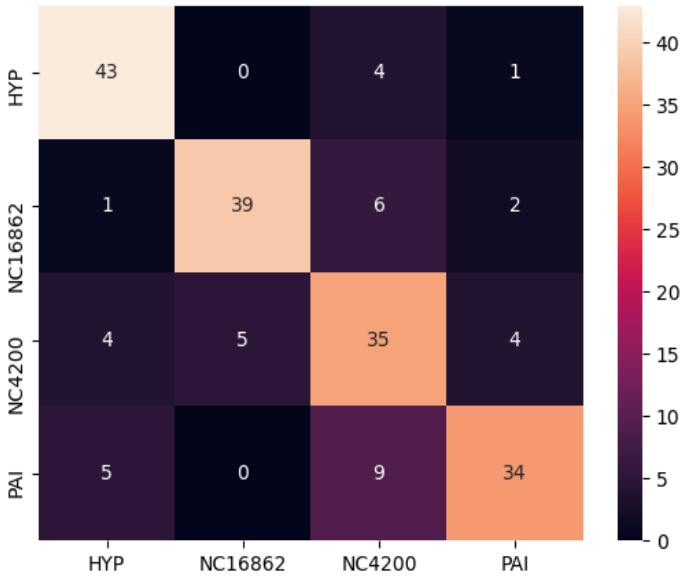

Figure 7: Confusion matrix for classification of fungal strains based on the second derivative of the difference spectra. Rows represent correct species of samples, while columns show how those samples were classified.

Table 3 and Figure 8 show classification validation accuracies of different normalization procedures. For the difference data it looks like normalization procedures do not affect the validation accuracy of the classifier, while for the first derivative and the second derivative spectra make a difference. We can see that normalizing the data by the mean significantly lowers the validation accuracy. Furthermore, normalizing the same data by the maximum value yields slightly better results (for the second derivative data increase is negligible).

Table 3: The mean and the standard deviation of classification validation accuracies of fungal strains based on differently normalized difference, the first derivative, and the second derivative spectra. Mean represents normalization of spectra by the mean value, while max represents normalization of spectra by the max value.

Normalization / Spectra	Mean	Max
Difference	69.0 % (σ = 10.3 %)	68.0 % (σ = 10.0 %)
First derivative	78.5 % (σ = 9.6 %)	88.3 % (σ = 7.6 %)
Second Derivative	65.8 % (σ = 12.0 %)	81.2 % (σ = 9.2 %)

