On Investigation of Irradiation Effect on Biosensor Counterparts Interaction Based on Advanced Spectrophotometric Studies

V. Martsenyuk and A. Kłos-Witkowska*

Department of Computer Science and Automatics, University of Bielsko-Biala, Bielsko-Biala, Poland

Due to their ease of use, spectrophotometric studies are traditionally applied for the qualitative analysis of solutions and the effects of external influences. In brief, the analysis of spectrographs is based on the analysis of maximum (minimum) absorption values and the corresponding wavelengths. All other (non-extremal) absorption values are eliminated from the consideration, which in turn involves the risk of losing significant information, e.g., on protein conformational changes caused by external factors. Such effects are of importance for bovine serum albumin, which plays the role of a stabiliser for the biosensor receptor layer. In the given work, we develop an advanced numerical method, taking into account the entire range of spectrum wavelengths, which enables us to study more complicated effects and relations within the biochemical system. For the purpose of proving the effectiveness of the method, we study here the effect of ultraviolet radiation on biosensor components with the help of ultraviolet–visible spectroscopy. The primary processing of the spectra was carried out on the basis of absorption-extrema analysis. Further, we apply the dimension reduction technique in order to isolate the wavelengths of greatest importance in the process of the conformation of the studied protein and its complexes. The conducted studies allowed us to prove the accelerated conformational changes in the protein and its complexes under the influence of ultraviolet radiation.

topics: ultraviolet–visible (UV-Vis) spectroscopy, biosensor, analysis, frequency

1. Introduction

Biosensors are devices whose importance is increasing in the modern world due to their versatility and simplicity of use. There are widespread efforts to develop these devices by improving their design [1], minimizing their size [2], and enhancing their detection quality [3], as well as expanding their application spectrum. Among the work on designing biosensors and improving their construction, research is also being conducted on improving the stability and sensitivity of biosensor receptor layers under external influences. These layers are very important design elements that can deactivate with time due to the biological material used, which translates into the lifetime of the device. That is why the improvement and maintenance of the biosensor matrix is vital. Currently, the studies are being conducted in two parallel directions, namely experimental, where the use of new substances in the construction of biosensors is studied [4], and theoretical, where mathematical models are proposed [5, 6]. There are many literature reports on the effects of various external factors on the components of biosensor arrays, including the impact of temperature [7], the use of new substances [8] in the construction of arrays, as well as the effect of electromagnetic field on the substances contained in the matrix [9]. Most works on the effects of electromagnetic radiation on biological substances focus on thermal effects, while non-thermal effects are poorly understood and insufficiently explained due to their complex nature [10].

In the given work, an experiment focusing on non-thermal effects is described. An attempt was made to explain the observed effects. The impact of electromagnetic radiation in the ultraviolet (UV) range on the crosslinking component of biosensor matrices and its complexes, also important for biosensors, is investigated.

Ultraviolet–visible (UV-Vis) spectroscopy is an analytical method that deals with the analysis of spectra arising from the interaction of electromagnetic radiation with matter. This method is based on the phenomenon of selective absorption of radiation by the test sample. The effect caused by the absorbed electromagnetic radiation is the transition of valence electrons to higher energy levels $\pi \to \pi^*$, $n \to \pi^*$.

UV-Vis spectroscopy is used in the study of the structure of organic compounds [11] and is commonly used to quantify various compounds [12]. It is also used to track conformational changes in proteins [13–15]. Based on the analysis of radiation absorption at a certain wavelength, using spectra, it is possible to track changes in the activity of chemical compounds among proteins. Two parameters visible in the spectrum are commonly used for analytical purposes. The first is the absorption of radiation, the second is the wavelength. There are known methods of analysis based on the absorption maximum at a certain wavelength [16], as well as the method [17] using absorption maximum (A_{max} occurring at a certain wavelength (peaks in the spectrum) with the minimum absorption at a certain wavelength (A_{min}). Based on the obtained $A_{\mathrm{max}}/A_{\mathrm{min}}$ parameter changing with time, conclusions are made about the conformational changes occurring in the studied substances with time.

1.1. Research gap in processing spectroscopy data

Methods of dimension reduction are widely used (PCA — principal component analysis — is very often applied) when processing experimental data in various branches (unsupervised learning, clustering, data compression). As a rule, they are applied to tuples of datasets presenting object attributes. On the other hand, the application of PCA for the sequential data is outside the consideration. In turn, we could acquire significant results on the numerical values and curve shapes by applying PCA transformation to the data sequences. Moreover, the obtained visualisations in 2D or 3D allowed us to get clear insight into the corresponding clusters. The idea of applying PCA to the sequential data of absorption spectrograms is innovative and has not yet been studied properly.

The research gap that could be filled by the proposed work can be characterised by two aspects.

Firstly, the traditional usage of the UV spectroscopy method means analysing the peaks of the spectrogram chart only ("position", "intensity", and "width" [18, 19]). However, at the same time, we eliminate the whole spectrum of the other spectra's values. These omitted values bring a lot of the information on the sample, including quantitative and qualitative features (e.g., changes in the polypeptide chain).

Secondly, UV spectroscopy studies of bovine serum albumin (BSA in a traditional way (peak-oriented approach) do not allow us to investigate the qualitative behaviour of samples affected by various external conditions (e.g., irradiation). It is very important to take such influences into account when designing biosensors where BSA plays the role of stabiliser. These effects can be expressed through

the absorption at the corresponding spectra's values (not only peaks). In order to find such effects, the shape of the absorption curve should be analysed, i.e., the whole spectrum has to be taken into account. Time series classification is traditionally the keystone of machine learning. Most of the works are based on the models of deep learning for big datasets. We mention the recent results in [19] concerning the classification of local and global patterns on robust temporal feature networks, in [20] for federated distillation learning system, and in [21] based on convolutional neural networks for UV spectra classification of food additives. Here, we apply the models of classic machine learning accompanied by dimension reduction, which shows good efficiency on absorption spectra data.

In this work, the experiment was carried out using UV-Vis spectroscopy. The study was analysed by an absorption-extrema-based method (pick-oriented method — PO method), and a novel, more detailed spectra analysis based on dimension reduction (whole-spectrum absorption analysis method — WSAA method) was proposed. Using the proposed technique, it is also possible to track changes in the protein with time and determine the magnitude of these changes.

1.2. Research motivation

The motivation for the search for new methods of analysis came from the experience of ongoing studies on the influence of external factors on the stability of the components of the receptor layers of biosensors, which were conducted using UV-VIS spectroscopy [22]. Therefore, a new, more accurate analysis tool was needed, which would additionally allow for prediction of conformational phenomena.

To date, the problems with proving the effects of external factors on the component biosensor receptor layer exist. The evidence should be formulated in terms of parameter sensitivity and biosensor stability. On the other hand, the results should be presented in an interpretable manner, such as decision trees of if–then rules, whereas, as a rule, biochemists widely use linear regression (see paper [15]). Moreover, most of the investigation methods, such as UV-spectrograms, require comprehensive tools for their interpretation, whereas they usually analyse only the extrema, leaving the rest of the values out of consideration. Although recently there appeared works trying to apply deep learning for spectrogram analysis, the possibilities of classical machine learning have not been properly studied. Currently developed analysis methods using deep learning are based on the use of large data sets, whereas real data, like spectrograms, are not such (a few hundred). Due to it, we are able to analyse the spectrum not only on the basis of the maximum absorption value and the corresponding wavelength, but we are

able to achieve greater accuracy of analysis on the basis of the whole spectrum.

Throughout the years, PCA has been a well-known technique for dimension reduction. Here, we try to apply it for the first time to enhance the processing of spectrometric data. In recent research, artificial neural networks were applied to a classification problem with 2001 inputs; dimension reduction allowed us to determine the most important wavelengths for classification. Moreover, the dimension reduction allowed us to acquire the wavelengths related to the type of classification problem (with respect to the content of substances, external effects such as irradiation, etc.).

Our approach is based on the modification of PCA, called PCA with a reduced set of attributes on output (PCA-AO). Further, this reduced set of attributes will be used for solving the problems of classification (e.g., with the help of the C5.0 algorithm).

