PLANTGEN'23

PLANT GENETICS, GENOMICS, BIOINFORMATICS AND BIOTECHNOLOGY

Hyperspectral Image Analysis Pipeline

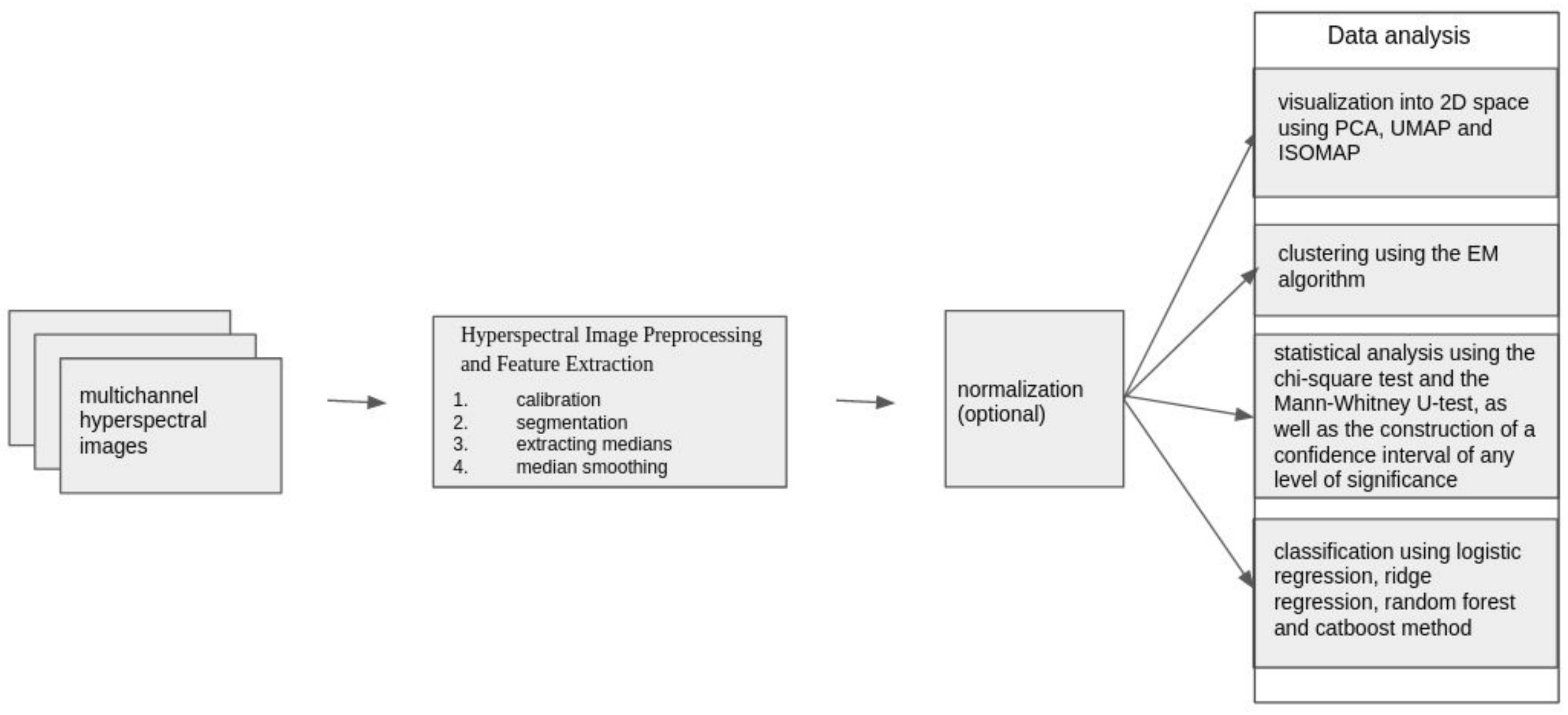
Busov I.D., Genaev M.A., Komyshev E. G., Zykova T. E., Afonnikov D.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Relevance

The analysis of hyperspectral images has a great interest in various problems. However, the development of algorithms for processing and analyzing such data is a laborious task. This work presents a pipeline for hyperspectral image analysis, which can significantly reduce the time to obtain the results of the research of hyperspectral images.

