ОПТИКА И СПЕКТРОСКОПИЯ. ЛАЗЕРНАЯ ФИЗИКА

Известия Саратовского университета. Новая серия. Серия: Физика. 2025. Т. 25, вып. 3. С. 305–315
Izvestiya of Saratov University. Physics, 2025, vol. 25, iss. 3. P. 305–315
https://fizika.sgu.ru https://doi.org/10.18500/1817-3020-2025-25-3-305-315, EDN: KWEXHY

Научная статья
УДК 519.688:543.424.2

Влияние малых концентраций гиалуроновой кислоты на структуру изолята сывороточного протеина при конъюгировании: разработка и оптимизация моделей машинного обучения на основе адаптивного бустинга для анализа спектроскопических данных

С. А. Шевцова, М. С. Савельева, О. А. Майорова, Е. С. Прихожденко✉

Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Россия, 410012, г. Саратов, ул. Астраханская, д. 83

Шевцова Светлана Александровна, институт физики, студент, Научный медицинский центр, лаборант лаборатории «Дистанционно управляемые системы для тераностики», sveta.shevtsova.0404@mail.ru, https://orcid.org/0009-0002-2533-4827
Савельева Мария Сергеевна, кандидат физико-математических наук, институт физики, ассистент кафедры материаловедения, технологии и управления качеством, Научный медицинский центр, младший научный сотрудник лаборатории «Дистанционно управляемые системы для тераностики», mssaveleva@yandex.ru, https://orcid.org/0000-0003-2021-0462, AuthorID: 938218
Майорова Оксана Александровна, кандидат химических наук, Научный медицинский центр, старший научный сотрудник лаборатории биомедицинской фотоакустики, oksanaamayorova@gmail.com, https://orcid.org/0000-0002-6440-3947, AuthorID: 1001358
Прихожденко Екатерина Сергеевна, кандидат физико-математических наук, институт физики, доцент кафедры инноватики, prikhozhdenkoes@gmail.com, https://orcid.org/0000-0003-2700-168X, AuthorID: 850345

Аннотация. В данном исследовании с помощью спектроскопии комбинационного рассеяния (КР) было изучено влияние малых концентраций гиалуроновой кислоты (ГК, 0.1–0.5%) на структуру изолята сывороточного протеина (ИСП) при конъюгировании. Анализ спектров КР выявил, что основное изменение происходит в области 1003 см$^{-1}$, соответствующей колебаниям фенилаланина. Для классификации и регрессионного анализа спектральных данных использовались ансамблевые методы машинного обучения, включая адаптивный бустинг (AdaBoost). Оптимальные параметры модели (глубина дерева принятия решений max_depth=3, количество деревьев в ансамбле 325) обеспечили высокую точность классификации (98.3%) и коэффициент детерминации (R^2 = 0.91) при объеме обучающей выборки 300 спектров на образец. Подбор параметров проводился с помощью решетчатого поиска (GridSearchCV). Было также изучено влияние объема обучающей выборки на эффективность модели адаптивного бустинга. Модель также позволила выявить ключевые волновые числа (763, 1003, 1240, 1400 см$^{-1}$), наиболее значимые для прогнозирования изменений в структуре ИСП при добавлении ГК. Результаты демонстрируют перспективность комбинации спектроскопии КР и машинного обучения для анализа белково-полисахаридных взаимодействий.

НАУЧНЫЙ ОТДЕЛ

Ключевые слова: гиалуроновая кислота, изолят сывороточного протеина, спектроскопия КР, адаптивный бустинг, машинное обучение, решётчатый поиск, классификация, регрессия

Благодарности: Работа выполнена при финансовой поддержке Российского научного фонда (проект № 22-79-10270, https://rscf.ru/project/22-79-10270/).

Для цитирования: *Шевцова С. А., Савельева М. С., Майорова О. А., Прихожденко Е. С.* Влияние малых концентраций гиалуроновой кислоты на структуру изолята сывороточного протеина при конъюгировании: разработка и оптимизация моделей машинного обучения на основе адаптивного бустинга для анализа спектроскопических данных // Известия Саратовского университета. Новая серия. Серия: Физика. 2025. Т. 25, вып. 3. С. 305–315. https://doi.org/10.18500/1817-3020-2025-25-3-305-315, EDN: KWEXHY

Article

Effect of low concentrations of hyaluronic acid on the structure of whey protein isolate during conjugation: Development and optimization of machine learning models based on adaptive boosting for spectroscopic data analysis

S. A. Shevtsova, M. S. Saveleva, O. A. Mayorova, E. S. Prikhozhdenko✉

Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia

Svetlana A. Shevtsova, sveta.shevtsova.0404@mail.ru, https://orcid.org/0009-0002-2533-4827
Mariia S. Saveleva, mssaveleva@yandex.ru, https://orcid.org/0000-0003-2021-0462
Oksana A. Mayorova, oksanaamayorova@gmail.com, https://orcid.org/0000-0002-6440-3947
Ekaterina S. Prikhozhdenko, prikhozhdenkoes@gmail.com, https://orcid.org/0000-0003-2700-168X

Abstract. ***Background and Objectives****:* Multicomponent mixtures with bioactive compounds, such as hyaluronic acid (HA) in protein matrices, are critical in pharmaceuticals, nutraceuticals, and cosmetics. However, detecting low-concentration additives (e.g., 0.1–0.5 wt.% HA in whey protein isolate, WPI) remains challenging due to signal interference and matrix complexity. Raman spectroscopy (RS) is a powerful tool for such analyses, but interpreting spectral data requires advanced computational methods. This study leverages adaptive boosting (AdaBoost), an ensemble ML algorithm, to (1) classify WPI-HA mixtures by HA concentration, (2) quantify HA content via regression, and (3) determine the minimal training dataset size needed for robust predictions. ***Materials and Methods:*** WPI (5 wt.%) was mixed with HA (0.1, 0.25, 0.5 wt.%) in saline, dialyzed, and dried into thin films. Renishaw inVia spectrometer equipped with a 532 nm laser was implemented to collect 600 spectra/sample (20×30-point maps). Preprocessing included cosmic-ray removal, baseline correction, and L_2 normalization. AdaBoost models (scikit-learn) were optimized via GridSearchCV (hyperparameters: DecisionTree max_depth, 1–3; n_estimators, 50–350). Performance was tested across training set sizes (50–500 spectra/sample). Metrics included accuracy (classification) and R^2/RMSE (regression). ***Results:*** Optimization: 325 DecisionTrees with max_depth = 3 have been found to be the best hyperparameters of AdaBoost. Classification: 50 spectra/sample have achieved 94.5% accuracy; 200/300 spectra have improved this to 97.9%/98.3%, respectively. The models have reliably distinguished WPI + 0.1% HA from WPI (>96% accuracy). Regression: 300 spectra/sample have yielded optimal results (R^2 = 0.910, RMSE = 0.061%). Larger datasets (400–500 spectra) have reduced performance (R^2 = 0.894), suggesting overfitting. Key bands for analysis: 763 cm^{-1} (tryptophan), 1003 cm^{-1} (phenylalanine), and 1240 cm^{-1} (amide III). Bands at 1450–1667 cm^{-1} (C–H/amide I/II) have shown negligible importance, indicating minimal HA-induced changes. ***Conclusion:*** AdaBoost models efficiently analyze trace HA in WPI with small training datasets (200 spectra for classification, 300 for regression). The method precision and speed make it ideal for industrial applications, while identified spectral markers have deepen understanding of HA-protein interactions. Future work could extend this framework to other multicomponent systems with low analyte concentrations.