The work addresses the problems mentioned above as follows. The aim is to evidence the effects of external factors, such as temperature and irradiation, by studying absorption spectrograms. For this purpose, the techniques of classic machine learning have been used. The advantages of the proposed approach are using the whole spectrum analysis, dimension reduction, and interpretable models, such as decision trees and classification rules.

In summary, the objective of this work is to study the effect of external influences, namely, UV-Vis radiation, on substances important to biosensors using an advanced dimension reduction method of analysis of UV-Vis spectra.

2. Materials and methods

2.1. Experimental studies

Aqueous solutions of bovine serum albumin (BSA) with a concentration of 2 mg/ml and two complexes: (i) bovine serum albumin with enzyme acetylcholinesterase (BSA+AChE) with concentration of 2 mg/ml (BSA) and concentration of 2 mg/ml (AChE) and (ii) bovine serum albumin (concentration of 2 mg/m) with enzyme acetylcholinesterase (concentration of 2 mg/m) and acetylcholine chloride (concentration of 2 mg/m) (BSA+AChE+AChCl) were tested.

The examination reagents have been obtained from Sigma Company. They are as follows: BSA in the form of a solid, crystallised and lyophilised powder (99% purity, lot: SLBK3063V); AChE type VI lyophilised powder 200–1000 units/mg protein 23 mg solid, 22 units/mg solid, 288 units/mg protein lot: SLCC7974; AChCl $\geq$ 99%(TLC), lot: BCCD7863.

Tests were performed on aqueous solutions on the basis of scientific reports [17] that such a solution

provides a better understanding of the physico-chemical properties of the test substances. The choice of BSA 2 mg/ml was a consequence of earlier studies [13].

During the study, the solutions have been divided into two groups. One of them consisted of control samples (BSA; BSA+AChE; BSA+AChE+AChCl), and the second one (BSA; BSA+AChE; BSA+AChE+AChCl) was exposed to an electromagnetic field with a wavelength $\lambda = 395$ nm during 60 min in the first day of the experiment. During radiation, the cuvettes with solutions were placed directly in front of the sources, while the illumination was carried out in a darkened system isolated from external light.

It is well known [10] that the effects of electromagnetic radiation on matter depend not only on the type of radiation, but also on the exposure time. In the case of our experiment, exposure times varied. The time of a single measurement in the spectrometer did not exceed 1 min, compared to the time of continuous exposure during radiation of 60 min. The power of radiation was also different — in the case of a UV spectrometer, the source power was 6 W compared to 3 W from source radiation. Nevertheless, taking into account the time of exposition, we get: in the case of UV-radiation — 6×1 min in single measurement mean 6 W/min (7 times in case of repetition) = 6 W/min = 42 W/min, and in the case of radiation source — 3 W $\times$ 60 min = 180 W min. It has also been proven [22] that continuous exposition to radiation gives stronger effects than fractional irradiation. In the case of 7 times repetition, the radiation will be fractional not continuous.

Next, on the 1st, 2nd, 3rd, 5th, 6th, 7th, and 8th day of experiment, UV-Vis spectra in the range of 240–350 nm were collected with the use of UV-Vis double beam Dynamica spectrophotometer (Halo DB-20R).The spectrum range was selected in such a way that it was possible to track changes in the absorption peak at 277 nm, originating from aromatic resistors present in the tested protein, such as tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe) [23]. The BSA UV-Vis spectrum presents two absorption peaks — one at 220 nm and the second at 277 mn. We concentrated on the second maximum peaks because of our research capabilities and consistency with previous research in this area. This is also the correct research approach, which is also used to analyse the BSA [24].

The instrument performance was checked over each day of experiments by making the basic measurement of the solvent (which was water); this was the strength line to make sure that the instrument did not drift. In addition, the equipment was switched on 15 min before the scheduled measurement in order to obtain adequate stability.

The spectra were collected in two formats, i.e., as a screenshot and as a CSV file. From the screenshots, A_{max} (value of maximum absorption
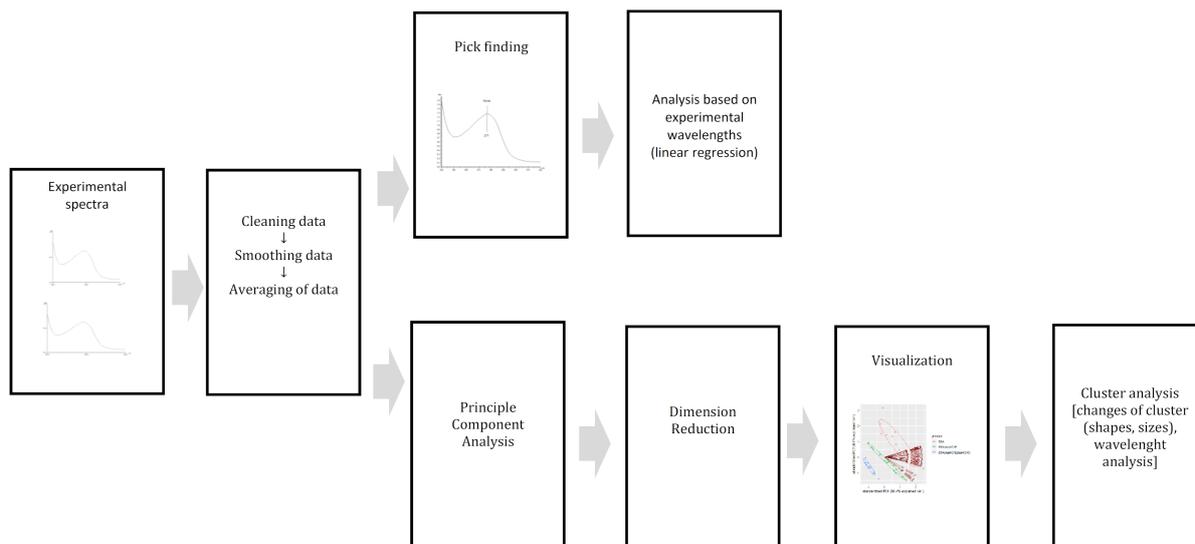
Fig. 1. Flowchart of the numerical method for advanced spectrophotometric studies (the upper branch is for the PO method, the lower one is for the WSAA method).

amplitude) and $A_{\min}$ (value of absorption valley) have been obtained and used for calculations using two models.

2.2. Description of the numerical method

The proposed approach for spectroscopic data includes three basic stages, enabling both the qualitative analysis of the solutions and the assessment of the impacts of the external effects (irradiation).

The numerical analysis is based on the following considerations.

(i) The wavelengths are known to be the characteristic features under investigation in spectroscopy. The absolute values of absorption vs wavelengths should be analysed entirely. This means that we should analyse the entire spectrum of wavelengths (not only for extremal values of the absorption).

(ii) The shape of the spectrum should be taken into account as a whole, as it also brings us a lot of "hidden" information.

(iii) Similarly to the content of the solution, there are some ranges of wavelengths responsible for certain external effects.

Based on the assumptions mentioned above, we developed the numerical methods, including both traditional approaches analysing "extremal" wavelengths and advanced ones based on dimension reduction with visualisation coming from unsupervised machine learning (Fig. 1).

The numerical method has been implemented within the RStudio environment.

In the first stage, we clean data acquired from UV-Vis in the form of a wavelength-absorption table. Here, we remove outliers and artefacts in absorption data. Further, we apply the smoothing of data using one of the mathematical functions (Savitsky–Golay, mean, or median). Next, we should average the data gathered from different experiments under the same conditions (the same substances of the solutions as well as external effects).

Then, the pick-oriented (PO) method, which is traditionally used, finds the picks using the integration of the absorption curve. For this purpose, one of the methods can be utilised (rectangular, trapezoid, Romberg). Besides that, some additional information can be obtained from the derivative of the absorption curve. The next steps are based on the analysis of the picks. The ratios of extremal values, as well as their visualisations, will be analysed in the next section. The pick-oriented method shows us the picks of absorptions corresponding to certain compounds of the solution. Changes in absolute values of the absorption are analysed.

