UDC 535.37:539.23:004.8

Yu. Bilak

MODELING, OPTIMIZATION AND AI-FORECASTING TECHNOLOGY IN RAMAN SPECTROMETRY

Yuriy Bilak

Uzhgorod National University,
https://orcid.org/0000-0001-5989-1643

yuriy.bilak@uzhnu.edu.ua

In the work, the technology of modeling, optimization and prediction of spectral characteristics of thin films based on Raman spectroscopy with the use of artificial intelligence was developed. Modern machine learning methods are implemented, including ensemble algorithms (random forest, gradient boosting) and neural networks, which ensures high accuracy of forecasts and automation of spectrum analysis. An innovative approach includes the use of the Voight profile, which combines Lorentzian and Gaussian components, allowing to describe accurately the width and shape of the peaks for approximation, taking into account the physicochemical parameters of the films and the influence of experimental conditions. The task is caused by the growing requirements for the accuracy of material analysis in the fields of optoelectronics, photocatalysis and sensor systems, where Raman spectroscopy is an indispensable tool. Traditional data processing methods are limited by the complexity of the interaction of light with the material, and the integration of AI allows you to overcome these difficulties, optimizing analysis and prediction. The proposed technology combines physical modeling of spectra with AI-prediction, which allows accurate consideration of the effects of defects, inhomogeneities, and absorption. Algorithms for model optimization with minimization of root mean square error and selection of the best model for specific problems have been implemented. Additional optimization of the model takes into account the influence of the film thickness due to the absorption coefficient and the suppression of unwanted reactions with the help of buffer gases (Ne, Ar). The developed approach provides reduction of time and resources for experimental research, automation of spectrum analysis and development of new materials. The application of AI methods allows obtaining highly accurate results even with a small amount of experimental data. The prospects for technology development include the integration of multilayer structures, consideration of material anisotropy, and detailed modeling of defects in films. Additionally, it can be adapted to analyze different types of materials, such as organic films or hybrid structures. Software enhancements can include automation of spectrum fitting, optimization of film parameters, and machine learning-based property prediction with large data sets. This opens up new opportunities for research in the physico-chemistry of materials and the development of intelligent analysis systems.

Keywords: Raman spectroscopy, thin films, modelling, machine learning, spectral analysis, optimization, neural networks, physicochemical properties.

Introduction

Raman spectroscopy is one of the most powerful methods for studying molecular structures and properties of materials, especially thin films. Its uniqueness lies in the ability to obtain information about the chemical composition, structure, defects and other physical and chemical characteristics without destroying the sample. Thin films of metal oxides are widely used in optoelectronics, photocatalysis and sensor technologies

due to their unique properties. At the same time, the formation of such films is accompanied by the appearance of defects, inhomogeneities and other structural features, which complicates their analysis.

Traditional methods of processing spectral data often do not take into account the complexity of the interaction of light with the material, as well as the influence of experimental conditions. This limits the accuracy of the results and requires the development of new approaches that can integrate analytical models with modern forecasting methods such as machine learning. The paper proposes a complex technology for modeling Raman spectra of thin films, which combines physical modeling and optimization with prediction based on artificial intelligence.

The main problem in the analysis of Raman spectra of thin films is their high complexity caused by several factors. First, the intensity of Raman lines depends on the energy distribution of laser radiation and the molecular structure of the sample, including features such as defects and vacancies. Secondly, the thickness of the film and its heterogeneity significantly affect the absorption, scattering and diffusion of light, which makes it difficult to describe accurately the processes. Third, experimental conditions, such as the type of gas environment or noise level, introduce additional errors that are difficult to account for with traditional methods. Solving these problems requires developing models that are able to combine physical understanding of processes with machine learning algorithms to improve accuracy and speed of analysis. Optimizing film formation processes and minimizing defects is also an important aspect, which requires taking into account experimental conditions and film parameters. The approach proposed in the work integrates modern machine learning methods for modeling and forecasting the spectral characteristics of thin films, which was previously implemented mostly through analytical models. The uniqueness lies in the use of three different algorithms — random forest, gradient boosting and neural network — which provide flexibility in choosing the best model depending on the type of data and the complexity of dependencies. For the first time, the automated determination of the optimal model based on the minimization of the mean square error (MSE) was implemented, which allows to increase the accuracy of forecasts without the need for manual adjustment of parameters. The use of a neural network provides the possibility of considering nonlinear relationships in the data, which is especially important for complex film structures and their spectra.

The developed model is a universal tool for spectroscopic analysis in Raman spectroscopy, but it can also be adapted for interferometry, infrared analysis and other types of spectrometry. Its flexibility is provided by the integration of physical modeling and machine learning methods, which allows analyzing spectra even in difficult conditions. For interferometry, it is necessary to take into account the specifics of interference signals, and for infrared analysis, the specifics of molecular vibrations and absorption.

The model demonstrates high accuracy of forecasts and automation of data processing, which are key advantages for the analysis of multicomponent mixtures, determination of physicochemical parameters of materials and development of new materials. Applications include optoelectronics, chemistry, medicine, and sensor technology. Adaptation to other types of spectrometry is possible through the modification of mathematical models and algorithms, which allows to expand the limits of its application.

The practical significance of the proposed approach lies in the ability to predict accurately the spectral characteristics of materials, which significantly reduces time and resources for experimental research. Automating model selection simplifies data analysis for non-machine learning experts. The use of algorithms allows efficient work even with small data sets, which is typical for spectral studies. The implementation of such an approach can find application in the optimization of thin film production technolo-

gies, the development of new materials, and the analysis of physicochemical properties of samples in the fields of optoelectronics, photocatalysis, and sensor technologies.

Literature review

The literature review for the development of a complex model-based technology using machine learning for the analysis of Raman spectra covers various aspects such as spectral data processing, application of machine learning for analysis, and data visualization.

