

Article

Not peer-reviewed version

Raman Spectroscopy of Protein-Polysaccharide Conjugates: A Comparative Study of Tree-Based Ensemble Models

Oksana A. Mayorova , [Mariia S. Saveleva](#) , [Ekaterina S. Prikhodzhenko](#) *

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2649.v1

Keywords: Raman spectroscopy; machine learning; ensemble models; whey protein isolate; hyaluronic acid



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Raman Spectroscopy of Protein-Polysaccharide Conjugates: A Comparative Study of Tree-Based Ensemble Models

Oksana A. Mayorova ¹, Mariia S. Saveleva ¹ and Ekaterina S. Prikhodzhenko ^{2,*}

¹ Science Medical Center, Saratov State University, 83 Astrakhanskaya str., 410012 Saratov, Russia

² Future Biophysics Institute, Moscow Center for Advanced Studies, 20 Kulakova str., 141700 Moscow, Russia

* Correspondence: prikhodzhenko.es@moscowcas.ru

Abstract

Proteins with additives, especially in small quantities, are of great interest as a subject of a study. Machine learning approaches implemented to Raman spectroscopy data could provide an insight into chemical structure of such mixtures or conjugates. Although, decision tree model could be powerful in solving either classification or regression task and could provide accessible predictions, it is prone to overfitting. Ensemble models that implement several decision trees could overcome the determined problem. Five different model types are discussed: RandomForest, GradientBoosting, AdaBoost, Voting, and Stacking. Raman spectroscopy data of whey protein isolate (5 wt. %) with different amounts of hyaluronic acid (0, 0.1, 0.25, and 0.5 wt. %) were used as datasets. Optimization established that ensembles of 200 decision trees with a maximum depth of four were optimal. AdaBoostClassifier found to be the most efficient in finding differences between whey protein isolate and its conjugates with hyaluronic acid: 99.5% accuracy, 100% sensitivity, and 98.0% specificity. Stacking of RandomForest, GradientBoosting, and AdaBoost regressors with final estimator of RidgeCV was the most effective approach in the regression task ($R^2 = 0.963$). According to the feature importance plots, the Raman bands that were most influential in predicting the results were 1003 cm^{-1} (phenylalanine, ring breath), 1206 cm^{-1} (C–C stretching), 1240 cm^{-1} (amide III (β -sheet), N–H in-plane bend, C–N stretch), and 1399 cm^{-1} (aspartic and glutamic acids, C=O stretch of COO[−]).

Keywords: Raman spectroscopy; machine learning; ensemble models; whey protein isolate; hyaluronic acid

1. Introduction

Multicomponent mixtures containing biological active substances are of great importance in the fields of pharmaceuticals, food industry, biotechnology, and tissue regeneration [1–3]. However, the analysis of these substances is difficult due to the low concentration of target components, complex matrix structures and the need for sensitive methods. This is particularly relevant when studying the interaction between macromolecules and the formation, for instance, of complexes and conjugates. Therefore, the development of new methods for detecting and assessing trace substances in these systems remains crucial.

Raman spectroscopy offers an excellent opportunity to investigate the chemical composition and structure of substances [4–7]. This has several advantages, such as being non-destructive and highly responsive to structural modifications. Various samples, including biological molecules, polymers, and pharmaceuticals can be examined using this technique [8–11]. Complex mixtures, macromolecule interactions, and massive data require the use of machine learning methods for analysis. If only slight variations are present in the Raman spectrum, these methods are highly advantageous.

Decision trees (DT) are a type of machine learning model that can be used to solve both classification and regression problems [12,13]. They provide a human-readable explanation of the reasoning behind the decision made, making them a valuable tool for understanding complex data. However, this estimator can potentially lead to overfitting of the training data. Therefore, the use of multiple DT in a single ensemble can help to overcome these limitations. Several estimators can be used either in parallel (bootstrap aggregating algorithms) or in sequence (boosting algorithms). Random forest (RF) is a special case of bootstrap aggregating algorithms in which DT are implemented as estimators [12–16]. The final prediction is made by either majority voting in solving a classification problem, or by averaging predictions in solving a regression problem. Gradient boosting (GB) is a method that involves fitting each individual estimator to the residual errors in a sequential manner [17–19]. AdaBoost (AB) algorithm involves sequential adjusting the weights of samples based on their level of inaccuracy [20,21]. Voting algorithm is more generalized approach in machine learning as models of different origin can be united into a single ensemble [22]. In the stacking algorithm, the predictions from each individual estimator are combined and used as input for a final estimator in order to produce the final prediction [23]. This final estimator is trained through cross-validation.

One of the most challenging tasks when employing Raman spectroscopy in conjunction with advanced machine learning techniques is the investigation of proteins. The incorporation of contaminants, particularly in small quantities, into protein formulations can have both beneficial and detrimental consequences. The latter category, in particular, encompasses substances such as melamine [24,25] and urea [26], which are added to milk, infant formula, and feed to artificially inflate the "pseudo-protein" metrics [27,28]. Additionally, there is the practice of "amino acid adulteration", which involves the addition of inexpensive amino acids to manipulate the outcomes of food protein analyses [29]. To more precisely identify hazardous additives, it is essential to develop novel methods for analyzing and detecting contaminants.