Methods



multichannel hyperspectral images

Hyperspectral Image Preprocessing and Feature Extraction

1. calibration
2. segmentation
3. extracting medians
4. median smoothing

normalization (optional)

Data analysis

- visualization into 2D space using PCA, UMAP and ISOMAP
- clustering using the EM algorithm
- statistical analysis using the chi-square test and the Mann-Whitney U-test, as well as the construction of a confidence interval of any level of significance
- classification using logistic regression, ridge regression, random forest and catboost method

At the input, the pipeline gets multichannel hyperspectral images, each channel of which corresponds to the reflection intensity in a certain range of wavelengths.

Hyperspectral Image Preprocessing and Feature Extraction

Firstly, the images were calibrated according to formula 1, and using the threshold function of the OpenCV library, the area of the Petri dish with grains was segmented on the images. Then, for each image, the medians for each channel were calculated from the pixel values in the segmented area. After that, the Savitzky-Golay filter was applied to smooth the median values, and each sample was subsequently identified with this median vector.

$$R = \frac{I_S - I_D}{I_W - I_D}$$

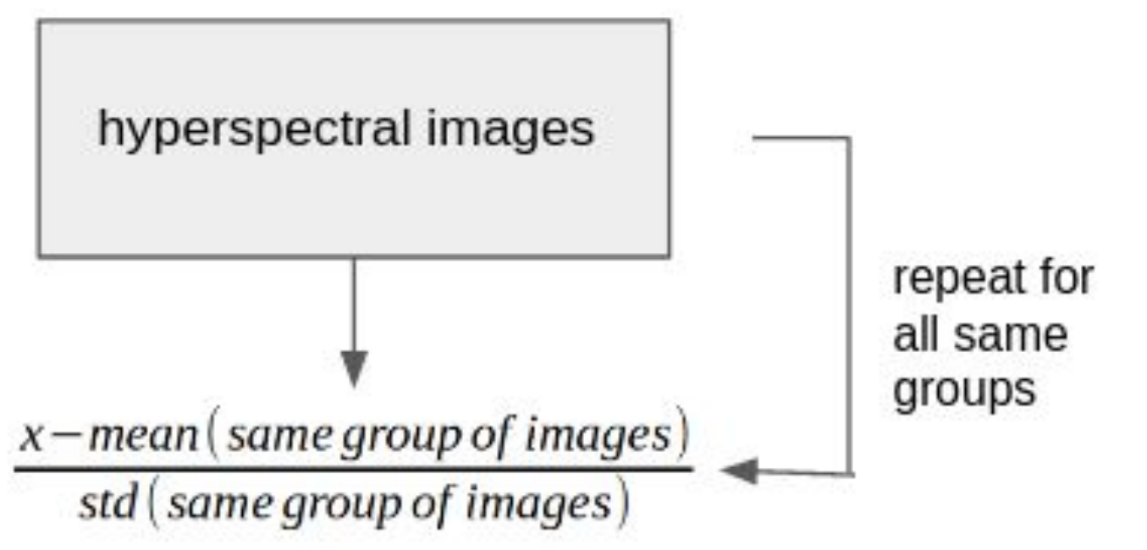
Formula 1. “IS“ is a hyperspectral image, “ID“ is a black background calibration image, “IW“ is a white background calibration image, “R” is an image obtained as a result of calibration.

Normalization

To level out the differences that arise between shooting series, 2 methods of image normalization were implemented in the pipeline. If the same samples were present in each series, and if the change in the hyperspectrum of these samples between the acquisition of the first and last series of surveys is insignificant, then it can be assumed that the filtered medians of the channels of hyperspectral images of the same samples in each series will have the same distribution, and use normalization by the same samples (formula 2). If we also assume that the same groups should have the same distribution in each series, then the second normalization method can be applied (Scheme 1).

$$\frac{x - mean(same\ images)}{std(same\ images)}$$

Formula 2. x is the image (vector of medians) in a fixed series, mean (same images) is the vector of sample means of images for the same samples, std is the root of the unbiased sample variance of the images for the same samples.



hyperspectral images

$$\frac{x - mean(same\ group\ of\ images)}{std(same\ group\ of\ images)}$$

repeat for all same groups

Scheme 1. Sequentially for each group within each series, the mean is subtracted and divided by the standard deviation. x is the image (vector of medians) in a fixed series, mean(same group of images) is the vector of sample means of images for samples of one common group, std is the root of the unbiased sample variance of images for samples of one common group.

Dimensionality reduction

Three dimensionality reduction methods were added to the pipeline: PCA, ISOMAP and UMAP. For visualization using one of these methods, a scatterplot is plotted in two-dimensional space.

