

Data augmentation of Raman spectra using the principal component analysis

© E.S. Prikhodzhenko

Moscow Institute of Physics and Technology (National Research University), Future Biophysics Institute, Laboratory of Medical Equipment in the Field of *in vitro* Diagnostics, Dolgoprudny, Moscow Region, Russia

e-mail: prikhodzhenko.es@mipt.ru

Received November 21, 2025

Revised December 17, 2025

Accepted December 17, 2025

An approach to the generation of synthetic Raman spectra using the principal component method is proposed. The Raman spectra obtained for adipose tissue before and after lipase exposure are used as initial data. The proposed data are characterized by the similarity of the spectra, since the studied materials have a similar chemical structure: adipose tissue consists of triglyceride molecules. After exposure to lipase, a mixture of di- and monoglycerides, as well as free fatty acids, is formed. The approach proposed in the work to generate spectral data is as follows: (1) applying the principal component analysis (PCA) to a set of real spectral data, (2) calculating the average value and standard deviation of the scores of a selected number of principal components (PC), (3) generating a new set of PC scores based on the average value and standard deviation, (4) generating spectra according to the PC loadings of the selected number of main components and the new set of PC scores. To confirm the similarity of the initial and synthetic data, a random forest classification model was trained on the initial data, followed by an accuracy assessment on the synthetic data. The proposed approach made it possible to perform augmentation process of Raman spectroscopy data with an average accuracy of $90.6 \pm 0.5\%$.

Keywords: Raman spectroscopy, data augmentation, principal component analysis, machine learning, random forest.

DOI: 10.61011/EOS.2026.05.63858.53-25

Introduction

Raman scattering (RS) spectroscopy is a powerful tool that is widely used in analytical chemistry, biomedicine and materials science, and provides detailed information about the molecular composition and structure of a substance [1,2]. However, the use of machine learning methods for RS spectra analysis is often associated with an issue of test data deficiency. This is due to the fact that data acquisition requires significant time and financial expenditures [3,4]. Data deficiency can lead to statistical unreliability of models, model re-training and reduced accuracy of predictions.

To overcome data deficiency in the RS spectroscopy, a wide range of methods has been developed that can be divided into two categories: simple deterministic and complex generative. Traditional approaches include elementary transformations such as addition of the Gaussian noise to the initial spectrum for instrument error simulation [5]. Affine transformations such as baseline shift, change of slope and intensity are also used to consider the test condition variability [6]. Although these methods are easy to implement and do not require large computational resources, they have a substantial disadvantage: generated data represent only small variations of initial samples and cannot significantly expand the model determination domain because they do not create any brand new spectral features.

Revolution in the field of data generation was achieved due to the use of deep learning, in particular, of generative

adversarial networks (GAN) [7] and variational autoencoders (VAE) [8]. These models can adapt to complex and diverse initial data distributions and generate high quality synthetic spectra which are visually indistinguishable from real ones [9,10]. However, despite the impressive results, these approaches have substantial disadvantages. They require significant computational capabilities and large amounts of data for learning, making them not always available. In addition, they are hard to tune and require special learning stabilization methods [7].

Principal component analysis is used to expand a multi-dimensional dataset in a set of successive orthogonal components (loads) which explain the maximum data variance percentage [11]. In the RS spectroscopy, the principal component analysis is used to decrease data dimensionality, effectively separating the useful signal (contained in the first principal components) from noise (in the last components) [12]. The principal component analysis is also used to identify patterns in spectroscopy data with focus on vibrational modes that are more typical of the given dataset [13,14]. This approach can be used for graphical RS spectra clustering through imaging each individual spectrum as a point in space of selected principal components: scores are coordinates in this case [15].

This work describes a new approach to generation of synthetic RS spectra on the basis of the principal component analysis. Real fat tissue spectra before ($n = 100$) and after ($n = 40$) the lipase-catalyzed enzymatic hydrolysis were used to prove the performance of this method.

This model system has a high spectral similarity which is associated with a closely related chemical structure of the given materials. The initial fat tissue predominantly consists of triglycerides [16], and its breakdown product is a complex mixture of diglycerides, monoglycerides and free fatty acids [17–19]. Thus, verification of the proposed method for spectral data generation on complex and similar biotissue spectra using a fat tissue before and after the lipase action is of great interest.

The proposed procedure includes the following steps: 1) applying the principal component analysis to the initial set of spectra to extract principal components (loadings) and corresponding scores, 2) statistical analysis of the calculated scores with determining their mean values and standard deviations, 3) generating new synthetic scores by random sampling from the normal distribution with the found parameters, 4) reconstruction of the full synthetic spectrum by linear combination of the principal component loadings and generated scores. The work describes comparative analysis data and evaluates the similarity between synthesized and real datasets using machine learning methods.