The addition of impurities to proteins can also provide positive properties to the final material. Hyaluronic acid (HA) is an important biopolymer, widely used in medicine and cosmetology due to its unique properties such as moisturizing and tissue regeneration [30–33]. Adding a small amount of hyaluronic acid to a whey protein isolate can significantly improve its properties without greatly increasing the cost. This property is especially valuable when creating targeted drug delivery systems. The use of the WPI+HA complex as a stabilizing agent instead of WPI can significantly increase the service life of microcarriers that are produced using it [34].

In this work, mixtures of whey protein isolate (WPI, 5 wt. %) with the addition of hyaluronic acid in various concentrations (0, 0.1, 0.25 and 0.5 wt%) are investigated. The aim of this research is to compare the performance of different ensemble machine learning models, including RF, GB, AB, voting and stacking, when applied to spectroscopic data in solving both classification and regression tasks. The findings of this study may contribute to the development of novel methods for quality control and analysis of complex multi-component systems in various industrial settings.

2. Materials and Methods

2.1. Materials

The whey protein isolate (WPI) was purchased from California Gold Nutrition® (Irvine, CA, USA). Sodium salt of hyaluronic acid (HA, purity of 99%, 404 Mw) was obtained from Macklin Biochemical Co., Ltd. (Shanghai, China). Sodium chloride was purchased in Sigma Aldrich (St. Louis, MO, USA). The deionized water (Millipore Milli-Q, resistance of 18.2 MΩ·cm⁻¹) was utilized during all sets of experiments.

2.2. Sample Preparation

WPI solution (10 wt.% in 0.15 M NaCl), and HA solutions (0.2 wt. %, 0.5 wt. %, and 1 wt.% in 0.15 M NaCl) were used in all experiments. In order to prepare WPI+HA conjugates, equal volumes

of WPI and HA solutions were mixed and shaken for 30 min at 22 °C. The resulting compositions were as follows: WPI + 0.1% HA (5 wt. % : 0.1 wt. %, 50:1 ratio), WPI + 0.25% HA (5 wt. % : 0.25 wt. %, 20:1 ratio), WPI + 0.5% HA (5 wt. % : 0.5 wt. %, 10:1 ratio). The WPI+HA conjugates obtained were washed to remove the free HA by dialysis against saline for 3 days at 4 °C. Initial WPI mixed with 0.15 M NaCl in equal volumes was used as control WPI sample (5 wt. %).

2.3. Raman Data Acquisition

WPI and WPI+HA samples were placed on a quartz substrate (10 μ L per sample). Raman measurements were performed after sample drying and forming a thin film without any coffee ring effect. Renishaw inVia spectrometer (Renishaw, Wotton-under-Edge, UK) with 532 nm laser focused through 50 $\times$ /0.5 N.A. objective was used for data collection. Raman maps (20 $\times$ 30 spectra with 2 μ m step) were collected from each sample.

2.4. Data Analysis

All Raman spectra were collected using Renishaw WiRE v. 4.2 software (Renishaw, UK). The Subtract Baseline tool of Renishaw WiRE software was implemented to remove background using eleven-degree polynomial. Further data processing was performed in Jupyter Notebook environment with Python 3.

Data import was carried out with renishawWiRE v. 0.1.16 library [35]. All machine learning approaches were conducted with sci-kit learn (sklearn) v. 1.7.2 library [36]. As the last preprocessing step, Raman spectra were normalized by the l^2 -norm using sklearn.preprocessing.normalize. All collected Raman spectra were split into train and test datasets in 3:1 ratio with sklearn.model_selection.train_test_split. There were two classes in classification task, WPI (600 spectra) and WPI+HA (1800 spectra). HA amount (0%, 0.1%, 0.25%, and 0.5%) was used as target value in the regression task. Model optimization in a sense of number of DT and maximum depth of each DT (n_estimators and max_depth parameters, respectively) was performed for RF, GB, and AB classifiers using sklearn.model_selection.GridSearchCV with 3-fold cross-validation and accuracy as metric. All discussed models were implemented using sklearn.ensemble. The accuracy_score, confusion_matrix (classification models), r2_score (regression models) metrics from sklearn were used to evaluate the performance of models.

3. Results and Discussion

In order to analyze the impact of HA addition onto chemical structure of WPI, three different conjugate compositions were prepared: WPI + 0.1% HA (5 wt. % : 0.1 wt. %, 50:1 ratio), WPI + 0.25% HA (5 wt. % : 0.25 wt. %, 20:1 ratio), WPI + 0.5% HA (5 wt. % : 0.5 wt. %, 10:1 ratio). Raman maps (20 $\times$ 30, 600 single spectra) were recorded for each WPI + HA conjugate as well as pure WPI (5 wt. %). The results of Raman spectroscopy are shown on Figure 1. All Raman bands of WPI are indicated with grey dotted lines (Figure 1a).

There were only two Raman bands on difference spectrum (Figure 1a), on which the impact of HA addition can be noted, namely 756 and 1003 cm^{-1} . However, upon closer examination, the dependence of normalized intensity of 756 cm^{-1} on HA amount is not linear (Figure 1b). Normalized intensities of Raman bands marked on Figure 1a are shown on Figure 1b. Raman band assignments are indicated in Table 1.