Clustering

The pipeline implements clustering* using the EM algorithm. We assumed that each sample does not strictly belong to one cluster, but has a certain probability of belonging to each cluster, and that each sample obeys the model of a mixture of Gaussian distributions. Distribution parameters were found using the maximum likelihood method using the EM algorithm. The pipeline also returns a table with information about the most frequently occurring group in each cluster and the sampling percentage of that most popular group in the cluster.

Statistical analysis

In the created pipeline, for the difference between the sample means of two groups of images, you can build a confidence interval of any level of confidence. To test the hypothesis about the coincidence of the distributions of the two groups, the Mann-Whitney U-test, as well as the chi-square test, were added to the pipeline.

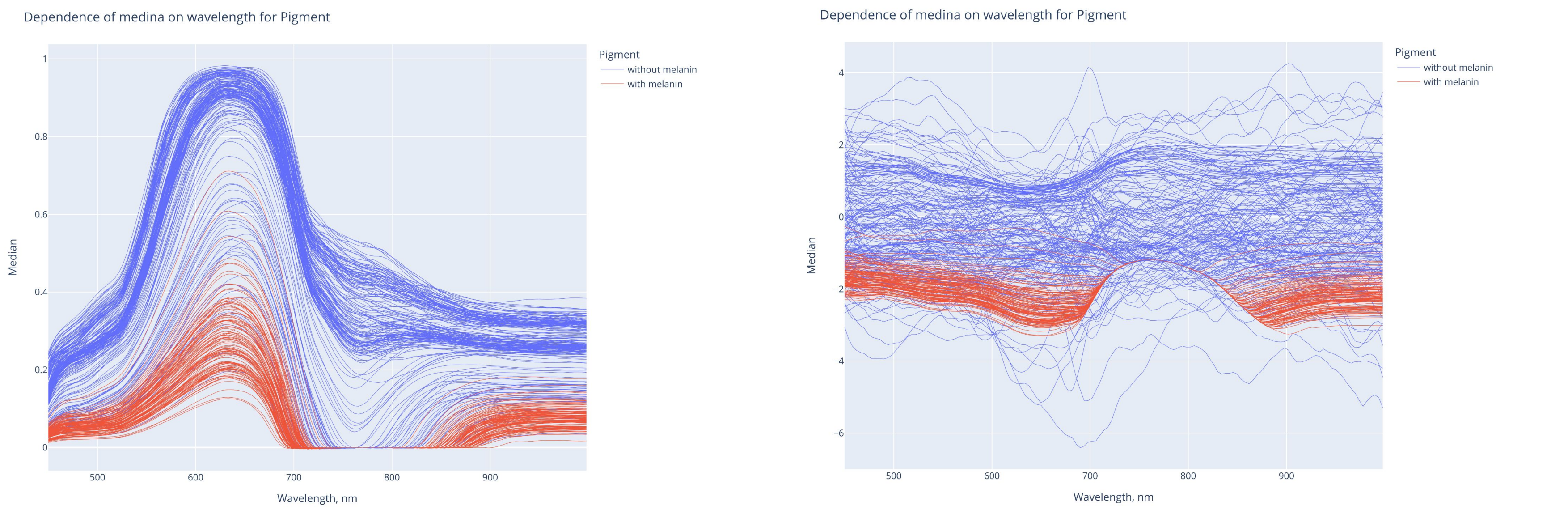
*To avoid the “curse of dimensionality”, clustering and classification is performed in a space of lower dimensionality (the dimensionality of the space and the dimensionality reduction method are hyperparameters).