Article [1] highlights the current achievements of deep learning (DL) in spectral analysis, focusing on solving the problems of insufficient data, interpretability of models and adaptation of DL to the specifics of spectroscopy. Transfer learning, data augmentation, and adversarial networks are reviewed, and the importance of combining DL with expert knowledge to improve outcomes is highlighted. The work [2] analyzes the use of support vector methods (SVM) for the classification of near infrared radiation (NIR). Attention is focused on optimization of regularization parameters and kernels to reduce classification errors. Methods of visualization of support vectors in the subspaces of the main components are also proposed, which improves the interpretation of models. Work [3] investigates joint decision-making systems by humans and artificial intelligence through numerical channels, developing rules to take into account the reliability of channels and increase their number in the system. [4] presents a platform for creating algorithms for the classification of spectroscopic data using long-wave infrared (LWIR) and Raman spectroscopy. It is based on 1D convolutional neural networks, which demonstrate a classification accuracy of over 90 %, even under low signal-to-noise conditions. The article [5] focuses on the application of machine learning for the identification and analysis of illegal substances using Raman spectroscopy. Genetic algorithms are used to reduce the dimensionality of the data and neural networks are used to predict concentrations, which provides higher accuracy compared to standard methods. In [6] convolutional neural networks (CNN) are used for automatic classification of spectra without pre-processing. On the basis of RRUFF spectra, high accuracy is achieved, which exceeds the results of other popular methods, in particular, SVM.

Paper [7] analyzes the application of DL to model spectral data, highlighting its ability to reveal complex relationships that are not available to classical methods. The advantages of DL, including the automation of analysis and modeling, are described, and the limitations, such as the need for large amounts of data and the difficulty of interpreting the results, are discussed. The advances in machine learning in the chemical sciences, including molecular property prediction, synthesis optimization, and reactivity analysis, are summarized in [8]. The prospects of creating more accurate models for the development of new materials and molecules are emphasized, which will contribute to increasing the efficiency of scientific research and reducing costs. Work [9] describes the application of modeling and simulation for spectroscopic studies, in particular Raman and infrared spectroscopy. The possibilities of predicting the behavior of systems and assessing the impact of changes based on atomistic methods are studied, which allows linking the structure of materials with their spectral characteristics. An automatic curve fitting algorithm for parametric spectral models that does not require manual tuning and has potential for applications such as calibration transfer is reviewed in [10]. Paper [11] describes indirect hard modeling (IHM) for the analysis of complex spectra with overlapping components. The method automatically determines peak parameters related to molecular interactions, simplifying analysis even for inexperienced users. This allows IHM to be used in ATR-IR and Raman spectroscopy spectra.

Article [12] discusses the use of CNNs for parallel processing tasks such as image segmentation and high-speed computing. The analysis of the spectral characteristics of the network graph, including the Kirchhoff index, graph energy, and spectral radius, is emphasized, which optimizes segmentation and improves diagnostic accuracy. In [13] it is described a neural network for evaluating the quality of alpha spectra, which allows detection of degradation due to calibration or extraneous impurities. The model uses expert knowledge to automate the evaluation process. The paper [14] presents a platform for the classification of spectroscopic data from LWIR and Raman spectroscopy using 1D-CNN. High classification accuracy (more than 90 %) is achieved even under low signal-to-noise conditions. In [15], a synthetic data set was developed for testing machine learning models in spectroscopy. Analysis of eight neural network architectures showed an accuracy of more than 98 %, although complex artifacts such as peak overlap caused errors. The importance of ReLU-activations in classification was highlighted, and the more complex components of the network turned out to be ineffective. The paper [16] investigated the emission spectra of an overvoltage discharge between zinc electrodes in air and nitrogen at different pressures. Microexplosions on the surface of the electrodes introduce zinc vapor into the discharge, which contributes to the formation of molecules and clusters of zinc, its oxides and nitrides in the plasma, which allows the synthesis of nanostructured films of zinc, oxide and nitride of zinc on glass or quartz substrates. In work [17], the plasma parameters in the nanosecond range based on a mixture of helium and zinc vapor for the synthesis of micro- and nanostructures were studied. Numerical modeling of the Boltzmann kinetic equation made it possible to calculate plasma parameters such as mobility, temperature, electron density, and energy losses.

The literature review proves that Raman spectroscopy and machine learning for spectrum processing are actively developing and finding applications in science and industry. Raman spectroscopy is an important method for analyzing the molecular structure and defects of materials, in particular thin films of metal oxides. The development of a complex model that combines physical modeling of processes with modern AI algorithms will improve analysis methods and allow optimizing the composition and structure of films for optoelectronics, photocatalysis and sensor technologies.

To implement the tasks, [18] was additionally developed — a guide for data analysis using Python, which focuses on the use of Pandas, NumPy and IPython libraries for data processing and analysis. The book contains examples of real data and methods of data wrangling. And [19], which evaluated the effectiveness of libraries for data visualization in Python, such as Matplotlib, Seaborn, Plotly, Bokeh, Altair, and ggplot. The results show that Matplotlib, Seaborn, and Plotly are the most popular, with different preferences for complex graphs, simplified plotting, and interactive visualizations.

Research methods

For the precise analysis of thin films using Raman spectroscopy, the mathematical model must take into account:

- the intensity of the Raman lines is determined by the energy distribution of the laser radiation and the molecular structure of the sample;
- film parameters — thickness, chemical composition, structure, presence of vacancies and other defects;
- the interaction of light with the material — includes the phenomena of scattering, absorption and diffusion of light in the film.