Conventional approach in analysis of Raman spectra (both by spectrum of difference and dependences of intensities on HA amount) has not provided a sufficient answer on which chemical bonds of WPI changed the most upon HA addition and whether it is possible to reliably differentiate between WPI and WPI + HA. The latter can be crucial not only in the discussed conjugates but considering wider issues of mixtures with small amounts of additives. To improve the analysis, the ensemble models based on DT were implemented. Previously, RF, GB, and AB models were discussed [37,38]. In all of these models, DT were used as individual estimators combined into

ensemble. As current work is dedicated to thorough analysis of implementation of ensemble models to Raman spectroscopy data, voting and stacking models would also be discussed. To obtain higher performance of models, number of DT in ensemble ($n_{estimators}$) and maximum depth of each DT (max_depth) had been optimized using cross-validation via GridSearchCV [39,40] with accuracy as metric. Train dataset was divided into three subsets for cross-validation. Model metrics are shown on Figure 2. Summary (red line) indicates average metrics of RF, GB, and AB models.

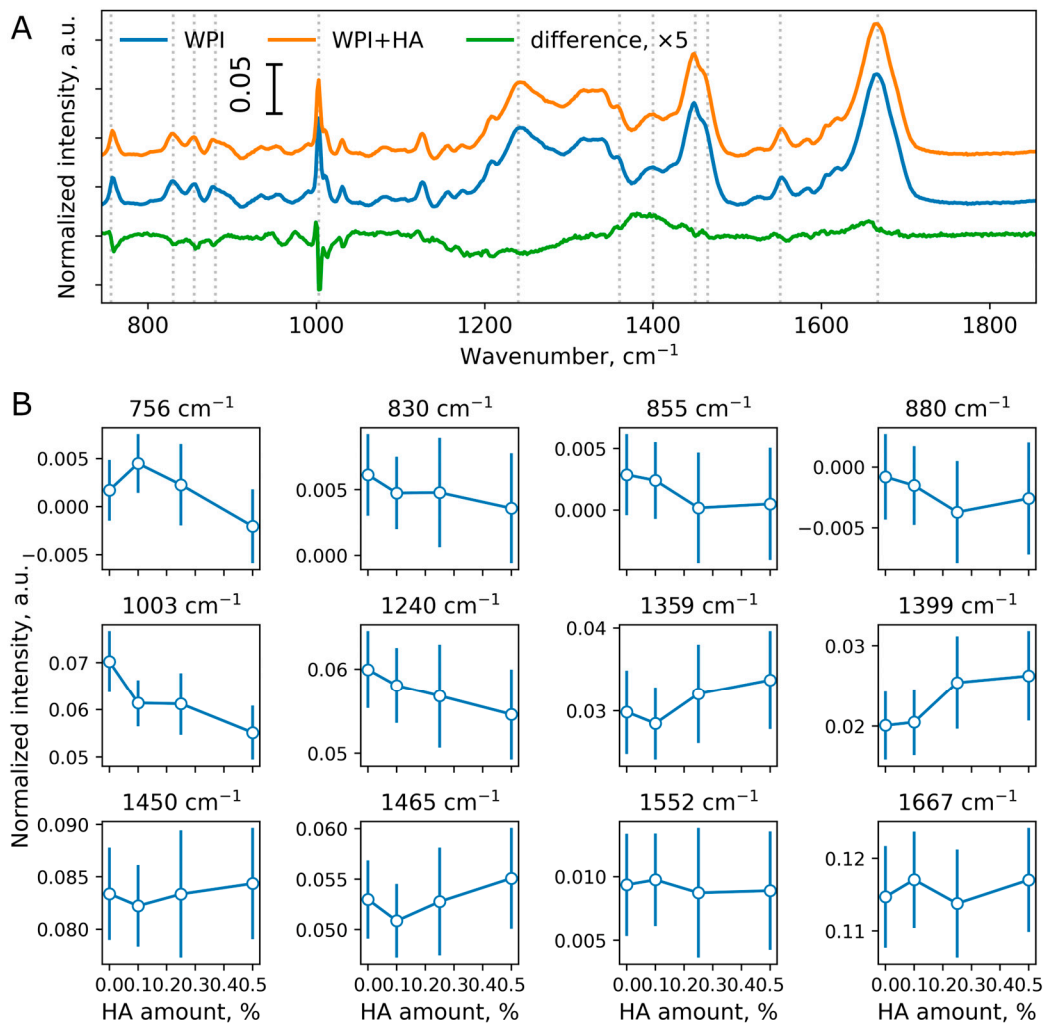


Figure 1. Raman spectroscopy data on WPI and WPI + n% HA ($n = 0.1, 0.25, 0.5$). (a) Average spectra of WPI (600 spectra) and WPI + HA (600 spectra per HA amount; 1800 spectra overall). Raman bands of WPI marked with grey dotted lines. Spectrum of difference was magnified by 5 for clarity. (b) Dependence of Raman band normalized intensities on HA amount (mean $\pm$ standard deviation).

Table 1. Raman bands of WPI with their assignments [34,41,42].

