UNIVERSITATEA “BABEŞ-BOLYAI” CLUJ-NAPOCA

FACULTATEA DE FIZICĂ

SPECIALIZAREA BIOFIZICĂ ŞI FIZICĂ MEDICALĂ

LUCRARE DE DISERTAŢIE

Coordonatori științifici

Prof. dr. Leopold Nicolae

Lect. Dr. Ștefania-Dana Iancu

Absolvent

Cimpoeșu Angela-Georgiana

2024

UNIVERSITATEA “BABEŞ-BOLYAI” CLUJ-NAPOCA

FACULTATEA DE FIZICĂ

SPECIALIZAREA BIOFIZICĂ ȘI FIZICĂ MEDICALĂ

LUCRARE DE DISERTAŢIE

REDUCEREA DIMENSIONALITĂŢII DATELOR SPECTRALE SERS PE BIOFLUIDE: O ABORDARE METODOLOGICĂ ŞI UN MODEL DE CLASIFICARE PENTRU IDENTIFICAREA PACIENŢILOR CU CANCER

DIMENSIONALITY REDUCTION OF SERS SPECTRAL DATA ON BIOFLUIDS: A METHODOLOGICAL APPROACH AND A CLASSIFICATION MODEL FOR IDENTIFYING CANCER PATIENTS

Coordonatori științifici

Prof. dr. Leopold Nicolae

Lect. Dr. Ștefania-Dana Iancu

Absolvent

Cimpoeșu Angela-Georgiana

2024

Abstract

This thesis introduces a novel analytical approach to enhance the clinical interpretation of serum Surface-Enhanced Raman Scattering (SERS) spectra for cancer screening. SERS liquid biopsy has gained significant attention in cancer screening due to its ease of use, speed, and low sensitivity. However, the analysis of SERS spectra is often more mathematical than spectrally meaningful, which diminishes the trust medical doctors place in classification results, making cancer screening with SERS seem like a black box. Thus, this thesis aims to develop an alternative to the Principal Component Analysis (PCA) dimensionality reduction technique, specifically designed based on the molecular signatures of metabolites known to adsorb on silver surfaces and produce SERS signals from serum.

A thorough theoretical overview of SERS and its use in biomedical diagnostics is given in Chapter 1. Chapter 2 details the development of a model that focuses on important purine metabolites: uric acid, hypoxanthine, xanthine, urea, and creatinine, in order to reduce dimensionality and increase specificity in cancer classification. To support the development of machine learning classifiers, this model uses Gram-Schmidt orthogonalization to separate metabolite contributions. Chapter 3 discusses the application of this model across three cancer datasets: colorectal, gastrointestinal, and prostate cancers. The results demonstrate notable improvements in prostate cancer classification compared to PCA-based methods, while for colorectal and gastrointestinal cancer, we observed a decrease in classification accuracy. Although PCA performed better for these two datasets, it is possible that contaminants rather than real sample signals were used for the classification, as shown by an analysis of the loading plot of the used PCs, which showed the presence of Crystal Violet, a known contaminant in SERS spectroscopy. The results highlight the potential of the model to improve SERS-based diagnostics by guaranteeing that classifications are made exclusively on the basis of purine metabolites. In order to further improve cancer screening procedures, future research directions include broadening the model application, incorporating cutting-edge machine learning strategies, and investigating unknown spectral contributors.

Table of Contents

Introduction

SERS has a great potential to change the field of liquid biopsy diagnostics by providing sensitive and specific detection of biomolecules, especially in the field of cancer screening, where lots of studies showed performance higher than 80%. However, a major obstacle is the direct assignment of the molecule that is responsible for the prediction of the patients class. In order to address this challenge, this thesis proposes a novel approach to decompose SERS spectra, enabling a more precise and meaningful interpretation of the data. The original contribution of this work is the development and optimization of a decomposition model that effectively reduces the SERS spectra's dimensionality while retaining the chemical information related to purine metabolites. By incorporating the characteristic bands for each metabolite and by using orthogonalization, the model's performance demonstrates similar or improved classification accuracy when compared to traditional PCA-based methods.

The thesis is divided into three main chapters. A thorough overview of the theoretical underpinnings of SERS technology and its application in biochemical analysis is given in the first chapter. It covers the fundamentals of Raman scattering SERS enhancement mechanisms and the role purine metabolites play in biological systems. The chapter also discusses previous studies and methods related to SERS-based metabolite detection, which helps set the scene for the development of the proposed model. The second chapter details the materials and experimental methods used in the study. It explains how to prepare SERS substrates, get SERS spectra from serum samples and pure metabolites, and preprocess the spectral data. The algorithmic approach for spectral decomposition, which makes use of constraints and orthogonalization techniques, is also covered in this chapter.

The third chapter presents the results of the thesis and provides a detailed discussion of the findings. The decomposition model is first validated on pure metabolite spectra, and then it is applied to serum SERS spectra from cancer patients. By contrasting the classification accuracy, the suggested model's performance is evaluated against conventional PCA-based techniques. The implications of the findings for SERS-based diagnostics are also covered in this chapter, along with a chemical interpretation of the decomposed variables. A synopsis of the key findings and the significance of the findings are included in the thesis conclusion. It draws attention to the proposed model's potential impact on metabolite quantification and the advancement of SERS-based diagnostics. The dissertation concludes with two sections: an extensive bibliography that includes a list of all the references cited in the thesis and acknowledgments, which are used to thank people who contributed to the research.

Chapter 1: Theoretical background

1.1 Surface-Enhanced Raman Scattering (SERS) and its clinical applications

The molecules in all matter possess unequal charge distributions along their chemical bonds, resulting in the formation of dipoles. When exposed to light, these charge distributions can be altered, inducing or modifying the molecular dipole moment and generating oscillating electromagnetic fields [1]. This type of interaction provides qualitative and quantitative information on molecule composition [2].

Molecules can remain transparent to light or interact with photons via absorption or scattering [1]. Absorption occurs when the photon energy matches the energy difference between the electronic, vibrational, and rotational states of the molecule, whereas scattering does not require a resonance condition [3]. Scattering efficiency is proportional to the incident light frequency and the molecule's cross-section, with higher frequencies increasing the scattering probability [4].

Raman spectroscopy uses a single wavelength of incident radiation and collects scattered photons. Incident photons distort the molecule's electron cloud, altering its polarizability, which is required for observing the Raman effect [2]. Polarizability refers to the energy needed to distort an electron cloud with an electromagnetic wave, resulting in an induced dipole moment μ.

The induced dipole moment consists of three components: Rayleigh radiation, Stokes scattering, and anti-Stokes scattering. While the majority of incident photons scatter elastically (Rayleigh radiation), a few lose energy to molecular vibrations (Stokes scattering) and gain energy (anti-Stokes scattering). It is important to note that only around one in every million photons is inelastically scattered, making Raman spectroscopy a low-sensitivity method [2]. However, sensitivity can be enhanced by using techniques such as Surface-Enhanced Raman Scattering (SERS) [5].

Raman spectroscopy, especially its advanced method, Surface-Enhanced Raman Spectroscopy (SERS), has garnered considerable interest in the medical field due to its high sensitivity and versatility. This interest stems from SERS's potential to serve as an alternative to current diagnostic tools, especially for cancer detection. Among the various SERS-based diagnostic approaches, SERS liquid biopsy has emerged as the most favored option because it is non- or minimally invasive, requires low sample volumes, and offers quicker turnaround times. Moreover, technological advancements have made spectrometers more accessible and compact, making SERS feasible for practical application in real-world hospital settings [6].

However, the translation of surface-enhanced Raman scattering (SERS) from the optical bench to the clinical environment presents several challenges. These include a lack of understanding of clinical pathways, the absence of standard operating procedures (SOPs), regulatory hurdles, and a limited number of samples in proof-of-principle studies. Nevertheless, among the biofluids analyzed using SERS, urine is the preferred choice. It has been demonstrated that the effectiveness of SERS urine biopsy as a screening tool applies to various cancers [7, 8, 9, 10]. In addition to screening, SERS urine biopsy shows promise as a support tool for reducing unnecessary biopsies in individuals with clinically suspected cancer.

1.2 Detection of cancer using SERS liquid biopsy: purine metabolites

SERS liquid biopsy has potential in medical diagnostics due to its ability to detect purine metabolites, which are indicative of cellular turnover rate, inflammation, and oxidative stress, all of which are disrupted in numerous diseases ranging from cancer to autoimmune disorders.

Purine metabolites are derived from the breakdown of purine nitrogen bases like adenine and guanine and proceed through a metabolic pathway involving inosine, hypoxanthine, xanthine, and ultimately uric acid (Figure 1.1). In contrast, pyrimidine bases such as thymine, cytosine, and uracil are metabolized to simpler compounds like water and carbon dioxide.

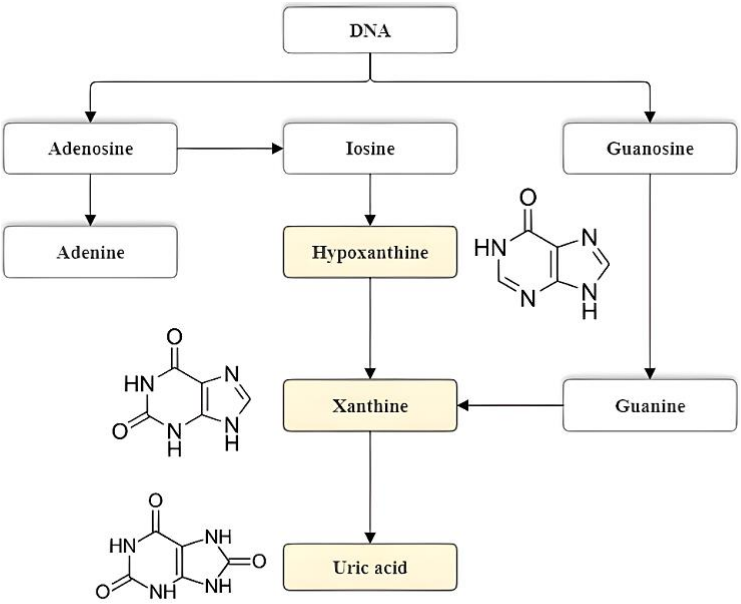

***Figure 1.1**: A simplified diagram of the purine metabolism decomposition*

Numerous medical conditions are associated with elevated purine metabolite levels [11]. For example, cancer patients often show increased uric acid and hypoxanthine levels compared to healthy individuals due to accelerated cellular turnover [11]. However, the biomedical implications of purine metabolism are nuanced, as elevated uric acid levels can be detrimental in conditions like gout, chronic kidney disease, or metabolic syndrome [12, 13, 14], yet potentially protective against diseases such as Parkinson's, multiple sclerosis, and Alzheimer's [15, 16].

Beyond uric acid, variations in xanthine and hypoxanthine levels are also linked to specific diseases. Studies have indicated lower serum levels of xanthine and hypoxanthine in patients with colorectal cancer and lower urinary hypoxanthine in those with non-Hodgkin lymphoma [17, 18]. Moreover, the balance between these purine metabolites is crucial; reduced activity of xanthine oxidoreductase in cancer leads to imbalances among hypoxanthine, xanthine, and uric acid [19].

While the precise relationship between purine metabolism and certain diseases is not fully understood, these findings underscore the potential of methods like SERS liquid biopsy to identify purine metabolite imbalances, aiding in the diagnosis of various diseases.

1.3 Methods for analyzing SERS spectra: machine-learning (ML)

Traditional analytical techniques no longer meet current demands. SERS enables the detection of biofluid metabolites that adsorb to metal nanoparticles, while the free, bulk molecules remain undetected in the SERS spectrum [20, 21]. From the approximately 4000 metabolites typically present in human serum [22], only those with a high affinity for the metal substrate are detected by SERS, with purine metabolites being the most prominent [23, 24]. Therefore, the SERS spectra of biofluids will be dominated by purine metabolite bands, with other contributions often obscured by overlapping bands or high noise. Due to this, using SERS in clinical applications and as a diagnosis technique is of high interest because, as mentioned in the previous chapter, purine metabolites are linked to a variety of diseases. However, some unidentified metabolites can also contribute to the SERS spectrum, and given the large size of SERS spectroscopy datasets, visual analysis is impractical. But precise spectral signature assignment is not necessary when employing machine learning (ML) algorithms for classification based on spectral shapes [25], and ML models can be used in order to process these signals effectively. Consequently, in medical diagnostics, SERS functions as a "black box" system where spectra serve as inputs and ML algorithms classify them without disclosing the contributing spectral features. However, for the ML approach to be trustworthy, a medical and chemical interpretation is needed.