1. Methods

1.1. Materials and methods

Lipase B *Candida antarctica* was purchased from Sigma Aldrich (Saint Louis, Missouri, USA). Study of fat tissue taken *post-mortem* from subcutaneous fat tissue was approved by the Ethics Committee of Saratov State Medical University named after V.I. Razumovsky (minutes № 1 as of 03.09.2013).

1.2. RS spectra recording

The Renishaw inVia (UK) confocal RS microscope equipped with 785 nm laser was used for measurements. Laser beam was transmitted through the Leica N PLAN L 50× lens with a numerical aperture of 0.5. SynchroScan mode in WiRE 4.2 continuous scanning software was used to obtain RS spectra ranging from 150 to 3200 cm⁻¹.

The raw fat tissue (initial sample) was examined using 3 mW laser radiation. The number of RS spectra was equal to 100. After fat tissue processing in lipase B *Candida antarctica* (0.75 mg/ml, aqueous solution) for 30 min at 37°C, the sample represented a micellar solution of fatty acids in water. RS spectra of the processed sample were measured after cooling the sample to room temperature. Laser radiation power was 15 mW, number of spectra was 40. RS integration time for both types of samples was 10 s.

1.3. Data pre-processing

Polynomial background described by an equation of the 11th degree was removed from the measured spectra using the Subtract Baseline tool included in WiRE 4.2. Further data processing and employment of machine learning methods

were performed in Python version 3 in the *Jupyter Notebook* using *scikit-learn* (*sklearn*) library version 1.7.2 [20]. The *sklearn.preprocessing.normalize* method was applied to the dataset for data normalization using the l^2 norm:

$$||x||_2 = \sqrt{\sum_{i=1}^m |x_i|^2},$$

where $||x||_2$ is the divisor when using the l^2 norm, x_i is the RS intensity at the i th wavenumber, m is the number of wavenumbers in the spectrum ($m = 3047$ for the given dataset). Thus, the intensities were divided by the obtained divisor ($||x||_2$) for each spectrum. As a result, normalized intensities were calculated.

1.4. RS spectra generation

The principal component analysis was implemented using the *sklearn.decomposition.PCA* method with varying $n_{components}$ within 5–130 in increments of 5. *fit_transform* was used to transform the 140×3047 initial spectral data array to the $140 \times n_{components}$ array. This array contained the desired coefficients for the linear combination (scores).

Mean values and standard deviations of scores were further calculated. Then, using the *numpy* library and *random.normal* method, 1000 scores each were generated with respect to the mean and standard deviation for fat tissue spectra before and after the lipase action.

The generated scores and principal component (loadings) spectra were used for *inverse_transform*. Thus, sets of synthetic RS spectra were obtained for the fat tissue before and after the lipase catalysis. Dimensions of the generated data arrays were 1000×3047 .

1.5. Evaluating the generation using a random forest classification model

Real data were preliminary divided into the training and test samples at the ratio of 3:1 with a maintained proportion between the spectra before and after the lipase catalysis using *scikit-learn* (*sklearn*) and *model.selection.train_test_split*. *sklearn.ensemble.RandomForestClassifier*(*max_depth*=3, *n_estimators*=10) was used as a classification model [21]. Accuracy (*sklearn.metrics.accuracy_score*), precision (*sklearn.metrics.precision_score*), recall (*sklearn.metrics.recall_score*) and confusion matrix (*sklearn.metrics.confusion_matrix*) were used as metrics. Due to the confusion matrix, a percentage of correctly classified synthetic datasets for the fat tissue before and after the lipase catalysis was obtained.