Wavenumber, cm ⁻¹	Assignment
756, 880, 1359	Tryptophan, indole ring
830, 855	Tyrosine, Fermi resonance between ring fundamental and overtone
1003	Phenylalanine, ring breath
1206	C–C stretching
1240	Amide III (β -sheet), N–H in-plane bend, C–N stretch
1399	Aspartic and glutamic acids, C=O stretch of COO–
1450, 1465	Aliphatic residues, C–H bending
1552	Tyrosine, ring stretching

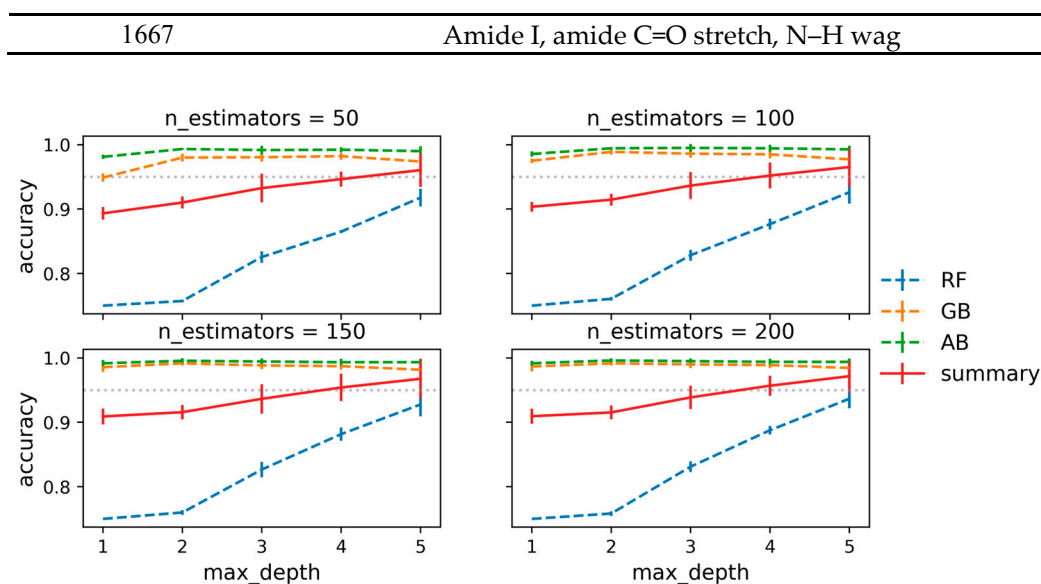


Figure 2. Accuracies of RF, GB, AB models on maximum depth (max_depth) and number of DT (n_estimators). Summary represents average metrics of RF, GB, AB models. Mean values and standard deviations were calculated in cross-validation (n = 3). Horizontal grey dotted line indicates accuracy = 0.95.

According to summary, max_depth parameter should be greater than 3 as average accuracies became greater than 0.95. Accuracies (mean and standard deviation values) for models with max_depth greater than 3 are shown in Table S1 (Supplementary information). Evaluating the performances of models with different parameters, the best parameters were found to be 4 for max_depth and 200 for n_estimators. These parameters were further used in training other classification and regression models.

Six different classification models were compared: RF, GB, AB, Voting (hard), Voting (soft), and Stacking. In both Voting models, RF, GB, and AB were used as estimators, majority prognosis was implied in hard version and average of class probabilities was used in soft version. In Stacking, predictions of RF, GB, and AB were used as input data to logistic regressor, which was applied as final estimator in the ensemble. The following metrics were calculated to evaluate the performance of each model: confusion matrix, accuracy, sensitivity, and specificity. The results of classification of WPI and WPI + HA Raman data are shown on Figures 3 and S1 and in Table 2.

Classification models based on DT (namely, RF, GB, and AB) can also provide information on importance of features. As intensities at different wavenumbers are considered as features in the analysis, this information can be visualized in the same wavenumber range as the Raman spectra being analyzed (Figure S1). The classification model developed using the RF algorithm was determined to be less precise and less specified than the other models that were evaluated. Although WPI + HA spectra were correctly labeled by the RF model in 100% of the test dataset, 41.3% of the WPI spectra in the test dataset were also considered to be WPI + HA by the RF. Presumably, the lack of specificity observed could be due to the different sizes of the WPI and WPI + HA datasets used (600 and 1800 spectra, respectively). Only wavenumbers with an importance to the analysis greater than 1% were considered valuable and were noted. (Figures 3 and S1). The most important for RF, GB, and AB analyses were found to be 1003 cm^{-1} (ring breath of phenylalanine), 1206 cm^{-1} (C–C stretching), and 1399 cm^{-1} (aspartic and glutamic acids, C=O stretch of COO^-). Most classification models, with the exception of RF, have demonstrated accuracies greater than 98%. Overall, the AB model performed best out of all the models considered for binary classification tasks.

Although samples WPI could be distinguished from WPI + HA using classification models fitted to Raman spectroscopy data, other tasks remained equally important. Consequently, the possibility of creating regression models that could be used on the same dataset was also investigated. Five different regression model have been implemented, namely, RF, GB, AB, Voting, and Stacking. The

additional model optimization had not been performed, the trained RF, GB, and AB regression models consisted of 200 DT with max_depth equal to 4. Similarly to classification task, Voting regressor generated the final prediction by averaging RF, GB, AB results. In Stacking regressor, Ridge model with cross-validation utilized RF, GB, AB prognoses as inputs. In the constructed regression task, the amount of HA in WPI + HA samples (0, 0.1, 0.25, and 0.5 wt. %) was used as target value. The results of solving the regression task are demonstrated on Figures 4 and S2. In order to evaluate the performance of each model, R^2 (determination coefficient) metric was applied.