ML can be applied to various tasks, but this thesis focuses on ML algorithms' clustering and classification applications. ML algorithms learn from a training dataset to perform well on a new, unseen test set. The task, typically a classification problem, is evaluated by performance metrics. ML classifiers learn from the training set by extracting relevant features and thresholds and applying this experience to new datasets [26].

The performance of ML models can be adjusted by altering error thresholds, discrimination functions, or model hyperparameters. Hyperparameters, which control ML model behavior, are optimized through a trial-and-error process. Performance metrics are determined using a validation set, distinct from the training and test sets [27].

ML involves three key steps: gaining experience from a training set, task-specific hyperparameter tuning analyzed by a validation set, and evaluating the model's classification accuracy on a test set. Training and test sets should be independent yet identically distributed [27]. For classification models, different samples should be used for training and testing, maintaining proportional label distribution. Validation and test sets should constitute 10–40% of the training set samples [28]. When sample numbers are small, cross-validation, where the dataset is repeatedly split into training and validation sets, is an alternative. ML algorithms are optimized on the training set, with hyperparameters adjusted to minimize training error or maximize classification accuracy. Optimal hyperparameters are determined outside the learning algorithm using another dataset [27].

Significant discrepancies between training and test errors indicate underfitting (high training error) or overfitting (low training error). Adjusting hyperparameters or selecting optimal models can mitigate these risks. ML algorithms are categorized as unsupervised or supervised based on the information available during the learning phase. Unsupervised algorithms learn from feature-only datasets, analyzing probability distributions or clustering tendencies, while supervised algorithms have access to datasets with features and associated labels or targets. Unsupervised algorithms, like Principal Component Analysis (PCA), which will be discussed later, are more reliable than supervised ones. Supervised algorithms extract relevant features to discriminate labels or targets, applying this experience to predict labels or targets on new datasets [29].

ML algorithms are also classified as parametric or nonparametric based on their mathematical approach. Parametric models use a finite parameter vector to describe a function based on input features, whereas nonparametric models use algorithms based on all possible probability distributions, not directly correlated with input features [27]. ML algorithm performance is generally evaluated by overall classification accuracy, F1 score, precision, recall, and specificity. Precision is the ratio of true positive outputs to the sum of true positive and false positive results. Recall, or sensitivity, is the ratio of true positive results to the sum of true positive and false negative results. Specificity is the ratio of true negatives to the sum of true negatives and false negatives. The F1 score is the harmonic mean of precision and recall [29]. A receiver operating characteristic curve, based on recall and specificity, provides an overview of the ML model's classification performance. The area under the curve (AUC) shows the trade-off between recall and specificity. The most reliable performance analysis involves constructing a confusion matrix, which includes the exact numbers of false and true negatives and positives [29].

In the next subsections, the different ML models used in this thesis will be described.

1.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a foundational multivariate analysis technique. It is an unsupervised method that reduces the number of variables while preserving as much variation as possible from the original dataset. This reduction is achieved by calculating principal components (PCs), which are uncorrelated vectors that maximize the variance in the dataset [30]. Each PC is a linear combination of the original variables, with coefficients indicating their significance and relationships [31].

PCA's main advantage is dimensionality reduction while maintaining the ability to interpret the data in the context of the original variables [32]. When using PCA, it's important to note that it operates as an unsupervised technique, meaning it does not consider class labels in the dataset. The selection of PCs is based purely on statistical criteria, specifically maximizing variance across the entire dataset. PCs are prioritized based on the amount of variance they explain; those explaining greater variance are considered more significant than those explaining less. However, lower variance data might erroneously be disregarded as "noise," potentially overlooking valuable information. In classification tasks, the features that maximize discrimination between groups may differ from those that maximize overall dataset variance [30]. Thus, PCA's application may inadvertently eliminate features crucial for discrimination while retaining those with a broader data spread.

1.3.2 Gaussian Naive Bayes (GNB)

Gaussian Naïve Bayes (GNB) algorithms leverage Bayes' theorem, assuming feature independence, to perform probabilistic classification effectively. This method calculates the likelihood of an event based on prior knowledge, adjusting probabilities as new evidence is introduced [28]. GNB evaluates the probability of each feature given a hypothesis (target class), multiplying these probabilities to determine the likelihood of the hypothesis given the observed data. The hypothesis with the highest probability is chosen as the predicted output. GNB assumes feature independence and a normal distribution of data, making it straightforward and quick for classification tasks. Unlike many classifiers, GNB requires no hyperparameter tuning, simplifying its application in various contexts [33].

1.3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) applies the concept of determining a boundary to divide classes, applicable in higher-dimensional spaces. SVM transforms features into points in an n-dimensional space and selects a hyperplane that best divides the classes. If the classes are not linearly separable, a kernel function (e.g., linear or polynomial) transforms the data into a higher dimension where separation is possible [34].

In SVM, a hyperplane is sought in the transformed feature space to delineate classes, maximizing the margin to the nearest points of each class, known as support vectors. These critical samples are weighted to optimize the separation between classes [35], enhancing the model's resilience to outliers [36]. The selection of support vectors is automated based on boundary conditions, affecting SVM's sensitivity to sample size [37]. Choosing an appropriate kernel involves iterative testing, necessitating distinct training, validation, and test datasets or cross-validation. Hyperparameters such as cost (C), determining the trade-off between classification accuracy and margin width, and complexity bound (v), setting the error threshold, are crucial for SVM optimization. The optimal C value hinges on data noise and variance, inversely affecting the hyperplane's width [29].

1.3.4 Logistic Regression (LR)

Logistic regression (LR) is widely utilized in medical contexts for binary outcomes [38]. It models the probability of belonging to a class using a sigmoid function, which outputs values between 0 and 1 based on a weighted sum of input features. Predictions are made by comparing these probabilities: samples with a probability less than 0.5 are classified as class 0, and those with a probability greater than 0.5 are classified as class 1. LR allows for the assessment of feature importance in outcome classification [28].

The training of an LR model involves determining optimal weights or coefficients for each feature through iterative testing of different combinations until the best fit is achieved. Once the optimal coefficients have been established, the conditional probabilities for each observation are computed, logged, and summed to produce predicted probabilities. LR captures the relationship between variables and outcomes via logarithmic odds, offering flexibility beyond simple linear relationships [28].

However, LR faces challenges with datasets containing a high number of features, where it can overfit. Regularization techniques are employed in such cases to penalize large coefficients. This regularization helps prevent overfitting by adjusting the model's complexity, thereby improving generalization to new data [28]. It's important to note that LR assumes input features are independent, which may not hold true in all practical scenarios.

1.3.5 AdaBoost (Adaptive Boosting)

AdaBoost (Adaptive Boosting) is a machine learning ensemble method that focuses on improving the classification accuracy of weak classifiers. It sequentially trains a series of classifiers, with each subsequent classifier giving more weight to incorrectly classified instances from the previous classifier. This adaptive weighting allows AdaBoost to prioritize difficult-to-classify instances, thereby enhancing overall model performance. The final prediction is determined through a weighted majority vote of all classifiers, with higher weights assigned to more accurate classifiers [39].

AdaBoost is particularly effective in scenarios where datasets are complex and contain noise, as it can adjust to emphasize problematic instances during training. By iteratively reweighting data based on previous classifier performance, AdaBoost ensures that subsequent models focus more on previously misclassified instances, progressively improving the overall classification accuracy [40].

1.3.6 Random Forest (RF)

Random Forest (RF) comprises multiple decision trees (DTs) that operate independently and collectively. Each tree analyzes a random subset of the dataset, extracting features that best distinguish between groups [23]. To elaborate, RF uses a technique called bagging, where each tree is trained on a bootstrap sample (a random selection with replacement from the original data). This ensemble method aggregates predictions from all trees via voting. Hyperparameters such as the number of trees can be adjusted to control model complexity and performance, with more trees potentially enhancing training accuracy at the risk of overfitting [28].

RF offers robustness against outliers and noisy data, requiring minimal preprocessing [41]. It mitigates overfitting compared to individual decision trees [42] and avoids significant user intervention by being less sensitive to preprocessing variations [43]. However, RF benefits from large datasets for effective training and may encounter challenges when applied directly to spectral data due to the inherent structure of its decision trees [36, 38].

1.3.7 k-Nearest Neighbors (kNN)

The k-Nearest Neighbors (kNN) model is a nonparametric learning approach that relies on the assumption that similar objects are located close to each other in the initial feature space. By calculating the distances to the k nearest neighbors, kNN determines the class closest to an unknown sample. These distances are not limited to Euclidean measures; various distance metrics can be employed depending on the task. Additionally, tuning the number of neighbors evaluated per sample is crucial for optimal performance. However, outliers can significantly impact the kNN model since all samples contribute to distance calculations [28].

In summary, most machine learning classifiers assume feature independence, which may not hold true for datasets like vibrational spectra, where features describing the same vibrational band are highly correlated. Consequently, employing ML models directly on spectral features violates this assumption. Therefore, techniques such as PCA should precede ML classifiers to transform initial features into orthogonal ones. Careful selection of the reduction technique is essential to avoid losing critical information. For instance, in complex datasets containing predominantly noise with occasional significant signals, applying dimensionality reduction methods risks losing valuable signals [44].

Chapter 2: Materials and methods

2.1 Patient population and cancer datasets

The study was conducted using three distinct datasets, each representing a different type of cancer: prostate cancer (29 patients and 22 controls), colorectal cancer (91 patients and 31 controls), and gastrointestinal cancer (53 patients and 25 controls). Gastrointestinal cancers encompass a range of malignancies affecting the stomach, esophagus, and other parts of the digestive system.

The cancers selected for analysis were chosen due to their prevalence as some of the top five most frequent cancers, as reported by the International Agency for Research on Cancer's (IARC) Global Cancer Observatory. In Romania, prostate cancer is the most frequently diagnosed cancer among males, with colorectal cancer ranking third and stomach cancer ranking fifth, as indicated in the 2022 Factsheet. Additionally, colorectal cancer ranks second among female cancers, making it the most prevalent cancer in both sexes. In 2022, there were 13,541 cases of colorectal cancer [45]. In terms of mortality, colorectal cancer holds the second position on the list, with a total of 7,381 deaths reported in Romania in 2022. Stomach cancer ranks sixth, and prostate cancer ranks seventh during the same year [45]. Lastly, it is essential to recognize that colorectal cancer ranks third globally in terms of prevalence, accounting for an estimated 10% of all cancer cases, and holds the position of the second leading cause of cancer-related fatalities worldwide [46].

2.2 Sample collection and preparation procedures

The study involved analyzing existing data (SERS spectra) on serum samples from gastrointestinal and prostate cancer patients acquired using the same portable Raman spectrometer. Additionally, we collected new data from colorectal cancer patients. The latest study was approved by the Ethical Committee of the 1st Surgical Clinic, County Emergency Clinical Hospital in Cluj-Napoca, and informed consent was obtained from all patients enrolled in the study. Blood samples were collected in tubes that promote clotting and then stored at 4°C for one hour. This allows the serum to separate from the clot. Next, centrifugation (spinning) at high speed (2000xg) for 15 minutes further isolated the serum. The separated serum was then frozen at -80°C for preservation. Before analysis, the serum underwent protein removal using a precipitation technique. This involved mixing the serum with methanol (at a 1:9 ratio) and centrifuging again at an even higher speed (5800xg) for 15 minutes. Finally, the protein-free liquid (supernatant) was collected for further testing.

