ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

http://pubs.acs.org/journal/acsodf

Article

Power of Light: Raman Spectroscopy and Machine Learning for the Detection of Lung Cancer

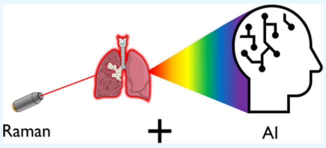
Harun Hano,* Charles H. Lawrie, Beatriz Suarez, Alfredo Paredes Lario, Ibone Elejoste Echeverría, Jenifer Gómez Mediavilla, Marina Izaskun Crespo Cruz, Eneko Lopez, and Andreas Seifert*

Cite This: *ACS Omega* 2024, 9, 14084−14091

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Lung cancer is the leading cause of cancer-related deaths worldwide, emphasizing the urgent need for reliable and efficient diagnostic methods. Conventional approaches often involve invasive procedures and can be time-consuming and costly, thereby delaying the effective treatment. The current study explores the potential of Raman spectroscopy, as a promising noninvasive technique, by analyzing human blood plasma samples from lung cancer patients and healthy controls. In a benchmark study, 16 machine learning models were evaluated by employing four strategies: the combination of dimensionality reduction with classifiers; application of feature selection prior to classification; stand-alone classifiers; and a unified predictive model. The models showed different performances due to the inherent complexity of the data, achieving accuracies from 0.77 to 0.85 and areas under the curve for receiver operating characteristics from 0.85 to 0.94. Hybrid methods incorporating dimensionality reduction and feature selection algorithms present the highest figures of merit. Nevertheless, all machine learning models deliver creditable scores and demonstrate that Raman spectroscopy represents a powerful method for future in vitro diagnostics of lung cancer.



INTRODUCTION

Early detection of diseases has become increasingly important, with lung cancer being the leading cause of cancer-related deaths worldwide.[1] Timely and accurate diagnosis of lung cancer is crucial for effective treatment and better survival rates. However, conventional methods are often expensive and time-consuming and have limited sensitivity in the early stages.[2] In contrast, Raman spectroscopy has emerged as a promising diagnostic technique that enables noninvasive, label-free, and real-time analysis.[3,4]

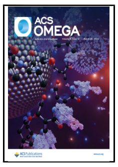
Raman spectroscopy is based on inelastic scattering of light, where a small fraction of the photons interact with the sample, resulting in a gain or loss of energy and thus a shift in the wavelength of the scattered light. This shift in wavelength is called the Raman shift and is proportional to the frequency of the molecular vibration. This highly effective and nondestructive approach can provide insight into the molecular composition of biological fluids.[5] In particular, human blood plasma, a complex biological fluid composed of proteins, lipids, nucleic acids, carbohydrates, etc., is an excellent source for identifying biochemical changes.[6] Therefore, Raman spectroscopy can be used to analyze the spectral signatures of blood plasma and provide valuable diagnostic information.[7,8] Raman spectroscopy and other vibrational spectroscopy methods have been used by several groups to investigate their capacity as new diagnostic technologies for a variety of cancers.[9,10]

This research mainly focused on the performance of several machine learning models used to discriminate the spectral signatures of human blood plasma samples between lung cancer patients and healthy controls. An ensemble of 16 different machine learning models was examined, including different combinations with a particular feature selection method, transformation techniques, and classifiers. Principal component analysis (PCA), a commonly used technique for dimensionality reduction, was applied along with a set of classifiers such as linear discriminant analysis (LDA), support vector machine (SVM), Naïve Bayes (NB), logistic regression (LR), and random forest (RF). The models were extended in conjunction with the Fisher score (FS) feature selection method in various configurations. Standalone classifiers and partial least-squares discriminant analysis (PLS-DA) were studied independently.

EXPERIMENTAL SECTION

Sample Collection. Eighteen blood samples were collected from patients with nonsmall cell lung carcinoma (NSCLC) in the Oncology Department of Hospital University Donostia (San Sebastián, Spain). Fourteen out of the 18 samples were obtained from NSCLC patients diagnosed in the advanced stage with metastasis detected in other organs. For 11 out of the 18 patients, blood collection was performed before any treatment administration. Blood samples were collected in ethylenediaminetetraacetic acid (EDTA) tubes, and plasma

Received: November 29, 2023
Revised: February 22, 2024
Accepted: February 27, 2024
Published: March 15, 2024



was prepared within 1 h of phlebotomy according to standard protocols. In addition, plasma was obtained retrospectively from 18 healthy donors from the Basque Biobank (Bioef). The samples were collected in accordance with the Declaration of Helsinki and with approval by local ethics committees (CEIC Euskadi approval number: PI2019170).

Sample Preparation and Data Collection. Prior to analysis, 1 μL of samples from a total of 36 subjects was deposited on aluminum foil (Alu-Labor-Folie) attached to the microscope slide (Superfrost Plus Adhesion Microscope Slides by Epredia) and dried for 5 min.

Aluminum foil offers high reflectivity that increases the Raman signal by the excitation laser light reflected from the sample;[11] stability that makes it a good choice for holding and stabilizing samples during analysis; flexibility regarding easy shaping or molding to fit the sample holder; a low background signal that helps minimize interference from unwanted signals during analysis;[12] and cost-effectiveness that makes it a practical option for routine Raman analysis.