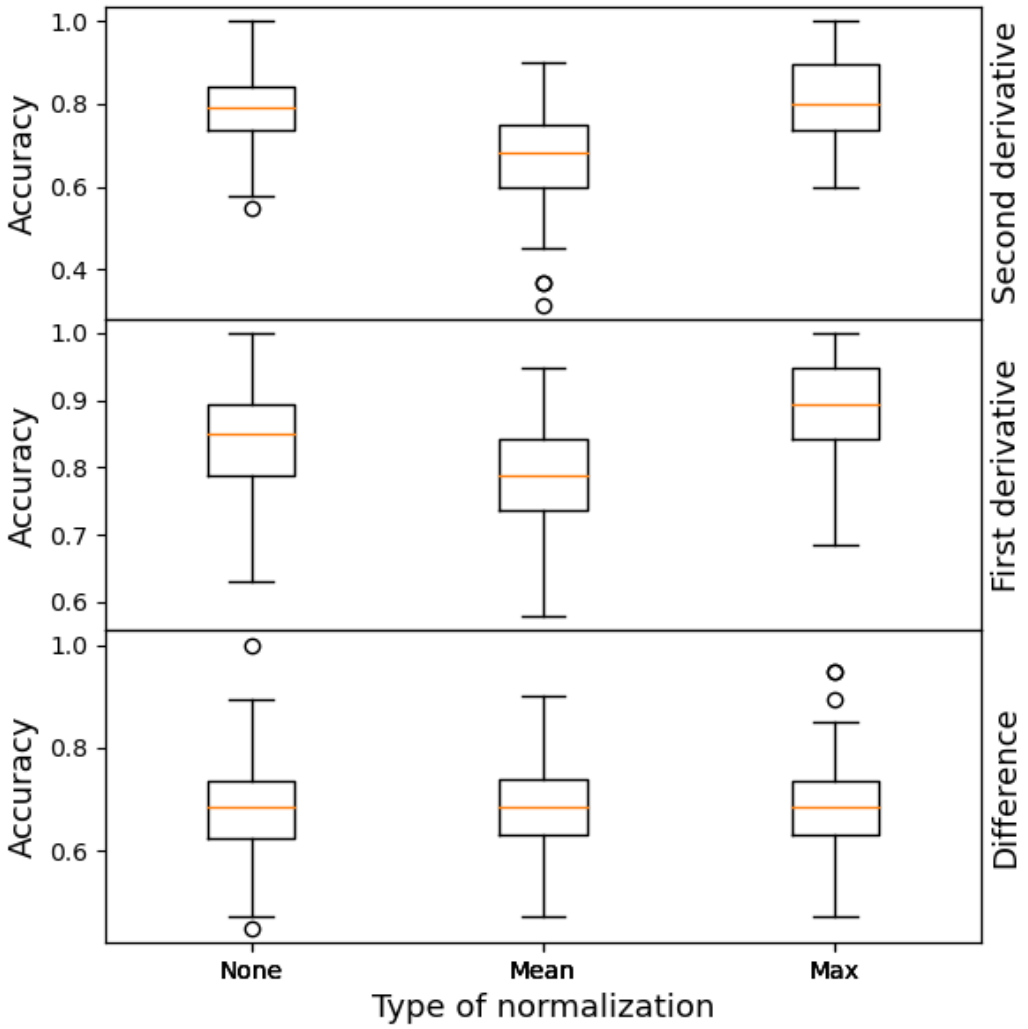

Figure 8: Box plot of validation accuracies of classifications of fungal strains based on differently normalized difference, the first derivative, and the second derivative spectra. Mean represents normalization of spectra by the mean value, while max represents normalization of spectra by the max value.

In Table 4 and Figure 9 are given validation accuracies of classifications of species based on different substrates. One thing to note is that classifiers used a smaller number of data points, around 86, and a smaller number of points for testing, around 10, for each cross-validation's step. However there is an obvious trend of the ligning-casein mix scoring higher than the casein alone.

Table 4: The mean and the standard deviation of validation accuracies of classifications of fungal strains grown on different substrates, pure casein and a mix of casein and lignin, based on the difference, the first derivative, and the second derivative spectra.

Spectra \ Substrate	Casein	Lignin-Casein
Difference	61.8 % (σ = 16.2 %)	73.1 % (σ = 14.1 %)
First derivative	82.8 % (σ = 10.7 %)	85.9 % (σ = 10.2 %)
Second Derivative	72.1 % (σ = 15.8 %)	80.3 % (σ = 13.1 %)

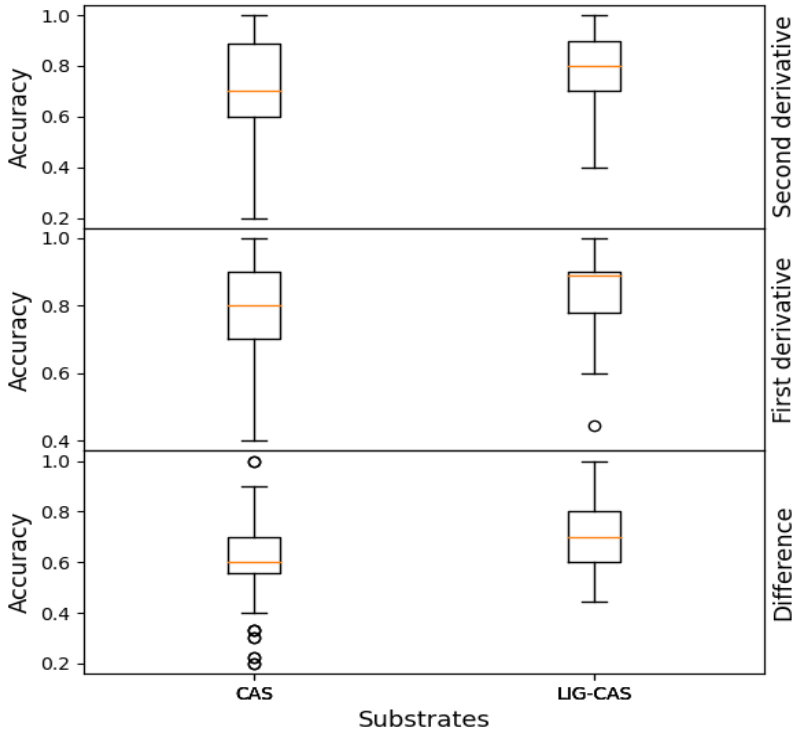

Figure 9: Box plot of classification validation accuracies of fungal strains grown on different substrates, pure casein and a mix of casein and lignin, based on the difference, the first derivative, and the second derivative spectra.