2.3 Silver nanoparticles synthesis

Silver nanoparticles were synthesized by the reduction of silver nitrate using hydroxylamine hydrochloride (hya-AgNPs). To summarize the procedure, 17 mg of silver nitrate was dissolved in 90 mL of ultrapure water (Millipore). Simultaneously, 17 mg of hydroxylamine hydrochloride was dissolved in 8.8 mL of ultrapure water and mixed with 1.2 mL of 1% sodium hydroxide solution. The hydroxylamine solution was swiftly added to the silver nitrate solution under vigorous stirring, resulting in the rapid formation of colloidal Ag nanoparticles, evident from the solution's change in color to brown-grey.

2.4 SERS spectra acquisition

A portable Raman spectrometer (i-Raman BWS415-532H, BWTek) with an attached Raman video microsampling module (BAC151, BWTek) was used to acquire the SERS spectra of serum samples. 5 µl of deproteinized serum were mixed with 45 µl hya-AgNPs and $5x10^{-4}$ M $Ca(NO_3)_2$. 5 µl of the mixture were deposited onto a microscope glass slide, covered with aluminum foil. A 20x objective (N.A. 0.4) was used for focusing the 532 nm laser line (~6 mW power) on the sample. For each spectrum, we acquired four acquisitions, five seconds each.

2.5 SERS spectra analysis

2.5.1 Purine metabolites contributions to SERS spectra of serum

As indicated in the previous chapter, the main contributors to the SERS spectra of serum are purine metabolites. Therefore, we know the correlation between the SERS bands and their meaning. To highlight the contributions of purine metabolites in the SERS spectrum of biofluids, the SERS spectra of uric acid, hypoxanthine, xanthine, urea and creatinine were acquired (Figure 2.1). As observed in Figure 2.1. the major SERS bands in the serum SERS spectra overlap with the SERS bands characteristic to purine metabolites (533, 637, 811, 889, 1017, 1136, 1365 cm^{-1} uric acid, 724 cm^{-1} hypoxanthine, 1206 cm^{-1} xanthine, 1518, 1575 cm^{-1} urea and 1616 cm^{-1} creatinine).

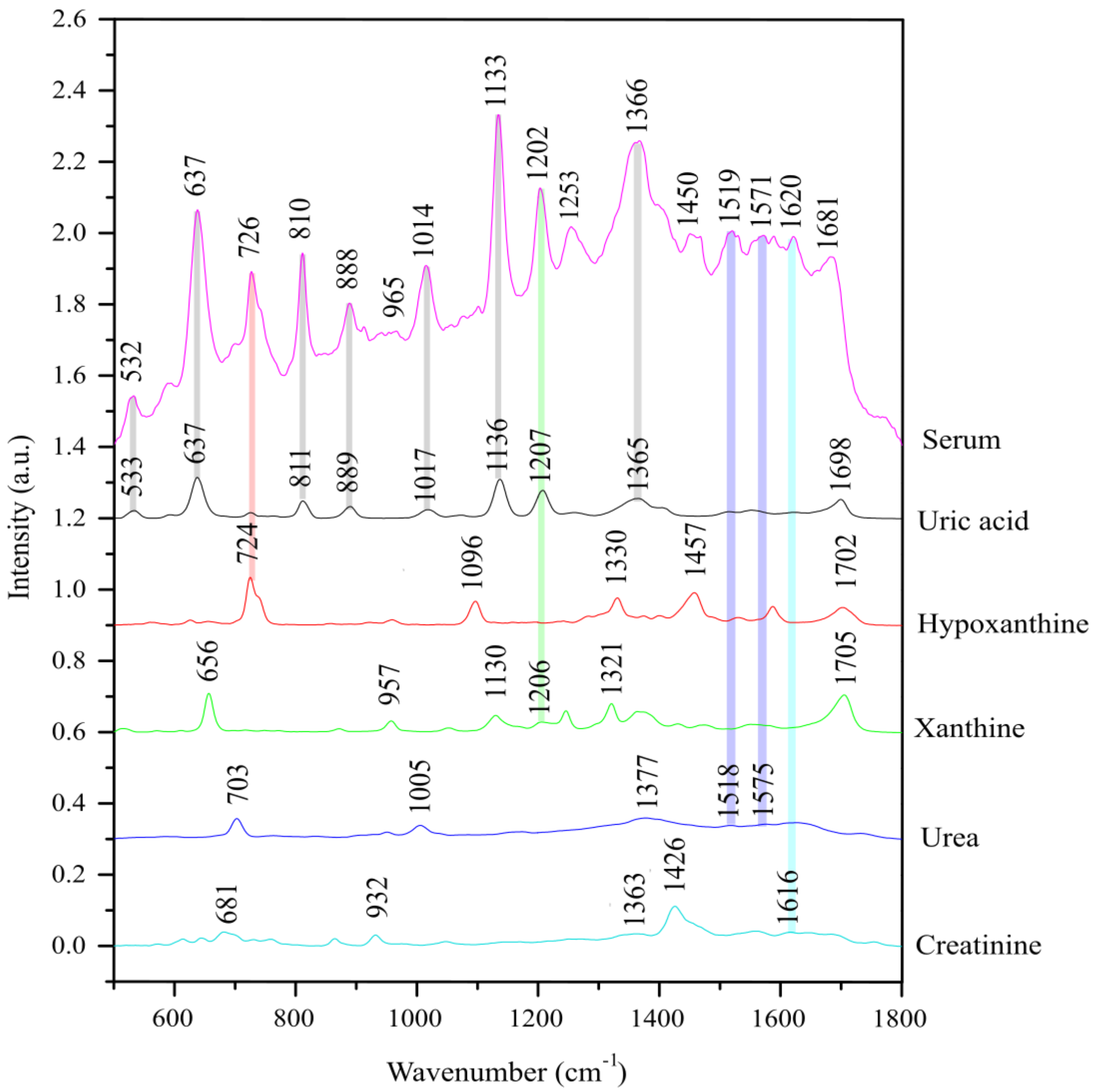

Figure 2.1*: SERS spectra of the purine metabolites: uric acid, hypoxanthine, xanthine, urea, and creatinine and mean SERS spectra of serum*

2.5.2 Dimensionality reduction of SERS spectral data

Our first objective was to devise an effective method for reducing the dimensionality of SERS spectral data in order to be able to construct a classification approach centered on the purine metabolites. As a first step, we developed an algorithm in GNU Octave to decompose the serum SERS spectra onto relevant purine metabolites in three distinct scenarios:

- **Scenario 1**: Decomposition of the serum SERS spectra onto relevant purine metabolites without imposing constraints;
- **Scenario 2**: Decomposition of the serum SERS spectra onto relevant purine metabolites with constraints;
- **Scenario 3**: Decomposition of the serum SERS spectra onto orthogonalized relevant purine metabolites.

In each scenario, the purine metabolites SERS spectra presented in Figure 2.1 were employed as separate metabolite vectors in the algorithm, and they will be subsequently referred to as metabolite vectors.

The angle between two vectors reveals the similarity between them. To project one vector onto another, the dot product is calculated after each vector is normalized to the unit. In essence, spectra are vectors of different magnitudes, and calculating the dot product of the biofluid's SERS spectra with a vector of a particular metabolite will later reveal the metabolite contribution to the biofluid SERS spectra. The SERS spectra projected onto the individual metabolite vectors lead to the score for every vector.

In the last step, comparing composed and original spectra requires the inverse transformation of the result back into the initial space using a linear combination of the scores and metabolite vectors. The algorithms were tested and validated using a data set containing SERS spectra of the purine metabolites. This data set was projected onto the metabolite's vectors, and the sample scores on the new feature were calculated. The scores obtained for the individual samples with respect to the various vectors represented the contribution of each metabolite in the SERS spectrum of the analyzed data set.

After testing and validating the algorithms in the three aforementioned scenarios, we proceeded to apply the optimal algorithm (corresponding to the third scenario) to datasets that contained SERS spectra of serum from patients with prostate, colorectal, and gastrointestinal cancer and controls.

2.6 Dimensionality reduction model (proposed model) creation and flow

The usual pipeline for building a classification model is SERS acquisition, dimensionality reduction using PCA, and employing machine learning classifiers, with the ultimate goal of classifying cancer patients and controls using the SERS spectra of serum.

Following this pipeline, for our proposed model, the serum SERS spectra were analyzed using Quasar-Orange software (Bioinformatics Laboratory of the University of Ljubljana) [47]. Next, we processed the serum SERS spectra by projecting them onto the relevant orthogonalized purine metabolites. In this way, we reduced the dimensionality of the data sets from thousands to five variables. Using the new variables, we build several Machine Learning (ML) techniques, including Naive Bayes, Support Vector Machine (SVM), Logistic Regression, AdaBoost, Random Forest, and k-nearest neighbors (kNN), to classify cancer patients and controls.

We compared the performance of our proposed model with a classic model using PCA, the well-known method for dimensionality reduction of the data set, while retaining as much as possible of the variation in the data set. For both models, the hyperparameters were tuned to achieve the best classification accuracy and kept to allow the comparison. Figure 2.2 depicts the models created using Quasar-Orange software for prostate cancer. The same classification models were used for the other data sets.

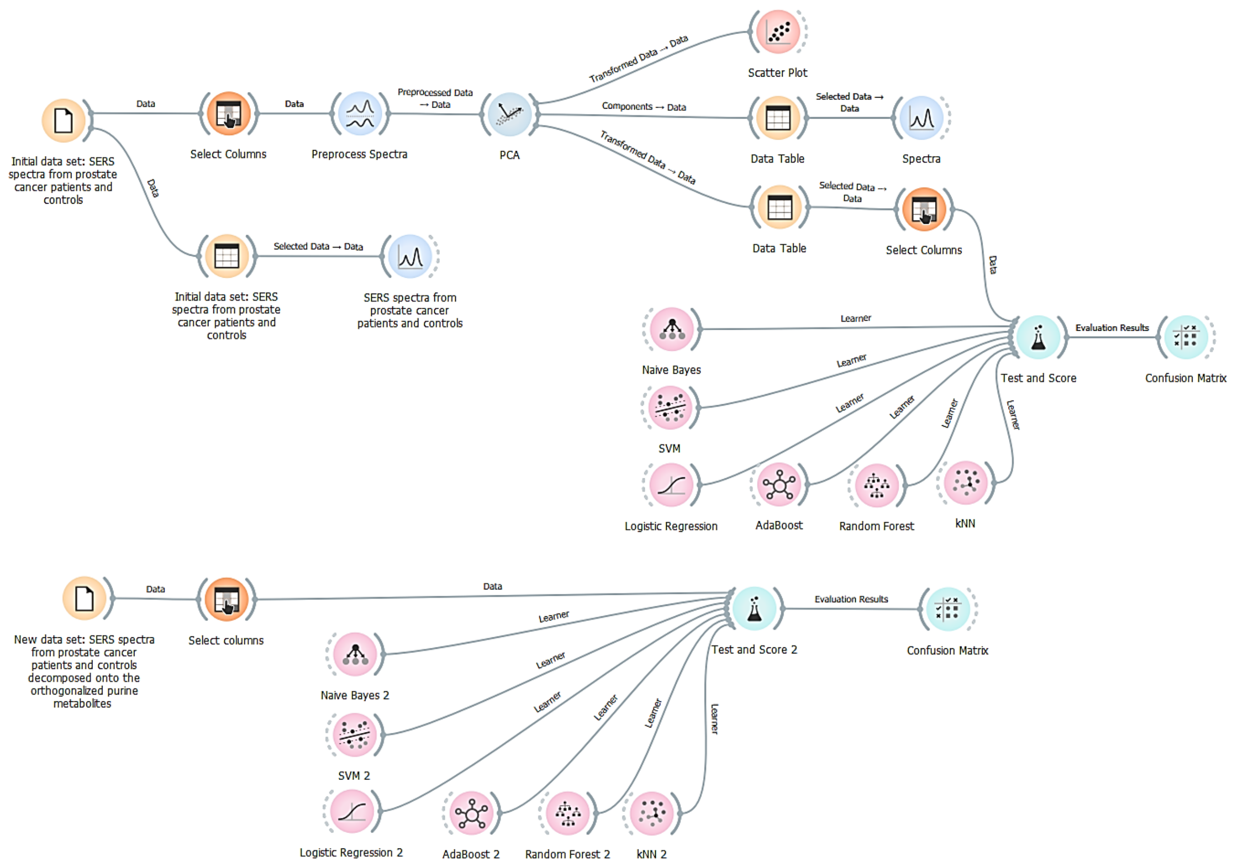

***Figure 2.2**: Classification models created using Quasar-Orange for prostate cancer. **(Top)**: Classification model using PCA **(Bottom)**: The proposed classification model based on the data set obtained by decomposing the serum SERS spectra onto the orthogonalized vectors of the purine metabolites*

2.7 Enhancing classification accuracy with residual analysis

Following the analysis of the three datasets containing serum SERS spectra from patients with prostate, colorectal, and gastrointestinal cancer, along with the comparison of MLs classification accuracy between the proposed model and the PCA-based model, we wanted to further improve the classification accuracy. In order to achieve this objective, we investigated potential contributions that were not considered or were excluded during the decomposition of spectra onto relevant orthogonalized purine metabolites.