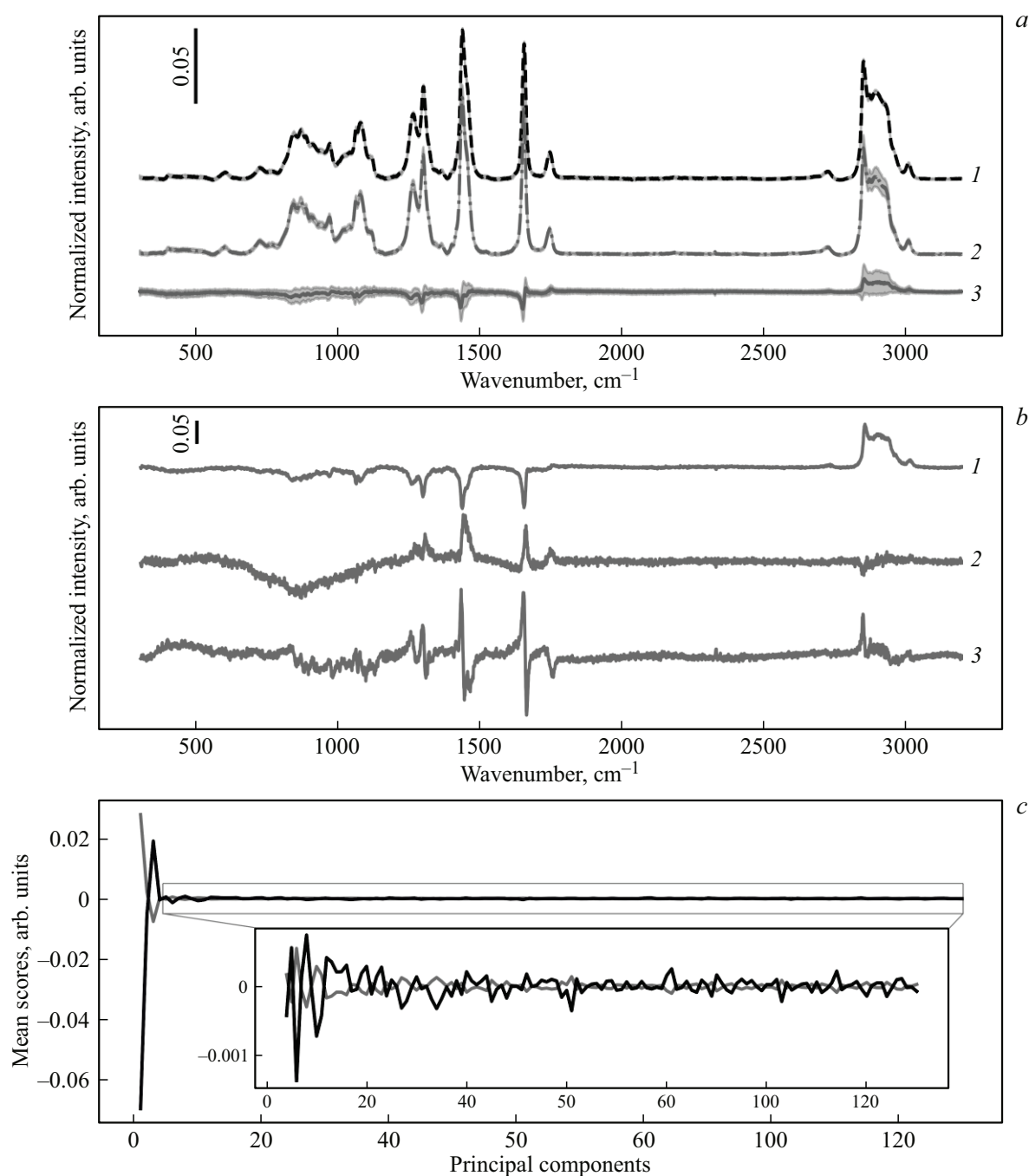


Figure 1. *a*) Averaged normalized RS spectra of the fat tissue: 1 — before lipase action ($n = 100$), 2 — after lipase action ($n = 40$), 3 — difference spectrum, the fill area characterizes the standard deviation of normalized intensities. *b*) — Loadings of the first three principal components from the given dataset. *c*) Mean principal component scores for fat tissue spectra before (grey curve) and after (black curve) the lipase catalysis, the inset shows the curves for principal components starting from the fourth one.

2. Results and discussion

2.1. Initial dataset description

To demonstrate the data generation approach in the RS spectroscopy, a set of previously measured spectra corresponding to the fat tissue before ($n = 100$) and after ($n = 40$) the lipase action was selected [22]. Thus, the initial

dataset is a 140×3047 table where 140 corresponds to the number of individual spectra and 3047 corresponds to a length of each spectrum. Averaged normalized spectra and spectra difference are shown in Figure 1, *a*. RS spectra of the given samples have minor difference, which, in turn, makes them perfect candidates for testing the proposed data generation strategy. The principal component analysis was proposed as the key tool [15]. When using this method,