On the other hand, the whole-spectrum absorption analysis (WSAA) method proposed in our work starts with dimension reduction implemented with the help of the principal component analysis with clear determining of decisive attributes (PCA-CDDA). The traditional PCA method proposes as an answer the principal components presenting the transformed values of the attributes with the help of the corresponding affine transform. The advantage of the method is that a few principal components include the most of variance within the data. On the other hand, we change the initial set of attributes to the modified ones. However, the researchers are interested in presenting the results using the names of initial attributes. The fact that using the inverse

Algorithm 1 Principal component analysis with clear determining of decisive attributes (PCA-CDDA).

Require: (Input data) D

Ensure: (Result) Principal components PC1, PC2 together with explained variations, principal attributes.

1: Calculate mean values $\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_j^i$, for $\overline{j = 1, p}$

2: Calculate $\mathrm{Var}(x_j)$, for $\overline{j = 1, p}$

3: Calculate $\mathrm{Var}(D) = \sum_{j=1}^{p} \mathrm{Var}(x_j)$

4: $D' = \{x_j^i - \bar{x}_j\}_{j=\overline{1,p}}$, for $i = \overline{1, N} \in \mathbb{R}^{p \times N}$

5: $C = \frac{1}{p} D'(D')^T \in \mathbb{R}^p$

6: Calculate eigenvalues and eigenvectors of C : $\lambda_1 \ll \lambda_2 \ll \lambda_p$ and $w_1, w_2, \ldots, w_p$, respectively

7: Calculate PC1 $:= Dw_p$, PC2 $:= Dw_{p-1}$

8: Calculate variances $\mathrm{Var}(\mathrm{PC1})$ and $\mathrm{Var}(\mathrm{PC2})$

9: ExplainedVar (PC1) $:= \frac{\mathrm{Var(PC1)}}{\mathrm{Var}(D)}$ and ExplainedVar (PC2) $:= \frac{\mathrm{Var(PC2)}}{\mathrm{Var}(D)}$

10: Calculate $\pi(w_p)$ and $\pi(w_{p-1})$, where $\pi(\cdot)$ is the permutation ordering the vector in decreasing order of absolute values of its elements

11: Return (PC1), (PC2), ExplainedVar (PC1), ExplainedVar (PC2), the names for ExplainedVar (PC1) $\times$ 100% and ExplainedVar (PC2) $\times$ 100%, attributes from $\pi(w_p)$ and $\pi(w_{p-1})$, respectively.

---

transform is computationally complicated is also related to the weak conditionality of such a problem. In order to overcome such a shortcoming, we present the algorithm described below.

Let

$$D = \{(x_1^i x_2^i, \ldots, x_p^i)^T\}_{i=1}^{N} \qquad (1)$$

be the cleaned, smoothed, and averaged of N spectra in the form $X^i = (x_1^i, x_2^i, \ldots, x_p^i)$, where x_j^i is the absorption value, a result of i-th experiment corresponding to j-th wavelength; i is the spectrum identificatory; p is the number of wavelengths in the spectrum. Then the modified algorithm can be presented as Algorithm 1.

As a result of Algorithm 1, we obtained the presentation of the spectra in the form of principal components

$$x_{\mathrm{PC1}}^i = w_p^1 x_1^i + w_p^2 x_2^i + \ldots + w_p^p x_p^i, \qquad (2)$$

$$x_{\mathrm{PC2}}^i = w_{p-1}^1 x_1^i + w_{p-1}^2 x_2^i + \ldots + w_{p-1}^p x_p^i. \qquad (3)$$

The next exploration of the spectra is based on the visualisation of X^i as points $(x_{\mathrm{PC1}}^i, x_{\mathrm{PC2}}^i)$ in the corresponding scatter plots.

Cluster analysis is implemented in two stages. Firstly, we utilise scatter plots and construct 95% ellipses around the clusters of data points.

Secondly, we will construct decision trees based on the C5.0 algorithm. Nodes and the edges present the splitting conditions, whereas the leaves display the classification histograms.

We construct clusters of types:

(i) Clusters involving the spectra for the solutions of the same content.

(ii) Clusters involving the spectra for the solutions under the same external effects.

(iii) Clusters of spectra of the solutions with respect to conformation time.

Note that we can analyse the computational complexity of WSAA. It incorporates two counterparts. Firstly, the computational complexity of PCA-CDDA which can be expressed as $\mathcal{O}(\min(p^3, N^2))$. Secondly, the computational complexity of C5.0, which is $\mathcal{O}(pN \log(N))$ (see p. 335 in [25]).

In turn, at the second step, when using the one-hidden-layer neural network or generalised linear model, we have the computational complexity $\mathcal{O}(\frac{4}{3} i p^2 N)$ [26], where i is the number of iterations for the learning algorithm of the backpropagation type. For the PO model, we have $\mathcal{O}(4 i pN)$.

3. Results and discussion

3.1. Experimental studies analysed by the PO method

The experimental results have been analysed using two methods. The first one is well known, and the model of analysis has been presented in publication [17]. It uses the ratio of the maximum ($A_{\max}$) and minimum ($A_{\min}$) values of the absorption spectra obtained from the UV-Vis spectrometer. Using the compiled ratio of $A_{\max}/A_{\min}$ values over time, we are able to determine the dynamics of changes in the conformation of the protein over time. Using the filtering of the obtained values with a linear function in the form of $y = ax + b$, by analysing the parameter a, we infer the stability of the studied protein and its complexes.

Figure 2 shows an example UV-Vis spectrum for BSA, BSA+AChE, and BSA+AChE+AChCl, with $A_{\max}$ and $A_{\min}$ magnitude marked on the graph.

It can be seen that the spectra have a similar shape coming mainly from atheromatous amino acids contained in BSA, such as tyrosine (Tyr), tryptophan, (Thr), and phenylalanine (Phe), but differ in the value of the absorption coefficient ($A_{\max}$) and valley amplitude ($A_{\min}$).

Previous studies [27] have shown that as the stability of BSA changes over time, absorption increases as a result of structural changes in the protein, which are the consequence of denaturing changes in the protein itself. These may be the result of protein aggregates formed with time and may also result from changes in electrostatic Van der Waals interactions, as reported in earlier work. The protein's secondary structure has changed, and
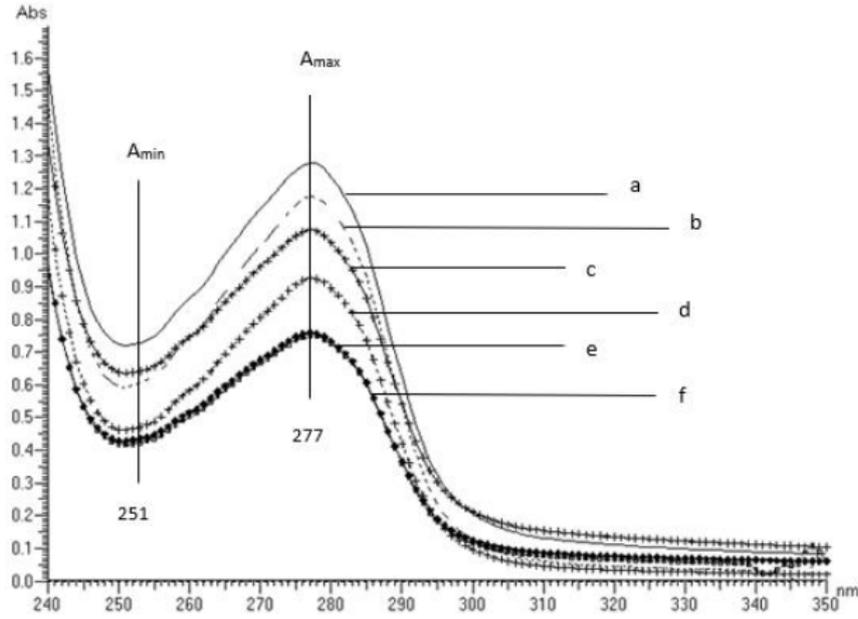
Fig. 2. UV-Vis spectrum for BSA (a), BSA+AChE (c), BSA+AChE+AChCl (e) on the 8th day and for BSA (b), BSA+AChE (d), BSA+AChE+AChCl, (f) on the 1st day, with $A_{\max}$ and $A_{\min}$ values marked on the graph.
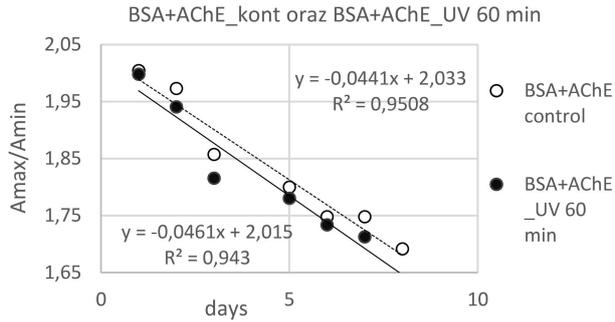

Fig. 3. Changes of the parameter $(A_{\max}/A_{\min})$ on the following days of the experiments for BSA+AChE control samples (empty circles) and BSA+AChE after 60 min UV irradiations (filled circles).