Keywords: hyaluronic acid, whey protein isolate, Raman spectroscopy, adaptive boosting, machine learning, GridSearchCV, classification, regression

Acknowledgements: This work was supported by the Russian Science Foundation (project No. 22-79-10270. https://rscf.ru/project/22-79-10270/).

For citation: Shevtsova S. A., Saveleva M. S., Mayorova O. A., Prikhozhdenko E. S. Effect of low concentrations of hyaluronic acid on the structure of whey protein isolate during conjugation: Development and optimization of machine learning models based on adaptive boosting for spectroscopic data analysis. *Izvestiya of Saratov University. Physics*, 2025, vol. 25, iss. 3, pp. 305–315 (in Russian). https://doi.org/10.18500/1817-3020-2025-25-3-305-315, EDN: KWEXHY

Введение

Исследование многокомпонентных смесей, включающих биологически активные вещества, представляет значительный интерес для различных областей науки и промышленности, таких как фармакология, пищевая индустрия, косметология и биотехнологии [1–3]. Эти смеси часто содержат малые концентрации дополнительных компонентов, которые могут оказывать существенное влияние на их функциональные свойства. Однако анализ таких систем сопряжен с рядом трудностей, связанных с низкой концентрацией целевых компонентов, сложностью матрицы и необходимостью использования

высокочувствительных методов исследования. В связи с этим разработка новых подходов для обнаружения и количественного анализа малых добавок в сложных смесях остается актуальной задачей.

Спектроскопия комбинационного рассеяния (КР) является одним из наиболее перспективных методов для изучения химического состава и структуры многокомпонентных систем [4]. Этот метод обладает рядом преимуществ, включая неразрушающий характер анализа, высокую чувствительность к изменениям в химической структуре и возможность работы с минимальной пробоподготовкой [5–7]. Спектроскопия КР успешно применяется для анализа биологических молекул, полимеров, фармацевтических препаратов и пищевых продуктов [8–11]. Однако интерпретация спектроскопических данных, особенно в случае сложных смесей, требует применения современных методов обработки и анализа, к их числу относится машинное обучение.

В последние годы методы машинного обучения активно внедряются в спектроскопические исследования для решения задач классификации, регрессии и прогнозирования [12–14]. Одним из наиболее эффективных подходов является адаптивный бустинг (AdaBoost), который позволяет комбинировать слабые (по надежности) модели (например, деревья принятия решений) в сильные ансамбли, значительно повышая точность и устойчивость прогнозов [15, 16]. Адаптивный бустинг успешно применяется для анализа спектроскопических данных, включая задачи идентификации компонентов и количественного определения их концентраций [17, 18]. Однако эффективность таких моделей во многом зависит от правильного выбора гиперпараметров, что требует использования методов оптимизации, таких как GridSearchCV [19–21].

Важным аспектом при построении моделей машинного обучения является объем и качество обучающей выборки. Исследования показывают, что количество образцов в обучающей выборке может существенно влиять на производительность модели, особенно в случае глубокого обучения [22–24]. Поэтому изучение зависимости точности модели от размера выборки является важным этапом в разработке надежных аналитических методов.

Внесение примесей, в особенности в малых концентрациях, в белковый смеси может нести как положительное, так и отрицательное влияние. К последней категории, в частности, относятся такие примеси, как например меламин [25, 26] и мочевина [27], которые добавляют в молоко, детские смеси и корма для искусственного завышения показателей «псевдобелка» [28, 29]. Также практикуется «аминокислотный спайкинг» – добавление дешевых аминокислот для манипуляции результатами анализов продуктов питания на содержание белков [30]. Разработка новых подходов к анализу и обнаружению примесей необходима для более точного выявления опасных добавок.

Добавление примесей к белкам может привносить и дополнительные положительные свойства в итоговый материал. В данной работе исследуются смеси изолята сывороточного протеина (ИСП, 5 мас. %) с добавлением гиалуроновой кислоты в различных концентрациях (0, 0.1, 0.25 и 0.5 мас. %). Гиалуроновая кислота (ГК) является важным биополимером, широко используется в медицине и косметологии благодаря своим уникальным свойствам, таким как увлажнение и регенерация тканей [31–34]. Добавление небольшого количества гиалуроновой кислоты в изолят сывороточного белка может значительно улучшить его свойства без сильного увеличения стоимости. Это свойство особенно ценно при создании систем адресной доставки лекарств. Применение комплекса ИСП-ГК в качестве стабилизирующего агента вместо ИСП может значительно увеличить срок службы микроносителей, которые производятся с его использованием [35]. Комбинация ИСП и ГК в различных соотношениях, а также характеристики получаемого микрогеля и наночастиц ранее были исследованы Weigang Zhong и соавторами [36–38]. В этих работах наименьшее исследуемое соотношение ИСП : ГК составило 10 : 1.

Целями работы ставились изучение влияния малых концентраций гиалуроновой кислоты (ГК, 0.1–0.5%) на структуру изолята сывороточного протеина (ИСП) при конъюгировании посредством спектроскопии комбинационного рассеяния, а также разработка и оптимизация моделей машинного обучения на основе адаптивного бустинга для анализа спектроскопических данных, а также изучение влияния размера обучающей выборки на производительность моделей. Полученные результаты могут быть полезны для разработки новых методов контроля качества и анализа сложных многокомпонентных систем в различных отраслях промышленности.

1. Материалы и методы

1.1. Материалы

Изолят сывороточного протеина (ИСП; whey protein isolate, WPI) произведен компанией California Gold Nutrition® (Ирвайн, Калифорния, США). Натриевая соль гиалуроновой кислоты (ГК; англ. – hyaluronic acid, HA; чистота 99%, молярная масса M_w = 404 г/моль) приобретена у компании Macklin Biochemical Co., Ltd. (Шанхай, Китай). Хлорид натрия (NaCl, чистота >99%) произведен компанией Merck (Дармштадт, Германия). В качестве водной среды во всех сериях экспериментов использовали воду, очищенную системой Milli-Q (Merck Millipore, Германия) (18.2 МОм·см$^{-1}$).

1.2. Приготовление конъюгатов ИСП-ГК (WPI-HA)

В процессе создания конъюгатов ИСП-ГК (WPI-HA) использовалась одинаковая концентрация ИСП (WPI) во всех случаях – 10% по массе. Образцы ИСП-ГК были получены путем смешивания растворов ИСП и ГК в физиологическом растворе (с массовой долей 0.15 NaCl). Для получения конъюгатов к растворам ИСП добавляли равные объемы раствора ГК с определенной концентрацией. После этого смесь энергично перемешивали в течение 30 минут при температуре 22°C. Полученные комплексы ИСП-ГК содержали различные концентрации ГК: 0.1%, 0.25% и 0.5% по массе. Для удаления несвязанной ГК конъюгаты ИСП-ГК промывали с помощью диализа в физиологическом растворе в течение 3 дней при температуре 4°C. В качестве контроля был приготовлен образец ИСП путем двукратного разбавления исходного раствора ИСП физиологическим раствором. Таким образом, было получено четыре образца: ИСП, ИСП + 0.1% ГК, ИСП + + 0.25% ГК и ИСП + 0.5% ГК.