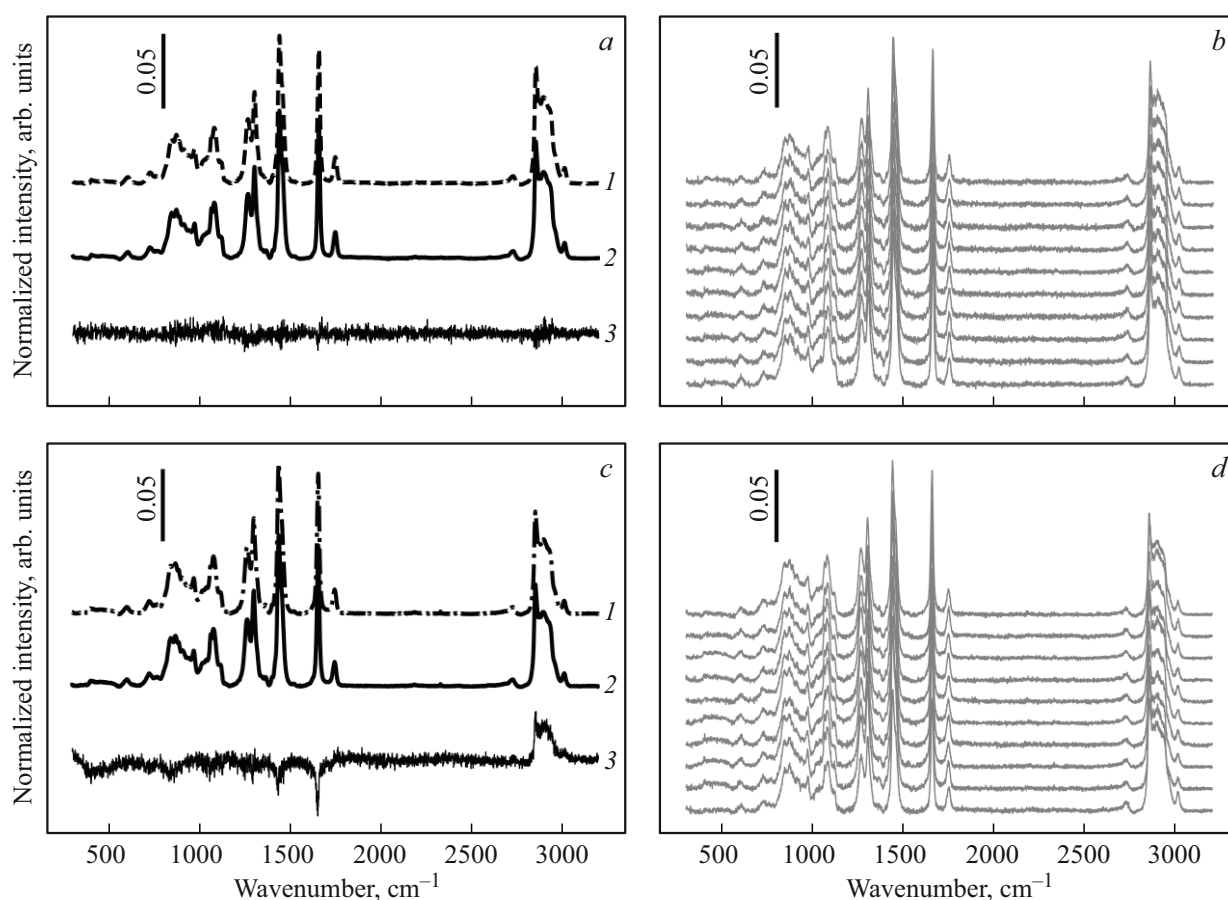


Figure 2. RS spectra generation using the principal component method ($n_{components} = 130$) for fat tissue samples before (*a, b*) and after (*c, d*) lipase action: (*a, c*) 1 — averaged RS spectrum of real data; 2 — averaged RS spectrum of synthetic data ($n = 1000$); 3 — difference of averaged spectra of real and synthetic data with 100x magnification for clarity. (*b, d*) Ten random spectra from the synthetic data set.

100% data variances appeared to be explained by 139 principal components. Loadings corresponding to the first three principal components are shown in Figure 1, *b*. These components explain 53.3%, 5.1% and 4.2% data variances, respectively. When implementing the principal component analysis, 140×3047 initial data are transformed into the $140 \times$ number of principal components array. In the thus obtained array, each sample is characterized by its own set of scores in a linear combination of calculated principal components, which allows the mean scores for the fat tissue before and after the lipase action to be determined and plotted (Figure 1, *c*). Thus, in terms of the principal component analysis, spectra of the given samples have obvious differences.

Besides the direct transformation, reduction of the initial data dimensions in the given case from 3047 wavenumbers to 139 principal components, an inverse operation can be performed. Thus, by using the principal component analysis on a RS spectroscopy dataset, calculating typical loadings and corresponding score distribution, an additional set of scores can be generated and inverse

transformation into synthetic RS spectra can be performed.

2.2. Synthetic dataset generation using the principal component analysis

For principal component score generation, this study used an assumption that principal component scores corresponding to a certain class of samples (fat tissue before and after the lipase action) are characterized by a normal distribution. Thus, by employing *numpy* and *numpy.random.normal*, and indicating the mean and standard deviations of scores, additional scores can be obtained for the linear combination of principal component loadings followed by reconstruction to synthetic RS spectra. Two parameters affect the generation process: *random_state* or *numpy, seed* — any integer allowing random variables to be generated in *Python3* (it is important to have a constant value throughout the code); *n_components* is the number of principal components.