We inspected the images that were misclassified but we were not able to notice anything special in regards to those images that would set them apart (i.e. the quality of those images did not differ from the quality of other images). However, we noticed that strange periodic patterns show up on most images. We are not sure whether this was artifact of experiments or of the image processing in OCTAVVS.

Figure 10 shows correctly and incorrectly classified spectra. From the figure we can see that NC4200 stays the most misclassified species with the normalized first derivative data.

Another thing we can see is that it looks like there are two types of NC4200 first derivative spectra. To be more specific, it seems that one type of spectra are almost flat above 1600 cm^{-1}, while the other type has a dip in the same region. It appears that spectra with a dip got mostly misclassified.

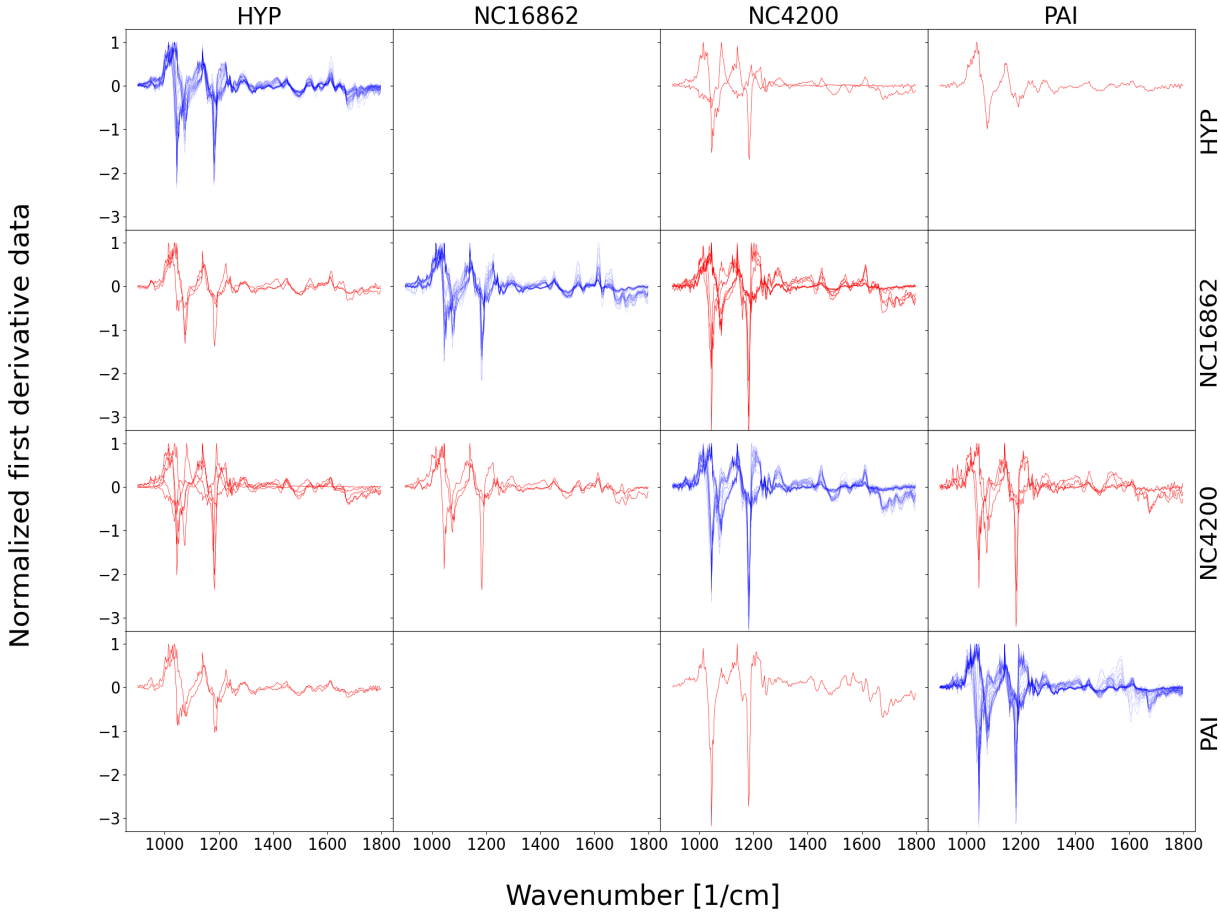

Figure 10: Plots of correctly and incorrectly classified normalized first derivative spectra (normalization by the maximum value). Rows represent correct species of samples, while columns show how those samples were classified.