1.3. Спектроскопия КР

На кварцевую подложку было нанесено по 10 мкл ИСП и конъюгатов ИСП-ГК, после чего образцы были высушены на воздухе. После высыхания образцов на подложке образовались тонкие пленки. Для сбора спектров комбинационного рассеяния (КР) света высушенных образцов ИСП и конъюгатов ИСП-ГК использовался конфокальный спектрометр Renishaw inVia (Renishaw, Уоттон-андер-Эдж, Великобритания), оснащенный лазером с длиной волны 532 нм. Все измерения проводились с использованием объектива 50×/0.5 N. A. при мощности лазера 2.5 мВт. Для каждого образца были получены карты КР (600 отдельных спектров, 20×30 точек с шагом 2 мкм, регистрация единичного спектра занимала 5 с).

1.4. Анализ данных

В процессе работы были использованы данные, полученные с помощью программы Renishaw WiRE v.4.2 (Renishaw, Уоттон-андер-Эдж, Великобритания). При необходимости к этим данным применялся инструмент для удаления космических лучей (Cosmic Ray Removal) из Renishaw WiRE. Для удаления полиномиального фона из собранных карт КР использовался инструмент Subtract Baseline. В качестве функции для удаления фона был выбран полином десятой степени.

Последующая обработка данных проводилась с использованием Python 3.6 в среде Jupyter Notebook. Загрузка данных спектроскопии КР производилась с помощью библиотеки renishawWiRE. Имплементация моделей машинного обучения, предобработка данных осуществлялись с помощью библиотеки scikit learn [39]. Спектры были нормированы с помощью L_2-нормы, реализованной с использованием sklearn.preprocessing.normalize. Деление данных на обучающую и проверочную выборки осуществлялось с помощью train_test_split из sklearn.model_selection. В качестве моделей классификации и регрессии использовались модели адаптивного бустинга AdaBoostClassifier и AdaBoostRegressor из sklearn.ensemble соответственно. Для изначального подбора параметров моделей использовали решетчатый поиск (sklearn.model_selection.GridSearchCV), оптимизировали параметр max_depth (диапазон целочисленных значений 1–3) единичной модели DecisionTreeClassifier из sklearn.tree и количество единичных моделей (диапазон значений 50–350 с шагом 25) в AdaBoostClassifier. Объем обучающей выборки при оптимизации был равен 500 спектров КР на тип образца: ИСП, ИСП + 0.1% ГК, ИСП + 0.25% ГК и ИСП + 0.5% ГК.

Затем с использованием подобранных параметров производилось обучение моделей классификации и регрессии на основе адаптивного бустинга при разном объеме обучающей выборки: 50, 100, 200, 300, 400 и 500 спектров КР на тип образца. В качестве метрик использовали матрицу неточностей (confusion_matrix), точность прогнозов (accuracy_score) для моделей классификации;

коэффициент детерминации (r2_score), среднеквадратичную ошибку (root_mean_squared_error) из sklearn.metrics. Все рисунки были получены с помощью библиотеки matplotlib.

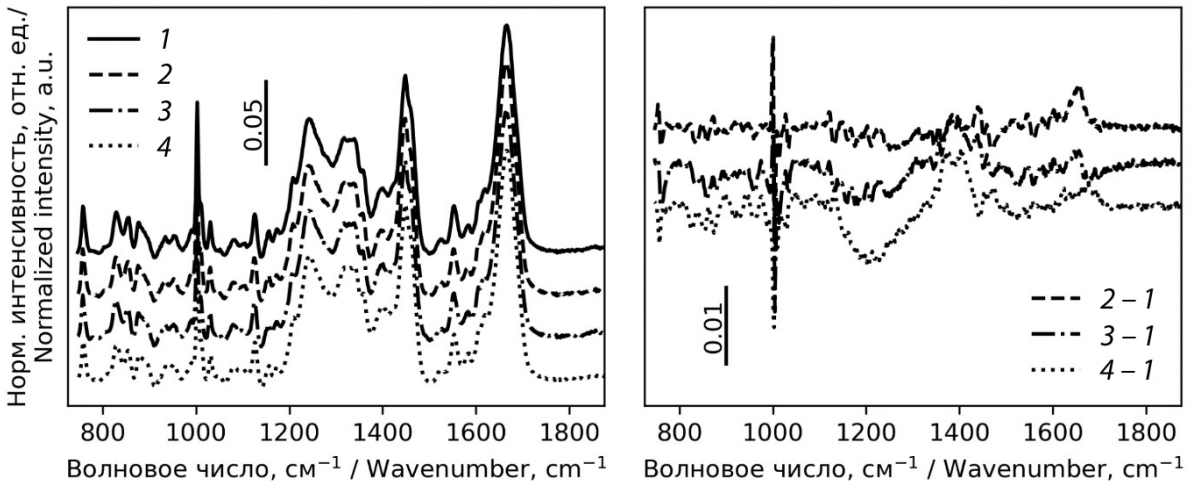
2. Результаты и их обсуждение

Для изучения влияния различных концентраций гиалуроновой кислоты на молекулу изолята сывороточного протеина, были созданы конъюгаты ИСП-ГК. Концентрация ИСП оставалась постоянной во всех образцах и составляла 5% по массе. Для создания конъюгатов использовали три различных количества ГК: 0.1% (соотношение белков и полисахаридов 50 : 1, смесь ИСП + 0.1% ГК); 0.25% (соотношение белков и полисахаридов 20 : 1, смесь ИСП + 0.25% ГК); 0.5% (соотношение белков и полисахаридов 10 : 1, смесь ИСП + 0.5% ГК). В качестве контроля применялся образец ИСП с концентрацией белка 5% по массе. Чтобы определить влияние небольшого количества ГК на молекулу белка, были измерены спектры комбинационного рассеяния света (рис. 1).

Используемая в исследуемых конъюгатах концентрация ГК достаточно мала, средние нормированные спектры полученных образцов слабо отличаются от контроля – спектра КР образца ИСП) (см. рис. 1, *б*). В ходе анализа стало очевидным, что единственной характеристикой, которая претерпевает заметные изменения, является дыхательная мода колебаний фенилаланина 1003 $см^{-1}$ [35, 40].

Ранее для анализа данного набора спектров КР были предложены ансамблевые методы на основе деревьев принятия решений: случайный лес и градиентный бустинг [41]. Данные методы показали высокую точность (свыше 95%) при решении задачи классификации и относительно высокое значение (свыше 0.8 для случайного леса и свыше 0.9 для градиентного бустинга) коэффициента детерминации (R^2) при решении регрессионной задачи. При этом вопрос, какого количества образцов спектров достаточно для построения точной модели классификации и регрессии, требует дополнительного изучения.