For fat tissue before and after the lipase action, 1000 RS spectra each were generated with a different number of principal components. 5–130 in increments of 5. Spectra

generation results for $n_components = 130$ are shown in Figure 2. On the curves of difference between the averaged RS spectra of real and synthetic data (curves 3 in Figure 2, *a, c*), it can be seen that when using the fat tissue spectra before the lipase action (Figure 2, *a*), the generation gave results which were quite close to real data: the difference curve with 100x magnification represented random noise that did not contain individual RS bands. Similarity between the synthetic dataset for the fat tissue after lipase action and the real prototype is less pronounced, the difference spectrum between the real and generated data with 100x magnification has some RS bands: in particular, 1456 cm^{-1} (CH_2 stretching), 1657 cm^{-1} ($\text{C}=\text{C}$ stretching) and in the $2800\text{--}3000\text{ cm}^{-1}$ area (C-H stretching) [23,24].

This effect can be probably caused by large standard deviations compared with the fat tissue sample before the lipase action (spectrum 2, Figure 1, *a*). Simplicity of the proposed approach and no need for additional spectra generation steps, in particular, random noise introduction, should be also noted. Thus, when using 130 components for the generation process, the obtained RS spectra are characterized by a noticeable noise level (Figure 2, *b, d*). The effect of the number of principal components on the resulting spectrum generation is more vividly shown in Figure 3. This figure shows the generation of an RS spectrum with a different number of principal components. 10–130 in increments of 10. Noise enhancement in the synthesized spectra as the number of principal components increases is particularly noticeable in the case of fat tissue spectra generation before the lipase catalysis (Figure 3, *a*). Figure 4, *a* shows loadings for some principal components. Noise on the shown curves occurs starting from the second principal component. Starting from component 20, the load curve predominantly contains noise. The explained data variance percentage depending on the number of principal components is shown in Figure 4, *b*.

Since each next component is responsible for a lower explained data variance percentage, their effect on the RS spectra generation process is nonuniform (Figure 4, *b*). It is also important to understand a data loss percentage that can be neglected in the generation. Thus, for the real dataset of interest with $n_components = 105$, a loss percentage is less than 5%. If a higher accuracy is required, a larger percentage of stored data set information, with $n_components = 120$ the loss percentage is less than 2.5%, and with $n_components = 130$ the loss is less than 1%. If the number of components is not an important criterion, and the stored data variance percentage is more informative, then the former can be specified directly when using the *decomposition.PCA* module from the *sci-kit learn* library in Python3. Thus, with $n_components = 0.95$ a number of principal components will be selected such that the explained variance percentage exceeds 95%.

Evaluation of dataset similarity using the classification model

To evaluate the synthetic and real dataset similarity, the random forest classification model was used [21,25]. This is an ensemble machine learning model that combines the predictions of multiple decision trees working in parallel. Each decision tree is a hierarchical structure, and the final ensemble prediction is formed via majority voting. The main random forest model parameters are the number of trees used ($n_estimators$) and „branching“ of each tree (max_depth). This work used the random forest model with the following parameters: $n_estimators = 10$, $max_depth = 3$.

For model training, only real data were used, for which they had been preliminary divided into the training and test samples at the ratio of 3 : 1. Model prediction accuracy on the test sample was 100%. The obtained model was then used for synthetic dataset classification. For similarity evaluation, metrics were used such as accuracy, percentage of correctly classified synthetic data for the fat tissue before ($n = 1000$) and after ($n = 1000$) lipase action, and percentages of correctly classified synthetic data separately. Results are listed in the table.

The calculated accuracy of synthetic dataset classification was limited by RS spectra simulating the fat tissue signal after lipase action. Mean percentage of correctly classified data in this category was $(81.3 \pm 1.0)\%$ regardless of the number of principal components used for generation. While the maximum percentage reached 83.4% with $n_components = 95$ and an explained variance percentage of 93.1%. Total accuracy for two sets of generated RS spectra was $(90.6 \pm 0.5)\%$.

For the shown dataset, a weak dependence of spectra generation accuracy on the number of principal components is observed. Starting from fifteen principal components, synthetic RS spectra of the fat tissue after the lipase action were classified correctly with an accuracy more than 80%. The explained variance percentage and noise in synthetic spectra are important parameters rather than the number of principal components. Thus, in terms of classification accuracy, an explained variance percentage of 70% (number of components more than 15) is sufficient. However, for better simulation of noise in data, an explained variance percentage of 90% should be better used via *sklearn.decomposition.PCA* ($n_components = 0.9$) or higher up to 0.99.

