

DOI 10.18699/vjgb-24-50

Конвейер обработки гиперспектральных изображений на примере исследования зерен ячменя, содержащих меланин

И.Д. Бусов^{1, 2} , М.А. Генаев^{1, 2}, Е.Г. Комышев¹, В.С. Коваль¹, Т.Е. Зыкова^{1, 2}, А.Ю. Глаголева¹, Д.А. Афонников^{1, 2} 

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский государственный университет, Новосибирск, Россия

 i.busov@gs.nsu.ru

Аннотация. Анализ гиперспектральных изображений представляет большой интерес при изучении растений. В настоящее время такой анализ используется все более широко, поэтому создание методов обработки гиперспектральных изображений является актуальной задачей. В статье представлен конвейер для работы с гиперспектральными изображениями, который включает: предварительную обработку, базовый статистический анализ, визуализацию многоканального гиперспектрального изображения, а также решение задач классификации и кластеризации с применением методов машинного обучения. В текущей версии пакета программ реализованы следующие методы: построение доверительного интервала произвольного уровня для разницы выборочных средних; проверка сходства распределений интенсивности линий спектра для двух наборов гиперспектральных изображений на основе U-критерия Манна–Уитни и критерия согласия Пирсона; визуализация в двухмерном пространстве с применением методов понижения размерности PCA, ISOMAP и UMAP; классификация с использованием линейной или гребневой регрессии, случайного леса и градиентного бустинга; кластеризация образцов с помощью EM-алгоритма. Программный конвейер реализован на языке Python с использованием библиотек Pandas, NumPy, OpenCV, SciPy, Sklearn, Umap, CatBoost и Plotly. Исходный код доступен по адресу: https://github.com/igor2704/Hyperspectral_images. Данный конвейер был применен для идентификации пигмента меланина в оболочке зерен ячменя на базе гиперспектральных данных. Визуализация на основе методов PCA, UMAP и ISOMAP, а также использование алгоритмов кластеризации показали, что на базе гиперспектральных данных с высокой точностью можно провести линейное разделение образцов зерен с пигментацией и без нее. Анализ выявил статистически значимые различия в распределении медиан интенсивности для выборок изображений зерен с пигментом и без него. Таким образом, продемонстрировано, что с помощью гиперспектральных изображений с большой точностью можно определить наличие или отсутствие меланина в зернах ячменя. Созданный в данной работе гибкий и удобный инструмент позволит существенно повысить эффективность анализа гиперспектральных изображений.

Ключевые слова: гиперспектральные изображения; машинное обучение; статистический анализ; зерна ячменя; пигментный состав.

Для цитирования: Бусов И.Д., Генаев М.А., Комышев Е.Г., Коваль В.С., Зыкова Т.Е., Глаголева А.Ю., Афонников Д.А. Конвейер обработки гиперспектральных изображений на примере исследования зерен ячменя, содержащих меланин. *Вавиловский журнал генетики и селекции*. 2024;28(4):443-455. DOI 10.18699/vjgb-24-50

Финансирование. Разработка структуры конвейера, алгоритмов и программ выполнена при поддержке РНФ, проект № 22-74-00122.

Благодарности. Тестирование конвейера и обработка данных проведены с использованием вычислительных ресурсов ЦКП «Биоинформатика» (при поддержке бюджетного проекта № FWNR-2022-0020).

A pipeline for processing hyperspectral images, with a case of melanin-containing barley grains as an example

I.D. Busov^{1, 2} , M.A. Genaev^{1, 2}, E.G. Komyshev¹, V.S. Koval¹, T.E. Zyкова^{1, 2}, A.Y. Glagoleva¹, D.A. Afonnikov^{1, 2} 

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

 i.busov@gs.nsu.ru

Abstract. Analysis of hyperspectral images is of great interest in plant studies. Nowadays, this analysis is used more and more widely, so the development of hyperspectral image processing methods is an urgent task. This paper presents a hyperspectral image processing pipeline that includes: preprocessing, basic statistical analysis, visualization of a multichannel hyperspectral image, and solving classification and clustering problems using machine learning methods. The current version of the package implements the following methods: construction of a confidence interval of an arbitrary level for the difference of sample averages; verification of the similarity of intensity distributions of spectral lines for two sets of hyperspectral images on the basis of the Mann–Whitney U-criterion and Pearson's

criterion of agreement; visualization in two-dimensional space using dimensionality reduction methods PCA, ISOMAP and UMAP; classification using linear or ridge regression, random forest and catboost; clustering of samples using the EM-algorithm. The software pipeline is implemented in Python using the Pandas, NumPy, OpenCV, SciPy, Sklearn, Umap, CatBoost and Plotly libraries. The source code is available at: https://github.com/igor2704/Hyperspectral_images. The pipeline was applied to identify melanin pigment in the shell of barley grains based on hyperspectral data. Visualization based on PCA, UMAP and ISOMAP methods, as well as the use of clustering algorithms, showed that a linear separation of grain samples with and without pigmentation could be performed with high accuracy based on hyperspectral data. The analysis revealed statistically significant differences in the distribution of median intensities for samples of images of grains with and without pigmentation. Thus, it was demonstrated that hyperspectral images can be used to determine the presence or absence of melanin in barley grains with great accuracy. The flexible and convenient tool created in this work will significantly increase the efficiency of hyperspectral image analysis.

Key words: hyperspectral images; machine learning; statistical analysis; barley grains; pigment composition.

For citation: Busov I.D., Genaev M.A., Komyshev E.G., Koval V.S., Zyкова T.E., Glagoleva A.Y., Afonnikov D.A. A pipeline for processing hyperspectral images, with a case of melanin-containing barley grains as an example. *Vavilovskii Zhurnal Genetiki i Selekcii* = *Vavilov Journal of Genetics and Breeding*. 2024;28(4):443-455. DOI 10.18699/vjgb-24-50

Введение

Наличие пигментов в оболочке зерна влияет на его различные технологические свойства. Например, флавоноиды, антоцианы и каротиноиды обладают рядом ценных свойств, являются антиоксидантами и влияют на пищевую ценность зерна. Добавление в муку отрубей пшеницы с пурпурным перикарпием или голубым алейроновым слоем позволяет улучшить качество хлебопекарной продукции за счет вкусовых, текстурных и цветовых характеристик (Machálková et al., 2017). Флабафены, придающие окраске перикарпия зерна красный цвет, положительно воздействуют на продолжительность состояния покоя зерна и препятствуют его предуборочному прорастанию (Flintham et al., 2002). Ввиду этого генотипы пшеницы с красной окраской зерен используются в селекции в качестве доноров генов устойчивости к предуборочному прорастанию зерна (Крупнов, 2013; Fakthongphan et al., 2016).

Генетический контроль формирования окраски как зерен, так и других органов растений осуществляется генами, кодирующими ферменты, вовлеченными в биосинтез пигментов, а также регуляторными генами (Khlestkina, 2014; Lachman et al., 2017; Шоева и др., 2018). Для ряда пигментов эти гены исследованы достаточно хорошо, вплоть до полной расшифровки их нуклеотидных последовательностей и расположения в геноме. Однако для некоторых пигментов, например меланина, определяющего черную окраску зерен ячменя, молекулярные механизмы биосинтеза еще полностью не известны (Glagoleva et al., 2017; Шоева и др., 2018).

Высокопроизводительные, неразрушающие и точные методы измерения играют важную роль при оценке качества семян и улучшении сельскохозяйственного производства (Афонников и др., 2016; Afonnikov et al., 2021). Технологии гипер- и мультиспектральной визуализации, охватывающие видимый, ближний инфракрасный диапазоны длин волн, предоставляют спектральную и пространственную информацию для каждого пикселя изображения. Гиперспектральные изображения – это значения интенсивностей отраженного излучения для сотен интервалов длин волн, что существенно больше, чем для мультиспектральных изображений с несколькими диапазонами волн (Gowen et al., 2007).

За счет уменьшения общего объема данных мультиспектральные системы визуализации нацелены на быстрое

получение изображений с относительно низким пространственным разрешением и могут использоваться в режиме реального времени. Гиперспектральные же изображения обычно применяются в качестве наборов данных, из которых можно определить оптимальные диапазоны волн, которые будут в дальнейшем использоваться при мультиспектральной визуализации в решении конкретной прикладной задачи (Qin et al., 2013). Такие технологии позволяют получать более точную информацию о характеристиках отраженного излучения объектов по сравнению с цифровыми изображениями RGB (трехканальные изображения. Red, green, blue – красный, зеленый, синий).

Анализ гиперспектральных данных успешно применяется для оценки и предсказания урожайности сельскохозяйственных культур. L. Serrano с коллегами прогнозировали биомассу и урожайность озимой пшеницы с помощью спектральных индексов (Serrano et al., 2000). W.S. Weber с коллегами (2012) предсказывали урожайность зерна, используя спектры (495–1853 нм) отражения кроны и листьев растений кукурузы, выращенных при различных водных режимах, и получили наиболее подходящие длины волн для прогнозирования урожайности. X. Zhang и Y. He (2013) разработали метод ранней и быстрой оценки урожайности семян с помощью гиперспектральных изображений листьев масличного рапса в видимой и ближней инфракрасной областях (380–1030 нм). Урожайность семян сои (*Glycine max*) прогнозировалась на основе гиперспектральных данных (395–1005 нм) и алгоритмов машинного обучения: многослойного персептрона, метода опорных векторов и случайного леса, что позволило также выявить наиболее значимый спектр отражения (395 нм) (Yoosefzadeh-Najafabadi et al., 2021).