During the drying process on the aluminum substrate, the typical coffee ring effect occurred due to capillary forces, leading to a higher concentration of molecular components at the periphery of the droplet. Subsequent Raman measurements targeted these rings where proteins and other components were anticipated to be the most dense. This technique, previously documented, augments Raman measurements by amplifying analyte concentration giving potentially better insight into sample properties.[13−15]

To optimize the strength of the Raman signals while minimizing damage to the sample, the 785 nm laser wavelength of the Renishaw inVia confocal Raman microscope with a grating of 1200 L/mm was selected for the analysis of the dried biological samples. Overall, the choice of 785 nm as excitation wavelength is advantageous for biological samples, as it provides sufficiently strong Raman signals while minimizing sample damage, reducing fluorescence, and providing a larger penetration depth.[16−18] The laser was operated at 73 mW output power, and the light was focused onto the sample through a 50× long distance objective. A total of 20 accumulations were performed with an exposure time of 1 s.

Data Preprocessing and Exploratory Data Analysis. First, 25 spectra were taken from each subject at different points on the periphery of the droplet. The embedded software of the commercial Raman system removes cosmic rays, which no longer affects further data processing methods such as background (baseline) reduction or averaging. Two preprocessing methods were employed: asymmetric Whittaker baseline correction and standard normal variate (SNV) transformation. Baseline correction was employed to address baseline drifts and distortions in spectral data and to enhance the accuracy and reliability of the quantitative analysis by handling asymmetric features and noise.[19,20] Besides, SNV transformation removed multiplicative baseline variations due to sample thickness, scattering, instrumental response, etc., without altering the shape of the spectra.[21,22] Then, 25 spectra belonging to one subject were averaged to create a single representative spectrum per subject. This step is essential for reducing random noise, improving signal-to-noise ratio, and capturing the overall spectral signature.[23]

Various machine learning models were used for data analysis, all of which are suitable for handling complex data sets. The first five models integrated PCA with classifiers such as LDA, SVM, NB, LR, and RF. In these models, PCA reduces the dimensionality while preserving as much variance as possible. Another set of five models used PCA, however, prior to applying the classifiers as before, the Fisher score feature selection method was applied considering class labels. This method helps to select the most discriminative features for classification tasks. A third set of models solely relies on the previously stated classifiers without a PCA or Fisher score. This makes it possible to examine the performance of the classifiers on the raw data set directly after preprocessing. Finally, PLS-DA is employed as a stand-alone model for handling multicollinearity in data by incorporating class labels directly into the model fitting process. The benefits of using those machine learning models include their ability to handle high-dimensional data sets and to identify the most relevant features for classification.

RESULTS AND DISCUSSION

Comparative Spectral Analysis. Averaged spectra of lung cancer and healthy control were analyzed, as illustrated in Figure 1, as well as three discrete Raman shift regions in Figure S1A,C in the Supporting Information, allowing a comprehensive study of specific molecular vibrations. These three regions allow a more targeted analysis of specific vibrations and corresponding molecular alterations associated with lung cancer, further supporting the utility of Raman spectroscopy as a diagnostic technique.

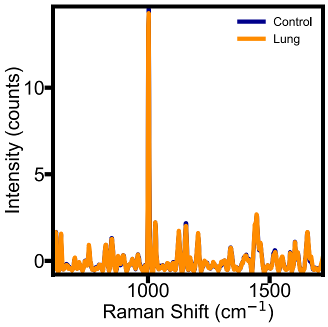


Figure 1. Averaged Raman spectra per class.

Figure S1A emphasizes the spectral range of 610−990 cm^{-1} of the Raman shift, which corresponds to ring vibrations, ring breathings, and skeletal stretching of chemical groups such as tryptophan, tyrosine, or nucleic acids. The range between 990 and 1016 cm^{-1}, corresponding to the symmetric ring breathing mode of phenylalanine, was intentionally excluded. Due to its pronounced and sharp intensity, its presence can overshadow and suppress other relevant peaks in the selected regions. Figure S1B focuses on the vibrational frequency range of 1016−1360 cm^{-1}, emphasizing the occurrence of distinct vibrational stretching and bending modes as well as deformation modes and the amide III region. Here, spectral features provide information about the secondary structure of proteins and conformational changes in nucleic acids, which are also essential for understanding the molecular alterations associated with lung cancer. Figure S1C targets the range of 1360−1720 cm^{-1} which encompasses mostly C═C stretching and the amide I region offering insights into the secondary structure of proteins, such as α-helices.

Comprehensive Visualization of Selected Components and Variables. Figure 2 embodies an in-depth representation of the multivariate structure within the data