Conclusions

A simple RS spectroscopy data generation method is proposed on the basis of real measurements and the principal component analysis. Using the principal component analysis, the real dataset is expanded in a linear combination of loadings with corresponding scores. New random scores were generated on the assumption about normal score distribution (with known mean values and standard

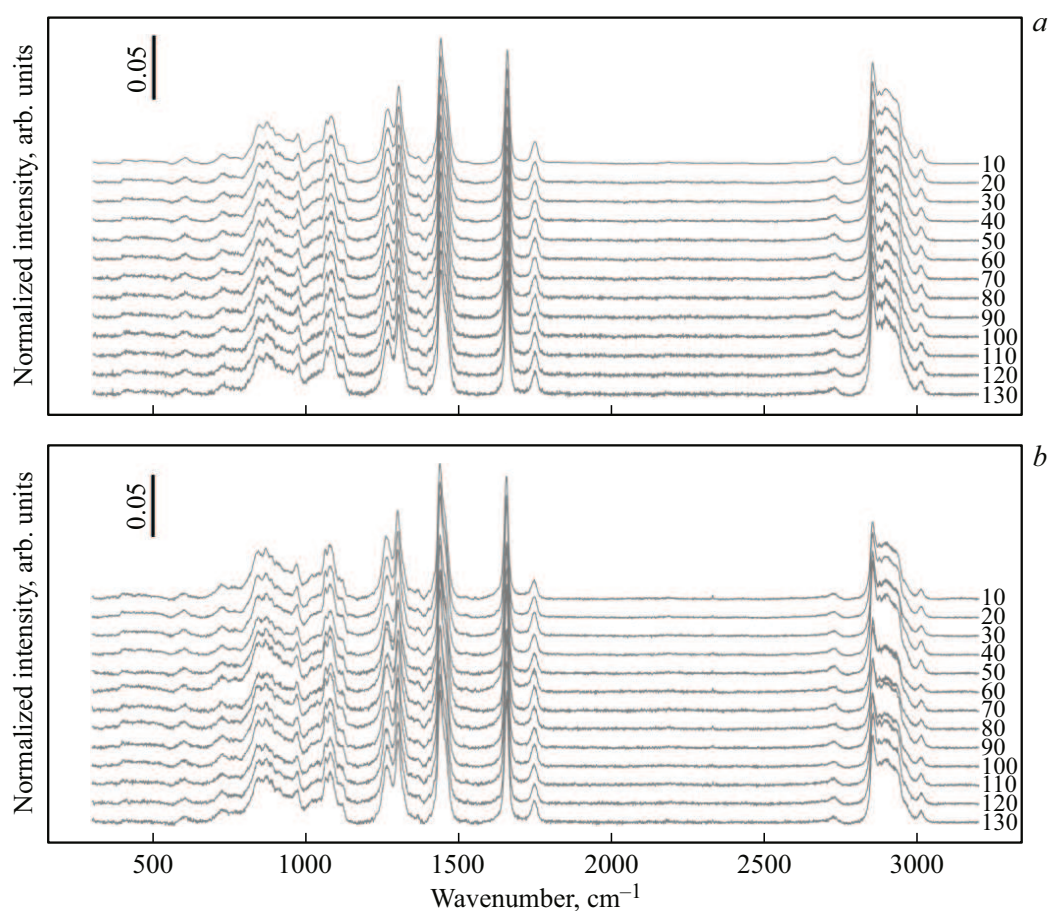


Figure 3. RS spectrum generation for the fat tissue before (a) and after (b) the lipase action with $n_{components}$ 10–130 ranging from 10 to 130.