Classification

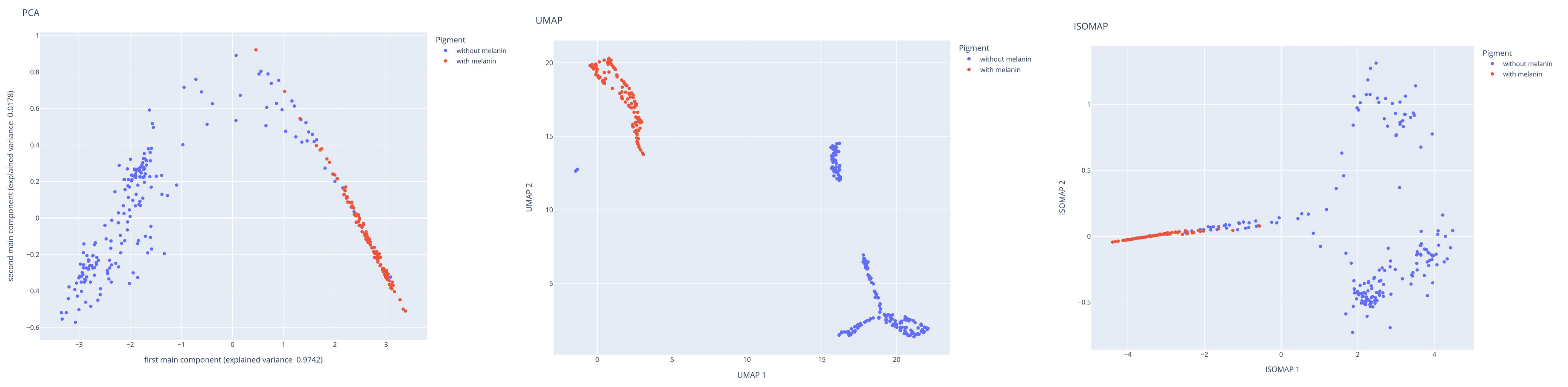
The developed pipeline allows classifying hyperspectral images using methods such as logistic regression, ridge regression, random forest, and catboost. The pipeline returns 2 or 3 tables with classification results for metrics such as accuracy, f1, precision, and recall, as well as error matrices for each classifier. The first table contains the results of classification by macro-metrics, the second by micro-metrics, and if a function is passed to the pipeline that converts a group into a vector according to a special rule, the pipeline will also return the third table with the average results of binary classification* for each individual component of the vector.

Results

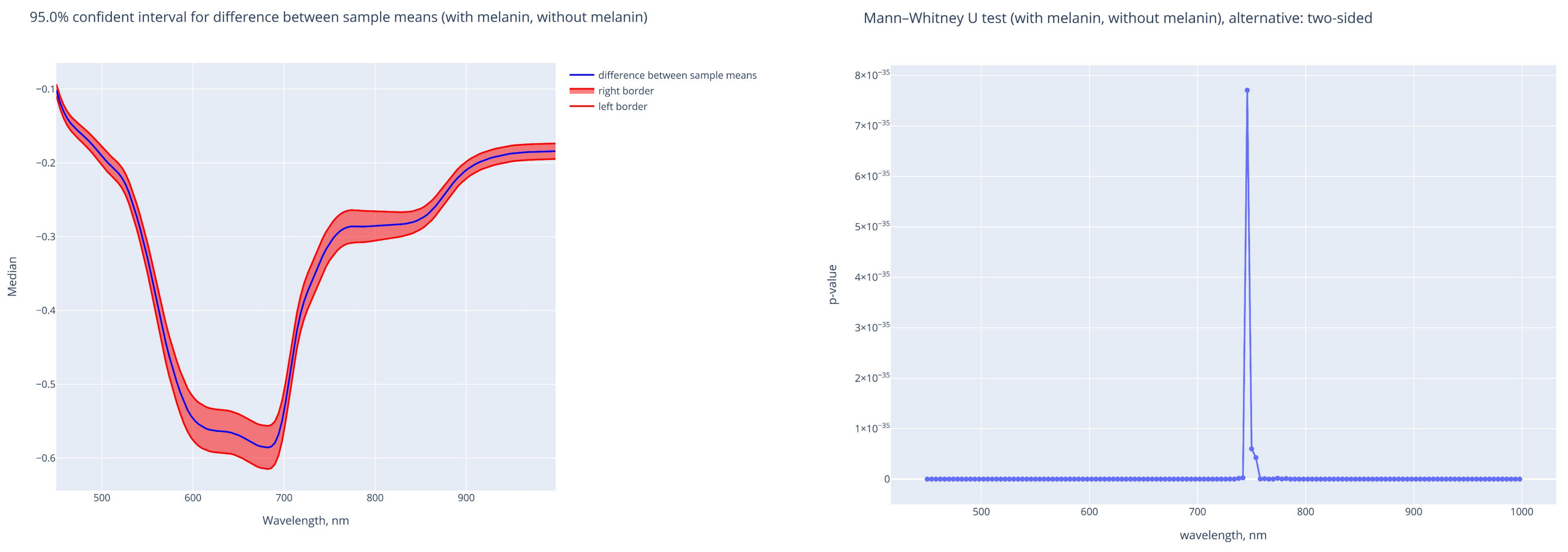
The developed pipeline was tested on the task of analyzing barley grains containing melanin. Seeds of 313 samples of barley (Hordeum vulgare) were selected for the study, of which 117 samples contained melanin, and the remaining 196 samples did not have this pigment. Samples were obtained from 3 series of surveys. In two series there were no grains containing melanin. There were no identical samples in different series of surveys. Hyperspectral images of grains were obtained using a Cubert S185 camera with Cinegon 1.8 16 lenses installed.