TABLE I

The value of parameter a for samples subjected to UV radiation (60 min) and control samples.

Parameter a		
Samples	Control	UV 60 min
BSA	0.0305±0.0003	0.0361±0.0002
BSA+AChE	0.0441±0.0002	0.0461±0.0004
BSA+AChE+AChCl	0.0052±0.0003	0.0518±0.0003

the order of the polypeptide chain has been rearranged in the local environment. The modifications show a decrease in beta-sheet structure and an increase in helix conformation, which results in a loosening of the protein skeleton [28, 29].

The reported studies showed a similar trend of changes for the protein complexes BSA + enzyme and BSA enzyme AChCl, thus, it can be seen that with time, as the protein ages and at the same time loses stability, the absorption increases in both BSA and complexes.

The primary goal of the experiment was to answer the question of how electromagnetic radiation affects the stability of substances important to the receptor layers of biosensors, namely BSA and BSA+AChE and BSA+AChE+AChCl complexes.

Figure 3 shows the results of the analyses performed using the first well-known model for tracking changes in protein dynamics with time by compiling the $A_{\max}/A_{\min}$ quotient over time.

Over the following days of the studies, the parameter $A_{\max}/A_{\min}$ varies. As shown in Fig. 3, the parameter $A_{\max}/A_{\min}$ drops. According to the model's presumptions, the protein's aromatic amino acids may undergo damaging reactions over time. The parameter of an adjusted linear trend, $y = ax + b$, shows that the rate of change is slightly different. The a parameter shows the kinetics of the changes that are occurring. For control samples (BSA, BSA+AChE, BSA+AChE+AChCl), the value of parameter a is lower than for the same samples subjected to 60 min UV irradiation (Table I). This indicates that the denaturation process in irradiated samples occurs faster. UV radiation causes accelerated changes in the protein. The protein and its complexes under the influence of an external factor, such as UV radiation, lose the stability of the structure faster.
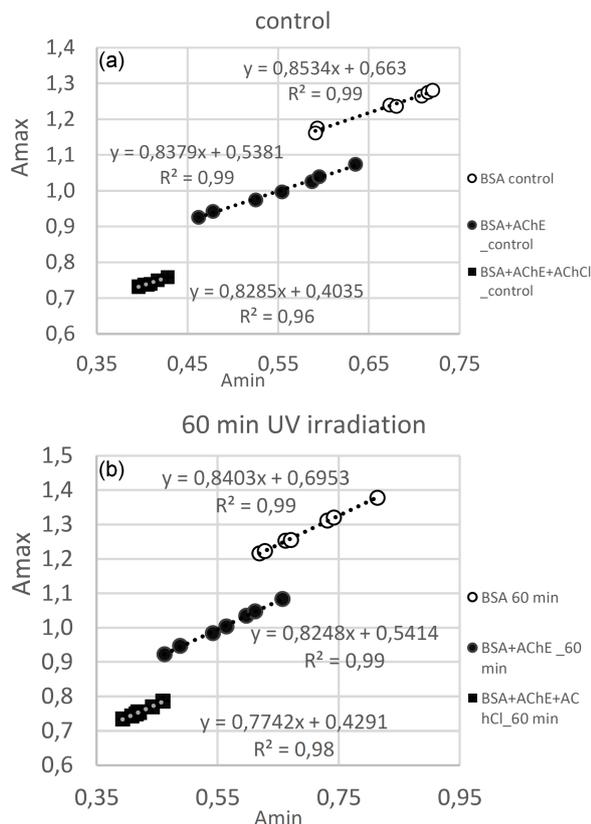
69

Fig. 4. Analysis of conformational changes using the second method of analysis.

Figure 4 shows the analysis of conformational changes using the second method of analysis. This method involves compiling the parameters $A_{\min}$ and $A_{\max}$ and fitting them with a linear function $y = ax + b$. In this method, we infer the rate of change in the dynamics of the system on the basis of parameter b. Comparing the measurement values of parameter b in the control samples and in the radiated samples, it can be concluded that a higher value of parameter b appears in the radiated samples (both in BSA and in the complexes) than in the control samples. Thus, it can be seen that the second proposed method of analysis confirmed that conformational changes in the protein occur more dynamically under UV radiation.

Protein denaturation is a consequence of the loss of stability. The strength of the interaction between amino acids changes with variations in distance along the polypeptide chain. Under the influence of UV radiation, the tertiary structure of the biosensor receptor layer component (BSA and BSA complexes) is altered, which causes the loss of biological activity of the protein. The solvent, which is water in the research that has been done, also affects how stable proteins are. It is common knowledge that the hydrophobic hydration of polar surfaces causes the denaturation of globular proteins. UV exposure may be associated with less buried non-polar side chains

coming into contact with water during unfolding. It is possible that the variations in parameter changes between samples that were exposed to UV radiation and those that were not exposed to electromagnetic field are related to the reorganization of water within the UV-damaged albumin macromolecule.

3.2. Experimental studies analysed by the WSAA method

Application of Algorithm 1 to the experimental dataset resulted in the following values of the explained variation

$$\text{ExplainedVar}(\text{PC1}) = 0.833, \tag{4}$$

$$\text{ExplainedVar}(\text{PC1}) = 0.152. \tag{5}$$

Figures 5–8 present bi-plots in PCA (score plot overlays with the loading plot). The ellipses around score plots (clusters) have a similar meaning to the ellipses around any other scatter plots. The uses of such ellipses are confidence ellipses. Namely, you can view them as the contours of the closest bivariate normal distribution (with 95% confidence, the true centre is located within the ellipse).

When analysing all 110 wavelengths, in Fig. 5 we see that the content of the solution is characterised and can be differed at some wavelengths. Namely, we see that the wavelengths (in nm) 296, 297, 295, 298, 294, 299, 300, 293, 301, 302, 292, 250, 251, 249, 252, 303, and 253 are the most significant with respect to the content of the solution. We note that the transfers between the clusters of the corresponding samples (approximated with the ellipses) are seen best in the corresponding directions. On the other hand, the wavelengths 340, 341, 347, 339, and 346 nm are less informative, as they show the transfers within the clusters.

When executing Step 11 of Algorithm 1, we get the most informative attributes (in the order of decreasing absolute values of weights, in nm): 296, 297, 295, 298, 294, 299, 300, 293, 301, 302, 292, 250, 251, 249, 252, 303, 253 and 340, 347, 341 from PC1 and PC2, respectively. Hence, we can determine 3 of the most decisive intervals: 249–253 nm (corresponds to $A_{\min}$), 292–303 nm (absorption descending after reaching $A_{\max}$), and 340–341 and 347 nm (tail of the spectrum).

The essence of PCA-CDDA (modification of PCA) is that the method gives us a set of "important" wavelengths on output (contrary to PCA, which gives us the principal components). The "important" wavelengths are selected based on the PCA transformations. Namely, the loadings (wavelengths) corresponding to the biggest magnitudes of the rotations for PC1 and PC2 (proportionally to the explained variance) have been chosen.