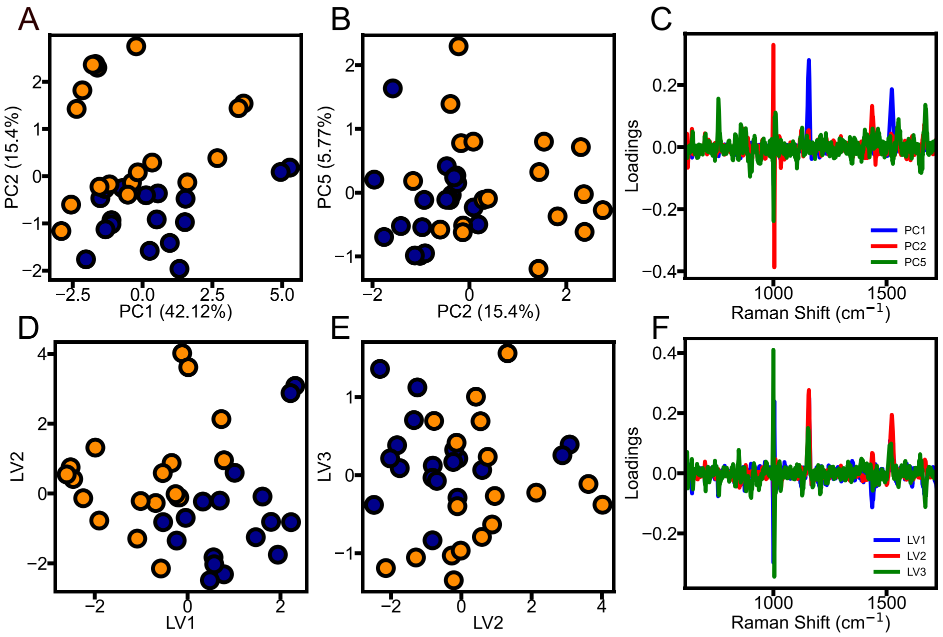


Figure 2. 2D visualization of the discrimination between lung cancer patients (orange) and healthy controls (blue). (A) Score plots of PC1−PC2, (B) PC2−PC5, (D) LV1−LV2, and (E) LV2−LV3. (C) Corresponding loadings of PC1, PC2, and PC5, and (F) loadings of LV1, LV2, and LV3.

set while providing interconnected insights into PC and latent variable (LV) spaces.

Figure 2A,B depicts score plots derived from the first and second PCs (PC1−PC2) and the second and fifth PCs (PC2−PC5), respectively. They accounted for 42.12, 15.4, and 5.77% of the total variance. The first PC (PC1) captured the highest amount of variation, with subsequent components (PC2, PC5) explaining the remaining variance in a decreasing order. The presented choice of PCs goes beyond the traditional approach. A high Fisher score associated with PC2 solidifies its role as a dominant axis, achieving a clear separation between lung cancer and the control group. PC3 and PC4, although not initially considered, produced results of 11.84 and 8.67%, respectively. These components are categorized as less significant according to the Fisher Score, or in other words less significant for classification.

Interestingly, our analysis also flagged PC5, a subsequent component, based on its elevated score. Thus, PC2 and PC5 were selected using a robust, data-driven approach that provides intriguing insight into the potential multivariate structure of lung cancer. Furthermore, Figure 2D,E portrays a score plot for LV1−LV2 and LV2−LV3, respectively, determined through PLS-DA. The purpose of PLS-DA is to find the multivariate relationship between a data set (X) and response variables (y). Here, LVs are linear combinations of predictors that explain the maximum covariance with the response variable and thus enable efficient classification.

The loading plots shown in Figure 2C for PC1, PC2, and PC5 and in Figure 2F for LV1, LV2, and LV3 show the influence of each original variable on the derived characteristic features. Loadings are essentially the coefficients or correlations between original features and selected components or variables. They indicate how much each feature contributes to or detracts from selected features, offering insights into spectral signatures.[24] On the other hand, the regression vector (RV) for three LVs shown in Figure S2 in the Supporting Information offers a numerical representation of the degree and direction of influence that each LV has on the dependent variable. This condensed information on variable interplay accurately reflects the impact of each LV on the model outcome.

Table 1 elucidates the relevant features by detailing peak assignments from loading plots of PCs and the RV of LVs. It indicates the molecular groups responsible for each dominant peak that contributes to the observed differences. Relevant features are PC1, PC2, PC5, and RV, associated with vibrational modes detected in Raman spectra. For instance, the vibrational mode associated with the C−C twisting mode of phenylalanine is detected by all four features, which indicates its importance and high variance in the data set. Conversely, C−C stretching mode backbone (α—helix conformation) and C═C stretching mode of tyrosine are uniquely captured by PC5, indicating that it represents a feature with lower variance. Moreover, the inclusion of regression coefficients, as opposed to the loadings of the PCs, is driven by our objective to quantitatively assess the relative influence of each spectral feature on the distinction between healthy and lung cancer samples. Regression coefficients reflect the influence of each feature on classification, with their absolute values indicating the strength of distinction, regardless of whether they are positive or negative. The instances of "none" signify the absence of certain vibrational modes in RV. This indicates that these specific modes do not significantly contribute to or are not detected in the differentiation process captured by the RV.

Performance of the Models. The evaluation of the models focuses on two main aspects: (1) accuracy for quantifying the proportion of correct predictions made by the model relative to the total number of input samples and (2) receiver operating characteristic (ROC) curve for visualizing and measuring a trade-off between true positive rate (sensitivity) and false positive rate (1-specificity).

A comprehensive evaluation strategy was employed with four distinct approaches to assess the performance of five machine learning classifiers LDA, SVM, NB, LR, and RF alongside a separate evaluation for PLS-DA. Four specific approaches were executed: the first incorporated PCA for data

Table 1. Raman Spectral Band Assignments for Human Blood Plasma as Reported in the Literature