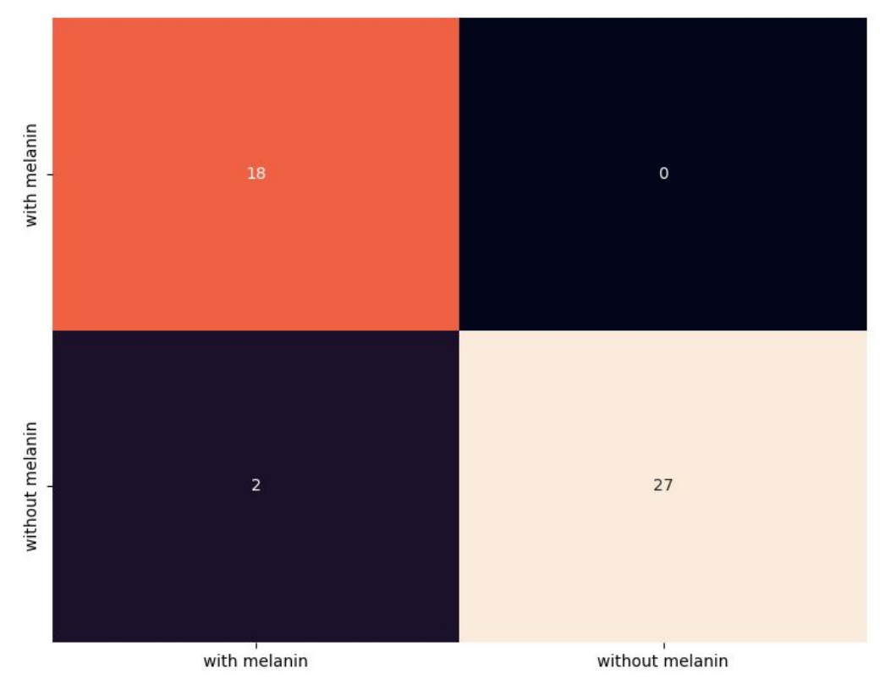
Plots of dependence of medians on wavelength. The blue lines correspond to the medians of images of barley grains that did not contain melanin, and the red lines correspond to the medians of images of grains with melanin. On the left without normalization, on the right with normalization for the same groups. All graphs in the pipeline are interactive. Already on these graphs, one can see that the hyperspectrum of grains containing melanin is very different from the hyperspectrum of grains without this pigment.



Visualization in 2D space with PCA, UMAP and ISOMAP (left PCA, center UMAP, right ISOMAP) without normalization. Blue dots correspond to samples without melanin, and red dots with melanin. All graphs visualized in two-dimensional space show that the data is linearly separable with high accuracy.



95 percent confidence interval for the difference in the sample means. The blue line is the values of the difference between the sample means. The red area is the 95 percent confidence interval. And a plot of p-value versus wavelength for the Mann-Whitney U test. The images were not normalized. Statistical analysis revealed statistically significant differences in the distribution of channel medians throughout the hyperspectrum.



	with melanin	without melanin
with melanin	18	0
without melanin	2	27

Logistic regression confusion matrix for the test sample in dimensional space 15, using PCA for downsizing. The test set size is 47 samples, the training set size is 266 samples. The test sample was stratified. With the help of hyperspectral images, the presence or absence of melanin in barley grains can be determined with high accuracy.

The software pipeline is implemented in Python using the following libraries: Pandas, NumPy, OpenCV, SciPy, Sklearn, UMAP CatBoost, and Plotly. The source code is available at: https://github.com/igor2704/Hyperspectral_images.

Conclusion

In this work, a convenient and flexible tool was created that can significantly reduce the time for developing algorithms for processing and analyzing hyperspectral images. The developed pipeline was tested on the task of studying the effect of melanin on the hyperspectrum of barley grains. The results of the analysis revealed significant differences between the target groups.

The work was supported by the Russian Science Foundation, the project 22-74-00122, https://rscf.ru/project/22-74-00122

10-15 July | Kazan, Tatarstan, Russia

https://plantgen2023.ofr.su