Figures 6–8 present the same bi-plots. The only difference is the clusters (groups) constructed and the 95% confidence ellipses. The locations of the
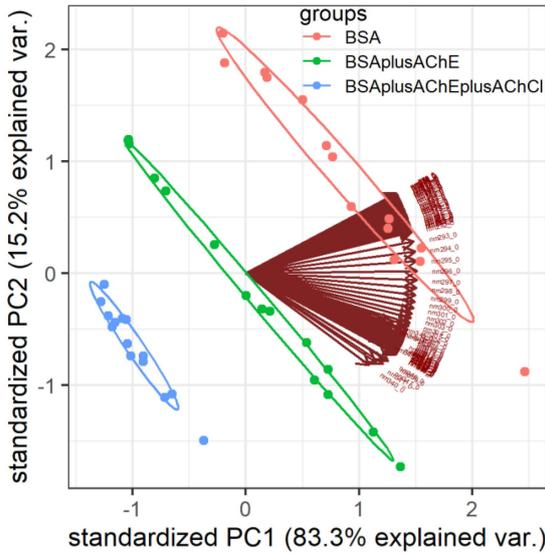
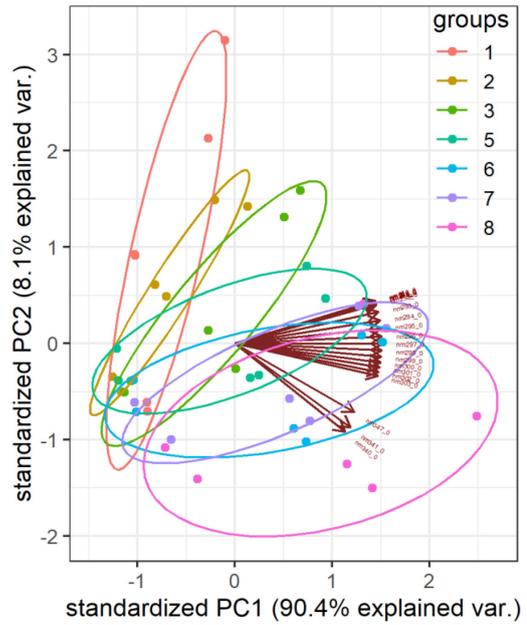Fig. 5. Visualisation of the WSAA-method results for the analysis of the content of the solution.
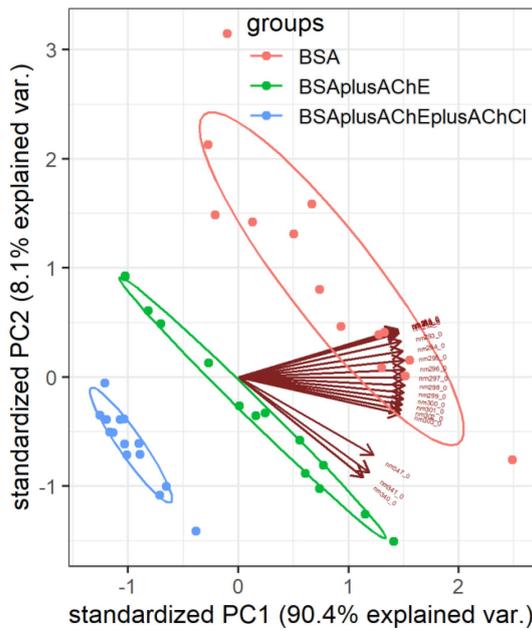

Fig. 6. Spectra in PCA coordinates after the dimension reduction; clusters correspond to the content of the solutions.
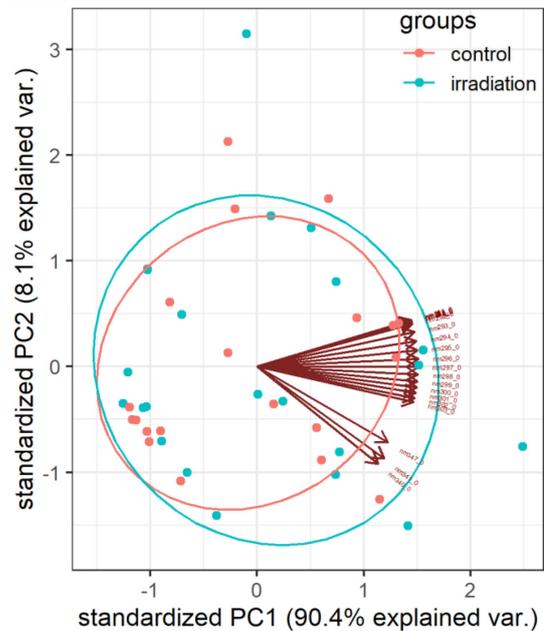

Fig. 7. Spectra in PCA coordinates after the dimension reduction; clusters correspond to the day of the experiment.

Fig. 8. Spectra in PCA coordinates after the dimension reduction; clusters correspond to the irradiation applied.

ellipses and loading plots allow us to get a primary interpretation (in general) of the impact of external effects on absorption at the "important" wavelengths.

The wavelengths on the output of PCA-CDDA are crucial for advanced spectra analysis, as we are interested in knowing the wavelengths responsible for external impacts, conformational changes, etc. The corresponding thorough interpretation can be obtained by further using the decision tree.

Figure 6 shows the selected 20 of the 110 wavelengths mentioned above that were chosen based on Algorithm 1 and are the most significant from the point of view of the analyses.

In Fig. 7, we see that the wavelengths 340, 341, and 347 nm display the duration of the denaturation the most. On the other hand, the wavelengths in
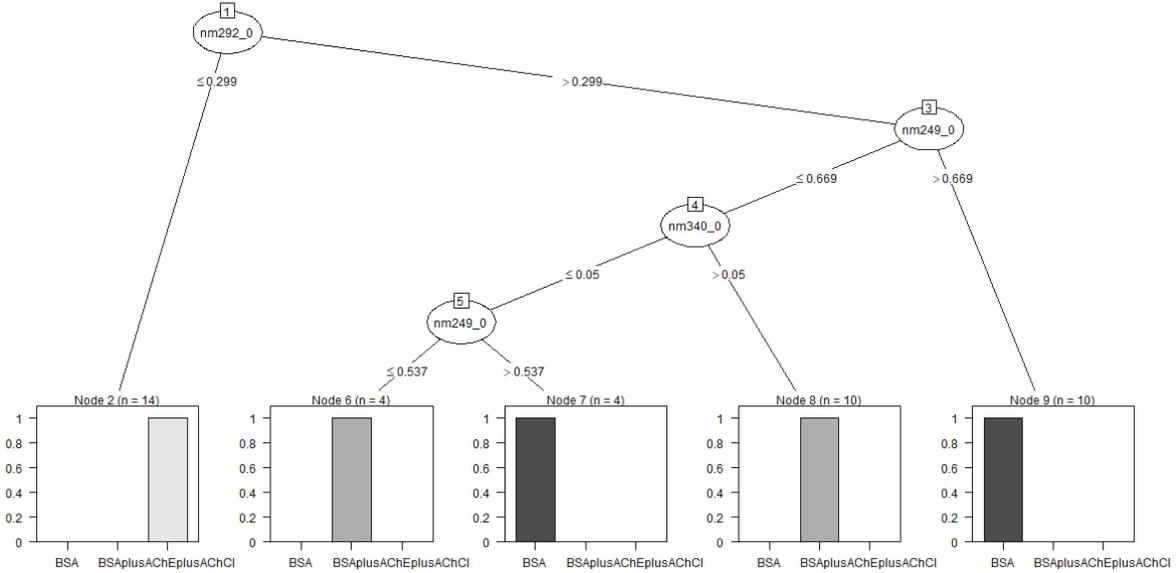
Fig. 9. Decision tree for content analysis.

the range 249–253 nm show the denaturation to a lesser extent and are more "responsible" for the qualitative content of the solution. In the figure, we can also see a decrease in the activity of the protein with time, as evidenced by the change in the position of the clusters from vertical to horizontal. From the width of the ellipses, we can infer the dynamics of the conformational changes, i.e., it can be seen that the greatest changes occur on day 8, when the ellipse is the widest. The observed effect also agrees with the theoretical basis for protein conformations and experimental studies carried out with a spectrophotometer.

Analysing Fig. 8, the differences between the irradiated samples and control samples are noticeable. It can be seen that the dynamics of the process of protein conformation in the radiated samples occurs faster than in the control samples, as evidenced by the shift of the cluster (blue) to the horizontal side. The smallest changes in absorption as a result of irradiation will be observed at the wavelengths range 249–253 nm, while the biggest ones at wavelengths of 340, 341, and 347 nm.