peak positions (cm^{-1})[a]	vibrational modes	PCs	regression coefficients
619−624	C−C twisting mode of phenylalanine	PC1, PC2, PC5	−0.021
641−643	C−C twisting mode of tyrosine	PC1, PC5	none
698−701	*n*(C−S) *trans*(amino acid methionine)	PC2, PC5	0.031
756−758	symmetric ring breathing of tryptophan	PC1, PC2, PC5	0.051
822	out of plane ring breathing tyrosine	PC5	−0.021
855−856	ring breathing mode of tyrosine	PC1, PC5	none
874−878	arginine	PC2	0.041
897−901	monosaccharides (b-glucose), (C−O−C) skeletal mode	PC2, PC5	none
939	C−C stretching mode backbone *α*—helix	PC5	none
1000−1004	symmetric ring breathing mode of phenylalanine	PC1, PC2, PC5	−0.067
1029−1033	C−H in-plane bending mode of phenylalanine	PC1, PC2	−0.017
1104	C−C vibration mode of the gauche-bonded chain	PC5	−0.017
1123−1127	proteins; C−C phospholipids stretching	PC2, PC5	0.031
1156−1157	C−C/C−N stretching mode	PC1, PC2	−0.024
1204−1210	tryptophan and phenylalanine *n*(C−C_6H_5) mode	PC1, PC5	0.029
1232−1269	amide III	PC5	0.023, 0.014
1397−1404	glutathione	PC5	−0.025
1436−1438	C−H deformation	PC2, PC5	0.068
1513−1528	carotenoids (C=C)	PC1, PC2, PC5	−0.042
1548−1553	tryptophan	PC5	0.028
1587−1589	C=C stretching	PC2, PC5	−0.022
1604−1606	C=C stretching mode of phenylalanine and tryptophan	PC2	−0.016
1619	C=C stretching mode of tyrosine and tryptophan	PC5	none
1666−1671	amide I: *α*—helix	PC2, PC5	0.066

[a] Peak positions are reported concerning the following features: PC1, PC2, and PC5, and the RV of the first three LVs LV1, LV2, and LV3.[27−29]

reduction; the second combined PCA with Fisher score to select the most prominent PCs; the third focused solely on classifiers; and the fourth was devoted to PLS-DA. The data set was split into training and test sets with an 85:15 ratio. Hyperparameter tuning was done using RandomizedSearchCV, with 6-fold cross-validation on the training set. This was executed across 20 iterations to find the best hyperparameters. After these optimal settings were confirmed, each classifier was iteratively tested 100 times to further scrutinize model stability and performance. Cross-validation helps assess the performance and generalization ability of the models by minimizing the risk of overfitting or underfitting. By evaluating the model on unseen data in each iteration, cross-validation can provide a reliable estimate of how likely the model is to perform well on new, unseen samples.[25,26]

PCA + Classifiers. This approach is based on an iterative assessment that sequentially incorporates the first 10 PCs. Models were evaluated on five classifiers, which are LDA, SVM, NB, LR, and RF, using the output of PCA as input variables.

The results presented in Figure 3A and Table S1 show that the optimal range for the number of PCs to be considered is between 6 and 8. This range effectively captures the most significant variance within the data set, thereby contributing to the predictive power of the models. Model performance showed no significant improvement beyond these selected numbers, reinforcing the fact that subsequent PCs contain less information that is critical for classification.

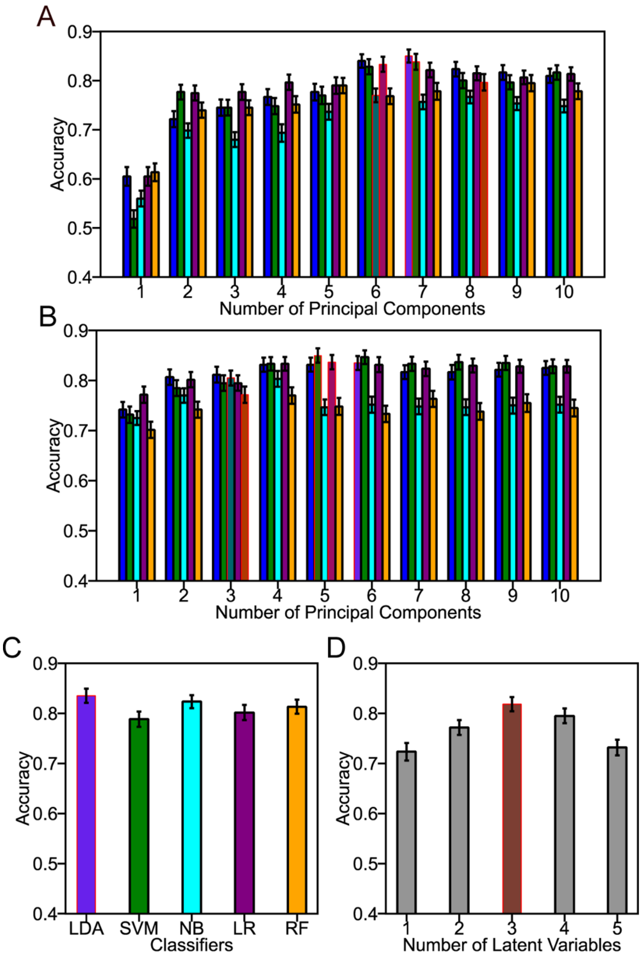


Figure 3. Accuracy of comparative classification performance. Hatched bars indicate the highest accuracy presented in each graph per classifier. The error bars reflect the standard errors. (A) PCA with the first 10 PCs as inputs for various classifiers: LDA (blue), SVM (green), NB (cyan), LR (purple), and RF (orange). (B) Same procedure as in (A), but here with selected PCs by Fisher score feature selection before applying the classifiers. (C) Only classifiers without dimensionality reduction. (D) The first 5 selected LVs.

Based on the observations, PCA + LDA and PCA + SVM maintain their edge, delivering mean accuracies of 0.85 ± 0.13 and 0.84 ± 0.16, respectively, at 7 PCs. This suggests that LDA and SVM are particularly proficient at distinguishing between classes in a feature space constrained to the first 7 PCs. The prevalence of linearly separable features within the data set is evident. LDA and SVM, which excel under such conditions, therefore perform notably well. Conversely, PCA + NB returns a lower mean accuracy of 0.77 ± 0.14 with 6 PCs, while PCA +

RF scores 0.80 ± 0.17 at 8 PCs. These results imply that RF may necessitate a slightly higher dimensionality, i.e., more PCs, for a more diverse feature set to build its ensemble of decision trees effectively. Besides, these algorithms either require more complex features or are not as effective in a reduced feature space. PCA + LR performs competitively with a mean accuracy of 0.83 ± 0.15 at 6 PCs, illustrating its capability but still slightly trailing behind PCA + LDA and PCA + SVM.