Анализ гиперспектральной отражательной способности может предоставить достоверную информацию о жизнеспособности семян как сорных (Matzrafi et al., 2017), так и культурных растений: риса (He et al., 2019; Jin et al., 2022), пшеницы (Zhang et al., 2018), кукурузы (Ambrose et al., 2016; Wakholi et al., 2018), арахиса (Zou et al., 2023), дыни (Kandpal et al., 2016), японской шпинатной горчицы (Ma et al., 2020).

На основе гиперспектральных технологий разрабатываются инновационные методы диагностики болезней растений (Чешкова, 2022). Технология гиперспектральных изображений, охватывающих видимый и ближний инфра-

красный диапазон длин волн (400–1000 нм), применялась для анализа риса с целью обнаружения обесцвеченных, больных семян, зараженных бактериальным ожогом метелок (*Burkholderia glumae*). Было показано, что для точной (более 90 %) классификации пораженных патогеном и здоровых растений достаточно определить интенсивность отраженного излучения в небольшом количестве диапазонов длин волн (Baek et al., 2019).

Гиперспектральные изображения используются для определения химического состава семян культурных растений. Показано, что анализ отраженного излучения в ближнем инфракрасном диапазоне (895–2504 нм) имеет потенциал в прогнозировании содержания антоцианов в зернах черного риса (Amanah et al., 2021). С. Liu с коллегами (2020) продемонстрировали возможность использования технологии анализа гиперспектральных данных в ближнем инфракрасном диапазоне (930–2500 нм) для определения содержания крахмала в зернах кукурузы. В работе G. Yang с коллегами (2018) была применена рамановская гиперспектральная технология с линейным сканированием для определения химического состава семян кукурузы. Было обнаружено, что характеристические пики комбинационного рассеяния, идентифицированные при 477, 1443, 1522, 1596 и 1654 нм в спектре от 380 до 1800 нм, были связаны с кукурузным крахмалом, смесью масла и крахмала, зеаксантином, лигнином и маслом в семенах кукурузы соответственно.

Метод неразрушающей оценки концентраций и пространственного распределения влаги, белка и сахаров на разных стадиях развития семян вигны был предложен на основе мультиспектральных данных 20 дискретных длин волн в ультрафиолетовой, видимой и ближней инфракрасной областях (ElMasry et al., 2022). Методы портативной спектроскопии ближнего инфракрасного диапазона и гиперспектральных изображений применяли для количественного определения содержания масла и жирных кислот, а также классификации видов семян рода *Brassica* (da Silva Medeiros et al., 2022). Гиперспектральные изображения применяются для решения задачи классификации для зерен риса (Díaz-Martínez et al., 2023), райграса (Reddy et al., 2023) и многих других важных для аграрной промышленности культур.

Для предоставления гиперспектральной информации семян разрабатываются платформы, такие как HyperSeed, включающая высокопроизводительный спектрограф линейного сканирования (600–1700 нм), а также программное обеспечение с открытым исходным кодом на основе графического пользовательского интерфейса. Система использовалась для классификации семян риса (с точностью до 97.5 %), выращенных в условиях теплового стресса и контрольной среды, с помощью как традиционных моделей машинного обучения, так и моделей нейронных сетей (3D CNN) (Gao et al., 2021).

Таким образом, анализ гиперспектральных изображений представляет большой интерес в различных задачах, связанных с исследованием растений. Однако разработка алгоритмов анализа таких данных является трудоемкой задачей.

Мы применили данный конвейер для определения содержания меланина в зернах ячменя. Несмотря на то что

присутствие меланина обуславливает темную окраску зерна, на практике визуальное определение его наличия затруднительно. Темный цвет зерна может быть связан с увеличением количества пигментов антоцианов, которые, накапливаясь в алейроне зерновки, придают созревшим зернам серый цвет. Зерна ячменя также могут темнеть в процессе хранения. Поэтому точное определение наличия меланина требует дополнительного анализа, например погружения зерен в раствор щелочи для его экстракции.

В данной работе мы представляем инструмент для изучения гиперспектральных изображений – конвейер, использование которого может существенно сократить временные затраты в подобных исследованиях. На примере задачи определения содержания меланина в зернах ячменя продемонстрированы возможности этого конвейера. Для тестирования была выбрана задача исследования спектра зерен, содержащих и не содержащих меланин, так как известно, что в их спектре есть существенные различия. Проведенный нами анализ также показал значимые различия спектра зерен, содержащих меланин, и образцов без этого пигмента. В отличие от других работ в этой области, помимо классификации образцов, перед нами стояла задача – реализовать пакет для облегчения и автоматизации получения результатов исследования гиперспектральных изображений. Разработанный конвейер позволяет визуализировать и кластеризовать входные данные, а также проводить их статистический анализ.

Материалы и методы

Растительный материал. Для исследования были выбраны семена 313 образцов ячменя (*Hordeum vulgare*), из которых 117 содержали меланин, а у остальных 196 образцов данный пигмент отсутствовал (Приложение)¹. Материал был получен из коллекции ячменя Всероссийского института генетических ресурсов растений им. Н.И. Вавилова (<https://www.vir.nw.ru>), коллекции ячменя Института цитологии и генетики СО РАН (<http://www.bionet.nsc.ru>). Также был использован материал популяции Oregon Wolfe Barleys (<https://barleyworld.org/owb>). Биохимический анализ образцов с окрашенным зерном, а также подробное описание метода детекции меланина выполнены А.Ю. Глаголевой с коллегами (2022).

Химический метод определения пигментного состава зерен. Для определения качественного присутствия меланина в зерне была проведена экстракция 2 % NaOH с последующим почернением раствора. На основе этого метода каждому из образцов был присвоен тип пигментации по наличию пигмента: «содержат меланин» либо «не содержат меланин».

Получение изображений. Гиперспектральные изображения зерен получены с помощью камеры Cubert S185 с объективом Cinegon 1.8/16. Для этого на белый матовый лист бумаги формата А3 помещали пластиковую чашку Петри диаметром 55 мм, заполненную зернами без промежутков. По бокам располагался рассеивающий свет, камеру фиксировали на штативе сверху, объективом вертикально вниз. На выходе камера выдавала 138-канальное гиперспектральное изображение, каждый канал которого

¹ Приложение см. по адресу:
<https://vavilovj-icg.ru/download/pict-2024-28/appx16.pdf>

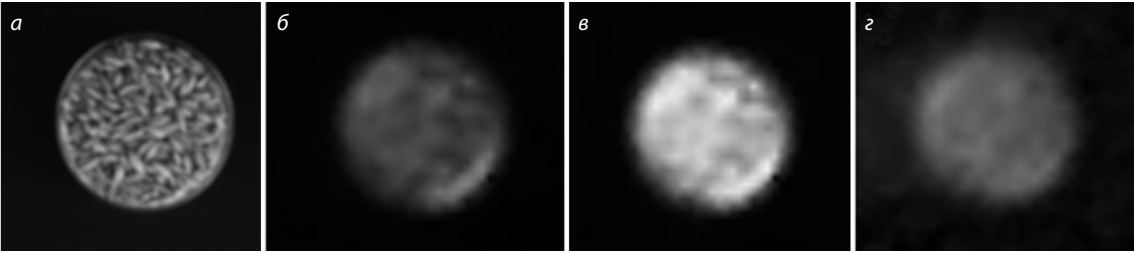


Рис. 1. Изображение зерен ячменя в чашке Петри в оттенках серого (а) и визуализация интенсивности отраженного излучения в интервалах длин волн 450 нм (б), 554 нм (в) и 986 нм (г).



Рис. 2. Схема конвейера для анализа гиперспектральных изображений.

соответствовал интенсивности отражения в определенном диапазоне длин волн (рис. 1). Размер гиперспектрального изображения: 50×50 пикселей, спектральный диапазон: 450–998 нм, ширина спектрального канала: 4 нм. Изображения сохраняли в формате tiff.

Таким образом, гиперспектральное изображение, полученное камерой Cubert S185, представляет собой гиперкуб, в котором индексы i, j ($i, j = 1, \dots, 50$) соответствуют пространственным координатам (пикселям изображения), индекс $k = 1, \dots, 138$ – линиям гиперспектра с определенной длиной волны. Каждый элемент этого гиперкуба соответствует интенсивности отраженного излучения от объекта съемки для пикселя на изображении с пространственными координатами i, j и спектральной линии с порядковым номером k .

Изображения для исследования пигментного состава зерен ячменя получены в результате нескольких серий съемок в течение нескольких дней.

Описание конвейера. Входными данными для конвейера являются гиперспектральные изображения в формате tiff, описанные в предыдущем разделе, и калибровочные гиперспектральные изображения (изображения черного и белого фонов в формате tiff).

Анализ многоканального гиперспектрального изображения производится в несколько этапов, включающих предобработку, извлечение признаков, нормировку и непосредственно анализ данных (рис. 2).

Предобработка гиперспектральных изображений и извлечение признаков. Характер внешнего освещения может влиять на интенсивность отраженного спектра (Zahavi et al., 2019). Для того чтобы отраженные интенсивности излучения на разных линиях спектра можно было сравнивать для разных условий съемки, мы использовали калибровку изображений по следующей формуле:

$$R_{ijk} = \frac{S_{ijk} - D_{ijk}}{W_{ijk} - D_{ijk}},$$

где S_{ijk} – элемент гиперкуба гиперспектрального изображения ячменя; D_{ijk} – элемент калибровочного изображения черного фона; W_{ijk} – элемент калибровочного изображения белого фона; R_{ijk} – элемент изображения после калибровки.