Figure 9 presents the decision tree. This method is commonly used for classifying the excitability depending on the relationship between initial conditions [30]. It can be seen that changes at wavelength 292, 340, and 249 nm influenced complex BSA+AChE+AChCl the most; the changes at 249 nm wavelength seem to be the most important for BSA. Table II shows the attribute usage with the corresponding wavelength nomenclature.

On the other hand, the outcomes of the algorithm C5.0 can be presented in the form of classification rules (please see Appendix A). Rule 1 and Rule 2 allow us to determine the wavelength for BSA classification. Rule 3 and Rule 4 refer to the

TABLE II

Attribute usage for the content analysis of the solutions based on the decision tree (see Fig. 9).

Probability	Wavelengths nomenclature
100.00%	nm292_0
66.67%	nm249_0
42.86%	nm340_0

TABLE III

Attribute usage for the content analysis of the solutions based on the classification rules

Probability	Wavelengths nomenclature
66.67%	nm292_0
66.67%	nm249_0
35.71%	nm340_0

classification of BSAplusAChE, Rule 5 refers to the classification of BSAplusAChEplusAChCl. The attribute usage is presented in Table III.

The rules can be interpreted in the following way:

- Rule 1. If the wavelength is greater than 249 nm for an absorption of 0.669, then the sample is classified as BSA with a probability of 0.917.
- Rule 2. If the wavelength is greater than 249 nm for an absorption of 0.537 and the wavelength is less than 340 nm for an absorption of 0.05, then the sample is classified as BSA with a probability of 0.857.
- Rule 3. If the wavelength is greater than 292 nm for an absorption of 0.299 and the wavelength is less than 249 nm for an absorption of 0.669 and the wavelength is less

than 340 nm for an absorption of 0.05, then the sample is classified as BSAplusAChE with a probability of 0.917.

- Rule 4. If the wavelength is greater than 292 nm for an absorption of 0.299 and the wavelength is less than 249 nm for an absorption of 0.537, then the sample is classified as BSAplusAChE with a probability of 0.857.

- Rule 5. If the wavelength is less than 292 nm for absorption of 0.2995, then the sample is classified as BSAplusAChplusAChCl with a probability of 0.938.

Figure 10 shows the decision tree based on the reduced number of wavelengths. Due to it, we can distinguish the wavelengths whose absorption change will be crucial due to the effect of radiation. Among them, we can distinguish wavelengths 298, 296, 292, 299, 253, 251, 347, 300, 295, 294, 341, and 340 nm.

Based on the decision tree and the number of samples analysed, we can construct the order of the importance of the wavelengths for the purpose of the sample classification (Table IV).

On the other hand, the outcomes of the C5.0 algorithm can be presented in the form of classification rules. As a result of the algorithm, we obtained 15 classification rules (please see Appendix B), i.e., 8 for control samples and 7 for radiated samples. Consideration of the rules should follow the order corresponding to the decrease in the probability of decision-making.

We start the analysis of the control class with Rule 1. If the wavelength is greater than 296 nm for an absorption of 0.256 and the wavelength is less than 298 nm for an absorption of 0.24, then the sample is classified as a control class with a probability of 0.995, and then we consider the Rules from 2 to 8.

For radiation class, we start from Rule 9. If the wavelength is less than 296 nm for an absorption of 0.256 and the wavelength is greater than 292 nm for an absorption of 0.393, then the sample is classified as class irradiation with a probability of 0.995, and then we consider the Rules from 10 to 15.

Parameters for other rules are placed in Appendix B.

Based on the classification rules and the number of samples analysed, we can construct the order of the importance of the wavelengths for the purpose of the sample classification with respect to irradiation (Table V).

3.3. Comparison of numerical methods

For the purpose of splitting the dataset, a 10-fold cross-validation (CV-10) technique was used.

We have compared the proposed technique WSAA (based on PCA-CDDA and C5.0) with the PO method. Moreover, we have studied and checked

TABLE IV

Attribute usage for the sample classification.

Probability	Wavelengths nomenclature
100.00%	nm298_0
90.48%	nm296_0
76.19%	nm292_0
61.90%	nm299_0
57.14%	nm253_0
45.24%	nm251_0
35.71%	nm347_0
33.33%	nm300_0
16.67%	nm294_0
9.52%	nm341_0
7.14%	nm340_0

TABLE V

Attribute usage for classification rules with respect to irradiation.

Probability	Wavelengths nomenclature
57.14%	nm296_0
28.57%	nm292_0
23.81%	nm298_0
21.43%	nm253_0
19.05%	nm299_0
16.67%	nm294_0
14.29%	nm347_0
11.90%	nm300_0
11.90%	nm251_0
11.90%	nm341_0
7.14%	nm340_0

the modifications of WSAA using the generalised linear model (lm) and one hidden layer neural network (nn). The results of the experiments are shown in Figs. 11–12.

Throughout the training, we have used two measures, namely, error rate and training time. At each step, error rate ERR was based on the confusion matrix $C = \{c_{ij}\}_{i,j=1}^{N}$ and computed as

$$\text{ERR} = \frac{\sum_{i=1}^{N} c_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij}}, \tag{6}$$

where N is the number of classes, c_{ij} is the number of samples from class i classified with the corresponding model as from class j. Number of classes N has to be set to 2 for the problem of binary classification of irradiated samples and 3 for the problem of classification of the content (BSA; BSA+AChE; BSA+AChE+AChCl).

Training time was obtained with the help of the systemic function "system.time()".
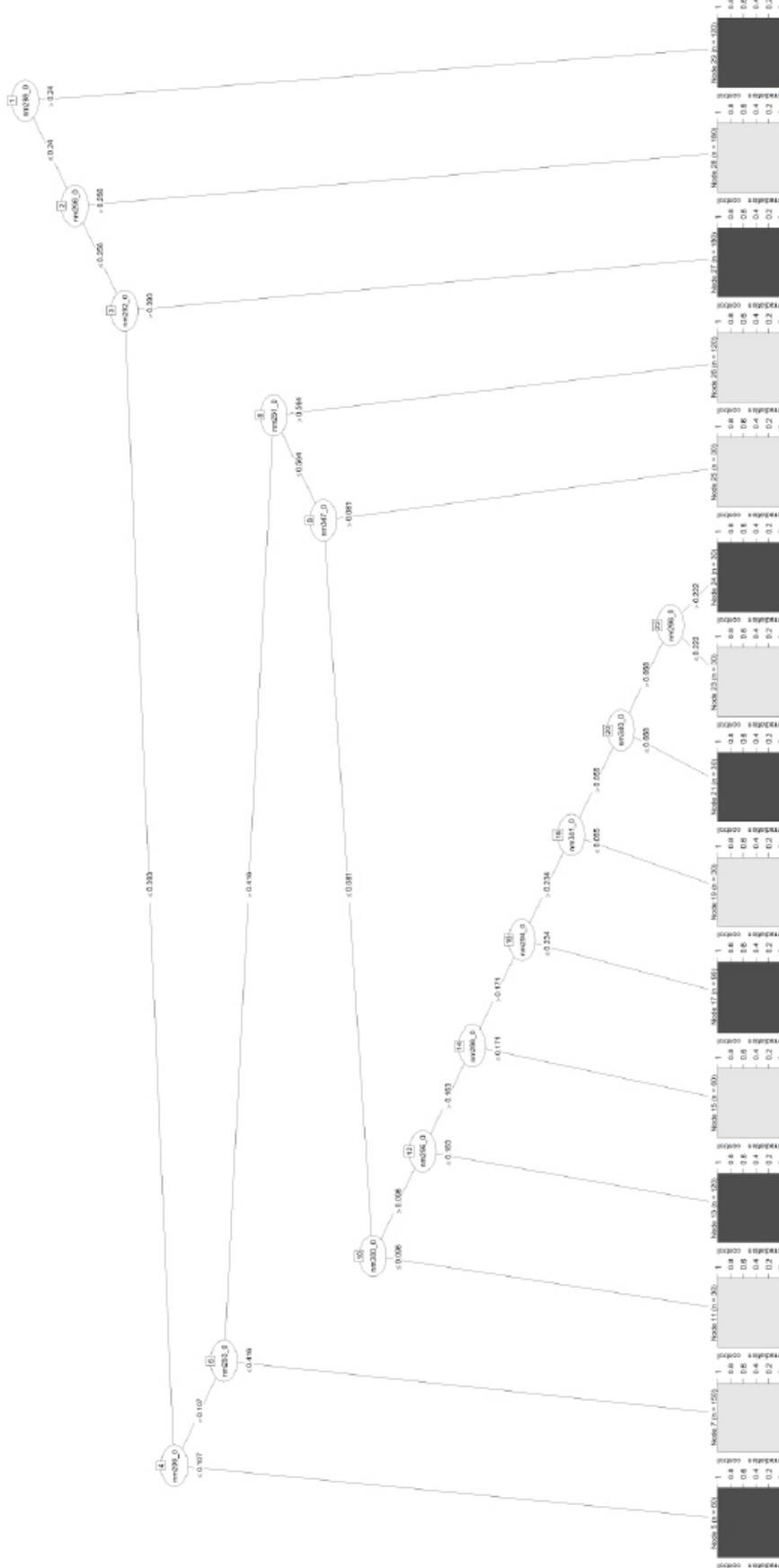
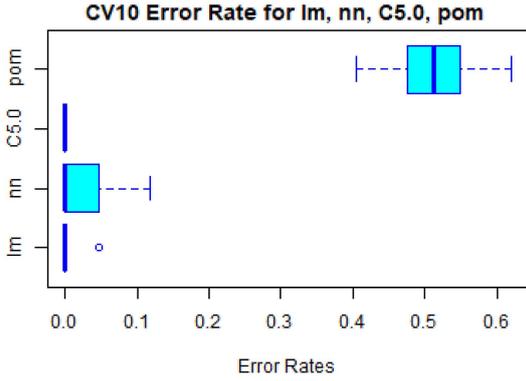Fig. 10.   Decision tree for irradiation analysis.