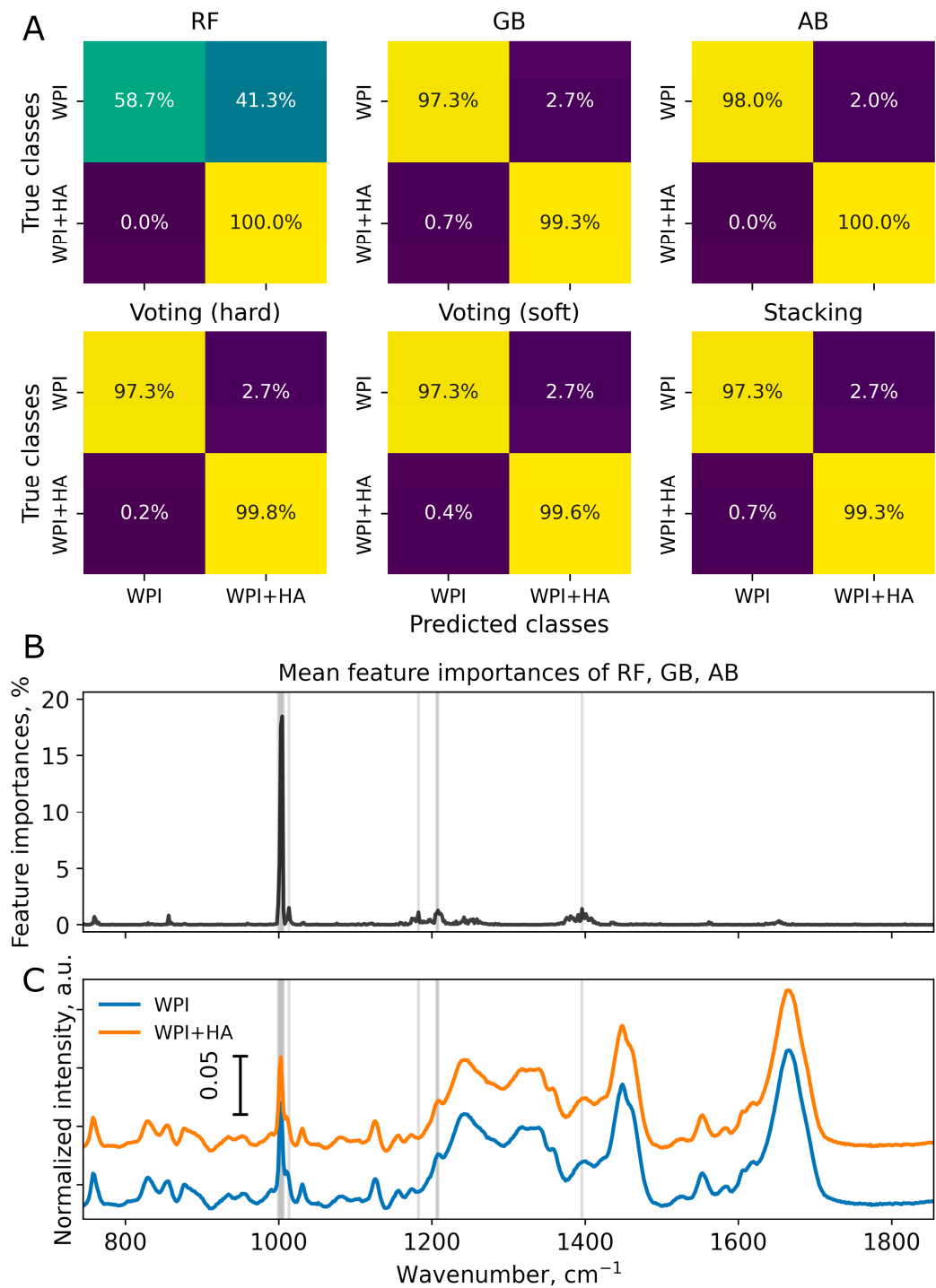


Figure 3. (a) Confusion matrices of each classifier, test dataset consists of 150 spectra of WPI and 450 spectra of WPI + HA (150 spectra per each HA amount: 0.1, 0.25, and 0.5 wt.%). (b) Mean feature importances of RF, GB,

and AB. Wavenumbers with importance greater than 1% are marked with vertical grey lines. (c) Normalized mean spectra of WPI and WPI + HA.

Table 2. Metrics of classification models.