Калибровочные изображения с помощью порогового преобразования (функции `threshold()` библиотеки OpenCV (Howse, 2013)) преобразуются в трехканальное изображение, аппроксимирующее RGB на основе интенсивностей для длин волн 450 нм (синий цвет), 510 нм (зеленый цвет), 630 нм (красный цвет). Данное изображение преобразуется в изображение в оттенках серого (функция `cvtColor()` библиотеки OpenCV) и бинаризуется, чтобы выделить область чашки Петри с зернами. При необходимости конвейер позволяет использовать собственную реализацию сегментации, но для решения рассматриваемой задачи достаточно сегментации по пороговому значению.

Затем для каждого изображения по значениям пикселей в сегментированной области, занятой зернами, рассчитываются медианы для каждого гиперспектрального канала. Для сглаживания значений медиан применяется фильтр Савицкого–Голая (Savitzky, Golay, 1964). Полученный вектор медиан характеризует гиперспектральные данные для каждого исследованного образца.

Нормировка. Чтобы избавиться от различий, возникающих между сериями съемки, в конвейере были реализованы два способа нормализации изображений. Первый способ – стандартизация (вычитание выборочного среднего и деление на стандартное отклонение) по одинаковым образцам каждого изображения (вектора медиан). Второй способ – стандартизация по одинаковым группам (образцы, содержащие/не содержащие меланин), внутри каждой серии.

Анализ данных. Методы понижения размерности. Для наглядной визуализации образцов в пространстве гиперспектральных данных конвейер использует три метода снижения размерности: PCA (метод главных компонент) (Jolliffe, 2002), ISOMAP (isometric mapping) (Balasubramanian, Schwartz, 2002) и UMAP (uniform manifold approximation and projection) (McInnes et al., 2018). PCA – линейный метод снижения размерности, который сохраняет наибольший процент дисперсии.

ISOMAP, UMAP – нелинейные методы снижения размерности. Метод UMAP строит взвешенный граф, где ребрами соединены лишь ближайшие соседи (число соседей задается как параметр конвейера). Метод ISOMAP вначале строит разреженный граф, где так же, как и в графе для UMAP, ребрами соединены лишь ближайшие соседи (число соседей задается как параметр конвейера). Затем с помощью алгоритма Дейкстры (Cormen et al., 2002) либо алгоритма Флойда–Уоршалла (Cormen et al., 2002) вычисляются расстояния между объектами в разреженном графе для метода ISOMAP. После построения графов и матрицы расстояний для них с помощью методов UMAP и ISOMAP определяется положение образцов в пространстве меньшей размерности (обычно 2 или 3), которые сохраняют расстояния между объектами. Методы понижения размерности были реализованы с помощью библиотек Sklearn (Hao et al., 2019) и Umap (Becht et al., 2019).

Визуализация. После этапов предобработки и извлечения признаков каждый образец (гиперспектральное изображение) представляется в виде вектора, лежащего в пространстве размерности, равной числу каналов гиперспектрального изображения. Элементы вектора описывают интенсивность отраженного излучения для соответствующего канала. После получения координат образцов в пространствах меньшей размерности проводили визуализацию в виде диаграммы рассеяния с помощью функции `plotly.express.scatter` библиотеки Plotly (Stančin et al., 2019).

Кластеризация. В конвейере реализована кластеризация с помощью EM-алгоритма (Dempster et al., 1977). Предполагалось, что каждый образец может принадлежать к каждому кластеру с вероятностью, подчиняющейся модели смеси распределений Гаусса. Параметры распределений находили с помощью метода максимального правдоподобия с использованием EM-алгоритма. Основные

гиперпараметры кластеризации: размерность пространства, в котором происходит кластеризация, метод понижения размерности, способ инициализации весов (случайная инициализация, инициализация методом k -средних). Конвейер возвращает таблицу с информацией о наиболее часто встречающейся группе в каждом кластере и проценте образцов в ней. При реализации кластеризации использовали библиотеку Sklearn.

Статистический анализ. В созданном конвейере для разницы выборочных средних двух групп изображений можно определить доверительный интервал при заданном уровне значимости, который строится на основе центральной предельной теоремы (ЦПТ). По ЦПТ, при достаточном объеме выборки можно предположить, что разница выборочных средних нормально распределена. Для этой случайной величины вычисляются выборочное среднее и выборочная дисперсия и, таким образом, строятся доверительные интервалы произвольного уровня.

Для проверки гипотезы о совпадении распределений двух групп в конвейер были добавлены тесты на основе U -критерия Манна–Уитни (Wilcoxon, 1992) и критерия хи-квадрат (Greenwood, Nikulin, 1996). Статистический анализ реализован с использованием библиотеки SciPy (Nunez-Iglesias et al., 2017).

Классификация. Разработанный конвейер позволяет классифицировать гиперспектральные изображения с помощью таких методов, как логистическая регрессия (Norman, Harry, 2007), гребневая регрессия (Norman, Harry, 2007), случайный лес (Ho, 1995) и градиентный бустинг (Prokhorenkova et al., 2018). Конвейер возвращает таблицы с результатами классификации по таким метрикам, как accuracy, F1, precision и recall, а также матрицы ошибок для каждого классификатора. Первая таблица содержит результаты классификации по макрометрикам, вторая – по микрометрикам. Если в конвейер передать функцию, которая преобразовывает группу в вектор, то конвейер вернет третью таблицу с усредненными результатами бинарной классификации для каждой отдельной компоненты вектора. Классификация реализована с использованием библиотек Sklearn и CatBoost (Hancock, Khoshgofaar, 2020).

Результаты

Изображения образцов для анализа пигментного состава были получены из трех серий съемок. В двух сериях отсутствовали зерна, содержащие меланин. В одной серии присутствовали как зерна с меланином, так и зерна без этого пигмента. Одинаковых образцов в различных сериях съемок не было. Для каждого образца получено два изображения: гиперспектральное и изображение высокого разрешения. Так как во всех сериях съемки присутствовали образцы без пигмента, выполнялась нормализация по образцам зерен с отсутствием пигмента.

График медиан

По полученным медианам были построены графики зависимости интенсивности от длины волны для каждого изображения (рис. 3). Как можно отметить, гиперспектр зерен, содержащих меланин, заметно отличается от гиперспектра зерен без этого пигмента.

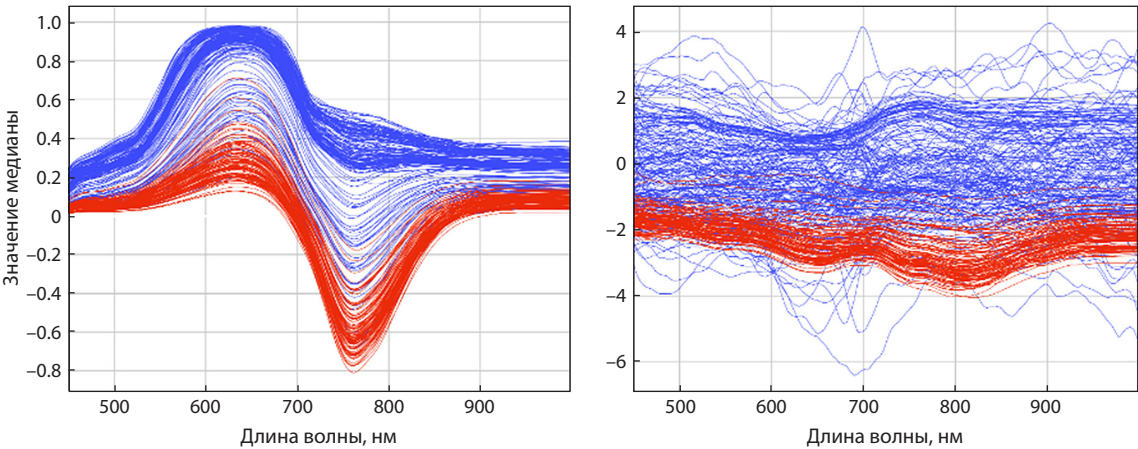


Рис. 3. График зависимости медианы интенсивности отраженного излучения для образцов зерен ячменя в зависимости от длины волны.
Слева – график без нормализации, справа – график медиан после нормировки по одинаковым группам. Синие линии соответствуют медианам изображений зерен ячменя, не содержавших меланина, красные – с меланином.