In this work, the intensities of Raman lines using the Voight profile, which describes the width and shape of the spectral lines, as well as the influence of the energy distribution of the laser beam through the Gaussian profile, are taken into account

during the development of the modeling technology. The film parameters include the chemical composition, the corresponding spectral peaks, the film thickness due to the exponential absorption of the intensity, and defects affecting the amplitude of the spectral lines. The model accounts for the interaction of light with the material due to the absorption effect, which depends on the absorption coefficient, and the effect of buffer gases (Ne, Ar), which suppress unwanted chemical reactions. Machine learning methods have been implemented to predict spectral characteristics and optimize model parameters using random forests, gradient boosting, and neural networks, with the minimization of root mean square error.

The code does not take into account the distribution of laser energy in three-dimensional space and the effect of its power on the spectrum. Film anisotropy, modeling of multilayer structures, vacancies or dislocations are not included. Also, light scattering at film boundaries, light diffusion, temperature effect on the spectrum and the effect of fluctuations in external parameters such as pressure or temperature are not considered. In addition, there are no methods for noise correction in experimental data.

Mathematical foundations. Basic mathematical models for describing Raman spectroscopy contain a number of steps. Let's consider popular approaches in more detail.

1. Basic equations of Raman spectroscopy.

• Raman signal intensity. The intensity of IR Raman scattering is described by the expression:

$$I_{\mathrm{R}} = I_0 K(\upsilon_0 - \upsilon_s)^4 \frac{d\sigma}{d\Omega} n,$$

where I_0 — intensity of incident laser radiation; K — a constant that depends on the experimental setup; υ_0 — the frequency of laser radiation; υ_s — the frequency of the Raman line; $\frac{d\sigma}{d\Omega}$ — differential cross section of Raman scattering; n — the concentration of molecules interacting with the laser.

• Calculation of Raman peaks shift. The frequency υ_s for oscillations is defined as:

$$\upsilon_s = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}},$$

where k — strength of elastic connection; μ — the reduced mass of atoms participating in oscillations.

• Absorption of laser radiation in a film. The laser radiation intensity $I(z)$ at the depth z in the film is described by:

$$I(z) = I_0 e^{-\alpha z},$$

where α is the absorption coefficient of the film.

2. Model for a multilayer structure. If the film consists of several layers (for example, defect zones or different crystallinity), the IR signal is summed:

$$I_{\mathrm{R}} = \sum_{i=1}^{N} I_{\mathrm{R},i} = \sum_{i=1}^{N} I_0 K(\upsilon_0 - \upsilon_{s,i})^4 \frac{d\sigma_i}{d\Omega} n_i\, e^{-\alpha_i z},$$

where N — number of layers; α_i — the absorption coefficient for the *i*-th layer; n_i — the concentration of active molecules in the *i*-th layer.

3. Influence of defects and nanostructures.

• Model of defects in the film. Defects, such as oxygen vacancies in WO_3, change vibrational frequencies due to local changes in elastic forces:

$$\Delta\upsilon_s = \upsilon_s^{\text{ideal}} - \upsilon_s^{\text{defect}},$$

where $\upsilon_s^{\text{ideal}}$ — frequency for ideal structure; $\upsilon_s^{\text{defect}}$ — the frequency for the defect region.

• Peak width calculation. The width of the Raman peaks Γ depends on the size of the nanostructures:

$$\Gamma = \Gamma_0 + \frac{C}{d},$$

where Γ_0 — peak width for a large crystal; C — a constant that takes into account the material of the film; d — the average size of nanoclusters.

4. Numerical modeling and implementation. To calculate the intensity, the Monte Carlo method is often used to simulate laser penetration and absorption in a multilayer film. Fitting of experimental data is usually implemented using nonlinear regression to determine parameters α, n, $\Delta\upsilon_s$ and Γ from experimental spectra. In order to visualize the results, maps of the distribution of intensities in the film are constructed for the analysis of the local structure.

Modeling technology and software implementation. In this work, computer modeling technology is implemented in the Python programming environment (SciPy, NumPy, Scipy, Sklearn, Pandas, and Matplotlib libraries), and the program structure contains a number of components. Let's consider in detail the main components of the model.

• Raman spectrum generation. The total intensity of the spectrum consists of individual peaks, which are described by Lorentz or Voight profiles:

$$I(\upsilon_s) = \sum_{i=1}^{N} I_i(\upsilon_s),$$

where $I_i(\upsilon_s)$ is the intensity of the i-th peak.

Profiles of Lorents and Voits:

$$I_i(\upsilon_s) = \frac{A_i}{1 + \left(\dfrac{\upsilon_s - \upsilon_{0,i}}{\gamma_i}\right)^2}; \quad I_i(\upsilon_s) = A_i \,\mathrm{Re}\left[\frac{\mathrm{wofz}(z)}{\sigma\sqrt{2\pi}}\right],$$

where

$$z = \frac{(\upsilon_s - \upsilon_{0,i}) + j\gamma_i}{\sigma\sqrt{2}}.$$

These profiles are implemented programmatically with corresponding functions:

```
def lorentzian(nu_s, nu_0, gamma, amplitude):
    return amplitude / (1 + ((nu_s - nu_0) / gamma) ** 2)
def voigt_profile(nu_s, nu_0, gamma, sigma, amplitude):
    z = ((nu_s - nu_0) + 1j * gamma) / (sigma * np.sqrt(2))
    return amplitude * np.real(wofz(z)) / (sigma * np.sqrt(2 * np.pi))
def total_spectrum(nu_s):
    sigma = 10e12 # Gaussian width for Voigt profile
    return (voigt_profile(nu_s, nu_peak_1, gamma_1, sigma, amplitude_1) +
```

```
voigt_profile(nu_s, nu_peak_2, gamma_2, sigma, amplitude_2) +
voigt_profile(nu_s, nu_peak_3, gamma_3, sigma, amplitude_3))
```