Figure 11 shows the performance of the three methods for finding the most important wavenumbers (for method ordering see the last part of the Methods section). From the figure it can be observed that the third method is slightly better and reaches peak validation accuracy faster than the other two methods. Interestingly, we can see that using the third method we get a validation accuracy around 90% while using only ten wavenumbers. This suggests that chemical substances by which these species differ are probably infrared active at some of those wavenumbers. In Table 5 are given those top ten wavenumbers together with the iteration at which they were obtained.

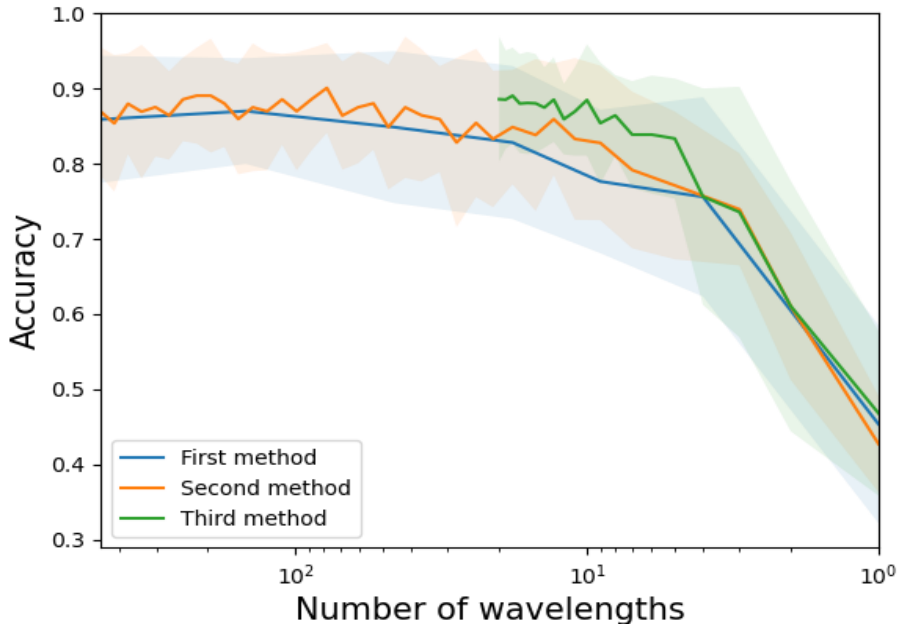

Figure 11: Comparison of three procedures (described in the Methods section) for obtaining the most important wavenumbers. The shaded regions represent obtained standard deviations for each method.

Table 5: Top ten wavenumbers for classifying fungal strains based on the third method (described in the last paragraph of the Methods section).

Iteration	Wavenumber [1/cm]
1	977.8
2	1656.7
3	1512.0
4	1618.1
5	1652.8
6	1010.6
7	943.1
8	1353.9
9	1731.9
10	1645.1

Figures 12 and 13 show the dynamics of feature importance as number of iteration increases for the first and second method, respectively. In Figure 13 not all values are given on the y-axis for the sake of clarity. Using those figures and Table 5, we can compare the important wavenumbers (regions) obtained by those three methods. We can see that most of the wavenumbers locations are around 1500 cm^{-1} and 1650 cm^{-1} and that they are the same for all three procedures. However, the third method selected wider range of wavenumbers

(several wavenumbers around 970 cm^{-1} and one at 1353.9 cm^{-1} and at 1731.9 cm^{-1}) while the other two methods centered only around three regions (the third region being around 1175 cm^{-1}). Additionally, from Figure 11 we can see that the third method attains the higher validation accuracy with the smaller standard deviation with smaller number of wavenumbers than the other two methods. This indicates that a wider range of wavenumbers is better for the validation accuracy of the classification.