Модель случайного леса является частным случаем моделей бэггинга, для которых характерно параллельное обучение единичных моделей с последующим усреднением полученных прогнозов для формирования итогового решения [42–44]. Модель градиентного бустинга как частный случай бустинга характеризуется последовательным использованием единичных моделей, где каждая следующая модель уточняет итоговый прогноз [45–47]. Поскольку модель



а/*a* *б*/*b*

Рис. 1. Средние нормированные спектры КР образцов: *1* – ИСП, *2* – ИСП + 0.1% ГК, *3* – ИСП + 0.25% ГК, *4* – ИСП + 0.5% ГК. Усреднение проводилось по 600 спектрам (карта 20×30 точек) (*а*). Разности между средними нормированными спектрами КР (из спектров конъюгатов ИСП + ГК различных концентраций вычитается спектр ИСП) (*б*). Спектры приведены со смещением. Значения нормированной интенсивности отмечены масштабными отрезками 0.05 (*а*) и 0.01 (*б*)

Fig. 1. Mean normalized Raman spectra of the following samples: *1* – WPI, *2* – WPI + 0.1% HA, *3* – WPI + 0.25% HA, *4* – WPI + 0.5% HA. The averaging was carried out over 600 spectra (a map of 20×30 points) (*a*). Differences between mean normalized spectra: the WPI Raman spectrum was subtracted from the spectra of WPI + HA conjugates of various concentrations (*b*). The spectra are offset. The values of the normalized intensity are marked by scale bars 0.05 (*a*) and 0.01 (*b*)

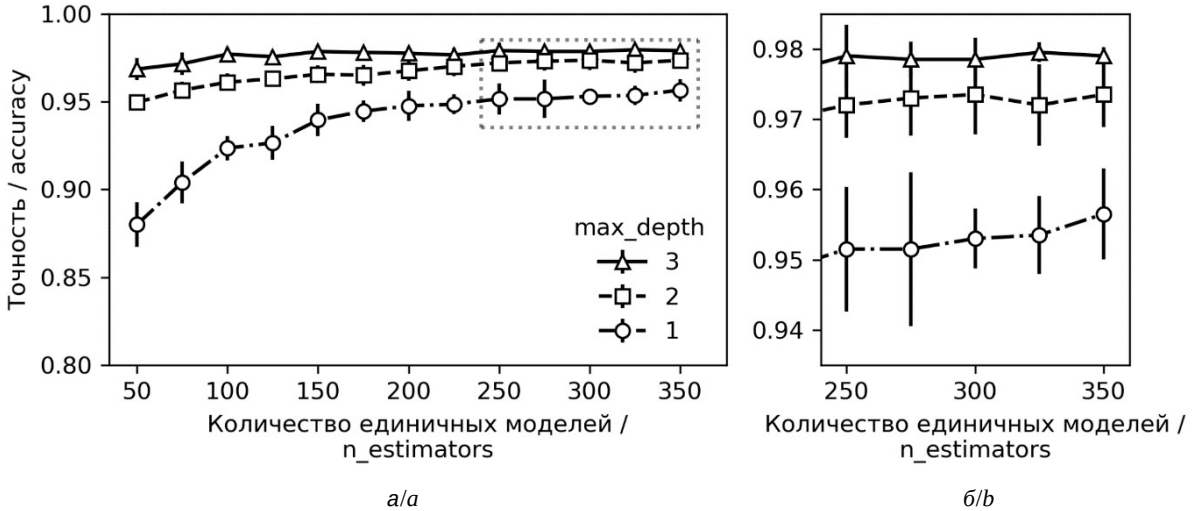
градиентного бустинга продемонстрировала большую точность и коэффициент детерминации по сравнению со случайным лесом, для данного исследования был использован вариант бустинга – адаптивный бустинг.

Адаптивный бустинг также представляет собой ансамблевую модель, в качестве единичной модели в которой используются деревья принятия решения, а для получения итогового прогноза единичные модели обучаются последовательно [15, 16]. Для адаптивного бустинга характерно изменение веса образцов, чей прогноз оказался менее точным.

В работе первоначально был осуществлен подбор параметра max_depth единичного дерева принятия решения и оптимальное количество таких деревьев в ансамблевой модели адаптивного бустинга. Подбор осуществлялся с помощью решетчатого поиска (GridSearchCV) для модели классификации при использовании для обучения 500 спектров КР на каждый тип образца: ИСП, ИСП + 0.1% ГК, ИСП + 0.25% ГК и ИСП + + 0.5% ГК. Параметр max_depth характеризует уровень детализации каждого дерева и в данном случае принимает целочисленные значения 1–3. Количество деревьев в ансамбле изменялось в диапазоне 50–350 с шагом 25. Метрикой производительности модели выступала точность – доля верно классифицированных образцов. Решетчатый поиск позволяет проводить кросс-валидацию, в данном случае 500 спектров на образец делились на 3 отдельных поднабора данных: 2 – на обучение, 1 – на поверку. Процесс повторялся трижды, с изменением поднабора для поверки модели. Усредненные оценки для точности моделей приведены на рис. 2.

Таким образом, был осуществлен подбор оптимальных параметров модели адаптивного бустинга: max_depth = 3, количество единичных моделей = 325. Далее эти параметры были взяты за основу для моделей классификации и регрессии, обучаемых на разном количестве спектров КР в обучающей выборке на каждый тип образца: 50, 100, 200, 300, 400 и 500. На спектрах КР, не вошедших в обучающую выборку, осуществлялась оценка производительности полученных моделей.

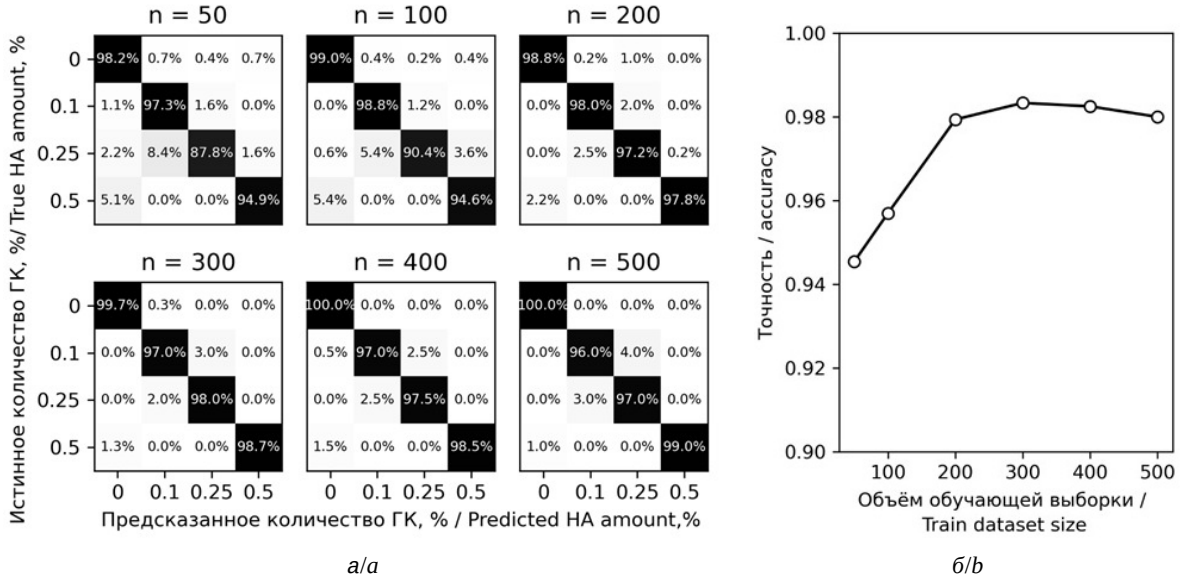
В качестве метрик для оценки работы моделей классификации на основе адаптивного бустинга используются матрицы неточностей (рис. 3, *а*) и точность классификации (рис. 3, *б*). На главной диагонали матриц неточностей отображается доля верно классифицированных спектров. В качестве меток классов на осях отображается количество ГК в каждом типе образца. Точность моделей даже при малом объеме обучающей выборки составила свыше 94%, с увеличением количества спектров КР в процессе обучения до 200 точность моделей возрастает до 98% и далее, с увеличением объема выборки, колеблется около достигнутого значения. Также стоит отметить, что обученные модели клас-