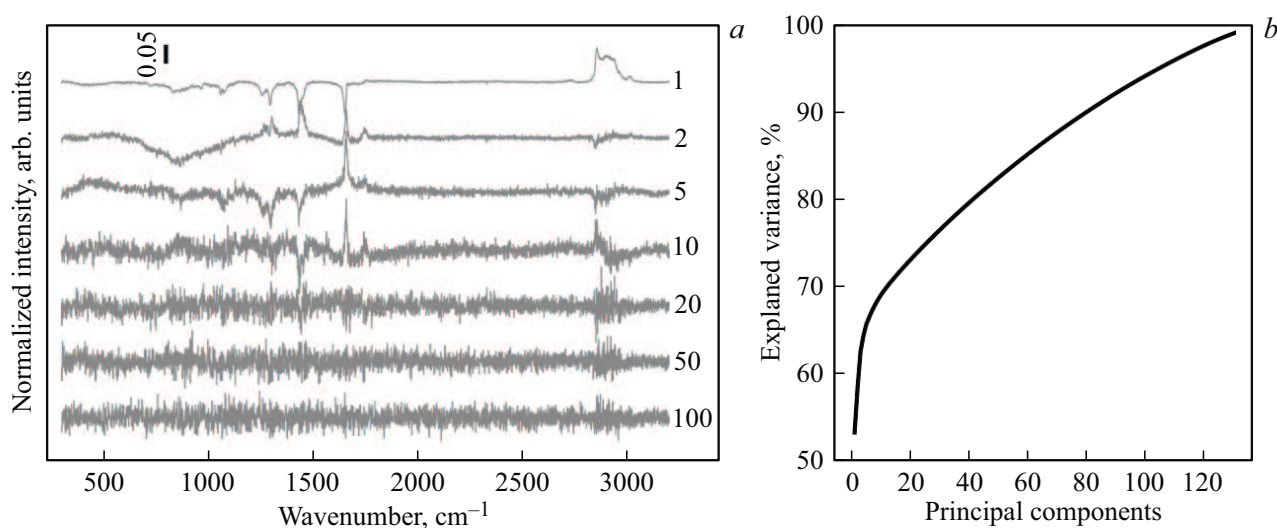


Figure 4. (a) Loading curves for the principal components (1, 2, 5, 10, 20, 50, 100) as shown in the figure; (b) cumulative curve of the explained variance percentage for 1–130 principal components.

deviations). The obtained synthetic set and loadings were used for inverse transformation. The proposed approach was tested on the RS spectroscopy data for the fat tissue before

and after the lipase action. Regardless of materials with similar chemical bond sets, the synthesized sets of spectra were different, which was demonstrated using the random

Correctly classified synthetic data percentages, accuracy, precision and recall depending on the number of principal components

Number of principal components	Explained data variance, %	Percentage of correctly classified synthetic spectra before lipase action, %	Percentage of correctly classified synthetic spectra after lipase action, %	Accuracy (accuracy), %	Precision (precision), %	Recall (recall), %
5	65.75	100.0	79.3	89.65	100.00	100.00
10	69.13	100.0	79.6	89.80	100.00	100.00
15	71.17	100.0	82.4	91.20	100.00	100.00
20	73.00	100.0	82.2	91.10	100.00	100.00
25	74.71	100.0	82.0	91.00	100.00	100.00
30	76.31	100.0	81.3	90.65	100.00	100.00
35	77.92	100.0	80.2	90.10	100.00	100.00
40	79.38	100.0	81.3	90.65	100.00	100.00
45	80.88	100.0	81.6	90.80	100.00	100.00
50	82.28	100.0	81.2	90.60	100.00	100.00
55	83.66	100.0	81.7	90.85	100.00	100.00
60	84.99	100.0	81.5	90.75	100.00	100.00
65	86.30	100.0	80.0	90.00	100.00	100.00
70	87.55	100.0	81.0	90.50	100.00	99.88
75	88.75	99.9	80.6	90.25	99.88	100.00
80	89.90	100.0	81.0	90.50	100.00	99.88
85	91.01	99.9	81.8	90.85	99.88	100.00
90	92.11	100.0	79.3	89.65	100.00	99.88
95	93.14	99.9	83.4	91.65	99.88	99.88
100	94.16	99.9	82.4	91.15	99.88	100.00
105	95.11	100.0	82.2	91.10	100.00	99.76
110	96.04	99.8	82.4	91.10	99.76	99.76
115	96.99	99.8	81.8	90.80	99.76	100.00
120	97.80	100.0	80.9	90.45	100.00	99.88
125	98.53	99.9	80.6	90.25	99.88	99.76
130	99.18	99.8	82.1	90.95	99.76	100.00

forest classification model. The mean accuracy for synthetic data for the fat tissue before and after the lipase action was $(99.9 \pm 0.1)\%$ and $(81.3 \pm 1.0)\%$, respectively. The total generation accuracy was $(90.6 \pm 0.5)\%$.

Funding

The study was supported by the Ministry of Science and Higher Education of the Russian Federation (state order topic FSMG-2025-0054).

Conflict of interest

The author declares no conflict of interest.