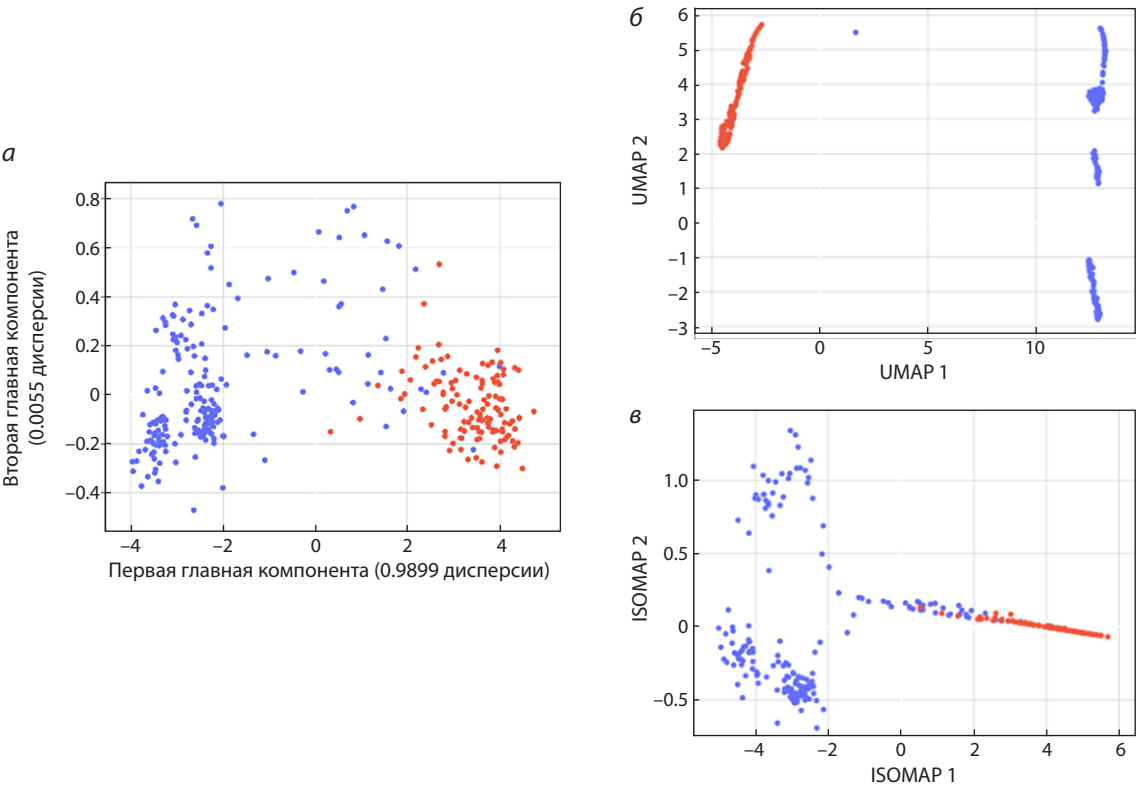


Рис. 4. Графики распределений в двухмерном пространстве, полученные с помощью PCA (а), UMAP (б) и ISOMAP (в).
Синие точки соответствуют образцам без меланина, красные – с меланином.

На графике без нормализации для кривых медиан наблюдаются локальные максимумы в диапазоне 600–700 нм и локальные минимумы в диапазоне 700–800 нм. Большинство кривых медиан зерен с меланином сгруппированы плотнее (средняя по длинам волн дисперсия меньше) и имеют меньшие средние значения, чем кривые образцов без пигмента на всем диапазоне длин волн. Несмотря на частичное перекрытие, большая часть кривых медиан образцов с пигментом отличима от кривых медиан образцов без пигмента.

Визуализация в двухмерном пространстве

На графиках PCA (рис. 4, а) и ISOMAP (см. рис. 4, в) можно заметить, что дисперсия у зерен без меланина больше, чем у зерен с этим пигментом.

Результаты кластеризации

Была выполнена кластеризация на два кластера, которые представляли образцы с меланином и образцы без пигмента. Кластеризация подтверждает, что медианы гиперспектральных изображений разделимы с высокой точностью (рис. 5, табл. 1).

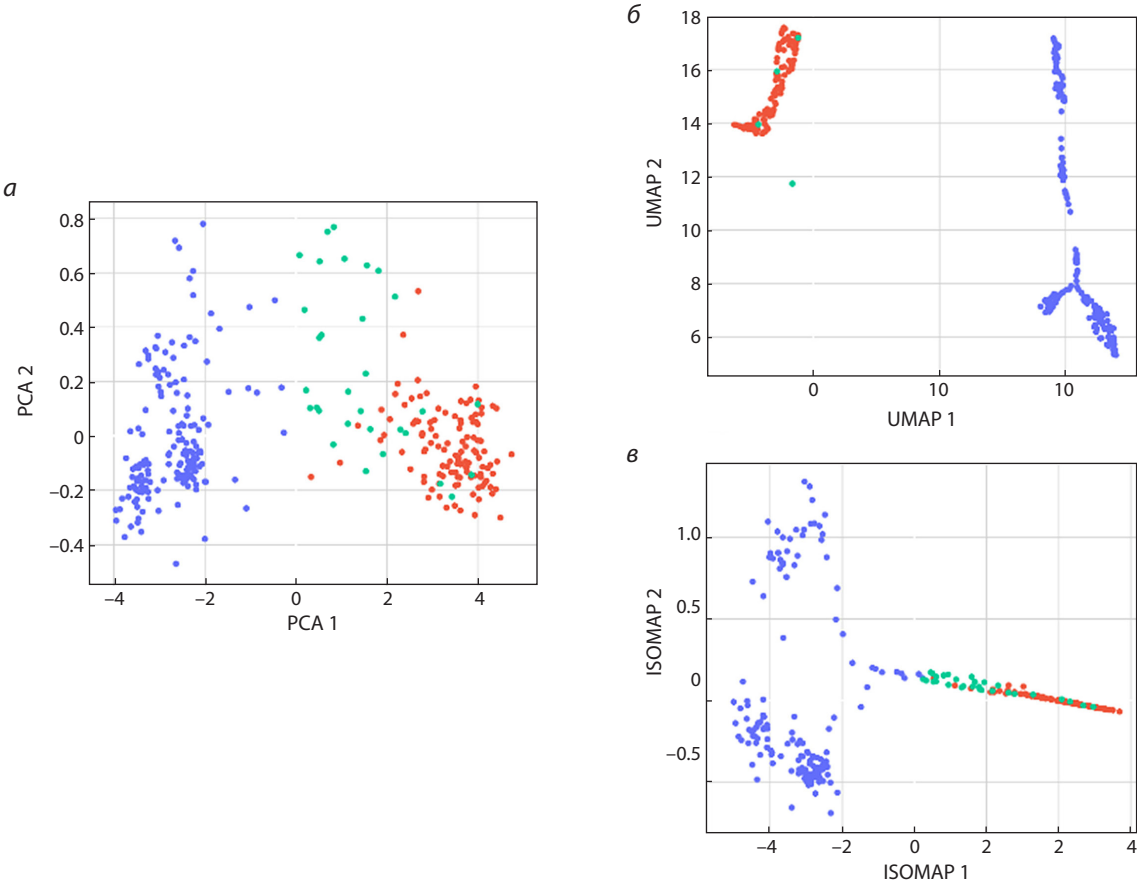


Рис. 5. Визуализация результатов кластеризации с помощью EM-алгоритма. Инициализация выполнена методом *k*-средних в пространстве размерности 15 с использованием методов понижения размерности PCA (а), UMAP (б) и ISOMAP (в). Синие точки соответствуют зернам, которые не содержат меланин и находятся в первом кластере; красные точки – зернам, содержащим меланин и находящимся во втором кластере; зеленые точки – зернам, которые не содержат меланин и принадлежат ко второму кластеру. Образцы, содержащие меланин, но отнесенные к первому кластеру, отсутствовали.

Таблица 1. Точность кластеризации EM-алгоритмом со случайной инициализацией в пространстве размерности 15 с использованием UMAP-метода понижения размерности

Кластер	Превалирующий класс	Метод понижения размерности	Частота наиболее часто встречаемого класса в кластере
2	С меланином	PCA	0.79
1	Без меланина	PCA	1.00
2	С меланином	UMAP	0.96
1	Без меланина	UMAP	1.00
2	С меланином	ISOMAP	0.80
1	Без меланина	ISOMAP	1.00

Примечание. Превалирующий класс – наиболее часто встречаемый класс образцов в кластере.

Наименее явным образом разделились образцы с пигментом и без него на графике PCA (см. рис. 4, а): в районе кластера образцов с меланином (красные точки справа) присутствуют образцы без пигментов (синие точки). Данные образцы при кластеризации (см. рис. 5, а) были отнесены ко второму кластеру (зеленые точки). На графике UMAP, напротив, все образцы с пигментом расположились изолированно (см. рис. 4, б), слева сверху, в то время как образцы без пигментов образовали группы точек справа. Тем не менее на графике кластеризации (см. рис. 5, б) единичные образцы слева были отнесены ко

второму кластеру. График ISOMAP (см. рис. 4, в) демонстрирует хорошую группировку образцов с меланином, в то время как образцы без пигмента распределились в левой и в меньшей степени в правой части графика, частично перекрываясь с образцами с меланином. При кластеризации (см. рис. 5, в) часть этих образцов была отнесена ко второму кластеру (зеленые точки в правой части графика). Данные табл. 1 численно подтверждают то, что векторы медиан гиперспектральных изображений зерен разных классов (содержащих и не содержащих меланин) при

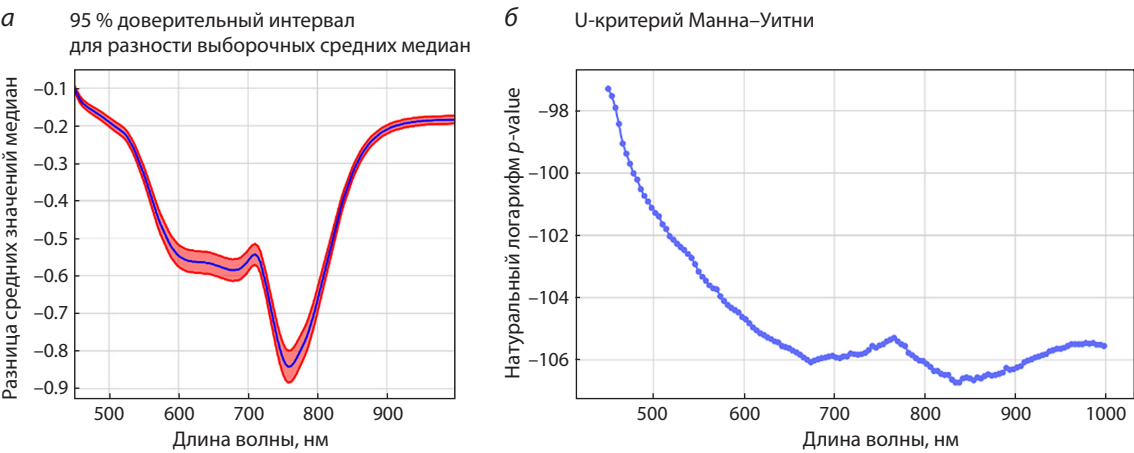


Рис. 6. 95 % доверительный интервал для разности выборочных средних медиан: синяя линия – значения разности выборочных средних, красная область – 95 % доверительный интервал (а); график зависимости логарифма p -value для U-критерия Манна–Уитни при оценке разницы средних значений интенсивности отраженного спектра для образцов зерен с меланином и без него для разных интервалов длин волн (б).