а/*a* *б*/*b*

Рис. 2. Средняя точность ($n = 3$) модели адаптивного бустинга (объем выборки 500 спектров КР на образец) при разном количестве единичных моделей решающих деревьев и разных значений параметра max_depth (*а*). Участок графика, выделенный из рис. (*а*) в рамку (*б*)

Fig. 2. Average accuracy ($n = 3$) of the adaptive boosting model (the sample size is 500 Raman spectra per sample) with different number of single models of decision trees and different values of the max_depth parameter (*a*). Selected section of the graph *a*, highlighted in the frame in Fig. *b*



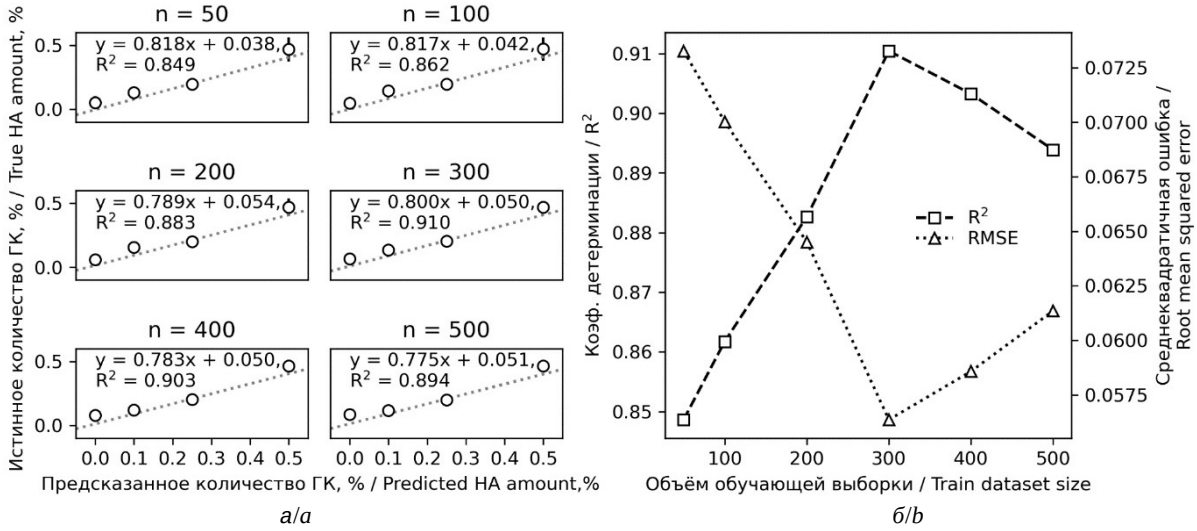
а/*a* *б*/*b*

Рис. 3. Матрицы неточностей (*а*) и точность модели классификатора на основе адаптивного бустинга (*б*) при разном объеме обучающей выборки на каждый образец: 50, 100, 200, 300, 400 и 500

Fig. 3. Confusion matrices (*a*) and accuracy of the adaptive boosting classification models (*b*) trained with different train dataset sizes: 50, 100, 200, 300, 400, and 500

сификации позволяют отличить спектры ИСП от ИСП + 0.1% ГК с точностью свыше 96% вне зависимости от объема обучающей выборки.

Ключевым отличием решения задачи регрессии от классификации является возможность получения калибровочной прямой (рис. 4, *а*), которая потенциально может быть использована для определения образца с неизвестной концентрацией ГК. Для оценки эффективности обученных моделей регрессии на основе адаптивного бустинга использовались коэффициент детерминации R^2 калибровочной прямой и среднеквадратичная ошибка прогнозов RMSE (рис. 4, *б*). Как и в моделях классификации, при увеличении объема обучающей выборки до 300 спектров на тип образца наблюдается увеличение коэффициента детерминации (до 0.910), сопровождающееся уменьшением среднеквадратичной ошибки (до 0.056). После 300 спектров КР на образец данные показатели несколько ухудшаются (R^2 до 0.894,



а/*a* *б*/*b*

Рис. 4. Калибровочные прямые (*а*), коэффициент детерминации R^2 и среднеквадратичная ошибка (*б*) моделей регрессии на основе адаптивного бустинга при разном объеме обучающей выборки на каждый образец: 50, 100, 200, 300, 400 и 500

Fig. 4. Calibration lines (*a*), R^2, and root mean squared error (RMSE) (*b*) of the adaptive boosting regression models trained with different train dataset sizes: 50, 100, 200, 300, 400, and 500

RMSE до 0.061). Вероятно, данный эффект связан с явлением переобучения – слишком точного подстраивания к обучающим данным с потерей эффективности на проверочных данных.

Модели адаптивного бустинга, как и модели градиентного бустинга и случайного леса, позволяют получить график важности признака. При обучении модель воспринимает волновые числа в спектрах КР как отдельные признаки, а нормированную интенсивность как значение признака. График важности признаков показывает, интенсивности на каких волновых числах оказывают большее влияние на итоговый прогноз модели. На рис. 5, *а*, *б* указаны графики важности при-