PCA + Fisher Score + Classifiers. In contrast to the previous section, here, PCA was implemented with the maximum number of components. Fisher's score was subsequently applied to rank the relevance of features. The most significant features were selected across iterations, creating a cumulative list of the important features. Then, the top 10 most frequently recurring features were selected for further model evaluation. Notably, even as the feature list was extended to encompass these 10 components, the model consistently exhibited high accuracy. This aligns with the expectation that the Fisher score effectively identifies the most discriminative features, thereby enabling the achievement of stable and notable accuracy with even a limited set of components.

Highlighting the results compiled in Figure 3B and Table S1, PCA + FS + LDA and PCA + FS + SVM achieve their peak performance at 6 and 5 PCs, respectively, with mean accuracies of 0.84 ± 0.14 and 0.85 ± 0.14. Intriguingly, SVM maintains a near-identical performance with fewer PCs compared to the PCA-only approach, hinting at the model's resilience to the reduction in dimensionality. LDA, however, experiences a minor decrement in performance, suggesting that the extra features isolated by FS may create a slightly more complex decision boundary.

It is noteworthy that the performance of PCA + FS + NB and PCA + FS + RF accomplishes mean accuracies of 0.81 ± 0.15 and 0.77 ± 0.16, while utilizing only 3 PCs. Naïve Bayes appears to benefit from Fisher score feature selection more than it did with just PCA. This could indicate that FS succeeds in isolating features that encapsulate class-discriminative information on NB more effectively.

Conversely, the effectiveness of RF decreases, possibly indicating that RF as an ensemble method requires a higher level of feature complexity than the top 3 PCs can provide through FS. PCA + FS + LR, in contrast, sustains its performance, securing a mean accuracy of 0.84 ± 0.14 with 5 PCs. This consistency indicates its robustness and adaptability to feature spaces curated by both the PCA and the FS.

In summary, the integration of the Fisher score as a feature ranking has varying degrees of impact on the classifier performance. While SVM and LR exhibit stability or slight improvement, LDA, NB, and RF demonstrate nuanced shifts in performance in this feature selection method. The observations accentuate the utility of the Fisher score when paired with PCA in optimizing classifier performance, particularly when feature relevance is not uniformly distributed across the dimensions.

Only Classifiers. In contrast to the prior approaches that utilized PCA and PCA + FS for dimensionality reduction, this section bypasses data transformation techniques to evaluate classifiers in the original feature space. This approach offers a more straightforward and unfiltered assessment of the performance of each classifier. Direct application of the classifiers to high-dimensional data sets provided insightful results. As indicated in Figure 3C and Table S1, LDA emerged as the top performer with a mean accuracy of 0.84 ± 0.14, demonstrating its robust handling of high-dimensional data. Surprisingly, NB, which is often considered a simple classifier, also performed admirably, achieving 0.82 ± 0.13.

On the other hand, SVM and LR, both relying on finding optimal hyperplanes for classification, recorded slightly lower accuracies of 0.79 ± 0.15 and 0.80 ± 0.15, respectively, suggesting potential challenges when dealing with high-dimensional data, especially without the assistance of any feature selection or extraction techniques. RF, an ensemble method, demonstrated solid performance, as expected, given its aptitude for high-dimensional data, yielding a value of 0.81 ± 0.14.

The findings point out that certain classifiers, attributable to their inherent algorithmic properties, can exhibit robust performance, even in high-dimensional spaces, without the aid of dimensionality reduction techniques. This can be particularly valuable if it is desirable to retain the original characteristics for the sake of interpretability or other analytical considerations.

PLS-DA. Unlike PCA, PLS-DA considers class labels directly during the extraction of LVs. The optimum number of LVs, as shown in Figure 3D, was chosen with respect to the model accuracy.

The performance of PLS-DA was commendable, achieving a mean accuracy of 0.82 ± 0.14 using only three components. This result highlights the effectiveness of PLS-DA in utilizing class-specific information for classification, making it a potent tool in high-dimensional data analysis. It also emphasizes the efficiency of class-guided dimensionality reduction techniques as they can produce more class-relevant features leading to improved classifier performance.

ROC Curve. In conjunction with accuracy scores, ROC curves and their corresponding area under the curve (AUC) scores offer comprehensive performance metrics. ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system when its discrimination threshold is varied. AUC score, ranging from 0 to 1, serves as a comprehensive measure of classification performance; an AUC score closer to 1 indicates a better classification performance.

As depicted in Figure 4A, models combining PCA with various classifiers show considerable variations in their AUC values. Notably, PCA + LDA and PCA + LR exhibit remarkable AUC scores of 0.93 and 0.92, respectively, thereby highlighting their excellent discrimination power. In comparison, PCA + SVM performs commendably but slightly trails behind with an AUC of 0.91. PCA + NB and PCA + RF register lower AUC values of 0.90 and 0.89, hinting at their lower efficiency in balancing the sensitivity and specificity.