Таблица 2. Результаты классификации на тестовой выборке в пространстве размерности 15 с использованием PCA, UMAP и ISOMAP для понижения размерности

Модель	Accuracy	F1	Precision	Recall
Логистическая регрессия	0.979	0.971	0.944	1.000
Гребневая регрессия	0.979	0.971	0.944	1.000
Случайный лес	0.979	0.971	0.944	1.000
Градиентный бустинг (catboost)	0.979	0.971	0.944	1.000

Примечание. Для полученной тренировочной и тестовой выборки результаты при использовании разных методов понижения размерности на тестовой выборке были одинаковыми.

кластеризации в основном попадают в разные кластеры, что свидетельствует о существовании значимых различий в спектре зерен с пигментом и без него. Нужно отметить, что в первый кластер попали образцы исключительно без меланина.

Статистический анализ

На рис. 6, а приведены разности выборочных средних значений интенсивности отраженного излучения для образцов ячменя для всех интервалов длин волн. Как можно отметить, средние значения у разных групп зерен статистически значимо различаются во всем рассматриваемом интервале длин волн. На рис. 6, б приведен график зависимости логарифма достоверности различий (p -value) от длины волны для U-критерия Манна–Уитни. Данный критерий (с учетом поправки Бонферрони) позволил обнаружить статистически значимые различия для всего изучаемого гиперспектра.

Результаты классификации

Классификация гиперспектральных изображений зерен по содержанию меланина – это задача бинарной классификации. В табл. 2 приведены оценки точности классификации по метрикам accuracy, F1-мера, precision и recall для каждого метода понижения размерности. Размер тестовой выборки – 47 образцов, размер тренировочной выборки – 266 образцов. При обучении использовалась k -fold

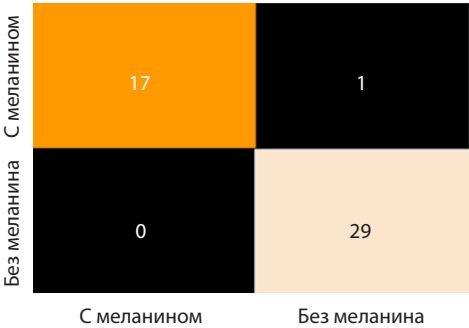


Рис. 7. Матрица ошибок для тестовой выборки в пространстве размерности 15.

кросс-валидация ($k = 4$); 18 образцов в тестовой выборке содержали меланин; 29 были без меланина; 99 в тренировочной выборке были с меланином, 167 образцов – без этого пигмента.

Исследуемые образцы зерен, помимо меланина, содержали антоцианы, что позволило изучить возможность дифференцировки меланинов и антоцианов. Образцы классифицировали в 15-мерном пространстве, полученном ранее с помощью PCA, с использованием логистической регрессии (266 образцов для обучающей выборки и 47 – для тестовой выборки). В результате ошибки классификации возникали в основном между

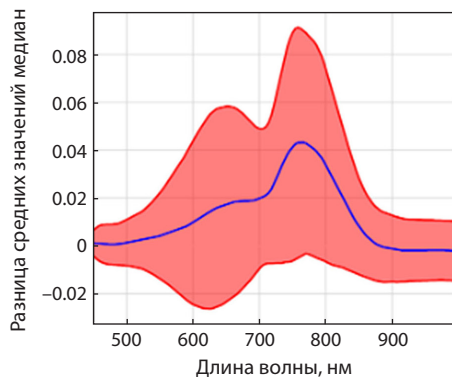
классами «без пигментов» и «только с антоцианами», а также при определении образцов, содержащих оба пигмента, и зерен, содержащих только меланин (рис. 7, 8).

На основе результатов статистического анализа для зерен, содержащих только меланин, и зерен с обоими пигментами не было обнаружено статистически значимых различий ($p\text{-value} < 0.05/138$, с учетом поправки Бонферрони) по всему спектру. Наименьшее значение $p\text{-value}$ для критерия Манна–Уитни для этих групп было достигнуто при длине волны 774 нм и составило 0.0438 (рис. 9, а). Для зерен, содержащих только антоцианы, и зерен без пигментов были обнаружены статистически значимые различия ($p\text{-value} < 0.05/138$ с учетом поправки Бонферрони) в волнах, попадающих в красный и инфракрасный диапазоны (> 714 нм) (см. рис. 9, б).

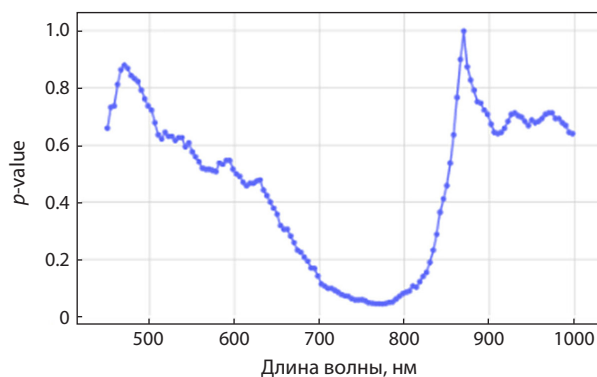
Меланин и антоцианы	0	0	4	0
Только антоцианы	0	13	1	2
Только меланин	0	0	14	0
Без пигментов	0	5	0	8
	Меланин и антоцианы	Только антоцианы	Только меланин	Без пигментов

Рис. 8. Матрица ошибок классификации на основе логистической регрессии образцов зерен на четыре класса: содержащих меланин и антоцианы, только антоцианы, только меланин и без пигментов.

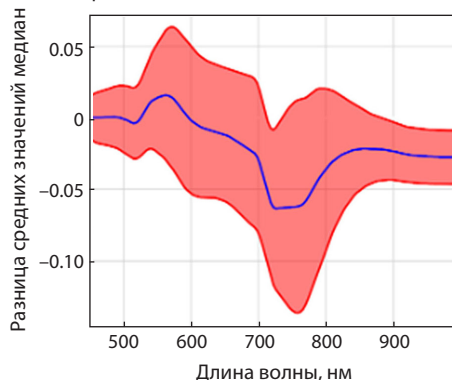
а 95 % доверительный интервал для разности выборочных средних медиан между зернами, содержащими только меланин, и зернами с обоими пигментами



U-критерий Манна–Уитни для зерен, содержащих только меланин, и зерен с обоими пигментами



б 95 % доверительный интервал для разности выборочных средних медиан между зернами, содержащими только антоцианы, и зернами без пигментов



U-критерий Манна–Уитни для зерен, содержащих только антоцианы, и зерен без пигментов

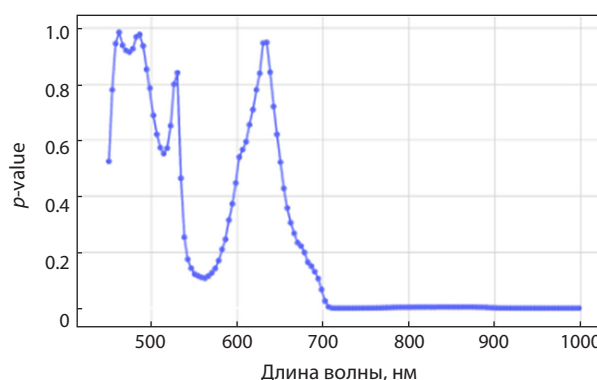


Рис. 9. Графики зависимости логарифма $p\text{-value}$ для U-критерия Манна–Уитни при оценке разности средних значений интенсивности отраженного спектра и 95 % доверительные интервалы для разности выборочных средних медиан.

Обсуждение

Конвейеры в области обработки гиперспектральных изображений

Существует множество современных подходов к автоматизации процесса анализа гиперспектральных данных. Они используют широкий спектр машинного обучения, компьютерного зрения и передовых методов обработки данных. Гиперспектральные изображения характеризуются высокой размерностью, большим объемом данных, подвержены влиянию шума, требуют калибровки и нормализации, в сравнении с RGB-изображениями их сложнее визуализировать. Помимо этого, существует проблема объема обучающих выборок. Для ее решения применяются различные методы увеличения размера обучающих наборов (аугментация). С другой стороны, высокая размерность гиперспектральных данных может легко привести к высокому уровню избыточности данных. Для решения этой проблемы используются алгоритмы ранжирования и фильтрации значимых признаков, а также выбора групп спектров.