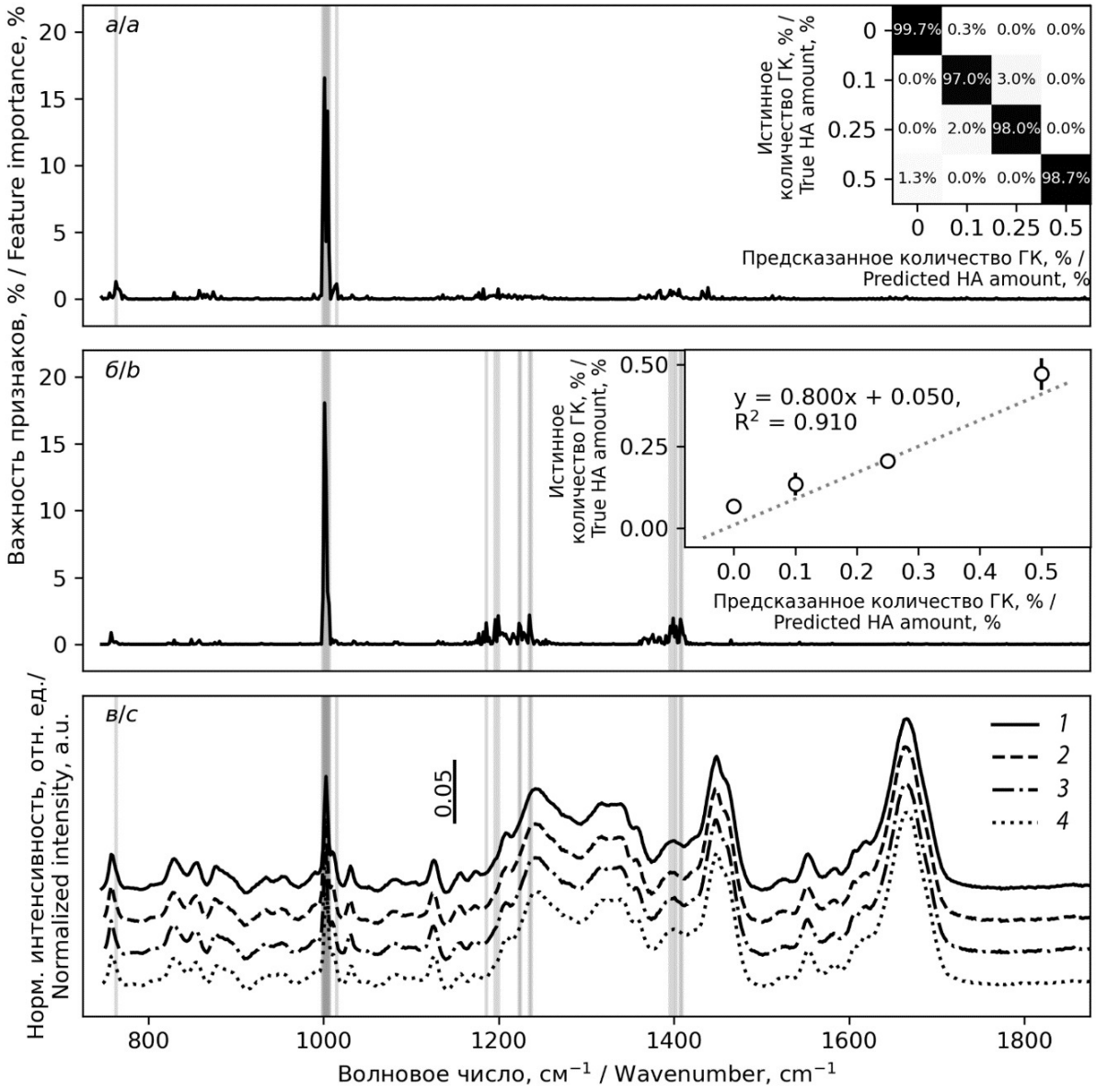


Рис. 5. Графики важности признаков для моделей классификации (*а*) и регрессии (*б*) на основе адаптивного бустинга. Объем обучающей выборки равен 300 спектров КР на тип образца. На вставках приведены соответствующие моделям матрица неточностей (*а*) и калибровочная прямая (*б*). Средние нормированные спектры КР образцов: *1* – ИСП, *2* – ИСП + 0.1% ГК, *3* – ИСП + 0.25% ГК, *4* – ИСП + 0.5% ГК. Усреднение проводилось по 600 спектрам (карта 20×30 точек) (*в*). Серые вертикальные линии указывают волновые числа с важностью более 1%

Fig. 5. Feature importances plots calculated for the adaptive boosting classifier (*a*), and regressor (*b*). The train dataset size is 300 spectra per sample type. The inserts show the confusion matrix (*a*) and the calibration line (*b*) corresponding to the models. Mean normalized Raman spectra of the following samples: *1* – WPI, *2* – WPI + 0.1% HA, *3* – WPI + 0.25% HA, *4* – WPI + 0.5% HA. The averaging was carried out over 600 spectra (a map of 20×30 points) (*c*). The gray vertical lines indicate wavenumbers with importance greater than 1%

знаков для моделей классификации и регрессии соответственно. Объем обучающей выборки был равен 300 спектров КР на тип образца. Волновые числа, важность которых превышала 1%, выделены серым. Такой подход позволяет выявить те колебания химических связей, которые подверглись большим изменениям при добавлении ГК к ИСП. Полоса, соответствующая дыхательной моде колебаний фенилаланина 1003 $см^{-1}$, является наиболее заметной как для модели классификации (рис. 5, *а*), так и для модели регрессии (рис. 5, *б)*. В остальном имеются различия. Так, на модель классификации оказало большее влияние мода колебаний, соответствующая индольному кольцу триптофана 763 $см^{-1}$ [34, 35]. В модели регрессии выделились полоса ~1240 $см^{-1}$, соответствующая амиду III (изгибу в плоскости N–H, растяжению в плоскости C–N), а также ~1400 $см^{-1}$, соответствующая аспарагиновой и глутаминовой кислотам (C=O часть COO^-) [34, 35]. Также стоит отметить, что заметные на усредненных спектрах (рис. 5, *в*) полосы 1450, 1465 $см^{-1}$ (C–H– изгиб алифатических остатков), 1540 $см^{-1}$ (амид II, N–H– деформация), 1667 $см^{-1}$ (амид I, амидное растяжение C=O, N–H– колебание) не оказывают влияния на итоговый прогноз модели, а значит по этим химическим связям не проходит изменений при формировании конъюгата ИСП + ГК при малых концентрациях ГК [36].

Заключение

В работе исследовалось влияния малых количеств гиалуроновой кислоты (ГК, 0.1, 0.25 и 0.5% по массе) на изолят сывороточного протеина (ИСП, 5% по массе) методом спектроскопии комбинационного рассеяния. Для каждого типа конъюгата ИСП-ГК было зарегистрировано по 600 спектров. При сравнении усредненных нормированных спектров были обнаружены отличия только в дыхательной моде колебаний фенилаланина 1003 $см^{-1}$. Для проведения более детального анализа были разработаны модели классификации и регрессии, основанные на алгоритме адаптивного бустинга. Оптимизация гиперпараметров с использованием GridSearchCV позволила добиться высокой производительности при минимальных вычислительных затратах. Исследование влияния размера обучающей выборки показало, что для задачи классификации уже при 50 образцах достигается удовлетворительная точность (94.5%). Для повышения качества классификации (точность 97.9% или 98.3%) рекомендуется увеличить объем выборки до 200 или 300 спектров соответственно, что по-прежнему считается небольшим набором данных. При определении регрессии оптимальный размер обучающей выборки составлял 300 образцов, при этом достигнуты значения: для коэффициента детерминации (R^2) 0.910 и для среднеквадратичной ошибка (RMSE) 0.061%. Таким образом, модели демонстрируют высокую эффективность даже при ограниченном наборе данных, что делает их особенно ценными в условиях, когда сбор больших выборок затруднителен. Модели адаптивного бустинга также позволяют выделить те колебательные моды, изменения которых повлияли сильнее на прогнозы и эффективность обученных моделей. Добавление и конъюгирование ГК с ИСП приводит к изменениям на полосах: 763 $см^{-1}$ (индольное кольцо триптофана); 1003 $см^{-1}$ (дыхательная мода колебаний фенилаланина); ~1240 $см^{-1}$ (амид III, изгиб в плоскости N–H, растяжение в плоскости C–N); ~1400 $см^{-1}$ (аспарагиновая и глутаминовая кислоты, C=O часть COO^-).

Разработанные модели универсальны при анализе данных спектроскопии комбинационного рассеяния и применимы как для классификации образцов с известными концентрациями, так и для регрессионного анализа смесей с неизвестными значениями, обеспечивая надежные результаты при ограниченных ресурсах.