Therefore, we developed a code using GNU Octave to obtain residues for each cancer dataset. This was achieved by subtracting the recomposed spectra after decomposition onto relevant orthogonalized datasets from the initial dataset. Running the code on each dataset yielded the residual data and spectra.

After obtaining and visualizing the residuals and their bands, we followed a similar approach to the previous section. Using Quasar-Orange and the same MLs, we built a residual-based model and a model using the initial datasets. This allowed for a comparison between the two models with comparable dataset sizes, ensuring a fair comparison. As for the previous models, the hyperparameters were adjusted to obtain the highest classification accuracy and kept for comparison.

In Figure 2.3, we depict the initial data-based model and the residuals-based model created using Quasar-Orange for prostate cancer. The same classification models were used for the other datasets.

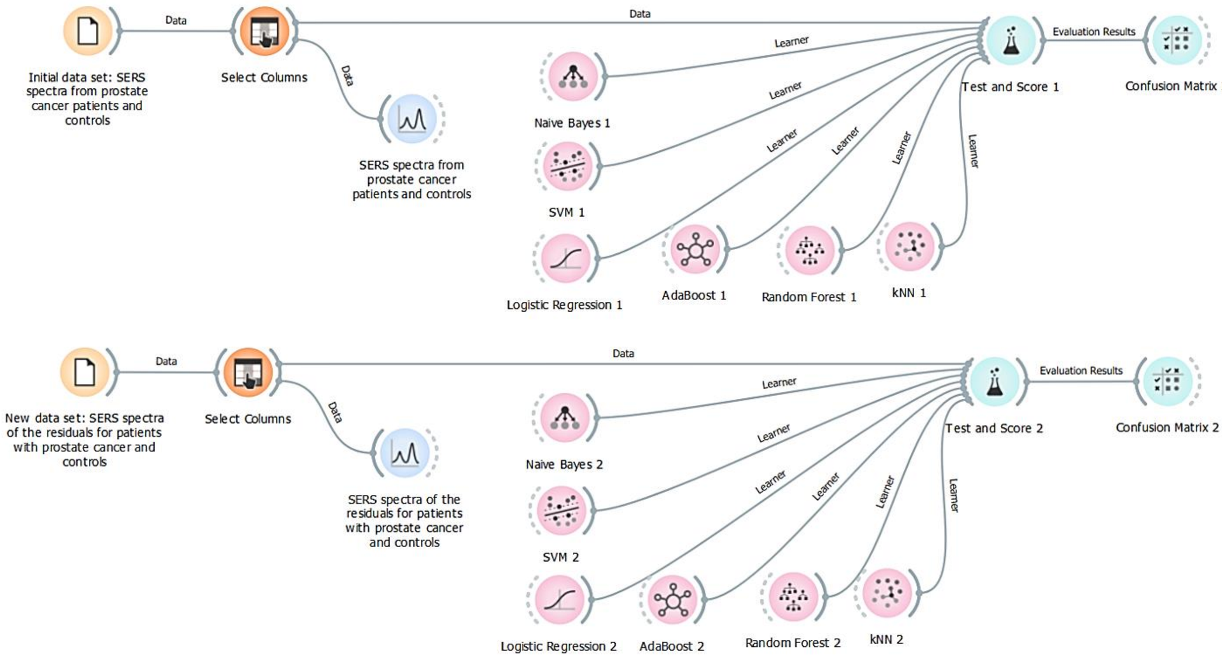

Figure 2.3*: Classification models created using Quasar-Orange for prostate cancer.* ***(Top)****: Classification model using the initial data set* ***(Bottom)****: The classification model based on the serum SERS of the residuals*

Chapter 3: Results and discussion

Our objective is to create a model for reducing the dimensionality of SERS spectra, incorporating the known correlations of SERS bands associated with detected purine metabolites. Given that purine metabolites are key contributors to serum SERS spectra, we map the SERS spectra onto the known spectra of each purine metabolite to derive a score indicating the 'concentration' of each metabolite in the samples. This approach aims to provide a chemical interpretation for SERS liquid biopsy diagnostics and enhance the quantitative detection of metabolites using SERS.

3.1 Decomposition model optimization. Validation on SERS spectra of purine metabolites

Initially, we evaluated the proposed scenarios using the SERS spectra of pure metabolites. We expected that when the SERS spectrum of a metabolite is decomposed onto itself, it would yield a score of 1, while its decomposition onto other metabolites would result in a score of 0.

3.1.1 Scenario 1: No constraints imposed

In the first scenario, the algorithm was utilized without imposing any constraints.

Table 3.1: *The dot product of the SERS spectra of uric acid, hypoxanthine, xanthine, urea, creatinine, and themselves for Scenario 1*

Actual/predicted	**Uric acid**	**Hypoxanthine**	**Xanthine**	**Urea**	**Creatinine**
Acid uric	**1**	0.43387	0.68439	0.59855	0.52086
Hypoxanthine	0.43387	**1**	0.58484	0.66552	0.69032
Xanthine	0.68439	0.58484	**1**	0.67809	0.61917
Urea	0.59855	0.66552	0.67809	**1**	0.87424
Creatinine	0.52086	0.69032	0.61917	0.87424	**1**

The results obtained for this scenario do not achieve the desired goals (Table 3.1). We have expected the dot product between two different metabolites to be 0, and 1 for the dot product of the metabolite with itself. This discrepancy may be explained by the fact that the purine metabolites have many common bands with each other (vectors are not orthogonal). Figure 2.1 highlights numerous regions in the spectra that have overlapping bands. Therefore, the model will have difficulties accurately recognizing each purine metabolite contribution to the spectra.

3.1.2 Scenario 2: Constraints imposed

Considering the results for the first scenario, in this scenario we included constraints in the algorithm with the aim of achieving a better distinction of the purine metabolites in the spectra by the model. Therefore, we analyzed the purine metabolites SERS spectra and searched for isolated, characteristic bands for each of the metabolites. We identified and selected the 811 cm^{-1} band for uric acid, the 1246 cm^{-1} band for xanthine, the 703 cm^{-1} band for urea, and the 681 cm^{-1} band for creatinine. For hypoxanthine, an isolated band couldn't be identified; however, the 724 cm^{-1} band was selected as it's the most intense. The SERS spectra of the purine metabolites with the selected bands are illustrated in Figure 3.1.

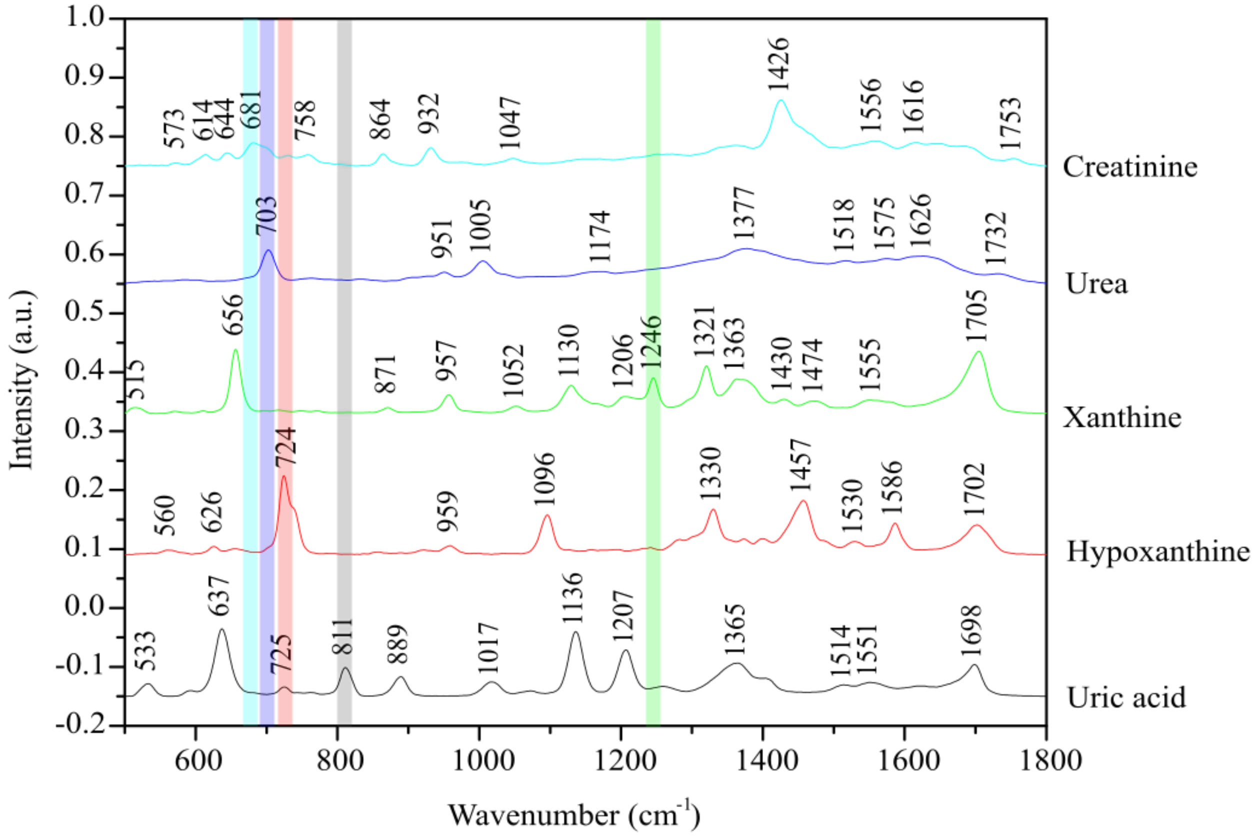

Figure 3.1*: SERS spectra of the purine metabolites with the selected bands for scenario 2*

In this second scenario, the algorithm will calculate the score of the SERS spectra on the metabolite's vectors only if the selected bands of the respective metabolites are present in the spectra. Table 3.2 presents the results for Scenario 2.

***Table 3.2**: The dot product of the SERS spectra of uric acid, hypoxanthine, xanthine, urea, creatinine, and themselves for Scenario 2*

Actual/predicted	Uric acid	Hypoxanthine	Xanthine	Urea	Creatinine
Acid uric	**1**	0.43387	0	0	0
Hypoxanthine	0	**1**	0	0	0.69032
Xanthine	0	0	**1**	0.67809	0
Urea	0	0	0	**1**	0
Creatinine	0	0	0	0	**1**

The results obtained for this scenario are improved. However, some overlapping between hypoxanthine and uric acid, as well as between xanthine and urea, cannot be avoided due to common bands.

3.1.3 Scenario 3: Decomposition onto orthogonalized purine metabolites

In the last scenario, a Python script was employed as a preprocessing step using the Gram-Schmidt process in order to orthogonalize the purine metabolites vectors to uric acid SERS spectrum. The orthogonalized spectra are represented in Figure 3.2.