Полученные гиперспектральные сырые данные подвергаются предобработке: обнаружению выбросов с помощью анализа главных компонент (PCA), усреднению по группам, масштабированию и центрированию (Yoosefzadeh-Najafabadi et al., 2021), калибровке полученных изображений с использованием эталонных изображений (темного и белого), нормализации, фильтрации Savitzky–Golay, ранжированию и фильтрации параметров для классификации с целью повышения точности и общности моделей (Amanah et al., 2021).

Применение методов понижения размерности может приводить к снижению точности классификации, тем не менее это может быть оправданно в целях повышения общности моделей, чтобы избежать их переобучения. Таким образом, при разработке подходов к решению отдельных задач с использованием гиперспектральных данных требуется многостадийная обработка, реализация которой возможна в архитектуре программных конвейеров, где каждая отдельная стадия заменяема и может быть тщательно настроена и адаптирована.

В настоящее время активно развиваются конвейерные подходы. К примеру, F. Zhu с коллегами (2024) изучали способы предварительной обработки спектральных данных для эффективного снижения влияния различного освещения на оценку хлорофилла в посевах базилика, выращенных при различной интенсивности освещения. Эти исследователи определили оптимальный конвейер анализа данных гиперспектральной визуализации ближнего действия, оценив производительность регрессионного моделирования и получив удовлетворительные карты распределения хлорофилла, согласующиеся с наблюдаемыми различиями в уровнях хлорофилла.

H. Feng с коллегами (2017) создали интегрированный конвейер анализа изображений для автоматической обработки больших объемов гиперспектральных данных. Были построены модели для точного количественного определения четырех пигментов (хлорофилла а, хлорофилла b, общего хлорофилла и каротиноидов) из листьев риса и определены важные группы длин волн (700–760 нм), связанные с этими пигментами. На стадии кущения значения R^2

и средние абсолютные процентные ошибки моделей составляли 0.827–0.928 и 6.94–12.84 % соответственно.

Путем создания четырехэтапного конвейера управления обработкой изображений и анализа данных была оценена применимость гиперспектрального дистанционного зондирования для раннего выявления стресса в результате засухи и заражения галловыми нематодами (root-knot nematodes, RKN) у растений томата (Žibrat et al., 2019). Конвейер включал получение изображений, извлечение данных, предобработку и анализ. Объединив дискриминантный анализ на основе частичных наименьших квадратов и машины опорных векторов с анализом временных рядов, U. Žibrat с коллегами достигли 100 % успеха классификации при определении режима полива и степени заражения. Таким образом, разработка конвейерных решений анализа гиперспектральных данных является активно развивающимся направлением на сегодняшний день.

В представленном в настоящей статье примере анализа гиперспектральных данных также использован конвейерный подход, включающий предобработку и анализ данных с понижением размерности (анализ главных компонент, усреднение по группам, калибровка с использованием эталонных изображений, нормализация, фильтрация Savitzky–Golay). Структура конвейера позволяет применять различные методы понижения размерности: PCA (Jolliffe, 2002), ISOMAP (Balasubramanian, Schwartz, 2002) и UMAP (McInnes et al., 2018) в комбинации с разными методами классификации: логистической регрессией (Norman, Harry, 2007), гребневой регрессией (Norman, Harry, 2007), случайным лесом (Ho, 1995), градиентным бустингом (Prokhorenkova et al., 2018).

Методы классификации изображений растений на основе гиперспектральных данных

Гиперспектральные изображения используются для классификации физиологического состояния растений. T. Zhang с коллегами (2018) исследовали возможность применения методов гиперспектральной визуализации в видимом и ближнем инфракрасном диапазоне (VIS/NIR, 400–1000 нм) для распознавания жизнеспособных и нежизнеспособных семян пшеницы. Для этого они использовали модели классификации, дискриминантный анализ частичных наименьших квадратов (PLS-DA) и машины опорных векторов (SVM) в сочетании с некоторыми методами предварительной обработки и алгоритмом последовательных проекций (SPA). Результаты показали, что стандартная модель нормальной вариации (SNV)-SPA-PLS-DA имела высокую точность классификации для цельных семян (> 85.2 %) и жизнеспособных семян (> 89.5 %).

Y. Lu с коллегами (2022) смогли достичь точности до 99.6 % при дифференциации пяти сортов конопли и 100 % точности при различении пяти стадий роста и двух органов растения (листьев и цветов) с помощью настольной системы гиперспектральной визуализации в спектральном диапазоне 400–1000 нм и машинного обучения, основанного на регуляризованном линейном дискриминантном анализе.

В работе (da Silva et al., 2024) оценена эффективность пяти алгоритмов ML и чувствительность 90 спектров в задаче прогнозирования содержания азота и пигментов (хлорофилла и каротиноидов) в листьях кукурузы на раз-

ных фенологических стадиях для оптимизации азотных удобрений. При прогнозе содержания хлорофиллов а и б значение коэффициента корреляции Пирсона между предсказанными и наблюдаемыми данными составляло около 0.6, а средняя абсолютная ошибка (MAE) была ниже 0.5. При прогнозировании содержания флавоноидов значение коэффициента корреляции между предсказанными и наблюдаемыми данными составляло около 0.6, а MAE – 0.07. При прогнозировании содержания азота значение коэффициентов корреляции были выше 0.35, а MAE – ниже 2.75.

С. Мо с коллегами (2014) разработали метод оценки жизнеспособности семян перца (*Capsicum annuum* L.) на основе гиперспектральной визуализации в диапазоне 400–700 нм, полученных с использованием синего, зеленого, красного и RGB светодиодного освещения. Для этого была создана модель частичного дискриминантного анализа наименьших квадратов (PLS-DA), основанная на стандартном нормальном варианте светодиодного освещения RGB (400–700 нм), которая обеспечивала точность распознавания от 96.7 до 100 %.

R. Falcioni с коллегами (2023) разработали метод оценки пигментов, таких как хлорофиллы, каротиноиды, антоцианы и флавоноиды, в шести агрономических культурах: кукурузе, сахарном тростнике, кофе, рапсе, пшенице и табаке, на базе гиперспектральных данных. Кластеризация, основанная на анализе главных компонент (PCA), и анализ коэффициента Каппа позволили получить точность (ассигасу и precision) от 92 до 100 % в области ультрафиолета (UV-VIS), ближнем инфракрасном (NIR) и коротковолновом инфракрасном (SWIR) диапазонах.

В нашем исследовании мы получили довольно высокие значения точности: ассигасу = 0.979, F1-мера = 0.971 с precision = 0.944 и recall = 1.000 для всех моделей предсказания, что является сравнимым с аналогичными показателями в других работах, в частности, описанных выше. Полученные оценки были одинаковыми для всех моделей, вероятно, в силу того, что выборка имела небольшой размер и была однородной. В результате чего при получившемся разбиении все модели на тестовой выборке совершили одну ошибку, неверно классифицировав один образец. С другой стороны, это свидетельствует о высокой устойчивости предсказаний на основе гиперспектральных данных и предложенных моделей.

В нашей предыдущей статье (Комышев и др., 2023) разработан метод оценки присутствия антоцианов и меланина в оболочке зерна ячменя на основе анализа цифровых RGB изображений с помощью компьютерного зрения и алгоритмов машинного обучения. Мы задействовали схожий протокол получения изображений с использованием чашек Петри для зерен, но съемка производилась обычной RGB камерой. Образцы были взяты из аналогичной коллекции. При этом лучшую точность (ассигасу = 0.821) показала модель сети U-Net, основанная на топологии EfficientNetB0. Таким образом, даже при применении методов глубокого машинного обучения точность классификации оказалась ниже, чем в настоящей работе. Можно заключить, что гиперспектральные изображения позволяют проводить более точную классификацию зерен растений по содержанию пигментов с помощью менее ресурсозатратных методов «мелкого» машинного обучения.

Мы изучили влияние присутствия антоцианов на точность определения меланина в образцах ячменя, содержащих антоцианы. Она оказалась меньше (ассигасу = 0.95) в сравнении с образцами без этого пигмента (ассигасу = 1) (см. рис. 8). Таким образом, присутствие антоцианов незначительно снижает точность определения меланина в образцах.

Возможность дифференцировать образцы только с меланином от образцов с меланином и антоцианами оказалась слабой (см. рис. 9, а). Определение антоцианов на основе полученных гиперспектральных данных, по-видимому, вероятно с высокой точностью благодаря спектру в волнах, попадающих в красный и инфракрасный диапазоны (> 714 нм) (см. рис. 9, б). Таким образом, данный подход дает возможность дифференцировать зерна без пигментов от зерен с антоцианами, но не позволяет определить присутствие антоцианов в образцах с меланином.

Наша цель заключалась в изучении возможности различения образцов семян, содержащих и не содержащих меланин, с помощью только гиперспектральных данных. Мы также опробовали несколько подходов, состоящих из взаимозаменяемых методов, которые формируют типичный конвейер обработки гиперспектральных данных, и сформировали его в виде программного инструмента. Этот программный инструмент может быть использован для быстрого построения алгоритма анализа гиперспектральных данных, включающих основные этапы обработки данных, такие как загрузка изображений, предобработка, анализ и визуализация.