Список литературы / References

1. Vaou N., Stavropoulou E., Voidarou C., Tsakris Z., Rozos G., Tsigalou C., Bezirtzoglou E. Interactions between medical plant-derived bioactive compounds: Focus on antimicrobial combination effects. *Antibiotics*, 2022, vol. 11, iss. 8, art. 1014. https://doi.org/10.3390/antibiotics11081014
2. Mehta N., Kumar P., Verma A. K., Umaraw P., Kumar Y., Malav O. P., Sazili A. Q., Domínguez R., Lorenzo J. M. Microencapsulation as a noble technique for the application of bioactive compounds in the food Industry: A comprehensive review. *Appl. Sci.*, 2022, vol. 12, no. 3, art. 1424. https://doi.org/10.3390/app12031424
3. Senthilkumar K., Vijayalakshmi A., Jagadeesan M., Somasundaram A., Pitchiah S., Gowri S. S., Alharbi S. A., Ansari M. J., Ramasamy P. Preparation of self-preserving personal care cosmetic products using multifunctional ingredients and other cosmetic ingredients. *Sci. Rep.*, 2024, vol. 14, no. 1, art. 19401. https://doi.org/10.1038/s41598-024-57782-9
4. Saletnik A., Saletnik B., Puchalski C. Overview of Popular Techniques of Raman Spectroscopy and Their Potential in the Study of Plant Tissues. *Molecules*, 2021, vol. 26, no. 6, art. 1537. https://doi.org/10.3390/molecules26061537

5. Rebrosova K., Samek O., Kizovsky M., Bernatova S., Hola V., Ruzicka F. Raman spectroscopy – A novel method for identification and characterization of microbes on a single-cell level in clinical settings. *Front. Cell. Infect. Microbiol.*, 2022, vol. 12, art. 866463. https://doi.org/10.3389/fcimb.2022.866463
6. Pezzotti G. Raman spectroscopy in cell biology and microbiology. *J. Raman Spectrosc.*, 2021, vol. 52, no. 12, pp. 2348–2443. https://doi.org/10.1002/jrs.6204
7. Kočišová E., Kuižová A., Procházka M. Analytical applications of droplet deposition Raman spectroscopy. *Analyst*, 2024, vol. 149, iss. 12, pp. 3276–3287. https://doi.org/10.1039/D4AN00336E
8. Dodo K., Fujita K., Sodeoka M. Raman Spectroscopy for Chemical Biology Research. *J. Am. Chem. Soc.*, 2022, vol. 144, no. 43, pp. 19651–19667. https://doi.org/10.1021/jacs.2c05359
9. Koronaki E. D., Kaven L. F., Faust J. M., Kevrekidis I. G., Mitsos A. Nonlinear manifold learning determines microgel size from Raman spectroscopy. *AIChE J.*, 2024, vol. 70, no. 10, art. e18494. https://doi.org/10.1002/aic.18494
10. Zhang Y., Gao P., Zhang N., Hong H., Ruan J., Gao X. Efficient detection of specific pharmaceutical components in compound medications based on Raman spectroscopy. *Opt. Commun.*, 2025, vol. 577, art. 131470. https://doi.org/10.1016/j.optcom.2024.131470
11. Sun Y., Tang H., Zou X., Meng G., Wu N. Raman spectroscopy for food quality assurance and safety monitoring: A review. *Curr. Opin. Food Sci.*, 2022, vol. 47, art. 100910. https://doi.org/10.1016/j.cofs.2022.100910
12. Fernández-Manteca M. G., Ocampo-Sosa A. A., de Alegría-Puig C. R., Roiz M. P., Rodríguez-Grande J., Madrazo F., Calvo J., Rodríguez-Cobo L., López-Higuera J. M., Fariñas M. C., Cobo A. Automatic classification of Candida species using Raman spectroscopy and machine learning. Spectrochim. *Acta Part A Mol. Biomol. Spectrosc.*, 2023, vol. 290, art. 122270. https://doi.org/10.1016/j.saa.2022.122270
13. Guo F., Yang X., Zhang Z., Liu S., Zhang Y., Wang H. Rapid Raman spectroscopy analysis assisted with machine learning: A case study on Radix Bupleuri. *J. Sci. Food Agric.*, 2025, vol. 105, iss. 4, pp. 2412–2419. https://doi.org/10.1002/jsfa.14012
14. Tang J.-W., Li F., Liu X., Wang J. T., Xiong X. S., Lu X. Y., Zhang X.-Y., Si Y.-T., Umar Z., Tay A. C. Y., Marshall B. J., Yang W.-X., Gu B., Wang L. Detection of Helicobacter pylori Infection in Human Gastric Fluid Through Surface-Enhanced Raman Spectroscopy Coupled With Machine Learning Algorithms. *Lab. Investig.*, 2024, vol. 104, iss. 2, art. 100310. https://doi.org/10.1016/j.labinv.2023.100310
15. Freund Y., Schapire R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 1997, vol. 55, no. 1, pp. 119–139. https://doi.org/10.1006/jcss.1997.1504
16. Zhu J., Zou H., Rosset S., Hastie T. Multi-class adaboost. *Stat. Interface*, 2009, vol. 2, no. 3, pp. 349–360. https://doi.org/10.4310/SII.2009.v2.n3.a8
17. Wang P., Li Y., Wang K., Qu H. Research on the application of ensemble learning methods for rapid diagnosis of osteoarthritis. Ensemble learning-assisted rapid diagnosis methods. Practical research on the application of serum Raman spectroscopy combined with ensemble learning methods. In: *ICBAR'24: Proceedings of the 2024 4th International Conference on Big Data, Artificial Intelligence and Risk Management*. New York, ACM, 2024, pp. 421–427. https://doi.org/10.1145/3718751.3718818
18. Poth M., Magill G., Filgertshofer A., Popp O., Großkopf T. Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing. *J. Raman Spectrosc.*, 2022, vol. 53, no. 9, pp. 1580–1591. https://doi.org/10.1002/jrs.6402
19. Mishra D. P., Gupta H. K., Saajith G., Bag R. Optimizing heart disease prediction model with gridsearch CV for hyperparameter tuning. In: *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC–CGU)*. IEEE, 2024, pp. 1–6. https://doi.org/10.1109/IC-CGU58078.2024.10530772
20. Muzayanah R., Pertiwi D. A. A., Ali M., Muslim M. A. Comparison of gridsearchcv and bayesian hyperparameter optimization in random forest algorithm for diabetes prediction. *J. Soft Comput. Explor.*, 2024, vol. 5, no. 1, pp. 86–91. https://doi.org/10.52465/joscex.v5i1.308
21. Kurniasih A., Previana C. N. Implementation of GridSearchCV to find the best hyperparameter combination for classification model flgorithm in predicting water potability. *J. Artif. Intell. Eng. Appl.*, 2025, vol. 4, no. 2, pp. 1174–1182. https://doi.org/10.59934/jaiea.v4i2.844
22. Rajput D., Wang W.-J., Chen C.-C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, 2023, vol. 24, no. 1, art. 48. https://doi.org/10.1186/s12859-023-05156-9
23. Ramezan C. A., Warner T. A., Maxwell A. E., Price B. S. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. *Remote Sens.*, 2021, vol. 13, iss. 3, art. 368. https://doi.org/10.3390/rs13030368
24. Stahlschmidt S. R., Ulfenborg B., Synnergren J. Multimodal deep learning for biomedical data fusion: A review. *Brief. Bioinform.*, 2022, vol. 23, iss. 2, art. bbab569. https://doi.org/10.1093/bib/bbab569
25. Bates F., Busato M., Piletska E., Whitcombe M. J., Karim K., Guerreiro A., del Valle M., Giorgetti A., Piletsky S. Computational design of molecularly imprinted polymer for direct detection of melamine in milk. *Sep. Sci. Technol.*, 2017, vol. 52, iss. 8, pp. 1441–1453. https://doi.org/10.1080/01496395.2017.1287197
26. Lu Y., Xia Y., Liu G., Pan M., Li M., Lee N. A., Wang S. A Review of methods for detecting melamine in food samples. *Crit. Rev. Anal. Chem.*, 2017, vol. 47, iss. 1, pp. 51–66. https://doi.org/10.1080/10408347.2016.1176889
27. Einkamerer O. B., Ferreira A. V., Fair M. D., Hugo A. The effect of dietary non-protein nitrogen content on the meat quality of finishing lambs. *S. Afr. J. Anim.*, 2024, vol. 54, no. 3, pp. 340–357. https://doi.org/10.4314/sajas.v54i3.05
28. Alizadeh Sani M., Jahed-Khaniki G., Ehsani A., Shariatifar N., Hadi Dehghani M., Hashemi M., Hosseini H.,