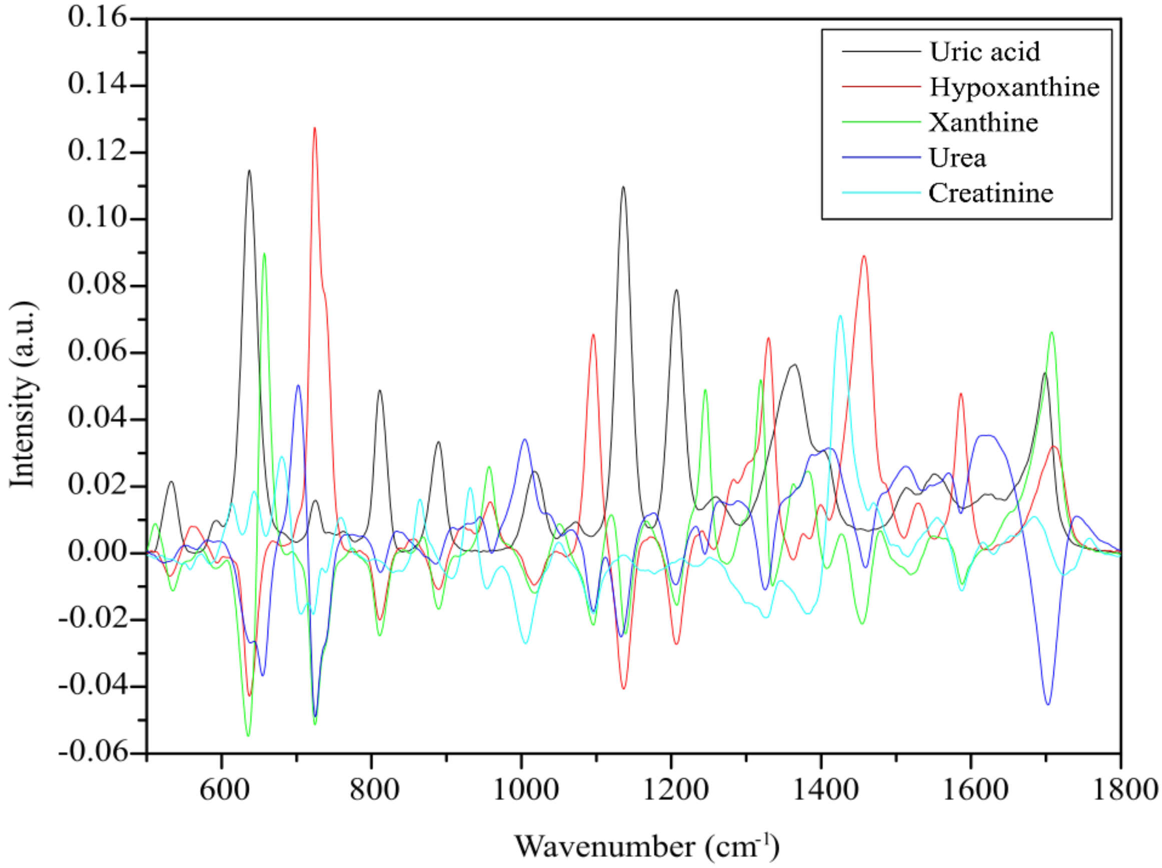

***Figure 3.2**: SERS spectra of the orthogonalized purine metabolites to uric acid*

The orthogonalization process ensured the elimination of correlations and overlaps between the metabolites and facilitated differentiation between them in the spectra. The algorithm was validated and tested using the SERS spectra of the orthogonalized purine metabolites. The results are included in Table 3.3.

***Table 3.3**: The dot product of the SERS spectra of uric acid, hypoxanthine, xanthine, urea, creatinine, and themselves for Scenario 3*

Actual/predicted	Uric acid	Hypoxanthine	Xanthine	Urea	Creatinine
Acid uric	**1**	0	0	0	0
Hypoxanthine	0	**1**	0	0	0
Xanthine	0	0	**1**	0	0
Urea	0	0	0	**1**	0
Creatinine	0	0	0	0	**1**

As it was expected in the beginning, the dot product between two different metabolites is 0, and 1 for the dot product of the metabolite with itself. This scenario provided the desired results and offered the possibility of an accurate and distinct identification of purine metabolites, making it the optimal algorithm to be used in the analysis of serum SERS spectra from cancer patients in subsequent applications.

3.2 Proposed model and PCA-based model classification accuracy on data sets consisting of SERS spectra of serum from patients with cancer

As described in Chapter 2, after the testing and validation of the algorithms in the three aforementioned scenarios, the optimal algorithm (corresponding to the third scenario) was applied to the three studied datasets, obtaining new datasets reduced in dimensionality (five variables with chemical interpretation).

The mean serum SERS spectra of the three analyzed data sets are depicted in Figure 3.3. As anticipated, upon analyzing the mean SERS spectra from cancer patients and control subjects, it is evident that the spectra are primarily composed of peaks originating from purine metabolites (see Figure 2.1).

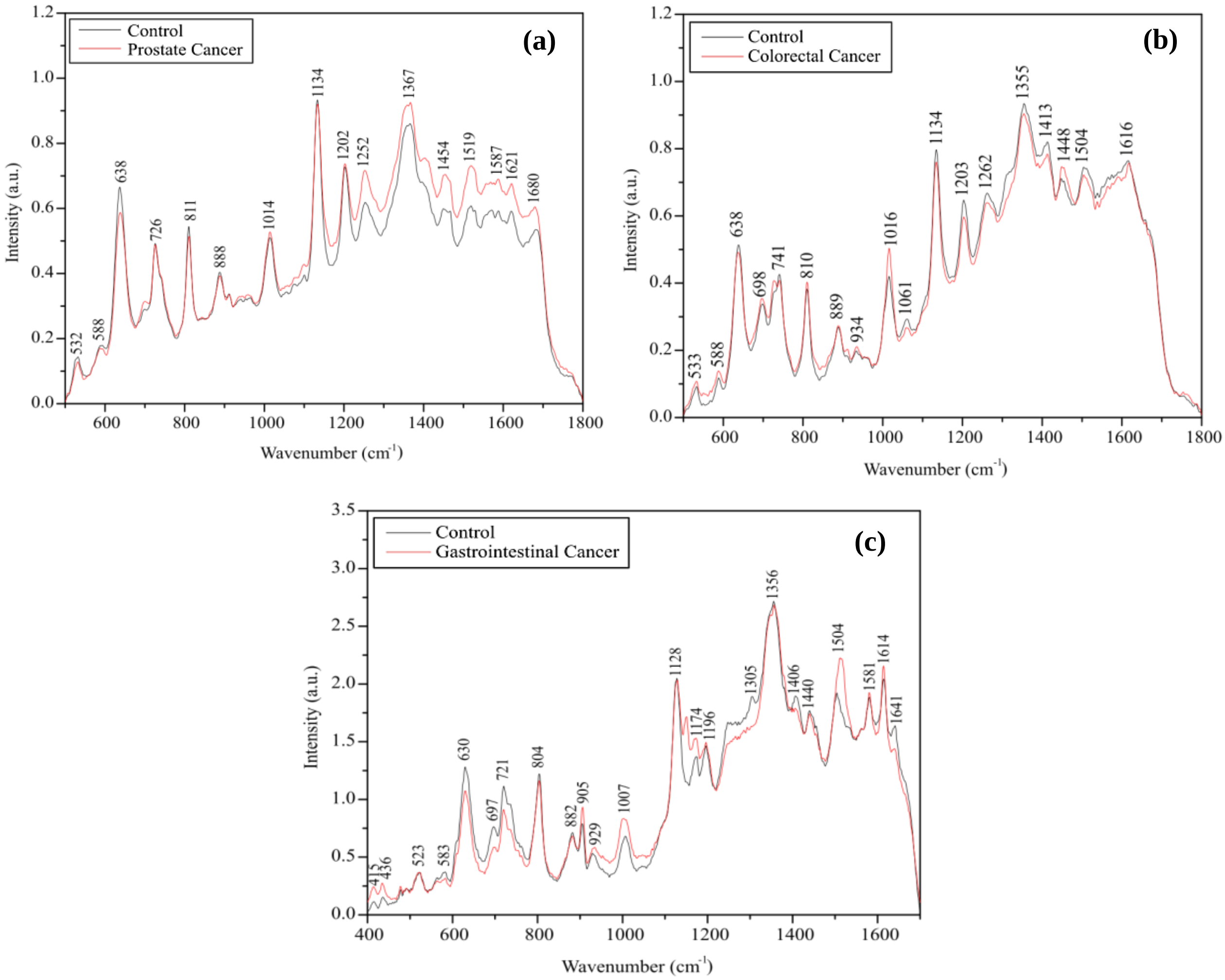

Figure 3.3*: Mean serum SERS spectra of patients with* ***(a)*** *prostate cancer and controls,* ***(b)*** *colorectal cancer and controls, and* ***(c)*** *gastrointestinal cancer and controls*

Using the decomposed variables (five variables for each dataset), we compared our proposed dimensionality reduction model to the classic model using PCA in terms of the resultant performance in the classification task. To ease the review of the results, in the subsequent tables, the cells are colored yellow if the classification accuracy has the same value for both models, green if the value for the classification accuracy is better for the proposed dimensionality reduction model, and red if the classification accuracy has decreased for the proposed model.

3.2.1 Classification accuracy of classification models for prostate cancer

The classification accuracy of the MLs employed on the purine metabolites score, and on the PCs score for prostate cancer can be found in Table 3.4.

Table 3.4: *Classification accuracy of the MLs for the PCA-based model and our proposed model for prostate cancer*

ML		PCA-based model	Proposed model
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		66.7	66.7
SVM	Linear	72.2	72.2
	Polynomial	66.7	66.7
	RBF	66.7	88.9
	Sigmoid	72.2	50
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	61.1	61.1
	Lasso, BCD	66.7	66.7
	Ridge, NBCD	61.1	66.7
	Ridge, BCD	61.1	61.1
	None, NBCD	50	61.1
	None, BCD	50	50
AdaBoost		38.9	61.1
Random Forest		61.1	77.8
kNN	Metric: Euclidian	72.2	88.9
	Metric: Manhattan	83.3	88.9
	Metric: Chebyshev	83.3	77.8

For prostate cancer, for the majority of ML, the results obtained on the purine metabolites score indicated consistency or even an improvement in classification accuracy compared to the PCA-based model.

3.2.2 Classification accuracy of classification models for colorectal cancer

The classification accuracy of the MLs employed on the purine metabolites score, and on the PCs score for colorectal cancer is detailed in Table 3.5.

***Table 3.5**: Classification accuracy of the MLs for the PCA-based model and our proposed model for colorectal cancer*

ML		PCA-based model	Proposed model
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		82.1	74.4
SVM	Linear	79.5	74.4
	Polynomial	82.1	76.9
	RBF	89.7	76.9
	Sigmoid	76.9	76.9
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	82.1	76.9
	Lasso, BCD	79.5	71.8
	Ridge, NBCD	84.6	76.9
	Ridge, BCD	84.6	76.9
	None, NBCD	79.5	71.8
	None, BCD	74.4	61.5
AdaBoost		71.8	61.5
Random Forest		87.2	87.2
kNN	Metric: Euclidian	92.3	84.6
	Metric: Manhattan	92.3	84.6
	Metric: Chebyshev	92.3	82.1

Despite obtaining high classification accuracy values for some ML algorithms, mostly above 70% (with a maximum of 87.2% for random forest), for colorectal cancer, the proposed model has a decrease in classification accuracy compared to the PCA-based model.

3.2.3 Classification accuracy of classification models for gastrointestinal cancer

The classification accuracy of MLs employed on the purine metabolites score, and on the PCs score for gastrointestinal cancer can be found in Table 3.6.

Table 3.6: *Classification accuracy of the MLs for the PCA-based model and our proposed model for gastrointestinal cancer*

ML		**PCA-based model**	**Proposed model**
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		62.5	50
SVM	Linear	70.8	66.7
	Polynomial (order: 0.5)	62.5	62.5
	RBF	79.2	58.3
	Sigmoid	70.8	62.5
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	58.3	83.3
	Lasso, BCD	58.3	58.3
	Ridge, NBCD	58.3	66.7
	Ridge, BCD	58.3	58.3
	None, NBCD	54.2	83.3
	None, BCD	58.3	58.3
AdaBoost		58.3	37.5
Random Forest		79.2	75
kNN	Metric: Euclidian	87.5	62.5
	Metric: Manhattan	87.5	62.5
	Metric: Chebyshev	83.3	66.7

Similar to the results obtained for colorectal cancer, for gastrointestinal cancer, we observed a decrease in classification accuracy. Although the classification accuracy achieved for some MLs (SVM, polynomial regression, and logical regression) remains constant or improves, a comparison of the reduced values reveals a substantial difference between the values obtained for our proposed model and the PCA-based model (e.g., 62.5% for the PCA-based model vs. 50% for the proposed model).