In Figure 4B, model scores incorporating PCA, FS and classifiers present more consistent performances, ranging from 0.85 to 0.94. Interestingly, the AUC score of PCA + FS + RF stands at 0.85, which is comparatively lower than those of other classifiers like PCA + FS + LDA and PCA + FS + SVM, which have AUC scores of 0.94. However, it is crucial to note that PCA + FS + RF accomplishes this with only three PCs, indicating a level of efficiency in capturing the essential characteristics of the data. Moreover, its accuracy of 0.77 ± 0.16 is quite respectable and adds another layer to its value as a classifier. This indicates that although it does not outperform other classifiers in terms of AUC, it is still a competitive, resource-efficient alternative that maintains a high degree of accuracy.

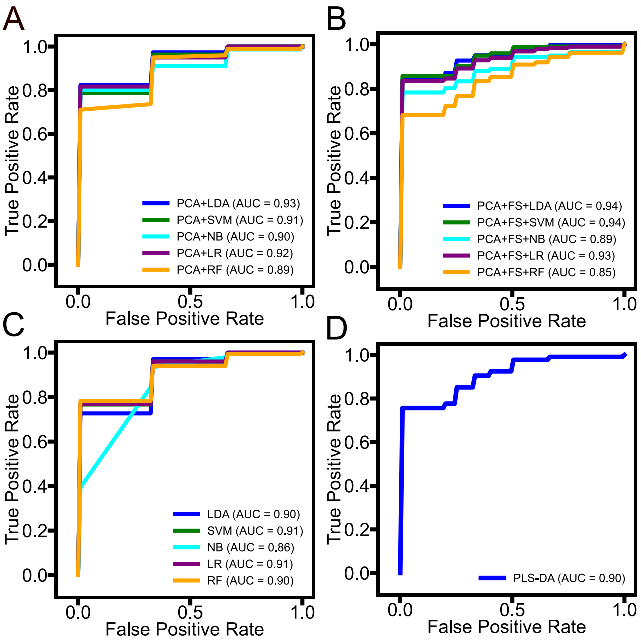


Figure 4. ROC curves based on various classification strategies. (A) PCs with classifiers. (B) Selected PCs by Fisher score with classifiers. (C) Only classifiers. (D) Selected LVs from PLS.

Without feature extraction or selection, standalone classifiers, as illustrated in Figure 4C, exhibit robust AUC scores. LDA, LR, and RF achieve AUC scores between 0.90 and 0.91, testifying to their inherent strengths in managing high-dimensional data spaces. Although NB lags slightly with an AUC score of 0.86, it still represents a commendable performance given the complexity of the data set. SVM displays a respectable AUC score of 0.91 but suggests room for potential optimization. The analysis continues in Figure 4D with PLS-DA, a distinct method that incorporates class labels into the feature extraction process. AUC value of 0.90 affirms its effective implementation of class-specific information for achieving high classification performance. This suggests that the inherent features of PLS-DA, which take class labels into account when generating LVs, allow for more precise identification of true positives and true negatives.

In conclusion, a holistic evaluation of model performance, integrating accuracy and AUC scores from ROC curves, reveals distinct patterns in model efficacy. Specifically, models such as PCA + LDA, PCA + FS + SVM, LDA alone, and PLS-DA consistently demonstrate superior performance, while NB and RF show enhanced results with the application of feature selection techniques. These insights are crucial for choosing appropriate models for particular tasks and data types, leading to more precise and reliable predictions. It is important to note, however, that these conclusions are intrinsically linked to the unique structure of our data set and may not be directly transferable to other data sets or applications.

■ CONCLUSIONS

This comprehensive study reaffirms the potential of Raman spectroscopy as a promising tool for lung cancer detection. By comparison of the Raman spectra of lung cancer patients and healthy controls, significant differences in spectral features were identified, highlighting the considerable potential to provide insights into the molecular alterations associated with lung cancer. For identifying these changes and elucidating compositional and structural modifications that occur in proteins, carbohydrates, lipids, nucleic acids, and other biomolecules, it turns out that the entire spectral range of the Raman spectra from human blood plasma is important.

In the presented analysis, PCA + LDA and PCA + FS + SVM are leading in terms of accuracy, both falling within 0.85 ± 0.14 and featuring AUC scores above 0.93. LDA stands its ground with similar performance metrics, even without feature extraction methods. PLS-DA, although slightly behind in accuracy, holds a respectable AUC score of 0.90, signaling its reliability. Among standalone classifiers, NB distinguishes itself with a competitive accuracy of 0.82 ± 0.13. Overall, the findings indicate that while PCA-enhanced models offer the highest accuracy and AUC scores, simpler models like LDA and PLS-DA remain robust choices depending on the specific requirements of a given application. However, it turns out that the inner structure of our data is very robust with respect to different machine learning algorithms applied to Raman spectra from dried blood plasma samples. Generally, the inner structure and the intraclass and interclass variability of the presented data set offer flexibility and freedom concerning the choice of machine learning strategies.

In summary, this study highlights the potential of Raman spectroscopy as a diagnostic tool for lung cancer detection and emphasizes the benefits of employing machine learning models to analyze spectral data for classification purposes. Furthermore, it highlights the role of model selection and the importance of multivariate analysis methods in attaining superior performance. It was shown that different models could be optimally applied based on the specific needs of the task, leading to more accurate and effective diagnostic tools, which could lead to earlier detection, improved treatment, and better patient outcomes. Using Raman spectroscopy data supported by artificial intelligence offers a rapid and low-cost technology for in vitro diagnostics. Once the model is validated and calibrated for specific disease patterns, the proposed technology can replace complex chemical analyses and, in addition to classifying the disease, provide detailed insight into biochemical changes in physiology in real time. The technology is not limited to lung cancer and therefore has the potential for a paradigm shift in medical diagnostics. With the potential to revolutionize cancer diagnosis, these findings are a significant step forward in medical research, offering new hope to millions of people worldwide.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c09537.