Заключение

Визуализация на основе методов PCA, UMAP и ISOMAP, а также кластеризация в пространстве размерности 15 показали, что на основе гиперспектральных изображений образцы ячменя с меланином и без него можно разделить на два соответствующих класса с высокой точностью. Анализ выявил статистически значимые различия в распределении интенсивности отраженного излучения для этих образцов для всех линий гиперспектра.

Преимущества применения разработанного конвейера перед классическими и более точными биохимическими методами решения задачи классификации: малые временные и трудовые затраты, объективность полученных результатов. В данной версии пакета для классификации не использованы нейронные сети/методы глубокого машинного обучения. Недостатками нейросетевых подходов перед реализованными в конвейере методами могут быть сложная интерпретируемость результатов предсказаний, а также необходимость в обучающей выборке очень большого объема.

В настоящей работе был создан вычислительный конвейер на языке Python с открытым исходным кодом для анализа гиперспектральных изображений, который включает в себя визуализацию в двумерном пространстве, кластеризацию, базовый статистический анализ и классификацию. Предложенный программный пакет может существенно снизить временные затраты в исследованиях, включающих анализ гиперспектральных изображений. Конвейер был протестирован в задаче исследования влияния меланина на гиперспектр зерен ячменя.

Список литературы / References

- Афонников Д.А., Генаев М.А., Дорошков А.В., Комышев Е.Г., Пшеничникова Т.А. Методы высокопроизводительного фенотипирования растений для массовых селекционно-генетических экспериментов. *Генетика*. 2016;52(7):788-803. DOI 10.7868/S001667581607002X
- [Afonnikov D.A., Genaev M.A., Doroshkov A.V., Komyshev E.G., Pshenichnikova T.A. Methods of high-throughput plant phenotyping for large-scale breeding and genetic experiments. *Russ. J. Genet.* 2016;52(7):688-701. DOI 10.1134/S1022795416070024]
- Глаголева А.Ю., Новокрещёнов Л.А., Шоева О.Ю., Ковалева О.Н., Хлесткина Е.К. Изучение разнообразия окраски зерновки ячменя в коллекции ВИР. *Труды по прикл. ботанике, генетике и селекции*. 2022;183(3):76-84. DOI 10.30901/2227-8834-2022-3-76-84
- [Glagoleva A.Y., Novokreschyonov L.A., Shoeva O.Y., Kovalova O.N., Khlestkina E.K. Studying grain color diversity in the barley collection of VIR. *Trudy po Prikladnoy Botanike, Genetike i Selektii = Proceedings on Applied Botany, Genetics, and Breeding*. 2022;183(3):76-84. DOI 10.30901/2227-8834-2022-3-76-84 (in Russian)]
- Комышев Е.Г., Генаев М.А., Бусов И.Д., Кожекин М.В., Артеменко Н.В., Глаголева А.Ю., Коваль В.С., Афонников Д.А. Определение содержания меланина и антоцианов в зернах ячменя на основе анализа цифровых изображений методами машинного обучения. *Вавиловский журнал генетики и селекции*. 2023;27(7):859-868. DOI 10.18699/VJGB-23-99
- [Komyshev E.G., Genaev M.A., Busov I.D., Kozhekin M.V., Artemenko N.V., Glagoleva A.Y., Koval V.S., Afonnikov D.A. Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods. *Vavilovskii Zhurnal Genetiki i Selektii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):859-868. DOI 10.18699/VJGB-23-99 (in Russian)]
- Крупнов В.А. Генетическая сложность и контекст-специфичность признаков урожая пшеницы в засушливых условиях. *Вавиловский журнал генетики и селекции*. 2013;17(3):524-534
- [Krupnov V.A. Genetic complexity and context specificity of traits improving wheat yield under drought conditions. *Vavilovskii Zhurnal Genetiki i Selektii = Vavilov Journal of Genetics and Breeding*. 2013;17(3):524-534 (in Russian)]
- Чешкова А.Ф. Обзор современных методов обнаружения и идентификации болезней растений на основе анализа гиперспектральных изображений. *Вавиловский журнал генетики и селекции*. 2022;26(2):202-213. DOI 10.18699/VJGB-22-25
- [Cheshkova A.F. A review of hyperspectral image analysis techniques for plant disease detection and identification. *Vavilovskii Zhurnal Genetiki i Selektii = Vavilov Journal of Genetics and Breeding*. 2022;26(2):202-213. DOI 10.18699/VJGB-22-25 (in Russian)]
- Шоева О.Ю., Сtryгина К.В., Хлесткина Е.К. Гены, контролирующие синтез флавоноидных и меланиновых пигментов ячменя. *Вавиловский журнал генетики и селекции*. 2018;22(3):333-342. DOI 10.18699/VJ18.369
- [Shoeva O.Yu., Strygina K.V., Khlestkina E.K. Genes determining the synthesis of flavonoid and melanin pigments in barley. *Vavilovskii Zhurnal Genetiki i Selektii = Vavilov Journal of Genetics and Breeding*. 2018;22(3):333-342. DOI 10.18699/VJ18.369 (in Russian)]
- Afonnikov D.A., Komyshev E.G., Efimov V.M., Genaev M.A., Koval V.S., Gierke P.U., Börner A. Relationship between the characteristics of bread wheat grains, storage time and germination. *Plants*. 2021;11(1):35. DOI 10.3390/plants11010035
- Amanah H.Z., Wakholi C., Perez M., Faqeerzada M.A., Tunny S.S., Masithoh R.E., Choung M.G., Kim K.H., Lee W.H., Cho B.K. Near-infrared hyperspectral imaging (NIR-HSI) for nondestructive prediction of anthocyanins content in black rice seeds. *Appl. Sci.* 2021;11(11):4841. DOI 10.3390/app11114841
- Ambrose A., Kandpal L.M., Kim M.S., Lee W.H., Cho B.K. High speed measurement of corn seed viability using hyperspectral imaging. *Infrared Phys. Technol.* 2016;75:173-179. DOI 10.1016/j.infrared.2015.12.008
- Baek I., Kim M.S., Cho B.K., Mo C., Barnaby J.Y., McClung A.M., Oh M. Selection of optimal hyperspectral wavebands for detection of discolored, diseased rice seeds. *Appl. Sci.* 2019;9(5):1027. DOI 10.3390/app9051027
- Balasubramanian M., Schwartz E.L. The isomap algorithm and topological stability. *Science*. 2002;295(5552):7. DOI 10.1126/science.295.5552.7a
- Becht E., McInnes L., Healy J., Dutertre C.A., Kwok I.W., Ng L.G., Ginhoux F., Newell E.W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 2019;37(1):38-44. DOI 10.1038/nbt.4314
- Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. Introduction to Algorithms. Cambridge, Massachusetts: The MIT Press, 2022
- da Silva B.C., de Mello Prado R., Baio F.H.R., Campos C.N.S., Teodoro L.P.R., Teodoro P.E., Santana D.C., Fernandes T.F.S., da Silva J.C.A., de Souza Loureiro E. New approach for predicting nitrogen and pigments in maize from hyperspectral data and machine learning models. *Remote Sens. Appl. Soc. Environ.* 2024;33:101110. DOI 10.1016/j.rsase.2023.101110
- da Silva Medeiros M.L., Cruz-Tirado J.P., Lima A.F., de Souza Netto J.M., Ribeiro A.P.B., Bassegio D., Godoy H.T., Barbin D.F. Assessment oil composition and species discrimination of *Brassica* seeds based on hyperspectral imaging and portable near infrared (NIR) spectroscopy tools and chemometrics. *J. Food Compos. Anal.* 2022;107:104403. DOI 10.1016/j.jfca.2022.104403
- Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*. 1977;39(1):1-22. DOI 10.1111/j.2517-6161.1977.tb01600.x
- Díaz-Martínez V., Orozco-Sandoval J., Manian V., Dhath B.K., Walia H. A deep learning framework for processing and classification of hyperspectral rice seed images grown under high day and night temperatures. *Sensors*. 2023;23(9):4370. DOI 10.3390/s23094370
- ElMasry G., Mandour N., Ejeez Y., Demilly D., Al-Rejaie S., Verdier J., Belin E., Rousseau D. Multichannel imaging for monitoring chemical composition and germination capacity of cowpea (*Vigna unguiculata*) seeds during development and maturation. *Crop J.* 2022;10(5):1399-1411. DOI 10.1016/j.cj.2021.04.010
- Fakhthongphan J., Graybosch R.A., Baenziger P.S. Combining ability for tolerance to pre-harvest sprouting in common wheat (*Triticum aestivum* L.). *Crop Sci.* 2016;56(3):1025-1035. DOI 10.2135/cropsci2015.08.0490
- Falcioni R., Antunes W.C., Demattê J.A.M., Nanni M.R. Reflectance spectroscopy for the classification and prediction of pigments in agronomic crops. *Plants*. 2023;12(12):2347. DOI 10.3390/plants12122347
- Feng H., Chen G., Xiong L., Liu Q., Yang W. Accurate digitization of the chlorophyll distribution of individual rice leaves using hyperspectral imaging and an integrated image analysis pipeline. *Front. Plant Sci.* 2017;8:1238. DOI 10.3389/fpls.2017.01238
- Flintham J., Adlam R., Bassoi M., Holdsworth M., Gale M. Mapping genes for resistance to sprouting damage in wheat. *Euphytica*. 2002;126:39-45. DOI 10.1023/A:1019632008244
- Gao T., Chandran A.K.N., Paul P., Walia H., Yu H. HyperSeed: an end-to-end method to process hyperspectral images of seeds. *Sensors*. 2021;21(24):8184. DOI 10.3390/s21248184
- Glagoleva A.Y., Shmakov N.A., Shoeva O.Y., Vasiliev G.V., Shatskaya N.V., Börner A., Afonnikov D.A., Khlestkina E.K. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the *Black lemma* and *pericarp (Blp)* gene. *BMC Plant Biol.* 2017;17(Suppl. 1):182. DOI 10.1186/s12870-017-1124-1
- Gowen A.A., O'Donnell C.P., Cullen P.J., Downey G., Frias J.M. Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends Food Sci. Technol.* 2007;18(12):590-598. DOI 10.1016/j.tifs.2007.06.001
- Greenwood P.E., Nikulin M.S. A Guide to Chi-Squared Testing. New York: Wiley, 1996;196-202