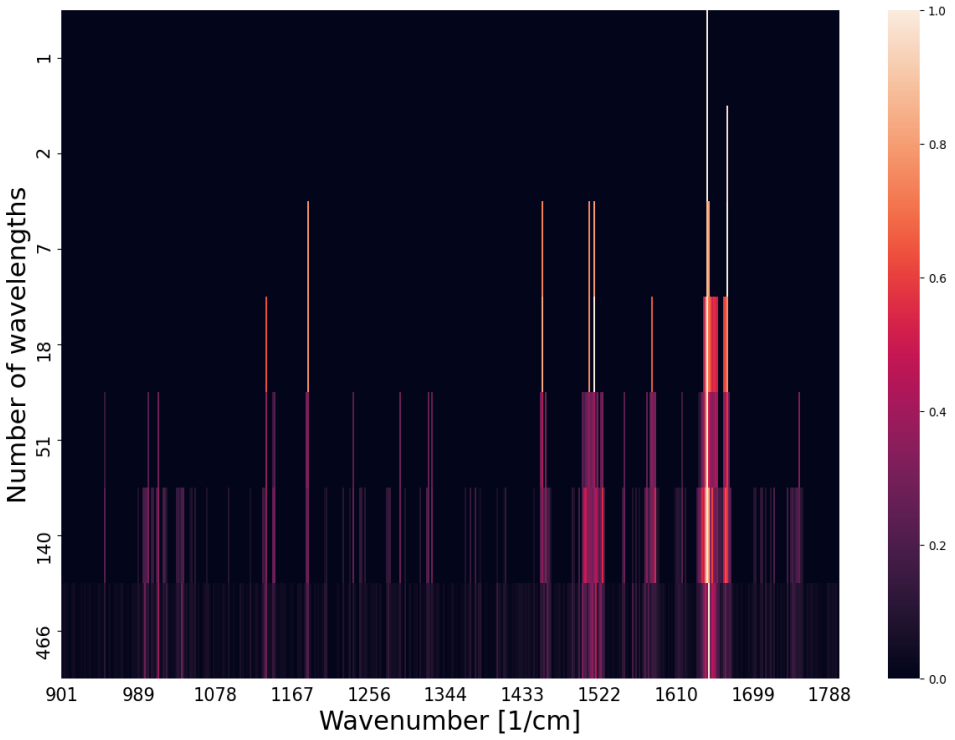

Figure 12: Heat map of feature importance for each wavenumber at each iteration obtained using the first method (described in the last paragraph of the Methods section). Importance is normalized by the maximum importance in each iteration, meaning that relative importances for wavenumbers at each iteration are shown.

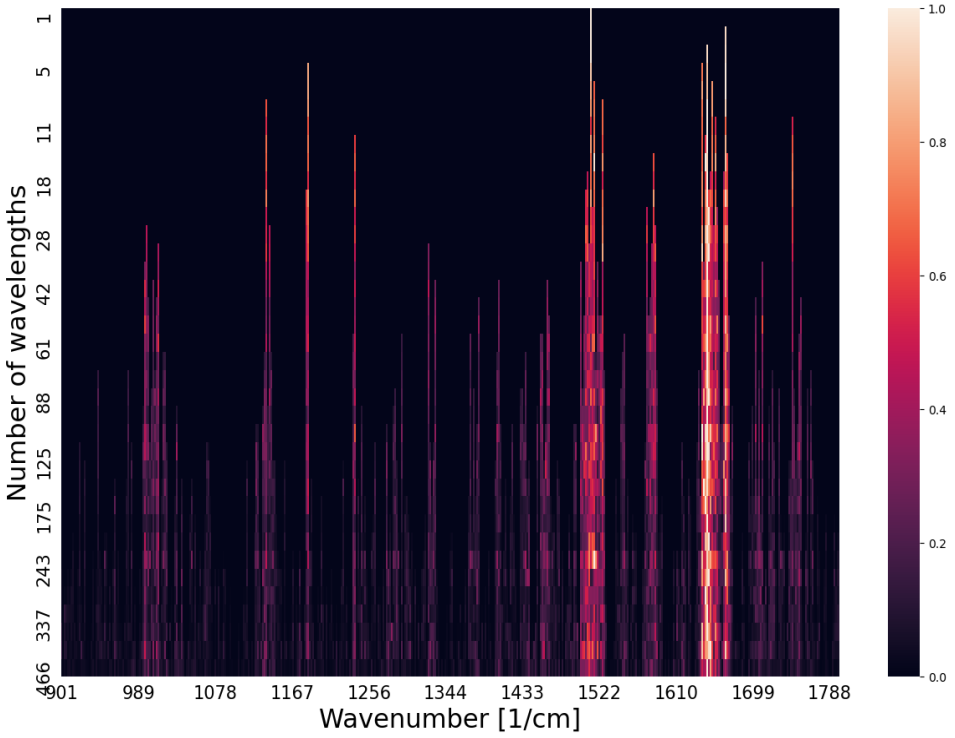

Figure 13: Heat map of feature importance for each wavenumber at each iteration obtained using the second method (described in the last paragraph of the Methods section). Importance is normalized by the maximum importance in each iteration, meaning that relative importances for wavenumbers at each iteration are shown.