3.3 Discussion on the chemical interpretation of PCs loadings and purine metabolite decomposition

The appearance of anomalous bands and artifacts is a common issue in SERS spectra, which can lead to the incorrect assignment of analytes or metabolites if they are not properly identified [48]. These bands can arise from contaminants, reagents used in substrate preparation, or byproducts of reactions and are selectively enhanced even at low concentrations due to resonance Raman scattering. Notably, specific anomalous bands, such as those from rhodamine-like species, have been observed in SERS spectra obtained with silver colloids and can interfere with the analysis, possibly being mistaken for sample-related bands [49, 50].

Taking this into account and due to the differences in classification accuracy the loading plots of the principal components (PCs) for the datasets were examined, with a particular focus on identifying any peaks associated with crystal violet, given its use as a substrate. Crystal violet exhibits distinct spectral bands in the spectral regions around 1586-1590 cm^{-1} and 1619-1622 cm^{-1} as well as a characteristic peak at 916 cm^{-1} [51], which can be used to differentiate it from other purine metabolites.

We examined the loading plots of the 19 PCs for the three datasets, which accounted for 95% of the variance in prostate cancer, 92% in colorectal cancer, and 97% in gastrointestinal cancer. Our aim was to determine whether these PCs contained informative signals or noise and if any bands from crystal violet were present. For the sake of simplicity, we present only the loading plots for the PCs where crystal violet bands were detected (PC7 for prostate cancer, PC2 and PC5 for colorectal cancer, and PC1 and PC2 for gastrointestinal cancer), as well as the SERS spectra of crystal violet with its distinct peaks at 910, 1583, and 1617 cm^{-1} (Figure 3.4).

In light of the consistent or enhanced classification accuracy achieved using the proposed model for prostate cancer, it was important to evaluate the PC spectra and determine whether the PCA-based model's classification relied primarily on purine metabolites. For the prostate cancer data set, the spectra of the PC appear to be predominantly characterized by peaks associated with purine metabolites. However, the loading plot for PC7 contains bands characteristic of crystal violet.

For colorectal and gastrointestinal cancer datasets, we found that bands attributable to purine metabolites were less frequent than those attributable to crystal violet, which appeared prominently in the spectra of PC2 and PC5 for colorectal cancer and PC1 and PC2 for gastrointestinal cancer.

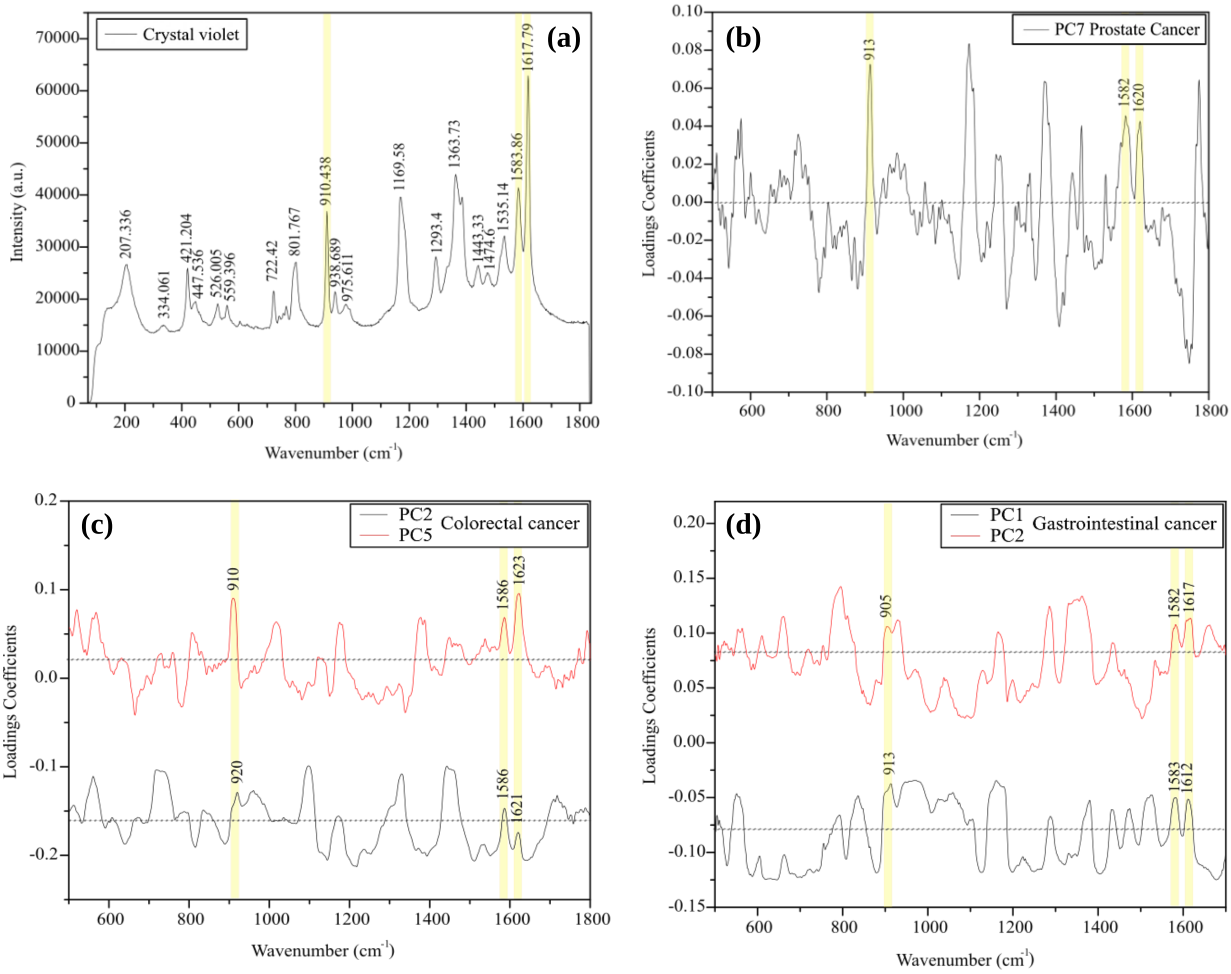

Figure 3.4: ***(a)*** *SERS spectra of crystal violet; Loading plots for* ***(b)*** *PC7 for prostate cancer,* ***(c)*** *PC2 and PC5 for colorectal cancer, and* ***(d)*** *PC1 and PC2 for gastrointestinal cancer*

While the hypoxanthine SERS spectra do present a band at 1586 cm^{-1} (see Figure 3.1) in the same region as crystal violet, if bands are also present in the 1619–1622 cm^{-1} region and around 910–915 cm^{-1}, they can be attributed to crystal violet. This observation underscores an explanation for the decrease in classification accuracy observed for the proposed model compared to the PCA-based model for colorectal and gastrointestinal cancer.

Considering these findings, we deduce that the proposed model exclusively accounts for the contribution of purine metabolites in the serum SERS spectra. In contrast, the PCA-based model incorporates bands from various contributors, including crystal violet. Consequently, the results obtained for the proposed model are more accurate as it avoids incorporating bands that are not of interest, thus mitigating the misclassification caused by sample contamination, anomalous bands and artifacts.

3.4 Residual-based model and initial data-based model classification accuracy

As a last step, we wanted to further improve the classification accuracy by investigating other contributions that may be excluded during the decomposition of the spectra and analyzing the residuals.

After obtaining the residuals for the three data sets, we visualized the mean serum SERS spectra of patients with prostate, colorectal, and gastrointestinal cancer and controls from the initial dataset alongside the mean spectra of the residuals. Figures 3.5, 3.6, and 3.7 emphasize the distinctive bands from the three data sets attributable solely to the residuals.

Using the residual datasets, we investigated their potential contributions to classification accuracy by comparing the performance of the residual-based model to the initial data-based model for the prostate, colorectal, and gastrointestinal cancer datasets.

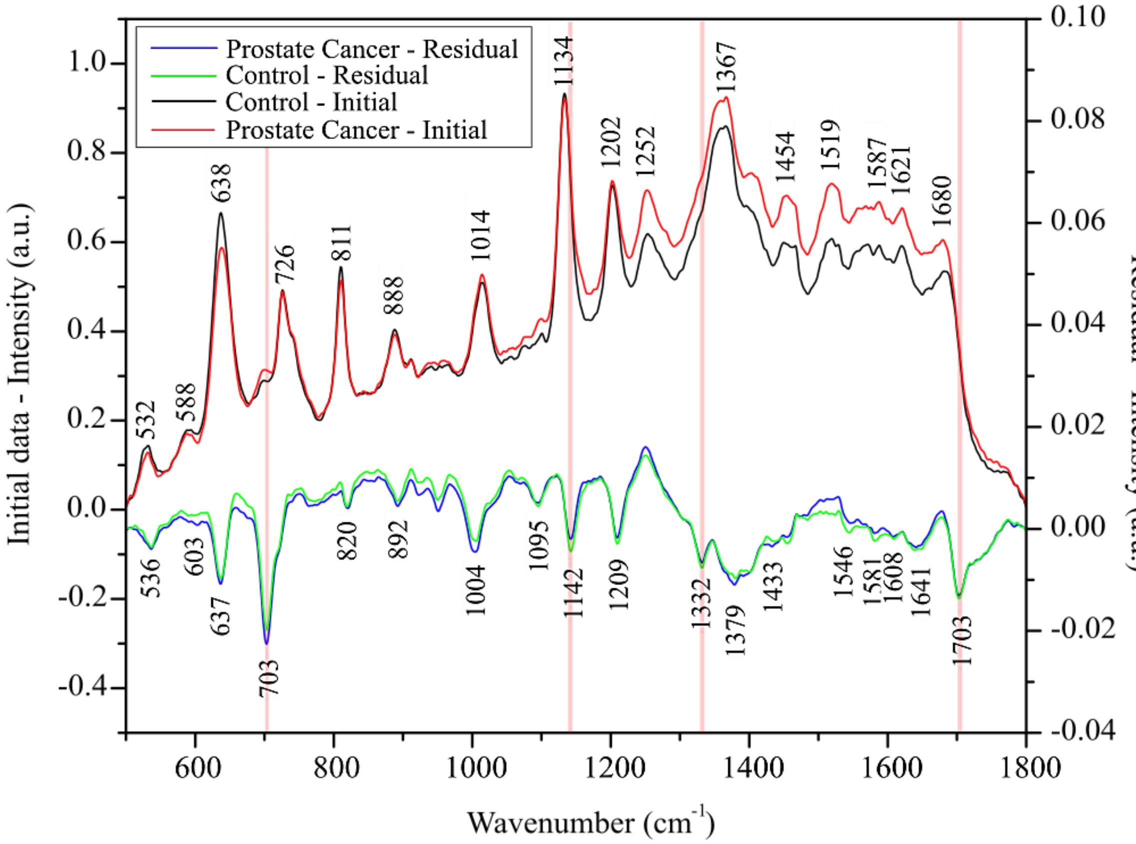

Figure 3.5: *Mean serum SERS spectra of patients with prostate cancer and controls, alongside the mean SERS spectra of the residuals*