Selected spectral regions of interest revealing subtle differences between lung cancer patients and healthy controls (Figure S1); RV of the first three LVs (Figure S2); Comparative assessment of classification performance in terms of accuracy (Table S1) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Harun Hano − *CIC nanoGUNE BRTA, 20018 San Sebastián, Spain; Department of Physics, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain;* orcid.org/0000-0001-8559-8563; Email: h.hano@nanogune.eu

Andreas Seifert − *CIC nanoGUNE BRTA, 20018 San Sebastián, Spain; IKERBASQUE—Basque Foundation for Science, 48009 Bilbao, Spain;* orcid.org/0000-0001-5849-4953; Phone: +34 943574045; Email: a.seifert@nanogune.eu

Authors

Charles H. Lawrie − *IKERBASQUE—Basque Foundation for Science, 48009 Bilbao, Spain; Biogipuzkoa Health Research Institute, 20014 San Sebastián, Spain; Sino-Swiss Institute of Advanced Technology (SSIAT), University of Shanghai, 201800 Shanghai, China; Radcliffe Department of Medicine, University of Oxford, OX3 9DU Oxford, U.K.*

Beatriz Suarez − *Faculty of Nursing and Medicine, University of the Basque Country (UPV/EHU), 20014 San Sebastián, Spain; Biogipuzkoa Health Research Institute, 20014 San Sebastián, Spain*

Alfredo Paredes Lario − *Servicio de Oncología Médica, Hospital Universitario Donostia, 20014 San Sebastián, Spain*

Ibone Elejoste Echeverría − *Servicio de Oncología Médica, Hospital Universitario Donostia, 20014 San Sebastián, Spain*

Jenifer Gómez Mediavilla − *Servicio de Oncología Médica, Hospital Universitario Donostia, 20014 San Sebastián, Spain*

Marina Izaskun Crespo Cruz − *Servicio de Oncología Médica, Hospital Universitario Donostia, 20014 San Sebastián, Spain*

Eneko Lopez − *CIC nanoGUNE BRTA, 20018 San Sebastián, Spain; Department of Physics, University of the Basque Country (UPV/EHU), 20018 San Sebastián, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c09537

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financially supported by grant CEX2020-001038-M funded by MICIU/AEI/10.13039/501100011033; further financial support by the grant for the Requalification of Doctor Staff at the UPV/EHU (Code: MARSA22/38) financed by the Spanish Ministry of Universities and the European Union with the Next Generation EU funds.

■ REFERENCES

(1) Leblond, F.; Dallaire, F.; Tran, T.; Yadav, R.; Aubertin, K.; Goudie, E.; Romeo, P.; Kent, C.; Leduc, C.; Liberman, M. Subsecond Lung Cancer Detection within a Heterogeneous Background of Normal and Benign Tissue Using Single-Point Raman Spectroscopy. *J. Biomed. Opt.* **2023**, *28*, 90501.

(2) Field, J. K.; Oudkerk, M.; Pedersen, J. H.; Duffy, S. W. Prospects for Population Screening and Diagnosis of Lung Cancer. *Lancet* **2013**, *382*, 732−741.

(3) Olaetxea, I.; Valero, A.; Lopez, E.; Lafuente, H.; Izeta, A.; Jaunarena, I.; Seifert, A. Machine Learning-Assisted Raman Spectroscopy for PH and Lactate Sensing in Body Fluids. *Anal. Chem.* **2020**, *92*, 13888−13895.

(4) Zheng, Q.; Li, J.; Yang, L.; Zheng, B.; Wang, J.; Lv, N.; Luo, J.; Martin, F. L.; Liu, D.; He, J. Raman Spectroscopy as a Potential Diagnostic Tool to Analyse Biochemical Alterations in Lung Cancer. *Analyst* **2020**, *145*, 385−392.

(5) Cialla-May, D.; Krafft, C.; Rösch, P.; Deckert-Gaudig, T.; Frosch, T.; Jahn, I. J.; Pahlow, S.; Stiebing, C.; Meyer-Zedler, T.; Bocklitz, T.; Schie, I.; Deckert, V.; Popp, J. Raman Spectroscopy and Imaging in Bioanalytics. *Anal. Chem.* **2022**, *94*, 86−119.

(6) Sole, C.; Arnaiz, E.; Manterola, L.; Otaegui, D.; Lawrie, C. H. The Circulating Transcriptome as a Source of Cancer Liquid Biopsy Biomarkers. *Semin. Cancer Biol.* **2019**, *58*, 100−108.

(7) Kuhar, N.; Sil, S.; Verma, T.; Umapathy, S. Challenges in Application of Raman Spectroscopy to Biology and Materials. *RSC Adv.* **2018**, *8*, 25888−25908.

(8) Chen, C.; Wu, W.; Chen, C.; Chen, F.; Dong, X.; Ma, M.; Yan, Z.; Lv, X.; Ma, Y.; Zhu, M. Rapid Diagnosis of Lung Cancer and Glioma Based on Serum Raman Spectroscopy Combined with Deep Learning. *J. Raman Spectrosc.* **2021**, *52*, 1798−1809.

(9) Zhang, S.; Qi, Y.; Tan, S. P. H.; Bi, R.; Olivo, M. Molecular Fingerprint Detection Using Raman and Infrared Spectroscopy Technologies for Cancer Detection: A Progress Review. *Biosensors* **2023**, *13*, 557.