Model\metric	RF	GB	AB	Voting (hard)	Voting (soft)	Stacking
Accuracy	0.897	0.988	0.995	0.992	0.990	0.988
Sensitivity	1.000	0.993	1.000	0.998	0.996	0.993
Specificity	0.587	0.973	0.980	0.973	0.973	0.973

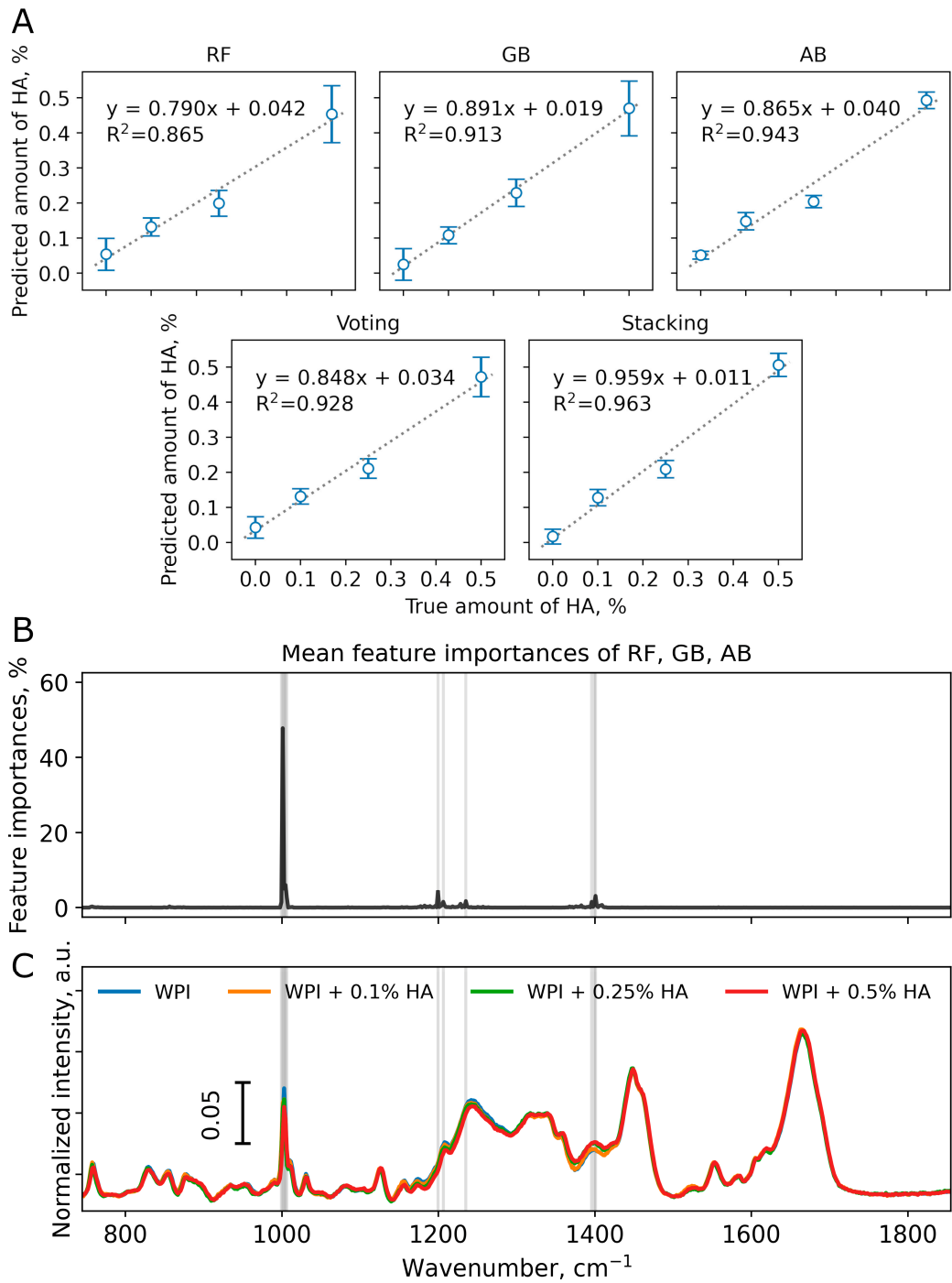


Figure 4. (a) Linear regression plots between true and predicted HA amounts calculated for test dataset (150 spectra per HA amount). Corresponding regression equations with R^2 values are indicated for each regression

model. (b) Mean feature importances of RF, GB, and AB. Wavenumbers with importance greater than 1% are marked with vertical grey lines. (c) Normalized mean spectra of WPI, WPI + 0.1% HA, WPI + 0.25% HA, WPI + 0.5% HA.

According to feature importance plots, the wavenumbers with most impact on the regression models were found to be 1003 cm^{-1} (phenylalanine, ring breath), 1206 cm^{-1} (C–C stretching), 1240 cm^{-1} (amide III (β -sheet), N–H in-plane bend, C–N stretch), and 1399 cm^{-1} (aspartic and glutamic acids, C=O stretch of COO[−]) (Figure S2). In most of regression models discussed, all four spectra groups were clearly distinguishable. However, the best regression model for the discussed dataset was found to be Stacking regressor ($R^2 = 0.963$).

4. Conclusions

A thorough analysis of the implementation of ensemble models in Raman spectroscopy data was performed. The dataset consisted of WPI (5 wt. %), WPI + 0.1% HA (5 wt. % : 0.1 wt. %, 50:1 ratio), WPI + 0.25% HA (5 wt. % : 0.25 wt. %, 20:1 ratio), WPI + 0.5% HA (5 wt. % : 0.5 wt. %, 10:1 ratio) Raman spectra (600 spectra per sample type).