• Laser beam energy distribution. The laser_profile function models the laser energy distribution using a Gaussian profile, using the equation to describe the laser radiation intensity as a function of the distance r from the beam center:

$$P(r) = P_0 \cdot e^{-\left(\frac{r}{\omega}\right)^2},$$

where P_0 — laser power, ω — beam width. The parameter r is used to model the energy distribution of the laser beam using a Gaussian profile. Programmatically implemented by a function

```
def laser_profile(r):
    return laser_power * np.exp(-(r / beam_width) ** 2)
```

• Effect of film thickness. The intensity is modified using the absorption coefficient:

$$I(d) = I_0 \cdot e^{-\alpha d},$$

where α — absorption coefficient, d is the film thickness. Implemented by function

```
def intensity_with_absorption(I_0, d):
    return I_0 * np.exp(-absorption_coefficient * d)
```

• Generation of experimental data. The load_spectral_data function loads spectral data from a file. Reads a CSV file containing two columns: nu_s (Raman shift (frequency, Hz)) and intensity (measured intensity). Returns two variables: nu_s array (frequencies) and intensity array. The corresponding function has the form:

```
def load_spectral_data(file_path):
    data = pd.read_csv(file_path)
    nu_s = data['nu_s'].values # Raman shift frequencies (Hz)
    intensity = data['intensity'].values # Measured intensities
    return nu_s, intensity
```

Added logic to check if the file exists to load or generate data. If the file is found, loads the data using load_spectral_data. If the file is not found, generates synthetic data according to the following algorithm:

```
experimental_data = total_spectrum(nu_s) + np.random.normal(0, 0.02, len(nu_s))
```

• Fitting (approximation) of the spectrum. The fitting uses non-linear methods to find parameters A_i, $\upsilon_{0,i}$, γ_i, σ_i that minimize the difference between experimental and simulated data.

The function for optimization has the form:

$$\chi^2 = \sum_k (I_{\text{exp}}(\upsilon_{s,k}) - I_{\text{fit}}(\upsilon_{s,k}))^2.$$

The software is implemented as follows:

```
def fit_func(nu_s, a1, c1, g1, a2, c2, g2, a3, c3, g3):
    sigma = 10e12 # Consistent Gaussian width
    return (voigt_profile(nu_s, c1, g1, sigma, a1) + voigt_profile(nu_s, c2, g2, sigma, a2) +
    voigt_profile(nu_s, c3, g3, sigma, a3))
initial_guess = [1.0, nu_peak_1, gamma_1, 0.8, nu_peak_2, gamma_2, 0.7, nu_peak_3,
gamma_3]
popt, pcov = curve_fit(fit_func, nu_s, experimental_data, p0=initial_guess)
fitted_spectrum = fit_func(nu_s, *popt)
```

• Machine learning for prediction. Since spectral data often have non-linear relationships between different spectral features, machine learning algorithms can effectively model such relationships, unlike traditional linear methods such as linear regression. In addition, spectral data usually have a large set of dimensions (large spectra), and these algorithms can handle a large number of features, taking into account the importance of each one. Ensemble algorithms (Random Forest and Gradient Boosting) are able to work with data containing noise, which is typical for spectral measurements. In terms of automating the process, neural networks, particularly deep neural networks, can automatically detect important patterns in large spectral data sets, reducing the need for manual feature selection. Overall, these algorithms are powerful tools for spectral analysis because they are able to process complex, multidimensional data with high accuracy and efficiency.

In connection with the above, the work uses a comprehensive approach to forecasting spectral characteristics based on ensemble algorithms and neural networks. Three algorithms based on machine learning have been implemented, namely Random Forest, GradientBoostingRegressor — for more accurate dependency analysis, and neural network (MLPRegressor) — for non-linear forecasting. In addition, the selection of the best model based on the MSE is implemented.

• The Random Forest algorithm predicts the intensity of the spectrum. The work of the method is based on an ensemble of decision trees. The final prediction is calculated as the average of the tree predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x),$$

where T — the number of trees, $f_t(x)$ is the prediction of t-th tree. The software implementation has the form:

```
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
mse_rf = mean_squared_error(y_test, y_pred_rf)
```

• Gradient Boosting. Gradient boosting builds an ensemble of trees step by step, minimizing the loss function (MSE):

$$\hat{y}_{m+1} = \hat{y}_m + \eta h_m(x),$$

where η is the learning rate, $h_m(x)$ is a new tree that approximates the negative gradient of the loss function. The software implementation has the form:

```
gb_model = GradientBoostingRegressor(n_estimators=100, random_state=42)
gb_model.fit(X_train, y_train)
y_pred_gb = gb_model.predict(X_test)
mse_gb = mean_squared_error(y_test, y_pred_gb)
```

• Neural network (MLPRegressor). A neural network consists of layers of neurons that use an activation function (ReLU). The forecast is calculated as:

$$\hat{y} = \sigma(W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2),$$

where W_1, W_2 are weights, b_1, b_2 are displacements, and σ is an activation function. The software is implemented as follows:

```
nn_model = MLPRegressor(hidden_layer_sizes=(100, 50), max_iter=1000, random_state=42)
nn_model.fit(X_train, y_train)
y_pred_nn = nn_model.predict(X_test)
mse_nn = mean_squared_error(y_test, y_pred_nn)
```

• Comparison of models. The MSE of each model is calculated:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

```
best_model = min([(rf_model, mse_rf), (gb_model, mse_gb), (nn_model, mse_nn)],
key=lambda x: x[1])[0]
```

• Forecast by the best model. It is implemented on the basis of forecast uncertainty analysis. Mathematically, the standard deviation of the forecast has the form:

$$\Delta I_{\text{AI}}(\upsilon_s) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(I_{\text{AI},t}(\upsilon_s) - \bar{I}_{\text{AI}}(\upsilon_s))^2}.$$

Software forecasting is implemented as follows:

```
prediction_std = np.std([tree.predict(X) for tree in rf_model.estimators_], axis=0)
plt.fill_between(nu_s.flatten() * 1e-12, y_pred_full - prediction_std, y_pred_full + predic-
tion_std, color="gray", alpha=0.3)
y_pred_full = best_model.predict(X)
```

These processes make it possible to compare the performance of different machine learning approaches for the analysis and prediction of Raman spectra.