References

- [1] G. Pezzotti. *J. Raman Spectrosc.*, **52**, 2348.2443 (2021). DOI: 10.1002/jrs.6204
- [2] D. Cialla-May, C. Krafft, P. Rösch, T. Deckert-Gaudig, T. Frosch, I.J. Jahn, S. Pahlow, C. Stiebing, T. Meyer-Zedler, T. Bocklitz, I. Schie, V. Deckert, J. Popp. *Anal. Chem.*, **94**, 86–119 (2022). DOI: 10.1021/acs.analchem.1c03235
- [3] J.S. Prell, J.T. O'Brien, E.R. Williams. *J. Am. Soc. Mass Spectr.*, **21**, 800–809 (2010). DOI: 10.1016/j.jasms.2010.01.010
- [4] L. Gao, R.T. Smith. *J. Biophotonics*, **8**, 441–456 (2015). DOI: 10.1002/jbio.201400051
- [5] A. Conlin, E. Martin, A. Morris. *Chemom. Intell. Lab. Syst.*, **44**, 161–173 (1998). DOI: 10.1016/S0169-7439(98)00071-9
- [6] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø. *Chemom. Intell. Lab. Syst.*, **118**, 62–69 (2012). DOI: 10.1016/j.chemolab.2012.07.010
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. *Commun. ACM*, **63**, 139–144 (2020). DOI: 10.1145/3422622
- [8] D. Papadopoulos, V.D. Karalis. *Appl. Sci.*, **13**, 8793 (2023). DOI: 10.3390/app13158793
- [9] S. Yu, H. Li, X. Li, Y.V. Fu, F. Liu. *Sci. Total Environ.*, **726**, 138477 (2020). DOI: 10.1016/j.scitotenv.2020.138477
- [10] Y. Du, D. Han, S. Liu, X. Sun, B. Ning, T. Han, J. Wang, Z. Gao. *Talanta*, **237**, 122901 (2022). DOI: 10.1016/j.talanta.2021.122901
- [11] A. Coccato, M.C. Caggiani. *J. Raman Spectrosc.*, **55**, 125–147 (2024). DOI: 10.1002/jrs.6621
- [12] J.E. Guimares, R. Nadas, W. Zhang, T. Endo, K. Watanabe, T. Taniguchi, R. Saito, Y. Miyata, A. Jorio. *Phys. St. Solidi*, 2500291 (2025). DOI: 10.1002/pssb.202500291
- [13] C. Huang, X. Xu, J. Fu, D.-G. Yu, Y. Liu. *Polymers*, **14**, 3266 (2022). DOI: 10.3390/polym14163266

- [14] N. Jin, Y. Song, R. Ma, J. Li, G. Li, D. Zhang. *Anal. Chim. Acta*, **1197**, 339519 (2022). DOI: 10.1016/j.aca.2022.339519
- [15] A. Massei, N. Falco, D. Fissore. *Eur. J. Pharm. Biopharm.*, **200**, 114342 (2024). DOI: 10.1016/j.ejpb.2024.114342
- [16] D.R. Body. *Prog. Lipid Res.*, **27**, 39–60 (1988). DOI: 10.1016/0163-7827(88)90004-5
- [17] U.M. Lankage, S.A. Holt, S. Bridge, B. Cornell, C.G. Cranfield. *ACS Appl. Mater. Interfaces*, **15** (45), 52237–52243 (2023). DOI: 10.1021/acsami.3c11767
- [18] J. Xuan, Z. Wang, Q. Xia, T. Luo, Q. Mao, Q. Sun, Z. Han, Y. Liu, S. Wei, S. Liu. *Foods*, **11**, 3664 (2022). DOI: 10.3390/foods11223664
- [19] H. hai Wang, Q. Zhang, X. Yu, J. Liang, Y. Zhang, Y. Jiang, W. Su. *Ind. Eng. Chem. Res.*, **62**, 15733–15751 (2023). DOI: 10.1021/acs.iecr.3c02132
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. *J. Mach. Learn. Res.*, **12**, 2825–2830 (2011).
- [21] L. Breiman/ *Random Forests Mach. Learn.*, **45**, 5–32 (2001). DOI: 10.1023/A:1010933404324
- [22] I.Y. Yanina, Y.I. Svenskaya, E.S. Prikhodzhenko, D.N. Bratashov, M.V. Lomova, D.A. Gorin, G.B. Sukhorukov, V.V. Tuchin. *J. Biophotonics*, e201800058 (2018). DOI: 10.1002/jbio.201800058
- [23] F. Gao, D. Ben-Amotz, S. Zhou, Z. Yang, L. Han, X. Liu. *LWT*, **134**, 110105 (2020). DOI: 10.1016/j.lwt.2020.110105
- [24] N.N. Yazgan Karacaglar, T. Bulat, I.H. Boyaci, A. Topcu. *J. Food Drug Anal.*, **27**, 101–110 (2019). DOI: 10.1016/j.jfda.2018.06.008
- [25] M. Poth, G. Magill, A. Filgertshofer, O. Popp, T. Großkopf. *J. Raman Spectrosc.*, **53**, 1580–1591 (2022). DOI: 10.1002/jrs.6402

Translated by E.Ilyinskaya