Abdollahi M., Hassani S., Bayrami Z., McClements D. J. Metal-organic framework fluorescence sensors for rapid and accurate detection of melamine in milk powder. *Biosensors*, 2023, vol. 13, no. 1, art. 94. https://doi.org/10.3390/bios13010094

29. Lukacs M., Zaukuu J. L. Z., Bazar G., Pollner B., Fodor M., Kovacs Z. Comparison of multiple NIR spectrometers for detecting low-concentration nitrogen-based adulteration in protein powders. *Molecules*, 2024, vol. 29, no. 4, art. 781. https://doi.org/10.3390/molecules29040781
30. Lukacs M., Bazar G., Pollner B., Henn R., Kirchler C. G., Huck C. W., Kovacs Z. Near infrared spectroscopy as an alternative quick method for simultaneous detection of multiple adulterants in whey protein-based sports supplement. *Food Control*, 2018, vol. 94, pp. 331–340. https://doi.org/10.1016/j.foodcont.2018.07.004
31. Marinho A., Nunes C., Reis S. Hyaluronic acid: A key ingredient in the therapy of inflammation. *Biomolecules*, 2021, vol. 11, no. 10, art. 1518. https://doi.org/10.3390/biom11101518
32. Yasin A., Ren Y., Li J., Sheng Y., Cao C., Zhang K. Advances in hyaluronic acid for biomedical applications. *Front. Bioeng. Biotechnol.*, 2022, vol. 10, art. 910290. https://doi.org/10.3389/fbioe.2022.910290
33. Juncan A. M., Moisă D. G., Santini A., Morgovan C., Rus L. L., Vonica-Țincu A. L., Loghin F. Advantages of hyaluronic acid and Its combination with other bioactive ingredients in cosmeceuticals. *Molecules*, 2021, vol. 26, no. 15, art. 4429. https://doi.org/10.3390/molecules26154429
34. Iaconisi G. N., Lunetti P., Gallo N., Cappello A. R., Fiermonte G., Dolce V., Capobianco L. Hyaluronic Acid: A powerful biomolecule with wide-ranging applications – A comprehensive review. *Int. J. Mol. Sci.*, 2023, vol. 24, no. 12, art. 10296. https://doi.org/10.3390/ijms241210296
35. Wang N., Zhao X., Jiang Y., Ban Q., Wang X. Enhancing the stability of oil-in-water emulsions by non-covalent interaction between whey protein isolate and hyaluronic acid. *Int. J. Biol. Macromol.*, 2023, vol. 225, pp. 1085–1095. https://doi.org/10.1016/j.ijbiomac.2022.11.170
36. Zhong W., Li C., Diao M., Yan M., Wang C., Zhang T. Characterization of interactions between whey protein isolate and hyaluronic acid in aqueous solution: Effects of pH and mixing ratio. *Colloid. Surf. B: Biointerfaces*, 2021, vol. 203, art. 111758. https://doi.org/10.1016/j.colsurfb.2021.111758
37. Zhong W., Zhang T., Dong C., Li J., Dai J., Wang C. Effect of sodium chloride on formation and structure of whey protein isolate/hyaluronic acid complex and its ability to loading curcumin. *Colloid. Surf. A: Physicochem. Eng. Asp.*, 2022, vol. 632, art. 127828. https://doi.org/10.1016/j.colsurfa.2021.127828
38. Zhong W., Li J., Wang C., Zhang T. Formation, stability and *in vitro* digestion of curcumin loaded whey protein/hyaluronic acid nanoparticles: Ethanol desolvation vs. pH-shifting method. *Food Chem.*, 2023, vol. 414, art. 135684. https://doi.org/10.1016/j.foodchem.2023.135684
39. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 2011, vol. 12, iss. 85, pp. 2825–2830. Available at: http://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (accessed April 22, 2025).
40. Zhao Y., Ma C. Y., Yuen S. N., Phillips D. L. Study of succinylated food proteins by Raman spectroscopy. *J. Agric. Food Chem.*, 2004, vol. 52, iss. 7, pp. 1815–1823. https://doi.org/10.1021/jf030577a
41. Mayorova O. A., Saveleva M. S., Bratashov D. N., Prikhozhdenko E. S. Combination of machine learning and Raman spectroscopy for determination of the complex of whey protein isolate with hyaluronic acid. *Polymers*, 2024, vol. 16, no. 5, art. 666. https://doi.org/10.3390/polym16050666
42. Breiman L. Random Forests. *Mach. Learn.*, 2001, vol. 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324
43. Becker T., Rousseau A. J., Geubbelmans M., Burzykowski T., Valkenborg D. Decision trees and random forests. *Am. J. Orthod. Dentofac. Orthop.*, 2023, vol. 164, iss. 6, pp. 894–897. https://doi.org/10.1016/j.ajodo.2023.09.011
44. Sun Z., Wang G., Li P., Wang H., Zhang M., Liang X. An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Syst. Appl.*, 2024, vol. 237, pt. B, art. 121549. https://doi.org/10.1016/j.eswa.2023.121549
45. Friedman J. H. Greedy Function Approximation: A gradient boosting machine. *Ann. Stat.*, 2001, vol. 29, no. 5, pp. 1189–1232. Available at: http://www.jstor.org/stable/2699986 (accessed April 22, 2025)
46. Friedman J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 2002, vol. 38, iss. 4, pp. 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
47. Wang M., Zhang J. Surface enhanced Raman spectroscopy Pb2+ Ion Detection based on a gradient boosting decision tree algorithm. *Chemosensors*, 2023, vol. 11, no. 9, art. 509. https://doi.org/10.3390/chemosensors11090509

Поступила в редакцию 02.05.2025; одобрена после рецензирования 23.05.2025;
принята к публикации 12.06.2025; опубликована 29.08.2025

The article was submitted 02.05.2025; approved after reviewing 23.05.2025;
accepted for publication 12.06.2025; published 29.08.2025