The next key stage was system optimization. This stage is implemented by adding the optimization logic of the gas environment. This made it possible to select a buffer gas (for example, Ne or Ar) using the selected_gas variable; the buffer gas affects the reduction of unwanted reactions due to the suppression factor specified in the dictionary BUFFER_GASES; the result is integrated into a Voigt function that takes into account the effect of the buffer gas. Let's dwell in more detail on the system optimization process.

The implementation of the suppression of unwanted chemical reactions by the buffer gas is modeled by scaling the intensity of the spectrum through the suppression factor F_g, which is specific for each buffer gas:

$$I_{\text{buffered}}(\upsilon_s) = I(\upsilon_s)F_g,$$

where $F_g \in [0,1]$ — suppression factor defined for each buffer gas. According to the dictionary

$$F_g = \begin{cases} 0.8 \text{ for Ne,} \\ 0.9 \text{ for Ar.} \end{cases}$$

• Integration into the Voigt profile. The buffer gas affects the result of the Voigt function:

$$I_{\text{buffered}}(\upsilon_s) = \frac{A \cdot \text{Re}\left[\text{wofz}(z)\right]}{\sigma\sqrt{2\pi}} \cdot F_g, \quad z = \frac{(\upsilon_s - \upsilon_0) + j\gamma}{\sigma\sqrt{2}}.$$

Programmatically, these steps were implemented by adding functions and making appropriate changes to the code.

Research results

The designed and developed technology illustrates the full process from mathematical modeling and optimization to AI prediction and analysis of results.

Scientific papers [20–22] combine studies of the physicochemical processes of formation of thin films and micro- and nanostructures of materials (metal oxides and

sulfides) using laser or electric discharge in gas-vapor mixtures. The conditions of synthesis of materials are considered, in particular, due to the effect of the discharge on metal vapors in the presence of gases. All works focus on the optical and spectral characteristics of the films, which opens up the possibilities of their application in plasma chemical reactors, synthesis of microstructured materials, as well as in bactericidal and optical technologies. Work [20] describes a high-voltage nanosecond discharge in a gas-vapor mixture «Air-Tungsten» at different pressures, which promotes the synthesis of thin films of tungsten oxide (WO_3) on a glass substrate. The main components of the plasma were determined and the optical properties of the discharge were investigated. In [21], the formation of films on the glass surface during the irradiation of aqueous solutions of copper sulfate with nanosecond laser radiation was studied, which leads to the formation of ordered and disordered films with transparency in the range of 300–1200 nm. In [22], a pulsed source of ultraviolet fluxes of silver atoms and ions and micro-nanostructures of silver sulfide formed by a nanosecond discharge are investigated, which can be used as a source of bactericidal radiation and for the synthesis of silver sulfide films. For all the above-analyzed scientific works, the actual task is mathematical and computer modeling, as well as optimization of the experiments. A detailed comparative analysis of the studied processes, modeling results and AI-prediction is no less interesting. Experimental data from [20] were taken to test the model.

In the first stage, a mathematical model for analyzing Raman spectra was tested. It includes calculating the Raman scattering intensity for a single or multilayer material, modeling the Raman spectrum peaks using a Lorentz profile, and fitting experimental data [20] to determine the peak parameters. The result of this step is shown in Fig. 1.

The simulated spectrum (solid line) for Raman scattering based on parameters from [20], the experimental data (dots) with the addition of random noise to simulate real conditions, and the approximated spectrum (dashed line) based on the fitting results. The curves overlap because the approximated model shows a high agreement with both the theoretical spectrum and the experimental data. This indicates that the model effectively describes the physical processes that affect the intensity of the spectrum and takes into account the noise of real measurements. Small discrepancies between the curves may be due to the influence of additional factors, such as film inhomogeneities or measurement errors.

The next step was to apply an artificial intelligence component to the program (random forest model) for predicting the Raman spectrum. At this stage, the system uses experimental data to train the model, predicts the spectral intensity for the entire range of Raman shifts, and visualizes the prediction results on a single graph along with simulated, experimental, and approximated data.

Fig. 2 shows the simulated spectrum (dashed-dotted line) — theoretical signal without noise, experimental data (dots) — with noise added to simulate real conditions, spectrum fitting (dashed line) — approximated peak parameters and AI prediction (solid line) — prediction result made by the random forest model. The overlap of the spectra indicates the high accuracy of the theoretical model and prediction. The fact that the predicted spectrum almost perfectly matches the approximated and experimental data indicates the successful application of machine learning, which considers complex dependencies in the data. The difference between the experimental and predicted spectra, although minimal, is explained by the presence of noise and measurement errors that are not fully taken into account by the model. The MSE of the model is 0.00058, which indicates high prediction accuracy. Additionally, the figure shows the uncertainty intervals for the forecast (filled areas around the forecast line). They illustrate the level of confidence in the results, allowing an assessment of the possible variability of the model forecasts.