5 Conclusion and Outlook

In this section, we repeat the main points of this thesis and suggest possible avenues for future research. The main goal of this thesis was to train a random forest classifier to classify several fungal strains based on hyperspectral data and from that to acquire more information about the samples.

Several modifications of spectra (e.g. taking derivatives), as well as the unmodified spectra, were used for the classification. Interestingly, the first derivative of the difference between the foreground and the background spectra performed better than the other types of spectra. Additionally, it was observed that dividing the first derivative spectra by their maximum value gave the best results (slightly better than the first derivative without normalization). Looking at the confusion matrix of the classification, it was observed that *Neurospora crassa* without mutation got misclassified noticeably more than the other strains. We have discussed several possible reasons for this in the text.

Exploring the data further, the random forest classifier was applied to the normalized first derivative spectra of strains grown only on casein or on casein-lignin substrates. From that analysis, interestingly, it was discovered that the classifier performed better for a casein-lignin mixture. However, for more definite results more data points should be used.

In the last part of the project, the most important wavenumbers (features) for the classification were determined by three different methods. The third method, the one that starts with an empty set and adds wavenumbers to the set, starts performing better earlier than the other two methods, achieving a validation accuracy of around 90% with only ten wavenumbers. By looking at the most important wavenumbers obtained by the third method, we can see that there are wavenumbers from a wider range of the spectrum. In the future, it would be interesting to use OPTIR spectroscopy to probe the samples at the wavenumbers from Table 5.

There are several other directions that would be compelling to explore further. One of them is the classification of fungal strains using a combination of two or more different types of spectra together. As a suggestion, it might be beneficial to combine the difference spectra, that possess more global features, together with the second derivative spectra, that zoom in more on the local features, giving a better overall picture of the differences among the fungal strains.

Another suggestion we have is to see exactly how substrates affect the classification. It is possible that some metabolic pathways for degrading casein of these species have more in common than respective pathways for degrading lignin causing classification to be more difficult for one type of substrates than the other. We believe that a more thorough investigation of this should be performed with more data. Also, we suspect that growing the same strains on the lignin substrate would provide more illuminating results.

References

[1]M. Op De Beeck, C. Troein, S. Siregar, L. Gentile, G. Abbondanza, C. Peterson, P. Persson, and A. Tunlid, "Regulation of fungal decomposition at single-cell level", The ISME journal **14**, 896–905 (2020).

[2]C. Troein, S. Siregar, M. Op De Beeck, C. Peterson, A. Tunlid, and P. Persson, "Octavvs: a graphical toolbox for high-throughput preprocessing and analysis of vibrational spectroscopy imaging data", Methods and protocols **3**, 34 (2020).

[3]P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining* (Pearson Education India, 2016).

[4]*Decision Tree Classification in Python*, `https://www.datacamp.com/community/tutorials/decision-tree-classification-python` (visited on 01/17/2022).

[5]M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, et al., "Using Fourier transform IR spectroscopy to analyze biological materials", Nature protocols **9**, 1771–1791 (2014).

[6]*O-PTIR non-contact submicron visible probe infrared spectroscopy*, `https://www.photothermal.com/o-ptir/` (visited on 12/28/2021).

[7]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research **12**, 2825–2830 (2011).

[8]C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy", Nature **585**, 357–362 (2020).

[9]The pandas development team, "Pandas-dev/pandas: pandas", version latest, `10.5281/zenodo.3509134` (2020).

[10]M. L. Waskom, "Seaborn: statistical data visualization", Journal of Open Source Software **6**, 3021 (2021).

[11]J. D. Hunter, "Matplotlib: A 2D graphics environment", Computing in Science & Engineering **9**, 90–95 (2007).

[12]A. J. Owen, *Uses of derivative spectroscopy* (Agilent Technologies, 1995).