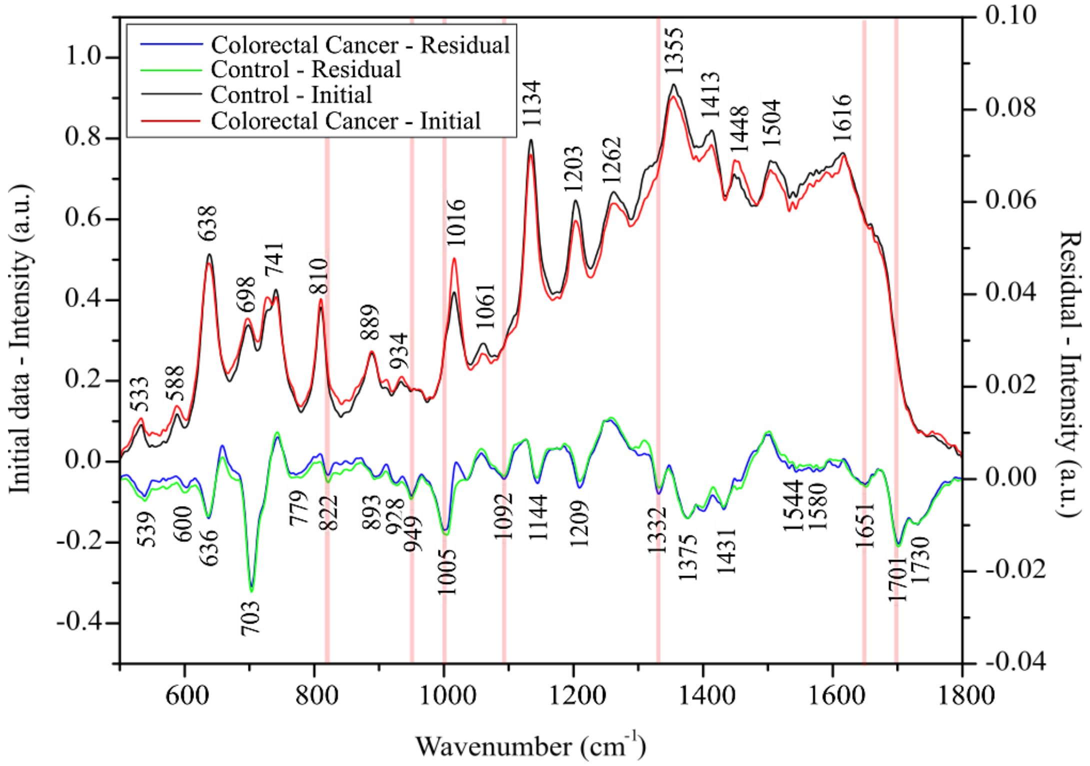

Figure 3.6: *Mean serum SERS spectra of patients with colorectal cancer and controls, alongside the mean SERS spectra of the residuals*

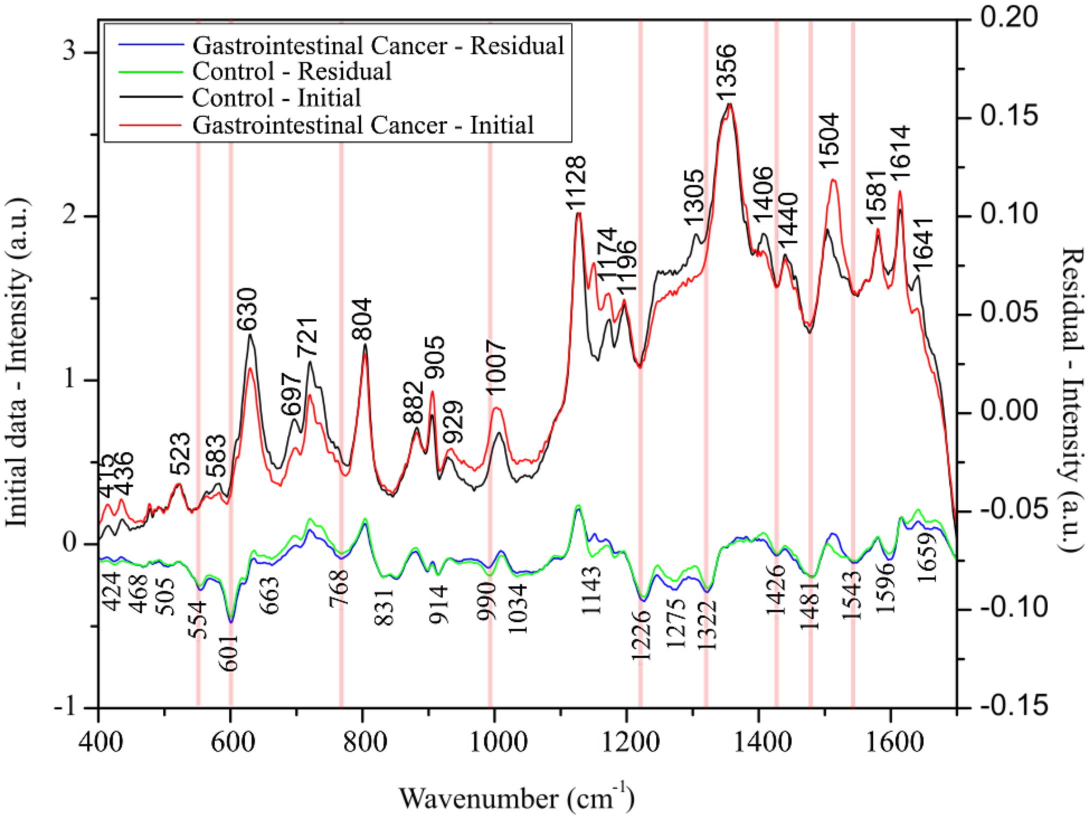

Figure 3.7: *Mean serum SERS spectra of patients with gastrointestinal cancer and controls, alongside the mean SERS spectra of the residuals*

3.4.1 Classification accuracy of classification models prostate cancer

The classification accuracy values of various MLs for prostate cancer are summarized in Table 3.7.

***Table 3.7**: Classification accuracy of the MLs for the initial data-based model and the residual-based model for prostate cancer*

ML		Initial data-based model	Residual-based model
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		61.1	55.6
SVM	Linear	55.6	61.1
	Polynomial	61.1	77.8
	RBF	61.1	66.7
	Sigmoid	72.2	66.7
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	66.7	61.1
	Lasso, BCD	66.7	66.7
	Ridge, NBCD	72.2	61.1
	Ridge, BCD	66.7	66.7
	None, NBCD	61.1	72.2
	None, BCD	55.6	66.7
AdaBoost		55.6	55.6
Random Forest		72.2	77.8
kNN	Metric: Euclidian	94.4	83.3
	Metric: Manhattan	88.9	77.8
	Metric: Chebyshev	83.3	88.9

Analyzing the results, we observe that for prostate cancer, the classification accuracy of the residuals-based model is comparable to or higher than that of the initial data-based model. Although the highest classification accuracy obtained for the residuals-based model (88.9%) is slightly lower than the highest classification accuracy obtained for the initial data-based model (94.4%), this aligns (see Table 3.4) with the proposed model's highest classification accuracy (88.9%) and is an improvement over the PCA-based model (83.3%).

3.4.2 Classification accuracy of classification models for colorectal cancer

Table 3.8 presents the classification accuracy of different ML techniques for colorectal cancer.

***Table 3.8**: Classification accuracy of the MLs for the initial data-based model and the residual-based model for colorectal cancer*

ML		Initial data-based model	Residual-based model
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		61.5	71.8
SVM	Linear	82.1	84.6
	Polynomial	84.6	89.7
	RBF	89.7	92.3
	Sigmoid	76.9	76.9
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	87.2	89.7
	Lasso, BCD	82.1	87.2
	Ridge, NBCD	92.3	89.7
	Ridge, BCD	84.6	87.2
	None, NBCD	79.5	84.6
	None, BCD	79.5	84.6
AdaBoost		79.5	69.2
Random Forest		89.7	92.3
kNN	Metric: Euclidian	89.7	92.3
	Metric: Manhattan	92.3	92.3
	Metric: Chebyshev	89.7	89.7

The residuals-based model exhibited comparable or higher classification accuracy compared to the initial data-based model, with both achieving a maximum classification accuracy of 92.3%. Moreover, these results represent an improvement (see Table 3.5) over the proposed model's highest classification accuracy (87.2%) and match the performance of the PCA-based model.

3.4.3 Classification accuracy of classification models for gastrointestinal cancer

Similarly, Table 3.9 displays the classification accuracy of different ML techniques for gastrointestinal cancer.

***Table 3.9**: Classification accuracy of the MLs for the initial data-based model and the residual-based model for gastrointestinal cancer*

ML		**Initial data-based model**	**Residual-based model**
		Classification accuracy (%)	**Classification accuracy (%)**
Naïve Bayers		41.7	41.7
SVM	Linear	79.2	79.2
	Polynomial	79.2	79.2
	RBF	70.8	79.2
	Sigmoid	87.5	62.5
Logic Regression (Regularization type: Lasso, Ridge, None) No balance class distribution (NBCD) Balance class distribution (BCD)	**Lasso**, NBCD	83.3	87.5
	Lasso, BCD	75	79.2
	Ridge, NBCD	79.2	79.2
	Ridge, BCD	75	79.2
	None, NBCD	75	83.3
	None, BCD	75	75
AdaBoost		66.7	66.7
Random Forest		79.2	83.3
kNN	Metric: Euclidian	75	75
	Metric: Manhattan	79.2	75
	Metric: Chebyshev	66.7	75

The residuals-based model demonstrated comparable or higher classification accuracy than the initial data-based model, both achieving a maximum accuracy of 87.5%. This represents an enhancement (see Table 3.6) over the proposed model's highest classification accuracy (83.3%) and matches the performance of the PCA-based model.

Conclusions

In conclusion, we developed a model to reduce the dimensionality of serum SERS spectra by focusing on known contributors such as uric acid, hypoxanthine, xanthine, urea, and creatinine. This model enhances the clinical interpretation of patient screening using SERS liquid biopsy by identifying the molecular origin of the classification. We first optimized the model to ensure that the contributions of these metabolites were uncorrelated. During testing on pure metabolite SERS spectra, we observed overlap between some purine metabolite bands. Therefore, we applied Gram-Schmidt orthogonalization to the SERS spectra of purine metabolites and used these as representative vectors for uric acid, hypoxanthine, xanthine, urea, and creatinine. Next, we decomposed each serum SERS spectrum onto these vectors to obtain scores reflecting the ‘concentration’ of each metabolite in the sample. This dimensionality reduction resulted in five variables that could be used to build machine learning classifiers for cancer screening.

To test the model, we built ML classifiers using these five variables and tested them on three datasets: one for colorectal cancer detection, one for gastrointestinal cancer detection, and one for prostate cancer detection. Since the general workflow uses PCA as a dimensionality reduction technique, we compared our model's cancer classification performance with results obtained by building the model on the first 19 PCs of each dataset. For prostate cancer, our model showed an overall improvement in classification performance. However, for colorectal and gastrointestinal cancer, we observed a decrease in classification accuracy. Analysis of the loading plot of the used PCs revealed that some PCs contained the signal of Crystal Violet (a known contaminant in SERS spectroscopy), suggesting that even though PCA showed better performance, the classification might have relied on contaminants rather than actual sample signals. In contrast, our model ensures the classification is based on molecules present in serum samples, enhancing the reliability and specificity of the results.

We then reconstructed the serum SERS spectra from the scores of each metabolite and their spectra. By subtracting the reconstructed spectrum from the initial spectrum, we identified some SERS bands not characteristic of purine metabolites. These unidentified contributors in the serum SERS spectra could aid classification. However, caution is needed as these bands could include contaminants, leading to false results.

The results of this thesis highlight the suggested model potential for improving SERS-based diagnostics. In addition to improving the accuracy and interpretability of SERS data, the spectral decomposition and dimensionality reduction techniques created pave the way for more exact and dependable cancer screening procedures.

Future research could explore several promising directions. The classification accuracy of the model might be improved by including additional metabolites and contaminants. Examining the potential for integrating sophisticated machine learning methods like deep learning models could improve the predictive capacity and resilience of the model. To evaluate the model's generalizability and clinical applicability, it would also be essential to validate it on larger and more varied clinical datasets. Investigating the mechanisms underlying the unidentified SERS bands could also unveil new biomarkers or advance knowledge of the spectral traits of different cancer types. These future directions hold the promise of further advancing the field of SERS-based diagnostics and contributing to the development of more effective cancer screening and diagnosis tools.