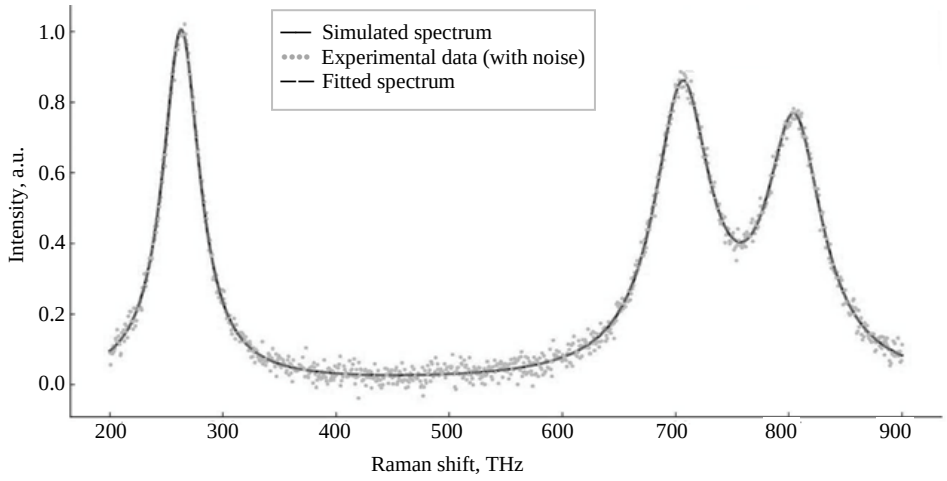

Fig. 1

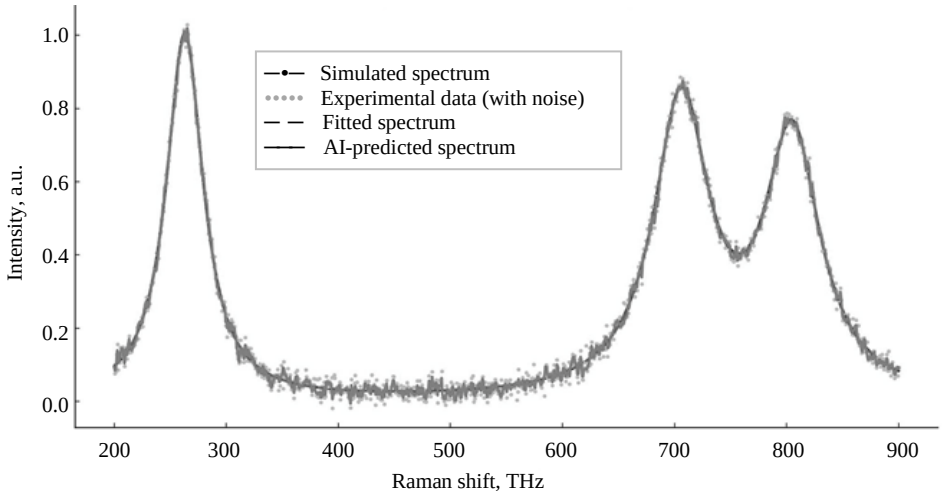

Fig. 2

At the final stage, a Voight profile was added for peak approximation, which provides better analysis accuracy than the Lorenz profile, an analysis of AI prediction uncertainty through visualization of confidence intervals of the random forest model and three machine learning algorithms to improve scientific modeling and prediction of spectra. Additionally, the selection of the best model based on MSE was implemented.

Fig. 3 shows the simulated spectrum (dashed-dotted line) — theoretical signal without noise, experimental data (dots) — with noise added to simulate real conditions, spectrum fitting (dashed line) — approximated peak parameters and AI prediction (solid line) — prediction result made by the random forest model. The overlap of the spectra indicates the high accuracy of the theoretical model and prediction. The fact that the predicted spectrum almost perfectly matches the approximated and experimental data indicates the successful application of machine learning, which considers complex dependencies in the data. The difference between the experimental and predicted spectra, although minimal, is explained by the presence of noise and measurement errors that are not fully taken into account by the model. The MSE of the model is 0.00058, which indicates high prediction accuracy. Additionally, the figure shows the uncertainty intervals for the forecast (filled areas around the forecast line). They illustrate the level of confidence in the results, allowing an assessment of the possible variability of the model forecasts.

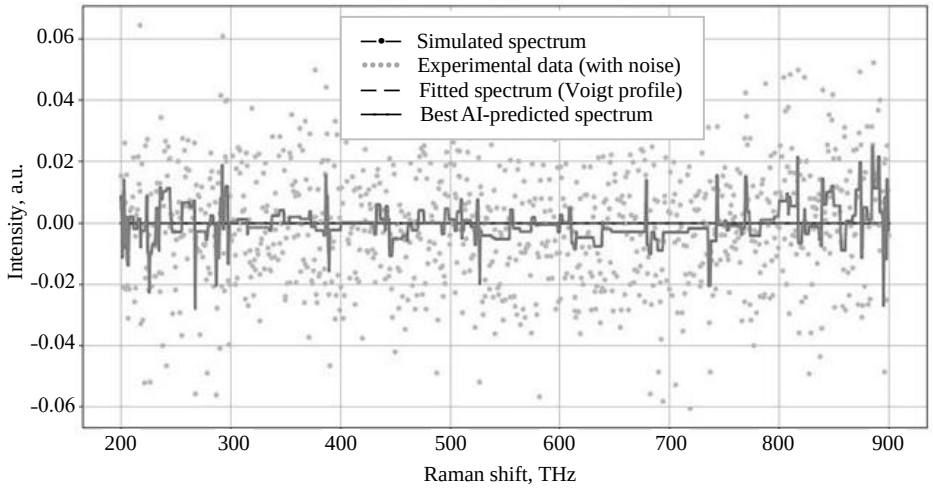

Fig. 3

The result shows that machine learning methods are effective for spectral data analysis, allowing to increase the accuracy of the prediction and take into account complex relationships. The width and shape of the peaks are in good agreement between the fitting and the prediction, indicating the accuracy of the constructed model.