(10) Liu, Y.; Chen, C.; Tian, X.; Zuo, E.; Cheng, Z.; Su, Y.; Chang, C.; Li, M.; Chen, C.; Lv, X. A Prospective Study: Advances in Chaotic Characteristics of Serum Raman Spectroscopy in the Field of Assisted Diagnosis of Disease. *Expert Syst. Appl.* **2024**, *238*, 121787.

(11) Aitekenov, S.; Sultangaziyev, A.; Boranova, A.; Dyussupova, A.; Ilyas, A.; Gaipov, A.; Bukasov, R. SERS for Detection of Proteinuria: A Comparison of Gold, Silver, Al Tape, and Silicon Substrates for Identification of Elevated Protein Concentration in Urine. *Sensors* **2023**, *23*, 1605.

(12) Cui, L.; Butler, H. J.; Martin-Hirsch, P. L.; Martin, F. L. Aluminium Foil as a Potential Substrate for ATR-FTIR, Transflection FTIR or Raman Spectrochemical Analysis of Biological Specimens. *Anal. Methods* **2016**, *8*, 481−487.

(13) Filik, J.; Stone, N. Analysis of Human Tear Fluid by Raman Spectroscopy. *Anal. Chim. Acta* **2008**, *616*, 177−184.

(14) Chen, R.; Zhang, L.; Zang, D.; Shen, W. Blood Drop Patterns: Formation and Applications. *Adv. Colloid Interface Sci.* **2016**, *231*, 1−14.

(15) Barman, I.; Dingari, N. C.; Kang, J. W.; Horowitz, G. L.; Dasari, R. R.; Feld, M. S. Raman Spectroscopy-Based Sensitive and Specific Detection of Glycated Hemoglobin. *Anal. Chem.* **2012**, *84*, 2474−2482.

(16) Synytsya, A.; Judexova, M.; Hoskovec, D.; Miskovicova, M.; Petruzelka, L. Raman Spectroscopy at Different Excitation Wavelengths (1064, 785 and 532 Nm) as a Tool for Diagnosis of Colon Cancer. *J. Raman Spectrosc.* **2014**, *45*, 903−911.

(17) Kerr, L. T.; Byrne, H. J.; Hennelly, B. M. Optimal Choice of Sample Substrate and Laser Wavelength for Raman Spectroscopic Analysis of Biological Specimen. *Anal. Methods* **2015**, *7*, 5041−5052.

(18) Bonnier, F.; Ali, S. M.; Knief, P.; Lambkin, H.; Flynn, K.; McDonagh, V.; Healy, C.; Lee, T. C.; Lyng, F. M.; Byrne, H. J. Analysis of Human Skin Tissue by Raman Microspectroscopy: Dealing with the Background. *Vib. Spectrosc.* **2012**, *61*, 124−132.

(19) Eilers, P. H. C. A Perfect Smoother. *Anal. Chem.* **2003**, *75*, 3631−3636.

(20) Baek, S.-J.; Park, A.; Ahn, Y.-J.; Choo, J. Baseline Correction Using Asymmetrically Reweighted Penalized Least Squares Smoothing. *Analyst* **2015**, *140*, 250−257.

(21) Rinnan, Å.; Berg, F. v. d.; Engelsen, S. B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC, Trends Anal. Chem.* **2009**, *28*, 1201−1222.

(22) Afseth, N. K.; Segtnan, V. H.; Wold, J. P. Raman Spectra of Biological Samples: A Study of Preprocessing Methods. *Appl. Spectrosc.* **2006**, *60*, 1358−1367.

(23) Blake, N.; Gaifulina, R.; Griffin, L. D.; Bell, I. M.; Thomas, G. M. H. Machine Learning of Raman Spectroscopy Data for Classifying Cancers: A Review of the Recent Literature. *Diagnostics* **2022**, *12*, 1491.

(24) Bro, R.; Smilde, A. K. Principal Component Analysis. *Anal. Methods* **2014**, *6*, 2812−2831.

(25) Lever, J.; Krzywinski, M.; Altman, N. Points of Significance: Model Selection and Overfitting. *Nat. Methods* **2016**, *13*, 703−704.

(26) Lopez, E.; Etxebarria-Elezgarai, J.; Amigo, J. M.; Seifert, A. The Importance of Choosing a Proper Validation Strategy in Predictive

Models. A Tutorial with Real Examples. *Anal. Chim. Acta* **2023**, *1275*, 341532.

(27) Poon, K. W. C.; Lyng, F. M.; Knief, P.; Howe, O.; Meade, A. D.; Curtin, J. F.; Byrne, H. J.; Vaughan, J. Quantitative Reagent-Free Detection of Fibrinogen Levels in Human Blood Plasma Using Raman Spectroscopy. *Analyst* **2012**, *137*, 1807−1814.

(28) Nargis, H. F.; Nawaz, H.; Ditta, A.; Mahmood, T.; Majeed, M. I.; Rashid, N.; Muddassar, M.; Bhatti, H. N.; Saleem, M.; Jilani, K.; Bonnier, F.; Byrne, H. J. Raman Spectroscopy of Blood Plasma Samples from Breast Cancer Patients at Different Stages. *Spectrochim. Acta, Part A* **2019**, *222*, 117210.

(29) Carota, A.; Campanella, B.; del carratore, R.; Bongioanni, P.; Giannelli, R.; Legnaioli, S. Raman Spectroscopy and Multivariate Analysis as Potential Tool to Follow Alzheimer's Disease Progression. *Anal. Bioanal. Chem.* **2022**, *414*, 4667−4675.