Fig. 11. Error rates for the models: Im — linear model, nn — neural network, C5.0 — decision tree, pom — pick oriented model.
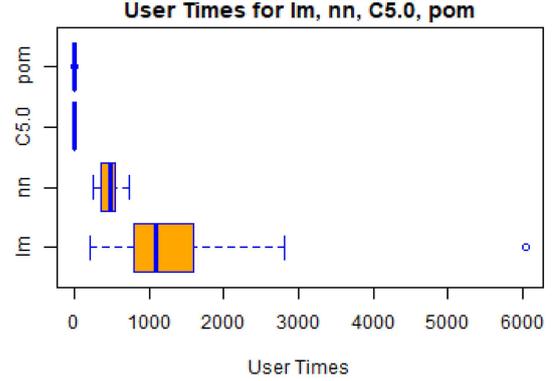

Fig. 12. Training times for the models: Im — linear model, nn — neural network, C5.0 — decision tree, pom — PO-oriented model.

To study different models, we have compared the PO model and the WSAA method (using linear model, neural network, and the decision tree induced with the help of algorithm C5.0). Both linear model and neural network architectures involve one hidden layer with 12 neurons (it was experimentally proven that the value used for the hidden layer was close to 2/3 of the number of inputs). The neural network model uses a logistic function for activation.

A comparison of the values of those measures when using different models is shown in Figs. 11 and 12 using box plots.

The methods were compared from the viewpoint of two criteria. Namely, when regarding the error rate, WSAA combined with the C5.0 algorithm for the decision tree appears to be the most accurate with 100% accuracy. When applying a linear model, despite leaving the same mean value of accuracy, some outliers for error rates appeared. The situation has to be worse for the one-hidden-layer neural network, where we have some positive average values of the error rate. Analysis of the training time (Fig. 12) shows us the computational cost of the algorithms. We see that methods WSAA with C5.0 and the PO model are the most effective computationally. It agrees with the expressions of the computational complexities of the algorithms presented at the end of Sect. 2.2. The computational complexities for neural network and linear model in particular are bigger, since the value of i iterations is required for the training.

For the reasons given, we see that the models based on C5.0 seem the most successful from the viewpoint of both error rate and training time. In turn, the linear model appeared to be more accurate when compared with the neural network, requiring more training time.

Presentation of the peak-oriented model, linear model, and neural network that are optimal from the view-point of the error rate, are shown in Figs. 13–15, respectively.

3.4. Limitations of WSAA

To be effective from the viewpoint of performance measurements, the WSAA method has to overcome a few limitations.

One of them is data cleaning and smoothing. The removal of outliers and artefacts within absorption data might be subjective and dependent on the chosen smoothing method. It may lead to the loss of important information or introduce bias. Hence, we have to clearly define the criteria for outlier removal and choose smoothing methods based on the characteristics of the absorption data. We should document the impact of outlier removal and smoothing on the results.

Averaging absorption data from different experiments assumes that the conditions are exactly the same, which might not always be the case. Variations in experimental conditions can affect the accuracy of the averaged data. So, we should provide detailed information about the experimental conditions and consider statistical measures (such as standard deviation) to quantify the variability among experiments. We have to explore ways to account for experimental variations in the analysis.

The inverse transform of principal components is computationally expensive. PCA-CDDA offers only an approach to represent results in terms of initial attribute names based on the number of attributes and the value of the explained variance. When analysing the clusters for the respective class attribute in PCA coordinates, we have to consider the projections of the attributes to the corresponding axes, which is a more relevant approach for the specific problems.

Eigenvalue calculation in the PCA technique may lead to numerical instability, especially if the covariance matrix is ill-conditioned. So, we need to implement numerical stability checks and regularisation techniques during eigenvalue calculations to ensure
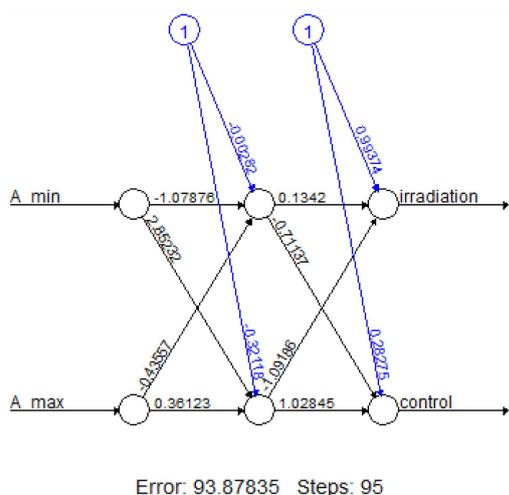
Error: 93.87835 Steps: 95

Fig. 13. Optimal peak-oriented model based on CV-10.

Error: 25.003603 Steps: 505684

Fig. 14. Optimal neural network constructed based on CV-10.

Error: 2.381453 Steps: 898424

Fig. 15. Optimal linear model constructed based on CV-10.

using other classification algorithms and evaluate their performance in comparison to decision trees.

Defining clusters based on specific criteria (content, external effects, conformation time) assumes prior knowledge and might not capture all relevant patterns. Hence, we should explore unsupervised clustering methods that can reveal hidden patterns in the data without predefined criteria.

The interpretation of clusters and decision trees may be challenging, especially if they are complex and involve multiple factors. Here, we study the influence of single factors. In order to extend the results for multiple factors, we need to provide clear explanations of the criteria used for cluster construction and decision tree nodes. Considering visual aids or additional analyses to enhance the interpretability of the results is required.

By addressing these limitations and documenting the steps taken, the reliability and applicability of the numerical method for absorption spectrogram analysis can be improved.

4. Conclusions

The results of the study on the effect of an electromagnetic field in the UV range on the substances used in the construction of biosensors are presented in the paper. The conformational changes occurring in the protein on successive days of the experiment, after the exposure performed on the first day, were tracked. UV spectroscopy was used for the study. The resulting spectra were analysed using standard

robustness. Documenting any measures taken to handle ill-conditioned matrices is required.

Constructing clusters based on scatter plots of PCs and ellipses might not be optimal for all types of data distributions. It may not capture non-linear relationships. So, we should explore alternative clustering methods that can handle non-linear relationships.