Classification models were used to answer the question of whether it is possible to detect differences between pure protein (WPI) and its conjugate with polysaccharide (WPI + HA). Prior to solving, Random Forest, Gradient Boosting, AdaBoost models were optimized in number of decision trees utilized and maximum depth of each decision tree. The optimal parameters were evaluated using grid search with cross-validation, and found to be 200 decision trees with maximum depth of 4. Out of six discussed models, namely Random Forest, Gradient Boosting, AdaBoost, Voting soft and hard, and Stacking, AdaBoost performed the best with overall accuracy of 99.5%, specificity of 98% and sensitivity of 100%. Moreover, RF, GB, and AB models provided additional information on the Raman bands of most importance in obtained predictions: 1003 cm^{-1} (ring breath of phenylalanine), 1206 cm^{-1} (C–C stretching), and 1399 cm^{-1} (aspartic and glutamic acids, C=O stretch of COO[−]).

Regression models were implemented not only to distinguish between Raman spectra of protein and its conjugate, but also to specifically determine HA amount. The best model in solving regression task was found to be Stacking regressor, which utilized Ridge model with cross-validation and RF, GB, and AB predictions as inputs. The determination coefficient of this model was found to be 0.963. RF, GB, and AB regression models had also determined additional important in analysis Raman band at 1240 cm^{-1} (amide III (β -sheet), N–H in-plane bend, C–N stretch).

Implementation of ensemble models in analysis of Raman spectroscopy could be perceived as an intermediate solution between classical machine learning and neural networks. The training process of such models is less time consuming; however, the obtained performance is quite high. The discussed approach could be implemented to analyses of various Raman spectroscopy data.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/doi/s1>, Table S1: Accuracies across 3 cross-validations of RF, GB, and AB models with different number of DT ($n_{\text{estimators}}$) and maximum depth of each DT (max_depth); Figure S1: Feature importances of RF, GB, and AB classification models (black plots). Wavenumbers with importance greater than 1% are marked with vertical grey lines. Normalized mean spectra of WPI (blue line) and WPI+HA (orange line); Figure S2: Feature importances of RF, GB, and AB regression models (black plots). Wavenumbers with importance greater than 1% are marked with vertical grey lines. Normalized mean spectra of WPI (blue line), WPI + 0.1% HA (orange line), WPI + 0.25% HA (green line), WPI + 0.5% HA (red line).

Author Contributions: Conceptualization, O.A.M. and E.S.P.; methodology, E.S.P.; software, E.S.P.; validation, E.S.P.; formal analysis, E.S.P.; investigation, O.A.M., M.S.S. and E.S.P.; resources, E.S.P.; data curation, E.S.P.; writing—original draft preparation, E.S.P.; writing—review and editing, O.A.M., M.S.S. and E.S.P.; visualization, E.S.P.; supervision, E.S.P.; project administration, E.S.P.; funding acquisition, E.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Russian Ministry of Science and Higher Education (state assignment no. FSMG-2025-0054).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AB	AdaBoost
DT	Decision tree
GB	Gradient boosting
HA	Hyaluronic acid
RF	Random forest
WPI	Whey protein isolate

References

1. Vaou, N.; Stavropoulou, E.; Voidarou, C. (Chrysa); Tsakris, Z.; Rozos, G.; Tsigalou, C.; Bezirtzoglou, E. Interactions between Medical Plant-Derived Bioactive Compounds: Focus on Antimicrobial Combination Effects. *Antibiotics* **2022**, *11*, 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**, *12*, 1424. <https://doi.org/10.3390/app12031424>.
3. Senthilkumar, K.; Vijayalakshmi, A.; Jagadeesan, M.; Somasundaram, A.; Pitchiah, S.; Gowri, S.S.; Ali Alharbi, S.; Javed Ansari, M.; Ramasamy, P. Preparation of self-preserving personal care cosmetic products using multifunctional ingredients and other cosmetic ingredients. *Sci. Rep.* **2024**, *14*, 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**, *26*, 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**, *12*. <https://doi.org/10.3389/fcimb.2022.866463>.
6. Pezzotti, G. Raman spectroscopy in cell biology and microbiology. *J. Raman Spectrosc.* **2021**, *52*, 2348–2443. <https://doi.org/10.1002/jrs.6204>.
7. Kočíšová, E.; Kuižová, A.; Procházka, M. Analytical applications of droplet deposition Raman spectroscopy. *Analyst* **2024**, *149*, 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**, *144*, 19651–19667. <https://doi.org/10.1021/jacs.2c05359>.
9. Koronaki, E.D.; Kaven, L.F.; Faust, J.M.M.; Kevrekidis, I.G.; Mitsos, A. Nonlinear manifold learning determines microgel size from Raman spectroscopy. *AIChE J.* **2024**, *70*. <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**, *577*, 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**, *47*, 100910. <https://doi.org/10.1016/j.cofs.2022.100910>.
12. Becker, T.; Rousseau, A.-J.; Geubbelmans, M.; Burzykowski, T.; Valkenborg, D. Decision trees and random forests. *Am. J. Orthod. Dentofac. Orthop.* **2023**, *164*, 894–897. <https://doi.org/10.1016/j.ajodo.2023.09.011>.