Acknowledgements

I want to extend my sincere appreciation to my coordinators, Prof. dr. Leopold Nicolae and Lect. Dr. Ștefania-Dana Iancu, for their outstanding guidance, constant support, and unmatched expertise during this research project. Their feedback, encouragement, and unwavering commitment were indispensable in finishing this thesis. I truly appreciate their patience, readiness to share their knowledge, and guidance, which have greatly influenced my academic path and enhanced my understanding of the subject matter. This thesis undoubtedly exists thanks to their steadfast support and priceless contributions.

Bibliography

[1] N. Rivera and I. Kaminer, "Light–matter interactions with photonic quasiparticles," *Nature Reviews Physics,* vol. 2, no. 10, p. 538–561, 2020.

[2] N. Leopold, Surface-enhanced Raman Spectroscopy: Selected Applications, Napoca Star, 2009.

[3] E. Fermi, "Quantum Theory of Radiation," *Reviews of Modern Physics,* vol. 4, no. 1, pp. 87-132, 1932.

[4] J. A. Koningstein, Introduction to the Theory of the Raman Effect, Springer Science & Business Media, 2012.

[5] M. Fleischmann, J. Robinson and P. R. Graves, "Enhanced and normal Raman scattering from pyridine adsorbed on rough and smooth silver electrodes," *Journal of electroanalytical chemistry and interfacial electrochemistry,* vol. 182, no. 1, pp. 73-85, 1985.

[6] L. Avram, S. D. Iancu, A. Stefancu, V. Moisoiu, A. Colnita, D. Marconi, V. Donca, E. Buzdugan, R. Craciun and N. Leopold, "SERS-based liquid biopsy of gastrointestinal tumors using a portable Raman device operating in a clinical environment," *Journal of Clinical Medicine,* vol. 9, no. 1, p. 212, 2020.

[7] V. Moisoiu, A. Stefancu, D. Gulei, R. Boitor, L. Magdo, L. Raduly, S. Pasca, P. Kubelac, N. Mehterov and V. Chiș, "SERS-based differential diagnosis between multiple solid malignancies: Breast, colorectal, lung, ovarian and oral cancer," *International journal of nanomedicine,* pp. 6165-6178, 2019.

[8] J. B. Phyo, A. Woo, H. J. Yu, K. Lim, B. H. Cho, H. S. Jung and M.-Y. Lee, "Label-free SERS analysis of urine using a 3D-stacked AgNW-glass fiber filter sensor for the diagnosis of pancreatic cancer and prostate cancer," *Analytical chemistry,* vol. 93, no. 8, pp. 3778-3785, 2021.

[9] Y. Zhang, X. Mi, X. Tan and R. Xiang, "Recent progress on liquid biopsy analysis using surface-enhanced Raman spectroscopy," *Theranostics,* vol. 9, no. 2, p. 491, 2019.

[10] S. Cervo, E. Mansutti, G. Del Mistro, R. Spizzo, A. Colombatti, A. Steffan, V. Sergo and A. Bonifacio, "SERS analysis of serum for detection of early and locally advanced breast cancer," *Analytical and bioanalytical chemistry,* vol. 407, pp. 7503-7509, 2015.

[11] R. El Ridi and H. Tallima, "Physiological functions and pathogenic potential of uric acid: A review," *Journal of advanced research,* vol. 8, no. 5, pp. 487-493, 2017.

[12] G. Ragab, M. Elshahaly and T. Bardin, "Gout: An old disease in new perspective–A review," *Journal of advanced research,* vol. 8, no. 5, pp. 495-511, 2017.

[13] R. J. Johnson, T. Nakagawa, D. Jalal, L. G. Sánchez-Lozada, D. H. Kang and E. Ritz, "Uric acid and chronic kidney disease: which is chasing which?," *Nephrology Dialysis Transplantation,* vol. 28, no. 9, pp. 2221-2228, 2013.

[14] M. Kanbay, T. Jensen, Y. Solak, M. Le, C. Roncal-Jimenez, C. Rivard, M. A. Lanaspa, T. Nakagawa and R. J. Johnson, "Uric acid in metabolic syndrome: from an innocent bystander to a central player," *European journal of internal medicine,* vol. 29, pp. 3-8, 2016.

[15] Z. Yu, S. Zhang, D. Wang, M. Fan, F. Gao, W. Sun, Z. Li and S. Li, "The significance of uric acid in the diagnosis and treatment of Parkinson disease: An updated systemic review," *Medicine,* vol. 96, no. 45, p. e8502, 2017.

[16] J. Massa, E. O'reilly, K. Munger, G. Delorenze and A. Ascherio, "Serum uric acid and risk of multiple sclerosis," *Journal of neurology,* vol. 256, pp. 1643-1648, 2009.

[17] Y. Long, B. Sanchez-Espiridion, M. Lin, L. White, L. Mishra, G. S. Raju, S. Kopetz, C. Eng, M. A. Hildebrandt, and D. W. Chang, "Global and targeted serum metabolic profiling of colorectal cancer progression," *Cancer,* vol. 123, no. 20, pp. 4066-4074, 2017.

[18] B. C. Yoo, S. Y. Kong, , S. G. Jang, K. H. Kim, S. A. Ahn, W. S. Park, S. Park, T. Yun and H. S. Eom, "Identification of hypoxanthine as a urine marker for non-Hodgkin lymphoma by low-mass-ion profiling," *BMC cancer,* vol. 10, pp. 1-9, 2010.

[19] N. Linder, J. Lundin, J. Isola, M. Lundin, K. O. Raivio and H. Joensuu, "Down-regulated xanthine oxidoreductase is a feature of aggressive breast cancer," *Clinical cancer research,*

vol. 11, no. 12, pp. 4372-4381, 2005.

[20] R. Moldovan, E. Vereshchagina, K. Milenko, B.-C. Iacob, A. E. Bodoki, A. Falamas, N. Tosa, C. M. Muntean, C. Farcău and E. Bodoki, "Review on combining surface-enhanced Raman spectroscopy and electrochemistry for analytical applications," *Analytica Chimica Acta,* vol. 1209, p. 339250, 2022.

[21] E. Le Ru and P. Etchegoin, Principles of Surface-Enhanced Raman Spectroscopy: and related plasmonic effects, Elsevier, 2008.

[22] S. D. Iancu, R. G. Cozan, A. Stefancu, M. David, T. Moisoiu, C. Moroz-Dubenco, A. Bajcsi, C. Chira, A. Andreica and L. F. Leopold, "SERS liquid biopsy in breast cancer. What can we learn from SERS on serum and urine?," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* vol. 273, p. 120992, 2022.

[23] A. Bonifacio, S. Dalla Marta, R. Spizzo, S. Cervo, A. Steffan, A. Colombatti and V. Sergo, "Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au nanoparticles: a systematic study," *Analytical and Bioanalytical Chemistry,* vol. 406, pp. 2355-2365, 2014.

[24] A. Bonifacio, S. Cervo and V. Sergo, "Label-free surface-enhanced Raman spectroscopy of biofluids: fundamental aspects and diagnostic applications," *Analytical and bioanalytical chemistry,* vol. 407, pp. 8265-8277, 2015.

[25] T. Rojalin, D. Antonio, A. Kulkarni and R. P. Carney, "Machine Learning-Assisted Sampling of Surfance-Enhanced Raman Scattering (SERS) Substrates Improve Data Collection Efficiency," *Applied spectroscopy,* vol. 76, no. 4, pp. 485-495, 2022.

[26] E. Alpaydin, Introduction to machine learning, MIT press, 2020.

[27] I. Goodfellow, Y. Bengio and A. Courville, Deep learning, MIT press, 2016.

[28] T. Hastie, R. Tibshirani and J. H. Friedman, , The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer, 2009.

[29] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly

Media, Inc., 2022.

[30] I. T. Jolliffe, Principal component analysis for special types of data, Springer, 2002.

[31] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences,* vol. 374, no. 2065, p. 20150202, 2016.

[32] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani and A. Hooman, "An overview of principal component analysis," *Journal of Signal and Information Processing,* vol. 4, 2020.

[33] C. D. Manning, Introduction to information retrieval, Syngress Publishing, 2008.

[34] W. S. Noble, "What is a support vector machine?," *Nature biotechnology,* vol. 24, no. 12, pp. 1565-1567, 2006.

[35] B. Mahesh, "Machine learning algorithms-a review.," *International Journal of Science and Research (IJSR),* vol. 9, no. 1, pp. 381-386, 2020.

[36] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner and R. Goodacre, "A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding," *Analytica chimica acta,* vol. 879, pp. 10-23, 2015.

[37] F. Lussier, V. Thibault, B. Charron, G. Q. Wallace and J. F. Masson, "Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering," *TrAC Trends in Analytical Chemistry,* vol. 124, p. 115796, 2020.

[38] E. W. Steyerberg, Applications of prediction models, Springer, 2009.

[39] Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences,* vol. 55, no. 1, pp. 119-139, 1997.

[40] R. E. Schapire, "The boosting approach to machine learning: An overview," *Nonlinear estimation and classification,* pp. 149-171, 2003.

[41] B. H. Menze, . M. B. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics,* vol. 10, pp. 1-16, 2009.

[42] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[43] M. Poth, G. Magill, A. Filgertshofer, O. Popp and T. Großkopf, "Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing," *Journal of Raman Spectroscopy,* vol. 53, no. 9, pp. 1580-1591, 2022.

[44] W. J. Thrift and R. Ragan, "Quantification of analyte concentration in the single molecule regime using convolutional neural networks," *Analytical chemistry,* vol. 91, no. 21, pp. 13337-13342, 2019.

[45] "Population factsheets," International Agency for Research on Cancer, [Online]. Available: https://gco.iarc.fr/today/en/fact-sheets-populations#countries. [Accessed 16 06 2024].

[46] "Colorectal cancer," World Health Organization (WHO), [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer. [Accessed 16 06 2024].

[47] M. Toplak, . S. T. Read, . C. Sandt and F. Borondics, "Quasar: Easy Machine Learning for Biospectroscopy," *Cells,* vol. 10, no. 9, p. 2300, 2021.

[48] S. Sánchez-Cortés and J. V. García-Ramos, "Anomalous Raman bands appearing in surface-enhanced Raman spectra," *Journal of Raman Spectroscopy,* vol. 29, no. 5, pp. 365-371, 1998.

[49] A. Bonifacio, S. Cervo and V. Sergo, "Label-free surface-enhanced Raman spectroscopy of biofluids: fundamental aspects and diagnostic applications," *Analytical and bioanalytical chemistry,* vol. 407, pp. 8265-8277, 2015.

[50] A. Bonifacio, S. Dalla Marta, R. Spizzo, S. Cervo, A. Steffan, A. Colombatti and V. Sergo, "Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au

nanoparticles: a systematic study," *Analytical and Bioanalytical Chemistry*, vol. 406, pp. 2355-2365, 2014.

[51] W. Meng, F. Hu, X. Jiang and L. Lu, "Preparation of silver colloids with improved uniformity and stable surface-enhanced Raman scattering," *Nanoscale research letters*, vol. 10, pp. 1-8, 2015.

DECLARAȚIE PE PROPRIE RĂSPUNDERE

Subsemnata Cimpoeșu Angela-Georgiana declar că Lucrarea de disertație pe care o voi prezenta în cadrul examenului de finalizare a studiilor la Facultatea de Fizică, din cadrul Universității Babeș-Bolyai, în sesiunea iulie 2024, sub îndrumarea Prof. dr. Leopold Nicolae și Lect. Dr. Ștefania-Dana Iancu, reprezintă o operă personală. Menționez că nu am plagiat o altă lucrare publicată, prezentată public sau un fișier postat pe Internet. Pentru realizarea lucrării am folosit exclusiv bibliografia prezentată și nu am ascuns nici o altă sursă bibliografică sau fișier electronic pe care să le fi folosit la redactarea lucrării.

Prezenta declarație este parte a lucrării și se anexează la aceasta.

Data
25.06.2024

Cimpoeșu Angela-Georgiana
Semnătură