Decision tree construction based on the C5.0 algorithm might be sensitive to noise in the data and may overfit the training data. Hence, performing thorough validation and pruning of decision trees to avoid overfitting is required. We have to consider

models and advanced numerical method, which allows for determining the wavelength that affects the conformational changes in the test substance and its complexes. The experiment showed accelerated deactivation of the protein and its complexes under the influence of an electromagnetic field in the UV range. The observed effects were the results of protein aggregation in secondary structure, local environmental changes, an increase in helix conformation, and a decrease in beta-sheet structure. Moreover, the examined protein may fragment and aggregate over time, and electrostatic interactions (van der Waals interactions) inside the protein and their complexes may alter. Faster protein degradation under the action of UV light was observed in the study.

As can be seen from the number of classification rules and the complexity of the decision tree (the number of levels), the UV-Vis spectrophotometry method is more suitable for qualitative analysis of the solution than for conducting analysis of the influence of external factors, although by using advanced numerical methods based on the entire spectrum, it can also be used for this type of research.

Acknowledgments

Appendix A

Classification roles applied to content analysis are the following

- Rule 1: (10, lift 2.8)
  nm249_0 > 0.669
  → class BSA [0.917]
- Rule 2: (5, lift 2.6)
  nm249_0 > 0.537
  nm340_ ≤ 0.05
  → class BSA [0.857]
- Rule 3: (10, lift 2.8)
  nm292_0 > 0.299
  nm249_0 ≤ 0.669
  nm340_0 > 0.05
  → class BSAplusAChE [0.917]
- Rule 4: (5, lift 2.6)
  nm292_0 > 0.299

nm249_0 ≤ 0.537
→ class BSAplusAChE [0.857]
- Rule 5: (14, lift 2.8)
  nm292_0 ≤ 0.299
  → class BSAplusAChEplusAChCl [0.938]

Appendix B

Classification roles applied to irradiation analysis are the following

- Rule 1:
  nm296_0 > 0.256
  nm298_0 ≤= 0.24
  → class control [0.995]
- Rule 2:
  nm299_0 > 0.107
  nm253_0 ≤ 0.416
  → class control [0.993]
- Rule 3:
  nm292_0 ≤ 0.393
  nm251_0 > 0.564
  → class control [0.992]
- Rule 4:
  nm294_0 > 0.234
  nm292_0 ≤ 0.393
  nm341_0 ≤ 0.055
  → class control [0.989]
- Rule 5:
  nm296_0 > 0.163
  nm296_0 ≤ 0.171
  → class control [0.984]
- Rule 6:
  nm292_0 ≤ 0.393
  nm347_0 > 0.081
  → class control [0.984]
- Rule 7:
  nm296_0 ≤ 0.222
  nm294_0 > 0.234
  nm340_0 > 0.068
  → class control [0.969]
- Rule 8:
  nm299_0 > 0.107
  nm300_0 ≤ 0.096
  → class control [0.969]
- Rule 9:
  nm296_0 ≤ 0.256
  nm292_0 > 0.393
  → class irradiation [0.995]
- Rule 10:
  nm296_0 ≤ 0.163
  nm300_0 > 0.096

nm253_0 > 0.416

nm347_0 ≤ 0.081

→ class irradiation [0.992]

- Rule 11:

  nm298_0 > 0.24

  → class irradiation [0.992]

- Rule 12:

  nm296_0 > 0.171

  nm294_0 ≤ 0.234

  → class irradiation [0.989]

- Rule 13:

  nm296_0 > 0.171

  nm340_0 ≤ 0.068

  nm341_0 > 0.055

  → class irradiation [0.984]

- Rule 14:

  nm299_0 ≤ 0.107

  → class irradiation [0.984]

- Rule 15:

  nm296_0 > 0.222

  nm251_0 ≤ 0.564

  → class irradiation [0.969]

References

[1] K. Theyagarajan, Y. Kim, *Biosensors* **13**(4), 424 (2023).

[2] C. Shiro, H. Nishikawa, X. Kong, H. Tomiyama, S. Yamashita, *Biosensors* **12**(5), 277 (2022).

[3] V. Martsenyuk, A Sverstiuk, O Bahrii-Zaiats, A. Kłos-Witkowska, *CEUR Workshop Proceedings* **134**, 3309 (2022).

[4] L. Mosinska, P. Popielarski, K. Fabisiak, A. Dychalska, *Opt. Mater.* **101**, 109676 (2020).

[5] V. Martsenyuk, A. Klos-Witkowska, *IEEE Access* **7**, 62325 (2019).

[6] A. Klos-Witkowska, V. Martsenyuk, V. Karpinskyi, *Acta Phys. Pol. A* **135**(3), 375 (2019).

[7] A. Michnik, A. Kłos, Z. Drzazga, *J. Therm. Anal. Calorim.* **77**(1), 269 (2004).

[8] H. Park, G. Kim, J. Seo, Y. Yoon, J. Min, Ch. Park, T. Lee, *Biosensors (Basel)* **11**(12), 525 (2021).

[9] A. Klos-Witkowska, V. Martsenyuk, M. Karpinski, I. Obeidat, *2019 Advances in Science and Engineering Technology International Conferences*, ASET 2019, (2019) 8714302.

[10] E. Polk, W. Postow, *Handbook of biological effects of electromagnetic fields* CRP Press (1996).

[11] A.M. Mohammed, *EJERS* **3**(3), 8 (2018).

[12] R.A. Pratiwi, A.B. Nandiyanto, *IJERT* **2**(1), 1 (2022).

[13] A. Klos-Witkowska, *Acta Phys. Pol. A* **133**(1), 101 (2018).

[14] A. Kłos-Witkowska, V. Martsenyuk, *ECS S* **26**(4), 665 (2020).

[15] A. Klos-Witkowska, V. Martsenyuk, *Acta Biochim. Pol.* **68**(2), 325 (2021).

[16] M. Picollo, M. Aceto, T. Vitorino, *Phys. Sci. Rev.* **4**, 20180008 (2018).

[17] A. Michnik, K. Michalik, Z. Drzazga, *J. Therm. Anal. Calorim.* **80**, 399 (2005).

[18] H. Förster, *Mol. Sieves* **4**, 337 (2004).

[19] Z. Xiao, X. Xu, H. Xing, S. Luo, P. Dai, D.Zhan, *Information Sciences* **571**, 65 (2021).

[20] H. Xing, Z. Xiao, R. Qu, Z. Zhu, B. Zhao, *IEEE Transactions on Instrumentation*, 71 (2024).

[21] I. Patarniche, C. Sarosi, R. Terebes, L. Szolga, R. Galatus, *Sensors* **23**(17), 7517 (2023).

[22] A. Kłos-Witkowska, K. Kajstura, *Acta Phys. Pol. A* **138**(6), 781 (2020).

[23] X. Hongliang, Y. Nan nan, X. Haoran, W. Tianshi, L. Guiying, L. Zhengqiang, *Int. J. Mol. Sci.* **14**, 14185 (2013).

[24] A. Baral, L. Satish, D. Das, H. Sahoo, M. Ghosh, *New J. Chem.* **41**(16), 8130 (2017).

[25] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., ISBN 978-0-12-381479-1, (2012).

[26] Neural network models (supervised).

[27] A. Klos-Witkowska, B. Mlynko, M. Karpinski, V. Martsenyuk, I. Ireifidzh, *ICISD 2020: International Conference on Information Systems Development* 1158 (2020).

[28] Ch. Zhenxing, H. Bowen, R. Xiulian, Ch. Kunyu, X.L. Yuqian, L. Xinning, *Spectroscopy Letters* **51**(6), 279 (2018).

[29] S. Servagent-Noinville, M. Revault, H. Quiquampoix, M.H. Baron, *J. Colloid Interface Sci.* **221**(5), 273 (2000).

[30] V. Martsenyuk, K. Warwas, K. Augustynek, A. Klos-Witkowska, V. Karpinskyi, N. Klymuk, Z. Mayhruk, *2016 16th International Conference on Control, Automation and Systems (ICCAS)*, 489 (2016).