Conclusion

In the work, a technology based on mathematical and software models for the analysis of thin films using Raman spectroscopy was developed, which considers the intensity of Raman lines, film parameters and the phenomenon of light interaction with the material. An adaptive approach was implemented for modeling Raman spectra using the Voight profile, which provides an accurate description of the width and shape of spectral lines. For the first time, it was proposed to include the influence of film thickness, absorption coefficient and laser radiation distribution to ensure realistic modeling. The developed modeling technology includes software integration of machine learning methods (random forests, gradient boosting, neural networks), which allows for effective prediction of spectral characteristics and determination of the best fitting parameters. Model optimization was implemented by introducing buffer gases (Ne, Ar), which reduces the influence of undesirable impurity reactions and improves the quality of experimental data.

The analysis showed high accuracy of spectrum prediction and agreement of theoretical models with experimental data. In general, the developed technology allows to take into account the physicochemical properties of the film, including defects, chemical composition and structure, and also allows to adapt the model to different materials and types of spectrometry. The application of this model will contribute to the development of thin film analysis technologies for optoelectronics, sensor systems, photocatalysis and other high-tech industries. In the future, the expansion of the model may include multilayer films, modeling of crystalline anisotropy and automation of spectral data large sets processing.

Ю.Ю. Білак

ТЕХНОЛОГІЯ МОДЕЛЮВАННЯ, ОПТИМІЗАЦІЇ ТА ШІ-ПРОГНОЗУВАННЯ У РАМАНІВСЬКІЙ СПЕКТРОМЕТРІЇ

Білак Юрій Юрійович

Ужгородський національний університет

yuriy.bilak@uzhnu.edu.ua

У роботі описано розробку технології моделювання, оптимізації та прогнозування спектральних характеристик тонких плівок на основі раманівської спектроскопії із застосуванням штучного інтелекту. Розглянуто впровадження сучасних методів машинного навчання, включно з ансамблевими алгоритмами (випадковий ліс, градієнтний бустинг) та нейронними мережами, що забезпечує високу точність прогнозів і автоматизацію аналізу спектрів. Інноваційний підхід передбачає використання профілю Войта, який поєднує лоренцівську та гаусівську складові, що дає змогу точно описувати ширину і форму піків для апроксимації з урахуванням фізико-хімічних параметрів плівок і впливу експериментальних умов. Задача актуальна у зв'язку зі зростанням вимог до точності аналізу матеріалів у галузях оптоелектроніки, фотокаталізу та сенсорних систем, у яких раманівська спектроскопія є незамінним інструментом. Традиційні методи обробки даних обмежені складністю взаємодії світла з матеріалом, а за допомогою інтеграції ШІ можна подолати ці труднощі завдяки оптимізації, аналізу і прогнозуванню. Запропонована технологія поєднує фізичне моделювання спектрів з ШІ-прогнозуванням, що дає змогу точно враховувати вплив дефектів, неоднорідностей і поглинання. Реалізовано алгоритми для оптимізації моделі з мінімізацією середньоквадратичної помилки і вибору найкращої моделі для вирішення специфічних задач. При додатковій оптимізації моделі враховується вплив товщини плівки з використанням коефіцієнта поглинання та придушення небажаних реакцій за допомогою буферних газів (Ne, Ar). Розроблений підхід забезпечує скорочення часу і ресурсів для експериментальних досліджень, автоматизацію аналізу спектрів і розробку нових матеріалів. Завдяки методам ШІ можна отримати високоточні результати навіть за невеликої кількості експериментальних даних. Серед перспектив розвитку — інтеграція багатошарових структур, урахування анізотропії матеріалів та детальне моделювання дефектів у плівках, а також адаптація для аналізу різних типів матеріалів, таких як органічні плівки чи гібридні структури. Розширення функціональних можливостей програмного забезпечення може передбачати автоматизацію фітингу спектрів, оптимізацію параметрів плівки та прогнозування властивостей на основі машинного навчання з великими наборами даних. Це відкриває нові можливості для досліджень фізико-хімічних властивостей матеріалів і розробки інтелектуальних систем аналізу.

Ключові слова: раманівська спектроскопія, тонкі плівки, моделювання, машинне навчання, спектральний аналіз, оптимізація, нейронні мережі, фізико-хімічні властивості.

REFERENCES

1. Deep learning in spectral analysis: modeling and imaging / L. Xuyang, A. Hongle, C. Wensheng, S. Xueguang. *Trends in Analytical Chemistry*. 2024. Vol. 172. Article 117612. DOI: https://doi.org/10.1016/j.trac.2024.117612
2. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation / O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.-P. Huvenne. *Chemometrics and Intelligent Laboratory Systems*. 2009. Vol. 96. N 1. P. 27–33. DOI: https://doi.org/10.1016/j.chemolab.2008.11.005
3. Collaborative human-AI decision-making systems with numerical channels / O. Mules, M. Kotsipak, S. Dolgikh, Yu. Bilak, T. Radivilova, O. Baranovskyi. *12th International Conference on Advanced Computer Information Technologies (ACIT)*. Slovakia, Ruzomberok, 2022. P. 5–8. DOI: https://doi.org/10.1109/ACIT54803.2022.9913201
4. One dimensional convolutional neural networks for spectral analysis / M.S. Primrose, J. Giblin, C. Smith, M.R. Anguita, G.H. Weedon. *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXVIII*. United States : Florida, Orlando : SPIE Defense + Commercial Sensing, 2022. Vol. 12094. P. 98–108. DOI: https://doi.org/10.1117/12.2618487
5. Madden M.G., Ryder A.G. Machine learning methods for quantitative analysis of Raman spectroscopy data. *Opto-Ireland 2002: Optics and Photonics Technologies and Applications*. Ireland : Galway : Society of Photo-Optical Instrumentation Engineers (SPIE), 2003. Vol. 4876. 11 p. DOI: https://doi.org/10.1117/12.464039