13. da Silva, L.P.; Oliveira, M.D.L.; Villa, J.E.L. A comparison of decision tree-based algorithms for food discrimination using vibrational spectroscopy. *Food Chem.* **2025**, *488*, 144909. <https://doi.org/10.1016/j.foodchem.2025.144909>.
14. Sevetlidis, V.; Pavlidis, G. Effective Raman spectra identification with tree-based methods. *J. Cult. Herit.* **2019**, *37*, 121–128. <https://doi.org/10.1016/j.culher.2018.10.016>.
15. 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**, *237*, 121549. <https://doi.org/10.1016/j.eswa.2023.121549>.
16. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>.
17. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
18. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
19. Wang, M.; Zhang, J. Surface Enhanced Raman Spectroscopy Pb²⁺ Ion Detection Based on a Gradient Boosting Decision Tree Algorithm. *Chemosensors* **2023**, *11*, 509. <https://doi.org/10.3390/chemosensors11090509>.
20. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. <https://doi.org/10.1006/jcss.1997.1504>.
21. Zhu, J.; Zou, H.; Rosset, S.; Hastie, T.; others Multi-class adaboost. *Stat. Interface* **2009**, *2*, 349–360.
22. Dietterich, T.G. Ensemble Methods in Machine Learning. In; 2000; pp. 1–15.
23. Wolpert, D.H. Stacked generalization. *Neural Networks* **1992**, *5*, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
24. 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**, *52*, 1441–1453. <https://doi.org/10.1080/01496395.2017.1287197>.
25. 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**, *47*, 51–66. <https://doi.org/10.1080/10408347.2016.1176889>.
26. Niafizafar, M.; Besharati, M.; Jabbar, M.; Ghazanfar, S.; Asad, M.; Palangi, V.; Eseceli, H.; Lackner, M. Slow-release non-protein nitrogen sources in animal nutrition: A review. *Heliyon* **2024**, *10*, e33752. <https://doi.org/10.1016/j.heliyon.2024.e33752>.
27. Alizadeh Sani, M.; Jahed-Khaniki, G.; Ehsani, A.; Shariatifar, N.; Dehghani, M.H.; Hashemi, M.; Hosseini, H.; Abdollahi, M.; Hassani, S.; Bayrami, Z.; et al. Metal–Organic Framework Fluorescence Sensors for Rapid and Accurate Detection of Melamine in Milk Powder. *Biosensors* **2023**, *13*, 94. <https://doi.org/10.3390/bios13010094>.
28. 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**, *29*, 781. <https://doi.org/10.3390/molecules29040781>.
29. 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**, *94*, 331–340. <https://doi.org/10.1016/j.foodcont.2018.07.004>.
30. Marinho, A.; Nunes, C.; Reis, S. Hyaluronic Acid: A Key Ingredient in the Therapy of Inflammation. *Biomolecules* **2021**, *11*, 1518. <https://doi.org/10.3390/biom11101518>.
31. Yasin, A.; Ren, Y.; Li, J.; Sheng, Y.; Cao, C.; Zhang, K. Advances in Hyaluronic Acid for Biomedical Applications. *Front. Bioeng. Biotechnol.* **2022**, *10*. <https://doi.org/10.3389/fbioe.2022.910290>.
32. 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**, *26*, 4429. <https://doi.org/10.3390/molecules26154429>.
33. 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**, *24*, 10296. <https://doi.org/10.3390/ijms241210296>.

34. 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**, *225*, 1085–1095. <https://doi.org/10.1016/j.ijbiomac.2022.11.170>.
35. Henderson, A. Renishaw File Reader 2017.
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Mayorova, O.A.; Saveleva, M.S.; Bratashov, D.N.; Prikhozhenko, E.S. Combination of Machine Learning and Raman Spectroscopy for Determination of the Complex of Whey Protein Isolate with Hyaluronic Acid. *Polymers (Basel)*. **2024**, *16*, 666. <https://doi.org/10.3390/polym16050666>.
38. Shevtsova, S.A.; Saveleva, M.S.; Mayorova, O.A.; Prikhozhenko, 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. *Izv. Saratov Univ. Phys.* **2025**, *25*, 305–315. <https://doi.org/10.18500/1817-3020-2025-25-3-305-315>.
39. Kurniasih, A.; Previana, C.N. Implementation of GridSearchCV to Find the Best Hyperparameter Combination for Classification Model Algorithm in Predicting Water Potability. *J. Artif. Intell. Eng. Appl.* **2025**, *4*, 1174–1182. <https://doi.org/10.59934/jaiea.v4i2.844>.
40. 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**, *5*, 86–91. <https://doi.org/10.52465/josce.v5i1.308>.
41. Zhang, S.; Zhang, Z.; Lin, M.; Vardhanabhuti, B. Raman Spectroscopic Characterization of Structural Changes in Heated Whey Protein Isolate upon Soluble Complex Formation with Pectin at Near Neutral pH. *J. Agric. Food Chem.* **2012**, *60*, 12029–12035. <https://doi.org/10.1021/jf303780c>.
42. Zhao, Y.; Ma, C.-Y.; Yuen, S.-N.; Phillips, D.L. Study of Succinylated Food Proteins by Raman Spectroscopy. *J. Agric. Food Chem.* **2004**, *52*, 1815–1823. <https://doi.org/10.1021/jf030577a>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.