- Hancock J.T., Khoshgoftaar T.M. CatBoost for big data: an interdisciplinary review. *J. Big Data*. 2020;7(1):94. DOI 10.1186/s40537-020-00369-8
- Hao J., Ho T.K. Machine learning made easy: a review of *Scikit-learn* package in python programming language. *J. Educ. Behav. Stat.* 2019;44(3):348-361. DOI 10.3102/1076998619832248
- He X., Feng X., Sun D., Liu F., Bao Y., He Y. Rapid and nondestructive measurement of rice seed vitality of different years using near-infrared hyperspectral imaging. *Molecules*. 2019;24(12):2227. DOI 10.3390/molecules24122227
- Ho T.K. Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition. 1995;1:278-282. DOI 10.1109/ICDAR.1995.598994
- Howse J. OpenCV Computer Vision with Python. Birmingham: Packt Publishing, 2013
- Jin B., Qi H., Jia L., Tang Q., Gao L., Li Z., Zhao G. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. *Infrared Phys. Technol.* 2022;122:104097. DOI 10.1016/j.infrared.2022.104097
- Jolliffe I.T. Principal component analysis for special types of data. In: Principal Component Analysis. Springer Series in Statistics. New York, NY: Springer, 2002;338-372. DOI 10.1007/0-387-22440-8_13
- Kandpal L.M., Lohumi S., Kim M.S., Kang J.S., Cho B.K. Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds. *Sens. Actuators B*. 2016;229:534-544. DOI 10.1016/j.snb.2016.02.015
- Khlestkina E.K. Current applications of wheat and wheat-alien precise genetic stocks. *Mol. Breed.* 2014;34(2):273-281. DOI 10.1007/s11032-014-0049-8
- Lachman J., Martinek P., Kotíková Z., Orsák M., Šulc M. Genetics and chemistry of pigments in wheat grain. A review. *J. Cereal Sci.* 2017;74:145-154. DOI 10.1016/j.jcs.2017.02.007
- Liu C., Huang W., Yang G., Wang Q., Li J., Chen L. Determination of starch content in single kernel using near-infrared hyperspectral images from two sides of corn seeds. *Infrared Phys. Technol.* 2020;110:103462. DOI 10.1016/j.infrared.2020.103462
- Lu Y., Young S., Linder E., Whipker B., Suchoff D. Hyperspectral imaging with machine learning to differentiate cultivars, growth stages, flowers, and leaves of industrial hemp (*Cannabis sativa* L.). *Front. Plant Sci.* 2022;12:810113. DOI 10.3389/fpls.2021.810113
- Ma T., Tsuchikawa S., Inagaki T. Rapid and non-destructive seed viability prediction using near-infrared hyperspectral imaging coupled with a deep learning approach. *Comput. Electron. Agric.* 2020;177:105683. DOI 10.1016/j.compag.2020.105683
- Machálková L., Janečková M., Hřivná L., Dostálová Y., Hernandez K., Joany L., Mrkvicová E., Vyhnaněk T., Trojan V. Impact of added colored wheat bran on bread quality. *Acta Univ. Agric. Silv. Mendeliana Brun.* 2017;65(1):99-104. DOI 10.1118/actaun201765010099
- Matzrafi M., Herrmann I., Nansen C., Kliper T., Zait Y., Ignat T., Siso D., Rubin B., Karnieli A., Eizenberg H. Hyperspectral technologies for assessing seed germination and trifloxysulfuron-methyl response in *Amaranthus palmeri* (Palmer amaranth). *Front. Plant Sci.* 2017;8:474. DOI 10.3389/fpls.2017.00474
- McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*. 2018;1802.03426. DOI 10.48550/arXiv.1802.03426
- Mo C., Kim G., Lee K., Kim M.S., Cho B.K., Lim J., Kang S. Non-destructive quality evaluation of pepper (*Capsicum annuum* L.) seeds using LED-induced hyperspectral reflectance imaging. *Sensors*. 2014;14(4):7489-7504. DOI 10.3390/s140407489
- Norman R.D., Harry S. Applied Regression Analysis. Williams, 2007
- Nunez-Iglesias J., Van der Walt S., Dashnow H. Elegant SciPy: The Art of Scientific Python. Sebastopol, CA: O'Reilly Media, 2017
- Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulina A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018; 6639-6649
- Qin J., Chao K., Kim M.S., Lu R., Burks T.F. Hyperspectral and multi-spectral imaging for evaluating food safety and quality. *J. Food Eng.* 2013;118(2):157-171. DOI 10.1016/j.jfoodeng.2013.04.001
- Reddy P., Panozzo J., Guthridge K.M., Spangenberg G.C., Rochfort S.J. Single seed near-infrared hyperspectral imaging for classification of perennial ryegrass seed. *Sensors*. 2023;23(4):1820. DOI 10.3390/s23041820
- Savitzky A., Golay M.J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 1964;36(8):1627-1639. DOI 10.1021/ac60214a047
- Serrano L., Filella L., Penuelas J. Remote sensing of biomass and yield of winter wheat under different nitrogen supplies. *Crop Sci.* 2000;40(3):723-731. DOI 10.2135/cropsci2000.403723x
- Stančin I., Jović A. An overview and comparison of free Python libraries for data mining and big data analysis. In: 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2019;977-982. DOI 10.23919/MIPRO.2019.8757088
- Wakholi C., Kandpal L.M., Lee H., Bae H., Park E., Kim M.S., Mo C., Lee W.H., Cho B.K. Rapid assessment of corn seed viability using short wave infrared line-scan hyperspectral imaging and chemometrics. *Sens. Actuators B*. 2018;255:498-507. DOI 10.1016/j.snb.2017.08.036
- Weber V.S., Araus J.L., Cairns J.E., Sanchez C., Melchinger A.E., Orsini E. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. *Field Crops Res.* 2012;128:82-90. DOI 10.1016/j.fcr.2011.12.016
- Wilcoxon F. Individual comparisons by ranking methods. In: Kotz S., Johnson N.L. (Eds.). Breakthroughs in Statistics. Springer Series in Statistics. New York, NY: Springer, 1992;196-202. DOI 10.1007/978-1-4612-4380-9_16
- Yang G., Wang Q., Liu C., Wang X., Fan S., Huang W. Rapid and visual detection of the main chemical compositions in maize seeds based on Raman hyperspectral imaging. *Spectrochim. Acta A. Mol. Biomol. Spectrosc.* 2018;200:186-194. DOI 10.1016/j.saa.2018.04.026
- Yoosefzadeh-Najafabadi M., Earl H.J., Tulpan D., Sulik J., Eskandari M. Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* 2021;11:624273. DOI 10.3389/fpls.2020.624273
- Zahavi A., Palshin A., Liyanage D.C., Tamre M. Influence of illumination sources on hyperspectral imaging. In: 20th International Conference on Research and Education in Mechatronics (REM). Wels, Austria, 2019;1-5. DOI 10.1109/REM.2019.8744086
- Zhang X., He Y. Rapid estimation of seed yield using hyperspectral images of oilseed rape leaves. *Ind. Crops Prod.* 2013;42:416-420. DOI 10.1016/j.indcrop.2012.06.021
- Zhang T., Wei W., Zhao B., Wang R., Li M., Yang L., Wang J., Sun Q. A reliable methodology for determining seed viability by using hyperspectral data from two sides of wheat seeds. *Sensors*. 2018;18(3):813. DOI 10.3390/s18030813
- Zhu F., Qiao X., Zhang Y., Jiang J. Analysis and mitigation of illumination influences on canopy close-range hyperspectral imaging for the *in situ* detection of chlorophyll distribution of basil crops. *Comput. Electron. Agric.* 2024;217:108553. DOI 10.1016/j.compag.2023.108553
- Žibrat U., Susič N., Knapič M., Širca S., Strajnar P., Razinger J., Vončina A., Urek G., Stare B.G. Pipeline for imaging, extraction, pre-processing, and processing of time-series hyperspectral data for discriminating drought stress origin in tomatoes. *MethodsX*. 2019;6:399-408. DOI 10.1016/j.mex.2019.02.022
- Zou Z., Chen J., Wu W., Luo J., Long T., Wu Q., Wang Q., Zhen J., Zhao Y., Wang Y., Chen Y., Zhou M., Xu L. Detection of peanut seed vigor based on hyperspectral imaging and chemometrics. *Front. Plant Sci.* 2023;14:1127108. DOI 10.3389/fpls.2023.1127108

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 23.07.2023. После доработки 22.03.2024. Принята к публикации 26.03.2024.