6. Deep convolutional neural networks for Raman spectrum recognition: a unified solution / J. Liu, M. Osadchy, L. Ashton, M. Foster, C.J. Solomone, S.J. Gibson. *Analyst*. 2017. Vol. 142. N 21. P. 4067–4074. DOI: https://doi.org/10.1039/C7AN01371J
7. Deep learning for near-infrared spectral data modelling: hypes and benefits / P. Mishra, D. Passos, F. Marini, J. Xu, J.M. Amigo, A.A. Gowen, J.J. Jansen, A. Biancolillo, J.M. Roger, D.N. Rutledge, A. Nordon. *TrAC Trends in Analytical Chemistry*. 2022. Vol. 157. N 2. Article 116804. DOI: https://doi.org/10.1016/j.trac.2022.116804
8. Machine learning for molecular and materials science / K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh. *Nature*. 2018. Vol. 559. P. 547–555. DOI: https://doi.org/10.1038/s41586-018-0337-2
9. Huang L. Modeling and simulation in spectroscopic study. *Spectral Analysis Review.* 2014. Vol. 2. N 3. P. 7–9. DOI: https://doi.org/10.4236/sar.2014.23023
10. Alsmeyer F., Marquardt W. Automatic generation of peak-shaped models. *Applied Spectroscopy*. 2004. Vol. 58. N 8. P. 986–994. https://opg.optica.org/as/abstract.cfm?uri=as-58-8-986
11. Fully automated indirect hard modeling of mixture spectra / E. Kriesten, F. Alsmeyer, A. Bardow, W. Marquardt. *Chemometrics and Intelligent Laboratory Systems.* 2008. Vol. 91. N 2. P. 181–193. DOI: https://doi.org/10.1016/j.chemolab.2007.11.004
12. Nithya S., Manju G. Spectral analysis of cellular neural network: unveiling network parameters and graph characteristics. *PREPRINT (Version 1) available at Research Square*. 2024. 18 p. DOI: https://doi.org/10.21203/rs.3.rs-4338706/v1
13. Alpha spectral analysis via artificial neural networks / L.J. Kangas, G.L. Troyer, P.E. Keller, S. Hashem, R.T. Kouzes. *Proceedings of 1994 IEEE Nuclear Science Symposium – NSS'94*. USA, VA, Norfolk : IEEE, 1994. Vol. 1. P. 418–421. DOI: https://doi.org/10.1109/NSSMIC.1994.474348.
14. Modeling infrared spectra: an algorithm for an automatic and simultaneous analysis / C. Butucea, J.-F. Delmas, A. Dutfoy, C. Hardy. In B. Castanier, M. Cepin, D. Bigaud, C. Berenguer (Eds.). *31st European Safety and Reliability Conference (ESREL)*. France : Angers, Research Publishing Services, 2021. P. 3359–3366. DOI: https://doi.org/10.3850/978-981-18-2016-8_732-cd
15. Schuetzke J., Szymanski N.J., Reischl M. Validating neural networks for spectroscopic classification on a universal synthetic dataset. *npj Computational Materials*. 2023. Vol. 9. N 100. 12 p. DOI: https://doi.org/10.1038/s41524-023-01055-y
16. Spectroscopic diagnostics of overstressed nanosecond discharge plasma between zinc electrodes in air and nitrogen / O.K. Shuaibov, R.V. Hrytsak, O.I. Minya, A.A. Malinina, Yu.Yu. Bilak, Z.T. Gomoki. *Journal of Physical Studies.* 2022. Vol. 26. N 2. 8 p. DOI: https://doi.org/10.30970/jps.26.2501
17. Numerical simulation of plasma parameters of a gas-discharging reactor on the synthesis of surface zinc nanostructures / A.O. Malinina, O.K. Shuaibov, O.M. Malinin, Yu.Yu. Bilak. *30 Years of the Institute of Electronic Physics of the NAS of Ukraine.* Uzhhorod : IEF of the NAS of Ukraine, 2022. 244 p.
18. McKinney W. Python for data analysis, 2nd ed. USA : Sebastopol, O'Reilly Media, Inc., 2017. 547 p. https://www.oreilly.com/library/view/python-for-data/9781491957653/
19. Assessing the performance of Python data visualization libraries: a review / A. Lavanya, L. Gaurav, S. Sindhuja, S. Hussain, M. Joydeep, V. Uppalapati, W. Ali, V. Sagar. *International Journal of Computer Engineering in Research Trends*. 2023. Vol. 10. N 1. P. 28–39. DOI: https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0104
20. Conditions for pulsed gas-discharge synthesis of thin tungsten oxide films from a plasma mixture of air with tungsten vapors / O.K. Shuaibov, R.V. Hrytsak, O.Y. Minya, A.O. Malinina, I.V. Shevera, Yu.Yu. Bilak, Z.T. Homoki. *Physics and Chemistry of Solid State.* 2024. Vol. 25. N 4. P. 684–688. DOI: https://doi.org/10.15330/pcss.25.4.684-688
21. Synthesis of surface structures during laser-stimulated evaporation of a copper sulfate solution in distilled water / I.I. Bondar, V.V. Suran, O.Y. Minya, O.K. Shuaibov, Yu.Yu. Bilak, I.V. Shevera, A.O. Malinina, V.N. Krasilinets. *Ukrainian Journal of Physics*. 2023. Vol. 68. N 2. P. 138. DOI: https://doi.org/10.15407/ujpe68.2.138
22. Gas discharge source of synchronous Ffows of UV radiation and silver sulphide microstructures / O.K. Shuaibov, O.Y. Minya, R.V. Hrytsak, Yu.Yu. Bilak, A.O. Malinina, Z.T. Homoki, M.M. Pop, O.M. Konoplyov. *Physics and Chemistry of Solid State.* 2023. Vol. 24. N 3. P. 417–421. DOI: https://doi.org/10.15330/pcss.24.3.417-421

Submitted 30.12.2024